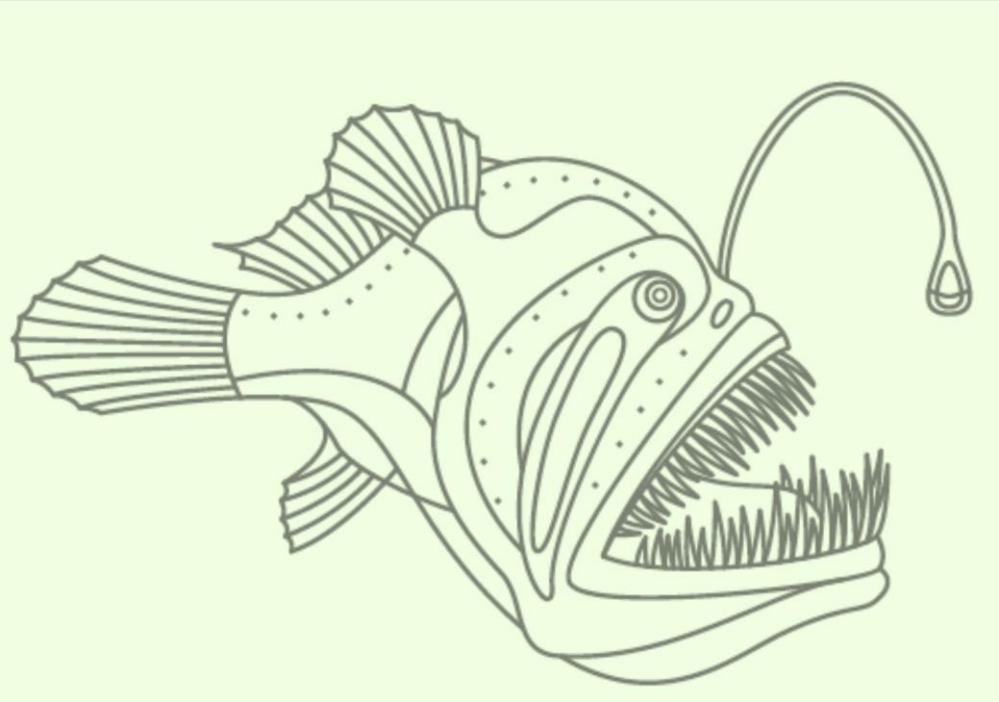
არჩილ ფრანგიშვილი, ოლეგ ნამიჩეიშვილი, ჟუჟუნა გოგიაშვილი

მანქანური სწავლების შესავალი

ლექციების კურსი



თბილისი
2024

საქართველოს ტექნიკური უნივერსიტეტი

არჩილ ფრანგიშვილი, ოლეგ ნამიჩეიშვილი,
ჟუჟუნა გოგიაშვილი

მანქანური სწავლების შესავალი

(ლექციების კურსი)

დამტკიცებულია:
სტუ-ს „IT კონსალტინგის სამეც-
ნიერო ცენტრის“ სარედაქციო
კოლეგიის მიერ, ოქმი N5 2.02.24

თბილისი
2024

უაკ 004.85

განხილულია მანქანური სწავლების ისეთი საკითხები, როგორიცაა სწავლება მასწავლებლით და მასწავლებლის გარეშე, მოდელის არჩევა და შეფასება, ბაიესური დანასკვი, პარამეტრული რეგრესიები, რეგულარიზაცია, ხელოვნური ნეირონული ქსელები, უახლოესი მეზობლების მეთოდი, ხეები და ტყეები, საყრდენი ვექტორების მანქანები და ბირთვული მეთოდები, განზომილების რედუქცია, კლასტერიზაცია, ამოზნექილი ოპტიმიზაციის პრინციპები, წერტილოვანი შეფასების კონცეფციები და სწავლება განმტკიცებით

გადმოცემული მასალა დაეხმარება მკითხველებს განსაზღვრონ პრობლემები, რომლებიც შეიძლება იყოს გადაჭრილი მანქანური სწავლების მიდგომით, განახორციელონ ამოცანათა ფორმალიზება, დაადგინონ თითოეული ამოცანისათვის ყველაზე უფრო შესატყვისი ალგორითმები, განახორციელონ მათი რეალიზება, დაბოლოს, ეზიარონ მიღებული შედეგების შეფასების მეთოდებს.

ლექციათა ამ კურსის კონცეფციები ილუსტრირებულია მრავალრიცხოვანი მაგალითებითა და ამოხსნილი სავარჯიშოებით. წიგნი განკუთვნილია ინფორმატიკისა და გამოყენებითი მათემატიკის, ასევე საინჟინრო მიმართულებათა სფეროს სტუდენტებისათვის ბაკალავრიატისა და მაგისტრატურის საფეხურებზე.

პროფ. ვახტანგ კვარაცხელიას რედაქციით

რეცენზენტები: პროფ. მარიამ ჩხაიძე,
პროფ. მაია გოგიაშვილი

რედკოლეგია:

ა. ფრანგიშვილი (თავმჯდომარე), ზ. აზმაიფარაშვილი, მ. ახობაძე, გ. გოგიჩაიშვილი, ზ. ბოსიკაშვილი, ე. თურქია, ლ. იმნაიშვილი, რ. კაკუბავა, ვ. კვარაცხელია, ნ. ლომინაძე, ჰ. მელაძე, ლ. პეტრიაშვილი, გ. სურგულაძე (რედაქტორი), ბ. შანშიაშვილი, ო. შონია, ზ. წვერაიძე

© სტუ-ს „IT კონსალტინგის სამეცნიერო ცენტრი", 2024

ISBN 978-9941-8-6373-8

ყველა უფლება დაცულია. ამ წიგნის ნებისმიერი ნაწილის (ტექსტი, ფოტო, ილუსტრაცია თუ სხვა) გამოყენება არც ერთი (ელექტრონული თუ მექანიკური) ფორმითა და საშუალებით არ შეიძლება გამომცემლის წერილობითი ნებართვის გარეშე. საავტორო უფლებების დარღვევა ისჯება კანონით.

წინასიტყვა

მანქანური სწავლება საფუძვლად უდევს მეცნიერებას მონაცემებზე და ხელოვნურ ინტელექტს. როცა კი წამოიჭრება საუბარი, კომპანიათა ციფრულ ტრანსფორმაციაზე, დიდ მონაცემებზე (big Data), ეროვნულ ან ევროპულ სტრატეგიაზე, მანქანური სწავლება აუცილებლად აღმოჩნდება ხოლმე ყურადღების ცენტრში. მისი გამოყენებები მრავალრიცხოვანია და მრავალფეროვანი, დაწყებული საძიებელი სისტემებით ან სახეთა ამოცნობით და დამთავრებული გენომური კვლევებით, სოციალური ქსელების ანალიზით, მიზნობრივი რეკლამით, კომპიუტერული ხედვით ავტომატური თარგმნით და ვაჭრობით ინტერნეტის საფუძველზე.

სტატისტიკისა და ინფორმატიკისა მიჯნაზე მანქანური სწავლება დაკავშირებულია მონაცემთა მოდელირებასთან. ამ სფეროს ძირითადი პრინციპები გაჩნდა სიხშირული ან ბაიესური სტატისტიკიდან, ხელოვნური ინტელექტიდან ან სიგნალების თეორიიდან. ლექციათა ამ კურსში ის აზრია გატარებული, რომ მანქანური სწავლება — ეს მეცნიერებაა მანქანურ სწავლებაზე, პროგნოსტიკული ფუნქცია იმ დაკვირვებათა ნაკრებიდან, რომლებიც ეხება მონიშნულ ან მოუნიშნავ მონაცემებს.

წინამდებარე კურსი იმ კონცეფციისა და ალგორითმების შესავალია, რომლებიც მანქანურ სწავლების საფუძველს წარმოადგენს. იგი თავაზობს სტუდენტებს ხედვას, რომელიც ფოკუსირებულია ემპირიული რისკის მინიმიზაციაზე პროგნოზირების ფუნქციათა მოცემულ კლასთან მიმართებაში.

პედაგოგიური მიზნები: ამ კურსის მიზანია - სტუდენტების თანხლება მანქანური სწავლების აღმოჩენისკენ მიმავალ გზაზე და აუცილებელი ინსტრუმენტების მიწოდებისას იმისათვის, რომ :

1. განისაზღვროს პრობლემები, რომლებიც შეიძლება გადაიჭრას მანქანური სწავლების მიდგომებით;
2. ჩატარდეს ამ პრობლემათა ფორმალიზება მანქანური სწავლების თვალსაზრისით;
3. დადგინდეს ყველაზე უფრო შესაბამისი კლასიკური ალგორითმები ამ ამოცანებისათვის და მოხდეს მათი რეალიზება;
4. განხორციელდეს დამოუკიდებლად ეს ალგორითმები მათ არსში და თავისებურებებში ჩასაწვდომად;
5. შეფასდეს და შედარდეს ყველაზე უფრო ობიექტური ხერხით მანქანური სწავლების რამდენიმე ალგორითმის ხარისხის მახასიათებელი კონკრეტული გამოყენებისათვის.

მიზნობრივი აუდიტორია: ლექციათა მოცემული კურსი განკუთვნილია სტუდენტებისათვის ინფორმატიკის და გამოყენებითი მათემატიკის მიმართულებებით ბაკალავრიატის გამოსაშვები ან მაგისტრატურის პირველი კურსის დონეზე, ასევე საინჟინრო მიმართულებათა სტუდენტებისათვის ამავე საფეხურებზე, რომლებსაც სურთ მანქანურ სწავლებაში გამოყენებული მთავარი ალგორითმების საფუძვლების გაგება. კურსი ეყრდნობა ჩვენს პედაგოგიურ გამოცდილებას და გულისხმობს დაშვების ისეთ წინაპირობებს, როგორიცაა :

- წრფივი ალგებრა (მატრიცათა ინვერსია, სპექტრული თეორემა, მატრიცის საკუთარი მნიშვნელობების და საკუთარი ვექტორების თეორია);
- ალბათობათა თეორიის ზოგიერთი ცნება (შემთხვევითი სიდიდე, განაწილებები, ბაიესის თეორემა).

კურსის მოკლე აღწერა: ეს კურსი იწყება მანქანური სწავლების და სხვადასხვა ტიპის იმ პრობლემის მიმოხილვით, რომლის გადაჭრა მას შეუძლია. მასში ნაჩვენებია, როგორ შეიძლება ეს ამოცანები მათემატიკურად იყოს ჩამოყალიბებული ოპტიმიზაციის ამოცანათა ფორმით (ლექცია 1), ხოლო დასკვნით ლექციებში საფუძველი ეყრება ამოზნექილი ოპტიმიზაციის საფუძვლებს, რაც აუცილებელია წარმოდგენილი ალგორითმების გაგებისათვის. ამ კურსის მნიშვნელოვანი ნაწილი ეძღვნება კონტროლირებადი სწავლების პრობლემებს. მე-2 ლექციაში დაწვრილებით არის აღწერილი მათი ფორმულირება და შემოტანილია ჰიპოთეზათა სივრცის, რისკის და დანაკარგის ცნებები, ასევე მოცემულია მათი ზოგიერთი განზოგადება. კონტროლირებადი სწავლების ყველაზე კლასიკური და ხშირად გამოყენებული ალგორითმების განხილვამდე მნიშვნელოვანია იმის გაგება, როგორ ხდება მოდელის შეფასება მონაცემთა ნაკრებზე და საუკეთესო მოდელის არჩევა რამდენიმე შესაძლებლობიდან, რასაც მე-3 ლექცია ეთმობა.

ამ ეტაპზე სავსებით გამართლებულია კონტროლირებადი პროგნოსტიკული მოდელების აგების საკითხის განხილვა. კურსში ჯერ განიხილება პარამეტრული მოდელები, რომლებშიც მონაცემთა განაწილების მამოდელირებელ ან მაპროგნოზირებელ ფუნქციას აქვს ცხადი ანალიზური სახე. ამას საფუძველი ეყრება დანასკვის ბაიესური ელემენტებით მე-4 ლექციაში, რომლებიც შემდეგ გამოყენებული აღმოჩნდება მე-5 ლექციაში კონტროლირებადი სწავლების პარამეტრულ მოდელებში. მომდევნო მე-6 ლექციაში წარმოდგენილია ამ ალგორითმების რეგულარიზებული ვარიანტები. ამის შემდეგ მე-7 ლექცია ხელოვნური ნეირონული ქსელების შესახებ აშუქებს საკითხს გაცილებით უფრო რთული პარამეტრული მოდელების აგების შესახებ და ღრმა სწავლების საფუძვლებისადმი მიმართვის აუცილებლობასაც ხაზგასმით აღნიშნავს.

შემდეგ კურსში განიხილება არაპარამეტრული მოდელები და ანალიზი იწყება ამ მიდგომებიდან ერთ-ერთი ყველაზე ინტუიციური მიდგომის - უახლოესი მეზობლების მეთოდის - აღწერით (ლექცია 8). მომდევნო მასალაში მოცემულია გადაწყვეტილებათა ხეზე დაფუძნებული მიდგომები, ხოლო შემდეგ წარმოდგენილია მეთოდების ის ნაკრები, რომელშიც ასახულია კონტროლირებადი მანქანური სწავლების ორი - ამჟამად ყველაზე მძლავრი - ალგორითმი : შემთხვევითი ტყეები და გრადიენტული ბუსტინგი (ლექცია 9). შეიძლება ითქვას, რომ გრადიენტული ბუსტინგი — ეს მანქანური სწავლების მეთოდია კლასიფიკაციისა და რეგრესიის ამოცანებისთვის. იგი აგებს წინასწარმეტყველების მოდელს სუსტი მაპროგნოზირებელი მოდელების ანსამბლის, ჩვეულებრივ, გადაწყვეტილებათა ხეების, ფორმით. მე-10 ლექცია ბირთვული მეთოდების შესახებ იყენებს საყრდენი ვექტორების მანქანათა კონცეფციას. იგი აღწერს არაწრფივი მოდელების აგებას წრფივი მოდელების საშუალებით მონაცემთა ხელახალი (განმეორებითი) აღწერის სივრცეში.

რაც შეეხება მე-11 ლექციას, აქ წარმოდგენილია განზომილების რედუქცია (შეკვეცა), როგორც კონტროლირებადი, ასევე არაკონტროლირებადი, ხოლო მე-12 ლექციაში განხილულია

არაკონტროლირებადი სწავლების ერთ-ერთი უმნიშვნელოვანესი პრობლემა : კლასტერიზაცია.

მე-13 ლექციაში განიხილება ისეთი საკითხები, როგორიცაა ამოზნექილობა და ოპტიმიზაციის პრობლემა, ამოზნექილი ოპტიმიზაცია შეზღუდვათა გარეშე და ამოზნექილი ოპტიმიზაცია შეზღუდვათა პირობებში. მე-14 ლექცია ეძღვნება წერტილოვანი შეფასების უმნიშვნელოვანეს კონცეფციებს, ხოლო დასკვნით, მე-15 ლექციაში დაწვრილებით არის გადმოცემული მანქანური სწავლება განმტკიცებით, ანუ წახალისებით — ალბათ, ყველაზე საინტერესო და დიდი იმედების მომცემი მიმართულება ხელოვნურ ინტელექტში.

პრაქტიკულად ყველა ლექცია შეიცავს მაგალითებსა და ამოხსნილ სავარჯიშოებს, დამატებით ინფორმაციას რჩევებით შემდგომი წინსვლისათვის და ბიბლიოგრაფიას.

როგორ უნდა წავიკითხოთ ლექციათა ეს კურსი: წინამდებარე კურსი შედგენილია თანამიმდევრული შესწავლისათვის. მაგრამ პირველი სამი ლექციის შემდეგ მომდევნო ლექციათა თანამიმდევრობა შეიძლება მკითხველის საჭიროებას დაექვემდებაროს. გამონაკლისს შეადგენს მხოლოდ მე-6 ლექცია, რომელიც ჩაფიქრებულია როგორც წინა მე-5 ლექციის ლოგიკური გაგრძელება. როგორც წესი, აუცილებლობის შემთხვევაში მკითხველს შეხვდება სათანადო მითითებები ლექციათა კურსის გარკვეულ განყოფილებებზე.

მადლობის გზავნილი:

ამ ნაშრომის მომზადებაში მრავალ პირს აქვს მიღებული მონაწილეობა ინტელექტუალური, მეცნიერული, ფინანსური თუ ტექნიკური თვალსაზრისით. მათი მეგობრული განწყობა და უანგარო დახმარება უდიდესი პატივია ჩვენთვის.

განსაკუთრებულ მადლიერებას გამოვხატავთ ამ სალექციო კურსის რედაქტორის საქართველოს ტექნიკური უნივერსიტეტის ნიკო მუსხელიშვილის სახელობის გამოთვლითი მათემატიკის ინსტიტუტის დირექტორის პროფესორ ვახტანგ კვარაცხელიას, ასევე რეცენზენტების პროფესორ მარიამ ჩხაიძის და პროფესორ მაია გოგიაშვილის მიმართ, რომელთა შენიშვნების საფუძველზე შესაძლებელი გახდა ნაშრომის საბოლოო სტრუქტურისა და შინაარსობრივი კონცეფციის ჩამოყალიბება ძალიან მაღალი სტატუსის მქონე ფრანგული სახელმძღვანელოს გათვალისწინებით : «Introduction au Machine Learning», 2^e édition, Dunod, 2022, 264 pages. იგი ეკუთვნის პარიზის უმაღლესი სამთო სკოლის გამოთვლითი ბიოლოგიის ცენტრის, კიურის ინსტიტუტის და ჯანმრთელობისა და სამედიცინო კვლევათა საფრანგეთის ეროვნული ინსტიტუტის პროფესორის ქლოე-აგათე აზენკოტის (Chloé-Agathe Azencott) კალამს.

სასიამოვნო მოვალეობად მიგვაჩნია გულწრფელი მადლობა გადავუხადოთ ასევე ყველა ჩვენს კოლეგას იმ ფასდაუდებელი შრომისათვის, რომელმაც არსებითად შეუწყო ხელი კურსის შინაარსისა და წარმოდგენის ფორმის გაუმჯობესებას.

სარჩევი

ლექცია 1 მანქანური სწავლების გაცნობა

შინაარსი

მანქანური სწავლება — ძალზე საინტერესო სფეროა. მრავალი დისციპლინიდან გამომდინარე, როგორიცაა სტატისტიკა, ოპტიმიზაცია, ალგორითმიკა ან სიგნალების დამუშავება, იგი კვლევათა მუდამ ცვალებადი არეა, რომელსაც დღეს წარმოუდგენლად დიდი როლი ენიჭება საზოგადოებაში. უკვე რამდენიმე ათწლეულია იგი გამოიყენება სიმბოლოთა ავტომატურ ამოცნობაში ან სპამ-ფილტრებში, ახლა კი გამოიყენება საბანკო თაღლითობისგან თავის დასაცავად, რეკომენდაციათა გასაწევად ჩვენს გემოვნებასთან მისადაგებულ წიგნებთან, ფილმებთან ან მრავალ სხვა პროდუქტთან დაკავშირებით, პირთა საიდენტიფიკაციოდ ჩვენი კამერის სამიზნებელში ან ტექსტების ერთი ენიდან მეორეზე სათარგმნელად ავტომატურ რეჟიმში.

უახლოეს წლებში მანქანური სწავლება, ალბათ, საშუალებას მოგვცემს მნიშვნელოვნად გავზარდოთ :

- საგზაო მოძრაობის უსაფრთხოება (სხვა ხერხთან ერთად, ავტონომიური სატრანსპორტო საშუალებების მასობრივად გამოყენებითაც) ;
- რეაგირება საგანგებო სიტუაციებზე, სტიქიურ უბედურებებზე ;
- ახალი სამკურნალო საშუალებების წარმოება ;
- ჩვენი შენობებისა და სამრეწველო დარგების ენერგოეფექტურობა.

ამ ლექციის დანიშნულება ისაა, რომ უფრო მკაფიოდ დავადგინოთ, რა მიეკუთვნება მანქანურ სწავლებას და რა არ მიეკუთვნება მას, ასევე მივუთითოთ ამ სფეროს ის შტოები, რომლებიც განხილვის საგანი გახდება ლექციათა ამ კურსში.

მიზნები

- მანქანური სწავლების განსაზღვრების ჩამოყალიბება
- მანქანური სწავლებისადმი გარკვეული პრობლემის მიკუთვნების საკითხის გარკვევა

— მანქანური სწავლების ამოცანათა ძირითადი კლასების კონკრეტული მაგალითების მოცემა.

1 რა არის მანქანური სწავლება?

რა არის სწავლება, როგორ ვასწავლით და რას ნიშნავს ეს მანქანისთვის? სწავლების საკითხი იტაცებს კომპიუტერულ მეცნიერებათა სპეციალისტებს და მათემატიკოსებს არანაკლებ, ვიდრე ნევროლოგებს, პედაგოგებს, ფილოსოფოსებს და მხატვრებს.

განსაზღვრება, რომელიც შეიძლება იყოს გამოყენებული, როგორც კომპიუტერული პროგრამის, ასევე რობოტის, შინაური ცხოველის ან ადამიანის მიმართ, მოცემული აქვს, მაგალითად, ფაბიენ ბენუროს : «*სწავლება არის ქცევის ცვლილება გამოცდილების საფუძველზე*».

კომპიუტერული პროგრამის შემთხვევაში, რომელიც ჩვენ ლექციათა ამ კურსში გვაინტერესებს, საუბარია *ავტომატურ სწავლებაზე, ხელოვნურ სწავლებაზე*, ანუ *მანქანურ სწავლებაზე* (ინგლისურად *machine learning*), როცა ამ პროგრამას აქვს უნარი ისწავლოს ქცევის ცვლილების ყოველგვარი ცხადი დაპროგრამების გარეშე. ეს არის განსაზღვრება, რომელიც ჯერ კიდევ 1959 წელს ჩამოაყალიბა არტურ სამუელმა (Arthur Samuel). ამრიგად, ჩვენ შეგვიძლია დავუპირისპიროთ ერთმანეთს *კლასიკური* პროგრამა, რომელიც იყენებს შესასვლელზე მიღებულ პროცედურას და მონაცემებს პასუხებისათვის გამოსასვლელზე და *მანქანური სწავლების* პროგრამა, რომელიც იყენებს მონაცემებს და პასუხებს პროცედურის შესაქმნელად, რომლითაც პასუხები ყალიბდება მონაცემებიდან.

მაგალითი

დავუშვათ, რომ კომპანიას სურს კლიენტის მიერ დახარჯული საერთო თანხის გაგება ამ კლიენტის ანგარიშ-ფაქტურებიდან. ამისათვის საკმარისია კლასიკური ალგორითმის, ე.ი. უბრალო შეკრების, განხორციელება : მანქანური სწავლების ალგორითმის გამოყენება აუცილებელი არ არის.

ახლა კი დავუშვათ, რომ ამ ანგარიშ-ფაქტურების გამოყენება უნდათ იმის დასადგენად, რა პროდუქტებს იყიდის, ყველაზე სავარაუდოდ, კლიენტი ერთი თვის განმავლობაში. თუმცა ეს ფაქტორები შეიძლება მართლაც იყოს კავშირში ერთმანეთთან, ჩვენ, ცხადია, არ გაგვაჩნია ამის გასაკეთებლად საჭირო მთელი ინფორმაცია. მაგრამ, თუ ჩვენ გვაქვს დიდი რაოდენობის ადამიანთა შენაძენების ისტორია, შესაძლებელი ხდება მანქანური სწავლების ალგორითმის გამოყენება პროგნოზის მოდელის მისაღებად პასუხისათვის დასმულ შეკითხვაზე.

1.1 რისთვის გამოიყენება მანქანური სწავლება ?

მანქანური სწავლება შეიძლება ემსახურებოდეს იმ პრობლემათა გადაჭრას, როცა :

— არ იციან, როგორ ამოხსნან (საყიდლების წინასწარმეტყველების შესახებ ზემოთ დასმული ამოცანის მსგავსად) ;

— იციან, როგორ ამოხსნან, მაგრამ არ იციან, როგორ განახორციელონ მათი ფორმალიზება ალგორითმიკის ტერმინებში (ასეთი მდგომარეობაა, მაგალითად, სახეთა ამოცნობის ან ბუნებრივი ენის გაგების ამოცანებში) ;

— იციან, როგორ ამოხსნან, მაგრამ კომპიუტერული რესურსებისადმი ძალიან მომთხოვნი პროცედურებით (ასეთი მდგომარეობაა, მაგალითად, დიდი მოლეკულების ურთიერთქმედებათა წინასწარმეტყველებისას, ვინაიდან სიძნელეები მოდელირების პროცესში პრაქტიკულად გადაულახავია).

ამიტომ მანქანური სწავლება გამოიყენება იმ შემთხვევაში, როცა *მონაცემთა* რაოდენობა შედარებით დიდია, მაგრამ *ცოდნა* საკმაოდ მიუწვდომელია ან განუვითარებელი.

ამრიგად, მანქანურ სწავლებას შეუძლია ასევე დახმარება გაუწიოს ადამიანს სწავლებაში : მანქანური სწავლების ალგორითმებით შექმნილ მოდელებს შეუძლია გარკვეული ინფორმაციის შედარებითი მნიშვნელობის ან სხვა ინფორმაციასთან თანამოქმედების ჩვენება კონკრეტული პრობლემის გადასაჭრელად. საყიდლების პროგნოზირებასთან დაკავშირებულ მაგალითში მოდელის გაგებას შეუძლია საშუალება მოგვცეს გავაანალიზოთ წინა შეძენათა ის მახასიათებლები, რომლებიც ახორციელებს მომავალი საყიდლების წინასწარმეტყველებას. მანქანური სწავლების ეს მხარე ფართოდ გამოიყენება სამეცნიერო კვლევებში : რომელი გენები და როგორ მონაწილეობს სიმსივნეთა გარკვეული ტიპის განვითარებაში ? ტვინის გამოსახულების რომელი არეები ახორციელებს ქცევის წინასწარმეტყველებას ? მოლეკულათა რომელი მახასიათებლები ანიჭებს მათ სამკურნალო თვისებებს კონკრეტული ჩვენებების-თვის ? ტელესკოპით დანაკვირვები სურათის რომელი ასპექტები იძლევა კონკრეტული ასტრონომიული ობიექტის იდენტიფიკაციის შესაძლებლობას ?

მანქანური სწავლების შედგენილობა:

მანქანური სწავლება ორ მთავარ საყრდენზეა აგებული. ესენია :

— ერთი მხრივ, მონაცემები იმ მაგალითების სახით, რომლებზეც ალგორითმი განახორციელებს სწავლებას ;

— და მეორეც, სწავლების ალგორითმი, რომელიც წარმოადგენს ამ მონაცემებზე განსახორციელებელ პროცედურას მოდელის შესაქმნელად. მონაცემთა კრებულზე სწავლების ალგორითმის ამოქმედებას სწავლება ეწოდება.

ეს ორი საყრდენი ერთნაირად მნიშვნელოვანია. ერთი მხრივ, სწავლების არც ერთ ალგორითმს არ შეუძლია კარგი მოდელის შექმნა არარელევანტური მონაცემების გამოყენებით - ეს არის ნაგვის კონცეფცია (*GIGO - garbage in, garbage out*) მოქმედებაში : «*ნაგავი შესასვლელზე → ნა-გავი გამოსასვლელზე*». დასკვნა ცალსახაა : მანქანური სწავლების ალგორითმს, რომელსაც უხარისხო მონაცემები მიეწოდება, ასეთივე დაბალი ხარისხის პროგნოზების გარდა, ვერაფერს მოახერხებს. მეორე მხრივ, არაადეკვატური ალგორითმის საშუალებით შესაბამის მონაცემებზე ნასწავლ მოდელს, კარგი ხარისხი არ ექნება.

ლექციათა ეს კურსი ეძღვნება მანქანური სწავლების ორი საყრდენის ხსენებული წყვილიდან მხოლოდ ერთ-ერთს - სწავლების ალგორითმებს. მაგრამ მხედველობიდან არ უნდა გავუშვათ ის გარემოება, რომ მონაცემთა შეგროვებისა და ანალიზის გაერთიანებისას პროგნოზულ ანალიზთან ამ სინთეზური დარგის სპეციალისტის (*machine learner*), ანუ, მოკლედ, მონაცემთა მკვლევარის (*data scientist*), საქმიანობის მნიშვნელოვანი ნაწილია საინჟინრო მუშაობა და მონაცემების მომზადება : აბერანტული (საერთო სურათიდან უკიდურესად ამოვარდნილი,

ძალზე გადახრილი) ერთეულების ამოსაღებად, მონაცემების არასაკმარისობის სამართავად, რელევანტური წარმოდგენის ასარჩევად და ა.შ.

ყურადღება

მიუხედავად იმისა, რომ ხშირად ეს ორი ცნება მიღებულია ერთისა და იმავე სახელით მოინათლოს, მნიშვნელოვანია განვასხვავოთ *მანქანური სწავლების ალგორითმი* და *დამსწავლელი მოდელი* : პირველი იყენებს მონაცემებს მეორის შესაქმნელად, რომელიც შემდეგ შეიძლება გამოვიყენოთ როგორც ჩვეულებრივი კლასიკური პროგრამა.

ამრიგად, სწავლების ალგორითმი მოვლენის მოდელირების საშუალებას იძლევა მაგალითების საფუძველზე. ითვლება, რომ აქ აუცილებელია მიზნის განსაზღვრა და მისი ოპტიმიზება. მაგალითად, საქმე შეიძლება შეეხოს სასწავლო მაგალითებზე მოდელის მიერ დაშვებული შეცდომების რაოდენობის მინიმიზებას. ლექციათა ამ კურსში სწორედ ასეთ ფორმაშია წარმოდგენილი ძირითადი კლასიკური და განსაკუთრებით პოპულარული ალგორითმები.

მაგალითი

განვიხილოთ რამდენიმე მაგალითი, რომელშიც ხდება მანქანური სწავლების პრობლემათა ხელმეორედ ჩამოყალიბება ოპტიმიზაციის ამოცანის სახით. ლექციათა ამ კურსის დარჩენილ ნაწილში კი სტუდენტები უფრო დაწვრილებით და ღრმად გაეცნობიან აქ ძალიან თავისუფლად ჩამოყალიბებულ საინტერესო პრობლემათა მკაცრ მათემატიკურ ფორმალიზაციას.

— ონლაინ-გამყიდველს შეიძლება მოუნდეს კლიენტთა რეპრეზენტატული (დამახასიათებელი) ტიპების *მოდელირება* წარსული ტრანზაქციების საფუძველზე ერთსა და იმავე ტიპს მიკუთვნებულ პირთა სიახლოვის *მაქსიმიზაციით* ;

— საავტომობილო კომპანიამ შეიძლება ეცადოს ავტომობილის მოძრაობის ტრაექტორიის *მოდელირება* გარემოში ავტომობილების ვიდეოჩანაწერების საფუძველზე ავარიათა რაოდენობის მინიმიზაციით :

— მკვლევარ-გენეტიკოსებს შეიძლება მოუნდეთ მუტაციის გავლენის *მოდელირება* პაციენტთა მონაცემებზე დაყრდნობით დაავადების შესახებ და თავისი მოდელის მაქსიმალურად შეთანხმებით თანამედროვე ცოდნასთან ;

— ბანკმა შეიძლება მოისურვოს რისკის ქცევის *მოდელირება* მთელი თავისი ისტორიიდან გამომდინარე, არაკრედიტუნარიანობის გავლენის პროცენტის *მაქსიმიზაციით*.

ამრიგად, მანქანური სწავლება ეფუძნება, ერთი მხრივ, მათემატიკას, სახელდობრ, სტატისტიკას, მოდელების ასაგებად და მათ გამოსაყვანად მონაცემებიდან, ხოლო მეორე მხრივ, ინფორმატიკაზე, მონაცემების წარმოსადგენად და ოპტიმიზაციის ალგორითმების ეფექტურად სარეალიზაციოდ. ხელმისაწვდომი მონაცემების სულ უფრო მეტი რაოდენობა მოითხოვს განაწილებული გამოთვლებისა და მონაცემთა ბაზების არქიტექტურების გამოყენებას. ეს ძალიან მნიშნელოვანი მომენტია, მაგრამ ლექციათა ამ კურსში ხსენებული საკითხების განხილვა გათვალისწინებული არ არის.

და რა როლი აქვს ხელოვნურ ინტელექტს ყველაფერ ამაში ?

მანქანური სწავლება შეიძლება იყოს განხილული როგორც ხელოვნური ინტელექტის შტო. მართლაც, სისტემა რომელსაც არ შეუძლია დასწავლა, საეჭვოა გახდეს ინტელექტუალური. სწავლების უნარი და გამოცდილებიდან გაკვეთილების გამოტანის ალღო უაღრესად მნიშვნელოვანია ცვალებად გარემოსთან საადაპტაციოდ განკუთვნილი სისტემისათვის.

ხელოვნური ინტელექტი, განსაზღვრული იმ მეთოდების კრებულად, რომლებიც გამოიყენება გონიერი - ინტელექტუალურად წოდებული - ქცევის დემონსტრირების შემძლე მანქანების შესაქმნელად, ასევე შეიცავს კოგნიტიურ მეცნიერებებს, ნეირობიოლოგიას, ლოგიკას, ელექტრონიკას, ინჟინერიასა და მრავალ სხვა რამეს.

შესაძლოა, რომ სწორედ ამიტომ ტერმინი «ხელოვნური ინტელექტი» უფრო მეტ სტიმულს ბადებს ადამიანთა საზოგადოების წარმოსახვაში და ამიტომ სულ უფრო ხშირად გამოიყენება «მანქანური სწავლების» ნაცვლად.

2 მანქანური სწავლების ამოცანათა ტიპები

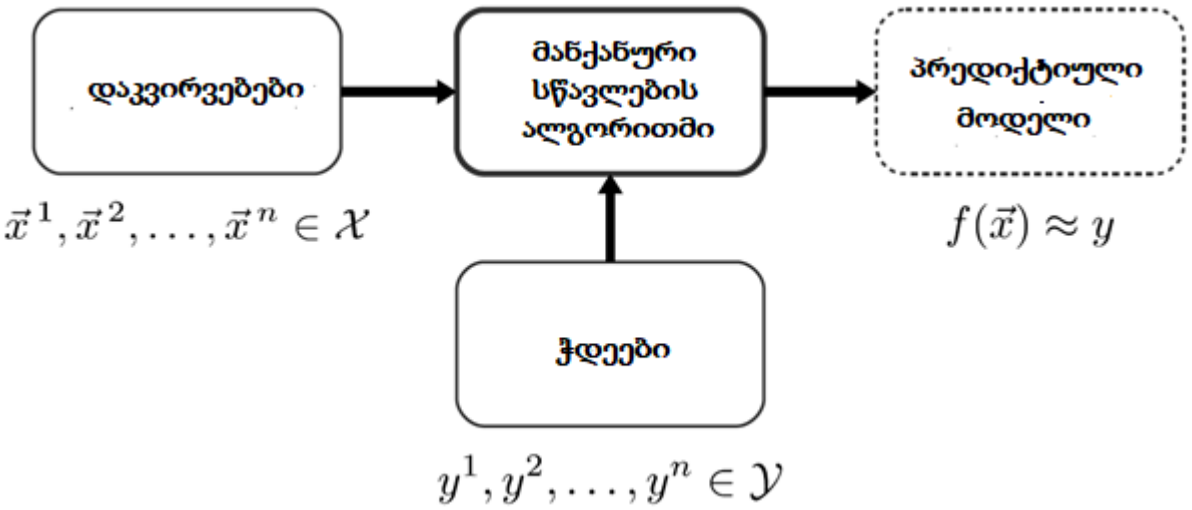
მანქანური სწავლება საკმაოდ ფართო სფეროა და ამ ქვედანაყოფში ჩვენ ჩამოვთვლით პრობლემათა მხოლოდ ყველაზე დიდ კლასებს, რომლებიც გვაინტერესებს კიდეც.

2.1 სწავლება მასწავლებლით

მანქანური სწავლების ამოცანათა შორის კონტროლირებადი სწავლება, ალბათ, ყველაზე მარტივი ტიპია გასაგებად : მისი მიზანია ვისწავლოთ *პროგნოზების* გაკეთება *მონიშნული* მაგალითების სიის საფუძველზე, ესე იგი იმ მნიშვნელობის საფუძველზე, რომელიც უნდა იყოს ნაწინასწარმეტყველები (იხ. ნახატი 1.1). ნიშნები, ანუ ჭდეები, თამაშობს «მასწავლებლის» როლს და კონტროლს უწევს ალგორითმის სწავლების პროცესს.

განსაზღვრება 1.1 (კონტროლირებადი სწავლება) *კონტროლირებადი სწავლება* ეწოდება მანქანური სწავლების შემდეგნაირად ფორმალიზებადი პრობლემებით დაინტერესებულ განშტოებას : მოცემულია n რაოდენობის $\left\{\vec{x}^i\right\}_{i=1,..,n}$ დაკვირვება, რომლებიც აღწერილია $\mathcal{X}$ სივრცეში, და მათი $\left\{y^i\right\}_{i=1,..,n}$ ჭდეები, რომლებიც აღწერილია $\mathcal{Y}$ სივრცეში, იგულისხმება, რომ ჭდეები შეიძლება იყოს მიღებული დაკვირვებებიდან ფიქსირებული და უცნობი $\phi : \mathcal{X} \to \mathcal{Y}$ ფუნქციის საშუალებით : $y^i = \phi\left(\vec{x}^i\right) + \epsilon_i$, სადაც ϵ_i არის შემთხვევითი ხმაური. შემდეგ აუცილებელია მონაცემების გამოყენება $f : \mathcal{X} \to \mathcal{Y}$ ფუნქციის დასადგენად, ისეთის. რომ ნებისმიერი წყვილისათვის $\left(\vec{x}, \phi(\vec{x})\right) \in \mathcal{X} \times \mathcal{Y}$, $f(\vec{x}) \approx \phi(\vec{x})$.

სივრცე, რომელზეც განისაზღვრება მონაცემები, ყველაზე უფრო ხშირად $\mathcal{X} = \mathbb{R}^p$ სახისაა. მაგრამ ჩვენ ასევე გავეცნობით მუშაობას წარმოდგენათა სხვა ტიპებთან, როგორიცაა ბინარული (ორობითი), დისკრეტული ან კატეგორიალური ცვლადები და თუნდაც სტრიქონები ან გრაფები.



ნახატი 1.1 – კონტროლირებადი სწავლება.

ბინარული (ორობითი) კლასიფიკაცია

იმ შემთხვევაში, როცა ჭდეები *ბინარულია*, ისინი უთითებს კუთვნილებას რომელიღაც ერთი *კლასისადმი*. მაშინ ლაპარაკობენ *ბინარულ კლასიფიკაციაზე*.

განსაზღვრება 1.2 (ბინარული კლასიფიკაცია) კონტროლირებადი სწავლების პრობლემას, რომელშიც ჭდეთა სივრცე ბინარულია, ანუ, სხვა სიტყვებით რომ ვთქვათ, $\mathcal{Y}=\{0,1\}$, ბინარული კლასიფიკაციის პრობლემა ეწოდება.

მაგალითი

ქვემოთ მოცემულია ბინარული კლასიფიკაციის ამოცანათა რამდენიმე მაგალითი :

- გაირკვეს, თუ წარმოადგენს ან არა წერილი სპამს ;
- გაირკვეს, თუ არის ან არა ტილო დახატული პიკასოს მიერ ;
- გაირკვეს, თუ არის ან არა ჟირაფი ნახატზე ;
- გაირკვეს, თუ შეუძლია ან არა მოლეკულას დეპრესიის მკურნალობა ;
- გაირკვეს, თუ არის ან არა ფინანსური ოპერაცია თაღლითური.

მრავალკლასიანი კლასიფიკაცია

იმ შემთხვევაში, როცა ჭდეები *დისკრეტულია* და, ამრიგად, შეესაბამება *ორზე მკაცრად მეტ* (ბურბაკის გაგებით, «*მრავალ*») კლასს, მაშინ ლაპარაკობენ *მრავალკლასიან კლასიფიკაციაზე*.

განსაზღვრება 1.3 (მრავალკლასიანი კლასიფიკაცია) კონტროლირებადი სწავლების ამოცანას, რომელშიც ჭდეთა სივრცე არის დისკრეტული და სასრული, სხვა სიტყვებით რომ ვთქვათ, $\mathcal{Y}=\{1,2,...,C\}$, *მრავალკლასიანი კლასიფიკაციის* პრობლემა ეწოდება. C — კლასების რაოდენობაა.

მაგალითი

მრავალკლასიანი კლასიფიკაციის პრობლემათა რამდენიმე მაგალითი შეიძლება ასეთი იყოს :

- დადგინდეს, რომელ ენაზეა დაწერილი ტექსტი ;
- დადგინდეს, რომელი ციფრია ხელნაწერი ათი არაბული ციფრიდან ;
- დადგინდეს სახის გამომეტყველება შესაძლებლობათა წინასწარ განსაზღვრული სიიდან (ბრაზი, სევდა, ბედნიერება, სიხარული და ა.შ.) ;
- დადგინდეს, რომელ სახეობას მიეკუთვნება ესა თუ ის მცენარე ;
- დადგინდეს ფოტოგრაფიაზე წარმოდგენილი ობიექტები.

რეგრესია

იმ შემთხვევაში, როცა ჭდეები წარმოდგენილია *ნამდვილი* სიდიდეებით, ლაპარაკობენ *რეგრესიაზე.*

განსაზღვრება 1.4 (რეგრესია) კონტროლირებადი სწავლების ამოცანას, რომელშიც $\mathcal{Y} = \mathbb{R}$ ჭდეთა სივრცეა გამოყენებული, *რეგრესიის* ამოცანა ეწოდება.

მაგალითი

ქვემოთ მოცემულია რეგრესიის ამოცანის რამდენიმე მაგალითი :

- ბმულზე დაწკაპუნების რაოდენობის წინასწარმეტყველება ;
- ონლაინ-სერვისის მომხმარებელთა რაოდენობის პროგნოზი დროის მოცემულ მომენტში ;
- აქციის (ფასიანი ქაღალდის) ღირებულების პროგნოზი საფონდო ბაზარზე ;
- ორ მოლეკულას შორის ბმის მსგავსების წინასწარმეტყველება ;
- სიმინდის მოსავლიანობის წინასწარმეტყველება.

სტრუქტურირებული რეგრესია

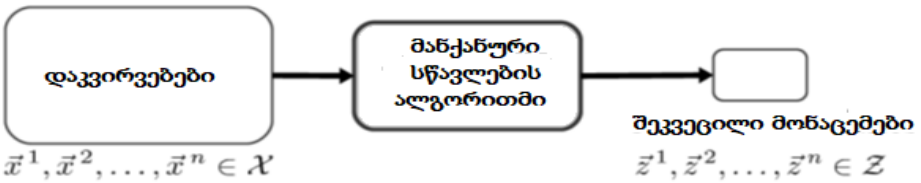
როცა ჭდეთა სივრცე წარმოადგენს უფრო რთულ სტრუქტურირებულ სივრცეს, ვიდრე ზემოთ მოხსენიებული სივრცეები, მაშინ ამას ეწოდება *სტრუქტურირებული რეგრესია* ან *სტრუქტურირებული გამოტანის წინასწარმეტყველება* (ინგლისურად *structured regression ან structured output prediction*). მისი გამოყენება შესაძლებელია, მაგალითად, ვექტორების, გამოსახულებების, გრაფიკების ან მიმდევრობების წინასწარმეტყველებისათვის. სტრუქტურირებული რეგრესია შეიძლება იყოს გამოყენებული მრავალი პრობლემის ფორმალიზაციისათვის. ისეთის. როგორიცაა მანქანური თარგმნის ან მეტყველების ამოცნობის პრობლემები (მაგალითად, ტექსტის გარდაქმნა მეტყველებად და მეტყველების გარდაქმნა ტექსტად). მაგრამ ეს საკითხები გამოდის ლექციათა მოცემული კურსის ფარგლებიდან.

ყურადღება აქ მახვილდება ბინარული და მრავალკლასიანი კლასიფიკაციის და ასევე კლასიკური რეგრესიის პრობლემებზე.

კონტროლირებადი სწავლება ლექციათა ამ კურსის ძირითადი თემაა და იგი დაწვრილებით იქნება განხილული მომდევნო მასალაში.

2.2 სწავლება მასწავლებლის გარეშე

არაკონტროლირებადი სწავლების დროს მონაცემების მონიშვნა არ ხდება. ასეთ შემთხვევაში საქმე ეხება დაკვირვებების მოდელირებას მათი უკეთ გაგებისათვის (იხ. ნახატი 1.2).



ნახატი1.2 – არაკონტროლირებადი სწავლება.

განსაზღვრება 1.5 (არაკონტროლირებადი სწავლება) *არაკონტროლირებადი სწავლება* ეწოდება მანქანური სწავლების განშტოებას, დაინტერესებულს იმ პრობლემებით, რომლებიც შეიძლება იყოს ფორმალიზებული შემდეგი სახით : მოცემულია $\mathcal{X}$ სივრცეში აღწერილი n რაოდენობის $\left\{\vec{x}^i\right\}_{i=1,\ldots,n}$ *დაკვირვება,* და მიზანია $\mathcal{X}$ სივრცეზე განსაზღვრული ისეთი ფუნქციის შესწავლა, რომელიც გარკვეული თვისებების შემოწმებას ახორციელებს.

ეს განსაზღვრება საკმაოდ ბუნდოვანია და, რა თქმა უნდა, უფრო ნათელი და მკაფიო გახდება მაგალითებზე.

კლასტერიზაცია

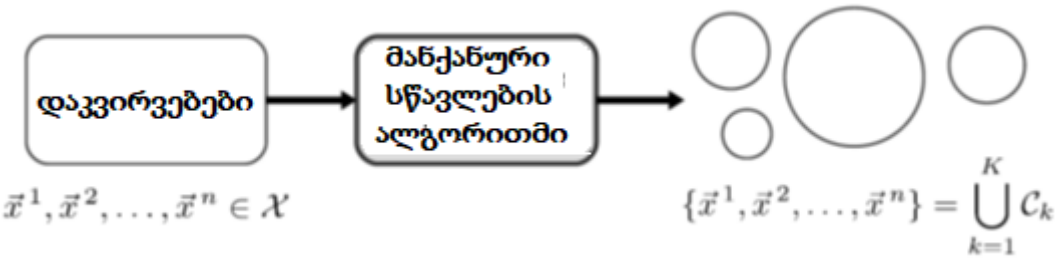
უპირველეს ყოვლისა, *კლასტერიზაცია*, ანუ *დანაწილება (ნაწილებად დაყოფა, ნაწილების გამოცალკევება)*, გულისხმობს *ჯგუფების* გამოვლენას მონაცემებში (იხ. ნახატი 1.3). ეს საშუალებას გვაძლევს გავერკვეთ მათ ზოგად მახასიათებლებში და, შესაძლოა, გავაკეთოთ დასკვნა დაკვირვების თვისებათა შესახებ იმ ჯგუფის საფუძველზე, რომელსაც ეს დაკვირვება მიეკუთვნება.

განსაზღვრება 1.6 (დანაწილება) *დანაწილება* ანუ *კლასტერიზაცია* არის არაკონტროლირებადი სწავლების ერთ-ერთი პრობლემა, რომელიც შეიძლება იყოს ფორმალიზებული როგორც n რაოდენობის $\left\{\vec{x}^i\right\}_{i=1,\ldots,n}$ *დაკვირვებათა* რაღაც $\bigcup_{k=1}^{K} C_k$ დანაწილების (დაჯგუფების) ძებნა. ეს დანაწილება (დაჯგუფება) ერთ ან რამდენიმე ზუსტად მითითებულ კრიტერიუმს უნდა შეესაბამებოდეს.

მაგალითი

ქვემოთ მოცემულია მონაცემთა დანაწილების (კლასტერიზაციის) პრობლემათა რამდენიმე მაგალითი :

- *ბაზრის სეგმენტაცია* გულისხმობს მსგავსი ქცევის მქონე მომხმარებლების ან კლიენტების ჯგუფთა განსაზღვრას. ეს ხელს უწყობს მათი პროფილის უკეთ გაგებას და სარეკლამო კამპანიის, კონტენტის ან მოქმედებების კონკრეტულად გარკვეულ ჯგუფებზე დამიზნების საშუალებასაც იძლევა. სამიზნე ბაზრის შერჩევას *თარგეთინგი* ეწოდება (ინგლისურად *targeting*).
- მსგავსი თემატიკის მქონე დოკუმენტთა ჯგუფების გამოვლენა მათი წინასწარი მონიშვნის გარეშე თემატიკის მიხედვით. ეს ტექსტების დიდი ბანკების ორგანიზების საშუალებას იძლევა.
- *შეკუმშვა* (*გამოსახულების კომპრესია*) შეიძლება იყოს ჩამოყალიბებული როგორც დანაწილების (კლასტერიზაციის) პრობლემა, რომელიც მდგომარეობს მსგავსი პიქსელების გადაჯგუფებაში მათი წარმოდგენისათვის უფრო ეფექტურად.
- *გამოსახულების სეგმენტაცია* მდგომარეობს გამოსახულებაზე იმ პიქსელების გამოვლენაში, რომლებიც ერთსა და იმავე უბანს მიეკუთვნება.
- ჯგუფების გამოვლენა ერთნაირი სიმპტომების მქონე პაციენტთა შორის დაავადების იმ ქვეტიპების გამოვლენის საშუალებას იძლევა, რომლებიც შეიძლება იყოს სხვადასხვანაირად ნამკურნალები.



ნახატი 1.3 – მონაცემთა დაჯგუფება, დანაწილება, ანუ კლასტერიზაცია

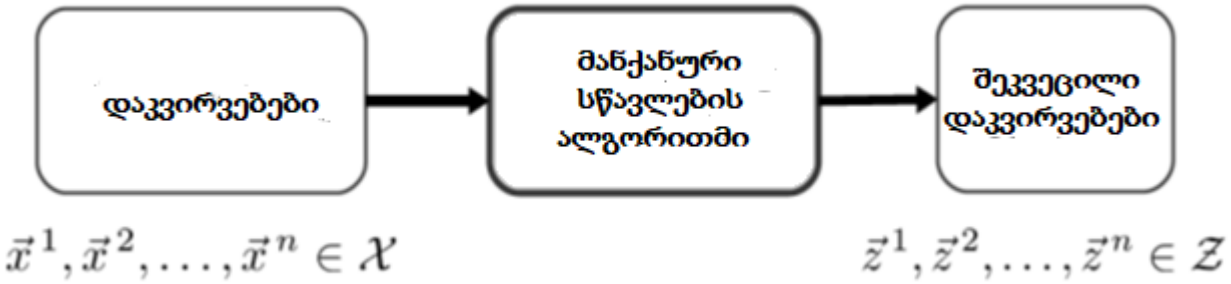
ეს თემა დაწვრილებით განიხილება მე-12 ლექციაში.

განზომილების რედუქცია (შემცირება)

განზომილების რედუქცია (*შემცირება*) - არაკონტროლირებადი სწავლების პრობლემათა კიდევ ერთი მნიშვნელოვანი ოჯახია. იგი გულისხმობს მონაცემთა წარმოდგენის პოვნას უფრო ნაკლები განზომილების სივრცეში, ვიდრე მათი თავდაპირველი წარმოდგენის სივრცეა (იხ. ნახატი 1.4). ეს ამცირებს გამოთვლათა დროს და მონაცემთა შესანახად აუცილებელი მეხსიერების მოცულობას, ასევე ხშირად აუმჯობესებს კონტროლირებადი სწავლების ამავე მონაცემებზე შემდგომ გაწვრთნილი ალგორითმის მახასიათებლებს.

განსაზღვრება 1.7 (განზომილების რედუქცია) *განზომილების რედუქცია* (*შემცირება. შეკვეცა*) ეწოდება არაკონტროლირებადი სწავლების პრობლემას, რომელიც შეიძლება იყოს

ფორმალიზებული როგორც უფრო ნაკლები განზომილების მქონე $\mathcal{Z}$ სივრცის ძებნა, ვიდრე $\mathcal{X}$ სივრცის განზომილებაა, რომელშიც n რაოდენობის $\left\{\vec{x}^i\right\}_{i=1,\ldots,n}$ დაკვირვებებია წარმოდგენილი. ამ მონაცემების $\left\{\vec{z}^i\right\}_{i=1,\ldots,n}$ ასახვები $\mathcal{Z}$ სივრცეზე უნდა აკმაყოფილებდეს გარკვეულ, ზუსტად მითითებულ, თვისებებს.



ნახატი 1.4 – განზომილების შემცირება

შენიშვნა

განზომილების შემცირების ზოგიერთი მეთოდი *კონტროლირებადია* : საქმე ეხება მაშასადამე, იმას, რომ ვიპოვოთ ყველაზე უფრო რელევანტური წარმოსახვა *მოცემული ჭდის წინასწარმეტყველებისთვის.*

განზომილების შემცირების (რედუქციის) უფრო დაწვრილებით განხილვას ჩვენ მე-11 ლექციას დავუთმობთ.

სიმკვრივის შეფასება

დაბოლოს, არაკონტროლირებადი სწავლების პრობლემათა დიდი ოჯახი სინამდვილეში არის სტატისტიკის ტრადიციული პრობლემა : საქმე ეხება ალბათობათა განაწილების შეფასებას იმ დაშვების გათვალისწინებით, რომ მონაცემთა კრებული არის შემთხვევითი ანარჩევი. მე-4 ლექციაში ეს თემა იქნება მოკლედ გაშუქებული.

2.3 ნახევრად-კონტროლირებადი სწავლება

როგორც შეიძლება ველოდოთ, *ნახევრად-კონტროლირებადი სწავლება* (*Semi-supervised learning*), ასევე წოდებული *ნახევრდ-ავტომატურ სწავლებად* ან *ნაწილობრივ სწავლებად*, მდგომარეობს ჭდეთა მიღებაში მონაცემთა ნაწილობრივ მონიშნული კრებულიდან. ამ მიდგომის პირველი უპირატესობა იმაში მდგომარობს, რომ ეს მთელი მასწავლებელი სიმრავლის მონიშვნის აუცილებლობის თავიდან აცილების საშუალებას იძლევა, რაც აქტუალურია იმ შემთხვევაში, როცა მონაცემთა დაგროვება სიძნელეს არ წარმოადგენს, მაგრამ მოითხოვს გარკვეული მოცულობის ადამიანურ მუშაობას მათ მოსანიშნად. მაგალითად, გამოსახულებათა კლასიფიკაციისას ადვილად მისაღებია მონაცემთა ბაზა, რომელიც ასობით ათასი გამოსახულების შემცველია, მაგრამ ჩვენთვის საინტერესო ჭდის დასმა ყოველ სათანადო ნიმუშზე შეიძლება წარმოუდგენლად შრომატევადი აღმოჩნდეს. უფრო მეტიც, ადამიანის მიერ გაკეთებული მონიშვნები, სავარაუდოა, ასახავდეს ადამიანურ

მიდრეკილებებსა და ცრურწმენებს, რომლებიც, თავის მხრივ, დიდი სიზუსტით აღწარმოებული აღმოჩნდება მთლიანად კონტროლირებადი ალგორითმით. ნახევრად-კონტროლირებადი სწავლება ზოგჯერ ამ წყალქვეშა ლოდისგან თავის არიდების საშუალებას იძლევა. საქმე ეხება ახალ და პერსპექტიულ თემას, რომლის დეტალური განხილვა ლექციათა ამ კურსში, გასაგები მიზეზების გამო, ვერ ხერხდება.

2.4 სწავლება წახალისებით

სწავლებაში წახალისებით (*განმტკიცებით*) მასწავლებელი სისტემა შეიძლება ურთიერთობდეს გარემოსთან და ასრულებდეს მოქმედებებს. საპასუხოდ ამ მოქმედებებზე იგი იღებს *ჯილდოს*, რომელიც შეიძლება იყოს დადებითი, თუ მოქმედება უკავშირდებოდა სწორ არჩევანს, ან უარყოფითი წინააღმდეგ შემთხვევაში. ზოგჯერ ჯილდოს მიღება შეიძლება მოხდეს მოქმედებათა ხანგრძლივი თანამიმდევრობის შედეგად ; ასე ხდება, მაგალითად, სისტემაში, რომელიც ასწავლის გოს ან ჭადრაკის თამაშს. ამრიგად, სწავლება მოცემულ შემთხვევაში მდგომარეობს *პოლიტიკის*, ე.ი. საუკეთესო შესაძლო ჯილდოს სისტემატური მიღების სტრატეგიის, განსაზღვრაში.

წახალისებით (განმტკიცებით) სწავლების მთავარი გამოყენებებია თამაშებსა (როგორიცაა, მაგალითად, ჭადრაკი ან გო) და რობოტიკაში. ეს თემა მნიშვნელოვნად გამოდის ლექციათა ამ კურსის ფარგლებიდან.

3 პრაქტიკული რესურსები

3.1 პროგრამული რეალიზაციები

მანქანური სწავლების ალგორითმების რეალიზებას იძლევა მრავალი პროგრამა და ბიბლიოთეკა ღია წვდომით. ქვემოთ მითითებულია ზოგიერთი მათ შორის :

- მაგალითები ამ წიგნში დაწერილია Python ენაზე ფართოდ ცნობილი scikit-learn (http://scikit-learn.org) ბიბლიოთეკის გამოყენებით, რომლის შექმნა, სხვა ბიბლიოთეკათა შორის, ჯერ კიდევ 2007 წელს დაიწყო Inria და Télécom Paris Tech დაწესებულებათა დახმარებით.
- მანქანური სწავლების მრავალი ინსტრუმენტი რეალიზებულია R ენაზე და მოცემულია https://cran.r-project.org/web/views/ გვერდზე.
- Weka (*Waikato environment for knowledge analysis*, *Waikato გარემო ცოდნის ანალიზისათვის,* https://www.cs.waikato.ac.nz/ml/weka/) - მანქანური სწავლების ინსტრუმენტების ერთობლიობა, რომლის შექმნა 1993 წელს დაიწყო ვაიკატოს უნივერსიტეტში (ახალი ზელანდია) და იგი დაწერილია Java ენაზე.
- Shogun (https://github.com/shogun-toolbox) - ეს მრავალი ენის ინტერფეისია, მასთან ურთიერთობაშია, კერძოდ, Python, Octave, R, და C# ენები. 1999 წელს შექმნილი Shogun კომპანია ატარებს თავისი დამაარსებლების სახელებს. ესენია : სიორენ ზონენბურგი (Søren Sonnenburg) და გუნარ რატში (Gunnar Rätsch).

– მრავალი სპეციალიზებული ინსტრუმენტი ღრმა სწავლებისათვის და განაწილებულ არქიტექტურებზე გამოთვლებისათვის გაჩნდა უკანასკნელ წლებში. მათ შორის TensorFlow (https://www.tensorflow.org/) მანქანური სწავლების მრავალ სხვა ალგორითმსაც ახორციელებს.

3.2 მონაცემთა საცავები

მონაცემთა მრავალი საცავი ღია წვდომაშია და მანქანური სწავლების ალგორითმების შესწავლის ან ხელახლა ტესტირების საშუალებას იძლევა. ზოგიერთი საკვანძო რესურსი ასეთია :

– კალიფორნიის უნივერსიტეტი, ირვაინი (UCI - University of California, Irvine), ცნობარი, UCI-ის რეპოზიტორიუმი (http://archive.ics.uci.edu/ml/index.php)

– KDNuggets საიტზე ჩამოთვლილი რესურსები : https://www.kdnuggets.com/datasets/-index.html. KDnuggets - წამყვანი საიტია, რომელიც ეხება მეცნიერებას მონაცემებზე, მანქანურ სწავლებაზე, ხელოვნურ ინტელექტსა და ანალიტიკაზე. KD აკრონიმი ნიშნავს ცოდნის აღმოჩენას - Knowledge Discovery.

– Kaggle პლატფორმა შეჯიბრებათა ჩასატარებლად მეცნიერებებში მონაცემების შესახებ (https://www.kaggle.com/)

4 აღნიშვნათა სისტემა

რამდენადაც ეს შესაძლებელია, ჩვენ ვიყენებთ ამ კურსში შემდეგ აღნიშვნებს :

– სტრიქონული ასოები (x) აღნიშნავს სკალარს ;

– სტრიქონული ასოები მათზე დასმული ისრით $(\vec{x})$ წარმოადგენს ვექტორს ;

– მთავრული ასოები (X) წარმოადგენს მატრიცას, ხდომილებას ან შემთხვევით ცვლადს ;

– კალიგრაფიული ასოები $(\mathcal{X})$ წარმოადგენს სიმრავლეს ან სივრცეს ;

– *სუბსკრიპტები* შეესაბამება ცვლადს, მაშინ როცა *ზედწერები* შეესაბამება დაკვირვებას : x_j^i არის $i-$ური დაკვირვების $j-$ური ცვლადი და შეესაბამება X მატრიცის X_{ij} შემავალ მონაცემს ;

– n არის დაკვირვებათა რიცხვი, p - ცვლადების რიცხვი, C - კლასების რიცხვი ;

– $[a]_+$ წარმოადგენს $a \in \mathbb{R}$ -ის დადებით ნაწილს, სხვა სიტყვებით, არის $\max(0.a)$;

– $\mathbb{P}(A)$ წარმოადგენს A ხდომილების ალბათობას ;

– $\mathbb{E}[X]$ არის X შემთხვევითი ცვლადის მათემატიკური ლოდინი ;

– $\mathbb{V}[X]$ არის X შემთხვევითი ცვლადის დისპერსია ;

— δ არის მაჩვენებელი (ინდიკატორული) ფუნქცია

$$\delta_A = \begin{cases} 1 & \text{თუ } A \text{ არის ჭეშმარიტი} \\ 0 & \text{წინააღმდეგ შემთხვევაში} \end{cases};$$

— $\langle .,. \rangle$ წარმოადგენს სკალარულ ნამრავლს $\mathbb{R}^p$ -ზე ;

— $\langle .,. \rangle_{\mathcal{H}}$ წარმოადგენს სკალარულ ნამრავლს $\mathcal{H}$ -ზე ;

— $M \geq 0$ ნიშნავს, რომ M არის სიმეტრიული ნახევრად-განსაზღვრული დადებითი მატრიცა.

5 საკვანძო მომენტები

— მანქანური სწავლების ალგორითმი არის ალგორითმი, რომელიც ასწავლის მოდელს მაგალითებზე პრობლემის გადაჭრის დაყვანით ოპტიმიზაციის ამოცანამდე.

— მანქანური სწავლება მაშინ გამოიყენება, როცა ძნელია ან შეუძლებელი ცხადი ინსტრუქციების განსაზღვრა კომპიუტერისთვის პრობლემის გადასაჭრელად, მაგრამ არსებობს მრავალი თვალსაჩინო მაილუსტრირებელი მაგალითი.

— მანქანური სწავლების ალგორითმები შეიძლება იყოს დაყოფილი იმ ამოცანის ხასიათის შესაბამისად, რომლის გადაჭრას ისინი ცდილობს. ესენია : კონტროლირებადი, არაკონტროლირებადი, ნახევრად-კონტროლირებადი და განმტკიცებით (წახალისებით) სწავლების ალგორითმები.

დამატებითი ინფორმაცია

- უფრო დაწვრილებითი ინფორმაციის მისაღებად სიმკვრივის შეფასების თაობაზე საინტერესო იქნება Scott-ის წიგნი (1992).
- სტრუქტურულ რეგრესიასთან დაკავშირებით სასარგებლო იქნება Bakir-ისა და თანაავტორების წიგნის (2007) გამოყენება.
- Barto-სა და Sutton-ის წიგნი (1998) კარგი ამოსავალი წერტილია წახალისებით (განმტკიცებით) სწავლების თემაში ჩასაღრმავებლად.

6 ბიბლიოგრაფია

1. Bakir, G., Hofmann, T., Schölkopf, B., Smola, A. J., Taskar, B., et Vishwanathan, S. V. N. (2007). *Predicting Structured Data*. MIT Press, Cambridge, MA.

https://mitpress.mit.edu/books/predicting-structured-data

2. Barto, R. S. et Sutton A. G. (1998). *Reinforcement Learning : An Introduction*. MIT Press, Cambridge, MA. http://incompleteideas.net/book/the-book-2nd.html

3. Benureau, F. (2015). *Self-Exploration of Sensorimotor Spaces in Robots*. Thèse de doctorat, Université de Bordeaux.

https://tel.archives-ouvertes.fr/tel-01259955/file/BENUREAU_FABIEN_2015.pdf

4. Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 44(1.2) : 206–226.

https://hci.iwr.uni-heidelberg.de/system/files/private/downloads/636026949/report_frank_gabel.pdf

5. Scott, D. W. (1992). Multivariate density estimation. Wiley, New York.

https://www.worldcat.org/title/multivariate-density-estimation-theory-practice-and-visualization/oclc/710751351

7 სავარჯიშოები

1.1 ელისოს სურს დაწეროს პროგრამა, რომელიც იყენებს «მეცნიერება», «საზოგადოება», «წვდომა», «უნივერსიტეტი», «მთავრობა», «დაფინანსება», «განათლება», «ბიუჯეტი», «მართლმსაჯულება» და «კანონი» სიტყვათა სიხშირეს იმის დასადგენად, თუ ეძღვნება სტატია მეცნიერების პოლიტიკას. მან დაიწყო ათასობით სტატიის ანოტირება მათი თემატიკის შესაბამისად. მანქანური სწავლების რა პრობლემის გადაჭრა მოუწევს ელისოს ?

1.2 შემდეგი პრობლემებიდან რომელი ამოცანები გადაწყდება კარგად მანქანური სწავლების საშუალებით?

1. ოპტიმალური დროის განრიგის (გრაფიკის) დადგენა კონტენტის განსათავსებელად ვებ-გვერდზე.
2. უმოკლესი მანძილის დადგენა ორ კვანძს შორის გრაფზე.
3. იმ ველოსიპედების რაოდენობის პროგნოზირება, რომლის აყვანა აღმოჩნდება შესაძლებელი გაქირავების საქალაქო სისტემის ყოველ სადგურზე.
4. იმ ღირებულების შეფასება, რომელსაც შეიძლება მიაღწიოს ოსტატის ტილომ აუქციონზე გაყიდვისას.
5. რადიო სიგნალის გაშიფვრა.

1.3 ბენჟამენს 10 000 საგაზეთო სტატია აქვს და მას სურს მათი კლასიფიცირება თემების მიხედვით. ამისათვის რომელი ალგორითმი უნდა გამოიყენოს მან : კონტროლირებადი თუ არაკონტროლირებადი?

1.4 სესილიას მონაცემები აღწერილია 10 ცვლადით. მაგრამ მას სურს მათი წარმოდგენა ორგანზომილებიან ერთ გრაფიკზე. სწავლების ალგორითმის რომელი ტიპის გამოყენება მოუწევს მას?

1.5 გიორგი მართავს ინსტრუმენტს, რომელიც შენახული HTML-დამოწმებების (HTML-ბმულების) მოწესრიგებას ახორციელებს. კატეგორიებს, რომლებსაც უნდა მიეკუთვნოს ახალი

დამოწმება, საფუძვლად უდევს სერვისის ყველა მომხმარებლის მიერ უკვე განსაზღვრულ კატეგორიათა გამოყენება. როგორი ალგორითმი უნდა იყოს გიორგის მიერ გამოყენებული ?

1.6 ელისოს გაუჩნდა თავის ფოსტაში სპამების შესწავლის სურვილი, სახელდობრ, სპამის სხვადასხვა ტიპის არსებობის დადგენის მიზნით. სწავლების რომელი ტიპის ალგორითმი უნდა გამოიყენოს მან?

1.7 ტომ მიტჩელი მანქანურ სწავლებას შემდეგნაირად განსაზღვრავს : «*ითვლება, რომ კომპიუტერული პროგრამა სწავლობს E გამოცდილებაზე T ამოცანისა და რაღაც ერთი მახასიათებლის P ზომისათვის, თუ ეს მახასიათებელი T -ზე, გამოსახული P მნიშვნელობით. უმჯობესდება E გამოცდილების ზრდასთან ერთად*». ფრედი წერს პროგრამას, რომელიც იყენებს ბანკის მონაცემებს საბანკო თაღლითობის გამოსავლენად. რას არის E, T და P ?

სავარჯიშოთა ამონახსნები

1.1 კონტროლირებადი სწავლება (ბინარული კლასიფიკაცია).

1.2 1, 3, 4 (2 წყდება გრაფზე ძებნის ალგორითმებით, 5 - სიგნალების დამუშავების ალგორითმებით).

1.3 არაკონტროლირებადი სწავლების ალგორითმი.

1.4 განზომილების რედუქციის (შემცირების, შეკვეცის) ალგორითმი.

1.5 კონტროლირებადი სწავლება (მრავალკლასიანი კლასიფიკაცია).

1.6 არაკონტროლირებადი სწავლება (კლასტერიზაცია).

1.7 E = ბანკის მონაცემები. T = თაღლითობის პროგნოზირება. P = თაღლითობის სწორად გამოვლენის უნარი (ვთქვათ, პროცენტობით).

ლექცია 2 სწავლება მასწავლებლით

შინაარსი

ლექციათა მოცემულ კურსში ჩვენ გვაინტერესებს, უმთავრესად, *კონტროლირებადი სწავლების* პრობლემები : მიზანია ალგორითმების შექმნა, რომლებსაც *პროგნოსტიკული,* ანუ *პრედიქტიული მოდელების* (*predictive models*) სწავლების უნარი აქვს. მონიშნული მაგალითების საფუძველზე ეს მოდელები შეძლებს ახალი ობიექტების ჭდეთა წინასწარმეტყველებას. მოცემული კურსის მიზანია იმ ზოგადი კონცეფციების ჩამოყალიბება, რომლებიც პრობლემათა ამ ტიპის ფორმალიზების საშუალებას იძლევა.

მიზნები

— პრობლემის ფორმალიზება კონტროლირებადი სწავლების ამოცანის სახით ;

— დანახარჯთა ფუნქციის არჩევა ;

— კავშირის დადგენა მოდელის განზოგადების უნარსა და მის სირთულეს შორის.

1 მასწავლებლით სწავლების პრობლემის ფორმალიზება

კონტროლირებადი სწავლების პრობლემა შეიძლება იყოს ფორმალიზებული შემდეგი სახით : მოცემულია n რაოდენობის $\left\{\vec{x}^1, \vec{x}^2, \ldots, \vec{x}^n\right\}$ დაკვირვებანი, სადაც ყოველი $\vec{x}^i$ დაკვირვება არის დაკვირვებათა $\mathcal{X}$ სივრცის ელემენტი, ასევე მათი $\left\{y^1, y^2, \ldots, y^n\right\}$ *ჭდეები,* სადაც ყოველი y^i ჭდე მიეკუთვნება ჭდეთა $\mathcal{Y}$ სივრცეს, კონტროლირებადი სწავლების მიზანია $f : \mathcal{X} \to \mathcal{Y}$ ფუნქციის პოვნა, ისეთის, რომ $f\left(\vec{x}\right) \approx y$ ყველა $\left(\vec{x}, y\right) \in \mathcal{X} \times \mathcal{Y}$ წყვილისთვის, სადაც მეწყვილენი ისეთივე თანაფარდობაშია ერთმანეთთან, როგორც დანაკვირვებ წყვილებში. $D = \left\{\left(\vec{x}^i, y^i\right)\right\}_{i=1,\ldots,n}$ სიმრავლე *მასწავლებელი ანარჩევის* ფორმირებას ახდენს.

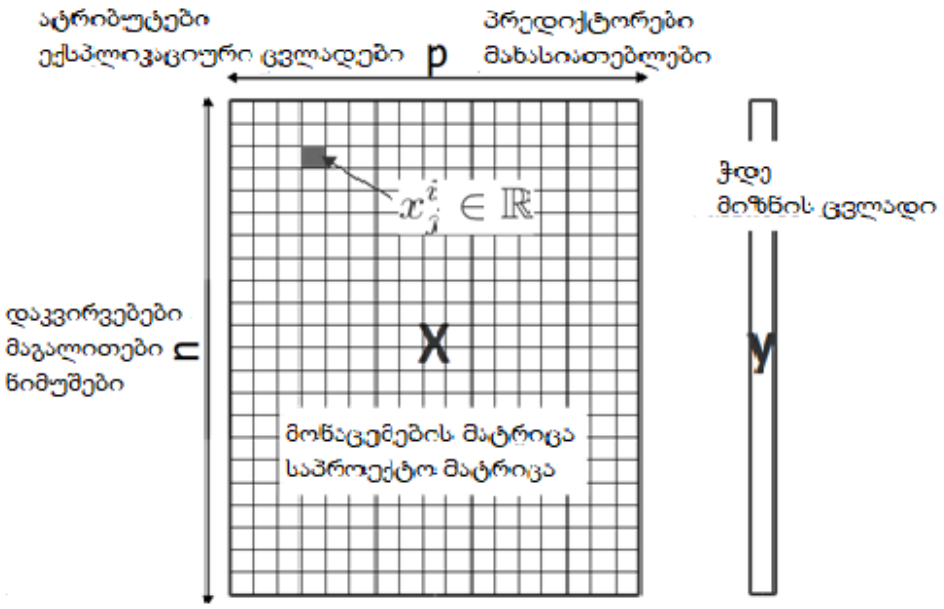
ლექციათა მოცემულ კურსში ჩვენ სამ განსაკუთრებულ შემთხვევას განვიხილავთ, როცა ჭდეთა $\mathcal{Y}$ სივრცისათვის გვაქვს :

— $\mathcal{Y} = \mathbb{R}$: ამ შემთხვევაში *რეგრესიის* პრობლემაზე ლაპარაკობენ ;

— $\mathcal{Y} = \{0,1\}$: ამ შემთხვევაში *ბინარული კლასიფიკაციის* ამოცანასთან აქვთ საქმე და 0--ით მონიშნულ დაკვირვებებს *უარყოფით* (*ნეგატიურ*) დაკვირვებებს უწოდებენ, ხოლო მაშინ, როცა დაკვირვებები მონიშნულია 1-ით, მათ *დადებით* (*პოზიტიურ*) დაკვირვებებს უწოდებენ. ზოგიერთ შემთხვევაში მათემატიკურად უფრო მარტივი იქნება $\mathcal{Y} = \{-1,1\}$ ჭდეთა სივრცის გამოყენება ;

— $\mathcal{Y} = \{1,2,...,C\}$, $C > 2$: ამ შემთხვევაში *მრავალკლასიანი კლასიფიკაციის* პრობლემაზე ლაპარაკობენ.

ზოგჯერ პრობლემა დაიყვანება შემთხვევამდე, როცა $\mathcal{X} = \mathbb{R}^p$. მაშინ ამბობენ, რომ დაკვირვებები წარმოდგენილია p *ცვლადით*. ამ შემთხვევაში $X \in \mathbb{R}^{n \times p}$ მატრიცას, ისეთს რომ $X_{ij} = x_j^i$ არის $j-$ური ცვლადი $i-$ურ დაკვირვებაში, უწოდებენ *მონაცემების მატრიცას* ან *საპროექტო მატრიცას* (ზოგჯერ, *დაპროექტების მატრიცას*).



ნახატი 2.1 – დაპროექტების (საპროექტო) მატრიცისა და ჭდეთა ვექტორის სახით ორგანიზებული კონტროლირებადი სწავლების ამოცანის მონაცემები და თავიანთი ექსპლიკაციური (ამხსნელი) ცვლადებით წარმოდგენილი დაკვირვებები.

ვინაიდან მანქანური სწავლების სათავე ძალიან ადვილად მოიპოვება რამდენიმე დისციპლინაში და გამოყენებით სფეროში, ხშირად ვაწყდებით სხვადასხვა დასახელებას ერთისა და იმავე ობიექტისათვის. ასე, მაგალითად, *ექსპლიკაციურ ცვლადებს* იმავდროულად უწოდებენ *დესკრიპტორებს*, *ატრიბუტებს*, *პრედიქტორებს* ან *თავისებურებებს / მახასიათებლებს* (ინგლისურად: *variables*, *descriptors*, *attributes*, *predictors* ან კიდევ *features*). *დაკვირვებები* ასევე ცნობილია სახელწოდებებით : *მაგალითები*, *ნიმუშები* ან *წერტილები მონაცემთა*

კრებულში (ინგლისურად : *samples* ან *data points*). დაბოლოს, *ჭდეებისათვის* ცნობილია ასეთი სახელებიც : *ეტიკეტები* / *ნიშნები*, *მიზნები* / *მიზნის ცვლადები* (ინგლისურად : *labels*, *targets* ან *outcomes*). სწორედ ეს კონცეფციებია ილუსტრირებული ნახატზე 2.1.

1.1 გადაწყვეტილება

კლასიფიკაციის ამოცანის შემთხვევაში პროგნოზირების მოდელი შეიძლება პირდაპირ იღებდეს f ფუნქციის ფორმას მნიშვნელობებით $\{0,1\}$ სიმრავლეზე, ან იყენებდეს შუალედურ g ფუნქციას ნამდვილი მნიშვნელობებით, რომელიც ანიჭებს დაკვირვებას მით უფრო მაღალ შეფასებას, რაც მეტია მოლოდინი, რომ იგი დადებითი იქნება. ეს შეფასება შეიძლება იყოს, მაგალითად, იმის ალბათობა, რომ მოცემული დაკვირვება მიეკუთვნება დადებით კლასს. ამრიგად, f ფუნქცია მიიღება g ფუნქციის *ზღურბლური მნიშვნელობით* ; g ფუნქციას *გადაწყვეტილების მიღების ფუნქცია* ეწოდება.

განსაზღვრება 2.1 (გადაწყვეტილების მიღების ფუნქცია) ბინარული კლასიფიკაციის ამოცანის ასპექტში *გადაწყვეტილების მიღების ფუნქცია*, ანუ *დისკრიმინანტული ფუნქცია* ეწოდება $g:\mathcal{X}\mapsto\mathbb{R}$ ფუნქციას, ისეთს რომ $f(\vec{x})=0$, თუ და მხოლოდ თუ $g(\vec{x})\le 0$ და $f(\vec{x})=1$. თუ და მხოლოდ თუ $g(\vec{x})>0$.

ეს განსაზღვრება შეიძლება იყოს განზოგადებული *მრავალკლასიანი კლასიფიკაციის* შემთხვევაში : მაშინ აქვთ გადაწყვეტილების მიღების C რაოდენობის ფუნქცია $g_c:\mathcal{X}\mapsto\mathbb{R}$, ისეთი რომ $f(\vec{x})=\arg\max_{c=1,\dots,C} g_c(\vec{x})$.

✿

გადაწყვეტილების მიღების ფუნქციის ცნება სივრცის დანაწილების საშუალებას იძლევა *გადაწყვეტილებათა მიღების არეებად* :

განსაზღვრება 2.2 (გადაწყვეტილების მიღების არე) ბინარული კლასიფიკაციის ამოცანის შემთხვევაში დისკრიმინანტული ფუნქცია ყოფს დაკვირვებათა $\mathcal{X}$ სივრცეს *გადაწყვეტილების მიღების* ორ ისეთ - $\mathcal{R}_0$ და $\mathcal{R}_1$ - არედ. რომ

$$\mathcal{R}_0 = \{\vec{x}\in\mathcal{X} \mid g(\vec{x})\le 0\} \text{ და } \mathcal{R}_1 = \{\vec{x}\in\mathcal{X} \mid g(\vec{x})>0\}.$$

მრავალი კლასის შემთხვევაში, ასევე, არსებობს გადაწყვეტილების მიღების C არე :

$$\mathcal{R}_c=\left\{\vec{x}\in\mathcal{X} \mid g_c(\vec{x})=\max_k g_k(\vec{x})\right\}.$$

✿

გადაწყვეტილებათა მიღების არეები განცალკევებულია ერთმანეთისგან *გადაწყვეტილებათა მიღების საზღვრებით* :

განსაზღვრება 2.3 (გადაწყვეტილების მიღების საზღვარი) კლასიფიკაციის პრობლემის კონტექსტში *გადაწყვეტილების მიღების საზღვარი,* ანუ *დისკრიმინანტი* ეწოდება

წერტილების $\mathcal{X}$ სიმრავლეს, სადაც გადაწყვეტილების მიღების ფუნქცია გაუქმებულია. ბინარული პრობლემის შემთხვევაში არის გადაწყვეტილების მიღების მხოლოდ ერთი საზღვარი ; მრავალკლასიანი პრობლემის შემთხვევაში C კლასით საზღვრების რიცხვიც ამდენივეა, ე.ი. $C-$ს შეადგენს.

1.2 მრავალკლასიანი კლასიფიკაცია

ლექციათა წინამდებარე კურსში წარმოდგენილ კონტროლირებადი სწავლების მეთოდებს შორის ზოგი მრავალკლასიანი პრობლემების უშუალოდ, პირდაპირ გადაჭრის საშუალებას იძლევა. მაგრამ ბინარული (ორობითი) კლასიფიკაციის ნებისმიერი ალგორითმი შეიძლება იყოს გამოყენებული კლასიფიკაციის პრობლემის გადასაწყვეტად C კლასზე «*ერთი ყველას წინააღმდეგ*» (ინგლისურად : One-versus-All) ან «*ერთი ერთის წინააღმდეგ*» (ინგლისურად : One-versus-One) მიდგომათა გამოყენებით.

განსაზღვრება 2.4 (ერთი ყველას წინააღმდეგ) როცა მოცემულია მრავალკლასიანი კლასიფიკაციის ამოცანა C კლასით, მაშინ «*ერთი ყველას წინააღმდეგ*» ანუ «*one-versus-all*» სიტყვებით გამოხატულ აზრს (პრინციპს, კონცეფციას) ეწოდება მიდგომა C ბინარულ კლასიფიკატორთა სწავლებისადმი. $c-$ური ამ კლასიფიკატორებიდან იყენებს c კლასის ყველა ეგზემპლარს როგორც დადებით მაგალითს და ყველა დანარჩენ დაკვირვებას როგორც უარყოფით მაგალითს გადაწყვეტილების g_c ფუნქციის მისაღებად. ამრიგად თითოეული კლასიფიკატორი სწავლობს ერთი კლასის გარჩევას ყველა სხვა კლასისგან. $\vec{x}$ ვექტორის ჭდე მოცემულია გადაწყვეტილების მიღების ფუნქციებიდან სწორედ ამ ერთი ფუნქციით. რომელიც გვიბრუნებს უდიდეს შეფასებას :

$$f\left(\vec{x}\right)=\arg\max_{c=1,..,C} g_c\left(\vec{x}\right).$$

⚙

განსაზღვრება 2.5 (ერთი ერთის წინააღმდეგ) როცა მოცემულია მრავალკლასიანი კლასიფიკაციის ამოცანა C კლასით, მაშინ «*ერთი ერთის წინააღმდეგ*» ანუ «*one-versus-one*» სიტყვებით გამოხატული აზრი (პრინციპი, კონცეფცია) არის $C\left(C-1\right)$ ბინარული კლასიფიკატორის შექმნის მიდგომა, რომელთა შორის თითოეული გამოაცალკევებს ერთ კლასს მეორისგან ყველა დანარჩენი მაგალითის იგნორირებით. დავუშვათ, რომ g_{ck} გადაწყვეტილების მიღების ფუნქციას წარმოადგენს იმ ბინარული კლასიფიკატორისთვის, რომელიც c კლასს k კლასისგან გამოყოფს. $\vec{x}$ ვექტორის ჭდე კი შემდეგნაირად განისაზღვრება :

$$f(\vec{x})=\arg\max_{c=1,..,C}\left(\sum_{k\neq c} g_{ck}(\vec{x})\right).$$

⚙

ძნელია იმის თქმა, ამ ორი მიდგომიდან რომელია უფრო ეფექტური. პრაქტიკაში არჩევანი ხშირად იქნება წარმართული ალგორითმული სირთულის მოსაზრებებით :

უფრო ეფექტურია C მოდელის სწავლება n დაკვირვებაზე, თუ $C(C-1)$ მოდელის $\frac{n}{C}$ დაკვირვებაზე (იმ პირობით, რომ კლასები დაბალანსებულია, ე.ი. $\mathcal{D}$ ყოველი კლასის მაგალითების, დაახლოებით, ერთნაირ რაოდენობას შეიცავს) ?

გარდა ამისა, მხედველობაში მიიღება ასევე ხელმისაწვდომი მაგალითების რაოდენობა ყოველი კლასისათვის : მიუხედავად იმისა, რომ მოდელის სწავლება უფრო ეფექტურია მონაცემთა მცირე რაოდენობაზე, თუ მონაცემების ეს რაოდენობა ზედმეტად მცირეა, მაშინ კარგი ხარისხის მოდელის მიღება ძნელი აღმოჩნდება.

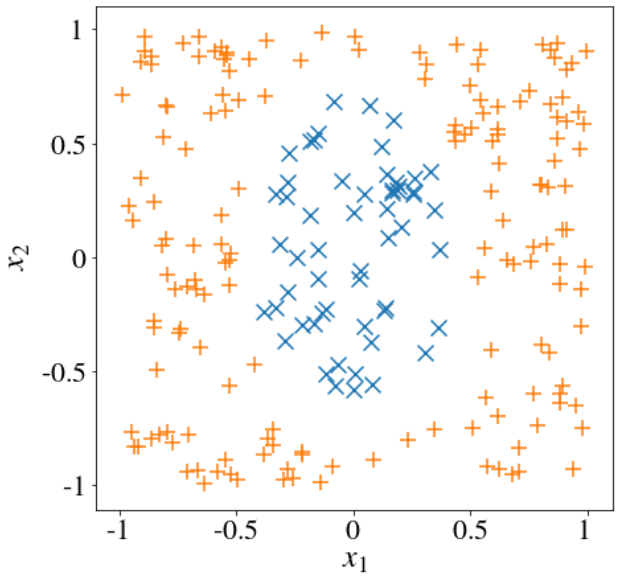
2 ჰიპოთეზათა სივრცე

კონტროლირებადი სწავლების ამოცანის დასმამდე ჩვენ უნდა მივიღოთ გადაწყვეტილება მამოდელირებელ ფუნქციათა იმ ტიპზე, რომლის განხილვას ვაპირებთ.

განსაზღვრება 2.6 (ჰიპოთეზათა სივრცე) *ჰიპოთეზათა სივრცე* ეწოდება ჩვენი ხედვის არეში მოხვედრილი მამოდელირებელი ფუნქციების აღმწერ $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ ფუნქციათა სივრცეს. ამ სივრცის არჩევა იმ *წარმოდგენათა* შესაბამისად ხდება, რომლებიც პრობლემასთან დაკავშირებით გვაქვს.

ამ ლექციის მეორე ნახატის მაგალითში შეიძლება შემოვიფარგლოთ დისკრიმინანტებით, რომლებიც წარმოადგენს ელიფსებს საკოორდინატო ღერძების პარალელური ღერძებით. ამრიგად ჰიპოთეზათა სივრცეს შემდეგი სახე ექნება :

$$\mathcal{F} = \left\{ \vec{x} \mapsto \alpha(x_1 - a)^2 + \beta(x_2 - b)^2 - 1 \right\}.$$



ნახატი 2.2 – დადებითი $(+)$ და უარყოფითი $(\times)$ მაგალითები გვაგონებს ელიფსით განცალკევებას.

დავუშვათ, რომ გვაქვს n მონიშნული დაკვირვების $\mathcal{D} = \{(\vec{x}^i, y^i)\}_{i=1,\dots,n}$ სიმრავლე და ჰიპოთეზათა $\mathcal{F}$ სივრცე. კონტროლირებადი სწავლების ამოცანა მდგომარეობს ვარაუდში, რომ ჭდეები გამოთვლილი იყო $\phi : \mathcal{X} \to \mathcal{Y}$ ფუნქციის საშუალებით, და ჰიპოთეზის პოვნაში, რომელიც მიზნის ϕ ფუნქციის უკეთეს აპროქსიმაციას იძლევა.

ასეთი ამოცანის განსახორციელებლად ორი დამატებითი ინსტრუმენტი (საშუალება) დაგვჭირდება :

1. *ჰიპოთეზის ხარისხის რაოდენობრივი შეფასების* ხერხი, რომელიც საშუალებას იძლევა განისაზღვროს თუ იყო ნაპოვნი დამაკმაყოფილებელი (ან. შესაძლოა. ოპტიმალური) ჰიპოთეზა. ამისათვის განყოფილებაში 2.4 განვსაზღვრავთ *ღირებულების* (*ხარჯების*) *ფუნქციის* (ინგლისურად : *cost function*) ცნებას.
2. $\mathcal{F}$ *სივრცეში ოპტიმალური ჰიპოთეზის ძებნა.* ლექციათა ამ კურსში ყურადღება გამახვილებული იქნება *სწავლების მეთოდებზე ოპტიმიზაციით* : კონტროლირებადი სწავლების ალგორითმებს, რომლებსაც შევუდგებით, მიზნად ექნება $\mathcal{F}$ სივრცეში ღირებულების (ხარჯების) ფუნქციის თვალსაზრისით (იხ. განყოფილება 2.3) *ოპტიმალური ჰიპოთეზის პოვნა.* სხვადასხვა ალგორითმები ჰიპოთეზათა სხვადასხვა $\mathcal{F}$ სივრცეს განსაზღვრავს და, რეალიზებული შემთხვევის შესაბამისად, ეს ძებნა იქნება ზუსტი ან მიახლოებითი.

ჰიპოთეზათა სივრცის არჩევას ფუნდამენტური მნიშვნელობა აქვს. მართლაც, თუ ეს სივრცე არ შეიცავს «წესიერ» ფუნქციას, მაგალითად, თუ ჰიპოთეზათა სივრცე ამ ლექციის მეორე ნახატის მონაცემებისათვის არჩეულია როგორც წრფეთა სიმრავლე, მაშინ გადაწყვეტილების მიღების კარგი ფუნქციის პოვნა შეუძლებელი იქნება.

მაგრამ მაშინაც კი, როცა სივრცე ნებისმიერია (მაგალითად, სრულიად შემთხვევითი სახისაა), ხარისხიანი მამოდელირებელი ფუნქციის პოვნა რთულდება და უზარმაზარ გამოთვლით სიმძლავრეს მოითხოვს.

3 ემპირიული რისკის მინიმიზაცია

კონტროლირებადი სწავლების ამოცანის ამოხსნა ნიშნავს $f \in \mathcal{F}$ ფუნქციის პოვნას, რომლის წინასწარმეტყველებები მაქსიმალურად უახლოვდება ჭეშმარიტ ჭდეებს მთელ $\mathcal{X}$ სივრცეზე. ამის ფორმალიზებისათვის გამოიყენება *ხარჯების* (*ღირებულების*) *ფუნქციის* (ინგლისურად : *cost function*) ცნება.

განსაზღვრება 2.7 (ხარჯების/ღირებულების ფუნქცია) *ხარჯების/ღირებულების ფუნქცია* $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, ასევე ცნობილი როგორც *დანაკარგის ფუნქცია* ან *შეცდომის ფუნქცია* (ინგლისურად : *cost function* ან *loss function*) არის ფუნქცია, რომელიც გამოიყენება წინასწარმეტყველების ხარისხის რაოდენობრივი შეფასებისთვის : $L\big(y, f(\vec{x})\big)$ ფუნქცია მით

უფრო იზრდება, რაც უფრო შორდება $f(\vec{x})$ ჭდე ჭეშმარიტ y მნიშვნელობას.

✿

როცა ღირებულების (ხარჯების) L ფუნქცია მოცემულია, მასთან დაკავშირებით ვეძებთ f -ს, რომელიც მინიმუმს ანიჭებს ამ ღირებულებას (ხარჯებს) $\vec{x} \in \mathcal{X}$ ვექტორის შესაძლო მნიშვნელობათა სიმრავლეზე, რაც ფორმალიზებულია *რისკის* ცნებით.

განსაზღვრება 2.8 (რისკი) კონტროლირებადი სწავლების პრობლემის ჭრილში *რისკი* ეწოდება ღირებულების (ხარჯების) ფუნქციის მათემატიკურ ლოდინს (იხ. მე-4 ლექცია კონტროლირებადი სწავლების ამოცანის ალბათური მოდელირების შესახებ) :

$$\mathcal{R}(h) = \mathbb{E}_{\mathcal{X}}\Big[L\big(h(\vec{x}), y\big)\Big].$$

✿

ასე რომ, f ფუნქცია, რომელსაც ჩვენ ვეძებთ, $f = \arg\min_{h \in \mathcal{F}} \mathbb{E}[L\big(h(\vec{x}), y\big)$ პირობას ამოწმებს. როგორც წესი, ამ პრობლემის გადაჭრა ვერ ხერხდება დამატებითი დაშვებების გარეშე : ჩვენ რომ ვიცოდეთ $\mathcal{X}$ სივრცის ყველა წერტილის ჭდე, მაშინ მანქანური სწავლება არც დაგვჭირდებოდა. როცა მოცემულია n მონიშნული $\big\{(\vec{x}^i, y^i)\big\}_{i=1,\dots,n}$ დაკვირვება, ახდენენ რისკის აპროქსიმაციას მისი შეფასებით ამ დანაკვირვებ მონაცემებზე.

განსაზღვრება 2.9 (ემპირიული რისკი) კონტროლირებადი სწავლების პრობლემის ჭრილში, როცა მოცემულია n მონიშნული $\big\{(\vec{x}^i, y^i)\big\}_{i=1,\dots,n}$ დაკვირვება, *ემპირიული რისკი* ეწოდება შემდეგ შეფასებას :

$$R_n(h) = \frac{1}{n}\sum_{i=1}^{n} L\big(h(\vec{x}^i), y^i\big).$$

✿

ასე რომ, პრედიქტორს (პროგნოზირების საშუალებას) *ემპირიული რისკის მინიმიზაციით* შემდეგი სახე აქვს :

$$f = \arg\min_{h \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} L\big(h(\vec{x}^i), y^i\big). \tag{2.1}$$

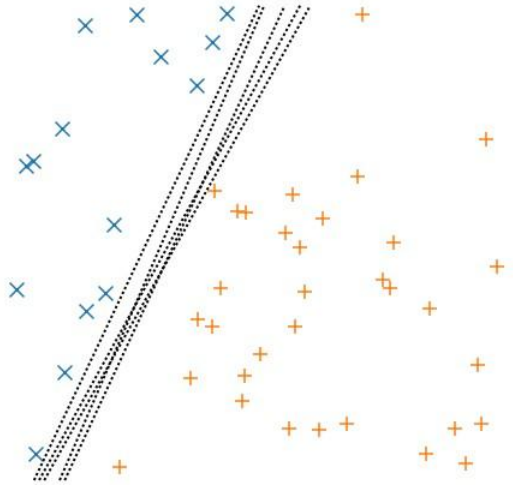
ჰიპოთეზათა $\mathcal{F}$ სივრცის არჩევის მიხედვით, განტოლებას 2.1 შეიძლება ჰქონდეს ერთი ცხადი ანალიზური ამონახსნი.

ეს ყოველთვის ეგრე როდია ; ამიტომ ხშირად ხდება ღირებულების (ხარჯების) *ამოზნექილი ფუნქციის* (ინგლისურად : *convex function*) არჩევა ოპტიმიზაციის ამ ამოცანის გადაჭრის გასაადვილებლად (იხ. ლექცია 13 ამოზნექილი ოპტიმიზაციის შესახებ).

ემპირიული რისკის მინიმიზაცია, ჩვეულებრივ, არის ჟაკ ადამარის (Jacques Hadamard) მიხედვით *ცუდად ჩამოყალიბებული* (არაკორექტული) *ამოცანა* (ინგლისურად : *an ill-posed problem in Hadamard's sense*), ესე იგი მას არ აქვს ერთადერთი ამონახსნი, რომელიც უწყვეტად

იქნება დამოკიდებული საწყის პირობებზე.

მაგალითად, შეიძლება აღმოჩნდეს, რომ ამონახსნთა უსასრულო რაოდენობამ ნულამდე დაიყვანოს მინიმიზაციისას ემპირიული რისკი (იხ. ნახ, 2.3).



ნახატი 2.3 – წრფეთა უსასრულო რაოდენობა იდეალურად გამოყოფს დადებით $(+)$ წერტილებს უარყოფითი $(\times)$ წერტილებისგან. ყოველ მათგანს აქვს ნულოვანი ემპირიული რისკი.

გარდა ამისა, პრედიქტორი. რომელიც ემპირიულ რისკს მინიმუმს ანიჭებს, არ არის სტატისტიკურად *ძალმოსილი* (ინგლისურად : *consistent*). გავიხსენოთ, რომ θ პარამეტრის n დაკვირვებაზე დამოკიდებული θ_n შეფასება არის *ძალმოსილი* (ინგლისურად : *consistent estimate*), თუ იგი კრებადია ალბათობის მიხედვით θ პარამეტრის მიმართ, როცა n მიისწრაფვის უსასრულობისკენ :

$$\forall \epsilon > 0, \lim_{n\to\infty} \mathbb{P}\left(\left|\theta_n - \theta\right| \geq \epsilon\right) = 0. \tag{2.2}$$

დიდი რიცხვების კანონი გარანტიას გვაძლევს. რომ ემპირიული რისკი კრებადია რისკის მიმართ

$$\forall h \in \mathcal{F}, R_n(h) \underset{n\to\infty}{\to} \mathcal{R}(h). \tag{2.3}$$

მაგრამ ეს საკმარისი არ არის იმის გარანტირებისათვის, რომ ემპირიული რისკის $\min_{h\in\mathcal{F}} R_n(h)$ მინიმუმი კრებადი აღმოჩნდეს რისკის მინიმუმის მიმართ.

მართლაც, თუ $\mathcal{F}$ განზომადი ფუნქციების სივრცეა, მაშინ $\min_{h\in\mathcal{F}} R_n(h)$, როგორც წესი, შეადგენს ნულს, რაც ასე არ არის $\mathcal{R}(h)$-ის შემთხვევაში. ასე რომ, რაიმე გარანტია იმისა, რომ f_n ფუნქცია, რომელიც $\mathcal{R}_n(h)$-ის მინიმიზაციას ახდენს, იქნება იმავდროულად $\mathcal{R}(h)$ რისკისათვის მინიმუმის მიმნიჭებელი f ფუნქციის კარგი შეფასებაც, ნამდვილად არ არსებობს.

ემპირიული რისკის მინიმიზაციის *ძალმოსილება* (ინგლისურად : *consistency,* რუსულად : *состоятельность*) დამოკიდებულია ვერსიათა $\mathcal{F}$ სივრცეზე. ამ ძალამოსილების შესწავლა

ვაპნიკისა (Vladimir Vapnik) და ჩერვონენკისის (Alexey Chervonenkis) სწავლების თეორიის ერთ-ერთი მთავარი ელემენტია, მაგრამ 1960 — 1990 წლებში ჩამოყალიბებული ეს თეორია ლექციათა წინამდებარე კურსის მისწრაფებებსა და მიზნებს ცდება.

4 დანახარჯების ფუნქციები

არსებობს *დანახარჯების*, ანუ *ღირებულების* მრავალი *ფუნქცია* (ინგლისურად : *cost function*). დანახარჯების ფუნქციის არჩევანი დამოკიდებულია, ერთი მხრივ, თავად ამოცანაზე, ე.ი. იმაზე, თუ რა არის აქტუალური განსახილველ პრაქტიკულ შემთხვევაში, ხოლო მეორე მხრივ, პრაქტიკულ მოსაზრებებზე : შესაძლებელი თუ არის ოპტიმიზაციის მიღებული ამოცანის საკმარისად ზუსტად და სწრაფად ამოხსნა ? ამ ლექციაში წარმოდგენილია დანახარჯების ყველაზე გავრცელებულ სახეთა პოპულარული ფუნქციები და მათი დამოწმება შესაძლებელი იქნება ამ კურსის მომდევნო ლექციებშიც.

4.1 დანახარჯების (ღირებულების) ფუნქციები ბინარული კლასიფიკაციისათვის

ღირებულების (დანახარჯების) ფუნქციის დასადგენად ბინარული (ორობითი) კლასიფიკაციისათვის ხშირად განიხილება $\mathcal{Y} = \{-1,1\}$ ჭდეთა სივრცე. მართლაც, იდეალური (უშეცდომო) კლასიფიკაციის შემთხვევაში $yf(\vec{x})$ ნამრავლი ერთის ტოლია.

0/1 ღირებულება ბინარული (ორობითი) კლასიფიკაციისთვის

განსაზღვრება 2.10 (0/1 ღირებულება – ბინარული კლასიფიკაცია) ბინარულ სიდიდეთა f ფუნქციის შემთხვევაში *0/1 ღირებულების ფუნქცია,* ანუ *0/1 დანაკარგები* (ინგლისურად :*Loss*) ეწოდება შემდეგ ფუნქციას :

$$L_{0/1} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$$
$$y, f(\vec{x}) \mapsto \begin{cases} 1 & \text{თუ } f(\vec{x}) \neq y \\ 0 & \text{წინააღმდეგ შემთხვევაში.} \end{cases}$$

თუ გამოვიყენებთ $\mathcal{Y} = \{-1,1\}$ ჭდეთა სივრცეს, შესაძლებელია მისი ხელახლა ჩაწერა შემდეგი სახით :

$$L_{0/1}\big(y, f(\vec{x})\big) = \frac{1 - yf(\vec{x})}{2}.$$

ღირებულების (ხარჯების) ამ ფუნქციის გამოყენებისას ემპირიული რისკი არის პროგნოზირების (წინასწარმეტყველების) შეცდომათა საშუალო რიცხვი.

თუ f დამოკიდებულებად განიხილება ნამდვილ სიდიდეთა გადაწყვეტილების მიღების ფუნქცია და არა ბინარულ (ორობით) სიდიდეთა პროგნოზირების (წინასწარმეტყველების) ფუნქცია, მაშინ 0/1 ღირებულების (ხარჯების) ფუნქცია შეიძლება შემდეგი სახით განი-საზღვროს :

0/1 ღირებულება რეგრესიისათვის

განსაზღვრება 2.11 (0/1 ღირებულება / დანახარჯები – რეგრესია) როცა განიხილება ნამდვილ სიდიდეთა გადაწყვეტილების მიღების ფუნქცია, მაშინ *0/1 ღირებულების (დანახარჯების) ფუნქცია*, ან სხვანაირად, *0/1 დანაკარგების ფუნქცია* (ინგლისურად :*0/1 loss*) შემდეგ ფუნქციას ეწოდება :

$$L_{0/1} : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}$$
$$y, f(\vec{x}) \mapsto \begin{cases} 1 & \text{თუ } yf(\vec{x}) \leq 0 \\ 0 & \text{წინააღმდეგ შემთხვევაში.} \end{cases}$$

ღირებულების ამ ფუნქციის უარყოფითი მხარე ისაა, რომ იგი არ არის დიფერენცირებადი, რაც ართულებს ოპტიმიზაციის ამოცანებს მისი გამოყენებით. უფრო მეტიც, იგი არ არის ძალიან ზუსტი : შეცდომა ისეთივეა, როგორც $f(\vec{x})$-ის, დამოუკიდებლად იმისა, იმყოფება იგი ძალიან ახლოს თუ ძალიან შორს გადაწყვეტილების მიღების ზღურბლიდან. შეიძლება გავიხსენოთ, რომ იდეალური, უშეცდომო კლასიფიკაციის შემთხვევაში $yf(\vec{x}) = 1$, როცა ჭდეთა სივრცე მოცემულია $\mathcal{Y} = \{-1, 1\}$ ბინარული სახით. ამრიგად, შეიძლება განვსაზღვროთ *ღირებულების* (*ხარჯების*) *ფუნქცია*, რომელიც მით მეტია, რაც უფრო შორს იმყოფება $yf(\vec{x})$ სიდიდე მარცხნივ 1-დან (ე.ი. $yf(\vec{x}) < 1$) ; მიიჩნევა, რომ შეცდომას ადგილი არ აქვს, თუ $yf(\vec{x}) > 1$. ამას *სახსრული შეცდომის* (ინგლისურად : *hinge error*) განმარტებასთან მივყავართ. შეცდომის ეს დასახელება ჩნდება იმის გამო, რომ იგი ახდენს «იდაყვის», «მუხლის» ანუ «სახსრების» ფორმირებას (იხ. ნახ. 2.4). ინგლისური *hinge* სიტყვა ქართულად ითარგმნება როგორც ანჯამა (კარისა, ფანჯრისა); სახსარი.

სახსრული შეცდომა

განსაზღვრება 2.12 (სახსრული შეცდომა) *სახსრული შეცდომის ფუნქცია*, ანუ *hinge loss* (*სახსრული დანაკარგი* / *ზარალი*), ეწოდება შემდეგ ფუნქციას :

$$L_{\text{hinge}} : \{-1, 1\} \times \mathbb{R} \to \mathbb{R}$$
$$y, f(\vec{x}) \mapsto \begin{cases} 0 & \text{თუ } yf(\vec{x}) \geq 0 \\ 1 - yf(\vec{x}) & \text{წინააღმდეგ შემთხვევაში.} \end{cases}$$

უფრო კომპაქტური სახით სახსრული შეცდომა შეიძლება ასევე შემდეგნაირად ჩაიწეროს :

$$L_{\text{hinge}}\big(f(\vec{x}), y\big) = \max\big(0, 1 - yf(\vec{x})\big) = [1 - yf(\vec{x})]_+.$$

შეიძლება ასევე ჩაითვალოს, რომ $f(\vec{x})$ უნდა იყოს მაქსიმალურად ახლოს 1-თან დადებითი დაკვირვებებისთვის (და -1-თან უარყოფითი დაკვირვებებისთვის). ამრიგად, დაჯარიმდება ასევე შემთხვევები, როცა $yf(\vec{x})$ გადაიხრება 1-დან მარჯვნივ, რაც შეიძლება გაკეთდეს *კვადრატული ღირებულების* (*ხარჯების*) საშუალებით (ინგლისურად : *quadratic* / *squared cost* ან კიდევ *square loss*).

კვადრატული ღირებულება (დანახარჯები) ბინარული კლასიფიკაციისათვის

განსაზღვრება 2.13 (კვადრატული ღირებულება/დანახარჯები) ბინარული კლასიფიკაციის ამოცანაში *კვადრატული ღირებულება* (*ხარჯები*), ანუ *square loss* ეწოდება შემდეგ ფუნქციას :

$$L_{\text{square}} : \{-1,1\} \times \mathbb{R} \to \mathbb{R}$$
$$y,\ f(\vec{x}) \mapsto \left(1 - yf(\vec{x})\right)^2 .$$

✿

დაბოლოს, შეიძლება ვეცადოთ განვსაზღვროთ გადაწყვეტილების მიღების ფუნქცია, რომლის აბსოლუტური მნიშვნელობა რაოდენობრივად ასახავს ჩვენს ნდობას მისი წინასწარმეტყველებისადმი. ამ შემთხვევაში ყველაფერს აკეთებენ იმისათვის, რომ $yf(\vec{x})$ გახდეს რაც შეიძლება უფრო დიდი და მიმართავენ *ლოგისტიკურ ღირებულებას* (*დანახარჯებს*).

ლოგისტიკური ღირებულება (დანახარჯები)

განსაზღვრება 2.14 (ლოგისტიკური ღირებულება/დანახარჯები) *ლოგისტიკური ღირებულების* (*დანახარჯების*) *ფუნქცია,* ანუ *ლოგისტიკური დანაკარგები* (ინგლისურად : *logistic loss*) ეწოდება შემდეგ ფუნქციას :

$$L_{\log} : \{-1,1\} \times \mathbb{R} \to \mathbb{R}$$
$$y,\ f(\vec{x}) \mapsto \log\left(1 + \exp\left(-yf(\vec{x})\right)\right)$$

✿

თუ უპირატესობა ჭდეთა $\mathcal{Y} = \{0,1\}$ სივრცეს ეძლევა, მაშინ ლოგისტიკური ღირებულება (ხარჯები) *ჯვარედინი ენტროპიის* ექვივალენტურია.

ჯვარედინი ენტროპია

განსაზღვრება 2.15 (ჯვარედინი ენტროპია) ბინარულ შემთხვევაში *ჯვარედინი ენტროპია*, ანუ *კროს-ენტროპია* (ინგლისურად : *cross-entropy*) შემდეგ ფუნქციას ეწოდება :

$$L_H : \{0,1\} \times]0,1[\to \mathbb{R}$$
$$y,\ f(\vec{x}) \mapsto -y \log f(\vec{x}) - (1-y)\log\left(1 - f(\vec{x})\right).$$

✿

შენიშვნა

ჯვარედინი ენტროპიის სათავე ინფორმაციის თეორიაში მოიპოვება, საიდანაც მისი დასახელებაც არის წარმოშობილი. თუ მხედველობაში მივიღებთ, რომ $\vec{x}$ ვექტორის ჭეშმარიტი კლასის მოდელირება Q განაწილებით ხდება, ხოლო მისი ნაწინასწარმეტყველი კლასის – P განაწილებით, ჩვენ შევეცდებით P-ს მოდელირება ისეთნაირად ჩავატაროთ, რომ იგი რაც შეიძლება ახლოს იმყოფებოდეს Q-თან. ამისათვის გამოიყენება *კულბაკისა და ლეიბლერის მანძილი / შეუსაბამობა / დივერგენცია* (ინგლისურად : *Kullback–Leibler divergence*), ანუ სხვანაირად, *ფარდობითი ენტროპია.* იგი წარმოადგენს *განსხვავების ზომას ალბათობათა ორ*

განაწილებას შორის. ამ ზომის შემოღება დაკავშირებულია ორი ამერიკელი კრიპტოანალიტიკოსის – სოლომონ კულბაკისა (Solomon Kullback, 1907–1994)) და რიჩარდ ლეიბლერის (Richard A. Leibler, 1914–2003) – სახელებთან.

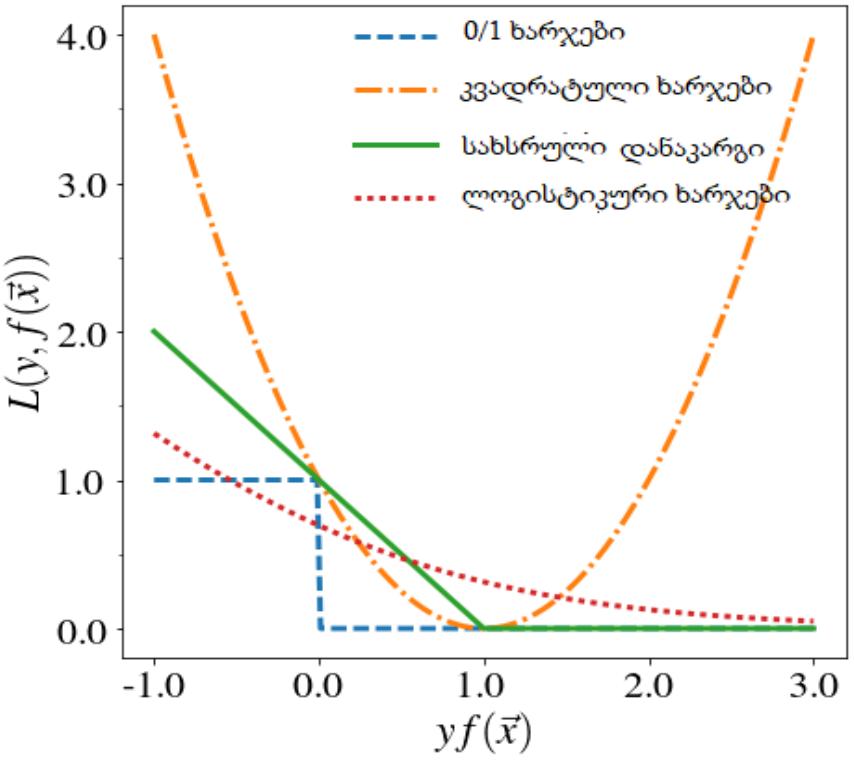
თუ კულბაკისა და ლეიბლერის მანძილისთვის (ფარდობითი ენტროპიისათვის) შემოვიღებთ $KL(Q \| P)$ აღნიშვნას, გვექნება :

$$KL(Q \| P) = \sum_{c=0,1} Q(y=c \mid \vec{x}) \log \frac{Q(y=c \mid \vec{x})}{P(y=c \mid \vec{x})} =$$

$$= -\sum_{c=0,1} Q(y=c \mid \vec{x}) \log P(y=c \mid \vec{x}) + \sum_{c=0,1} Q(y=c \mid \vec{x}) \log Q(y=c \mid \vec{x})$$

ვინაიდან $Q(y=c \mid \vec{x})$ არის ან 0 (c არ წარმოადგენს $\vec{x}$ ვექტორის კლასს) ან 1 (წინააღმდეგ შემთხვევაში), ამ გამოსახულების მეორე წევრი ნული იქნება და, ამრიგად, ჯვარედინი ენტროპიის ზემოთ მოცემული განსაზღვრება შენარჩუნებული დარჩება.

დანაკარგის (ზარალის) ფუნქციები ბინარული კლასიფიკაციისათვის ნაჩვენებია ქვემოთ (იხ. ნახ. 2.4).



ნახატი 2.4 – ხარჯების (დანაკარგის/ზარალის) ფუნქციები ბინარული კლასიფიკაციისათვის.

4.2 დანახარჯები მრავალკლასიანი კლასიფიკაციისათვის

ჯვარედინი ენტროპია (კროს-ენტროპია)

ჯვარედინი ენტროპიის განსაზღვრება ბინარულ შემთხვევაში ბუნებრივად შეიძლება განზოგადდეს მრავალკლასიან შემთხვევაზე $f_c : \mathcal{X} \to \mathcal{Y}$ გადაწყვეტილების მიღების C

ფუნქციათა განხილვის გზით.

განსაზღვრება 2.16 (ჯვარედინი ენტროპია) მრავალკლასიან შემთხვევაში *ჯვარედინი ენტროპია*, ანუ *კროს-ენტროპია* (ინგლისურად : *cross-entropy*) ეწოდება შემდეგ ფუნქციას :

$$L_H : \{1, 2, \ldots, C\} \times \mathbb{R} \mapsto \mathbb{R}$$
$$y, f(\vec{x}) \mapsto -\sum_{c=1}^{C} \delta(y, c) \log f_c(\vec{x}) = -\log f_y(\vec{x}).$$

სახსრული შეცდომის ფუნქციის გაფართოება

არსებობს რამდენიმე მოსაზრება სახსრული შეცდომის ფუნქციის განზოგადების შესახებ მრავალკლასიან შემთხვევაზე.

მიზანი ერთია და ყოველთვის იმაში მდგომარეობს, რომ გადაწყვეტილების მიღების $f_y(\vec{x})$ ფუნქცია $\vec{x}$ ვექტორის ჭეშმარიტი კლასისათვის იძენდეს უფრო დიდ მნიშვნელობას, ვიდრე გადაწყვეტილების მიღების ყველა დანარჩენი $f_c(\vec{x}),\ c \neq y$ ფუნქცია.

ჯეისონ უესტონს (Jason Weston) და კრის უოტკინსს (Chris Watkins) ხელოვნური ინტელექტის ევროპულ სიმპოზიუმზე ჯერ კიდევ 1999 წელს შემოთავაზებული აქვთ შემდეგი განსაზღვრება :

$$L_{\text{hinge}}\left(y, f(\vec{x})\right) = \sum_{c \neq y} \left[1 + f_c(\vec{x}) - f_y(\vec{x})\right]_+.$$

ორი წლის შემდეგ (2001) კობი კრამერი (Koby Crammer) და იორამ სინგერი (Yoram Singer) იყენებენ უკვე $f_c(\vec{x})$-ის მაქსიმუმს და არა უბრალოდ $f_c(\vec{x})$-ს, ასე რომ, მათი განსაზღვრებით გვაქვს :

$$L_{\text{hinge}}\left(y, f(\vec{x})\right) = \left[1 + \max_{c \neq y} f_c(\vec{x}) - f_y(\vec{x})\right].$$

ხარჯების ამ ფუნქციებს იშვიათად ხმარობენ პრაქტიკაში და, როგორც წესი, უპირატესობა ეძლევა ორობითი სახსრული დანაკარგის გამოყენებას «ერთი ერთის წინააღმდეგ» ან «ერთი ყველას წინააღმდეგ» მიდგომით.

4.3 დახარჯები რეგრესიისათვის

რეგრესიის ამოცანის შემთხვევაში უკვე ჭდეთა $\mathcal{Y} = \mathbb{R}$ სივრცე განიხილება. ღირებულების (ხარჯების) ჩვენი ფუნქციის მიზანია მაპროგნოზირებელი f ფუნქციის დაჯარიმება, თუ მისი მნიშვნელობა შორს არის $\vec{x}$ ვექტორის მიზნობრივი მნიშვნელობიდან.

კვადრატული ღირებულება (დანახარჯები)

განსაზღვრება 2.17 (კვადრატული ღირებულება) *ღირებულების კვადრატული ფუნქცია*, ან *კვადრატული დანაკარგი* (ინგლისურად : *quadratic loss*), ან კიდევ *კვადრატული შეცდომა* (ინგლისურად : *squared error*) ეწოდება შემდეგ ფუნქციას :

$$L_{\text{SE}} : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$$
$$y, f(\vec{x}) \mapsto \frac{1}{2}\big(y - f(\vec{x})\big)^2 .$$

$\frac{1}{2}$ კოეფიციენტი საშუალებას იძლევა თავიდან ავიცილოთ მამრავლი კოეფიციენტების გაჩენა ემპირიული რისკის დიფერენცირებისას მისი მინიმიზაციის პროცესში.

ϵ – უგრძნობელი ღირებულება (დანახარჯები)

კვადრატული ღირებულება (ხარჯები), როგორც წესი, დომინირებს *აგერანტული* (ტიპურ სიდიდეთა არიდან ამოვარდნილი) *მნიშვნელობებით* :

როგორც კი რამდენიმე დაკვირვებას მონაცემთა კრებულში აქვს პროგნოზი, რომელიც ძალიან აცდენილია მის ფაქტობრივ ჭდეს, წინასწარმეტყველების ხარისხი დანარჩენი დაკვირვების მიმართ პრაქტიკულად არაფერს ნიშნავს.

ასე რომ, კვადრატულ ღირებულებას შეიძლება შევეშვათ და უპირატესობა ვეძიოთ, მაგალითად, ეგრეთ წოდებულ *აბსოლუტურ ღირებულებაში.*

განსაზღვრება 2.18 (აბსოლუტური ღირებულება) *აბსოლუტური ღირებულების ფუნქცია*, ან *აბსოლუტური შეცდომა* (ინგლისურად : *absolute error*) ეწოდება შემდეგ ფუნქციას :

$$L_{\text{AE}} : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$$
$$y, f(\vec{x}) \mapsto \big|y - f(\vec{x})\big|.$$

ღირებულების ასეთი ფუნქციის პირობებში მაშინაც კი, როცა პროგნოზები ძალიან უახლოვდება ჭეშმარიტ ჭდეებს, ეს წინასწარმეტყველებები ჯარიმდება (თუნდაც ოდნავ). მაგრამ რიცხობრივად თითქმის შეუძლებელია ზუსტი პროგნოზის მიღება.

ϵ – უგრძნობელი ღირებულების ფუნქცია (ინგლისურად : *the ϵ – insensitive cost function*) ამ შეზღუდვის თავიდან აცილების საშუალებას იძლევა.

განსაზღვრება 2.19 (ϵ - უგრძნობელი ღირებულება) დავუშვათ, რომ მოცემულია $\epsilon > 0$, *ϵ – უგრძნობელი ღირებულების ფუნქცია*, ანუ *ϵ-უგრძნობელი დანაკარგი* (ინგლისურად : *ϵ – insensitive loss*), ეწოდება შემდეგ ფუნქციას :

$$L_{\epsilon} : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$$
$$y, f(\vec{x}) \mapsto \max\big(0, \big|y - f(\vec{x})\big| - \epsilon\big) .$$

ჰუბერის ღირებულების ფუნქცია

ϵ – უგრძნობელი ღირებულება არც $(-\epsilon)$ – ზე არის დიფერენცირებადი და არც $(+\epsilon)$ – ზე, რაც

ართულებს ემპირიული რისკის ოპტიმიზაციას.

ჰუბერის ღირებულების ფუნქცია კარგი კომპრომისის მიღწევის საშუალებას იძლევა ($0-$ ზე დიფერენცირებად) კვადრატულ ღირებულებასა და აბსოლუტურ ღირებულებას შორის, (რომელიც ექსტრემუმების მნიშვნელობებზე წყვეტას არ განიცდის).

შემოღებულია შვეიცარიელი სტატისტიკოსის პეტერ ჰუბერის მიერ 1964 წელს და დღეს უკვე შეფასებულია გარღვევად სტატისტიკაში (იხ. Peter Jost Huber, *Robust estimation of a location parameter*. In : « Breakthroughs in statistics », pages 492–518, Springer, 1992).

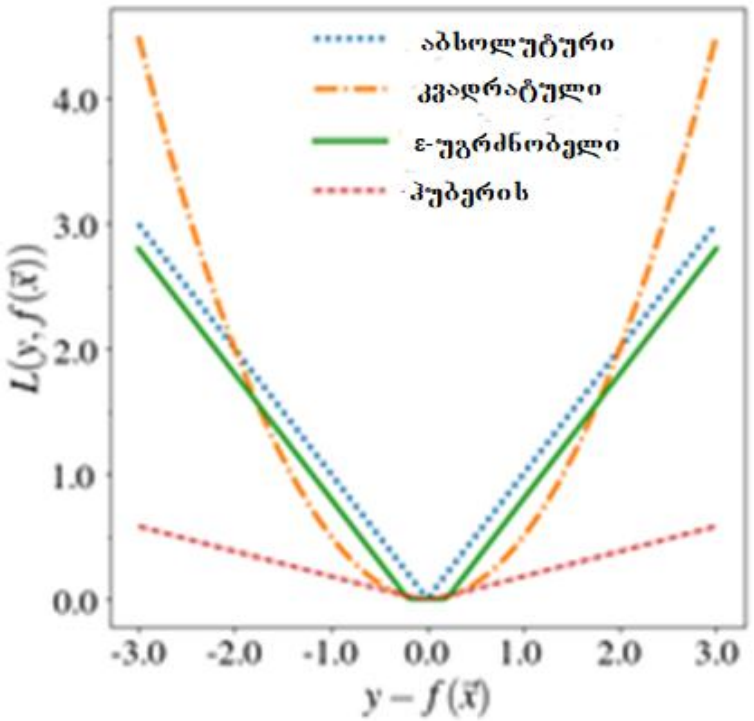
განსაზღვრება 2.20 (ჰუბერის ღირებულების ფუნქცია) *ჰუბერის ღირებულების ფუნქცია*, ან *ჰუბერის დანაკარგი/ჰუბერის ზარალი* (ინგლისურად : *Huber loss*) ეწოდება შემდეგ ფუნქციას :

$$L_{\text{Uber}} : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$$

$$y, f(\vec{x}) \mapsto \begin{cases} \frac{1}{2}\left(y - f(\vec{x})\right)^2 & \text{თუ } |y - f(\vec{x})| < \epsilon \\ \epsilon |y - f(\vec{x})| - \frac{1}{2}\epsilon^2 & \text{წინააღმდეგ შემთხვევაში.} \end{cases}$$

გამოსახულება $-\frac{1}{2}\epsilon^2$ უზრუნველყოფს ფუნქციის უწყვეტობას.

ღირებულების (ხარჯების) ფუნქციები რეგრესიისათვის ნაჩვენებია ნახატზე 2.5.



ნახატი 2.5 – ღირებულების (ხარჯების) ფუნქციები რეგრესიის ამოცანისთვის.

5 განზოგადება და ზედმეტად სწავლება

5.1 განზოგადება

წარმოვიდგინოთ ალგორითმი, რომელიც რაღაც $\vec{x}$ დაკვირვების ჭდის წინასწარმეტყველებისათვის გვიბრუნებს მის ჭდეს, თუ $\vec{x}$ მიეკუთვნება მონაცემებს ცნობილი ჭდეებით, და რომელიმე შემთხვევით სიდიდეს — წინააღმდეგ შემთხვევაში. ამ ალგორითმს ექნება მინიმალური ემპირიული შეცდომა ღირებულების ფუნქციის არჩევისგან დამოუკიდებლად, მაგრამ მოუწევს ძალიან ცუდი პროგნოზების გაკეთება ყველა ახალი დაკვირვებისათვის. მაგრამ ეს მთლად ის არ არის, რაც მხედველობაში აქვთ ხოლმე, როცა ლაპარაკობენ სწავლების შესახებ.

ყურადღება

ამრიგად, მანქანური სწავლების ალგორითმის შეფასება იმ მონაცემებზე, რომლებზეც იგი სწავლობდა, არაფერს გვეუბნება იმის შესახებ, როგორ მოიქცევა იგი ახალ მონაცემებზე, სხვა სიტყვებით რომ ვთქვათ, *განზოგადების* მის უნარზე, ეს არსებითი მომენტია !

განსაზღვრება 2.21 (განზოგადება) *განზოგადება* ეწოდება მოდელის უნარს აკეთოს კორექტული წინასწარმეტყველებები ახალ მონაცემებზე, რომლებიც არ ყოფილა გამოყენებული მის ასაგებად.

5.2 ზედმეტად სწავლება

ზემოთ მოტანილი, უდავოდ განსაკუთრებული, საგანგებოდ შერჩეული მაგალითი მოწმობს, რომ ადვილად შესაძლებელია სწავლების ისეთი პროცედურის ჩამოყალიბება, რომელიც მაპროგნოზირებელი სისტემის ასაგებად გამოყენებულ მონაცემებზე კარგი წინასწარმეტყველების შემძლე, მაგრამ ცუდად განმაზოგადებელ მოდელს ქმნის. ნაცვლად იმისა, რომ საინტერესო ობიექტების ჭეშმარიტი ბუნების მოდელირება მოახდინოს, ასეთი მოდელი ასევე (შეიძლება ითქვას, ძირითადად) აფიქსირებს და ასახავს ხმაურს, რომელსაც არავითარი კავშირი აქვს განხილვის საგანთან. მართლაც, მანქანური სწავლების ნებისმიერ ამოცანაში ჩვენი მონაცემები გარდაუვლად დამახინჯებულია

— გაზომვის ცდომილებებით, რომლებიც გამოწვეულია ჩვენი მონაცემების ამსახველი ცვლადების გასაზომად გამოყენებული სენსორების (გადამწოდების) უზუსტობით, ან იმ ოპერატორების ადამიანური შეცდომებით, რომლებმაც ეს გაზომვები მონაცემთა ბაზაში შეიტანეს ;

— ხშირად *მასწავლებლის ხმაურად* წოდებული (ინგლისურად : *teacher's noise*) *მონიშვნების შეცდომებით*, რომლებიც დაკავშირებულია მონაცემთა მონიშვნის განმხორციელებელი ოპერატორების ადამიანურ შეცდომებთან ;

— დაბოლოს, გასაზომ ცვლადთა რაოდენობის არასაკმარისობით ჩვენთვის საინტერესო მოვლენის მოდელირებისათვის, რაც შეიძლება იყოს გამოწვეული ან იმის გამო, რომ ისინი ცნობილი არ არის, ან იმის გამო, რომ მათი გაზომვა ძალიან ძვირი ჯდება).

მაგალითი

დავუშვათ, რომ გვინდა ფოტოგრაფიების კლასიფიცირება იმის მიხედვით, არის თუ არა ამ ფოტოგრაფიებზე მოცემული პანდები. ყოველი გამოსახულება წარმოდგენილია მისი შემადგენელი RGB პიქსელების მნიშვნელობებით. გვინდა დავრწმუნდეთ, რომ ჩვენ მიერ შექმნილი მოდელი ასახავს პანდის ჭეშმარიტ ბუნებას. მაგრამ ვერ გამოირიცხება გაზომვათა შეცდომები (ტექნიკური ცდომილებები ფოტოგრაფიული კამერის გადამწოდებში) და ასევე მონიშვნის შეცდომები (იმ ადამიანის შეცდომები, რომელსაც უნდა გადაეწყვიტა ყოველი ფოტოგრაფიისათვის, ეხებოდა თუ არა საქმე პანდას, და შეეძლო ან მცდარი არჩევანის ღილაკზე დაწკაპუნება ან კიდევ პანდის არევა დათვში). გარდა ამისა, შეზღუდულია ცვლადების ნაირსახეობაც : პიქსელები ვერ ასახავს უშუალოდ იმ ფაქტს, რომ პანდა მსუქანი ცხოველია თვალების გარშემო ნიღბით, რომელიც, ჩვეულებრივ, გარშემორტყმულია ბამბუკით.

შენიშვნა

როგორც ვხედავთ, მონაცემთა წარმოსადგენად გამოსაყენებელი ცვლადების არჩევა ძალიან მნიშვნელოვანი ეტაპია მოდელირების პროცესში. თუმცა მეშვიდე ლექციაში დავინახავთ, რომ ღრმა ნეირონული ქსელების სიმძლავრე, რომლებიც დღემდე ესოდენ პოპულარული რჩება, განპირობებულია მათი უნარით ამოიღონ მონაცემებიდან შესაბამისი წარმოდგენა. მეთოდები, რომლებიც აღწერილია მე-11 ლექციაში, შეიძლება იყოს გამოყენებული სახელმძღვანელოდ მაპროგნოზირებელი ცვლადების (პრედიქტორების) არჩევისას გამოყენების მიზნით.

განსაზღვრება 2.22 (ზედმეტად სწავლება) მოდელზე, რომელიც, ნაცვლად იმისა, რომ ასახავდეს მოსანიშნი ობიექტების ბუნებას, ასევე ამოდელირებს კიდეც ხმაურს და ვერ ახერხებს განზოგადებას, ამბობენ. რომ იგი *ზედმეტად დასწავლილია.* ინგლისურ ენაში ამას ეწოდება «*overfitting*» — *ზედმეტად მორგება.*

ზედმეტად დასწავლილი მოდელი, ჩვეულებრივ, არის *ძალიან რთული* მოდელი, რომელიც ძალიან «ეკვრის» მონაცემებს და ამიტომ მათი ხმაურის მიტაცებაც უწევს.

პირიქითაც, შესაძლებელია *ძალიან მარტივი მოდელის* აგებაც, რომლის მახასიათებლები კარგი არ იქნება არც ან მოდელის ასაგებად გამოყენებულ მონაცემებზე და არც განზოგადე-ბისთვის.

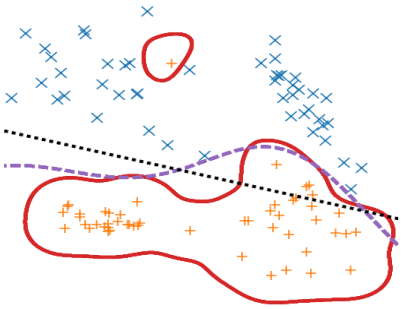
განსაზღვრება 2.23 (არასაკმარისად სწავლება) მოდელზე, რომელიც ძალიან მარტივია იმისათვის, რომ მას გააჩნდეს კარგი მახასიათებლები თუნდაც ამ მოდელის ასაგებად გამოყენებულ მონაცემებზე, ამბობენ, რომ იგი *არასაკმარისად დასწავლილია.* ინგლისურ ენაში ამას ეწოდება «*underfitting*» — *არასაკმარისად მორგება.*

ეს ცნებები ილუსტრირებულია ნახატზე 2.6.*a* ბინარული კლასიფიკაციის ამოცანისათვის და ნახატზე 2.6.*b* რეგრესიის ამოცანისთვის.

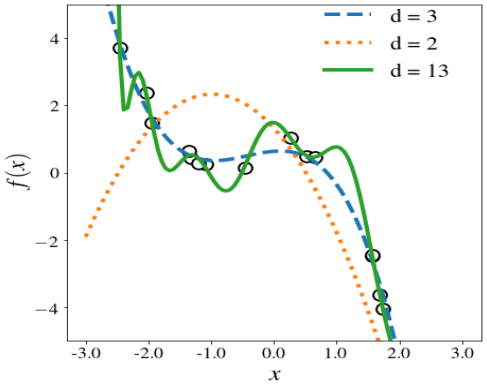
აქ შემდეგი განმარტებების გაკეთება იქნება სასარგებლო :

(a) ნეგატიური $(\times)$ დაკვირვებების გამოსაყოფად პოზიტიური $(+)$ დაკვირვებებისგან წერტილებით შედგენილი წყვეტილი წირით ნაჩვენებია *არასაკმარისად სწავლების* დონე. წითელი ფერის უწყვეტი წირით მოცემული გამყოფი საზღვარი გამორიცხავს შეცდომებს მონაცემებში, მაგრამ, უფრო მეტად სავარაუდოა, *ზედმეტად სწავლებასთან* არის დაკავშირებული. ტირეებით შედგენილი წყვეტილი წირით წარმოდგენილი გამყოფი საზღვარი კარგი *კომპრომისია.*

(b) დაკვირვებათა y ჭდეები, რომლებიც წარმოდგენილია წერტილებით, მიღებული იყო $d=3$ ხარისხის პოლინომისგან. $d=2$ ხარისხის მოდელი მონაცემთა ძალიან ცუდ აპროქსიმაციას იძლევა და *არასაკმარისად დასწავლილია.* მაშინ როცა $d=13$ ხარისხის მოდელი, რომლის ემპირიული რისკი უფრო მცირეა, *ზედმეტად დასწავლილია.*



(a)



(b)

ნახატი 2.6 – არასაკმარისად სწავლება და ზედმეტად სწავლება

5.3 კომპრომისი წანაცვლებასა და დისპერსიას შორის

უკეთ რომ ჩავწვდეთ $f:\mathcal{X}\to\mathcal{Y}$ მოდელის რისკს, ეს რისკი უნდა შევადაროთ მინიმალურ $\mathcal{R}^*$ შეცდომას, რომელიც შეიძლება იყოს მიღწეული ნებისმიერი განზომადი ფუნქციით დაკვირვებათა $\mathcal{X}$ სივრციდან ჭდეთა $\mathcal{Y}$ სივრცეში : ამას *ჭარბი შეცდომა* ეწოდება და იგი წარმოადგენს შეცდომას, რომელიც დასაშვებს აღემატება (ინგლისურად : *excessive error*). ჭარბი შეცდომის დეკომპოზიცია (დაშლა) შემდეგნაირად შეიძლება :

$$\mathcal{R}(f)-\mathcal{R}^{*}=\left[\mathcal{R}(f)-\min_{h\in\mathcal{F}}\mathcal{R}(h)\right]+\left[\min_{h\in\mathcal{F}}\mathcal{R}(h)-\mathcal{R}^{*}\right]. \quad (2.4)$$

პირველი წევრი, $\mathcal{R}(f)-\min_{h\in\mathcal{F}}\mathcal{R}(h)$, განსაზღვრავს მანძილს f მოდელსა და ოპტიმალურ მოდელს შორის ვერსიათა $\mathcal{F}$ სივრცეზე. ამას *შეფასების შეცდომა* ეწოდება.

მეორე წევრი, $\min_{h\in\mathcal{F}}\mathcal{R}(h)-\mathcal{R}^{*}$, განსაზღვრავს ოპტიმალური მოდელის ხარისხს $\mathcal{F}$-ზე, სხვანაირად რომ ვთქვათ, ჰიპოთეზათა სივრცის არჩევის ხარისხს. ამას *აპროქსიმაციის შეცდომა* ეწოდება. თუ $\mathcal{F}$ არის განზომად ფუნქციათა სიმრავლე, მაშინ აპროქსიმაციის შეცდომა ნულს უდრის.

ამრიგად, გამოსახულება 2.4 შეცდომის დაშლის საშუალებას იძლევა წევრად, რომელიც ჩნდება ჰიპოთეზათა სივრცის ხარისხის გამო, და წევრად, რომელიც ჩნდება ოპტიმიზაციის გამოყენებული პროცედურის გამო. პრაქტიკაში, განსაკუთრებულ შემთხვევათა გამოკლებით, როცა ამის შესაძლებლობას კონსტრუქცია იძლევა, შეცდომათა ხსენებული წევრების გამოთვლა შეუძლებელია. მაგრამ $\mathcal{R}(f)-\mathcal{R}^{*}$ სხვაობის ასეთი ფორმით ჩაწერა საშუალებას გვაძლევს გავიგოთ შემდეგი პრობლემა : ჰიპოთეზათა უფრო დიდი სივრცის არჩევა, ჩვეულებრივ, ამცირებს აპროქსიმაციის შეცდომას, ვინაიდან რეალობასთან უფრო ახლოს მყოფ მოდელს გაცილებით მეტი სავარაუდო შესაძლებლობაც აქვს იმყოფებოდეს ამ სივრცეში. მაგრამ, ვინაიდან ეს სივრცე ფართოა, ოპტიმალური გადაწყვეტილების პოვნაც მასში ასევე, როგორც წესი, რთულდება : შეფასების შეცდომა იზრდება. სწორედ ასეთ შემთხვევაში ხდება ზედმეტად სწავლება.

ჰიპოთეზათა უფრო ფართო სივრცე, ჩვეულებრივ, უფრო რთული მოდელების აგების საშუალებასაც იძლევა : მაგალითად, წრფეთა სიმრავლე მე-9 ხარისხის მრვალწევრთა სიმრავლის წინააღმდეგ (იხ. ნახ.2.6b). ეს *ოკამის საპარსის* (ინგლისურად : *Ockham's razor*) პრინციპის ვარიანტია, რომლის თანახმად ყველაზე მარტივი დაშვებები ყველაზე დამაჯერებელიც არის.

ფრანცისკანელი მონაზვნის, ყველა დროის უდიდესი ლოგიკოსის უილიამ ოკამის (William of Ockham, ასევე Occam, ლათინურიდან: Gulielmus Occamus) საპარსი (დანა-სამართებელი) მეთოდოლოგიური პრინციპია, რომელსაც ლათინურად ასახავს სიტყვები : «Entia non sunt multiplicanda praeter necessitatem». ქართულად კი ეს ნიშნავს : «არაფერი უნდა მრავლდებოდეს იმაზე მეტად, რაც საჭიროა».

სხვანაირად რომ ვთქვათ, ეს არის თეორიის აგების ან შეფასების პრინციპი, რომლის მიხედვითაც, სხვა თანაბარ პირობებში, ახსნა-განმარტებები, რომლებიც მოიცავს ნაკლები რაოდენობის სუბსტანციას ან სუბსტანციათა ნაკლები რაოდენობის ნაირსახეობას, უფრო სასურველია, ვიდრე ახსნა-განმარტებები, რომლებიც მოითხოვს მათ გაცილებით მეტი რაოდენობით.

ამიტომ არსებობს კომპრომისი აპროქსიმაციის შეცდომასა და შეფასების შეცდომას შორის : ძნელია შეამცირო ერთ-ერთი მათ შორის ისე, რომ არ გაიზარდოს მეორე. ამ კომპრომისს, ჩვეულებრივ, უწოდებენ *წანაცვლების* (*გადახრის*)*–დისპერსიის კომპრომისს* (ინგლისურად : *bias-variance trade-off*) : აპროქსიმაციის შეცდომა შეესაბამება სწავლების პროცედურის

წანაცვლებას (*გადახრას*), ხოლო შეფასების შეცდომა — მის *დისპერსიას*. ეს კომპრომისი შეიძლება ვიპოვოთ პარამეტრების ბაიესურ შეფასებაში (იხილე პუნქტი 4.3.3).

მაგალითი

განვიხილოთ რეგრესიის ამოცანისთვის გულუბრყვილო ჰიპოთეზათა სივრცე, რომელიც შეიცავს მხოლოდ კონსტანტურ ფუნქციებს. დავუშვათ, რომ ჭდეთა გენერირება ხდება ნორმალური განაწილებით, რომლის ცენტრია a წერტილზე. როგორიც არ უნდა იყოს დანაკვირვები მონაცემები, სწავლების პროცედურა ააგებს მოდელს, რომელიც დააბრუნებს a მნიშვნელობას, როგორიც არ უნდა იყოს შესაბამისი დაკვირვება : პროცედურის *დისპერსია* მონაცემთა სიმრავლის მიმართ ძალიან მცირეა. პირიქით, ვინაიდან წინასწარმეტყველების დასწავლილი ფუნქცია ძალზე ნაკლებად მგრძნობიარეა მონაცემთა კრებულის მიმართ, არსებობს მნიშვნელოვანი წანაცვლება (გადახრა), რაც იწვევს ისეთი პრედიქტორების აგებას, რომლებიც გვიბრუნებს a მნიშვნელობას ყველა დაკვირვებისათვის.

5.4 რეგულარიზაცია

რაც უფრო მარტივია მოდელი, მით ნაკლებია ზედმეტად სწავლების ალბათობა. ზედმეტად სწავლების რისკის შესამცირებლად სასურველია მოდელის სირთულის შეზღუდვა. ეს კეთდება *რეგულარიზაციით* — მეთოდით, რომლის უფრო დაწვრილებით განხილვას დავიწყებთ ამ კურსში მეექვსე ლექციიდან. აქ კი საკმარისი იქნება შემოვიფარგლოთ მოკლე მითითებით, რომ იგი მდგომარეობს წევრისადმი შეცდომით, რომლის მინიმიზაციას ვცდილობთ, პრობლემის სირთულის ამსახველი წევრის დამატებაში (წინა შემთხვევისათვის, მაგალითად, ეს შეიძლება იყოს პოლინომის ხარისხი ან მოდელის კოეფიციენტთა რაოდენობა). მაშასადამე, რთული მოდელი მცირე ემპირიული შეცდომით შეიძლება არამომგებიან მდგომარეობაში აღმოჩნდეს უფრო მარტივ მოდელთან შედარებით, მაშინაც კი, თუ ამ უკანასკნელს უფრო მაღალი ემპირიული შეცდომა აქვს.

6 საკვანძო მომენტები

1. კონტროლირებადი სწავლების ალგორითმს სამი მდგენელი აქვს და ესენია :

   - ჰიპოთეზათა (დაშვებათა) სივრცე,
   - ღირებულების (ხარჯების) ფუნქცია,
   - ოპტიმიზაციის ალგორითმი, რომელიც მონაცემებზე ღირებულების (გასაწევი ხარჯების) ფუნქციის აზრით ოპტიმალური ჰიპოთეზის პოვნის საშუალებას იძლევა (ემპირიული რისკის მინიმიზაცია).

2. კომპრომისი წანაცვლებასა და დისპერსიას შორის (ინგლისურად : *bias-variance trade-off*) ასახავს კომპრომისს სწავლების ალგორითმის ხარვეზით გამოწვეულ *აპროქსიმაციის შეცდომასა* და იმ *შეფასების შეცდომას* შორის, რომელიც მოდელის პარამეტრთა მნიშვნელობების დისპერსიას შეესაბამება.

სტატისტიკასა და მანქანურ სწავლებაში *კომპრომისი* (*დილემა*) *წანაცვლება–დისპერსია* შეცდომათა ორი წყაროს ერთდროულად მინიმიზაციის პრობლემაა, როცა ეს შეცდომები კონტროლირებადი სწავლების ალგორითმებს მონაცემთა განზოგადების საშუალებას არ აძლევს საწვრთნელი ანარჩევის მიღმა :

წანაცვლება – შეცდომაა, რომელიც ჩნდება მცდარი დაშვებების გამო სწავლების ალგორითმში. დიდი ცდომილება შეიძლება იყოს დაკავშირებული ალგორითმთან, რომელშიც წარმოდგენილი არ არის შესაბამისი კავშირები შემავალ მონაცემებსა და მოსალოდნელ შედეგებს შორის (არასაკმარისად სწავლება).

დისპერსია – შეცდომაა, რომელიც, განპირობებულია მგრძნობელობით მცირე რხევების მიმართ საწვრთნელ ანარჩევში. მაღალი დისპერსია შეიძლება ზედმეტად სწავლების მიზეზი აღმოჩნდეს, როცა საწვრთნელ მონაცემებში ხდება შემთხვევითი ხმაურის — და არა მოსალოდნელი შედეგების — მოდელირება.

3. განზოგადება და ზედმეტად სწავლება ძირითადი პრობლემებია მანქანურ სწავლებაში : როგორ უზრუნველვყოთ, რომ მოდელი. რომელიც ნასწავლია წინასწარმეტყველებას დანაკვირვებ მონაცემებზე შეცდომის მინიმალური რისკით, ჩვენთვის საინტერესო პროგნოზების შემძლე სხვა მონაცემებზეც აღმოჩნდეს განზოგადების საკმარისი უნარით ?

დამატებითი ინფორმაცია

- მოდელის სირთულის ცნება ვლადიმირ ვაპნიკის (Vladimir Vapnik) და ალექსეი ჩერვონენკისის (Alexey Chervonenkis) მიერ იყო ფორმალიზებული 1970-იან წლებში და დაწვრილებით არის აღწერილი, მაგალითად, ვაპნიკის ნაშრომში (1995).
- დამატებითი ინფორმაციის მისაღებად სწავლების თეორიის შესახებ შეიძლება Kearns and Vazirani (1994) წიგნის გამოყენება.
- კომპრომისის დაწვრილებითი ანალიზი წანაცვლებასა და დისპერსიას შორის მოიპოვება ნაშრომში Friedman (1997).

7 ბიბლიოგრაფია

1. Crammer, K. and Singer, Y. (2001). On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2 :265–292.

2. Friedman, J. H. (1997). On bias, variance, 0/1-loss and the curse of dimensionality. *Data Mining and Knowledge Discovery*, 1 :55–77.

3. Kearns, M. J. et Vazirani, U. V. (1994). *An Introduction to Computational Learning Theory*. MIT Press, Cambridge, MA.

4. Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer, New York.

5. Weston, J. and Watkins, C. (1999). Support vector machines for multi-class pattern recognition. In *European Symposium on Artificial Neural Networks*.

8 სავარჯიშოები

2.1 რა უპირატესობები და ნაკლოვანებები აქვს მარტივ მოდელს რთულ მოდელთან შედარებით?

2.2 ადელს უნდა ავტომატურად გაარკვიოს : არდადეგებთან დაკავშირებული მისი ფოტოებიდან რომელი იყო გადაღებული შიგნით, გარეთ, დღისით თუ ღამით, რა იქნება

უკეთესი : გამოიყენოს მან ოთხო ბინარული კლასიფიკატორი, 2 ბინარული კლასიფიკატორი თუ ერთი მრავალკლასიანი კლასიფიკატორი ?

2.3 ბილმა ასწავლა მოდელს კლასიფიცირება 1 000 დაკვირვებაზე. მისი მოდელი ამ დაკვირვებებიდან შეცდომას უშვებს 10-ზე. იმყოფება ეს მოდელი ნაკლებად დანასწავლ ან ზედმეტად დანასწავლ მდგომარეობაში, თუ ამ ორიდან არც ერთში და არც მეორეში ?

2.4 კლერს ბინარული კლასიფიკაციის ერთი ამოცანა აქვს. მისი მონაცემები წარმოდგენილია ორ განზომილებაში. იგი განიხილავს ყველა წრის ერთობლიობას როგორც ჰიპოთეზათა სივრცეს. რამდენი პარამეტრი უნდა განსაზღვროს მან? როგორ შეუძლია კლერს მათი დადგენა?

2.5 დენი ცდილობს ბინარული კლასიფიკაციის ამოცანის გადაჭრას მონაცემებისათვის ორ განზომილებაში. იგი განიხილავს ჰიპოთეზათა სივრცედ m მართკუთხედის გაერთიანებათა (კავშირების) ერთობლიობას (სიმრავლეს). რაში მდგომარეობს ჰიპოთეზათა ამ სივრცის უპირატესობები ? თუ არსებობს ზედმეტად სწავლების რისკი ?

2.6 ბინარული კლასიფიკაციის ამოცანისთვის წარმოადგინეთ (გრაფიკულად ააგეთ) ერთი დადებითი დაკვირვების შემთხვევაში : A. შეცდომა 0/1, B. კვადრატული შეცდომა, C. სახსრული შეცდომა, და D. ჯვარედინი ენტროპია როგორც ამ დაკვირვებასთან დაკავშირებული გადაწყვეტილების მიღების ფუნქციის მნიშვნელობაზე დამოკიდებულება. იგივე გააკეთეთ უარყოფითი დაკვირვებისთვის.

2.7 კროს-ენტროპია (ჯვარედინი ენტროპია) - დანახარჯის ზომაა ალბათობის ორ განაწილებას შორის. რას ნიშნავს ეს ? ენტროპია შემთხვევითობის ზომაა. მაგალითად, თუ ჩვენ განვიხილავთ სამ ტევადობას მყარი ტანით, სითხით და აირით, მაშინ ენტროპიები შემდეგი სახით ჩაიწერება : $E(\text{აირი}) > E(\text{სითხე}) > E(\text{მყარი ტანი})$. სხვა სიტყვებით რომ ვთქვათ, რამდენადაც შემთხვევითია მონაცემები, იმდენადვე ენტროპია უფრო დიდი აღმოჩნდება. ახლა კი იმის თქმაც შეიძლება, რომ კროს-ენტროპია (ჯვარედინი ენტროპია) - ეს უბრალოდ სხვაობაა ორ ან უფრო მეტ ენტროპიას შორის.

ბინარული ჯვარედინი ენტროპია წარმოადგენს ღირებულებას, დანახარჯს, დანაკარგს კლასიფიკაციის ამოცანებში, სადაც არის *ორი კატეგორია* ან *კლასი*. განტოლება შეიძლება იყოს მოცემული შემდეგი სახით :

$$loss = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i).$$

აქ N - ანარჩევების, ანუ მონაცემთა წერტილების საერთო რაოდენობაა, y - მოსალოდნელი გამოსასვლელი, $\hat{y}$ - ნაწინასწარმეტყველი გამოსასვლელი.

კლასიფიკაციისათვის ღირებულებად, დანახარჯად, დანაკარგად კატეგორიალური კროს-ენტროპიის (ჯვარედინი ენტროპიის) არჩევას არა ერთი მიზეზი აქვს. ვინაიდან კროს-ენტროპია ხშირად გამოიყენება გალოგარითმებით, ეს ფუნქცია ასევე ცნობილია log loss (*ლოგარითმული დანაკარგის*) სახელითაც.

ბინარული კლასიფიკაციისათვის წარმოადგინეთ და გაანალიზეთ კროს-ენტროპიული დანაკარგის (log loss ფუნქციის) გრაფიკები $y = 0$ და $y = 1$. პირობებში. აბსცისათა ღერძზე გადაზომეთ ალგორითმის მიერ y მნიშვნელობის ნაწინასწარმეტყველი p ალბათობის სიდიდე.

2.8 ხმაური შეცდომათა ერთ-ერთი წყაროა სწავლების პროცესში. ასე რომ ელისო, ბუნებრივია, ცდილობს აბერანტული (ნორმალური მდგომარეობიდან, აგებულებიდან ან განლაგებიდან გადახრილი) დაკვირვებების დეტექტირებას (გამოვლენას, პოვნას) თავის მონაცემთა მასივში (კრებულში). ხომ არ შეგიძლიათ შესთავაზოთ მას რაიმე პროცედურა ?

სავარჯიშოთა ამონახსნები

2.1 უპირატესობები : ზედმეტი სწავლების შეუძლებლობა ; უკეთესი განზოგადება ; გამოთვლის ნაკლები დრო. ნაკლოვანებები : თუ მოდელი ძალიან მარტივია, მახასიათებლე-ბიც ცუდია.

2.2 ვინაიდან ფოტოგადაღებათა კლასიფიკაციები შიგნით/გარეთ და დღისით/ღამით არ არის ურთიერთგამომრიცხავი, უკეთესია ორი ბინარული კლასიფიკატორის გამოყენება (ერთი - შიგნით/გარეთ წყვილისათვის, ხოლო მეორე - დღისით/ღამით წყვილისათვის).

2.3 ამის ცოდნა შეუძლებელია ესოდენ მწირი ინფორმაციის საფუძველზე.

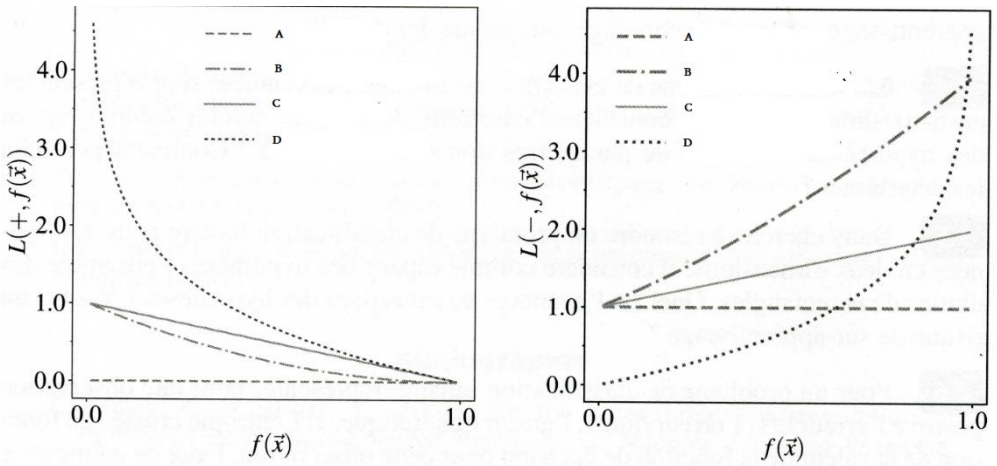
2.4

$$\mathcal{F} = \left\{ \vec{x} \mapsto \left(x_1 - a\right)^2 + \left(x_2 - b\right)^2 - r^2 \right\}.$$

დასადგენია სამი პარამეტრი : a. b და r. ამ პარამეტრების დასადგენად საჭიროა რომელიმე *ღირებულების ფუნქციის* არჩევა, მაგალითად, 0/1 დანაკარგის, და ემპირიული რისკის მინიმიზაციის განხორციელება.

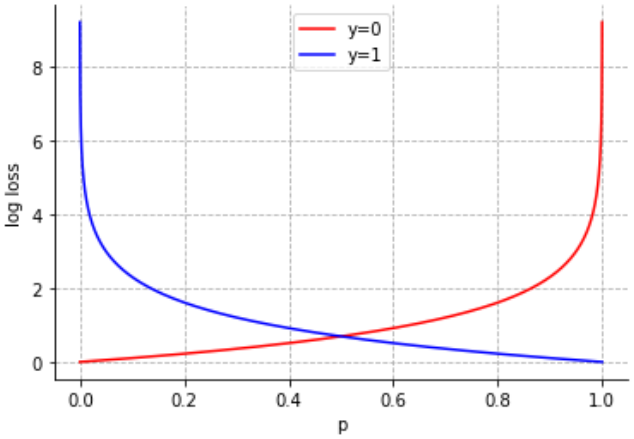
2.5 ჰიპოთეზათა ეს სივრცე ძალიან კარგი მოდელირების საშუალებას იძლევა, როცა m საკმარისად დიდია. ემპირიული რისკის მინიმიზაცია ყოველთვის შესაძლებელი იქნება. მაგრამ დიდი m -ის შემთხვევაში იზრდება ზედმეტად სწავლების საშიშროება.

2.6 გრაფიკები მოცემულია ქვემოთ და ასახვას ორივე შემთხვევას ოთხივე დამოკიდებულების-თვის.



ნახატი 2.7 გრაფიკი მოცემულია ქვემოთ და ასახვას ბინარული კროს-ენტროპიული დანახარჯების ქცევას.

თუ $y = 0$ მნიშვნელობას განვიხილავთ როგორც მოსალოდნელ გამოსასვლელს, მაშინ *log loss* დანახარჯები ნაკლებია მცირე p ალბათობისათვის, მაგრამ p ალბათობის გაზრდასთან ერთად *log loss* დანახარჯებიც სწრაფად იზრდება. აქ p წარმოადგენს ალგორითმის მიერ ნაწინასწარმეტყველ მნიშვნელობას ბინარული კლასიფიკატორის y გამოსასვლელისათვის.



p ალბათობის მცირე მნიშვნელობებისათვის გრადიენტები მცირეა, ხოლო p ალბათობის დიდი მნიშვნელობებისათვის გრადიენტები დიდია. (წითელ მრუდეზე არგუმენტის ერთსა და იმავე ნაზრდს ფუნქციის მკვეთრად განსხვავებული ნაზრდები შეესაბამება p ალბათობის მცირე და დიდი მნიშვნელობის არეში).

თუ ალგორითმი მცდარ წინასწარმეტყველებას აკეთებს და p ალბათობის მაღალ მნიშვნელობას იძლევა, მაშინ დანახარჯების ფუნქციამ უნდა დასაჯოს ასეთი ალგორითმი. სწორედ რომ ამას აკეთებს *log loss* ფუნქცია.

2.8 აბერანტული (ნორმალური მდგომარეობიდან, აგებულებიდან ან განლაგებიდან უკიდურესად გადახრილი) დაკვირვებების კორექტულად მონიშვნა ძნელია, ვინაიდან ისინი არ ემორჩილება იმავე კანონებს, რომლებიც სამართლიანია ყველა სხვა დაკვირვებისათვის. ასე რომ ელისოს შეუძლია ყველაზე დიდი ემპირიული რისკის მქონე დაკვირვებებით დაინტე-რესდეს.

ლექცია 3 მოდელის არჩევა და შეფასება

შინაარსი

მეორე ლექციაში კონტროლირებადი სწავლება ჩვენ მიერ იყო ფორმალიზებული როგორც მოდელის ძებნა, რომლის ემპირიული შეცდომა მინიმალურია დაკვირვებათა მოცემულ კრებულზე. მაგრამ ამ ემპირიული შეცდომის მინიმიზაცია არ იძლევა მოდელის შეცდომის მინიმიზაციის გარანტიას მონაცემთა მთელ სივრცეზე. მართლაც, ზედმეტად სწავლების მდგომარეობაში მოდელის შეცდომა (და აქედან გამომდინარე საფრთხე) არ იქნება სწორად შეფასებული. არადა სწორედ ეს შეცდომა, ან, სხვა სიტყვებით, ჩვენი უნარი ვიწინასწარმეტყველოთ უცნობი მოვლენები, წარმოადგენს განსაკუთრებულ ინტერესს. ამ ლექციაში საუბარი გვექნება იმის შესახებ, როგორ შეიქმნას ექსპერიმენტული სქემა, რომელიც მოდელის შეფასების საშუალებას გვაძლევს ისე, რომ ამასთან ერთად აცილებული იყოს ზედმეტად სწავლებასთან დაკავშირებული სისტემატური შეცდომა — ე.წ. წანაცვლება. ამ ჭრილში ყოველთვის მკაცრად მოხდება გარჩევა მოდელის *შეფასებასა* და მის *არჩევას* შორის : თუ პირველი გულისხმობს მოდელის ეფექტურობის, მუშობის ხარისხის განსაზღვრას მონაცემთა მთელ სივრცეზე, მეორე მდგომარეობს საუკეთესო მოდელის არჩევაში რამდენიმე ნიმუშიდან.

მიზნები

— ექსპერიმენტული სქემის დამუშავება, რომლის საფუძველზე შესაძლებელია კონტროლირებადი სწავლების მოდელის არჩევა ;

— კონტროლირებადი სწავლების მოდელისთვის შეფასების ერთი ან რამდენიმე კრიტერიუმის არჩევა ;

— კონტროლირებადი სწავლების მოდელის *განზოგადების* ეფექტურობის შეფასება.

დევიდ ვოლპერტისა (D.H. Wolpert) და უილიამ მაკრედის (W.G. Macready) თეორემა (1997) უფასო სადილების (ლანჩების) არარსებობის შესახებ («*NFL - No Free Lunch*»), რომელიც მანქანური სწავლების თანამედროვე ფორმათა ინდუქციის პრობლემას პირველად აყენებს სამეცნიერო ლიტერატურაში, გვეუბნება, რომ მანქანური სწავლების არც ერთ ალგორითმს არ შეუძლია კარგად მუშაობა სწავლების ყველა პრობლემისთვის : ალგორითმი, რომელიც კარგად მუშაობს პრობლემათა ერთ ტიპზე, ამას აკომპენსირებს იმით, რომ ნაკლებად ეფექტურად მუშაობს სხვა ტიპის პრობლემებზე. სხვანაირად რომ ვთქვათ, არ არსებობს *«ჯადოსნური ჯოხი»*, რომელსაც შეუძლია ყველა ჩვენი პრობლემის გადაჭრა მანქანურ სწავლებასთან დაკავშირებით და ამიტომ კონკრეტული ამოცანისათვის აუცილებელია რამდენიმე შესაძლებლობის შემოწმება ოპტიმალური მოდელის ასარჩევად. ურიგო არ იქნება იმის აღნიშვნაც, რომ ამ არჩევანში რამდენიმე კრიტერიუმის ჩარევაც შეიძლება ხდებოდეს : არა მხოლოდ პროგნოზების ხარისხთან დაკავშრებით, რაც ამ ლექციაში გვაინტერესებს, არამედ აუცილებელი გამოთვლითი რესურსების შესახებაც. სწორედ ამ რესურსებმა შეიძლება იჩინოს თავი შემზღუდველ ფაქტორად პრაქტიკაში.

1 განზოგადების შეცდომის ემპირიული შეფასება

ემპირიული შეცდომა, რომელიც გაზომილია მოდელის ასაგებად გამოყენებულ დაკვირვებებზე, *მოდელის შეცდომის*, ან *განზოგადების შეცდომის* ცუდი შეფასებაა შესაძლო მონაცემთა სიმრავლეზე : თუ მოდელის ზედმეტად სწავლება მოხდა, ეს ემპირიული შეცდომა შეიძლება ახლოს იყოს ნულთან ან უდრიდეს კიდეც მას, მაშინ როცა განზოგადების შეცდომა შეიძლება რაგინდ დიდი იყოს.

1.1 სატესტო ნაკრები

ამიტომ ძალზე მნიშვნელოვანია მოდელის შესაფასებლად მონიშნული მონაცემების გამოყენება, რომლებიც არ ყოფილა ნახმარი მის ასაგებად. ამის განსახორციელებლად ყველაზე მარტივი ხერხია დაკვირვებათა ნაწილის გადადება მოდელის შესაფასებლად და მხოლოდ დანარჩენ მონაცემთა გამოყენება მის ასაგებად.

განსაზღვრება 3.1 (სატესტო ნაკრები, საწვრთნელი ნაკრები) დავუშვათ, რომ გვაქვს მონაცემთა $\mathcal{D} = \left\{\left(\vec{x}^{\,i},\, y^i\right)\right\}_{i=1,\ldots,n}$ ნაკრები, რომელიც გაყოფილია ორ — $\mathcal{D}_{\text{tr}}$ და $\mathcal{D}_{\text{te}}$ — ნაკრებად, მაშინ პრედიქტიული (მაპროგნოზებელი) მოდელის საწვრთნელად (სწავლებისთვის) გამოსაყენებელ $\mathcal{D}_{\text{tr}}$ სიმრავლეს *საწვრთნელი ნაკრები* (ინგლისურად : *training set*) ეწოდება, ხოლო ამ მოდელის შესაფასებლად განკუთვნილ $\mathcal{D}_{\text{te}}$ სიმრავლეს — *სატესტო ნაკრები* (ინგლისურად : *test set*).

ვინაიდან მოდელის გასაწვრთნელად სატესტო ნაკრები გამოყენებული არ ყოფილა, შესაძლებელია მისი განხილვა *«ახალ»* მონაცემთა ნაკრებად. ამ სატესტო ნაკრებზე გამოანგარიშებული დანაკარგები არის განზოგადების შეცდომის შეფასება.

1.2 შემოწმების ნაკრები

ახლა განვიხილოთ ვითარება, როცა გასაკეთებელია არჩევანი K მოდელს შორის. უპირველეს ყოვლისა, შეგვიძლია ჩავუტაროთ სწავლება თითოეულ მოდელს საწვრთნელ მონაცემთა

ნაკრებზე, რათა ამ გზით დავადგინოთ გადაწყვეტილების მიღების K ფუნქცია $f_1, f_2, \ldots, f_K$, რის შემდეგ ყოველი ამ მოდელისთვის უნდა გამოითვალოს შეცდომა სატესტო ნაკრებზე. დაბოლოს, შეგვიძლია ავირჩიოთ მოდელად ის ნიმუში, რომელსაც უმცირესი შეცდომა აქვს სატესტო ნაკრებზე :

$$\hat{f} = \arg\min_{k=1,\ldots,K} \frac{1}{\left|\mathcal{D}_{\text{te}}\right|} \sum_{\vec{x},y\in\mathcal{D}_{\text{te}}} L\left(y, f_k\left(\vec{x}\right)\right). \tag{3.1}$$

მაგრამ როგორია მისი განზოგადების შეცდომა ? ვინაიდან მოდელის ასარჩევად ჩვენ გამოვიყენეთ $\mathcal{D}_{\text{te}}$ ნაკრები, იგი არ წარმოადგენს უკვე დამოუკიდებელ ნაკრებს, შედგენილს ახალი მონაცემებით, რომლებიც მოდელის განსაზღვრისთვის გამოყენებული არ ყოფილა.

ასე რომ. პრობლემის გადაჭრა მდგომარეობს მონაცემთა ჩვენი ნაკრების გაყოფაში სამ ნაწილად :

— $\mathcal{D}_{\text{tr}}$ საწვრთნელი ნაკრები(ინგლისურად : *training set*) , რომელზეც შევძლებთ ჩვენი K სწავლების ალგორითმის გაწვრთნას ;

— $\mathcal{D}_{\text{val}}$ შემოწმების (სავალიდაციო) ნაკრები (ინგლისურად : *Validation set*), რომელზეც ჩვენ შევაფასებთ ასეთნაირად მიღებულ K მოდელს საბოლოო მოდელის შესარჩევად, ზოგიერთ კონტექსტში შემოწმების ნაკრებს ასევე დაპროექტების ნაკრებსაც უწოდებენ — ინგლისურად : *Development (Dev) set* ;

— $\mathcal{D}_{\text{te}}$ სატესტო ნაკრები (ინგლისურად: *test set*), რომელზეც ბოლოს და ბოლოს შევაფასებთ შერჩეული მოდელის განზოგადების შეცდომას.

აქ ჩანს, რომ მნიშვნელოვანია განვასხვავოთ ერთმანეთისგან მოდელის *არჩევა* და მისი *შეფასება* : ორივე მოქმედების განხორციელებამ ერთსა და იმავე მონაცემებზე შეიძლება მიგვიყვანოს განზოგადების შეცდომის მცდარად შემცირებულ შეფასებამდე და არჩეული მოდელის ზედმეტად სწავლამდე. ასე რომ, სწავლება, ტესტირება და შემოწმება — საკვანძო ეტაპებია მანქანური სწავლების სამუშაო პროცესში.

შენიშვნა

მოდელის არჩევის შემდეგ მას შეიძლება განმეორებით ჩაუტარდეს სწავლება საწვრთნელი ნაკრებისა და შემოწმების ნაკრების გაერთიანებულ სიმრავლეზე საბოლოო მოდელის ასაგებად.

სავალიდაციო (შემოწმების) და სატესტო ნაკრებთა გამოყენება *ზრდის მოდელის განზოგადების უნარს ახალ უხილავ მონაცემებზე.* თუმცა საყურადღებოა, რომ სავალიდაციო ნაკრები ჭარბია და საჭირო არ არის, როცა მოდელის მორგება მისი სხვადასხვა პარამეტრის მოსინჯვით არ ხდება.

1.3 ჯვარედინი შემოწმება

მონაცემთა ნაკრების გაყოფა მწვრთნელ და სატესტო ნაკრებებად რაიმე წესებით შეზღუდული არ არის. ამიტომ ჩვენ ვრისკავთ შემთხვევით შევქმნათ არარეპრეზენტატულ მონაცემთა

ნაკრებები. ამ წყალქვეშა ლოდს რომ გვერდი ავუაროთ, სასურველია პროცედურის გამეორება რამდენჯერმე და შემდეგ მიღებული შედეგების გასაშუალოება, რათა მოხდეს შემთხვევითი ეფექტების დაგლუვება. ყველაზე ტრადიციული მიდგომა ასეთ ვითარებაში *ჯვარედინი შემოწმებაა*, რომელიც ნაჩვენებია ნახატზე 3.1.

განსაზღვრება 3.2 (ჯვარედინი შემოწმება) დავუშვათ, მოცემულია n დაკვირვების $\mathcal{D}$ ნაკრები და რაღაც K რიცხვი, მაშინ *ჯვარედინი შემოწმება* (სხვანაირად, *კროს-ვალიდაცია, მოსრიალე კონტროლი* ; ინგლისურად : *cross-validation*) ეწოდება პროცედურას, რომელიც მდგომარეობს იმაში, რომ :

1. გაიყოს $\mathcal{D}$ ნაკრები, დაახლოებით ერთნაირი ზომის, K ნაწილად $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_K$,

2. ყოველი $k = 1, \ldots, K$ მნიშვნელობისთვის

— გაიწვრთნას მოდელი $\bigcup_{l \neq k} \mathcal{D}_l$ – ზე,

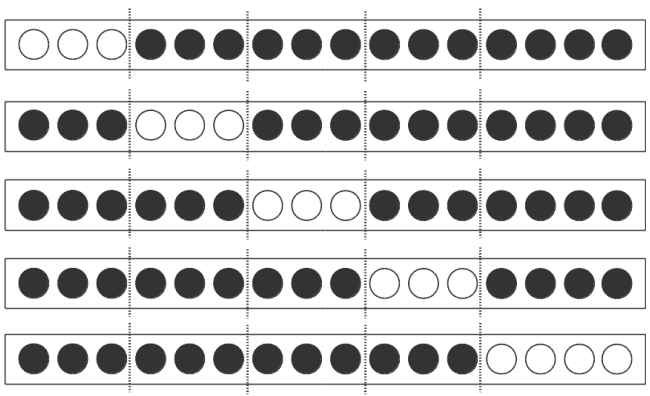
— შეფასდეს ეს მოდელი $\mathcal{D}_K$ – ზე.

$\mathcal{D}$ ნაკრების ყოველ დანაყოფს ორ — $\mathcal{D}_K$ და $\bigcup_{l \neq k} \mathcal{D}_l$ — სიმრავლედ ჯვარედინი შემოწმების (კროს-ვალიდაციის, მოსრიალე კონტროლის) *ნაკეცი* (ინგლისურად : *fold*) ან ზოგჯერ *ბლოკი* ეწოდება.

$\mathcal{D}$ ნაკრების ყოველი მონიშნული დაკვირვება მიეკუთვნება ერთადერთ სატესტო ნაკრებს და $(K-1)$ საწვრთნელ ნაკრებს. ასე რომ, ეს პროცედურა უზრუნველყოფს ერთ პროგნოზს დაკვირვებაზე $\mathcal{D}$ ნაკრებიდან. დასკვნის გამოსატანად მოდელის მუშობის ხარისხის და ეფექტურობის შესახებ, შეიძლება :

— ან შეფასდეს პროგნოზების ხარისხი $\mathcal{D}$ ნაკრებზე ;

— ან შეფასდეს K პრედიქტორიდან თითოეულის ხარისხი შესაბამის $\mathcal{D}_K$ სატესტო ნაკრებზე და განხორციელდეს მათი ეფექტურობების მაჩვენებლების გასაშუალოება. ეს მეორე მიდგომა ასევე ხსენებული მაჩვენებლების სტანდარტული გადახრის (საშუალო კვადრატული გადახრის) გაგების საშუალებას იძლევა, რაც უკეთეს წარმოდგენას გვიქმნის წინასწარმეტყველებათა ხარისხის ცვალებადობის ფუნქციური დამოკიდებულების კავშირზე საწვრთნელ მონაცემებთან.



ნახატი 3.1 - კროს-ვალიდაცია 5 ნაკეცით : ყოველი დაკვირვება ეკუთვნის ერთ-ერთს შემოწმების 5 ნაკრებიდან (თეთრი ფერი) და 4 დანარჩენ საწვრთნელ ნაკრებს (შავი ფერი).

სტრატიფიკაცია

განსაზღვრება 3.3 (სტრატიფიცირებული ჯვარედინი შემოწმება) დავუშვათ, რომ მოცემულია K რაოდენობის $\mathcal{D}_K$ ქვესიმრავლე. ითვლება, რომ ჯვარედინი შემოწმება *სტრატიფიცირებულია*, თუ დაკვირვებათა ჭდეების საშუალო მნიშვნელობა, დაახლოებით, ერთნაირია თითოეულ ამ ქვესიმრავლეში :

$$\frac{1}{|\mathcal{D}_1|}\sum_{i\in\mathcal{D}_1} y^i \approx \frac{1}{|\mathcal{D}_2|}\sum_{i\in\mathcal{D}_2} y^i \approx \cdots \approx \frac{1}{|\mathcal{D}_K|}\sum_{i\in\mathcal{D}_K} y^i \approx \frac{1}{|\mathcal{D}|}\sum_{i\in\mathcal{D}} y^i \,.$$

კლასიფიკაციის ამოცანის შემთხვევაში ეს ნიშნავს, რომ ყოველი კლასის მაგალითების წილი ერთნაირია ნებისმიერ ცალკეულ $\mathcal{D}_K$ ნაკრებზე. ასე რომ, აღნიშნული წილი ისეთივეა, როგორც მონაცემთა სრულ $\mathcal{D}$ ნაკრებზე.

ამ პროცედურის ინტერესია უზრუნველყოს ის, რომ დაკვირვებათა განაწილება ყოველი $\mathcal{D}_K$ ნაკრების შიგნით ისეთივე იყოს, როგორც მონაცემთა $\mathcal{D}$ ნაკრების შიგნით. წარმოვიდგინოთ, რომ, უიღბლობისა გამო თუ საუბედუროდ, ერთ-ერთი ნაკეცი შეიცავს მხოლოდ დადებით მაგალითებს თავის საწვრთნელ ნაკრებში და მხოლოდ უარყოფით მაგალითებს — თავის სატესტო ნაკრებში : სავსებით შესაძლებელია, რომ ამ ნაკეცზე ნებისმიერი მოდელი გაიწვრთნება იწინასწარმეტყველოს, რომ ყველაფერი დადებითია და, ამრიგად, ძალიან ცუდად იმუშავებს..

კონტროლი (შემოწმება) ცალკეულ ობიექტთა საშუალებით (Leave-one-out)

სწავლების ალგორითმი მით უკეთ დაისწავლის, რაც უფრო მეტი მონაცემია მისაწვდომი საწვრთნელად : რაც უფრო მეტი ჭდეა ცნობილი დაკვირვებებისთვის $\mathcal{X}$ სივრცეში, მით უფრო შესაძლებელია ვაიძულოთ მოდელი პატივისცემა გამოიჩინოს მათ მიმართ. მონაცემთა n მოცულობის ნაკრებისთვის, ჯვარედინი შემოწმების (კროს-ვალიდაციის) სატესტო ნაკრები K ნაკეცით (ბლოკით) შეიცავს $\frac{(K-1)n}{K}$ წერტილს : რაც უფრო მეტია ნაკეცების რიცხვი, მით უკეთ დაისწავლის გაწვრთნილი მოდელები თითოეულ ნაკეცზე, რაც $K=n$ შემთხვევის განხილვისკენ გვიბიძგებს.

განსაზღვრება 3.4 (LOOCV — Leave One Out Cross-Validation) კროს-ვალიდაციას (ჯვარედინ შემოწმებას), რომლის ნაკეცთა რიცხვი დაკვირვებათა რიცხვის ტოლია საწვრთნელ ნაკრებებზე და რომლის ყოველი ნაკეცი შედგება, ამრიგად, $n-1$ მოცულობის საწვრთნელი ნაკრებისგან და 1–ის ტოლი მოცულობის სატესტო ნაკრებისგან, ეწოდება *კონტროლი (შემოწმება) ცალკეულ ობიექტთა საშუალებით* (LOO — *leave one out*) : ყოველი ნაკეცისთვის გამოიყოფა მხოლოდ ერთი მაგალითი.

უფრო მარტივად რომ ვთქვათ, LOOCV (Leave One Out Cross-Validation) – ეს ჯვარედინი შემოწმების (კროს-ვალიდაციის) ნაირსახეობაა, რომლის დროსაც ყოველი დაკვირვება

განიხილება სატესტო ნაკრებად (ინგლისურად : *test set*), ხოლო დანარჩენი $n-1$ დაკვირვება — საწვრთნელ ნაკრებად (ინგლისურად : *training set*). ასე რომ, LOOCV–ში მოდელის მორგება და პროგნოზირება მხოლოდ ერთი დაკვირვების შემცველი სატესტო ნაკრების გამოყენებით ხორციელდება.

შეფასებას «*leave-one-out*» პრინციპით ორი ნაკლი აქვს. ჯერ ერთი, გამოთვლებისთვის იგი დიდ დროს მოითხოვს : n მოდელიდან თითოეულს უნდა ჩაუტარდეს სწავლება $(n-1)$ დაკვირვებაზე, ნაცვლად იმისა, რომ 10 მოდელიდან თითოეული გაიწვრთნას დაკვირვებათა 90 პროცენტზე (მაგალითად, იმ შემთხვევაში, თუ $K=10$). უფრო მეტიც, ასეთი გზით შედგენილი საწვრთნელი ნაკრებები ძალიან ჰგავს ერთმანეთს. ცხადია, ნასწავლი მოდელებიც ასევე იქნება ერთმანეთის მსგავსი და ნაკლებად განსხვავებული მონაცემთა მთელ ნაკრებზე გაწვრთნილი მოდელისგან. მეორე მხრივ, სატესტო ნაკრებები შეუსაბამო იქნება და მაჩვენებლებს შეიძლება მნიშვნელოვანი გაბნევა აღმოაჩნდეს, რაც გაართულებს მათ ინტერპრეტაციას.

1.4 ბუტსტრეპი

მონაცემთა თავიდან, მეორედ არჩევის სხვა ხერხი განზოგადების შეცდომის შესაფასებლად ცნობილია *ბუტსტრეპის* (*bootstrap*) სახელწოდებით.

განსაზღვრება 3.5 (ბუტსტრეპი).დავუშვათ, რომ მოცემულია n დაკვირვებათა $\mathcal{D}$ ნაკრები და რაღაც B რიცხვი. მაშინ *ბუტსტრეპი* (*bootstrap*) ეწოდება პროცედურას, რომელიც მდგომარეობს დაკვირვებათა $\mathcal{D}$ ნაკრებიდან B რაოდენობის $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_B$ ანარჩევთა შექმნაში, სადაც თითოეული აგებულია n მაგალითის ამოღებისას $\mathcal{D}$–დან *დაბრუნებით უკან*. ასე რომ, ყოველი მაგალითი შეიძლება რამდენჯერმე გვხვდებოდეს ან საერთოდ არ გვხვდებოდეს რომელიმე $\mathcal{D}_b$–ში.

ბუტსტრეპი — პროცედურაა, რომელიც ფართოდ გამოიყენება სტატისტიკაში პარამეტრის შესაფასებლად მისი შეფასებების საფუძველზე B რაოდენობის ანარჩევზე. ამ პროცედურის საშუალებით შეიძლება შესაფასებელი მოდელის გაწვრთნა თითოეულ $\mathcal{D}_b$ ანარჩევზე და შემდეგ მისი მუშაობის შეფასება მთელ $\mathcal{D}$ ანარჩევზე. მაგრამ ეს შეფასება წანაცვლებული იქნება მაგალითების ნაწილის მოხვედრის გამო $\mathcal{D}$ ანარჩევიდან $\mathcal{D}_b$ ანარჩევში. ამიტომ აუცილებელია შემოვიფარგლოთ მაგალითებით $\mathcal{D} \setminus \mathcal{D}_b$ *სხვაობიდან*. როგორც ცნობილია, მოცემული $\mathcal{D}$ და $\mathcal{D}_b$ *სიმრავლეთა სხვაობა* ეწოდება ისეთ სიმრავლეს, რომელიც შეიცავს $\mathcal{D}$ სიმრავლის ყველა იმ ელემენტს, რომელიც $\mathcal{D}_b$ სიმრავლეში არ შედის და $\mathcal{D} \setminus \mathcal{D}_b$ სახით ჩაიწერება. მიჩნეულია, რომ პრაქტიკაში ხშირად გამოყენებისთვის ეს პროცედურა ძალიან რთულია.

შენიშვნა

ალბათობა იმისა, რომ $\left(\vec{x}^{\,i}, y^i\right)$ წყვილი გაჩნდება $\mathcal{D}_b$–ში, შეიძლება იყოს გამოთვლილი ისეთი

ხდომილების ალბათობის დამატებად ერთისადმი, რომელშიც $\left(\vec{x}^{\,i},\, y^i\right)$ წყვილი არც ერთხელ არ ამოვა n–ჯერ გათამაშებისას. ხოლო $\left(\vec{x}^{\,i},\, y^i\right)$ წყვილის ერთხელ ამოსვლის ალბათობა $\frac{1}{n}$ სიდიდეს შეადგენს. ასე რომ,

$$\mathbb{P}\left[\left(\vec{x}^{\,i}, y^i\right) \in \mathcal{D}_b\right] = 1 - \left(1 - \frac{1}{n}\right)^n.$$

როცა n–ის მნიშვნელობები დიდია, ეს ალბათობა, დაახლოებით, უდრის $1 - e^{-1} \approx 0.632$ სიდიდეს, ვინაიდან $\left(1 + \frac{x}{n}\right)^n$ გამოსახულების ზღვარი, როცა $n \to +\infty$, შეადგენს e^x–ს და მოცემულ შემთხვევაში $x = -1$.

ასე რომ, $\mathcal{D}_b$ ნაკრები შეიცავს $\mathcal{D}$ სიმრავლიდან დაკვირვებათა თითქმის ორ მესამედს.

2 ჰიპერპარამეტრების ოპტიმიზაცია

ძალიან ზოგადი სახით ლექციის წინა ნაწილში საუბარი იყო არჩევანზე K მოდელს შორის. პრაქტიკაში ეს K მოდელი ხშირად მიიღება დაყრდნობით სწავლების ერთსა და იმავე ალგორითმზე, რომელიც იცვლება *ჰიპერპარამეტრით*.

განსაზღვრება 3.6 (ჰიპერპარამეტრი) *ჰიპერპარამეტრი* ეწოდება სწავლების ალგორითმის პარამეტრს (და არა მოდელის) : იქ, სადაც მოდელის პარამეტრები განსაზღვრულია სწავლების ალგორითმით, ჰიპერპარამეტრი არ შედის მოდელში (თუმცა გავლენას ახდენს მოდელის პარამეტრების მნიშვნელობებზე). შეიძლება, მაგალითად, საქმე ეხებოდეს მეზობლების რაოდენობას უახლოეს მეზობელთა ალგორითმში (იხ. პუნქტი 8.2.1), წრფივი რეგულარიზებული რეგრესიის რეგულარიზაციის კოეფიციენტს (იხ. ლექცია 6) ან ფარულ შრეთა რიცხვს ნეირონულ ქსელში (იხ. ლექცია 7).

იმისათვის. რომ ერთი ან რამდენიმე პარამეტრის ოპტიმალური მნიშვნელობა განისაზღვროს, ხშირად მიმართავენ ჯვარედინ შემოწმებას (კროს-ვალიდაციას), როგორც ეს დეტალურად აღწერილა ქვემოთ. ეს მოითხოვს შესაფასებელ ჰიპერპარამეტრთა K მნიშვნელობათა სიის დადგენას. ამისათვის რამდენიმე სტრატეგიის გამოყენება არის შესაძლებელი.

2.1 ძებნა მესრით

განსაზღვრება 3.7 (ძებნა მესრით) *ძებნა მესრით* (ინგლისურად : *grid search*) — ჰიპერპარამეტრების ოპტიმიზაციის სტრატეგიაა, რომელიც მდგომარეობს, ყველა შესაბამისი პარამეტრისთვის, იმ *მნიშვნელობათა სიის განსაზღვრაში*, რომლებიც აუცილებლად უნდა შემოწმდეს, ხოლო შემდეგ ჰიპერპარამეტრების მნიშვნელობათა ყველა შესაძლო კომბინაციის *განზოგადების შეცდომა* ემპირიულად უნდა შეფასდეს.

იმ შემთხვევაში, როცა ერთადერთი λ პარამეტრის ოპტიმიზაციას ცდილობენ, ჰიპერპარამეტრების «მესერი» უბრალოდ $\lambda_1, \lambda_1, \ldots, \lambda_K$ მნიშვნელობათა სიაა, რომლებიც

მოიცემა λ ჰიპერპარამეტრისთვის. საქმე დადის K სხვადასხვა მოდელის განზოგადების შეცდომის შეფასებამდე.

იმ შემთხვევაში, როცა საჭიროა ორი — λ და ν — ჰიპერპარამეტრის ოპტიმიზაცია, პროცედურა იწყება სატესტო $\lambda_1, \lambda_1, \ldots, \lambda_K$ მნიშვნელობათა სიის ფიქსირებით λ ჰიპერპარამეტრისთვის, ხოლო შემდეგ სატესტო $\nu_1, \nu_1, \ldots, \nu_K$ მნიშვნელობათა სიის ფიქსირებით ν ჰიპერპარამეტრისთვის. ამრიგად, ჩნდება ჰიპერპარამეტრთა (λ_k, ν_l) წყვილის KL რაოდენობა და, მაშასადამე, შესაფასებელი მოდელებიც.

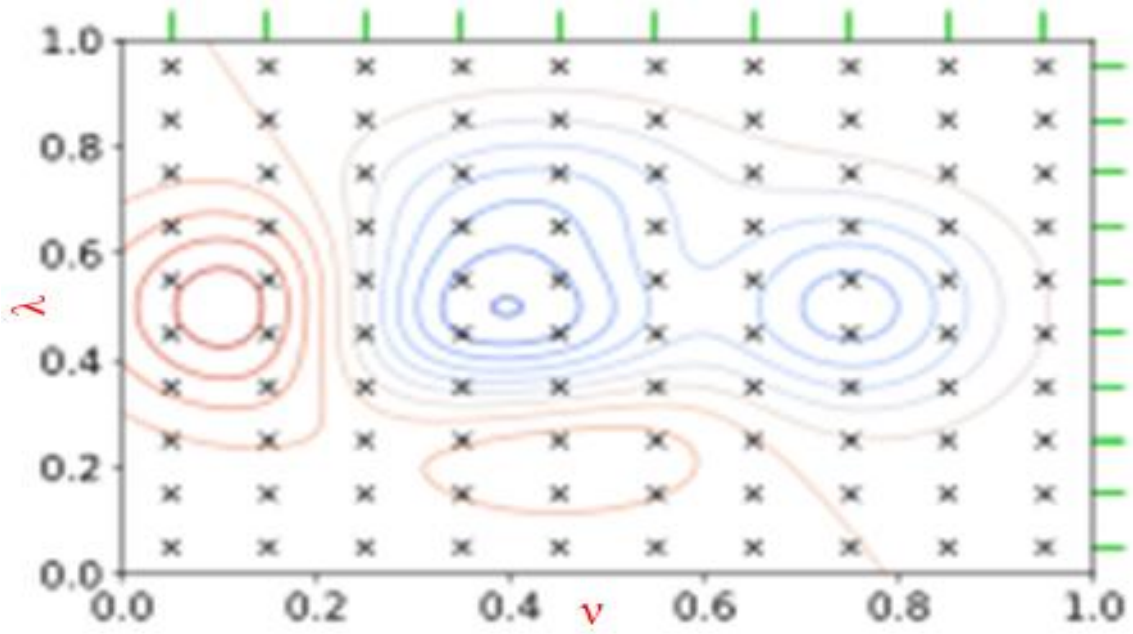
ყურადღება

ალგორითმული ხარჯები ჰიპერპარამეტრთა ძებნაზე მესრის გამოყენებით შეიძლება სწრაფად გაიზარდოს : თუ, მაგალითად, გამოიყენება ჯვარედინი შემოწმება (კროს-ვალიდაცია) 10 ნაკეცით (ბლოკით) ნებისმიერი იმ მოდელის შესაფასებლად, რომელიც გაწვრთნილია ალგორითმით, სადაც ჰიპერპარამეტრთა ყველა წყვილისთვის 10 შესაძლო მნიშვნელობა განიხილება, მაშინ საჭირო გახდება 1000 მოდელის გაწვრთნა და შეფასება.

მაგრამ ეს 1000 სწავლება და შეფასება მარტივად განპარალელებადია გამოთვლით კლასტერზე.

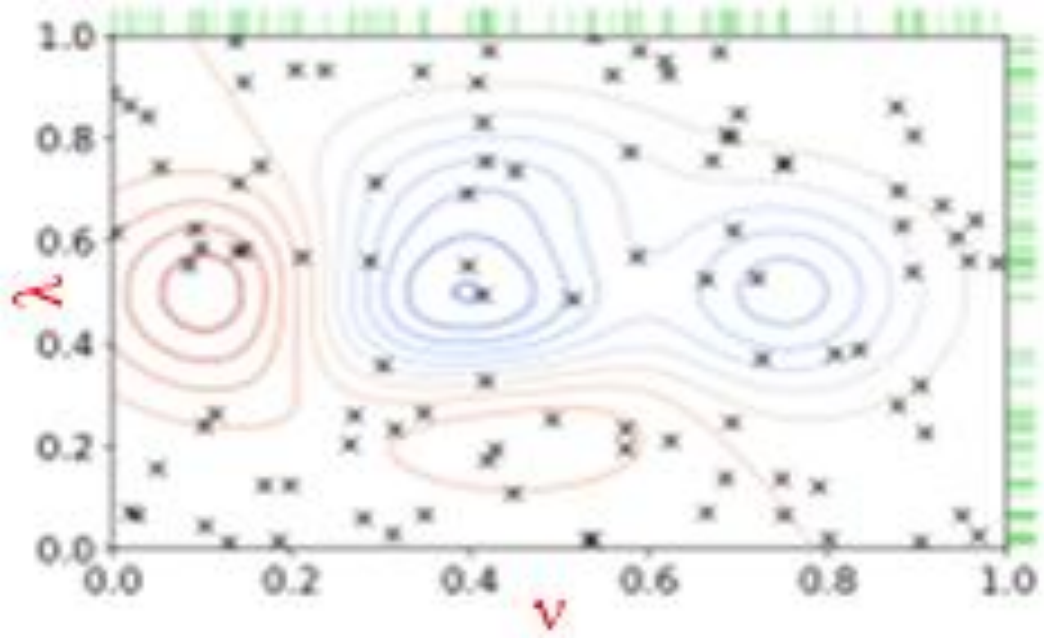
2.2 შემთხვევითი ძებნა

ძებნა მესრით, რა თქმა უნდა, ყველაზე ფართოდ გავრცელებული მეთოდია სწავლების ალგორითმთა ჰიპერპარამეტრების ოპტიმიზაციისათვის, როცა ეს ალგორითმები ერთ ან ორ ჰიპერპარამეტრს იყენებს. მაგრამ იგი არაეფექტური ხდება იმ შემთხვევაში, როცა დასადგენი ჰიპერპარამეტრების რიცხვი მეტია და, კერძოდ, ნეირონული ქსელებისთვის (იხ. ლექცია 7). მართლაც, სავარაუდოდ, შემოწმებული იქნება ჰიპერპარამეტრთა უამრავი კომბინაცია, რომლებიც, დაახლოებით, ერთმანეთის ექვივალენტურია და იმავდროულად იგნორირებული იქნება ჰიპერპარამეტრების სივრცის ყველაზე მნიშვნელოვანი არეები. ამრიგად, განიავებული აღმოჩნდება რესურსები, რომლებიც შეიძლება გაცილებით უკეთ ყოფილიყო გამოყენებული. ეს იდეა, რომელიც დაწვრილებით არის განხილული ჯეიმს ბერგსტრასა (James Bergstra) და



იოშუა ბენჯიოს (Yoshua Bengio) მიერ სტატიაში Bergstra, J. და Bengio, Y. (2012), ილუსტრირებულია ნახატით 3.2.

a)



b)

ნახატი 3.2 *a*) – ძებნა მესერზე და *b*) — შემთხვევითი ძებნა.

კერძოდ, *ძებნა მესრის გამოყენებით ორი ჰიპერპარამეტრის მნიშვნელობათა სხვადასხვა კომბინაციას შორის* ამ ნახატზე მოცემულია ზედა ნაწილში. თითოეული ჰიპერპარამეტრისთვის მოწმდება 10 მნიშვნელობა და, ამრიგად, სულ მიიღება 100 სხვადასხვა კომბინაცია გამოსაანგარიშებლად და შესადარებლად. ეფექტურობის დონეთა მრუდები მოცემულია ორი ფერით. ლურჯი კონტური აღნიშნავს არეებს უკეთესი შედეგებით, წითელი კონტური კი აღნიშნავს არეებს უარესი შედეგებით.

შემთხვევითი ძებნა ორი ჰიპერპარამეტრის მნიშვნელობათა სხვადასხვა კომბინაციას შორის ამავე ნახატზე მოცემულია მის ქვედა ნაწილში. ამ მაგალითში შემთხვევითი მნიშვნელობების 100 კომბინაცია ფასდება. მწვანე ჭდეებით ასახულია ის ფაქტი, რომ აქ განიხილება ყოველი ჰიპერპარამეტრის ინდივიდუალურ მნიშვნელობათა უფრო მეტი რაოდენობა, ვიდრე მესრით ძებნისას.

განსაზღვრება 3.8 (შემთხვევითი ძებნა) *შემთხვევითი ძებნა* (ინგლისურად : *random search*) ეწოდება ჰიპერპარამეტრების ოპტიმიზაციის სტრატეგიას, რომელიც მდგომარეობს შესაბამისი ჰიპერპარამეტრებიდან თითოეულისთვის მნიშვნელობათა დიაპაზონების განსაზღვრაში, რომლებიც ჯერ მოწმდება, ხოლო შემდეგ ამ დიაპაზონებიდან თანაბრად ამოღებული ჰიპერპარამეტრების მნიშვნელობათა K კომბინაციის განზოგადების შეცდომა ფასდება ემპირიული გზით.

შემთხვევითი ძებნის კიდევ ერთი უპირატესობა ისაა, რომ სტრატეგია ადვილად შეიძლება იყოს ადაპტირებული ცვლილებებთან რესურსებში : თუ გვიხდება გამოთვლების ნაადრევად შეწყვეტა, ჩატარებული ექსპერიმენტი მაინც რჩება შემთხვევით ძებნად და, პირიქით, თუ მეტი რესურსია ხელმისაწვდომი, ადვილად შეიძლება ძებნის წერტილების დამატება მესრის კორექტირების აუცილებლობის გარეშე.

3 ეფექტურობის კრიტერიუმები

კონტროლირებადი სწავლების მოდელის პრედიქტიული ეფექტურობის შეფასების მრავალი ხერხი არსებობს. ლექციის ეს ნაწილი უმთავრეს გავრცელებულ კრიტერიუმებს აღწერს.

3.1 დაბნეულობის მატრიცა და ნაწარმოები კრიტერიუმები

როგორც ვნახეთ, კლასიფიკაციის შეცდომათა რაოდენობა გამოიყენება პრედიქტიული მოდელის ხარისხის შესაფასებლად. აღსანიშნავია, რომ, ჩვეულებრივ, უპირატესობა ეძლევა შეცდომათა რაოდენობის აღწერას მაგალითების რაოდენობის წილის სახით : შეცდომათა წილის ასახვა მაგალითების 1%–ით უფრო მეტის მთქმელია, ვიდრე შეცდომათა აბსოლუტური რაოდენობის დასახელებით.

მაგრამ ყველა გაჩენილ შეცდომას აუცილებლად ერთნაირი ფასი არ აქვს. მოდით ავიღოთ ისეთი მოდელის მაგალითი, რომელმაც უნდა გვიწინასწარმეტყველოს, *არის* თუ *არ არის* რენტგენის სურათზე წარმოდგენილი საგანგაშო სიმსივნე : ცრუ განგაში, რომელიც შემდეგ ანულირებული იქნება დამატებითი გამოკვლევებით, ნაკლებად პრობლემატურია, ვიდრე სიმსივნის ვერდანახვა, გამოუვლენადობა რადიოგრაფიული მიდგომით და შესაბამისი პაციენტის დატოვება მკურნალობის გარეშე. როგორც ბინარული, ასევე მრავალკლასიანი საკლასიფიკაციო მოდელის ეფექტურობა შეიძლება განზოგადდეს ე.წ. *დაბნეულობის მატრიცაში.*

განსაზღვრება 3.9 (დაბნეულობის მატრიცა) კლასიფიკაციის ამოცანის განხილვისას, *დაბნეულობის მატრიცა* (ინგლისურად : *confusion matrix*) ეწოდება კვადრატულ M მატრიცას, რომლის რიგი კლასების რიცხვს უდრის, ხოლო მისი M_{ck} შესასვლელი იმ მაგალითების რაოდენობას ემთხვევა, რომლებშიც k ჭდეა პროგნოზირებული c კლასისთვის.

ბინარული კლასიფიკაციის შემთხვევაში გაუგებრობის მატრიცა შემდეგ ფორმას იძენს :

		რეალური/ფაქტობრივი კლასი	
		0	1
პროგნოზირებული კლასი	0	ჭეშმარიტად უარყოფითი (TN)	მცდარად უარყოფითი (FN)
	1	მცდარად დადებითი (FP)	ჭეშმარიტად დადებითი (TP)

ჭეშმარიტად დადებითი (ინგლისურად : *true positives*) ეწოდება უშეცდომოდ კლასიფიცირებულ დადებით მაგალითებს ; *მცდარად დადებითი* (ინგლისურად : *false positives*) ეწოდება მოდელის მიერ დადებითად მონიშნულ უარყოფით მაგალითებს და, პირიქით, *ჭეშმარიტად უარყოფითი* (ინგლისურად : *true negatives*) ეწოდება უშეცდომოდ კლასიფიცირებულ უარყოფით მაგალითებს და *მცდარად უარყოფითი* (ინგლისურად : *false negatives*) ეწოდება მოდელის მიერ უარყოფითად მონიშნულ დადებით მაგალითებს. *TP* სიმბოლოთი (ინგლისურად : *True Positive*), ჩვეულებრივ, აღნიშნავენ ჭეშმარიტად დადებით რიცხვებს, *FP სიმბოლოთი* (ინგლისურად : *False Positive*) – *მცდარად დადებით* რიცხვებს, *TN* სიმბოლოთი (ინგლისურად : *True Negative*) – ჭეშმარიტად უარყოფით რიცხვებს, ხოლო *FN* სიმბოლოთი (ინგლისურად : *False Negative*) – მცდარად უარყოფით რიცხვებს.

მცდარად დადებით შედეგებს ასევე უწოდებენ *მცდარ განგაშებს* ან *პირველი ტიპის შეცდომებს*, განსხვავებით *მეორე ტიპის შეცდომებისგან*, რომლებიც არის მცდარად უარყოფითი.

✿

გაუგებრობის მატრიცის საფუძველზე შესაძლებელია შეფასების მრავალი კრიტერიუმის მიღება. ზოგიერთი მაგალითი მოცემულია ქვემოთ :

განსაზღვრება 3.10 (სისავსე, მგრძნობელობა) ჭეშმარიტად დადებითი მაგალითების წილს, ესე იგი იმ დადებითი მაგალითების წილს, რომლებიც, როგორც ასეთი, სწორად იყო იდენტიფიცირებული, *სისავსე* (ინგლისურად : *recall*) ან *მგრძნობელობა* (ინგლისურად : *sensitivity*) ეწოდება :

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}.$$

✿

მაგრამ ძლიან ადვილია კარგი სისავსის მიღება, თუ გაკეთდა პროგნოზი, რომ *ყველა* მაგალითი დადებითია. ასე რომ, ეს კრიტერიუმი განმხოლოებულად არ შეიძლება იყოს გამოყენებული. ამიტომ ხშირად მას ე.წ. *სიზუსტე* ემატება.

განსაზღვრება 3.11 (სიზუსტე) უშეცდომო პროგნოზების წილს დადებით პროგნოზებს შორის *სიზუსტე* (ინგლისურად : *precision*) ან *დადებითი შედეგის პრედიქტიული (პროგნოსტიკული) ფასეულობა* (ინგლისურად : *positive predictive value, PPV*) ეწოდება :

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}.$$

✿

მსგავსად იმისა, როგორი ადვილია ძალიან კარგი სისავსის (მგრძნობელობის) მიღება (სიზუსტის ხარჯზე), ასევე ადვილია კარგი სიზუსტის მიღება (სისავსის ხარჯზე), თუ შემოვიფარგლებით დადებითი პროგნოზების ძალიან მცირე რაოდენობით, რაც ამცირებს იმის რისკს, რომ ისინი მცდარი აღმოჩნდება.

ყურადღება

ინგლისური ენა განასხვავებს ერთმანეთისგან სიტყვებს *precision* და *accuracy*. პირველი ნიშნავს სიზუსტეს, ხოლო მეორე — სწორად მონიშნული მაგალითების წილს, ე.ი. შეცდომათა წილის (პროცენტის) დამატებას ერთამდე. ამიტომ ზოგიერთ ენაში ტერმინი *accuracy* ასევე ითარგმნება როგორც *სიზუსტე*, ვთქვათ, ფრანგულ ენაში (*précision*). ამიტომ *precision* და *accuracy* ტერმინების გამოყენება სიფრთხილით უდნა ხდებოდეს. ლექციათა წინამდებარე კურსში *accuracy* ცნების გადმოცემა ხდება ტერმინით *სანდოობა*.

სანდოობის მეტრიკისთვის გვაქვს :

$$\text{Accuracy}=\frac{\text{TP+FN}}{\text{TP+TN+FP+FN}}=1-\frac{\text{TN+FP}}{\text{TP+TN+FP+FN}}$$

სანდოობა (accuracy) არის მაჩვენებელი, რომელიც აღწერს მოდელის წინასწარმეტყველების საერთო სიზუსტეს (ყველა კლასის მიხედვით). ეს განსაკუთრებით სასარგებლოა, როცა ყოველი კლასი ერთნაირად მნიშვნელოვანია. იგი დამოკიდებულია სწორი პროგნოზების $\left(\text{TP+FN}\right)$ ან $\left(\text{TN+FP}\right)$ რიცხვთა ფარდობაზე პროგნოზების საერთო $\left(\text{TP+TN+FP+FN}\right)$ რაოდენობასთან.

რომ მოხდეს *recall* და *precision* მაჩვენებლების გაერთიანება ერთ მახასიათებლად, გამოითვლება ე.წ. *F-ზომა* :

განსაზღვრება 3.12 (F-ზომა) *F-ზომა* (ინგლისურად : *F-score* ან *F1-score*) ეწოდება სიზუსტისა და სისავსის (მგრძნობელობის) საშუალო ჰარმონიულს :

$$F=2\frac{\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}}=\frac{2\,\text{TP}}{2\,\text{TP}+\text{FP}+\text{FN}}.$$

შენიშვნა

როგორც ცნობილია, საშუალო ჰარმონიულის განმარტება სტატისტიკაში ასეთია :

დავუშვათ, რომ მოცემულია n დადებითი $x_1,x_2,\ldots,x_n$ რიცხვი. მაშინ მათი *საშუალო ჰარმონიული* იქნება ისეთი $H\left(x_1,x_2,\ldots,x_n\right)$ რიცხვი, რომელიც არის შექცეული (შებრუნებული) სიდიდე $\frac{1}{x_1},\frac{1}{x_2},\cdots,\frac{1}{x_n}$ რიცხვთა საშუალო არითმეტიკულის მიმართ :

$$H\left(x_1,x_2,\ldots,x_n\right)=\frac{n}{\frac{1}{x_1}+\frac{1}{x_2}+\cdots+\frac{1}{x_n}}.$$

განსაზღვრება 3.13 (სპეციფიკურობა) *სპეციფიკურობა* (ინგლისურად : *specificity*), სხვანაირად *ჭეშმარიტად უარყოფითი პროპორცია,* ეწოდება ჭეშმარიტად უარყოფითი შედეგების წილს (პროცენტს) ან უფრო გასაგებად რომ ვთქვათ, უარყოფითი მაგალითების წილს, რომლებიც სწორად არის იდენტიფიცირებული ასეთებად :

$$\text{Specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}}.$$

მაგალითი

ამ ზემოთ ნახსენები სხვადასხვა კრიტერიუმის საილუსტრაციოდ განვიხილოთ კლინიკური ტესტის მაგალითი. ეს არ არის მანქანური სწავლების მოდელი და მხოლოდ კიბოზე აღებული საშვილოსნოს ყელის ნაცხის ანალიზია, რაც გაცილებით უფრო მარტივი და ნაკლებად ინვაზიური (უმნიშვნელოდ ტრავმირებადი) ტესტია, ვიდრე ჰისტოლოგიური (უჯრედებისა და ქსოვილების მიკროსკოპული აგებულების) გამოკვლევა, რომლის ინტერპრეტირება ექსპერტმა უნდა განახორციელოს და სწორედ მისი დასკვნა იქნება გამოყენებული საბაზო ჭეშმარიტებად.

ცხრილში 3.1 წარმოდგენილია ორმოცი წლისა და მეტი ასაკის ოთხ ათას ქალზე ჩატარებული ექსპერიმენტის შედეგები.

	კიბო	კიბო არ არის	სულ
ნაცხი + (დადებითი)	190	210	400
ნაცხი - (უარყოფითი)	10	3590	360
სულ	200	3800	4000

ცხრილი 3.1 - გაუგებრობის მატრიცა ექსპერიმენტისთვის, რომელიც ეხება საშვილოსნოს ყელის კიბოს სკრინინგს ნაცხის საშუალებით.

სისავსე (მგრძნობელობა) 95% არის, სპეციფიკურობა – $94,5\%$, მაგრამ სიზუსტე – მხოლოდ $47,5\%$. მართლაც, ეს ტესტი სკრინინგის (დაავადების შემთხვევათა გამოვლენის მიზნით კლინიკურად უსიმპტომო პირთა ჯგუფების გამოკვლევის) კარგი ინსტრუმენტია : კიბოს არარსებობის ალბათობა უარყოფითი ნაცხის დროს მაღალია $(3590/3600 \approx 99,7\%)$. მაგრამ ეს ცუდი დიაგნოსტიკური ინსტრუმენტია იმ აზრით, რომ ცრუ (მცდარი) განგაშის ალბათობა ძალზე მაღალია.

3.2 ბინარული კლასიფიკაციის ჭდის დამბრუნებელი მეთოდების ანალიზი

კლასიფიკაციის მრავალი ალგორითმი კი არ გვიბრუნებს პირდაპირ/უშუალოდ კლასის ჭდეს, არამედ იყენებს გადაწყვეტილების მიღების ფუნქციას, რომელიც ამ შემთხვევაში უნდა იყოს *ზღურბლური*, რომ იქცეს ჭდედ. გადაწყვეტილების მიღების ეს ფუნქცია შეიძლება ნებისმიერ შეფასებას წარმოადგენდეს (ვთქვათ, მოსანიშნი წერტილის k უახლოეს მეზობელს შორის დადებითი მაგალითების წილს — ჩვენ დაწვრილებით განვიხილავთ k უახლოესი მეზობლის ალგორითმს მერვე ლექციაში) ან დადებითი კლასისადმი მიკუთვნების ალბათობას (როგორც, მაგალითად, მეხუთე ლექციის მესამე პუნქტში წარმოდგენილი ლოგისტიკური რეგრესიის შემთხვევაში)

შენიშვნა

ბულის $f(x_1, x_2 \ldots, x_n)$ ფუნქციას *ზღურბლური* ეწოდება (ინგლისურად : *threshold function*), თუ შესაძლებელია მისი წარმოდგენა $f(x_1, x_2 \ldots, x_n) = \left(\sum_{i=1}^{n} w_i x_i \geq \vartheta \right)$ სახით, სადაც w_i არის x_i არგუმენტის *წონა* (ინგლისურად : *weight*), ხოლო ϑ არის f ფუნქციის *ზღურბლი* (ინგლისურად : *threshold*), ამასთან ერთად $w_i, \vartheta \in \mathbb{R}$.

არსებობს რამდენიმე კრიტერიუმი გადაწყვეტილების მიღების ფუნქციის ხარისხის შესაფასებლად ზღურბლის გამოთვლამდე.

ROC მრუდი

განსაზღვრება 3.14 (ROC მრუდი) *ROC*, *მრუდი* (ინგლისურად : *Receiver–Operator Characteristic — მიმღები ოპერატორის მახასიათებელი*) ეწოდება წირს, რომელიც აღწერს მგრძნობელობის ევოლუციას როგორც *სპეციფიკურობის 1-მდე დამატების* (სხვანაირად, *ანტისპეციფიკურობის*) ფუნქციას გადაწყვეტილების მიღების ზღურბლის ცვლილებისას.

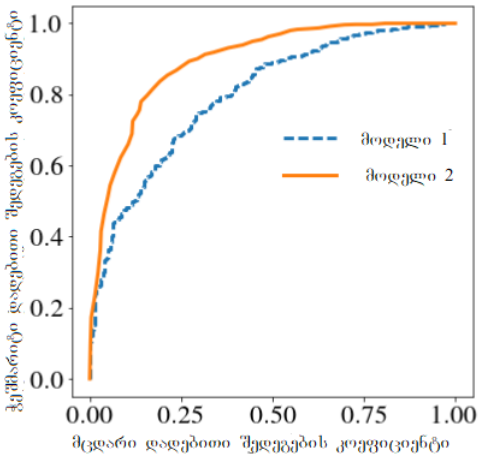
მანქანურ სწავლებაში ცნება შემოსულია ტელეკომუნიკაციიდან, სადაც ეს წირები გამოიყენება სისტემის მიერ სიგნალისა და ფონური ხმაურის ერთმანეთისგან განცალკევების შესაძლებლობის შესასწავლად.

შესაძლებელია ROC მრუდის განზოგადება ფართობით, რომელიც განთავსებულია ხსენებული მრუდის ქვეშ. ამ მახასიათებელს შემოკლებით *AUROC* აბრევიატურით აღნიშნავენ *Area Under the ROC* (ფართობი, არე, ზონა ROC–ის ქვეშ) ინგლისური ფრაზის შესაბამისად.

ROC მრუდის ერთ-ერთი მაგალითი წარმოდგენილია ნახატზე 3.3.

$(0,0)$ წერტილი მაშინ ჩნდება, როცა ზღურბლად გამოიყენება გადაწყვეტილების მიღების ფუნქციით დაბრუნებულ უდიდეს მნიშვნელობაზე მეტი რიცხვი : ასეთ შემთხვევაში ყველა მაგალითი მოინიშნება როგორც უარყოფითი.

და, პირიქით, $(1,1)$ წერტილი მაშინ ჩნდება, როცა ზღურბლად გამოიყენება გადაწყვეტილების მიღების ფუნქციით დაბრუნებულ უმცირეს მნიშვნელობაზე ნაკლები რიცხვი : ასეთ შემთხვევაში ყველა მაგალითი მოინიშნება როგორც დადებითი.



ნახატი 3.3 – ორი მოდელის ROC მრუდი : ჭეშმარიტი დადებითი შედეგების კოეფიციენტი (TRP – True Positive Rate) როგორც მცდარი დადებითი შედეგების კოეფიციენტის (FPR – False Positive Rate) ფუნქცია.

ROC მრუდის ასაგებად ზღურბლის როლი ენიჭება მონაცემთა კრებულზე გადაწყვეტილების მიღების ფუნქციის თანამიმდევრულ მნიშვნელობებს. ასე რომ, ყოველ ახალ ზღურბლურ მნიშვნელობაზე დაკვირვება, რომელიც ადრე იყო უარყოფითი, იცვლის თავის ჭდეს. თუ ეს დაკვირვება ნამდვილად დადებითია, მაშინ მგრძნობელობა (სისავსე) იზრდება $1/n_p$ – ით (სადაც n_p – დადებითი მაგალითების რიცხვია) ; წინააღმდეგ შემთხვევაში, ეს სწორედ ანტისპეციფიკურობაა, რომელიც იზრდება, $1/n_n$ – დან, სადაც n_n – უარყოფითი მაგალითების რიცხვია. ასე რომ, ROC მრუდი წარმოადგენს კიბისებრ მრუდს.

იდეალური კლასიფიკატორი, რომელიც შეცდომას არ უშვებს, სისტემატურად აკავშირებს უარყოფით მაგალითებთან უფრო დაბალ შეფასებებს, ვიდრე დადებით მაგალითებთან. ამიტომ მისი ROC-მრუდი მიჰყვება $[0,1]^2$ კვადრატის მარცხენა ზედა კუთხეს : ფართობი მრუდის ქვემოთ 1–ის ტოლია.

შემთხვევითი კლასიფიკატორის ROC მრუდი, რომელიც შეცდომებისა და სწორი კლასიფიკაციების, დაახლოებით, ერთნაირ წილს უზრუნველყოფს ნებისმიერ გამოყენებულ ზღურბლზე, ამ კვადრატის დიაგონალს მიჰყვება. ასე რომ, ზონის ფართობი შემთხვევითი კლასიფიკატორის ROC მრუდის ქვეშ 0.5–ის ტოლია.

მაგალითი

ROC მრუდის აგების საილუსტრაციოდ ავიღოთ მაგალითი, რომელიც ასახულია ცხრილში 3.2.

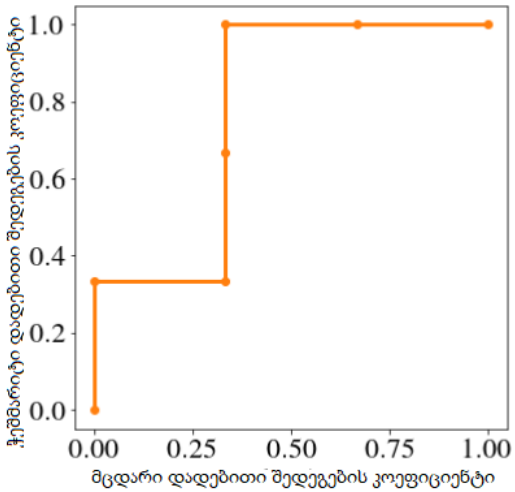
ჭდე	+	–	+	+	–	–
შეფასება	0.9	0.8	0.6	0.4	0.3	0.1

ცხრილი 3.2 – ექვს ნიმუშზე შეფასებული შედეგები ბინარული კლასიფიკაციის ექსპერიმენტში.

ზღურბლისთვის, რომელიც 0.9 –ზე მეტია, ექვსივე მაგალითი მონიშნულია როგორც უარყოფითი. ასე რომ, იწყებენ $(0,0)$ წერტილით. ზღურბლისთვის, რომელიც 0.95-სა და 0.9 -ის შორისაა, მხოლოდ პირველი დაკვირვებაა მონიშნული როგორც დადებითი. მგრძნობელობა არის, ამრიგად, 1, მაშინ როცა ანტისპეციფიკურობა რჩება ნულის ტოლი. ამის გაგრძელება მანამ შეიძლება, ვიდრე არ იქნება მიღწეული 0.1 –ზე ნაკლები ზღურბლი :

ზღურბლი	>0.9	0.8–0.9	0.6–0.8	0.4–0.6	0.3–0.4	0.1–0.3	<0
TP/P	0	1/3	1/3	2/3	1	1	1
FP/P	0	0	1/3	1/3	1/3	2/3	1

შესაბამისი ROC მრუდი ნაჩვენებია ნახატზე 3.4.



ნახატი 3.4 – ROC მრუდი ექსპერიმენტისათვის ცხრილში 3.2 : ჭეშმარიტი დადებითი შედეგების კოეფიციენტი (TRP – True Positive Rate) როგორც მცდარი დადებითი შედეგების კოეფიციენტის (FPR – False Positive Rate) ფუნქცია.

დაბოლოს, ROC მრუდი შეიძლება იყოს გამოყენებული გადაწყვეტილების მიღების ზღურბლის ასარჩევად იმ სისავსის/მგრძნობელობის (ინგლისურად : *recall* / *sensitivity*) ან სპეციფიკურობის (ინგლისურად : *specificity*) საფუძველზე, რომლის გარანტირება სასურველია.

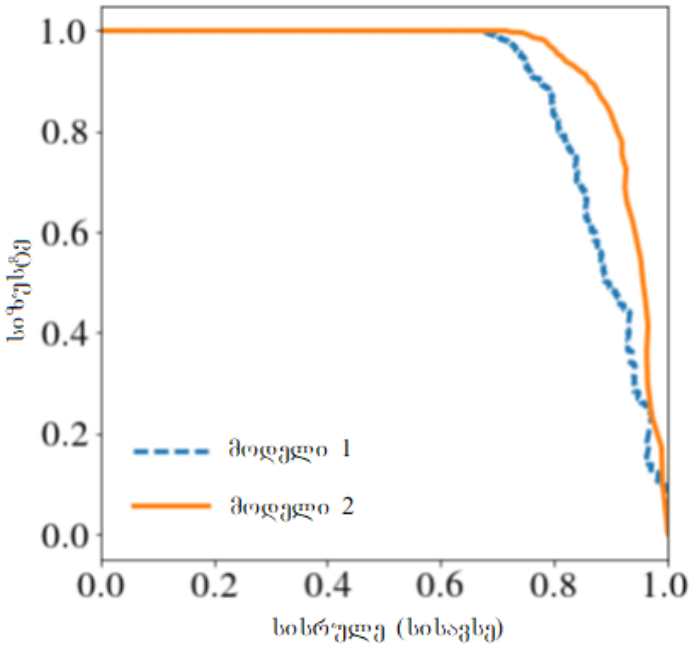
მრუდი სიზუსტე–სისავსე

ROC მრუდს ხშირად ემატება მრუდი სიზუსტე–სისავსე,

განსაზღვრება 3.15 (მრუდი precision-recall) *მრუდი სიზუსტე-სისავსე,* ანუ *precision-recall curve* ინგლისურად, ეწოდება მრუდს, რომელიც აღწერს სიზუსტის ევოლუციას როგორც სისავსის (სხვანაირად, მგრძნობელობის) ფუნქციას *გადაწყვეტილების მიღების ზღურბლის* ცვლილებისას.

ამ მრუდის წარმოსადგენად შეიძლება მის ქვეშ განთავსებული ზონის ფართობის გამოყენება, რომელსაც ხშირად შემოკლებით *AUPR* აბრევიატურით აღნიშნავენ *Area Under the Precision-Recall curve* ფრაზის გადმოსაცემად.

«სიზუსტე – სისავსე» მრუდის მაგალითი იმავე მონაცემებისათვის, რომლებიც ასახულია ნახატზე 3.3, წარმოდგენილია ნახატზე 3.5.



ნახატი 3.5 – მრუდები «სიზუსტე - სისავსე» (სხვანაირად, PR მრუდები — «Precision–Recall») ორი მოდელისთვის.

შენიშვნა

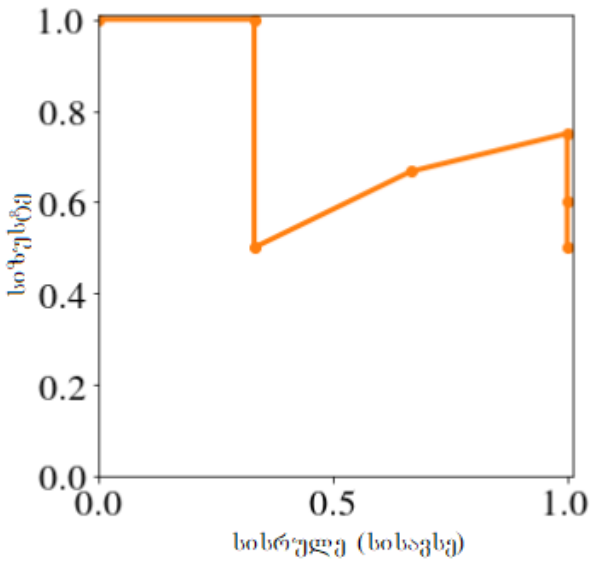
ყველაზე მაღალი ზღურბლისთვის არც ერთი მაგალითია მონიშნული როგორც დადებითი და სიზუსტე, ამრიგად, არ არის განსაზღვრული. შეთანხმების თანახმად, ჩვეულებრივ, გამოიყენება სიზუსტე, რომელიც 1–ის ტოლია, თუ პირველი განსახილველი დაკვირვება არის დადებითი, და სიზუსტე, რომელიც 0 –ის ტოლია, წინააღმდეგ შემთხვევაში.

მაგალითი

დავუბრუნდეთ წინა მაგალითს იმისათვის, რომ ავაგოთ მრუდი სიზუსტე–სისავსე. *სიზუსტის* (*precision*) და *სისავსის* / *მგრძნობელობის* (*recall* / *sensitivity*) პარამეტრთა მნიშვნელობები არის შემდეგი :

ზღურბლი	> 0.9	0.8–0.9	0.6–0.8	0.4–0.6	0.3–0.4	0.1–0.3	< 0.
სისავსე	0	1/3	1/3	2/3	1	1	1
სიზუსტე	–	1	1/2	2/3	3/4	3/5	3/6

ამრიგად, მიიღწევა « *სიზუსტე – სისავსე* » მრუდი, რომელიც ნაჩვენებია ნახატზე 3.6.



ნახატი 3.6 – მრუდი «სიზუსტე - სისრულე (სისავსე)», რომელიც შეესაბამება ექსპერიმენტს ცხრილიდან 3.2 .

3.3 რეგრესიის შეცდომები

რეგრესიის ამოცანის შემთხვევაში შეცდომათა რიცხვი არ არის შესაფერისი კრიტერიუმი მუშაობის შესაფასებლად. ერთი მხრივ, რიცხობრივი უზუსტობების (ცდომილებების) გამო, ადვილი არ არის რაოდენობრივად ნამდვილი რიცხვით შეფასებულ პროგნოზთან დაკავშირებით თქმა სწორია იგი თუ არა. მეორე მხრივ, მოდელი, რომლის პროგნოზების 50% კორექტულია 0.1 პროცენტით გადახრის ფარგლებში და დანარჩენი 50 პროცენტი ძალიან დაშორებულია ჭეშმარიტი მნიშვნელობებისაგან, უკეთესია მოდელზე, რომელიც მხოლოდ 0.1 პროცენტით გადახრის ფარგლებშია კორექტული, მაგრამ მაგალითების ასივე პროცენტი-სათვის ?

ასე რომ, რეგრესიული მოდელის მუშაობის ხარისხის შეფასება უკეთესია ხდებოდეს პროგნოზებსა და ფაქტობრივ მნიშვნელობებს შორის არსებული სხვაობის, შეუსაბამობის თვალსაზრისით.

ამიტომ პირველ კრიტერიუმს წარმოადგენს საშუალო კვადრატული შეცდომა :

განსაზღვრება 3.16 (საშუალო კვადრატული შეცდომა (MSE)) თუ მოცემულია n ფაქტობრივი ჭდე $y^1, y^2, \ldots, y^n$ და n პროგნოზი $f\left(\vec{x}^1\right), f\left(\vec{x}^2\right), \ldots, f\left(\vec{x}^n\right)$, მაშინ *საშუალო კვადრატული შეცდომა*, ანუ *MSE* (ინგლისურიდან : *mean squared error*) ეწოდება შემდეგ მნიშვნელობას :

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(\vec{x}^i - y^i\right)^2.$$

ამ შეცდომის გასაზომად მიზნისათვის მიღებული ერთეულით, ხშირად უფრო მიზანშეწონილია მისი კვადრატული ფესვის გამოყენება :

განსაზღვრება 3.17 ((საშუალო კვადრატული შეცდომის ფესვი (RMSE)) თუ მოცემულია n ფაქტობრივი ჭდე $y^1, y^2, \ldots, y^n$ და n პროგნოზი $f\left(\vec{x}^1\right), f\left(\vec{x}^2\right), \ldots, f\left(\vec{x}^n\right)$, მაშინ *საშუალო*

კვადრატული შეცდომის ფესვი, ანუ *RMSE* (ინგლისურიდან : *root mean squared error*) ეწოდება შემდეგ მნიშვნელობას :

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\vec{x}^i - y^i\right)^2}.$$

❁

იმ შემთხვევებში, როცა მიზნის მნიშვნელობებში წარმოდგენილია რამდენიმე სხვადასხვა რიგის სიდიდე, ზოგჯერ უპირატესობა ენიჭება ლოგარითმზე გადასვლას $f\left(\vec{x}^i\right)$-ის შედარებამდე y^i-თან, რათა არ მიეცეს მეტი წონადობა, არსებითობა შეცდომებს, რომლებიც იყო დაშვებული უფრო მაღალი მნიშვნელობებისთვის.

განსაზღვრება 3.18 (საშუალო კვადრატული ლოგარითმული შეცდომის ფესვი – RMSLE) თუ მოცემულია n ფაქტობრივი ჭდე $y^1, y^2, \ldots, y^n$ და n პროგნოზი $f\left(\vec{x}^1\right), f\left(\vec{x}^2\right), \ldots, f\left(\vec{x}^n\right)$, მაშინ *საშუალო კვადრატული ლოგარითმული შეცდომის ფესვი*, ანუ *RMSLE* (ინგლისურიდან : *root mean squared log error*) ეწოდება შემდეგ მნიშვნელობას :

$$\mathrm{RMSLE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\log\left(f(\vec{x}^i)+1\right) - \log\left(y^i+1\right)\right)^2}.$$

❁

თუმცა. ამ შეცდომების ინტერპრეტაცია მაინც მოითხოვს მიზნობრივი მნიშვნელობების განაწილების ცოდნას : RMSE საშუალო კვადრატული შეცდომის ფესვს, რომელიც ერთი სანტიმეტრის ტოლია, არ ექნება ერთნაირი მნიშვნელობა, ვთქვათ, ადამიანის და ხილის ბუზის (დროზოფილას) ზომების წინასწარმეტყველების მცდელობისას.

ამ პრობლემის გადასაწყვეტად შესაძლებელია ნაშთების კვადრატთა ჯამის ნორმალიზება არა გასაშუალოებით, არამედ მისი შედარებით იმ მანძილების ჯამთან, რომლებიც ცალკეულ მიზნობრივ მნიშვნელობებს მათი საშუალო მნიშვნელობიდან აშორებს.

განსაზღვრება 3.19 (დეტერმინაციის კოეფიციენტი) თუ მოცემულია n ფაქტობრივი ჭდე $y^1, y^2, \ldots, y^n$ და n პროგნოზი $f\left(\vec{x}^1\right), f\left(\vec{x}^2\right), \ldots, f\left(\vec{x}^n\right)$, მაშინ *ფარდობითი კვადრატული შეცდომა* (ინგლისურად : *RSE, relative squared error*) ეწოდება შემდეგ მნიშვნელობას :

$$\mathrm{RSE} = \frac{\sum_{i=1}^{n}\left(f\left(\vec{x}^i\right) - y^i\right)^2}{\sum_{i=1}^{n}\left(y^i - \frac{1}{n}\sum_{l=1}^{n} y^i\right)^2}.$$

❁

RSE პარამეტრის 1–მდე დამატება არის R^2 (*R-კვადრატ*) სიმბოლოთი აღნიშნული *დეტერმინაციის კოეფიციენტი* (ინგლისურად : *coefficient of determination*).

დეტერმინაციის კოეფიციენტი იმიტომ აღინიშნება როგორც R^2, რომ იგი წარმოადგენს $\vec{y}$ და $\left(f\left(\vec{x}^1\right), f\left(\vec{x}^2\right),\ldots, f\left(\vec{x}^n\right)\right)$ სიდიდეთა შორის არსებული კორელაციის კოეფიციენტის კვადრატს. ეს კოეფიციენტი კი მოიცემა შემდეგი ფორმულით :

$$R = \frac{\sum_{i=1}^{n}\left(y^i - \frac{1}{n}\sum_{l=1}^{n} y^l\right)\left(f\left(\vec{x}^i\right) - \frac{1}{n}\sum_{l=1}^{n} f\left(\vec{x}^l\right)\right)}{\sqrt{\sum_{i=1}^{n}\left(y^i - \frac{1}{n}\sum_{l=1}^{n} y^l\right)^2}\cdot\sqrt{\sum_{i=1}^{n}\left(f\left(\vec{x}^i\right) - \frac{1}{n}\sum_{l=1}^{n} f\left(\vec{x}^l\right)\right)^2}} \qquad (3.2)$$

ეს კოეფიციენტი აჩვენებს რამდენად კარგად კორელირებს საპროგნოზირებელი მნიშვნელობები ფაქტობრივ მნიშვნელობებთან ; ყურადღება უნდა მიექცეს იმასაც, რომ ეს კოეფიციენტი ასევე მაღალი იქნება, თუ საპროგნოზირებელი მნიშვნელობები ფაქტობრივ მნიშვნელობებთან ანტიკორელირებულია, ე.ი. იმყოფება შექცეულ (უარყოფით) კორელაცია-ში. უარყოფითი კავშირის დროს ერთ-ერთი ნიშნის დაბალ მნიშვნელობებს მეორე ნიშნის მაღალი მნიშვნელობები შეესაბამება.

3.4 შედარება გულუბრყვილო ალგორითმებთან

მანქანური სწავლების მოდელის ასაგებად ეყრდნობიან, ერთი მხრივ, მონაცემებს, ხოლო მეორე მხრივ, ჰიპოთეზებს მოდელის ფორმის შესახებ ; ეს ჰიპოთეზები განსაზღვრავს ჰიპოთეზათა სივრცეს. ამ ჰიპოთეზების სანდოობა, დასაბუთებულობა დამოკიდებულია შესასწავლ პრობლემაზე. ეს პრობლემა შეიძლება მეტ–ნაკლებად მარტივი იყოს და გაწვრთნილი, დანასწავლი მოდელის მუშაობის ხარისხი მხოლოდ ამ სირთულის გათვალისწინებით შეიძლება იყოს გაგებული.

ამის დასადგენად ძალიან სასარგებლო შეიძლება იყოს გულუბრყვილო სწავლების მიდგომათა გამოყენება, სხვანაირად რომ ვთქვათ, ძალიან მარტივი მიდგომების, რომლებიც იყენებს საწვრთნელი კრებულის გარკვეულ თვისებებს, მაგრამ არა მოსანიშნ დაკვირვებათა სიმრავ-ლის. ჩვენ არ ველით ამ მეთოდების მუშაობის კარგ მახასიათებლებს.

მაგრამ ისინი გვესახება ეტალონად იმისათვის, რომ უკეთ შევიცნოთ ეფექტურობის (გნებავთ, საერთოდ, მოდელის მუშაობის ხარისხის) განზომადი პარამეტრები (მახასიათებლები) და, ამასთან ერთად, ხსენებული მეთოდები გვატყობინებს აუცილებელ მინიმუმს, რომელსაც შეიძლება ველოდოთ აგებული მოდელებისგან.

კლასიფიკაციის გულუბრყვილო მეთოდები

კლასიფიკაციის ამოცანისთვის შეგვიძლია, მაგალითად, ერთ-ერთი შემდეგი მიდგომის განხილვა :

— მასწავლებელ სიმრავლეში უმრავლესობის ჭდეთა (მონიშვნათა) სისტემატური წინასწარმეტყველება.

— შემთხვევითი ჭდის წინასწარმეტყველება მონიშვნათა განაწილების შესაბამისად მასწავლებელ სიმრავლეზე.

— ბინარული კლასიფიკაციის შემთხვევაში შეფასებათა წინასწარმეტყველება თანაბრად მათ ზღურბლურ მნიშვნელობამდე. ეს მეთოდი განსაკუთრებით მაშინ არის რეკომენდებული, როცა ცდილობენ ROC ან PR მრუდის აგებას.

შენიშვნა

თუ საწვრთნელი კრებული *დაუბალანსებელია*, ე.ი. ერთი კლასი წარმოდგენილია უფრო ფართოდ, ვიდრე სხვა, მაშინ აქ აღწერილ პირველ გულუბრყვილო ალგორითმს შეიძლება ჰქონდეს შეცდომათა ძალიან დაბალი კოეფიციენტი. აუცილებელი იქნება ასევე ალგორითმის *სპეციფიკურობის* გათვალისწინებაც.

რეგრესიის გულუბრყვილო მეთოდები

რეგრესიის ამოცანისთვის შეიძლება შემდეგი გულუბრყვილო მიდგომების განხილვა :

— საწვრთნელი სიმრავლის ჭდეთა ყველაზე მცირე და ყველაზე დიდ მნიშვნელობებს შორის თანაბრად განაწილებული შემთხვევითი სიდიდის წინასწარმეტყველება ;

— საწვრთნელი სიმრავლის ჭდეთა საშუალოს ან მედიანის სისტემატური წინასწარ-მეტყველება.

4 საკვანძო მომენტები

— ზედმეტად სწავლების ასაცილებლად თავიდან, მოდელის არჩევის ეტაპზე მნიშვნელოვანია სხვადასხვა მოდელის შემოწმება, ეს მოდელები უნდა იყოს ტესტირებული მონაცემთა კრებულზე, რომელიც განსხვავდება სწავლებისთვის (წვრთნისთვის) გამოყენებულისგან.

— მოდელის მუშაობის ხარისხის დასადგენად აუცილებელია მისი შეფასება მონაცემებზე, რომლებიც არ ყოფილა გამოყენებული ამ მოდელის არც სწავლებისთვის და არც მისი არჩევისთვის.

— მოდელის პროგნოსტიკული ეფექტურობის შესაფასებლად მრავალი კრიტერიუმი არსებობს. მათი შერჩევა უნდა ხდებოდეს მოდელის გამოყენების სფეროდან გამომდინარე.

— მოდელის ეფექტურობის საინტერპრეტაციოდ შეიძლება სასარგებლო იყოს ამ მოდელის შედარება გულუბრყვილო მიდგომასთან.

დამატებითი ინფორმაცია

• ჰიპერპარამეტრების ოპტიმიზაცია ასევე შეიძლება იყოს განხორციელებული ბაიესური მიდგომით. თუ აღვნიშნავთ $f_{\vec{\theta}}$ სიმბოლოთი მოდელს, რომელიც გაწვრთნილია

ჰიპერპარამეტრების $\vec{\theta}$ ვექტორით, შეიძლება ობიექტური $\frac{1}{|\mathcal{D}_{te}|}\sum_{\vec{x},,y\in\mathcal{D}_{te}} L\left(y, f_{\vec{\theta}}\left(\vec{x}\right)\right)$ ფუნქციის მოდელირება $\vec{\theta}$ პირობისათვის და $\vec{\theta}$ ვექტორის შეფასებაც. უფრო დაწვრილებითი ინფორმაცია მოიპოვება ნაშრომში Bergstra და სხვ (2011).

· «*სიზუსტე–სისავსე*» მრუდის აგებისას ნახატზე 3.5 წერტილები შეერთებულია სეგმენტებით (წრფის მონაკვეთებით). ასეთი წრფივი ინტერპოლაცია სინამდვილეში ხშირად გაუმართლებელია (Jesse Davis და Mark Goadrich, ასევე Tom Fawcett). ამ საკითხთან უფრო დაწვრილებით გაცნობა შეიძლება, მაგალითად, სტატიებში Davis და Goadrich (2006), ასევე Fawcett (2006).

· ზოგიერთი მოდელისთვის, კერძოდ, საუბარია წრფივ მოდელებზე (იხ. ლექცია 5) შესაძლებელია *ოპტიმიზმის* შეფასება, ე.ი. სხვაობის სწავლებისა და ტესტირების შეცდომათა შორის, თეორიულად და არა კროს=ვალიდაციის (ჯვარედინი შემოწმების) გზით. წრფივი მოდელების შემთხვევაში შესაძლებელი იქნება, კერძოდ, *მალოუს* (Colin Lingwood Mallows) C_p *კოეფიციენტის*, აკაიკეს (Hirotugu Akaike) *საინფორმაციო კრიტერიუმის* ან ბაიესის (Thomas Bayes) და შვარცის (Gideon Schwarz) სახელებთან დაკავშირებული *საინფორმაციო კრიტერიუმის* გამოყენება. უფრო დაწვრილებითი ინფორმაციის მოპოვება შესაძლებელია იადოლაჰ დოდჯისა (Yadolah Dodge) და ვალენტინ რუსონის (Valentin Rousson) გამოყენებით რეგრესიულ ანალიზთან დაკავშირებულ ძალიან ცნობილ წიგნში Dodge და Rousson (2004).

· *აღწერის მინიმალური სიგრძე* (ინგლისურად : *minimum description length*, *MDL*) არის ინფორმაციის თეორიიდან ნასესხები ცნება, რომელიც შეიძლება იყოს გამოყენებული მოდელის ასარჩევად. იდეის დემონსტრირება ჯერ კიდევ 1978 წელს განახორციელა ცნობილმა ფინელმა მკვლევარმა ინფორმაციის თეორიის დარგში ჯორმა რისანენმა (Jorma Johannes Rissanen, 1932–2020), იხ. Rissanen (1978). ამ სტრუქტურაში მონაცემთა $\mathcal{D}$ კრებულის ჭდეები შეიძლება იყოს ასახული, ერთი მხრივ, მოდელის წარმოდგენით, ხოლო მეორე მხრივ, სხვაობის წარმოდგენით მოდელის წინასწარმეტყველებებსა და მათ ჭეშმარიტ ჭდეებს (მონიშვნებს) შორის $\mathcal{D}$–ში. საუკეთესოა ის მოდელი. რომლისთვისაც ამ წარმოდგენათა მნიშვნელობების ჯამი მინიმალურია : კარგი მოდელი მონაცემთა ეფექტურად შეკუმშვის საშუალებას იძლევა.

· ჯვარედინი შემოწმების (კროს-ვალიდაციის) პროცედურათა ღრმა და დაწვრილებითი ანალიზი მოდელის არჩევის ასპექტში მოიპოვება სილვენ არლოსა (Sylvain Arlot) და ალენ სელისის (Alain Celisse) მიმოხილვაში Arlot და Celisse (2010).

· მათემატიკურ ფოლკლორს და თეორემებს უფასო ლანჩის (სადილის) შეუძლებლობის შესახებ (NFL – No Free Lunch) ძებნისა და ოპტიმიზაციის პრობლემებისთვის მანქანურ სწავლებაში შეიძლება შევეხოთ დევიდ ვოლპერტისა (David Hilton Wolpert) და უილიამ მაკრიდის (William G. Macready) დახმარებით, იხ. Wolpert და Macready (1997).

5 ბიბლიოგრაფია

1. Arlot, S. და Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys,* 4:40–79.

2. Bergstra, J., Y., Bardenet, R., Bengio, Y. და Kégl, B. (2011). Algorithms for Hyper-Parameter Optimization. In *Advances in Neural Information Processing Systems.* 24:2546-2554.

3. Bergstra, J. და Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research,* 13(2):281-305.

4. Davis, J. და Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning* (ICML). 2006, Vol. 148, pages 233–240, New York, NY, USA, ACM Press.

5. Dodge, Y. და Rousson, V. (2004). *Analyse de régression appliquée*. Dunod, 279 pages

6. Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874.

7. Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, 14(5):465–471.

8. Wolpert, D. H. და Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82.

6 სავარჯიშოები

3.1 რა განსხვავებაა მოდელის სიზუსტესა და ეფექტურობას (მუშაობის ხარისხს) შორის ?

3.2 ანიკა წერს საბანკო თაღლითობის გამოვლენის ალგორითმს. გამოუვლენელი თაღლითობა ბანკს მეტი უჯდება, ვიდრე ისეთი ტრანზაქციის ხელით დამუშავება, რომელიც სინამდვილეში თაღლითური არ არის. რაზე უნდა გაამახვილოს ყურადღება ბანკმა : მცდარად დადებით, მცდარად უარყოფით თუ ჭეშმარიტად დადებით შედეგებზე ?

3.3 ბესო ბინარული კლასიფიკაციის ალგორითმის ტესტირებას ახორციელებს. ეს ალგორითმი შემთხვევითად აბრუნებს ჭდეებს «*უარყოფითი*» ან «*დადებითი*» 0.5–ის ტოლი ალბათობით ყოველი კლასისთვის. შეფასების სიმრავლე 85% დადებით და 15% უარყოფით მაგალითს შეიცავს. როგორი იქნება სანდოობა (ინგლ. accuracy), სისავსე (ინგლ. recall) და სიზუსტე (ინგლ. precision) ?

3.4 კარლო ბინარული კლასიფიკაციის ალგორითმის ტესტირებას ახორციელებს. ამ ალგორითმით ნებისმიერი დაკვირვება მოინიშნება როგორც დადებითი. შეფასების სიმრავლე 85% დადებით და 15% უარყოფით მაგალითს შეიცავს. როგორი იქნება სანდოობა (ინგლ. accuracy), სისავსე (ინგლ. recall) და სიზუსტე (ინგლ. precision) ?

3.5 დიტომ შექმნა სწავლების მოდელი მონაცემთა ისეთ სიმრავლეზე, რომელშიც კლასების უმრავლესობა უარყოფითია. შეფასების ინსტრუმენტად იგი F-ზომას იყენებს. რისი გამოყენებაა უკეთესი : დადებითი კლასის კონსტანტური კლასიფიკატორის თუ უარყოფითი კლასის კონსტანტური კლასიფიკატორის ?

3.6 როგორ წარვმართოთ კროს-ვალიდაცია (ჯვარედინი შემოწმება) დროითი მწკრივების შემთხვევაში, ე.ი. მაშინ, როცა $\left(\vec{x}^i, \vec{y}^i\right)$ წყვილები მოწესრიგებულია ქრონოლოგიურად ?

სავარჯიშოთა ამონახსნები

3.1 მოდელი, რომელსაც კარგი სიზუსტე აქვს, შეიძლება იყოს ნაკლებად ეფექტური, მაგალითად, იშვიათ მოვლენათა დეტექტირების შემთხვევაში : მოდელს. რომელიც წინასწარმეტყველებს, რომ ყველა დაკვირვება ეკუთვნის უმრავლესობათა კლასს, ძალიან კარგი სიზუსტე ექნება, მაგრამ იგი ნაკლებად ეფექტური აღმოჩნდება სრული უსარგებლობის თვალსაზრისით

3.2 მცდარი უარყოფითი შედეგები ყველაზე მვირადღირებულია ბანკისთვის და სწორედ ამ ასპექტზე უნდა მოხდეს ყურადღების გამახვილება უპირველეს ყოვლისა.

3.3 $TP=FN=0,5\times0,85$ და $FP=TN=0,5\times0,15$. ასე რომ :

$$\text{სანდოობა}(accuracy)=0,5\ ;\ \text{სისრულე}(recall)=0,5\ ;\ \text{სიზუსტე}(precision)=0,85.$$

3.4 $TP=0,85$, $FP=0,15$ და $FN=TN=0$. ასე რომ :

$$\text{სანდოობა}(accuracy)=0,85\ ;\ \text{სისრულე}(recall)=1\ ;\ \text{სიზუსტე}(precision)=0,85.$$

3.5 უარყოფითი კლასის კონსტანტურ კლასიფიკატორს ექნება $TP=0$ მაჩვენებელი და, ამრიგად, $0-$ის ტოლი $F-$ზომა. დადებითი კლასის კონსტანტური კლასიფიკატორი უფრო შესაფერისი ჩანს.

3.6 დროითი მწკრივების (ინგლისურად : *time series data*) დამუშავებისას ადამიანი ცდილობს მომავლის წინასწარმეტყველებას ნამყოს საფუძველზე. ამიტომ, შესაძლოა, მიზანშეწონილი იყოს მონაცემთა ე.წ, პირდაპირი გადაცემის განხორციელება (ინგლისურად : *forward chaining*), როდესაც $k-$ური სატესტო სიმრავლე შეიცავს $\left(\vec{x}^{k+1},\vec{y}^{k+1}\right)$ დაკვირვებას, ხოლო შესაბამისი საწვრთნელი სიმრავლე შეიცავს წინა k რაოდენობის დაკვირვებას.

ლექცია 4 ბაიესური დასკვნა (დანასკვი)

შინაარსი

ერთ-ერთ მნიშვნელოვან ფაქტორს იმ ფუნდამენტურ პრობლემებში, რომლებსაც ჩვენ ვაწყდებით მანქანური სწავლების დროს, წარმოადგენს განუსაზღვრელობა : სამყაროს ჩვენი გაგება შეზღუდულია ჩვენივე, გარდაუვალად არასრული დაკვირვებებით. ალბათური მოდელები ამ განუსაზღვრელობის ცხადად გათვალისწინების საშუალებას იძლევა.

ამ ლექციაში ჩვენ ვნახავთ, როგორ უნდა მოხდეს მოდელის სწავლების ამოცანის ფორმულირება დანასკვის გამოტანად დაკვირვებების და მოდელის პარამეტრების ერთობლივ განაწილებაზე. ასევე გავეცნობით ბაიესური სქემის გამოყენებას გადაწყვეტილებათა მისაღებად, რაც მოითხოვს დამატებით მოდელირებას, რომ დადგინდეს, რამდენად სასარგებლოა სწორი გადაწყვეტილების მიღება. საილუსტრაციოდ წარმოდგენილი იქნება გულუბრყვილო ბაიესური კლასიფიკაცია.

მიზნები

— კლასის ცნების ფორმალიზება ალბათური მოდელების საშუალებით ;

— დამაჯერებლობის (თანა)ფარდობის ტესტებზე დაფუძნებულ გადაწყვეტილებათა მიღების წესების განსაზღვრა ;

— განაწილების სიმკვრივის განსაზღვრა მაქსიმალური დამაჯერებლობის მეთოდით ან ბაიესური შემფასებლის საშუალებით ;

— გულუბრყვილო ბაიესური კლასიფიკაციის ალგორითმის რეალიზება.

1 წარმომშობი (გენერატიული) მოდელები ბინარული კლასიფიკაციისათვის

სტატისტიკური მიდგომა კლასიფიკაციისადმი კლასის ცნების ფორმალიზებას ახორციელებს ალბათური მოდელების საშუალებით. ამ სქემაში ჩვენ მივიჩნევთ, რომ n რაოდენობის $\vec{x}^1, \vec{x}^2, \ldots, \vec{x}^n$ დაკვირვებები არის $X \in \mathcal{X}$ შემთხვევითი ცვლადის რეალიზაციები. გარდა ამისა, ვუშვებთ, რომ მათი $y^1, y^2, \ldots, y^n$ ჭდეები არის $Y \in \{0,1\}$ შემთხვევითი ცვლადის რეალიზაციები. მრავალკლასიანი კლასიფიკაციის შემთხვევაში ჩვენ გამოვიყენებთ $Y \in \{1,2,\ldots,C\}$ შემთხვევით სიდიდეს, სადაც C – კლასების საერთო რიცხვია.

ახლა განვიხილოთ $\mathbb{P}(X,Y)$ ალბათობათა ერთობლივი განაწილება ყველა ცვლადისთვის მოდელში, კონკრეტულად, X და Y ცვლადისთვის. ამ მიდგომას ეწოდება *წარმომშობი* (*გენერატიული*) *მოდელირება* : იგი პასუხობს შეკითხვაზე «როგორ შეიძლება ყოფილიყო მიღებული მონაცემები, რომლებსაც ვაკვირდებით ?» რა თქმა უნდა, პრაქტიკაში ჩვენი მონაცემები ყოველთვის როდია შემთხვევითი პროცესის შედეგი ; მაგრამ მათი განხილვა შემთხვევითი ცვლადის რეალიზაციებად შეიძლება იყოს მათ მიერ შემცველი რთული ინფორმაციის წარმოდგენის ეფექტური ხერხი. მაგალითად, თუ ერთ-ერთი ჩვენი ცვლადი არის სამედიცინო სკრინინგული ტესტის შედეგი, უფრო მარტივია მისი წარმოდგენა ბერნულის განაწილების რეალიზაციად, ვიდრე მთელი იმ ბიოქიმიური ინფორმაციის მოდელირების ჩატარება, რომელსაც შეეძლო ასეთ შედეგამდე მივეყვანეთ. სწორედ ამ კონტექსტში წარიმართება ახლა ჩვენი საუბარი.

1.1 დასკვნა (დანასკვი) და წინასწარმეტყველება

ამრიგად, ალბათობა იმისა, რომ $\vec{x}$ დაკვირვება მიეკუთვნება c კლასს, განისაზღვრება $\mathbb{P}(Y=c \mid X=\vec{x})$ თანაფარდობით და ასეთ შემთხვევაში ჩნდება ორი პრობლემა :

– *დასკვნის* (*დანასკვის*) პრობლემა (ინგლისურად : *a problem of inference*), რომელიც მდგომარეობს $\mathbb{P}(Y=c \mid X=\vec{x})$ ალბათობის კანონების განსაზღვრაში ჩვენი დაკვირვებებიდან და ჰიპოთეზებიდან გამომდინარე ;

– *პროგნოზირების* (*წინასწარმეტყველების*) ან *გადაწყვეტილების მიღების* პრობლემა (ინგლისურად : *a prediction or decision problem*), რომელიც მდგომარეობს ამ კანონების გამოყენებაში $\vec{x}$ დაკვირვების კლასის დასადგენად.

$\vec{x}$ დაკვირვების კლასის *საწინასწარმეტყველოდ* გონივრულად გამოიყურება ყველაზე დიდი ალბათობის კლასის დადგენა ამ დაკვირვების გათვალისწინებით. შემდეგ უნდა მოხდეს $\mathbb{P}(Y=1 \mid X=\vec{x})$ და $\mathbb{P}(Y=0 \mid X=\vec{x})$ ალბათობათა შედარება. ფორმალურად, განიხილება შემდეგი სახის *გადაწყვეტილებათა მიღების წესი* (ინგლისურად : *the decision rule*) :

$$\hat{y} = \begin{cases} 1 & \text{თუ } \mathbb{P}(Y=1 \mid X=\vec{x}) > \mathbb{P}(Y=0 \mid X=\vec{x}) \\ 0 & \text{წინააღმდეგ შემთხვევაში} \end{cases} . \qquad (4.1)$$

მრავალკლასიან შემთხვევაზე გასავრცობად ეს წესი შეიძლება შემდეგნაირადაც ჩაიწეროს :

$$\hat{y} = \arg\max_{c=1,\ldots,C} \mathbb{P}(Y = c \mid X = \vec{x}).$$

გადაწყვეტილებათა მიღების ამ წესზე (ზოგიერთ სხვა წესთან ერთად), უფრო დაწვრილებითი საუბარი გვექნება პუნქტში 4.2.

ამ მომენტიდან — $\mathbb{P}\left(X = \vec{x}\right)$ აღნიშვნის ნაცვლად — ნახმარი იქნება უფრო კომპაქტური $\mathbb{P}\left(\vec{x}\right)$ ნოტაცია, თუ ამით ტექსტს ორაზროვანი ხასიათი არ მიეცემა.

1.2 ბაიესის კანონი

ნებისმიერი ალბათური დანასკვი მნიშვნელოვან წილად ეყრდნობა ბაიესის კანონს, რომელიც $\mathbb{P}\left(Y = c \mid \vec{x}\right)$ პირობითი განაწილების დადგენის საშუალებას იძლევა შემდეგი სახით :

თეორემა 4.1 (ბაიესის კანონი)

$$\mathbb{P}\left(Y = c \mid \vec{x}\right) = \frac{\mathbb{P}\left(Y = c\right) \cdot \mathbb{P}\left(\vec{x} \mid Y = c\right)}{\mathbb{P}\left(\vec{x}\right)}. \tag{4.2}$$

ამ განტოლების ყოველ ელემენტს აქვს გარკვეული დანიშნულება და სახელწოდებაც :

– $\mathbb{P}\left(Y = c\right)$ არის ჭდეთა *აპრიორული განაწილება* მონაცემთა დაკვირვებამდე ;

– $\mathbb{P}\left(\vec{x} \mid Y = c\right)$ არის *დამაჯერებლობა* (ინგლისურად : the *likelihood*). იგი განსაზღვრავს რამდენად დამაჯერებელია აკვირდებოდე X ვექტორის $\vec{x}$ რეალიზაციას და იცოდე, რომ საქმე გაქვს c კლასთან ;

– $\mathbb{P}\left(Y = c \mid \vec{x}\right)$ არის ჭდეთა *აპოსტერიორული განაწილება* (ინგლისურად : *post distribution*) მონაცემებზე დაკვირვებათა ჩატარების შემდეგ.

$\mathbb{P}\left(\vec{x}\right)$ მნიშვნელი არის მარგინალური (კერძო) ალბათობაა, რომლითაც ხდება $\vec{x}$ მაგალითის დაკვირვება, მისი კლასისგან დამოუკიდებლად. ამ ალბათობის გადაწერა შესაძლებელია ასეთი ფორმით : $\mathbb{P}\left(\vec{x}\right) = \mathbb{P}(\vec{x} \mid Y = 0)\mathbb{P}\left(Y = 0\right) + \mathbb{P}(\vec{x} \mid Y = 1)\mathbb{P}\left(Y = 1\right)$. მრავალკლასიან შემთხვევაში მნიშვნელი ჩაიწერება შემდეგი სახით : $\mathbb{P}\left(\vec{x}\right) = \sum_{c=1}^{C} \mathbb{P}\left(\vec{x} \mid Y = c\right)\mathbb{P}\left(Y = c\right)$.

მაგალითი

განვიხილოთ საშვილოსნოს ყელის კიბოს სისხლის ნაცხით სკრინინგის შემთხვევა. თუმცა ეს ტესტი შესასრულებლად მარტივია, იგი საკმაოდ უზუსტოა : მისი მგრძნობელობა (დაავადებული ქალების წილი, რომლებისთვისაც ტესტი დადებითია) შეადგენს, დაახლოებით, 70% (სწორედ ამიტომ ითხოვენ ტესტის რეგულარულ ჩატარებას), ხოლო სპეციფიკურობა (ჯანმრთელი ქალების წილი, რომლებისთვისაც ტესტი უარყოფითი) —

დაახლოებით 98%. (უფრო დაწვრილებითი ცნობების მისაღებად მგრძნობელობისა და სპეციფიკურობის შესახებ იხილეთ ლექციათა ამ კურსის განყოფილება 3.2.1.). გარდა ამისა, ავადობა შეადგენს 1 ქალს $10\,000$ ქალზე ($6,7$ ქალს ყოველ 100000 ქალზე 2012 წელს).

როგორია ალბათობა იმისა, რომ ქალი, რომელმაც გაიარა ტესტი, დაავადებულია კიბოთი, თუ ტესტი აღმოჩნდება დადებითი? სწორედ რომ ეს არის ბაიესური დანასკვის პრობლემა.

დავუშვათ, რომ X არის შემთხვევითი ცვლადი მნიშვნელობებით $\{0,1\}-$ზე, რომლებითაც ტესტის შედეგების მოდელირება ხდება (1 – დადებითისთვის, 0 – უარყოფითისთვის), ასევე დავუშვათ, რომ Y არის შემთხვევითი ცვლადი მნიშვნელობებით $\{0,1\}-$ზე, რომლებითაც ხდება იმ ქალის სტატუსის მოდელირება, რომელსაც გასავლელი აქვს ტესტი, (1 – ავადმყოფისთვის, 0 – ჯანმრთელისთვის).

დანასკვი : ჩვენ ვცდილობთ $\mathbb{P}(Y=1|\,X=1)$ ალბათობის გამოთვლას. მოდით გამოვიყენოთ ამისთვის ბაიესის კანონი :

$$\mathbb{P}(Y=1|\,X=1)=\frac{\mathbb{P}(X=1|\,Y=1)\cdot\mathbb{P}(Y=1)}{\mathbb{P}(X=1)}.$$

$\mathbb{P}(X=1|\,Y=1)$ ეს სხვა არაფერია, თუ არა ტესტის მგრძნობელობა, სხვანაირად რომ ვთქვათ, $\mathbb{P}(X=1|\,Y=1)=0,70$. $\mathbb{P}(Y=1)$ არის საშვილოსნოს ყელის კიბოს გავრცელებადობა ქალებში, ანუ 10^{-4} . დაბოლოს,

$$\mathbb{P}(X=1)=\mathbb{P}(X=1|\,Y=1)\mathbb{P}(Y=1)+\mathbb{P}(X=1|\,Y=0)\mathbb{P}(Y=0)\,.$$

ჩვენ უკვე ვიცით ამ განტოლების პირველი ორი წევრი. გარდა ამისა,

$$\mathbb{P}(X=1|\,Y=0)=1-0,90=0,02 \text{ და } \mathbb{P}(Y=0)=1-\mathbb{P}(Y=1)=0,9999\,.$$

ასე რომ,

$$\mathbb{P}(Y=1|\,X=1)=\frac{0,70\times10^{-4}}{0,70\times10^{-4}+0.02\times0,9999}=0,0035.$$

წინასწარმეტყველება : ამრიგად, ალბათობა იმისა, რომ ქალი დადებითი ტესტით ავადმყოფია, მხოლოდ $0,35\%$ შეადგენს. ყოველგვარი დამატებითი ინფორმაციის გარეშე :

$$\mathbb{P}(Y=0|\,X=1)>\mathbb{P}(Y=0|\,X=1)$$

და გადაწყვეტილების მიღების წესი 4.1 დაგვიბრუნებს უარყოფით პროგნოზს ყველა დადებითი ტესტისათვის.

1.3 პარამეტრული მოდელირება

წინა მაგალითში $\mathbb{P}(X\,|\,Y)$ ალბათობა მოცემული იყო. მაგრამ შესაძლებელია, რომ ჩვენ ასევე მოგვიხდეს ხსენებული განაწილების მოდელირებაც. ამ შემთხვევაში ჩვენ ვზღუდავთ მის

კუთვნილებას ისეთი ალბათური კანონების გარკვეული ოჯახით, რომლებიც პარამეტრიზებულია $\vec{\theta}$ ვექტორით და ამ უკანასკნელის მნიშვნელობები მოცემულია სასრული განზომილების Θ სივრციდან : ეს სწორედ ისაა, რასაც *პარამეტრული მოდელირება* ეწოდება.

დანასკვის (დასკვნის) და წინასწარმეტყველების (პროგნოზის) გარდა ახლა ჩვენ დამატებითი ამოცანაც გვაქვს : ეს არის *სწავლება*, რომელიც მდგომარეობს პარამეტრთა $\vec{\theta}$ ვექტორის მნიშვნელობის შეფასებაში.

როგორ უნდა გაკეთდეს ეს, ჩვენ დავინახავთ ქვედანაყოფში 4.3.

2 გადაწყვეტილების მიღების წესი

ამ ნაწილში ჩვენ დავინახავთ, როგორ უნდა იქნეს მიღებული გადაწყვეტილება ან გაკეთდეს პროგნოზი აუცილებელ ალბათურ კანონებზე დაყრდნობით.

2.1 დამაჯერებლობის ფარდობის ტესტები

გადაწყვეტილების მიღების წესი (განტოლება 4.1), რომელიც ყველაზე უფრო სავარაუდო კლასის პროგნოზს იძლევა მოცემული დაკვირვების გათვალისწინებით, მდგომარეობს აპოსტერიორული $\mathbb{P}(Y=\hat{y}\,|\,\vec{x})$ ალბათობის მნიშვნელობისთვის *მაქსიმუმის მიმნიჭებელი* $\hat{y}$ კლასის არჩევაში.

ასე რომ, ლაპარაკია *გადაწყვეტილების მიღებაზე აპოსტერიორული ალბათობის მაქსიმუმით.* ბაიესის კანონის გამოყენებით მას შეიძლება ასეთი სახეც მიეცეს :

$$\hat{y}=\begin{cases}1 & \text{თუ } \dfrac{\mathbb{P}\left(\vec{x}\,|\,Y=1\right)\cdot\mathbb{P}\left(Y=1\right)}{\mathbb{P}(\vec{x})}>\dfrac{\mathbb{P}\left(\vec{x}\,|\,Y=0\right)\cdot\mathbb{P}\left(Y=0\right)}{\mathbb{P}(\vec{x})}\\ 0 & \text{წინააღმდეგ შემთხვევაში}\end{cases}.$$

$\mathbb{P}(\vec{x})$ ალბათობა გავლენას არ ახდენს გადაწყვეტილებათა მიღების ამ წესზე და ამიტომ შეიძლება გამოირიცხოს. მაგრამ უნდა შევნიშნოთ, რომ $\mathbb{P}(\vec{x})$ ალბათობა დაგვჭირდება, თუ მოვინდომებთ $\mathbb{P}(Y=\hat{y}\,|\,\vec{x})$ აპოსტერიორული მნიშვნელობის შეფასებას ჩვენი გადაწყვეტილების ხარისხის რაოდენობრივად დახასიათებისთვის.

განსაზღვრება 4.1 (გადაწყვეტილების მიღება აპოსტერიორული ალბათობის მაქსიმუმით) ბინარულ შემთხვევაში გადაწყვეტილების მიღების

$$\hat{y}=\begin{cases}1 & \text{თუ } \mathbb{P}\left(\vec{x}\,|\,Y=1\right)\cdot\mathbb{P}\left(Y=1\right)>\mathbb{P}\left(\vec{x}\,|\,Y=0\right)\cdot\mathbb{P}\left(Y=0\right)\\ 0 & \text{წინააღმდეგ შემთხვევაში}\end{cases} \tag{4.3}$$

წესს ეწოდება *გადაწყვეტილების მიღების წესი აპოსტერიორული ალბათობის მაქსიმუმით.*

მრავალკლასიან შემთხვევაში ეს წესი შემდეგი სახით ჩაიწერება :

$$\hat{y}=\arg\max_{c=1,2,\ldots,C}\mathbb{P}(\vec{x}\,|\,Y=c)\cdot\mathbb{P}\left(Y=c\right).$$

შესაძლებელია ამ წესის ხელახალი ჩაწერაც დამაჯერებლობის ფარდობის გამოყენებით.

განსაზღვრება 4.2 (დამაჯერებლობის ფარდობა) $\Lambda\left(\vec{x}\right)$ აღნიშვნით წარმოადგენენ

$$\Lambda\left(\vec{x}\right)=\frac{\mathbb{P}\left(\vec{x}\,|\,Y=1\right)}{\mathbb{P}\left(\vec{x}\,|\,Y=0\right)}$$

დამაჯერებლობის ფარდობას. გადაწყვეტილების მიღების წესი აპოსტერიორული ალბათობის მაქსიმუმით შეიძლება იქნეს ხელახლა ფორმულირებული, როგორც *დამაჯერებლობის ფარდობის ტესტი* :

$$\hat{y}=\begin{cases}1 & \text{თუ } \Lambda\left(\vec{x}\right)>\dfrac{\mathbb{P}\left(Y=0\right)}{\mathbb{P}\left(Y=1\right)}\\ 0 & \text{წინააღმდეგ შემთხვევაში}\end{cases}. \qquad (4.4)$$

იმ შემთხვევაში, როცა ჰიპოთეზის სახით იგულისხმება აპრიორული ალბათობების ტოლობა, $\mathbb{P}\left(Y=0\right)=\mathbb{P}\left(Y=1\right)$ და დამაჯერებლობის ფარდობის შედარება მოხდება $1-$თან.

განსაზღვრება 4.3 (გადაწყვეტილების მიღება დამაჯერებლობის მაქსიმუმით) გადაწყვეტილების მიღების

$$\hat{y}=\begin{cases}1 & \text{თუ } \mathbb{P}\left(\vec{x}\,|\,Y=1\right)>\mathbb{P}\left(\vec{x}\,|\,Y=0\right)\\ 0 & \text{წინააღმდეგ შემთხვევაში}\end{cases}$$

წესს ან, ეკვივალენტური სახით,

$$\hat{y}=\begin{cases}1 & \text{თუ } \Lambda\left(\vec{x}\right)>1\\ 0 & \text{წინააღმდეგ შემთხვევაში}\end{cases}$$

დებულებას ეწოდება *გადაწყვეტილების მიღების წესი დამაჯერებლობის მაქსიმუმით* (ინგლისურად : *maximum likelihood decision rule*).

შენიშვნა

მრავალ შემთხვევაში უპირატესობა მიეცემა გალოგარითმებას და $\log\Lambda\left(\vec{x}\right)$ გამოსახულების ნიშნის დადგენას.

მაგალითი

დავუშვათ, რონ გვაქვს ანარჩევი ერთი სახის თევზების პოპულაციიდან და მათ შორის არის მამრი და მდედრი თევზები. ჩვენ ვცდილობთ განვსაზღვროთ მათი სქესი, რომლის მოდელირება ხდება ბინარული შემთხვევითი Y ცვლადით (მნიშვნელობით 0 მამრისთვის და მნიშვნელობით 1 მდედრისთვის) და უწყვეტი შემთხვევითი X ცვლადის მიერ

მოდელირებული თევზის მხოლოდ სიგრძით. ასევე დავუშვათ, რომ მამრი თევზის სიგრძე განაწილებულია ნორმალურად 4 სანტიმეტრის ტოლი ცენტრით და 1 სანტიმეტრის ტოლი სტანდარტული გადახრით, ხოლო მდედრი თევზის სიგრძე განაწილებულია ნორმალურად, მაგრამ 6 სანტიმეტრის ტოლი ცენტრით და იმავე 1 სანტიმეტრის ტოლი სტანდარტული გადახრით :

$$\left.\begin{aligned}\mathbb{P}(x\,|\,Y=0)\sim\mathcal{N}\left(4,1\right)\\ \mathbb{P}(x\,|\,Y=1)\sim\mathcal{N}\left(6,1\right)\end{aligned}\right\}.$$

დამაჯერებლობის ფარდობა ჩაიწერება

$$\frac{\mathbb{P}\left(x\,|\,Y=1\right)}{\mathbb{P}\left(x\,|\,Y=0\right)}=\frac{e^{-(x-6)^2/2}}{e^{-(x-4)^2/2}}$$

სახით და მის ლოგარითმს, ამრიგად, ექნება შემდეგი ფორმა :

$$\ln\Lambda(x)=-\left(x-6\right)^2+\left(x-4\right)^2=4\left(x-5\right).$$

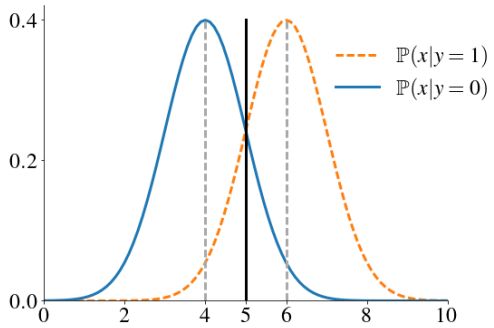
ამრიგად, თუ დავუშვებთ, რომ მამრთა და მდედრთა წილი თევზების ჩვენს სტატისტიკურ ანარჩევში ერთნაირია, მაშინ გადაწყვეტილების მიღების წესი მაქსიმალური აპოსტერიორული ალბათობით და გადაწყვეტილების მიღების წესი დამაჯერებლობის მაქსიმუმით ერთმანეთის ეკვივალენტურია (ტოლფასია) : ასე რომ, შეიძლება ვიწინასწარმეტყველოთ, რომ თევზი მდედრია, თუ მისი სიგრძე მეტია 5 სანტიმეტრზე, ხოლო წინააღმდეგ შემთხვევაში თევზი მამრია. გადაწყვეტილების მიღების ეს წესი ნაჩვენებია ნახატზე 4.1 *a*) .

ახლა, ვთქვათ, ცნობილია, რომ თევზების ჩვენს სტატისტიკურ ანარჩევში მდედრების რაოდენობა ხუთჯერ აღემატება მამრების რაოდენობას. ამიტომ აპრიორულ განაწილებათა ფარდობა $\frac{\mathbb{P}\left(Y=0\right)}{\mathbb{P}\left(Y=1\right)}=\frac{1}{5}$ სიდიდის ტოლია. ამრიგად, ჩვენ ვადარებთ დამაჯერებლობის ფარდობის ლოგარითმს $\ln\frac{1}{5}$ მნიშვნელობას. ჩვენ ვიწინასწარმეტყველებთ, რომ თევზი მდედრია, თუ მისი სიგრძე მეტია, ვიდრე $5-\ln\left(5\right)/4\approx4.58$: უწინდელზე უფრო სავარაუდო და ნათელი გახდა, რომ თევზი 5 სანტიმეტრზე ოდნავ ნაკლები სიგრძით - მდედრია. ხდება ზღურბლური მნიშვნელობის ძვრა (წანაცვლება) პრობლემის შესახებ ჩვენი აპრიორული ცოდნის შესაბამისად. გადაწყვეტილების მიღების ეს წესი ნაჩვენების ნახატზე 4.1 *b*) .

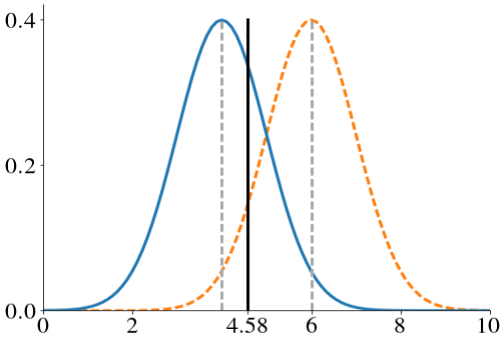
2.2 გადაწყვეტილების მიღების ბაიესური თეორია

ჩვენ მიერ ახლახან განხილული გადაწყვეტილებების მიღების წესები არის *გადაწყვეტილებათა მიღების თეორიის* უფრო ზოგადი სტრუქტურის ნაწილი. მის ფარგლებში $\mathcal{Y}$ -ზე განსაზღვრული შემთხვევითი Y ცვლადი წარმოადგენს არა ჭდეს, არამედ ფარულ «ჭეშმარიტებას», ანუ ბუნების რაღაც ერთ მდგომარეობას. $\mathcal{X}$ -ზე განსაზღვრული X — შემთხვევითი ცვლადია, რომელიც წარმოადგენს დაკვირვებათა მონაცემებს. გარდა ამისა, დავუშვათ, რომ გვაქვს A ცვლადი, განსაზღვრული $\mathcal{A}$ სივრცეზე, რომელიც წარმოადგენს

შესაძლო *გადაწყვეტილებათა* სიმრავლეს. ახლა კი შემოვიღოთ *დანახარჯების* (*დანაკარგების*, *ზარალის*, *წაგების*) *ფუნქცია* (ინგლისურად : *loss function*), $L : \mathcal{Y} \times \mathcal{A} \to \mathbb{R}$. დანახარჯების (დანაკარგების, ზარალის, წაგების) ეს ფუნქცია იმ ფუნქციათა ანალოგიურია, რომლებიც განისაზღვრა ამ კურსის მეორე ლექციის მე-4 პუნქტში. a მოქმედებისა და y ჭეშმარიტი ფარული მდგომარეობის გათვალისწინებით ეს ფუნქცია განსაზღვრავს ფასს, რომლის გადახდა მოგვიწევს a მოქმედების არჩევისათვის, მაშინ როცა ჭეშმარიტი ფარული მდგომარეობა იყო y.



a) თუ ორივე კლასი ერთნაირი ალბათობით ჩნდება, თევზები 5 სანტიმეტრზე ნაკლები ზომით მოინიშნება მამრებად, დანარჩენები კი - მდედრებად.



b) თუ ალბათობა იმისა, რომ თევზი მდედრია, აპრიორულად ხუთჯერ აღემატება საწინააღმდეგოს, მაშინ 4,58 სანტიმეტრზე ნაკლები ზომის თევზები მოინიშნება მამრებად, დანარჩენები კი - მდედრებად.

ნახატი 4.1 - აკვარიუმის თევზის — გუფის (ლათ. *poecilia reticulata*) — სქესზე გადაწყვეტილების მიღების წესი, თევზის ზომის მიხედვით.

მაგალითი

ღირს ამ დილით ქოლგის წაღება თუ არა ? ჩვენ შეგვიძლია ამ პრობლემის მოდელირება შემდეგი სახით : $\mathcal{A}$ შეიცავს ორ ქმედებას («ქოლგით გამოსვლა» და «უქოლგოდ გამოსვლა»). $\mathcal{Y}$ შეიცავს ჭეშმარიტებებს «არ წვიმს», «მოდის სუსტი წვიმა», «მოდის ძლიერი წვიმა», «ქრის ქარი». დაბოლოს, $\mathcal{X}$ არის სივრცე იმ ინფორმაციის აღსაწერად, რომელსაც შეიძლება დავეყრდნოთ, მაგალითად, ამინდის პროგნოზსა და ცის ფერს სახლიდან გამოსვლისას. შეიძლება დანახარჯების შემდეგი ფუნქციის არჩევა :

$\mathcal{A}$ \ $\mathcal{Y}$	არ წვიმს	სუსტი წვიმა	ძლიერი წვიმა	ქარი
დილით სახლიდან ქოლგით გამოსვლა	1	0	0	2
დილით სახლიდან უქოლგოდ გამოსვლა	0	2	4	0

ასეთ ვითარებაში, ცხადია, გონივრული იქნება ისეთი a ქმედების არჩევა, რომელიც მინიმუმს მიანიჭებს შეცდომის ალბათობას, ესე იგი დანახარჯების ფუნქციის მათემატიკურ ლოდინსაც.

განსაზღვრება 4.4 (ბაიესის გადაწყვეტილება) გადაწყვეტილების მიღების წესს ისეთი a^* ქმედების არჩევის შესახებ, რომელიც მინიმუმს ანიჭებს დანახარჯის ფუნქციის მათემატიკურ ლოდინს, *გადაწყვეტილების მიღების ბაიესის წესი* ეწოდება :

$$a^*(\vec{x}) = \arg\min_{a\in\mathcal{A}} \mathbb{E}\big[L(y,a)\big] = \arg\min_{a\in\mathcal{A}} \sum_{y\in\mathcal{Y}} \mathbb{P}(Y=y\mid\vec{x})L(y,a). \tag{4.5}$$

ამ შემთხვევაში ასევე ლაპარაკობენ *მოსალოდნელი დანაკარგების (ზარალის) მინიმიზაციის პრინციპის* შესახებ (ინგლისურად : *minimum expected loss*).

შენიშვნა

ეკონომიკაში დანახარჯების ფუნქციის ცნებას უპირატესობა ეძლევა *სარგებლობის* ცნებასთან შედარებით. ენციკლოპედიური განმარტებით, *სარგებლობა* იგივეა, რაც *სარგებელი* (თავისი ერთ-ერთი მთავარი გაგებით) და არის დადებითი, კეთილმყოფელი გავლენა. კარგი შედეგი, სიკეთე და კონკრეტულად ნიშნავს მოგებას, გამორჩენას, ხეირს, სარფს. მარტივად, სარგებლობა შეიძლება განიმარტოს როგორც დანახარჯების ფუნქციის საწინააღმდეგო ცნება. ასეთ შემთხვევაში ზემოთ მოცემული პრინციპი შეიძლება მივიჩნიოთ *მოსალოდნელი სარგებლობის მაქსიმიზაციად* (ინგლისურად : *maximum expected utility*).

ეს მიდგომა უნდა დაუპირისპირდეს რისკის ემპირიულ მინიმიზაციას (იხ. მე-2 ლექციის მე-3 პუნქტი), როდესაც $\mathbb{P}(X,Y)$ განაწილების ჩანაცვლება ხდება მისი *ემპირიული განაწილებით* და ეს უკანასკნელი მიიღება ალბათური მასის თანაბრად გადანაწილების გზით n დაკვირვებას შორის :

$$\mathbb{P}\big(X=\vec{x},Y=y\mid\mathcal{D}\big) = \frac{1}{n}\sum_{i=1}^{n}\delta\big(y,y^i\big)\delta\big(\vec{x},\vec{x}^i\big). \tag{4.6}$$

ამისგან განსხვავებით, ბაიესურ სისტემაში $\mathbb{P}(X,Y)$ განაწილება პარამეტრიზებულია ერთი $\vec{\theta}$ პარამეტრით, რომელიც ოპტიმიზებულია $\mathcal{D}-$ზე. ემპირიული სისტემის ფარგლებში ჰიპოთეზები მონაცემთა განაწილების შესახებ საკმაოდ გამარტივებულია : მაგრამ ბაიესურ სისტემაში ეს განაწილება შეისწავლება იმ პროცესის განხილვის გარეშე, რომელშიც მიღებული გადაწყვეტილება იქნება გამოყენებული.

მაშინ, როცა *ბაიესის გადაწყვეტილება* მდგომარეობს, ერთი მოცემული დაკვირვებისთვის, დანახარჯების ფუნქციის მოსალოდნელი მნიშვნელობის არჩევაში, *ბაიესის რისკი* უფრო ფართო ცნებაა და განისაზღვრება როგორც დანახარჯების ფუნქციის გლობალურად მოსალოდნელი მნიშვნელობა :

განსაზღვრება 4.5 (ბაიესის რისკი) *ბაიესის რისკი* არის დანახარჯის მოსალოდნელი მნიშვნელობა, რომელიც დადგენილია გადაწყვეტილების მიღების ბაიესის წესით :

$$r = \int_{\vec{x}\in\mathcal{X}} \left(\sum_{y\in\mathcal{Y}} L\big(y, a^*(\vec{x})\big)\mathbb{P}\big(\vec{x}, y\big) \right) \mathrm{d}\vec{x} \,. \tag{4.7}$$

ბაიესის რისკისთვის მინიმუმის მიმნიჭებელი სტრატეგიის განსაზღვრა გადაწყვეტილებათა მიღების ბაიესის წესის გამოყენების ტოლფასია.

ბინარული კლასიფიკაცია გადაწყვეტილების მიღების ბაიესის წესით

დავუბრუნდეთ ბინარული კლასიფიკაციის სისტემას. $y \in \mathcal{Y}$ არის დაკვირვების *ჭეშმარიტი* კლასი, მაშინ როცა $a \in \mathcal{A}$ წარმოადგენს *ნაწინასწარმეტყველებ* კლასს : $\mathcal{A} = \mathcal{Y} = \{0,1\}$. დანახარჯების ფუნქცია

$$\left.\begin{aligned} L : \mathcal{Y}\times\mathcal{Y} &\to \mathbb{R} \\ c, k &\mapsto \lambda_{ck} \end{aligned}\right\}$$

ახლა უკვე განსაზღვრავს k კლასის წინასწარმეტველების დანახარჯებს, მაშინ, როცა ჭეშმარიტი კლასია c (იხ. მე-2 ლექციის მე-4 ე ნაწილი). გადაწყვეტილებათა მიღების ბაიესის წესი (განტოლება 4.5) გადაწყვეტილებათა მიღების შემდეგი წესის ტოლფასია :

$$\hat{y} = \begin{cases} 1 & \text{თუ } \lambda_{11}\mathbb{P}\big(Y=1\,|\,\vec{x}\big) + \lambda_{10}\mathbb{P}\big(Y=0\,|\,\vec{x}\big) \le \lambda_{01}\mathbb{P}\big(Y=1\,|\,\vec{x}\big) + \lambda_{00}\mathbb{P}\big(Y=0\,|\,\vec{x}\big) \\ 0 & \text{წინააღმდეგ შემთხვევაში} \end{cases} . \tag{4.8}$$

მისი ჩაწერა დამაჯერებლობის ფარდობის ტესტის სახითაც შეიძლება :

$$\hat{y} = \begin{cases} 1 & \text{თუ } \dfrac{\mathbb{P}\big(\vec{x}\,|\,Y=1\big)}{\mathbb{P}\big(\vec{x}\,|\,Y=0\big)} > \dfrac{\big(\lambda_{01}-\lambda_{00}\big)\mathbb{P}\big(Y=0\big)}{\big(\lambda_{10}-\lambda_{11}\big)\mathbb{P}\big(Y=1\big)} \\ 0 & \text{წინააღმდეგ შემთხვევაში} \end{cases} . \tag{4.9}$$

მრავალი კლასის შემთხვევაში ეს წესი იძენს შემდეგ ფორმას :

$$\hat{y} = \arg\min_{c=1,\ldots,C} \sum_{k=1}^{K} \lambda_{kc}\mathbb{P}\big(Y=k\,|\,\vec{x}\big) .$$

ღირებულება 0/1

0/1 ღირებულებას (იხ. მე-2 ლექციის მე-4 ნაწილი) პოულობენ $\lambda_{ck} = 1-\delta(k,c)$ თანაფარდობის გამოყენებით. განტოლება 4.8 იძენს შემდეგ სახეს :

$$\hat{y} = \begin{cases} 1 & \text{თუ } \mathbb{P}\big(Y=0\,|\,\vec{x}\big) \le \mathbb{P}\big(Y=1\,|\,\vec{x}\big) \\ 0 & \text{წინააღმდეგ შემთხვევაში} \end{cases} . \tag{4.10}$$

ასე რომ, გადაწყვეტილების მიღების ბაიესის წესი და გადაწყვეტილების მიღების წესი აპოსტერიორული მაქსიმუმით ურთიერთტოლფასია. ეს სამართლიანია მრავალი კლასის შემთხვევაშიც.

0/1 ღირებულება არ არის ღირებულების ერთადერთი შესაძლო ფუნქცია ბინარული

კლასიფიკაციის ამოცანისთვისაც კი. კერძოდ, ყველა მცდარი კლასიფიკაცია ერთნაირად ძვირადღირებული როდია? მაგალითად, წინასწარმეტყველება, რომ კიბოთი დაავადებული ჯანმრთელია, შეიძლება გაცილებით უფრო დიდი საშიშროების შემცველი აღმოჩნდეს, ვიდრე პირიქით.

გადაწყვეტილებათა მიღების არეები (რეგიონები)

გადაწყვეტილებათა მიღების წესები ასევე შეიძლება იყოს გამოსახული გადაწყვეტილებათა მიღების არეების (რეგიონების) ტერმინებით (იხ. მე-2 ლექციის პირველი ნაწილი) : გადაწყვეტილებათა მიღების წესი უბრალოდ მდგომარეობს $\vec{x}$ დაკვირვების მონიშვნაში გადაწყვეტილებათა მიღების იმ არის შესაბამისად. რომელსაც იგი მიეკუთვნება :

$$\hat{y} = \begin{cases} 1 & \text{თუ } \vec{x} \in \mathcal{R}_1 \\ 0 & \text{წინააღმდეგ შემთხვევაში} \end{cases}. \tag{4.11}$$

მრავალი კლასის შემთხვევაში ეს წესი დაიყვანება შემდეგ თანაფარდობამდე :

$$\hat{y} = \sum_{c=1}^{C} \delta_{\vec{x} \in \mathcal{R}_c} .$$

გადაწყვეტილებათა მიღების ეს წესი გადაწყვეტილებათა მიღების ბაიესის წესის ეკვივალენტურია, თუ დისკრიმინანტულ ფუნქციად განსაზღვრული იქნება შემდეგი გამოსახულება :

$$f(\vec{x}) = \left(\lambda_{01}\mathbb{P}(Y=1|\vec{x}) + \lambda_{00}\mathbb{P}(Y=0|\vec{x})\right) - \left(\lambda_{11}\mathbb{P}(Y=1|\vec{x}) + \lambda_{10}\mathbb{P}(Y=0|\vec{x})\right), \tag{4.12}$$

ან, მრავალი კლასის პირობებში,

$$f_c(\vec{x}) = -\sum_{k=1}^{C} \lambda_{ck}\mathbb{P}(Y=k|\vec{x}).$$

0/1 დანახარჯების შემთხვევაში დისკრიმინანტული ფუნქცია $f(\vec{x}) = \mathbb{P}(Y=1|\vec{x}) - \mathbb{P}(Y=0|\vec{x})$ სახისაა და გადაწყვეტილების მიღების ბაიესის წესი გადაწყვეტილების მიღების ტოლფასია, მართლაც, აპოსტერიორული მაქსიმუმით (განტოლება 4.4).

ბაიესის რისკი (4.7) შეიძლება იყოს ჩაწერილი გადაწყვეტილებათა მიღების არეთა ფუნქციის სახითაც :

$$\begin{aligned} r &= \int_{\vec{x}\in\mathcal{X}} \left(\lambda_{0,a^*(\vec{x})}\mathbb{P}(\vec{x}|Y=0)\mathbb{P}(Y=0) + \lambda_{1,a^*(\vec{x})}\mathbb{P}(\vec{x}|Y=1)\mathbb{P}(Y=1)\right)\mathrm{d}\vec{x} \\ &= \int_{\vec{x}\in\mathcal{R}_0} \left(\lambda_{00}\mathbb{P}(\vec{x}|Y=0)\mathbb{P}(Y=0) + \lambda_{10}\mathbb{P}(\vec{x}|Y=1)\mathbb{P}(Y=1)\right)\mathrm{d}\vec{x} \\ &= \int_{\vec{x}\in\mathcal{R}_1} \left(\lambda_{01}\mathbb{P}(\vec{x}|Y=0)\mathbb{P}(Y=0) + \lambda_{11}\mathbb{P}(\vec{x}|Y=1)\mathbb{P}(Y=1)\right)\mathrm{d}\vec{x} \end{aligned} \tag{4.13}$$

გადაწყვეტილებათა მიღების არეთა ისეთნაირად განსაზღვრა, რომ მოხდეს ბაიესის რისკის მინიმიზება, 4.12 განტოლებით განსაზღვრული დისკრიმინანტული ფუნქციის გამოყენების ეკვივალენტურია. ეს ასევე სამართლიანია მრავალი კლასის შემთხვევაში, როცა

$$r = \sum_{k=1}^{C} \left(\int_{\vec{x} \in \mathcal{R}_k} \left(\sum_{c=1}^{C} \lambda_{ck} \mathbb{P}\left(\vec{x} \mid Y = c\right) \mathbb{P}\left(Y = c\right) \right) \mathrm{d}\vec{x} \right).$$

უარის თქმა გადაწყვეტილების მიღებაზე (ინგლისურად : Reject)

ზოგიერთ გამოყენებებში შეიძლება საინტერესო იყოს, რომ ალგორითმი უარს ამბობდეს გადაწყვეტილების მიღებაზე, როცა მისი ნდობა წინასწარმეტყველებისადმი ძალიან დაბალია, სხვა სიტყვებით, როცა დამაჯერებლობის ფარდობა ძალიან ახლოსაა $1-$ის ტოლ მნიშვნელობასთან. ამ შემთხვევაში შეიძლება *უარის თქმის* (ინგლისურად : *reject*) ერთი ხელოვნური ($C+1$ რიგითი ნომრის შესაბამისი) კლასის დამატება. 0/1 დანახარჯების მისადაგება ამ შემთხვევასთან შესაძლებელი გახდება, თუ შემოვიტანთ

$$\lambda_{ck} = \begin{cases} 0 & \text{თუ } k = c \\ \lambda & \text{თუ } k = C+1 \\ 1 & \text{წინააღმდეგ შემთხვევაში} \end{cases} \tag{4.14}$$

სიდიდეს, სადაც $0 < \lambda < 1$. გადაწყვეტილების მიღების წესი მოსალოდნელი დანაკარგების მინიმიზაციით იძენს შემდეგ სახეს :

$$\hat{y} = \begin{cases} c & \text{თუ } \mathbb{P}\left(Y = c \mid \vec{x} \quad\right) \geq \mathbb{P}\left(Y = k \mid \vec{x}\right) \quad \forall k \neq c \\ \text{reject} & \text{წინააღმდეგ შემთხვევაში} \end{cases} \tag{4.15}$$

ასეთ შემთხვევაში გადაწყვეტილებათა მიღების $\mathcal{R}_1, \ldots, \mathcal{R}_C$ არეები (რეგიონები) არ ფარავს $\mathcal{X}$ –ს, ხოლო მათი დამატებების გაერთიანება წარმოადგენს გადაწყვეტილების მიღებაზე უარის თქმის ზონას.

3 განაწილების სიმკვრივის შეფასება

დავუშვათ რომ გვაქვს $\mathcal{D} = x^1, x^2, \ldots, x^n$ ანარჩევი, რომელიც შედგება $\mathcal{X}-$ზე განსაზღვრული (მნიშვნელობების მქონე) X შემთხვევითი სიდიდის n დაკვირვებისგან. მომდევნო თხრობისას ნაგულისხმევი იქნება, რომ $X-$ის განაწილება მოცემული ტიპისაა და იგი პარამეტრიზებულია θ პარამეტრით. როგორ უნდა *შეფასდეს* θ ?

3.1 დამაჯერებლობის მაქსიმუმით შეფასება

განსაზღვრება 4.6 (მაქსიმალური დამაჯერებლობის შეფასება) θ პარამეტრის *მაქსიმალური დამაჯერებლობის შეფასება* (ინგლისურად : *maximum likelihood estimator*, *MLE*) არის $\hat{\theta}_{\text{MLE}}$ ვექტორი, რომელიც მაქსიმუმს ანიჭებს დამაჯერებლობას, სხვანაირად, $\mathcal{D}-$ს დაკვირვების ალბათობას მოცემული θ პარამეტრის პირობებში :

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} \mathbb{P}\left(\mathcal{D} \mid \theta\right). \tag{4.16}$$

⚙

იმისათვის, რომ ვიპოვოთ $\hat{\theta}_{\text{MLE}}$ სიდიდე, ჩვენ დავუშვებთ, რომ n დაკვირვება *დამოუკიდებელია* და ამასთან ერთად *განაწილებულია იდენტურად* (ინგლისურად :

independent and *identically distributediid*, *iid*). ეს დამაჯერებლობის დაშლის (დეკომპოზიციის) საშუალებას იძლევა შემდეგი სახით :

$$\mathbb{P}\left(\mathcal{D} \mid \theta\right) = \prod_{i=1}^{n} \mathbb{P}\left(X = \vec{x}^i \mid \theta\right). \tag{4.17}$$

დაბოლოს, უნდა აღინიშნოს, რომ გამოთვლათა გასამარტივებლად ხშირად მიმართავენ დამაჯერებლობის ლოგარითმის — და არა უშუალოდ დამაჯერებლობის — მაქსიმიზაციას :

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} \sum_{i=1}^{n} \log \mathbb{P}\left(X = \vec{x}^i \mid \theta\right). \tag{4.18}$$

მაგალითი

ავიღოთ მონეტის ზევით აგდების (ასროლის) მაგალითი. ამ შემთხვევაში მონეტის უკანა და წინა პირის დაკვირვების მოდელირება გარკვეული X შემთხვევითი სიდიდის რეალიზაციით ხდება. ეს შემთხვევითი სიდიდე განსაზღვრულია $\mathcal{X} = \{0,1\}$ სიმრავლეზე (0 შეესაბამება უკანა პირს, ხოლო 1 — წინა პირს) და ემორჩილება $\mathbb{P}$ ალბათობის კანონს. კლასიკურ არჩევანს ალბათობის ამ კანონისთვის *ბერნულის კანონის* გამოყენება წარმოადგენს :

$$\mathbb{P}\left(X = x\right) = \begin{cases} p & \text{თუ } x = 1 \\ (1-p) & \text{თუ } x = 0 \end{cases}, \tag{4.19}$$

ანუ, ეკვივალენტური სახით, $\mathbb{P}\left(X = x\right) = p^x (1-p)^{1-x}$. დავუშვათ, რომ $\mathcal{D} = \left\{x^1, x^2, \ldots, x^n\right\}$ შედგენილია *დამოუკიდებელი და ერთნაირად განაწილებული* n დაკვირვებით. (4.18) განტოლების თანახმად, $p-$ს მაქსიმალური დამაჯერებლობის შეფასებას შემდეგი სახე აქვს :

$$\begin{aligned} p_{\text{MLE}} &= \arg\max_{p\in[0,1]} \sum_{i=1}^{n} \log \mathbb{P}\left(X = x^i \mid p\right) = \arg\max_{p\in[0,1]} \sum_{i=1}^{n} \log\left(p^{x^i}\left(1-p\right)^{1-x^i}\right) \\ &= \arg\max_{p\in[0,1]} \left(\sum_{i=1}^{n} x^i \log p + \left(n - \sum_{i=1}^{n} x^i\right) \log(1-p)\right). \end{aligned}$$

$L : p \mapsto \sum_{i=1}^{n} x^i \log p + \left(n - \sum_{i=1}^{n} x^i\right) \log(1-p)$ ფუნქცია ჩაზნექილია და ამიტომ შესაძლებელია მისი მაქსიმიზაცია წარმოებულის განულებით :

$$\frac{\partial L}{\partial p} = \sum_{i=1}^{n} x^i \frac{1}{p} - \left(n - \sum_{i=1}^{n} x^i\right) \frac{1}{1-p},$$

რაც გვაძლევს, რომ

$$\left(1 - \hat{p}_{\text{MLE}}\right) \sum_{i=1}^{n} x^i - \hat{p}_{\text{MLE}} \left(n - \sum_{i=1}^{n} x^i\right) = 0$$

და, ამრიგად,

$$\hat{p}_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n} x^i \ . \tag{4.20}$$

მაქსიმალური დამაჯერებლობის შეფასება p სიდიდისათვის უბრალოდ ანარჩევის საშუალო მნიშვნელობას წარმოადგენს.

მაგალითი

კვლავ დავუბრუნდეთ მაგალითს, რომელიც ეხებოდა საშვილოსნოს ყელის კიბოს სკრინინგულ ტესტს. თუ ხსენებულ მაგალითში მოცემული იყო $\mathbb{P}(X \mid Y=0)$ და $\mathbb{P}(X \mid Y=1)$ სიდიდეები, ახლა ჩვენ შევუდგებით მათ შეფასებას დაკვირვების მონაცემთა საფუძველზე. დავუშვათ, რომ ჩვენ ვაკვირდებით ჯანმრთელად მიჩნეული n_0 ადამიანის $\mathcal{D}_0$ ნაკრებს, რომელთა შორის t_0 პირის ტესტი უარყოფითია, და ავადმყოფად მიჩნეული n_1 ადამიანის $\mathcal{D}_1$ ნაკრებს, რომელთა შორის t_1 პირის ტესტი დადებითია დაავადების გავრცელების მაჩვენებელი შეადგენს $\mathbb{P}\big(Y=1\big)=\pi$ სიდიდეს. $\mathbb{P}(X \mid Y=1)$ ალბათობის მოდელირება შეიძლება p_1 პარამეტრის მქონე ბერნულის კანონით, ხოლო $\mathbb{P}(X \mid Y=0)$ ალბათობის მოდელირება შეიძლება p_0 პარამეტრის მქონე ბერნულის კანონით.

ალბათობა იმისა, რომ დადებითი ტესტის მქონე პირს აქვს კიბო, შეადგენს შემდეგ სიდიდეს :

$$\mathbb{P}\big(Y=1 \mid X=1\big)=\frac{\mathbb{P}\big(X=1 \mid Y=1\big)\cdot\mathbb{P}\big(Y=1\big)}{\mathbb{P}\big(X=1\big)}.$$

ვიცით, რომ $\mathbb{P}(X=x \mid Y=0)=\big(p_9\big)^x\big(1-p_0\big)^{1-x}$ და $\mathbb{P}(X=x \mid Y=1)=\big(p_1\big)^x\big(1-p_1\big)^{1-x}$. ამიტომ,

$$\mathbb{P}\big(Y=1 \mid X=1\big)\frac{p_1\pi}{p_1\pi+p_0\big(1-\pi\big)}.$$

ჩვენ შეგვიძლია ჩავანაცვლოთ ამ განტოლებაში p_0 და p_1 სიდიდეები მათი მაქსიმალური დამაჯერებლობის შეფასებებით (განტოლება 4.20) :

$$p_0=1-\frac{t_0}{n_0} \ \text{ და } \ p_1=1-\frac{t_1}{n_1}. \tag{4.21}$$

საქმე ეხება შეფასებულ *ტესტის სპეციფიკურობასა* და *მგრძნობელობას,* შესაბამისად. $\frac{t_0}{n_0}=0,98,\ \ \frac{t_1}{n_1}=0,70$ და $\pi=10^{-5}$ მნიშვნელობათა გამოყენებით წინა შედეგები მიიღება.

3.2 ბაიესის შემფასებელი

უარი ვთქვათ θ პარამეტრის შესაძლო მნიშვნელობათა სრული უცოდნელობის ვარაუდზე და დავუშვათ, რომ ჩვენ — ექსპერტებს გამოყენების სფეროში — გვაქვს საკმაოდ კარგი წარმოდგენა მნიშვნელობებზე, რომლებიც ამ პარამეტრმა შეიძლება მიიღოს. ეს ინფორმაცია ძალიან სასარგებლოა, განსაკუთრებით მაშინ, როცა დაკვირვებათა რაოდენობა მცირეა. მის გამოსაყენებლად ჩვენ განვახორციელებთ θ პარამეტრის მოდელირებას შემთხვევითი ცვლადის სახით და განვსაზღვრავთ ამ ცვლადის აპრიორულ $\mathbb{P}(\theta)$ განაწილებას.

განსაზღვრება 4.7 (ბაიესის შეფასება) როცა მოცემულია ღირებულების L ფუნქცია, *ბაიესის შეფასება* $\hat{\theta}_{\text{Bayes}}$ θ პარამეტრისათვის განისაზღვრება როგორც

$$\hat{\theta}_{\text{Bayes}} = \arg\min_{\hat{\theta}} \mathbb{E}\Big[L\big(\theta, \hat{\theta}\big)\Big]. \tag{4.22}$$

თუ L ფუნქციისათვის გამოიყენება საშუალო კვადრატული შეცდომა, მაშინ

$$\hat{\theta}_{\text{Bayes}} = \arg\min_{\hat{\theta}} \mathbb{E}\Big[\big(\theta - \hat{\theta}\big)^2 \Big]. \tag{4.23}$$

$\hat{\theta}$ შეფასების განხილვისას დეტერმინისტულ სიდიდედ, გვაქვს :

$$\left.\begin{aligned} \hat{\theta}_{\text{Bayes}} &= \arg\min_{\hat{\theta}} \big(\hat{\theta}^2 - 2\hat{\theta}\mathbb{E}[\theta] + \mathbb{E}\big[\theta^2\big] \big) \\ &= \arg\min_{\hat{\theta}} \big(\hat{\theta} - \mathbb{E}[\theta] \big)^2 - \big(\mathbb{E}[\theta] \big)^2 + \mathbb{E}\big[\theta^2\big] \\ &= \mathbb{E}[\theta] \end{aligned}\right\}.$$

უკანასკნელი ტოლობა მიიღება, თუ შევნიშნავთ, რომ არც $\mathbb{E}[\theta]$ და არც $\mathbb{E}\big[\theta^2\big]$ დამოკიდებული არ არის $\hat{\theta}$ სიდიდეზე. ეს მათემატიკური ლოდინი აიღება θ და X ცვლადების განაწილებიდან, რომელსაც ჩვენ ვიყენებთ θ მნიშვნელობის შესაფასებლად ; ასე რომ,

$$\hat{\theta}_{\text{Bayes}} = \mathbb{E}[\theta \mid X] = \int \theta \cdot \mathbb{P}(\theta \mid X)\, d\theta. \tag{4.24}$$

როცა პარამეტრის აპრიორული განაწილება *თანაბარია* (სხვანაირად, *მართკუთხაა* — ინგლ. *uniform distribution* ან *rectangular distribution*), მაშინ ბაიესის შეფასება მაქსიმალური დამაჯერებლობის შეფასების ტოლფასია.

მაგალითი

კვლავ ავიღოთ საშვილოსნოს ყელის კიბოს სკრინინგის ჩვენი წინა მაგალითი და ახლა დავუშვათ, რომ p_0 და p_1 პარამეტრებიდან თითოეული ემორჩილება *ბეტა-განაწილებას* (α_0, β_0) და (α_1, β_1) პარამეტრებით შესაბამისად. ბეტა-განაწილების ალბათობის $0 \le u \le 1$ სიმრავლეზე განსაზღვრული სიმკვრივე $\alpha, \beta > 0$ პარამეტრებით შემდეგნაირად ჩაიწერება :

$$f_{\alpha,\beta}(u) = \frac{u^{\alpha-1}(1-u)^{\beta-1}}{B(\alpha,\beta)}, \tag{4.25}$$

სადაც $B(\alpha,\beta) = \dfrac{\Gamma(\alpha)\Gamma(\beta)}{B(\alpha,\beta)}$ და Γ არის გამა ფუნქცია. ამ კანონისთვის მათემატიკური ლოდინი $\dfrac{\alpha}{\alpha+\beta}$ სიდიდის ტოლია.

p_0 პარამეტრის ბაიესური შეფასების გამოსაანგარიშებლად საჭიროა $\mathbb{P}(p_0 \mid \mathcal{D}_0)$ კანონის ცოდნა.

ბაიესის კანონი მისი ჩაწერის საშუალებას გვაძლევს ქვემოთ მოცემული ფორმით :

$$\left.\begin{aligned}\mathbb{P}(p_0 \mid \mathcal{D}_0) &= \frac{\mathbb{P}(\mathcal{D}_0 \mid p_0)\cdot\mathbb{P}(p_0)}{\mathbb{P}(\mathcal{D}_0)} \\ &= \frac{1}{\mathbb{P}(\mathcal{D}_0)\cdot B(\alpha_0,\beta_0)}\prod_{i=1}^{n_0} p_0^{x^i}(1-p_0)^{1-x^i}\; p_0^{\alpha_0-1}(1-p_0)^{\beta_0-1} \quad \text{(ჰიპოთეზა iid)} \\ &= \frac{1}{\mathbb{P}(\mathcal{D}_0)\cdot B(\alpha_0,\beta_0)}\, p_0^{n_0-t_0+\alpha_0-1}(1-p_0)^{t_0+\beta_0-1}\end{aligned}\right\}.$$

ამრიგად, $p_0 \mid \mathcal{D}_0$ ემორჩილება ბეტა-განაწილებას $(n_0 - t_0 + \alpha_0)$ და $(t_0 + \beta_0)$ პარამეტრებით. ასე რომ, p_0 პარამეტრის ბაიესური შეფასებაა

$$\tilde{p}_0 = \mathbb{E}[p_0 \mid \mathcal{D}_0] = \frac{(n_0 - t_0 + \alpha_0)}{(n_0 - t_0 + \alpha_0) + (t_0 + \beta_0)} = \frac{n_0 - t_0 + \alpha_0}{n_0 + \alpha_0 + \beta_0}. \tag{4.26}$$

4.21 განტოლების გამოყენებით 4.26 განტოლების ხელახლა ჩაწერა შემდეგი სახით შეიძლება :

$$\tilde{p}_0 = \frac{n_0}{n_0 + \alpha_0 + \beta_0}\,\hat{p}_0 + \frac{\alpha_0 + \beta_0}{n_0 + \alpha_0 + \beta_0}\cdot\frac{\alpha_0}{\alpha_0 + \beta_0}.$$

როცა დაკვირვებათა n_0 რიცხვი დიდია, $\tilde{p}_0$ ბაიესის შეფასება ახლოსაა $\hat{p}_0$ მაქსიმალური დამაჯერებლობის შეფასებასთან და, პირიქით, თუ n_0 რიცხვი მცირეა, $\tilde{p}_0$ ბაიესის შეფასება ახლოსაა $\frac{\alpha_0}{\alpha_0 + \beta_0}$ სიდიდესთან, რომელიც არის p_0 პარამეტრის აპრიორული განაწილების მათემატიკური ლოდინი. ამრიგად, რაც უფრო მეტი მონაცემი გვაქვს, მით უფრო მეტად ვენდობით მათ და მით უფრო დაშორებულია შეფასება პარამეტრის აპრიორული მათემატიკური ლოდინისგან, რომელიც ახლოს იქნება დაკვირვებათა მცირე რაოდენობისას.

მსგავსი მსჯელობა გამოიყენება p_1 პარამეტრის მიმართაც, რომლის ბაიესის შეფასებას შემდეგი სახე აქვს :

$$\tilde{p}_1 = \frac{t_1 + \alpha_1}{n_1 + \alpha_1 + \beta_1} = \frac{n_1}{n_1 + \alpha_1 + \beta_1}\,\hat{p}_1 + \frac{\alpha_1 + \beta_1}{n_1 + \alpha_1 + \beta_1}\cdot\frac{\alpha_1}{\alpha_1 + \beta_1}. \tag{4.27}$$

3.3 კვადრატული შეცდომის წანაცვლება-დისპერსია წყვილის დეკომპოზიცია

რომელიღაც θ პარამეტრისთვის მისი $\hat{\theta}$ შეფასების საშუალო კვადრატული შეცდომა (ინგლისურად : Mean Square Error, MSE) არის

$$\left.\begin{aligned}\mathrm{MSE}(\hat{\theta}) &= \mathbb{E}[(\hat{\theta}-\theta)^2] \\ &= \mathbb{E}[(\hat{\theta}-\mathbb{E}[\hat{\theta}]+\mathbb{E}[\hat{\theta}]-\theta)^2] \\ &= \mathbb{E}[(\hat{\theta}-\mathbb{E}[\hat{\theta}])^2]+\mathbb{E}[(\mathbb{E}[\hat{\theta}]-\theta)^2]+\mathbb{E}[2(\hat{\theta}-\mathbb{E}[\hat{\theta}])(\mathbb{E}[\hat{\theta}]-\theta)] \\ &= \mathrm{Var}(\hat{\theta})+(\mathbb{E}[\hat{\theta}]-\theta)^2\end{aligned}\right\}.$$

ეს უკანასკნელი ტოლობა მიიღება, თუ გავითვალისწინებთ, რომ $\mathbb{E}[\hat{\theta}]-\theta$ არის დეტერმინისტული სიდიდე და, ამასთან ერთად, $\mathbb{E}[(\mathbb{E}[\hat{\theta}]-\theta)] = \mathbb{E}[\mathbb{E}[\hat{\theta}]]-\mathbb{E}[\theta] = 0$. ამრიგად, შეფასების საშუალო კვადრატული შეცდომა არის ამ შეფასების დისპერსიისა და წანაცვლების კვადრატის ჯამი. სწორედ ამიტომ წანაცვლებულ შეფასებას შეიძლება — თუ მისი დისპერსია ძალიან მცირეა — გააჩნდეს წაუნაცვლებელ შეფასებაზე ნაკლები საშუალო კვადრატული შეცდომა. აქ ჩვენ ვაწყდებით მეორე ლექციაში (პუნქტი 5.3) შემოღებულ ცნებას კომპრომისის შესახებ წანაცვლებასა და დისპერსიას შორის.

4 გულუბრყვილო ბაიესური კლასიფიკაცია

ახლა გამოვიყენოთ განხილული პრინციპები მანქანური სწავლების ჩვენი პირველი ალგორითმის ჩამოსაყალიბებლად.

4.1 პრინციპები

ჩვენ მიერ აქამდის განხილულ მაგალითებში გამოყენებული შემთხვევითი სიდიდეები იყო ერთგანზომილებიანი. მაგრამ გამოყენებათა უმრავლესობაში მონაცემები არის მრავალგანზომილებიანი. ამან შეიძლება მნიშვნელოვნად გაართულოს გამოთვლები, ცხადია, იმ შემთხვევის გამორიცხვით, როცა კეთდება *გულუბრყვილო დაშვება ცვლადების პირობითად დამოუკიდებლობის შესახებ.*

დავუშვათ, რომ გვაქვს p განზომილების n დაკვირვება : $\left\{\vec{x}^1, \vec{x}^2, \ldots, \vec{x}^n\right\}$ ისეთი, რომ $\vec{x}^i \in \mathbb{R}^p$, და $y^1, y^2, \ldots, y^n$. ჩვენ განვახორციელებთ p ცვლადიდან თითოეულის მოდელირებას X_j შემთხვევითი ცვლადის სახით. ჰიპოთეზა პირობითი დამოუკიდებლობის შესახებ ნიშნავს, რომ

$$\mathbb{P}\left(X_j = x_j \mid Y = y,\, X_m = x_m\right) = \mathbb{P}\left(X_j = x_j\, Y = y\right) \qquad \forall 1 \le j \ne m \le p\,. \tag{4.28}$$

ეს დაშვება დამაჯერებლობის შემდეგნაირად ჩაწერის საშუალებას გვაძლევს :

$$\mathbb{P}\left(Y = y \mid X_1 = x_1, X_2 = x_2, \ldots, X_p = x_p\right) = \frac{\prod_{j=1}^{p} \mathbb{P}\left(X_j = x_j \mid Y = y\right) \cdot \mathbb{P}\left(Y = y\right)}{\mathbb{P}\left(X_1 = x_1, X_2 = x_2, \ldots, X_p = x_p\right)}. \tag{4.29}$$

განსაზღვრება 4.8 (ბაიესური გულუბრყვილო კლასიფიკატორი) *გულუბრყვილო ბაიესური კლასიფიკატორი* (ინგლისურად : *naive Bayes classifier*) — ეს ისეთი კლასიფიკატორია, რომელიც აგებულია გადაწყვეტილების მიღების წესის გამოყენებით აპოსტერიორული ალბათობის მაქსიმუმიდან გამომდინარე მრავალგანზომილებიან დაკვირვებებში, სადაც

იგულისხმება, რომ ცვლადები პირობითად დამოუკიდებელია ჭდის მიხედვით.

გადაწყვეტილების მიღების წესი ამ შემთხვევაში შემდეგ სახისაა :

$$\hat{y} = \arg\max_{c=1,\ldots,C} \mathbb{P}\left(Y = c\right) \prod_{j=1}^{p} \mathbb{P}\left(X_j = x_j \mid Y = c\right). \tag{4.30}$$

4.2 სპამის ბაიესური ფილტრაცია

გულუბრყვილო ბაიესური კლასიფიკაციის კლასიკური მაგალითია ელექტრონული ფოსტის ფილტრაცია სპამის გამოსაყოფად ლეგიტიმური წერილებიდან. ჩვენ დავუშვებთ, რომ გვაქვს n ელექტრონული წერილი და მათი ორობითი ჭდეები : $y^i = 1$, თუ საქმე ეხება სპამს, და $y^i = 0$ — წინააღმდეგ შემთხვევაში.

ელექტრონული წერილის წარმოდგენა p - ცვლადით

ჩვენი უპირველესი ამოცანაა $p-$განზომილებიანი წარმოდგენის პოვნა ელექტრონული წერილებისთვის. ამისათვის ჩვენ ვირჩევთ p საკვანძო სიტყვას, ისეთს, როგორიცაა, მაგალითად, «მდიდარი», «უცოლო», «მოგება», «წკაპუნი» და სხვა. მათი არსებობა ინფორმაციის მატარებელი იქნება იმ ალბათობაზე, რომლითაც შეტყობინება შეიძლება იყოს სპამი. პრაქტიკულად, დასაშვებია ყველა იმ სიტყვათა სიმრავლის გამოყენება, რომლებიც ჩნდება მონაცემთა ჩვენი საწვრთნელი ბაზის ელექტრონულ წერილებში, ალბათ, მხოლოდ მომსახურე («*ფორმალური*») სიტყვების (კავშირების, ნაწილაკების, დამხმარე ზმნების და ა.შ.) გარდა.

მაშინ ელექტრონული წერილი წარმოდგენილი აღმოჩნდება p რაოდენობის ორობითი ცვლადით, რომელთა შორის თითოეული შეესაბამება ჩვენი სიის ერთ-ერთ სიტყვას და ფასდება $1-$ით, თუ ეს სიტყვა გვხვდება წერილში, და $0-$ით — წინააღმდეგ შემთხვევაში.

ამიტომ ჩვენ ყოველი ამ ცვლადის მოდელირებას განვახორციელებთ როგორც X_j ორობითი შემთხვევითი სიდიდის რეალიზაციას, რომელიც ემორჩილება *ალბათობათა განაწილების ბერნულის კანონს* p_j პარამეტრით. ჩვენ ასევე მივმართავთ ჭდის მოდელირებას როგორც Y ორობითი შემთხვევითი ცვლადის რეალიზაციას.

გულუბრყვილო ბაიესური კლასიფიკატორის გამოსაყენებლად ჩვენ მიერ მიჩნეული იქნება, რომ ცვლადები პირობითად არ არის დამოკიდებული ჭდეზე. ეს დაშვება მართლაც რომ გულუბრყვილოა : სპამს შორის ზოგიერთი სიტყვა ჩნდება მეტი ალბათობით, თუ ცნობილია, რომ მასში არსებობს სხვა სიტყვა. ვთქვათ, «უცოლო» ჩნდება მეტი ალბათობით, თუ ცნობილია, რომ შეტყობინებაში არის სიტყვა «პაემანი» ან «შეხვედრა». თუმცა, პრაქტიკულად, ეს ხელს არ უშლის კლასიფიკატორს საკმარისად ეფექტურად შეასრულოს თავისი ფუნქცია.

აპოსტერიორული (ალბათობის) მაქსიმუმი

გადაწყვეტილების მიღების წესს აპოსტერიორული ალბათობის მაქსიმუმით (იხ. განტოლება

4.30) შემდეგი სახე აქვს :

$$\hat{y} = \begin{cases} 1 & \text{თუ } \mathbb{P}(Y=1)\prod_{j=1}^{p}\mathbb{P}\left(X_j = x_j \mid Y=1\right) > \mathbb{P}(Y=0)\prod_{j=1}^{p}\mathbb{P}\left(X_j = x_j \mid Y=0\right) \\ 0 & \text{წინააღმდეგ შემთხვევაში} \end{cases}. \qquad (4.31)$$

დასკვნა

ამრიგად, ისღა დაგვრჩენია, რომ განვსაზღვროთ $\mathbb{P}(Y=1), \mathbb{P}(Y=0)$ უპირობო ალბათობები და ყოველი X_j ცვლადისთვის ვიპოვოთ $\mathbb{P}(X_j = x_j \mid Y=0)$ და $\mathbb{P}(X_j = x_j \mid Y=1)$ პირობითი ალბათობები.

$\mathbb{P}(Y=1)$ უბრალოდ არის სპამის სიხშირე ჩვენს საწვრთნელ მონაცემთა ნაკრებზე და $\mathbb{P}(Y=0) = 1 - \mathbb{P}(Y=1)$ — ლეგიტიმური ელექტრონული წერილების სიხშირე. ცხადია, შეგვიძლია ასევე ამ საკითხთან დაკავშირებული სტატისტიკური მასალის გამოყენებაც, რომელიც ასახელებს ელექტრონულ წერილებს შორის არსებული სპამისთვის, დაახლოებით, 80 პროცენტიან წილს.

ვინაიდან ავირჩიეთ მოდელი X_j შემთხვევითი ცვლადის სახით, რომელიც ალბათობათა განაწილების ბერნულის კანონს ემორჩილება, ამიტომ $\mathbb{P}\left(X_j = x_j \mid Y=1\right) = \left(p_j\right)^{x_j} \cdot \left(1-p_j\right)^{1-x_j}$. აქ p_j პარამეტრის შესაფასებლად შეიძლება ვისარგებლოთ მაქსიმალური დამაჯერებლობის შეფასებით (იხ. განტოლება 4.20) : $\hat{p}_j = \frac{S_j}{S}$ სადაც S_j — იმ სპამ-შეტყობინებათა რაოდენობაა, რომლებიც შეიცავს $j-$ურ სიტყვას, ხოლო S — სპამ-შეტყობინებათა რაოდენობაა მონაცემთა ჩვენს ნაკრებში.

მაგრამ ეს მაჩვენებელი ნაკლებად გამოსადეგია იშვიათი სიტყვების შემთხვევაში : თუ ჩვენი $j-$ური სიტყვა საერთოდ არ გვხვდება საწვრთნელ ნაკრებში, მაშინ $S_j = S = 0$. ამ მიზეზით უპირატესობა ეძლევა ე.წ. *ლაპლასის ადიტიური დაგლუვების* (ინგლისურად : *Laplace additive smoothing*) გამოყენებას შემდეგი შეფასების მისაღებად :

$$\hat{p}_j = \frac{S_j + 1}{S + 2}. \qquad (4.32)$$

სიტყვისთვის, რომელიც არ გვხვდება საწვრთნელ ნაკრებში, $\hat{p}_j = 0.5$.

5 ბაიესური მოდელის არჩევა

ბაიესური სქემა ახლებურად აშუქებს მოდელის არჩევის ცნებას (იხ. მე-3 ლექცია). მართლაც, ბაიესის წესი შეიძლება იყოს გამოყენებული $\mathcal{M}$ მოდელის და $\mathcal{D}$ მონაცემთა ნაკრების ერთობლივი ალბათობის მიმართ :

$$\mathbb{P}\left(\mathcal{M} \mid \mathcal{D}\right) = \frac{\mathbb{P}\left(\mathcal{D} \mid \mathcal{M}\right) \cdot \mathbb{P}\left(\mathcal{M}\right)}{\mathbb{P}\left(\mathcal{D}\right)}. \qquad (4.33)$$

გალოგარითმებით შესაძლებელი ხდება ეკვივალენტური სახითაც ჩაწერა :

$$\log \mathbb{P}(\mathcal{M} \mid \mathcal{D}) = \log \mathbb{P}(\mathcal{D} \mid \mathcal{M}) + \log \mathbb{P}(\mathcal{M}) - \log \mathbb{P}(\mathcal{D}) . \qquad \textbf{(4.34)}$$

4.34 განტოლების აპოსტერიორული მაქსიმიზაცია ტოლფასია ორი წევრის ჯამის მინიმიზაციის : ერთია ემპირიული შეცდომის დამახასიათებელი $\left(-\log \mathbb{P}(\mathcal{D} \mid \mathcal{M})\right)$ გამოსახულება, ხოლო მეორე წევრი — $\left(-\log \mathbb{P}(\mathcal{M})\right)$ — მოდელის სირთულეს ახასიათებს. ეს ფორმულირება *რეგულარიზაციის* (იხ. მე-2 ლექციის პუნქტი 5.4) ფორმულირების მსგავსია. ასე რომ, *რეგულარიზაცია* შეიძლება იყოს გაგებული როგორც ისეთი აპრიორული განაწილების არჩევა $\mathcal{M}-$თვის, რომელიც ნაკლებად რთული მოდელების ხელშემწყობია.

სწორედ ამ კონტექსტში თავსდება, მაგალითად, *ბაიესური საინფორმაციო კრიტერიუმი* (ინგლისურად : *BIC, bayesian information criterion*). ბაიესის საინფორმაციო კრიტერიუმი — ეს სტატისტიკური მოდელის არჩევის კრიტერიუმია გარკვეული სასრული სიმრავლიდან. უპირატესობა ენიჭება მოდელს კრიტერიუმის მინიმალური მნიშვნელობით. კრიტერიუმი ეფუძნება დამაჯერებლობის ფუნქციის გამოყენებას და მჭიდროდ არის დაკავშირებული იაპონელი სტატისტიკოსის ჰიროცუგუ აკაიკეს (Hirotsugu Akaike, 1927-2009) საინფორმაციო კრიტერიუმთან. ხსენებული კრიტერიუმის დაწვრილებითი აღწერა ლექციათა ამ კურსის ფარგლებს სცდება.

6 საკვანძო მომენტები

— სწავლების პრობლემის მოდელირება ალბათური მიდგომით ორ ამოცანად იყოფა :

— დანასკვის ამოცანა, რომელიც მდგომარეობს $\mathbb{P}(Y = c \mid X)$ ალბათობის კანონის დადგენაში ;

— გადაწყვეტილების მიღების წესის გამოყენების შემცველი პროგნოზირების ამოცანა.

— ბაიესის კანონზე დაფუძნებული ალბათური მსჯელობები.

— ჩვეულებრივ, $\mathbb{P}(Y = c \mid X)$ ალბათობის კანონის მოდელირება პარამეტრული მიდგომით ხდება და დანასკვის ამოცანის გადაჭრა, ერთი მხრივ, ბაიესის კანონზე დაყრდნობით ხორციელდება, ხოლო, მეორე მხრივ, ისეთი შეფასების გამოყენებით, როგორიცაა მაქსიმალური დამაჯერებლობის შეფასება ან ბაიესური შეფასება პარამეტრების დასადგენად.

— ბაიესური გადაწყვეტილების წესი, რომელიც მდგომარეობს დანაკარგის *მათემატიკური ლოდინის* მინიმიზაციაში, უპირისპირდება *ემპირიული* რისკის მინიმიზაციის (ინგლისურად : *Empirical risk minimization*, ERM) წესს, ქმნის კონტრასტს ამ წესთან.

დამატებითი ინფორმაცია

• ამ ლექციაში ჩვენ მხოლოდ კლასიფიკაციის პრობლემებს შევეხეთ. მაგრამ ბაიესის რისკის მინიმიზაციის სტრატეგია რეგრესიის შემთხვევაშიც ეფექტურია, თუ შესაბამისი დანაკარგების ფუნქცია იქნება შემოტანილი. ასეთ ვითარებაში ყველა ჯამი $y-$ით ჩანაცვლდება ინტეგრალებით.

- როსის ნაშრომი (Ross, 1987) უფრო დაწვრილებით განიხილავს შეფასებას დამაჯერებლობის მაქსიმუმით და ბაიესურ შეფასებას.

- კურსის მოცემულ ლექციაში არ იგეგმება მანქანური სწავლებისადმი ბაიესური მიდგომის უფრო დეტალური განხილვა. ამ პრობლემით დაინტერესებულ პირთ შეუძლიათ მიმართონ მერფის (Murphy, 2013) ან ბარბერის (Barber, 2012) ფუნდამენტურ წიგნებს.

- დევიდ ჰანდისა და კემინ იუს (Hand და Yu, 2001) სტატია დაწვრილებით აანალიზებს გულუბრყვილო ბაიესურ კლასიფიკაციას.

7 ბიბლიოგრაფია

1. Barber, D. (2012). *Bayesian Reasoning and Machine Learning*. Cambridge University Press.

2. Hand, D. J. and Yu, K. (2001). Idiot's Bayes — Not So Stupid After All? *International Statistical Review ;* 69(3), :385–398.

3. Murphy, K. P. (2013). *Machine Learning: A Probabilistic Perspective*. MIT Press, Cambridge, MA, 4th edition.

4. Ross, S. M. (1987). *Introduction to Probability and Statistics for Engineers and Scientists*. Wiley, New York.

8 სავარჯიშოები

4.1 თუ მიმართავენ აპრიორულ დაშვებებს (ჰიპოთეზებს) დამაჯერებლობის ფარდობის ტესტის გამოყენებისას ?

4.2 რა შეიძლება ითქვას აპოსტერიორული ალბათობის შესახებ, როცა დამაჯერებლობა იზრდება ?

4.3 წიგნების კოლექციის შედგენა – ალინას გატაცებაა. მან შექმნა ელექტრონული ფოსტის სპეციალური მისამართი და სურს გულუბრყვილო ბაიესური კლასიფიკატორის შექმნაც ისეთი შეტყობინებების ავტომატური იდენტიფიკაციისათვის, რომლებიც მას დააინტერესებს, მაგრამ ინფორმაცია არ უნდა ეხებოდეს ხელახალ გამოცემას ან ორიგინალური გამოცემის ასლს

საწვრთნელი სიმრავლე შემდეგი სახისაა :

კლასისათვის «საინტერესო შეტყობინება» :

- «შესანიშნავი ორიგინალური გამოცემა *სტატისტიკური სწავლების ელემენტები*. იდეალურ მდგომარეობაში.»

- «ორიგინალური გამოცემა *მანქანური სწავლება* : *ალბათური პერსპექტივა*.»

კლასისათვის «არასაინტერესო შეტყობინება» :

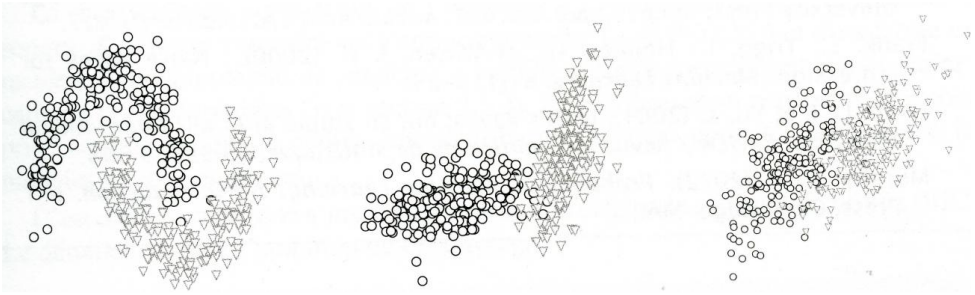
- «ახალი გამოცემა *მანქანური სწავლება* : *ალბათური პერსპექტივა*.»

- «ძალიან კარგი ასლი ორიგინალური გამოცემის *სტატისტიკური სწავლების ელემენტები.*»

1. შექმენით ალინასთვის საინტერესო შეტყობინება, რომელიც კლასიფიცირებული იქნება მისი კლასიფიკატორით როგორც არასაინტერესო.

2. და, პირიქით, შექმენით ალინასთვის არასაინტერესო შეტყობინება, რომელიც კლასიფიცირებული იქნება მისი კლასიფიკატორით როგორც საინტერესო.

4.4 ბესო განიხილავს კვადრატული დისკრიმინანტული ანალიზის (ინგლ. QDA - Quadratic Discriminant Analysis) ან წრფივი დისკრიმინანტული ანალიზის (ინგლ. LDA – Linear Discriminant Analysis) გამოყენების შესაძლებლობას იმისათვის, რომ წრეები გამოცალკევდეს სამკუთხედებისგან თითოეულ ნახატზე ქვემოთ მოცემულ სამ შემთხვევაში. ორი მიდგომიდან რომელს ურჩევდით მას გამოსაყენებლად ცალკეულ ვითარებაში ? რატომ თვლით, რომ ეს ასეა ?



4.5 შარლოტას აერია ერთმანეთში ღვინის ჭიქიანი ყუთები ორი სხვადასხვა ქარხნიდან. ასე რომ მას მოუნდა კლასიფიკატორის აგება, რომელიც საშუალებას მისცემს ხელახლა დაუდგინოს თითოეულ ყუთს მისი წარმომავლობა (წყარო) როგორც მოცემულ ყუთში ღვინის გატეხილ ჭიქათა წილის (პროპორციის) ფუნქცია. შარლოტამ იცის, რომ მიიღო B ქარხნის 5-ჯერ მეტი ჭიქა, ვიდრე A ქარხნის. ამიტომ მას ორჯერ უფრო ძნელი ეჩვენება იმის მტკიცება, რომ ყუთი მიღებულია A ქარხნიდან, მაშინ როცა უფრო სავარაუდოა, რომ იგი მოსულია B ქარხნიდან და არა პირიქით. ბოლოს და ბოლოს, შარლოტამ გადაწყვიტა ღვინის გატეხილი ჭიქების პროპორციის (წილის) მოდელირება თითოეულ ქარხანაში ე.წ. ბეტა კანონით (იხილეთ მე-14 ლექციის ფორმულა 23). ვინაიდან მან იცის, რომ A ქარხანა აწარმოებს ოდნავ უფრო ნაკლებად მყიფე ღვინის ჭიქებს, ვიდრე B ქარხანა, იგი ირჩევს ბეტა კანონს პარამეტრებით $\alpha = 2$ და $\beta = 11$ ქარხნისათვის A და იმავე ბეტა კანონს პარამეტრებით $\alpha = 2$ და $\beta = 10$ ქარხნისათვის B.

ზედმეტი არ იქნება იმის გახსენებაც, რომ $n \in \mathbb{N}^*$ შემთხვევაში სამართლიანია $\Gamma(n) = (n-1)!$ თანაფარდობა. პასუხი გაეცით შემდეგ ოთხ შეკითხვას :

1. დანახარჯების რომელი ფუნქცია უნდა იქნეს არჩეული ?

2. როგორია გადაწყვეტილების მიღების წესი დამაჯერებლობის მაქსიმუმით ?

3. როგორია გადაწყვეტილების მიღების წესი აპოსტერიორული ალბათობის მაქსიმუმით ?

4. როგორია ბაიესის გადაწყვეტილების მიღების წესი ?

4.6 დიმიტრის სურს ასწავლოს უნივერსიტეტის რობოტს იმის დადგენა, არის თუ არა კაბინეტში კალათა ქაღალდის შესაგროვებლად მეორადი გამოყენებისთვის — *რეციკლინგის-თვის*. კაბინეტები შეიძლება გამოეყოს მასწავლებლებს (მასწ), ან მეცნიერ მუშაკებს (მეცნ), ასევე სტუდენტებსაც (სტუდ). დიმიტრის მიერ შემდეგი მონაცემები იქნა მოპოვებული :

პოზიცია	მასწ	მასწ	სტუდ	მასწ	სტუდ	მეცნ	სტუდ	მეცნ
ფაკულტეტი	ინფო	მათ	ინფო	ინფო	მათ	ინფო	მათ	მათ
ზომა	საშუალო	საშუალო	დიდი	მცირე	დიდი	მცირე	საშუალო	მცირე
რეციკლინგი	კი	კი	კი	კი	კი	არა	არა	კი

გულუბრყვილო ბაიესური კლასიფიკატორის თანახმად, თუ არსებობს კალათა ქაღალდის რეციკლინგისთვის მათემატიკის ფაკულტეტის რომელიმე მცირე სტუდენტურ კაბინეტში ? გამოიყენეთ ლაპლასის დაგლუვება (ინგლისურად : *Laplace smoothing*).

სავარჯიშოთა ამონახსნები

4.1 არა

4.2 ისიც ასევე იზრდება

4.3

1. «ძალიან კარგი პირველი გამოცემა *სტატისტიკური სწავლების ელემენტები*» იქნება კლასიფიცირებული როგორც არასაინტერესო იმიტომ, რომ «ძალიან» და «კარგი» ჩნდება არასაინტერესო შეტყობინებებში, მაგრამ არა საინტერესო შეტყობინებათა ნიმუშებში ; «პირველი» არ ჩნდება არც ერთში და არც მეორეში ; ხოლო «გამოცემა», «სტატისტიკური», «სწავლების», «ელემენტები» ერთნაირად ხშირად გვხვდება ორივე კლასში.

2. და, პირიქით, «შესანიშნავი მე-3 გამოცემა საუცხოო მდგომარეობაში *მანქანური სწავლება* : *ალბათური პერსპექტივა*» კლასიფიცირებული იქნება როგორც საინტერესო შეტყობინება, თუმცა ასეთი არ არის.

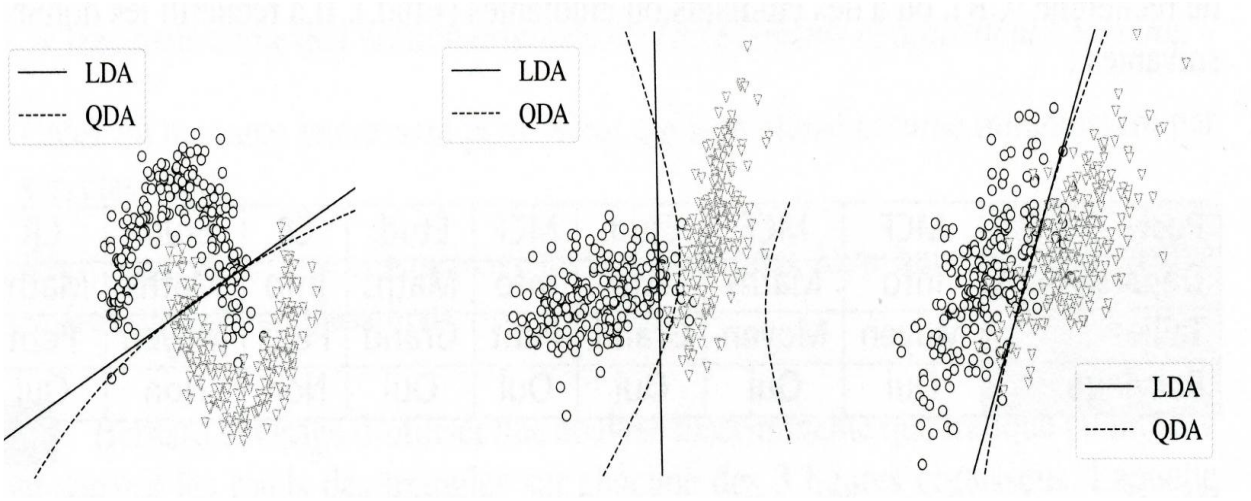
4.4 *მარცხენა ნახატი* : მონაცემები არ არის აგებული გაუსიანით (წერტილთა ღრუბლები ქმნის მთვარეებს და არა ელიფსებს). ამ მეთოდებიდან არცერთი არ გამოდგება და მათ ექნება ხარისხის მსგავსი მახასიათებლები : იმავე წარმატებით არის შესაძლებელი წრფივი დისკრიმინანტული ანალიზის გამოყენება, ვინაიდან სწავლებისთვის იგი (LDA) პარამეტრების ნაკლებ რაოდენობას მოითხოვს.

შუა ნახატი : მონაცემები აღებულია გაუსიანებიდან (წერტილთა ღრუბლები ქმნის ელიფსებს), რომლებსაც სხვადასხვა კოვარიაციული მატრიცა აქვს (ელიფსების ღერძები პარალელური არ არის). კვადრატული დისკრიმინანტული ანალიზი QDA უფრო მორგებულია ამოცანასთან, ვიდრე წრფივი დისკრიმინანტული ანალიზი LDA.

მარჯვენა ნახატი : მონაცემები აღებულია გაუსიანებიდან (წერტილთა ღრუბლები ქმნის ელიფსებს), რომლებსაც ერთნაირი კოვარიაციული მატრიცები აქვს (ორი ელიფსის ღერძები, სავარაუდოდ, პარალელურია). წრფივი დისკრიმინანტული ანალიზი LDA ისევე გამოდგება,

როგორც კვადრატული დისკრიმინანტული ანალიზი QDA. თუმცა LDA წრფივი დისკრიმინანტული ანალიზის გამოყენება მეტი წარმატებითაც შეიძლება, ალბათ, ვინაიდან სწავლებისთვის იგი პარამეტრების ნაკლებ რაოდენობას მოითხოვს.

ქვემოთ მოცემულია გადაწყვეტილებათა მიღების საზღვრები, რომლებიც დადგენილია წრფივი დისკრიმინანტული ანალიზით (ინგლისურად : LDA — *Linear Discriminant Analysis*) და კვადრატული დისკრიმინანტული ანალიზით (ინგლისურად : QDA — *Quadratic Discriminant Analysis*) თითოეული შემთხვევისთვის :



4.5

1. $\lambda_{AA} = \lambda_{BB} = 0; \lambda_{AB} = 2; \lambda_{BA} = 1.$

2. აღვნიშნოთ X – ით შემთხვევითი ნამდვილი ცვლადი, რომლითაც ხდება ყუთში გატეხილი ჭიქების წილის მოდელირება და Y – ით — შემთხვევითი ბინარული ცვლადი, რომლითაც ხდება ყუთის მწარმოებლის (წყაროს, წარმომავლობის) მოდელირება. შარლოტას ვარაუდების (ჰიპოთეზების) მიხედვით :

$$\mathbb{P}\left(X = x \mid Y = A\right) = \frac{\Gamma(13)\, x(1-x)^{10}}{\Gamma(2)+\Gamma(11)} \text{ და } \mathbb{P}\left(X = x \mid Y = B\right) = \frac{\Gamma(12)\, x(1-x)^{9}}{\Gamma(2)+\Gamma(10)}.$$

დისკრიმინანტული ფუნქცია დამაჯერებლობის მაქსიმუმით შემდეგნაირად ჩაიწერება :

$$f\left(X = x\right) = \frac{\mathbb{P}\left(X = x \mid Y = A\right)}{\mathbb{P}\left(X = x \mid Y = B\right)} - 1 = \frac{12}{10}(1-x) - 1.$$

შარლოტას შეფასებით ყუთის მწარმოებელია A ქარხანა, თუ გატეხილი ჭიქების წილი ნაკლებია $\frac{1}{6} \approx 0{,}17$ მნიშვნელობაზე, ხოლო წინააღმდეგ შემთხვევაში — B ქარხანა.

3. დისკრიმინანტული ფუნქცია აპოსტერიორული ალბათობის მაქსიმუმით არის :

$$f\left(X = x\right) = \frac{\mathbb{P}\left(X = x \mid Y = A\right)\mathbb{P}\left(Y = A\right)}{\mathbb{P}\left(X = x \mid Y = B\right)\mathbb{P}\left(Y = B\right)} - 1 = \frac{1}{5} \times \frac{12}{10}(1-x) - 1.$$

შარლოტას შეფასებით ყუთის მწარმოებელია A ქარხანა, თუ გატეხილი ჭიქების წილი ნაკლებია $\frac{19}{25} \approx 0,76$ მნიშვნელობაზე, ხოლო წინააღმდეგ შემთხვევაში — B ქარხანა.

4. ბაიესის დისკრიმინანტული ფუნქციაა $f\left(X=x\right)=2\mathbb{P}\left(Y=\mathrm{A}\,|\,X=x\right)-\mathbb{P}\left(Y=\mathrm{B}\,|\,X=x\right)$ ან, ეკვივალენტური სახით :

$$f\left(X=x\right)=\frac{2\mathbb{P}\left(X=x\,|\,Y=A\right)\mathbb{P}\left(Y=A\right)}{\mathbb{P}\left(X=x\,|\,Y=B\right)\mathbb{P}\left(Y=B\right)}-1=\frac{2}{5}\times\frac{12}{10}\left(1-x\right)-1.$$

შარლოტას შეფასებით ყუთის მწარმოებელია A ქარხანა, თუ გატეხილი ჭიქების წილი ნაკლებია $\frac{13}{25}=0.52$ მნიშვნელობაზე, ხოლო წინააღმდეგ შემთხვევაში — B ქარხანა.

დანახარჯების ფუნქცია აიძულებს მას უფრო ზომიერად გამოხატოს უპირატესობის მიცემა A კლასისთვის.

4.6 n –ით აღვნიშნოთ კაბინეტების რიცხვი, ხოლო n_1 – ით — იმ კაბინეტების რიცხვი, სადაც რეციკლინგის კალათა არსებობს. ამ უკანასკნელთა შორის n_{11} – ით აღვნიშნოთ მცირე ზომის კაბინეტების რიცხვი, n_{12} – ით — მათემატიკის დეპარტამენტის კაბინეტების რიცხვი, ხოლო n_{13} სიმბოლოთი — კაბინეტების რიცხვი სტუდენტებისთვის. ამავე მოდელში $n_{13}, n_{01}, n_{02}, n_{03}$ ნოტაციები შემოვიღოთ *«რეციკლინგის კალათა არსებობს»* კლასის ჩანაცვლებისას ალტერნატიული კლასით : «*რეციკლინგის კალათა არ არსებობს*».

იმისათვის, რომ გამოვიყენოთ წესი 4.17, $\frac{n_{11}+1}{n_1+1}\frac{n_{12}+1}{n_1+1}\frac{n_{13}+1}{n_1+1}\frac{n_1}{n}$ სიდიდე უნდა იქნეს შედარებული $\frac{n_{01}+1}{n_0+1}\frac{n_{02}+1}{n_0+1}\frac{n_{03}+1}{n_0+1}\frac{n_0}{n}$ მნიშვნელობასთან.

პირველი სიდიდე აისახება შემდეგი რიცხვით :

$$\frac{2+1}{6+2}\times\frac{3+1}{6+2}\times\frac{2+1}{6+2}\times\frac{6}{8}=\frac{27}{512}.$$

მეორე სიდიდე კი უფრო ნაკლებია :

$$\frac{1+1}{2+2}\times\frac{1+1}{2+2}\times\frac{1+1}{2+2}\times\frac{2}{8}=\frac{16}{512}.$$

ასე რომ, დიმიტრის რობოტის წინასწარმეტყველებით, კალათა ქაღალდის შესაგროვებლად მისი შემდგომი რეციკლინგისთვის არსებობს.

ლექცია 5 პარამეტრული რეგრესია

შინაარსი

რეგრესიის პარამეტრული მოდელი გულისხმობს, რომ გადაწყვეტილების მიღების ფუნქციის ანალიზური ფორმა ცნობილია. ამასთან დაკავშირებით მოცემულ ლექციაში ძირითადი ყურადღება ეთმობა *წრფივი რეგრესიის* ამოცანებს, ე.ი. იმ ამოცანებს, რომლებისთვისაც გადაწყვეტილების მიღების ფუნქცია არის დესკრიპტორების (მოვლენის აღმწერი დამოუკიდებელი ცვლადების) წრფივი ფუნქცია.

სტატისტიკაში წრფივ მოდელებს გრძელი ისტორია აქვს, რომელიც იწყება პერიოდით კომპიუტერების გაჩენამდე ; მაგრამ არა მხოლოდ ამ მიზეზით შეისწავლება ისინი.

მართლაც, მიუხედავად სიმარტივისა, მათ შეუძლია კარგი შედეგების ჩვენება, ზოგჯერ უკეთესად, ვიდრე უფრო პოპულარულ არაწრფივ მოდელებს (განსაკუთრებით მაშინ, როცა საწვრთნელი კრებული დიდი არ არის).

გარდა ამისა, ეს მოდელები ადვილად ინტერპრეტირებადია.

დაბოლოს, მათი გაგება კარგი საფუძველია არაწრფივი მოდელების ასაგებად.

ამ ლექციაში დაწვრილებით არის წარმოდგენილი წრფივი რეგრესია და მისი გამოყენება კლასიფიკაციის ამოცანებში ლოგისტიკური რეგრესიის საშუალებით.

ლექციის დასკვნითი ნაწილი კი ძალიან მოკლედ ეხება პოლინომურ რეგრესიასაც.

მიზნები

- სწავლების პარამეტრულ და არაპარამეტრულ მოდელებს შორის განსხვავების გაგება.
- წრფივი ან ლოგისტიკური რეგრესიის ამოცანის ფორმულირება.
- პარამეტრული რეგრესიის ამოცანის ამოხსნა.

1 პარამეტრული მოდელის კონტროლირებადი სწავლება

1.1 პარამეტრული მოდელები

პარამეტრულ მოდელზე მაშინ ლაპარაკობენ, როცა გამოიყენება სწავლების ალგორითმი, რომლის მიზანია დესკრიპტორების ფუნქციის (იხ. ქვედანაყოფი 4.1.3) ანალიზური ფორმით განსაზღვრული მოდელის პარამეტრთა ოპტიმალური მნიშვნელობების დადგენა.

პარამეტრული მოდელის სირთულე იზრდება გასაწვრთნელ პარამეტრთა რიცხვის გაზრდისას, სხვანაირად რომ ვთქვათ, ცვლადების რაოდენობის ზრდასთან ერთად.

და პირიქით, არაპარამეტრული მოდელის სირთულეს ზრდის ტენდენცია ექნება დაკვირვებათა რიცხვის გადიდებისას.

მაგალითი

სწავლების ალგორითმს, რომელიც α, β, γ კოეფიციენტების გაწვრთნის საშუალებას იძლევა გადაწყვეტილების მიღების $f : x \mapsto \alpha x_1 + \beta x_2 x_4^2 + \gamma e^{x_3 - x_5}$ ფუნქციაში, საქმე აქვს პარამეტრულ მოდელთან.

დაკვირვებათა რიცხვის განურჩევლად, ეს მოდელი უცვლელი რჩება.

და პირიქით, უახლოესი მეზობლის მეთოდი, რომელიც აკავშირებს $\vec{x}$-თან საწვრთნელი სიმრავლის იმ წერტილის ჭდეს, რომელთანაც იგი უახლოესია ევკლიდეს მანძილით, *არაპარამეტრული მოდელის* სწავლებას ახორციელებს : გადაწყვეტილების მიღების ფუნქციის ჩაწერა *პრედიქტორული*, ანუ *მაპროგნოზირებელი* ცვლადების (ინგლ. *predictor variables*), ფუნქციად შეუძლებელია.

რაც მეტია დაკვირვებები, გადაწყვეტილების მიღების მით უფრო რთული საზღვრის დასწავლას შეძლებს მოდელი (იხ. ლექცია 8).

უნდა გვესმოდეს, რომ პრედიქტორული ცვლადები — ეს ისეთი ცვლადებია, რომლებიც სხვა ცვლადის ან შედეგის პროგნოზირებისათვის გამოიყენება.

დავუშვათ, რომ მოცემულია p განზომილების n დაკვირვებისა და მისი რეალური ჭდის სიმრავლე :

$$\mathcal{D} = \left\{ \vec{x}^i, y^i \right\}_{i=1,\ldots,n}.$$

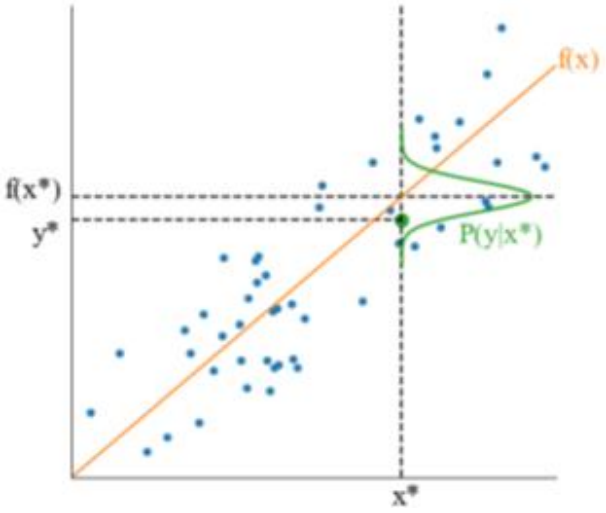
აქვე ასევე დავუშვათ, რომ გადაწყვეტილების მიღების f ფუნქცია პარამეტრიზებულია $\vec{\beta}$ ვექტორით, სადაც :

$$\vec{\beta} \in \mathbb{R}^m$$

ვისარგებლოთ ჰიპოთეზით, რომ შეცდომები, ე.ი. სხვაობები რეალურ ჭდეებსა და f ფუნქციის შესაბამის მნიშვნელობებს შორის, განაწილებულია ნორმალურად, ცენტრით 0 წერტილზე :

$$\left.\begin{aligned} y &= f(\vec{x} \mid \vec{\beta}) + \mathcal{E} \\ \mathcal{E} &\sim \mathcal{N}(0, \sigma^2) \end{aligned}\right\}. \tag{5.1}$$

ეს დაშვება ილუსტრირებულია ნახატზე 5.1 გადაწყვეტილების მიღების წრფივი ფუნქციის შემთხვევისათვის.



ნახატი 5.1 - მოცემული დაკვირვებისათვის (აქ ერთ განზომილებაში) შესაბამისი y^* ჭდის შესაძლო მნიშვნელობათა განაწილება არის ნორმალური (გაუსის) ცენტრით $f\left(x^*\right)$ წერტილზე.

ამ დაშვების შესაბამისად, $\vec{x}$ დაკვირვებები p ნამდვილი შემთხვევითი $X_1, X_2, \ldots, X_p$ ცვლადის რეალიზაციები არის, მათი y ჭდეები კი ნამდვილი შემთხვევითი Y ცვლადის რეალიზაციებს წარმოადგენს და ეს შემთხვევითი ცვლადები შემდეგ პირობას აკმაყოფილებს :

$$\mathbb{P}\left(Y = y \mid X = \vec{x}\right) \sim \mathcal{N}\left(f\left(\vec{x} \mid \vec{\beta}\right), \sigma^2\right). \tag{5.2}$$

აქ $\mathbb{P}\left(X_1 = x_1, X_2 = x_2, \ldots, X_p = x_p\right)$ ალბათობისათვის მიღებულია $\mathbb{P}\left(X = \vec{x}\right)$ აღნიშვნა.

1.2 მაქსიმალური დამაჯერებლობის შეფასება და უმცირეს კვადრატთა მეთოდი

5.2 დაშვების შესაბამისად, n დამოუკიდებელი და ერთნაირად განაწილებული დაკვირვების პირობებში, $\vec{\beta}$ პარამეტრის დამაჯერებლობის ლოგარითმი შემდეგი სახით ჩაიწერება :

$$\left.\begin{aligned} \log \mathbb{P}\left(\mathcal{D} \mid \vec{\beta}\right) &= \log \prod_{i=1}^{n} \mathbb{P}\left(X = \vec{x}^i \mid \vec{\beta}\right) \\ &= \log \prod_{i=1}^{n} \mathbb{P}\left(y^i \mid \vec{x}^i\right) + \log \prod_{i=1}^{n} \mathbb{P}\left(X = \vec{x}^i\right) \\ &= -\log \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y^i - f(\vec{x}^i \mid \vec{\beta})\right)^2 + C \end{aligned}\right\}.$$

უკანასკნელ განტოლებაში $\mathcal{C}$ არის მუდმივა $\vec{\beta}$ პარამეტრთან მიმართებაში და განპირობებულია, ერთი მხრივ, ნორმალური განაწილების $\frac{1}{\sqrt{2\pi}}$ კოეფიციენტით და, მეორე მხრივ, — $\mathbb{P}\left(X = \vec{x}^i\right)$ ალბათობით.

ამრიგად, დამაჯერებლობის მაქსიმიზაცია $\sum_{i=1}^{n}\left(y^i - f(\vec{x}^i \mid \vec{\beta})\right)^2$ სხვაობის კვადრატთა ჯამის მინიმიზაციის ეკვივალენტურია : ეს ცნობილია როგორც *მინიმიზაცია უმცირეს კვადრატთა მეთოდით*, რომელიც ჯერ კიდევ გერმანელი გაუსისა (Johann Carl Friedrich Gauss, 1777-1855) და ფრანგი ლეჟანდრის (Adrien-Marie Legendre, 1752-1833) დროიდან გამოიყენება მათემატიკაში. აუცილებლად უნდა აღინიშნოს ასევე, რომ ეს ემპირიული რისკის მინიმიზაციის ტოლფასია, როცა იგი 2.4.3 დანახარჯების კვადრატული ფუნქციის გამოყენებით ხდება.

2 წრფივი რეგრესია

დავიწყოთ *წრფივი მოდელების* განხილვით : ჩვენ გვინდა y მიზნობრივი ცვლადისათვის ნათელის მოფენა დესკრიპტორების *წრფივი კომბინაციით* — სხვა სიტყვებით რომ ვთქვათ, მათი შეწონილი ჯამით.

2.1 ფორმულირება

გადაწყვეტილების მიღების $f_{\vec{\beta}}$ ფუნქცია ავირჩიოთ შემდეგი სახით :

$$f_{\vec{\beta}} : \vec{x} \mapsto \beta_0 + \sum_{j-1}^{p} \beta_j x_j \,. \tag{5.3}$$

აქ $\vec{\beta} \in \mathbb{R}^{p+1}$ და, ამრიგად, $m = p+1$.

2.2 ამოცანის გადაწყვეტა

განსაზღვრება 5.1 (წრფივი რეგრესია) $f_{\vec{\beta}} : \vec{x} \mapsto \beta_0 + \sum_{j=1}^{p} \beta_j x_j$ სახის მოდელს, რომლის კოეფიციენტები მიღებულია უმცირეს კვადრატთა ჯამის მინიმიზაციის გზით, *წრფივი რეგრესია* ეწოდება, ესე იგი :

$$\arg\min_{\vec{\beta}\in\mathbb{R}^{p+1}} \sum_{i=1}^{n}\left(y^i - \left(\beta_0 + \sum_{j=1}^{p}\beta_j x_j\right)\right)^2 . \tag{5.4}$$

✪

ჩვენ შეგვიძლია გადავწეროთ ამოცანა 5.4 მატრიცული ფორმით, თუ დაკვირვებათა $X \in \mathbb{R}^p$ მატრიცას დავუმატებთ მარცხნიდან ერთიანების (1-ის) სვეტს :

$$X \leftarrow \begin{pmatrix} 1 & x_1^1 & \cdots & x_p^1 \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_1^n & \cdots & x_p^n \end{pmatrix}. \tag{5.5}$$

უმცირეს კვადრატთა ჯამი ჩაიწერება შემდეგი ფორმით :

$$\mathrm{RSS} = \left(\vec{y} - X\vec{\beta}\right)^{\mathrm{T}} \left(\vec{y} - X\vec{\beta}\right). \tag{5.6}$$

აქ შემოღებული RSS აკრონიმი შეესაბამება ინგლისურ ფრაზას ***R**esidual **S**um of **S**quares* — სხვაობათა (ნარჩენთა) კვადრატების ჯამი.

RSS არის მეთოდი, რომლითაც ფასდება სხვაობა მონაცემებსა და შესაფასებელ მოდელს შორის. რაც უფრო ნაკლებია სხვაობა, მით უკეთესია შეფასება. ზემოთ მოცემულ თანაფარდობაში 5.6 სწორედ ეს ნოტაცია არის გამოყენებული.

ამრიგად, საქმე გვაქვს $\vec{\beta}$ პარამეტრის ამოზნექილ კვადრატულ ფორმასთან, რომელიც შეიძლება იყოს მინიმიზირებული მისი $\nabla_{\vec{\beta}}\mathrm{RSS} = -2\mathrm{X}^{\mathrm{T}}\left(\vec{y} - X\vec{\beta}\right)$ გრადიენტის განულებით. შედეგად მიიღება :

$$X^{\mathrm{T}} X \vec{\beta}^{*} = X^{\mathrm{T}} \vec{y}. \tag{5.7}$$

თეორემა 5.1 *თუ X მატრიცის რანგი მისი სვეტების რაოდენობას უდრის, მაშინ უმცირეს კვადრატთა RSS ჯამი მინიმუმს აღწევს შემდეგ წერტილზე :*

$$\vec{\beta}^{*} = \left(X^{\mathrm{T}} X\right)^{-1} X^{\mathrm{T}} \vec{y}.$$

✿

მტკიცებულება. თუ დაკვირვებათა X მატრიცას სრული სვეტის რანგი აქვს, მაშინ $X^{\mathrm{T}}X$ ინვერტირებადია.

□

თუ $X^{\mathrm{T}}X$ არ არის ინვერტირებადი, მაშინ, მიუხედავად ამისა, შესაძლებელია $\vec{\beta}$ პარამეტრისათვის არაუნიკალური (არაერთადერთი) ამონახსნის პოვნა $\left(X^{\mathrm{T}}X\right)^{-1}$-ის ნაცვლად $X^{\mathrm{T}}X -$ის ფსევდოინვერსიის გამოყენებით, მაგალითად, ამერიკელი მათემატიკოსის მურის (Eliakim Hastings Moore, 1862-1932) და ბრიტანელი მათემატიკოსის პენროუზის (Sir Roger Penrose, 1931), ესე იგი ისეთი M მატრიცის, რომლისთვისაც $X^{\mathrm{T}}X\,M\,X^{\mathrm{T}}X = X^{\mathrm{T}}X$ პირობა სრულდება.

შენიშვნა

ალტერნატივის სახით (და ამას უპირატესობაც კი უნდა მიეცეს, თუ p დიდია და $X^{\mathrm{T}}X \in \mathbb{R}^{p\times p}$ მატრიცის ინვერსია დაკავშირებულია დიდ დანახარჯთან) $\vec{\beta}$ პარამეტრისათვის შეიძლება შეფასების მიღება მიმართული დაშვების ალგორითმით (იხ. ქვედანაყოფი 13.3.3).

ყურადღება

ერთმანეთში არ უნდა ავურიოთ, ერთი მხრივ, *ცვლადები*, რომლებიც წარმოადგენს *მონაცემების აღმწერ* p რაოდენობის $x_1, x_2, \ldots, x_p$ სიდიდეს და, მეორე მხრივ, *პარამეტრები*, რომლებიც

წარმოადგენს *მოდელის აღმწერ* $p+1$ რაოდენობის $\beta_0, \beta_1, \ldots, \beta_p$ სიდიდეს (მოდელის პარამეტრიზებისათვის).

2.3 ინტერპრეტირებადობა

წრფივი რეგრესია ინტერპრეტირებად მოდელს ქმნის იმ გაგებით, რომ β_j პარამეტრებით საშუალებას გვაძლევს გავიგოთ ცვლადების შედარებითი მნიშვნელოვნობა წინასწარმეტყველებისათვის. მართლაც, რაც უფრო დიდია $|\beta_j|$, მით უფრო მეტი გავლენა აქვს $j-$ურ ცვლადს წინასწარმეტყველებაზე და β_j პარამეტრის ნიშანი ამ გავლენის მიმართულებასაც კი გვიჩვენებს.

ეს ინტერპრეტაცია სამართლიანია მხოლოდ იმ შემთხვევაში, თუ ცვლადები არაკორელირებულია და x_j მონაცემები შეიძლება იყოს შეცვლილი სხვა ცვლადთა მოქმედების დაუზიანებლად. უფრო მეტიც, თუ ცვლადები კორელირებულია, მაშინ X მატრიცას არ აქვს სვეტის სრული რანგი და $X^{\mathrm{T}}X$ მატრიცა ამიტომ ინვერტირებადი არ არის. ამრიგად, წრფივ რეგრესიას რამდენიმე ამონახსნი აქვს. ინტუიციით ძნელი მისახვედრი არ არის, რომ დასაშვებია ამ ამონახსნებში ერთიდან მეორეზე გადასვლა, ვინაიდან ერთ-ერთი β_j წონის დარღვევა შეიძლება იყოს კომპენსირებული x_j-თან კორელირებული ცვლადების წონათა ცვლილებით.

გარდა ამისა, საფრთხილო საკითხია სხვადასხვა ცვლადის სიდიდეთა რიგი. მაგალითად, დავუშვათ, რომ გვაქვს x_1 ცვლადი მნიშვნელობებით ასსა და ორასს შორის, რომლისთვისაც რეგრესიის კოეფიციენტი $\beta_1 = 0.5$ სიდიდეს უდრის და x_2 ცვლადი მნიშვნელობებით ერთსა და ორს შორის, რომლისთვისაც რეგრესიის კოეფიციენტი არის $\beta_2 = 5.$ მიუხედავად იმისა, რომ $\beta_1 < \beta_2$, გადაწყვეტილების მიღების $f_{\hat{\beta}}$ ფუნქცია სინამდვილეში უფრო მგრძნობიარეა ცვლილებებისადმი x_1-ში, ვიდრე ცვლილებებისადმი x_2 ცვლადში, რომლის მნიშვნელობები ორი რიგით ნაკლებია პირველი ცვლადის შესაძლო მნიშვნელობებზე. ამიტომ, მოქმედებათა დაწყებამდე სასურველია ცვლადების ცენტრირება და მათი მნიშვნელობების შესაძლო არეთა შეზღუდვა, სხვა სიტყვებით რომ ვთქვათ, საჭიროა მათი გარდასახვა :

$$x_j^i \leftarrow \frac{x_j^i - \overline{x}_j}{\sqrt{\frac{1}{n}\sum_{l=1}^{n}\left(x_j^l - \overline{x}_j\right)^2}},$$

სადაც $\overline{x}_j = \frac{1}{n}\sum_{l=1}^{n} x_j^l$.

ყურადღება

განზოგადების შეცდომის ემპირიული შეფასების პროცედურის კონტექსტში, ყურადღებით და დიდი სიფრთხილით უნდა მოხდეს *მხოლოდ დამსწავლელ კრებულზე* გამოთვლილი საშუალო მნიშვნელობისა და სტანდარტული გადახრის გამოყენება, რათა ზედმეტად არ შეფასდეს მიღებული შედეგი, საექსპლუატაციო მახასიათებლების ხარისხი და საერთოდ მოდელის

ეფექტურობა. ისიც უნდა გვახსოვდეს, რომ აქ სწავლების მომენტისათვის არც ვალიდაცია (სანდოობის შემოწმება, სისწორის დადასტურება) და არც ტესტირება არ ივარაუდება.

2.4 გაუს-მარკოვის თეორემა

განსაზღვრება 5.2 (BLUE) დავუშვათ, რომ მოცემულია q განზომილების პარამეტრთა ω ვექტორი და $\vec{y}$ ვექტორი, რომლითაც მას აფასებენ. მაშინ ამბობენ, რომ ω ვექტორის ω^* *შემფასებელი* (სხვანაირად, *შემფასებელი ფუნქცია*) წარმოადგენს *საუკეთესო წაუნაცვლებელ შემფასებელს* (ინგლ. BLUE — ***B**est **L**inear **U**nbiased **E**stimator*), თუ :

1. ω^* არის წრფივი შემფასებელი, ე.ი. არსებობს A მატრიცა, ისეთი რომ $\omega^* = A\vec{y}$;
2. ω^* არის წაუნაცვლებელი, ე.ი. $\mathbb{E}[\omega^*] = \omega$;
3. როგორიც არ უნდა იყოს ω ვექტორის $\tilde{\omega}$ წაუნაცვლებელი წრფივი შემფასებელი, ამ შემფასებლის დისპერსია მეტია ω^* შემფასებლის დისპერსიაზე ან მისი ტოლია.

⚙

თეორემა 5.2 *დავუშვათ, რომ გვაქვს n რაოდენობის $\vec{x}^1, \vec{x}^2, \ldots, \vec{x}^n \in \mathbb{R}^p$ დაკვირვება და მათი $y^1, y^2, \ldots, y^n \in \mathbb{R}$ ჭდეები. ასევე დავუშვათ, რომ ნებისმიერო მნიშვნელობის i ინდექსისათვის $y^i = \beta_0 + \sum_j^p \beta_j x_j^i + \mathcal{E}^i$ და $\mathcal{E}^i$ ცვლადები განაწილებულია ნორმალურად და ცენტრირებულია ნულოვანი (0) მნიშვნელობის მიმართ. მაშინ $\vec{\beta}$ ვექტორის შემფასებელი უმცირეს კვადრატთა მეთოდით არის ერთადერთი საუკეთესო წაუნაცვლებელი BLUE შემფასებელი.*

⚙

მტკიცებულება. უპირველეს ყოვლისა, $\vec{\beta}^* = \left(X^{\mathrm{T}}X\right)^{-1} X^{\mathrm{T}}\vec{y}$ და $\vec{\beta}^*$, ამრიგად, *წრფივია.* მისი მათემატიკური ლოდინი არის $\mathbb{E}\left[\vec{\beta}^*\right] = \mathbb{E}\left[\left(X^{\mathrm{T}}X\right)^{-1} X^{\mathrm{T}}\left(X\vec{\beta} + \mathcal{E}\right)\right]$.

რამდენადაც $X, \vec{y}$ და $\vec{\beta}$ არ არის შემთხვევითი სიდიდეები და რამდენადაც $\mathcal{E}$ ცვლადის მათემატიკური ლოდინი $0-$ის ტოლია, ამიტომ $\mathbb{E}\left[\left(X^{\mathrm{T}}X\right)^{-1} X^{\mathrm{T}}X\vec{\beta}\right] = \vec{\beta}$ თანაფარდობა მიიღება. ამრიგად, $\vec{\beta}^*$ არის $\vec{\beta}$ ვექტორის *წაუნაცვლებელი* შემფასებელი.

$\vec{\beta}^*$ სიდიდის დისპერსია შეადგენს :

$$\left.\begin{aligned}\mathrm{Var}\left(\vec{\beta}^*\right) = \mathrm{Var}\left(\left(X^{\mathrm{T}}X\right)^{-1} X^{\mathrm{T}}\vec{y}\right) = \mathrm{Var}\left(\left(X^{\mathrm{T}}X\right)^{-1} X^{\mathrm{T}}\left(X\vec{\beta} + \mathcal{E}\right)\right)\\ = \mathrm{Var}\left(\left(X^{\mathrm{T}}X\right)^{-1} X^{\mathrm{T}}\mathcal{E}\right) = \sigma^2\left(X^{\mathrm{T}}X\right)^{-1}\end{aligned}\right\}.$$

დაბოლოს, დავუშვათ, რომ $\tilde{\beta} = A\vec{y}$ — ეს $\vec{\beta}$ ვექტორის კიდევ ერთი წრფივი და წაუნაცვლებელი შემფასებელია. განსაზღვრების თანახმად, $\mathbb{E}\left[\tilde{\beta}\right] = \vec{\beta}$. თუ $\vec{y}$ ვექტორს ჩავანაცვლებთ მისივე

მნიშვნელობით, მივიღებთ, რომ $\mathbb{E}\left[A\left(X\vec{\beta}+\mathcal{E}\right)\right]=\vec{\beta}$ და, ამრიგად, $AX\vec{\beta}=\vec{\beta}$. ვინაიდან ეს სამართლიანია ნებისმიერი $\vec{\beta}$ ვექტორისათვის, ვასკვნით, რომ $AX=I$.

ახლა დავუშვათ, რომ $D=A-\left(X^{\mathrm{T}}X\right)^{-1}X^{\mathrm{T}}$, მაშინ $\tilde{\beta}-\vec{\beta}^{*}=D\vec{y}$. $\tilde{\beta}$ სიდიდის დისპერსიისათვის კი გვექნება :

$$\mathrm{Var}\left(\tilde{\beta}\right)=\mathrm{Var}\left(A\vec{y}\right)=\mathrm{Var}\left(A\mathcal{E}\right)=AA^{\mathrm{T}}\sigma^{2}.$$

ჩვენ შეგვიძლია AA^{T} ნამრავლის ჩაწერა შემდეგი სახით :

$$\left.\begin{aligned}AA^{\mathrm{T}}&=\left(D+\left(X^{\mathrm{T}}X\right)^{-1}X^{\mathrm{T}}\right)\left(D+\left(X^{\mathrm{T}}X\right)^{-1}X^{\mathrm{T}}\right)^{\mathrm{T}}\\&=\left(D+\left(X^{\mathrm{T}}X\right)^{-1}X^{\mathrm{T}}\right)\left(D^{\mathrm{T}}+X\left(X^{\mathrm{T}}X\right)^{-1}\right)\\&=DD^{\mathrm{T}}+DX\left(X^{\mathrm{T}}X\right)^{-1}+\left(X^{\mathrm{T}}X\right)^{-1}X^{\mathrm{T}}D^{\mathrm{T}}+\left(X^{\mathrm{T}}X\right)^{-1}\end{aligned}\right\}.$$

რადგან $AX=I, DX=AX-(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}X=0$, ამიტომ $X^{\mathrm{T}}D^{\mathrm{T}}=0$ მაშასადამე, მივიღებთ, რომ $AA^{\mathrm{T}}=DD^{\mathrm{T}}+(X^{\mathrm{T}}X)^{-1}$. ამრიგად, $\mathrm{Var}(\tilde{\beta})=\sigma^{2}DD^{\mathrm{T}}+\sigma^{2}(X^{\mathrm{T}}X)^{-1}=\mathrm{Var}(\vec{\beta}^{*})+\sigma^{2}DD^{\mathrm{T}}$. ვინაიდან $AX=I$, $DX=AX-(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}X=0$ და DD^{T} დადებითად ნახევრადგანსაზღვრულია, ამიტომ $\mathrm{Var}(\tilde{\beta})>\mathrm{Var}(\vec{\beta}^{*})$, სახელდობრ, თუ გამოვრიცხავთ $D=0$ შემთხვევას, როცა $\tilde{\beta}=\vec{\beta}^{*}$.

□

შენიშვნა

ჰიპოთეზა იმის შესახებ, რომ $\mathcal{E}$ ცვლადი განაწილებულია ნორმალურად, არ არის აუცილებელი : საკმარისია, რომ $\{\mathcal{E}_i\}_{i=1,\ldots,n}$ შეცდომებს ჰქონდეს ნულოვანი მათემატიკური ლოდინი, ერთნაირი სასრული მნიშვნელობის σ^2 დისპერსია (ამას თვისებას ეწოდება *ჰომოსკედასტურობა*)[1] და ისინი არ იმყოფებოდეს კორელაციაში ერთმანეთთან $\left(\mathrm{Cov}\left(\mathcal{E}_i,\mathcal{E}_l\right)=0,\ \text{როცა}\ i\neq l\right)$.

3 ლოგისტიკური რეგრესია

ახლა დავუშვათ, რომ ჩვენ გვინდა ბინარული კლასიფიკაციის ამოცანის ამოხსნა წრფივად, ესე იგი $y\in\{0,1\}$ პირობის მოდელირება ცვლადთა წრფივი კომბინაციის დახმარებით.

გონივრული არ იქნება y ცვლადის უშუალო მოდელირება რამდენიმე ნამდვილი ცვლადის წრფივი კომბინაციის სახით : ასეთ კომბინაციას შეუძლია არა ორი, არამედ მნიშვნელობათა

[1] ჰომოსკედასტურობა (ინგლ. homoscedasticity) — თვისება, რომელიც ნიშნავს ვექტორის ან შემთხვევით სიდიდეთა მიმდევრობის პირობითი დისპერსიის მუდმივობას ; დაკვირვებათა მნიშვნელობების ერთგვაროვანი ვარიაციულობა, რომელიც გამოისახება რეგრესიული მოდელის შემთხვევითი შეცდომის დისპერსიის ერთგვაროვნების სტაბილობაში — დისპერსიები ერთნაირია გაზომვის ყველა მომენტში.

უსასრულო რაოდენობის მიღება.

ამ შემთხვევაში ჩვენ შეგვეძლო ალბათური მოდელის განხილვა, რომელშიც $\mathbb{P}\left(Y = y \mid X = \vec{x}\right)$ ალბათობის მოდელირება $\vec{x}-$ის ცვლადთა წრფივი კომბინაციით ხდება. მაგრამ $\mathbb{P}\left(Y = y \mid X = \vec{x}\right)$ უნდა იძენდეს მნიშვნელობებს ნულსა (0) და ერთს (1) შორის და ეს ფუნქცია ინტუიციურად არ უნდა იყოს წრფივი : თუ $\mathbb{P}\left(Y = 0 \mid X = \vec{x}\right)$ ძალიან ახლოსაა $1-$თან, ე.ი. დიდია იმის ალბათობა, რომ $\vec{x}$ უარყოფითია, $\vec{x}-$ის მნიშვნელობის მცირე შეშფოთება დიდ გავლენას არ უნდა ახდენდეს ამ ალბათობაზე ; და, პირიქით, თუ შეშფოთება ძალიან ახლოსაა მნიშვნელობასთან 0.5, სხვა სიტყვებით, ძალიან მცირე ნდობისას $\vec{x}-$ის ჭდის მიმართ, არ არსებობს მიზეზი, რატომ არ უნდა იმოქმედოს $\vec{x}-$ის მნიშვნელობის მცირე შეშფოთებამ ამ ალბათობაზე. სწორედ ამიტომ არის კლასიკურად მიღებული $\vec{x}-$ის მნიშვნელობის *ლოგისტიკური გარდასახვის* (ინგლ. *logit transformation*) მოდელირება ცვლადების წრფივი კომბინაციის სახით.

განსაზღვრება 5.3 (logit ფუნქცია) *logit ფუნქცია* ეწოდება შემდეგი სახის ფუნქციას :

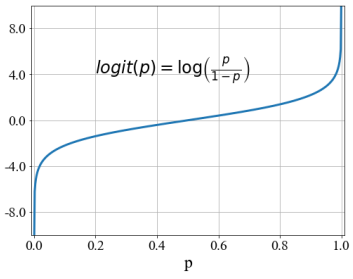
$$\left.\begin{aligned} \operatorname{logit} : [0,1] &\mapsto \mathbb{R} \\ p &\mapsto \log \frac{p}{1-p} \end{aligned}\right\}.$$

logit ფუნქცია (ნახატი 5.2*a*) შექცეული ფუნქციაა ლოგისტიკური ფუნქციის მიმართ (ნახატი 5.2*b*).

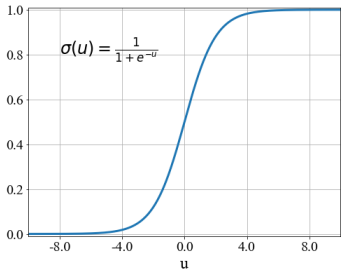
განსაზღვრება 5.4 (ლოგისტიკური ფუნქცია) *ლოგისტიკური ფუნქცია* ეწოდება შემდეგი სახის ფუნქციას :

$$\left.\begin{aligned} \sigma : \mathbb{R} &\to [0,1] \\ u &\mapsto \frac{1}{1+e^{-u}} = \frac{e^{u}}{1+e^{u}} \end{aligned}\right\}.$$

ფრთხილად იყავით და არ აურიოთ ერთმანეთში *ლოგისტიკური ფუნქცია* და *სტანდარტული* (*საშუალო კვადრატული*) *გადახრა* (ინგლ. *standard deviation*), თუმცა ორივე ხშირად წარმოდგენილია ერთისა და იმავე σ ნოტაციით.



(a) logit ფუნქცია.



(b) ლოგისტიკური ფუნქცია.

ნახატი 5.2 – logit და ლოგისტიკური ფუნქციები.

3.1 ფორმულირება

ამრიგად, ჩვენ ვცდილობთ $\log \frac{\mathbb{P}\left(Y=1 \mid \vec{x}\right)}{1-\mathbb{P}\left(Y=1 \mid \vec{x}\right)}$ გამოსახულების მოდელირებას $\vec{\beta}^{\mathrm{T}} \vec{x}$ სიდიდეთა წრფივი კომბინაციით, ან, ანალოგიურად, $\mathbb{P}\left(Y=1 \mid \vec{x}\right)$ ალბათობის მოდელირებას $\sigma\left(\vec{\beta}^{\mathrm{T}} \vec{x}\right)$ ლოგისტიკური ფუნქციით. აქ ჩვენ ვიყენებთ $\vec{x}$ სიდიდის გარდაქმნას 5.5.

დავუშვათ, რომ გვაქვს n დაკვირვებათა $\mathcal{D}=\left\{\vec{x}^i, y^i\right\}_{i=1,\ldots,n}$ კრებული. ასევე დავუშვათ, რომ ეს დაკვირვებები დამოუკიდებელია და ერთნაირად განაწილებული. მაშინ $\vec{\beta}$ სიდიდის დამაჯერებლობის ლოგარითმს შემდეგი სახე აქვს :

$$\left.\begin{aligned} \log \mathbb{P}\left(\mathcal{D} \mid \vec{\beta}\right) &= \log \prod_{i=1}^{n} \mathbb{P}\left(X=\vec{x}^i, Y=y^i \mid \vec{\beta}\right) = \log \prod_{i=1}^{n} \mathbb{P}\left(Y=y^i \mid \vec{x}^i, \vec{\beta}\right) + \log \prod_{i=1}^{n} \mathbb{P}\left(X=\vec{x}^i\right) \\ &= \sum_{i=1}^{n} \log \mathbb{P}\left(Y=1 \mid \vec{x}^i, \vec{\beta}\right)^{y^i} \left(1-\mathbb{P}\left(Y=1 \mid \vec{x}^i, \vec{\beta}\right)^{1-y^i}\right) + \mathcal{C} \\ &= \sum_{i=1}^{n} y^i \log \sigma\left(\vec{\beta}^{\mathrm{T}} \vec{x}^i\right) + \left(1-y^i\right) \log\left(1-\sigma\left(\vec{\beta}^{\mathrm{T}} \vec{x}^i\right)\right) + \mathcal{C} \end{aligned}\right\}, \quad (5.8)$$

სადაც $\mathcal{C}$ — კონსტანტაა $\vec{\beta}$ ვექტორის მიმართ. ჩვენ ვცდილობთ ამ დამაჯერებლობისათვის მაქსიმუმის მინიჭებას.

განსაზღვრება 5.5 (ლოგისტიკური რეგრესია) *ლოგისტიკური რეგრესია* ეწოდება $f: x \mapsto \sigma(\vec{\beta}^{\mathrm{T}} \vec{x})$ მოდელს, რომლის კოეფიციენტები მიღებულია ფორმულით

$$\arg \max_{\vec{\beta} \in \mathbb{R}^{p+1}} \sum_{i=1}^{ny} y^i \log \sigma\left(\vec{\beta}^{\mathrm{T}} \vec{x}^i\right) + \left(1-y^i\right) \log\left(1-\sigma\left(\vec{\beta}^{\mathrm{T}} \vec{x}^i\right)\right). \quad (5.9)$$

$\vec{\beta}$ ვექტორის ალბათობის მაქსიმიზაცია ამ მოდელის ფარგლებში ღირებულების ლოგისტიკური ფუნქციის გამოყენებით განსაზღვრული ემპირიული რისკის მინიმიზაციის ეკვივალენტურია (იხ. ქვედანაყოფი 2.4.1).

3.2 ამოცანის გადაწყვეტა

ლოგისტიკური რეგრესიის 5.8 დამაჯერებლობა არის ჩაზნექილი ფუნქცია.

თეორემა 5.3 *ლოგისტიკური რეგრესიის დამაჯერებლობის მაქსიმუმის გრადიენტი $\vec{\beta}$ ვექტორით უდრის :*

$$\sum_{i=1}^{n} \left(y^i - \frac{1}{1+e^{-\vec{\beta}^T \vec{x}^i}}\right) \vec{x}^i \,.$$

მტკიცებულება. ეს შედეგი ლოგისტიკური ფუნქციის ერთი არაჩვეულებრივი თვისების გამოყენებით მიიღება :

$$\sigma^{l}(u)=\sigma(u)\big(1-\sigma(u)\big).$$

□

ეს გრადიენტი არ შეიძლება იყოს განულებული ანალიზურად, ამიტომ ლოგისტიკურ რეგრესიას არ აქვს ცხადი ამონახსნი. ჩვეულებრივ, ასეთი რეგრესიის ამონახსნი მიიღება გრადიენტული ალგორითმით ან მისი ერთ-ერთი ნაირსახეობის გამოყენებით. ეს ალგორითმები იკრიბება — ისწრაფვის ოპტიმალური ამონახსნისკენ, ვინაიდან დამაჯერებლობა ჩაზნექილია და ამიტომ გამორიცხავს ლოკალური მაქსიმუმის არსებობას.

4 პოლინომური რეგრესია

d ხარისხის პოლინომური რეგრესიის შემთხვევაში შემდეგი სახის გადაწყვეტილების მიღების ფუნქცია იძებნება :

$$f:\vec{x}\mapsto\beta_0^0+\sum_{j=1}^{p}\beta_j^1x_j+\sum_{j=1}^{p}\sum_{k=1}^{p}\beta_{jk}^2x_jx_k+\cdots+\underbrace{\sum_{j=1}^{p}\cdots\sum_{\xi=1}^{p}}_{d\ \text{ჯამი}}\beta_{jk\ldots\xi}^{d}x_jx_{\,k}\ldots x_\xi. \tag{5.10}$$

ფაქტობრივად, საქმე ეხება გარკვეულ წრფივ რეგრესიას $\begin{pmatrix} d+p \\ p \end{pmatrix}$ რაოდენობის ცვლადზე.

5 საკვანძო მომენტები

- პარამეტრული რეგრესიული მოდელის კოეფიციენტები შეიძლება იყოს მიღებული დამაჯერებლობის მაქსიმიზაციით, რაც ემპირიული რისკის მინიმიზაციის ეკვივალენტურია კვადრატული ღირებულების გამოყენებისას დანაკარგთა ფუნქციად და უმცირეს კვადრატთა მეთოდის ტოლფასიც არის.
- წრფივ რეგრესიას ერთადერთი $\vec{\beta}^{*}=\left(X^{T}X\right)^{-1}X^{T}\vec{y}$ ამონახსნი აქვს მაშინ და მხოლოდ მაშინ, როცა $X^{T}X$ ინვერტირებადია. წინააღმდეგ შემთხვევაში ამონახსნების უსასრულო რაოდენობა არსებობს.
- პოლინომური რეგრესია დაიყვანება წრფივ რეგრესიამდე.
- ბინარული კლასიფიკაციის ამოცანისათვის ლოგისტიკური ღირებულების გამოყენება დანაკარგების ფუნქციად ემპირიული რისკის მინიმიზაციისას დამაჯერებლობის ისეთი მოდელის მაქსიმიზაციამდე დაიყვანება, რომელშიც დადებითი კლასისადმი კუთვნილების ალბათობა ცვლადთა წრფივი კომბინაციის ლოგისტიკურ გარდაქმნას წარმოადგენს.
- ლოგისტიკურ რეგრესიას არ აქვს ანალიზური ამონახსნი, მაგრამ მოდელის სწავლებისათვის შეიძლება მიმართული დაშვების ალგორითმის გამოყენება.

დამატებითი ინფორმაცია

- ლოგისტიკური რეგრესია, რა თქმა უნდა, შეიძლება იყოს გაფართოებული მრავალკლასიან კლასიფიკაციამდე ჯვარედინი ენტროპიის (კროს-ენტროპიის) გამოყენებით ღირებულების ფუნქციად (ინგლ. *cost function*).

- ჩვენ ვნახეთ, როგორ ავაგოთ პოლინომური მოდელები შემავალი ცვლადების პოლინომური გარდაქმნის გამოყენებით. ეს ერთადერთი გარდაქმნა არ არის, რომელიც შეიძლება ამ ცვლადებზე განხორციელდეს. შეიძლება ასევე განვიხილოთ, მაგალითად, უბან-უბან პოლინომური ფუნქციები (ინგლ. splines — სპლაინები,) ან ვეივლეტები (ინგლ. wavelets — პატარა ტალღები). ვეივლეტიც მათემატიკური ფუნქციაა, რომელიც მონაცემების სხვადასხვა სიხშირული კომპონენტის ანალიზის საშუალებას იძლევა. მისი გრაფიკი გამოიყურება ტალღურ რხევად მილევადი ამპლიტუდით კოორდინატთა სათავიდან გარბენის პროცესში. ამ მეთოდების ჯგუფი ცნობილია სახელწოდებით *ბაზისის გაფართოებები* (ინგლ. *basis expansions*). მათ ღრმა განხილვას ეძღვნება მეხუთე თავი ტრევორ ჰასტის, რობერტ ტიბშირანის და ჯერომ ფრიდმანის წიგნში — Trevor Hastie, Robert Tibshirani და Jerome Friedman (2009). უფრო მეტიც, ეს ნაშრომი ამ ლექციაში განხილულ კონცეფციათა გაცილებით ღრმა ასპექტებსაც ეხება და შეიძლება ძალიან სასარგებლო აღმოჩნდეს მკითხველისათვის.

6 ბიბლიოგრაფია

1. Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer-Verlag, New York, 2nd edition, 764 pages.

https://hastie.su.domains/Papers/ESLII.pdf

7 სავარჯიშოები

5.1 როგორი ალგორითმის გამოყენებაა საჭირო წრფივი რეგრესიის სწავლებისათვის მონაცემთა ნაკრებზე, რომელიც შეიცავს 10^5 დაკვირვებას და 5 ცვლადს? 10^5 ცვლადს და 5 დაკვირვებას? 10^5 ცვლადს და დაკვირვებათა ასეულ მილიონებს?

5.2 მაქსიმალური დამაჯერებლობის წრფივი რეგრესიის კოეფიციენტების დასადგენად საჭირო თუ არის იმის დაშვება, რომ დაკვირვებები დამოუკიდებელია და ერთნაირად განაწილებული? რომ ხმაური გაუსის ტიპისაა? რომ $\vec{x}^i$ დაკვირვებები დამოუკიდებელია y^i ჭდეებისგან?

5.3 პოლინომური რეგრესიის სწავლების ჩატარების შემდეგ ანიკა ხვდება, რომ სწავლების შეცდომა ბევრად ნაკლებია ვალიდაციის (შემოწმების) შეცდომაზე. რატომ ხდება ეს და რისი გაკეთება შეუძლია მას?

5.4 განვიხილოთ 5 დაკვირვება ორ განზომილებაში, რომლებსაც მიწერილი აქვს y-ით აღნიშნული ჭდე:

x_1	0.12	0.14	0.31	0.37	0.49
x_2	29.0	33.0	17.0	21.0	12.0
y	21.0	24.3	12.7	15.6	9.0

1. როგორია წრფივი რეგრესიის პარამეტრები ამ მონაცემებიდან გამომდინარე ?

2. როგორია მოდელის საშუალო კვადრატული შეცდომა ამ მონაცემებით ?

3. რა მოხდება, თუ გავამრავლებთ x_1 დაკვირვებას $10-$ზე და შევამცირებთ შედეგს $1-$ით ?

5.5 თუ ახდენს გავლენას წრფივ რეგრესიაზე მონაცემთა შესამჩნევი გადახრების არსებობა ?

5.6 წრფივი რეგრესიის გავლების შემდეგ ბენომ შეამჩნია, რომ ნაშთები, ე.ი. განსხვავება ნაწინასწარმეტყველ და ფაქტობრივ მნიშვნელობებს შორის, კორელირებულია ნაწინასწარმეტყველ სიდიდეებთან. უნდა ააღელვოს ბენო მისი მოდელის ამ თავისებურებამ ?

5.7 კატო მივიდა მე-4 ხარისხის პოლინომურ რეგრესიამდე მონაცემებზე, რომლებიც სინამდვილეში წარმოქმნილი (ნაწარმოები) იყო მე-3 ხარისხის პოლინომით. რა შეიძლება დამაჯერებლად მოხდეს?

5.8 დაჩიმ ჩაუტარა სწავლება წრფივ რეგრესიას თავის მონაცემებზე. იგი მივიდა დასკვნამდე, რომ მოდელი საკმარისად ნასწავლი არის. უნდა დაუმატოს მან ცვლადების რაოდენობა, თუ შეამციროს? ღირს პოლინომური რეგრესიის მოსინჯვა?

5.9 ელენიკო აპირებს ლოგისტიკური რეგრესიის გამოყენებას 3 კლასად კლასიფიცირების ამოცანის გადასაჭრელად. ეს როგორ უნდა განხორციელდეს?

სავარჯიშოთა ამონახსნები

5.1 როცა $X^{\mathrm{T}}X$ მატრიცა მცირეა (ცვლადების რაოდენობაა უმნიშვნელო), შეიძლება მატრიცის ინვერსიის ალგორითმის გამოყენება. წინააღმდეგ შემთხვევაში გაცილებით უკეთესი იქნება გრადიენტული ალგორითმი.

5.2 დაკვირვებები დამოუკიდებელი და ერთნაირად განაწილებულია, ხმაური კი — გაუსის ტიპისაა, მაგრამ ჭდეები დაკვირვებებისგან დამოუკიდებელი არ არის : სწორედ მათი დამოკიდებულების მოდელირების მცდელობა ხდება.

5.3 მოდელის სწავლება აშკარად გადამეტებით ხდება. ანიკას შეუძლია მოდელის პოლინომის ხარისხის შემცირება შეეცადოს ნაკლებად რთული მოდელის მისაღებად.

5.4

1. $\vec{\beta}=\left(X^{\mathrm{T}}X\right)^{-1}X^{\mathrm{T}}y$ თანაფარდობა გვაძლევს $y\approx -0.577+1.499x_1+0.744x_2$ გამოსახულებას.

2. $\frac{1}{5}\sum_{i=1}^{5}\left(y^i-\left(-0.577+1.499x_1^i+0.744x_2^i\right)\right)^2\approx 0.016$.

3. x_1-ის გამრავლებას $10-$ზე მოსდევს β_1-ის გაყოფა $10-$ზე. ამ შედეგიდან $1-$ის გამოკლება აისახება β_0 წანაცვლებაზე, ხოლო საშუალო კვადრატული შეცდომა იგივე რჩება.

5.5 დიახ (შესამჩნევად გადახრილი მონაცემები ეცდება წაანაცვლოს გადაწყვეტილების მიღების ფუნქცია თავისკენ).

5.6 წრფივ რეგრესიაში ხმაურის მოდელირება შემთხვევითი სიდიდით ხდება და იგი კორელირებული არ უნდა იყოს პროგნოზებთან. ასე რომ ბენოს აქვს დარდის მიზეზი.

5.7 არსებობს იმის დიდი ალბათობა, რომ მოხდება კატოს მოდელის ზედმეტად (გადამეტებულად) სწავლება.

5.8 დაჩის მოდელი საკმარისად რთული არ არის : შეიძლება ცვლადების დამატება და მოდელის პოლინომური ხარისხის გადიდება.

5.9 ელენიკოს შეუძლია *«ერთი ყველას წინააღმდეგ»* ან *«ერთი ერთის წინააღმდეგ»* მიდგომათა გამოყენება.

ლექცია 6 რეგულარიზაცია

შინაარსი

როცა ექსპლიციტური (აშკარად, განსაზღვრულად, არაორაზროვნად გამოხატული), არსის გადმომცემი ცვლადები კორელირებულია ან მათი რაოდენობა ძალიან დიდია, წრფივი რეგრესიული მოდელის სირთულე ხშირად ძალზე მაღალია, რაც ზედმეტად სწავლების მდგომარეობას ქმნის.

ლექციაში ჩვენ განვიხილავთ ამ მოდელების *რეგულარიზებას* რეგრესიის კოეფიციენტების მართვით ე.ი. იმ წონების მართვით, რომლებიც მიწერილი აქვს თითოეულს ამ ცვლადებიდან მათ წრფივ კომბინაციაში. ყურადღება კონცენტრირებული იქნება წრფივ რეგრესიაზე, მაგრამ ამ ლექციაში წარმოდგენილ მეთოდებს კიდევ უფრო ფართო გამოყენება აქვს ლოგისტიკური რეგრესიის შემთხვევაში.

მიზნები

- მოდელის სირთულის მართვა რეგულარიზაციის საშუალებით.
- LASSO-ს, თხემური რეგრესიისა და ელასტიკური ქსელის განსაზღვრება.
- ℓ_1 და ℓ_2 ნორმათა, როგორც *რეგულარიზატორების*, გაგება.
- *რეგულარიზაციის კოეფიციენტის* არჩევა.

1 რეგულარიზაციის არსი

თუ ცვლადები ძლიერ კორელირებულია ან მათი რიცხვი აჭარბებს დაკვირვებათა რაოდენობას, მაშინ $X \in \mathbb{R}^{p+1}$ მატრიცას, რომელიც ჩვენს მონაცემებს წარმოადგენს, არ შეუძლია იყოს

სრული სვეტის რანგის. ამიტომ $X^{\mathrm{T}}X$ მატრიცა არ არის ინვერტირებადი და წრფივი რეგრესიის უმცირეს კვადრატთა მინიმიზაციით მიღებული ერთადერთი ამონახსნი არ მოიძებნება. ამიტომ არსებობს ზედმეტად სწავლების (გაწვრთნის) რისკი : რაკი მოდელი ერთადერთი არ არის, როგორ შეგვიძლია იმის გარანტირება, რომ ჩვენ მიერ არჩეული მოდელი ყველაზე უკეთ ახორციელებს განზოგადებას ?

ზედმეტად სწავლების (წვრთნის) რისკის შესამცირებლად ჩვენ ვეცდებით ერთდროულად ვაკონტროლოთ მოდელის შეცდომა საწვრთნელ სიმრავლეზე და რეგრესიის კოეფიციენტების მნიშვნელობები, რომლებიც თითოეულ ცვლადს ენიჭება. ამ კოეფიციენტების მართვა — ეს მოდელის სირთულის მართვის ხერხია : შემდეგ ჩვენ ვნახავთ, რომ ეს მართვა მდგომარეობს იმაში, რომ კოეფიციენტები უნდა მიეკუთვნებოდეს $\mathbb{R}^{p+1}$ სივრცის მკაცრად განსაზღვრულ ქვესიმრავლეს, და არ შეიძლება ამ სივრცეში იძენდეს ნებისმიერ მნიშვნელობას, რაც ზღუდავს შესაძლო გადაწყვეტილებათა სივრცეს.

განსაზღვრება 6.1 (რეგულარიზაცია) *რეგულარიზაცია* ეწოდება მოდელის სწავლებას საწვრთნელ სიმრავლეზე ემპირიული რისკისა და Ω შემზღუდველი წევრის შესაძლო გადაწყვეტილებებზე ჯამის მინიმიზაციის გზით :

$$f = \arg\min_{h\in\mathcal{F}} \frac{1}{n}\sum_{i}^{n} L\left(h\left(\vec{x}^{\,i}\right), y^{i}\right) + \lambda\Omega(h), \tag{6.1}$$

სადაც რეგულარიზაციის $\lambda \in \mathbb{R}_{+}$ კოეფიციენტი მართავს თითოეული შესაკრები წევრის (მაჩვენებლის) ფარდობით მნიშვნელოვნობას.

წრფივი რეგრესიული მოდელის შემთხვევაში დანაკარგის ფუნქციად ჩვენ გამოვიყენებთ უმცირეს კვადრატთა ჯამს. რეგულარიზატორები, რომლებსაც განვიხილავთ, წარმოადგენს რეგრესიის კოეფიციენტების $\vec{\beta}$ ვექტორის ფუნქციას :

$$\arg\min_{\vec{\beta}\in\mathbb{R}^{p+1}} \left(\vec{y} - X\vec{\beta}\right)^{\mathrm{T}} \left(\vec{y} - X\vec{\beta}\right) + \lambda\Omega\left(\vec{\beta}\right),$$

ან, ანალოგიური გზით,

$$\arg\min_{\vec{\beta}\in\mathbb{R}^{p+1}} \left\|\left(\vec{y} - X\vec{\beta}\right)\right\|_2^2 + \lambda\Omega\left(\vec{\beta}\right). \tag{6.2}$$

აქ ჩვენ ვიყენებთ წინა ლექციის მე-5 წესით $\vec{x}$-ის გარდაქმნას, რომელიც წარმოადგენს ერთიანების სვეტის დამატებას X საანგარიშო (საპროექტო) მატრიცაში აღნიშვნათა გასამარტივებლად.

ჩანართი

რეგულარიზაციის კოეფიციენტი

რეგულარიზაციის λ კოეფიციენტი არის რეგულარიზებული წრფივი რეგრესიის ჰიპერპარა-

მეტრი. რეგულარიზაციის კოეფიციენტი აკონტროლებს ბალანსს საწვრთნელ ანარჩევთან მისადაგებასა და ჭარბი სირთულისათვის დასანიშნავ ჯარიმას შორის.

λ პარამეტრის მიახლოებისას $(+\infty)$ – ისადმი რეგულარიზაციის წევრის იზრდება, ვიდრე არ დაიწყებს დომინირებას შეცდომის წევრზე და მნიშვნელობას შეიძენს მხოლოდ რეგულარიზატორის მინიმიზაციას. უმრავლეს შემთხვევაში რეგულარიზატორის მინიმიზირება ხდება $\vec{\beta} = \vec{0}$ პირობის მიღწევისას და შემდგომი სწავლება (წვრთნა) წყდება.

და პირიქით, როცა λ მიისწრაფვის 0-კენ, რეგულარიზაციის წევრი ხდება უმნიშვნელო შეცდომის წევრთან შედარებით, მისი უგულებელყოფის შესაძლებლობა ჩნდება და $\vec{\beta}$ იღებს, მნიშვნელობად არარეგულარიზებული წრფივი რეგრესიის ამონახსნს.

ნებისმიერო სხვა ჰიპერპარამეტრის მსგავსად, λ შეიძლება იყოს არჩეული კროს-ვალიდაციის (ჯვარედინი შემოწმების) გზით. ჩვეულებრივ, გამოიყენება მნიშვნელობათა ლოგარითმული ბადე.

2 თხემური რეგრესია

რეგულარიზაციის ერთ-ერთ ყველაზე გავრცელებულ ფორმას, რომელსაც ხშირად მიმართავენ ცუდად დასმულ შებრუნებულ ამოცანებთან დაკავშირებულ მრავალ სფეროში, წარმოადგენს $\vec{\beta}$ ვექტორის ℓ_2 ნორმის გამოყენება რეგულარიზატორად :

$$\Omega_{\text{ridge}}\left(\vec{\beta}\right) = \left\|\vec{\beta}\right\|_2^2 = \sum_{j=0}^{p} \beta_j^2 \,. \tag{6.3}$$

აქ ისიც შეიძლება გავიხსენოთ, რომ ნორმა (ლათ. norma) მანქანურ სწავლებაში შეიძლება აღნიშნავდეს წესს ან რეგულაციას, რომელიც ეხება გარკვეულ სუბსტანციას და მოითხოვს თავის შესრულებას.

2.1 თხემური რეგრესიის ფორმულირება

განსაზღვრება 6.2 (თხემური რეგრესია) *თხემური რეგრესია,* ანუ *რიჯ-რეგრესია* (ინგლ. *ridge regression*) ეწოდება $f : x \mapsto \vec{\beta}^{\mathrm{T}} \vec{x}$ მოდელს, რომლის კოეფიციენტები მიღებულია შემდეგი თანაფარდობით :

$$\arg\min_{\vec{\beta} \in \mathbb{R}^{p+1}} \left\|\left(\vec{y} - X\vec{\beta}\right)\right\|_2^2 + \lambda \left\|\vec{\beta}\right\|_2^2 . \tag{6.4}$$

⚙

თხემური რეგრესია *ტიხონოვის რეგულარიზაციის* კერძო შემთხვევაა. ეს რეგულარიზაცია რუსი მათემატიკოსისა და გეოფიზიკოსის ა. ნ. ტიხონოვის (Андрей Николаевич Тихонов, 1906-1993) მიერ იყო შემოთავაზებული 1963 წელს ფრედგოლმის (Erik Ivar Fredholm, 1866–1927) ინტეგრალურ განტოლებათა ამოსახსნელად. იგი ასევე გამოიყენება ნეირონულ ქსელებშიც, სადაც მას *წონის დეგრადაცია* (ინგლ. weight decay) ეწოდება (იხ. ქვედანაყოფი 7.2.4).

2.2 ამოცანის გადაწყვეტა

უკანასკნელი 6.4 თანაფარდობით წარმოდგენილი პრობლემა ამოზნექილი ოპტიმიზაციის ამოცანაა (იხ. ლექცია 13) : იგი გულისხმობს კვადრატული ფორმის მინიმიზაციას და მისი ამოხსნა მიზნობრივი ფუნქციის გრადიენტის ($\vec{\beta}$ -თი) განულებით ხდება.

$$\nabla_{\vec{\beta}}\left(\left\|\left(\vec{y}-X\vec{\beta}\right)\right\|_2^2+\lambda\left\|\vec{\beta}\right\|_2^2\right)=0\,. \tag{6.5}$$

თუ აღვნიშნავთ $I_p\in\mathbb{R}^{p\times p}$ სიმბოლოთი p განზომილების იდენტურობის მატრიცას (ინგლ. Identity matrix), მივიღებთ :

$$\left(\lambda I_p+X^{\mathrm{T}}X\right)\vec{\beta}^{*}=X^{\mathrm{T}}\vec{y}\,. \tag{6.6}$$

ვინაიდან $\lambda>0$, ამიტომ $\lambda I_p+X^{\mathrm{T}}X$ მატრიცა ყოველთვის ინვერტირებადია (შექცევადია). მაშასადამე ჩვენს ამოცანას ყოველთვის ერთადერთი ცალსახა მნიშვნელობა აქვს. რეგულარიზაციამ ℓ_2 ნორმით აქცია ცუდად დასმული ამოცანა კარგად დასმულ ამოცანად, რომლის ამონახსნი ასეთია :

$$\vec{\beta}^{*}=\left(\lambda I_p+X^{\mathrm{T}}X\right)^{-1}X^{\mathrm{T}}\vec{y}\,. \tag{6.7}$$

შენიშვნა

თუ x_j ცვლადი გამრავლდება α კონსტანტაზე, მაშინ შესაბამისი კოეფიციენტი არარეგულარიზებულ წრფივ რეგრესიაში იყოფა α -ზე. მართლაც, თუ შემოვიღებთ X^{*} ნოტაციას მატრიცისათვის, რომელიც მიიღება x_j -ის ჩანაცვლებით αx_j -ით X მატრიცაში, მაშინ შესაბამისი წრფივი რეგრესიის $\vec{\beta}^{*}$ ამონახსნი აკმაყოფილებს $X^{*}\left(\vec{y}-X^{*}\vec{\beta}^{*}\right)=0$ განტოლებას, ხოლო წრფივი რეგრესიის $\vec{\beta}$ ამონახსნი X მატრიცაზე აკმაყოფილებს $X\left(\vec{y}-X\vec{\beta}\right)=0$ განტოლებას.

და პირიქით, თხემური რეგრესიის შემთხვევაში x_j -ის ჩანაცვლება αx_j -ით ასევე ახდენს გავლენას რეგულარიზაციის წევრზე და გაცილებით უფრო რთული ეფექტი აქვს. ამრიგად, სხვადასხვა ცვლადის შედარებით მასშტაბს არ შეუძლია ძლიერი გავლენის მოხდენა თხემური რეგრესიის შედეგებზე. ამიტომ სწავლების წინ სასურველია ცვლადების სტანდარტიზება, ესე იგი მათი სტანდარტული გადახრის შემცირება 1-ის ტოლ მნიშვნელობამდე ამ ცვლადების გაყოფით შესაბამის სტანდარტულ გადახრაზე :

$$x_j^i\leftarrow\frac{x_j^i}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_j^i-\frac{1}{n}\sum_{i=1}^{n}x_j^i\right)^2}}\,. \tag{6.8}$$

ყურადღება

ჭარბი სწავლების თავიდან ასაცილებლად მნიშნელოვანია, რომ ეს სტანდარტული გადახრა გამოითვლებოდეს მხოლოდ საწვრთნელ სიმრავლეზე და ამის შემდეგ გამოიყენებოდეს ტესტირების ან შემოწმების სიმრავლეზე.

თხემური რეგრესია «დამაჯგუფებელ» გავლენას ახდენს კორელირებულ ცვლადებზე იმ გაგებით, რომ ასეთ ცვლადებს მსგავსი კოეფიციენტები ექნება.

2.3 ალბათური მიდგომა

ისევე, როგორც $\vec{\beta}^*$ ვექტორის დადგენა კვადრატული ღირებულებით განსაზღვრული ემპირიული რისკის მინიმიზაციით, $\vec{\beta}^*$ ვექტორის შეფასების ეკვივალენტურია დამაჯერებლობის მაქსიმუმით ცენტრირებული გაისის ხმაურის დაშვების პირობებში (იხ. ქვედანაყოფი 5.1.2), ზუსტად ასევე ℓ_2 ნორმით რეგულარიზებული ემპირიული რისკის (განტოლება 6.4) მინიმიზაცია $\vec{\beta}^*$ ვექტორის შეფასების ეკვივალენტურია აპოსტერიორული მაქსიმუმით ცენტრირებული გაუსის ხმაურის დაშვებისა და კიდევ ერთი, დამატებითი, ვარაუდის პირობებში იმის შესახებ, რომ $\vec{\beta}^*$ აღებულია შემთხვევითი გაუსის ცენტრირებული B ვექტორიდან.

ეს ახალი ჰიპოთეზა ნიშნავს აპრიორულ დაშვებას იმის შესახებ, რომ $\vec{\beta}^*$ რეგრესიის კოეფიციენტი არ შეიძლება იყოს ძალიან დაშორებული 0 -გან (B ვექტორის მათემატიკური ლოდინისგან), ამასთან ერთად, B -ვექტორის დისპერსია უშვებს მხოლოდ მეტ-ნაკლებად მკაცრ დამოკიდებულებას ამ სიახლოვისადმი B -თან.

უფრო ზუსტად, დავუშვათ, რომ მონაცემების მონიშნული $\mathcal{D} = \left\{ \vec{x}^i, y^i \right\}_{i=1,\ldots,n}$ მასივი, რომელშიც p განზომილების n დაკვირვებაა მათი რეალური ჭდეებით, წარმოადგენს (X, Y) წყვილის სტატისტიკურ ანარჩევს. ამ წყვილში X არის ვექტორი p_X სიმკვრივით, ხოლო Y — შემთხვევითი ცვლადია გარკვეული ალბათური სიმკვრივით. მოცემული $\vec{x} \in \mathbb{R}^p$ პირობისათვის დავუშვათ, რომ $Y \mid X = \vec{x} \sim \mathcal{N}\left(f_{\vec{\beta}}(\vec{x}), \sigma^2 \right)$, რაც ასახავს ჰიპოთეზას გაუსის ხმაურის შესახებ, ხოლო $\vec{\beta}$ არის $B \sim \mathcal{N}\left(0, \sigma_B^2\right)$ შემთხვევითი ცვლადის რეალიზაცია.

$B \mid \left\{ \left(X = \vec{x}^i, Y = y^i \right) \right\}_{i=1,\ldots,n}$ შემთხვევით ცვლადს აქვს ალბათობათა განაწილების სიმკვრივე და ეს სიმკვრივე შემდეგი ფორმულით მოიცემა :

$$
\begin{aligned}
p_{B|\mathcal{D}}\left(\vec{b}\right) &= \prod_{i=1}^{n} \frac{p_{X,Y|B=\vec{b}}\left(\vec{x}^i, y^i\right) p_B\left(\vec{b}\right)}{p_{X,Y}\left(\vec{x}^i, y^i\right)} = \prod_{i=1}^{n} \frac{p_{Y|X=\vec{x}^i,B=\vec{b}}\left(y^i\right) p_X\left(\vec{x}^i\right) p_B\left(\vec{b}\right)}{p_{X,Y}\left(\vec{x}^i, y^i\right)} \\
&= \frac{1}{\mathcal{K}} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}\left(y^i - f_{\vec{b}}\left(\vec{x}^i\right)\right)^2 \right) \frac{1}{\sqrt{2\pi}\sigma_B} \exp\left(-\frac{1}{2\sigma_B^2}\left\langle \vec{b}, \vec{b} \right\rangle \right), \\
&= \frac{1}{\mathcal{K}_2} \exp\left(\sum_{i=1}^{n} -\left(y^i - f_{\vec{b}}\left(\vec{x}^i\right)\right)^2 - \frac{\sigma^2}{\sigma_B^2} \left\| \vec{b} \right\|_2 \right)
\end{aligned}
$$

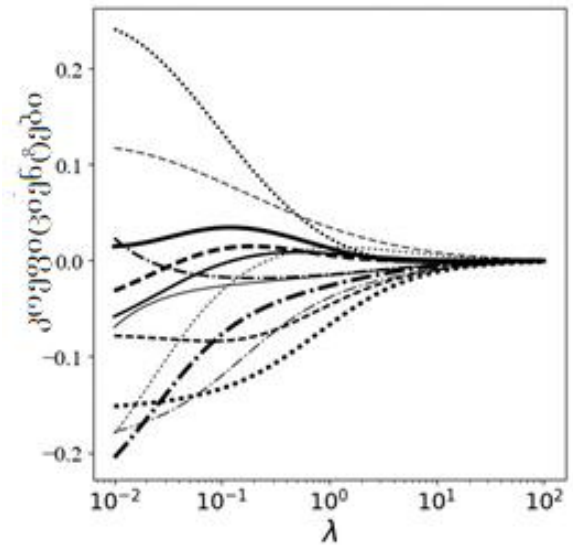
სადაც $\mathcal{K}$ და $\mathcal{K}_2$ დამოკიდებული არ არის $\vec{b}$ -ზე. ასეთ დაშვებათა პირობებში $\vec{\beta}$ სიდიდის მაქსიმალური აპოსტერიორული შეფასება მინიმუმს ანიჭებს ℓ_2 წევრით რეგულარიზებულ ემპირიულ რისკს რეგულარიზაციის $\lambda = \frac{\sigma^2}{\sigma_B^2}$ კოეფიციენტით.

უნდა აღინიშნოს, რომ ეს ეკვივალენტობა სამართლიანია $f_{\vec{\beta}}$ მოდელის პარამეტრული ფორმისაგან დამოუკიდებლად, რომელიც შეიძლება არც კი იყოს აუცილებლად წრფივი.

2.4 რეგულარიზაციის გზა

განსაზღვრება 6.3 (რეგულარიზაციის გზა) ცვლადის რეგრესიის კოეფიციენტის მნიშვნელობის ევოლუციის (ცვლილების, დინამიკის) ფუნქციურ დამოკიდებულებას რეგულარიზაციის λ კოეფიციენტზე *რეგულარიზაციის გზა* ეწოდება.

რეგულარიზაციის გზა საშუალებას გვაძლევს გავიგოთ რეგულარიზაციის გავლენა $\vec{\beta}$ ვექტორის მნიშვნელობებზე. ნახატზე 6.1 მოცემულია შესაბამისი მაგალითი.



ნახატი 6.1 - თხემური რეგრესიის რეგულარიზაციის გზა მონაცემების ნაკრებისათვის 12 ცვლადით. ყოველი წირი წარმოადგენს ამ ცვლადებიდან თითოეულის რეგრესიის კოეფიციენტის ცვლილებას λ პარამეტრის გაზრდისას : რეგრესიის კოეფიციენტი ინაცვლებს თავისი: მნიშვნელობიდან არარეგულარიზებულ რეგრესიაში ნულისკენ.

2.5 გეომეტრიული ინტერპრეტაცია

თეორემა 6.1 *დავუშვათ, რომ მოცემულია* $\lambda \in \mathbb{R}_+$, $X \in \mathbb{R}^{n\times p}$ *და* $\vec{y} \in \mathbb{R}^n$. *მაშინ არსებობს ერთადერთი* $t \in \mathbb{R}_+$ *სიდიდე, ისეთი რომ ამოცანა 6.4 შემდეგი ამოცანის ეკვივალენტურია :*

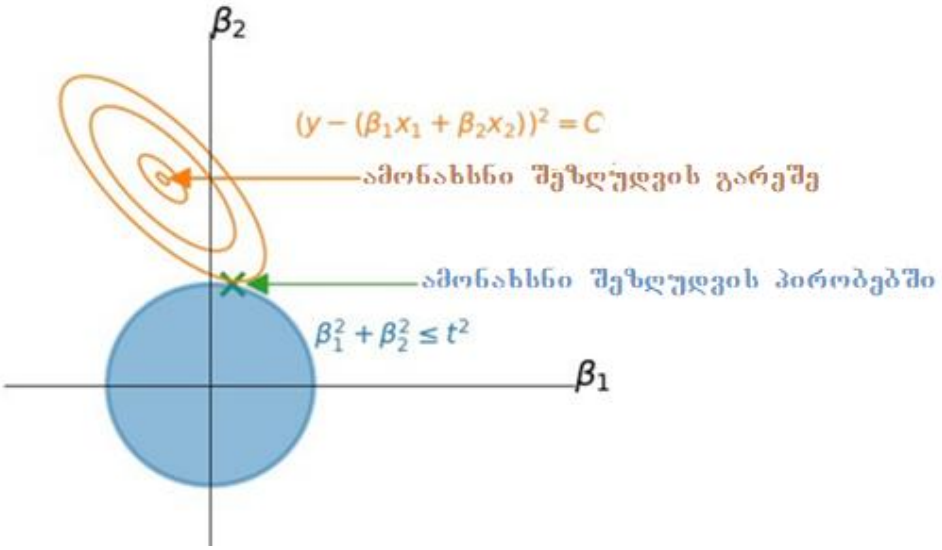
$$\arg\min_{\vec{\beta}\in\mathbb{R}^{p+1}} \left\|\left(\vec{y} - X\vec{\beta}\right)\right\|_2^2 \text{ ისეთი რომ } \left\|\vec{\beta}\right\|_2^2 \le t. \qquad 6.9)$$

⚙

მტკიცებულება. ეკვივალენტობა მიიღწევა დუალობით და არაწრფივი დაპროგრამების ამოცანის ამოსახსნელად აუცილებელი კარუშ-კუშ-ტაკერის (ინგლ. Karush — Kuhn — Tucker) პირობების ჩაწერით.

□

ამრიგად, თხემური რეგრესია შეიძლება იყოს ჩამოყალიბებული როგორც ოპტიმიზაციის ($\left\|\left(\vec{y} - X\vec{\beta}\right)\right\|_2^2$-ის მინიმიზაციის) კვადრატული ამოცანა $\left(\left\|\vec{\beta}\right\|_2^2 \le t\right)$ შეზღუდვებით : ამონახსნი უნდა იყოს მოთავსებული $\sqrt{t}$ რადიუსის ℓ_2 სფეროში. გარდა იმ შემთხვევისა, როცა ოპტიმიზაცია შეზღუდვათა გარეშე უკვე აკმაყოფილებს ამ პირობას, ამონახსნი განთავსებული იქნება ამ სფეროს საზღვარზე, როგორც ეს ნაჩვენებია ნახატზე 6.2.



ნახატი 6.2 - შეზღუდვებიანი ოპტიმიზაციის (აქ ორ განზომილებაში) 6.9 ამოცანის ამონახსნი : განთავსებულია $\sqrt{t}$ რადიუსის ℓ_2 სფეროს მხებ უმცირეს კვადრატთა ჯამის დონის წირზე.

3 LASSO

3.1 პარსიმონია, ანუ რაციონალური მომჭირნეობა და ყაირათიანობა

პარსიმონია (რაციონალური მომჭირნეობა და ყაირათიანობა) — მეცნიერების ზოგადი პრინციპია, რომელიც ამბობს, რომ სხვა თანაბარ პირობებში, მკვლევარმა უნდა ამჯობინოს გარკვეული ფენომენის უმარტივესი შესაძლო ახსნა, ან გარკვეული პრობლემის უმარტივესი შესაძლო გადაწყვეტა.

ზოგიერთ შემთხვევაში გონივრული შეიძლება იყოს დაშვება, რომ ჭდე, რომლის წინასწარმეტყველებას ვცდილობთ, დასაშვებია, ნაპოვნი აღმოჩნდეს ცვლადების მხოლოდ უმნიშვნელო რიცხვით. ასეთ შემთხვევაში სასურველია პარსიმონიალური, ანუ *მწირი, მეჩხერი, გამოხშირული, გაიშვიათებული* (ინგლ. *sparse*) მოდელის აგება, ე.ი. ისეთის, რომელშიც კოეფიციენტების გარკვეული რაოდენობა ნულის ტოლია : მაშინ შესაბამისი ცვლადების ამოღება გახდება შესაძლებელი მოდელიდან.

ამ მიზნის მისაღწევად 1966 წელს კანადური წარმომავლობის ამერიკელმა სტატისტიკოსმა, სტენფორდის უნივერსიტეტის სტატისტიკისა და ბიოსამედიცინო მონაცემთა კათედრების პროფესორმა რობ ტიბშირანიმ (Robert Tibshirani, 1956-) წამოაყენა რეგულარიზატორად $\vec{\beta}$ კოეფიციენტის ℓ_1 ნორმის გამოყენების ძალიან ნაყოფიერი იდეა :

$$\Omega_{\text{lasso}}(\vec{\beta}) = \left\|\vec{\beta}\right\|_1 = \sum_{j=0}^{p} \left|\beta_j\right|. \tag{6.10}$$

პასუხი ბუნებრივ შეკითხვაზე, რატომ იძლევა ეს რეგულარიზატორი საშუალებას «დავძრათ» ზოგიერთი კოეფიციენტი ნულისკენ, ადვილად მოიპოვება ნახატზე 6.4 და მასთან დაკავშირებულ ქვედანაყოფში 6.3.5.

3.2 LASSO-ს ფორმულირება

განსაზღვრება 6.4 (LASSO) ღრმა სწავლების თეორიული ენისათვის სპეციალურად შემოტანილი სიტყვით LASSO აღინიშნება $f : x \mapsto \vec{\beta}^{\mathrm{T}}\vec{x}$, სახის მოდელი, რომლის კოეფიციენტები შემდეგი გზით მიიღება :

$$\arg\min_{\vec{\beta}\in\mathbb{R}^{p+1}} \left\|\left(\vec{y} - X\vec{\beta}\right)\right\|_2^2 + \lambda\left\|\vec{\beta}\right\|_1. \tag{6.11}$$

ტერმინი LASSO — ეს აკრონიმია ინგლისური ფრაზიდან *Least Absolute Shrinkage and Selection Operator* (*უმცირესი აბსოლუტური შემცირების და არჩევის ოპერატორი*) და, ჟღერადობის გარდა, მას არაფერი აქვს საერთო *ლასოსთან*, რომლითაც ცხენოსანი მწყემსები და მონადირეები ცხოველებს იჭერენ ამერიკაში და რომლის მნიშვნელობა ქართულად გადმოიცემა სიტყვებით *მარყუჟი, ყულფი, ქამანდი* ან *საგდებელი*. LASSO — ეს მეთოდია, რომელიც იყენებს კოეფიციენტების აბსოლუტურ მნიშვნელობებს (ნორმა ℓ_1) ამ კოეფიციენტთა რიცხვის შესამცირებლად (ინგლ. *shrink*), რითაც ხდება ისეთი *ცვლადების არჩევა*, რომლებსაც ნულოვანი კოეფიციენტები არ ექნება. სიგნალების დამუშავების თეორიაში LASSO ასევე ცნობილია, მაგრამ როგორც *ძირითადი დადევნება/გამოდგომა* (ინგლ. *basis pursuit*).

შენიშვნა

ვინაიდან LASSO ქმნის პარსიმონიულ, გამოხშირულ, გამარტივებულ მოდელს და გამორიცხავს ცვლადებს ნულოვანი კოეფიციენტებით, იგი კონტროლირებადი ცვლადების შერჩევის ხერხია. ამიტომ იგი იმავდროულად წარმოადგენს განზომილების შემცირების მეთოდსაც.

3.3 ამოცანის გადაწყვეტა

LASSO-ს მოდელის კოეფიციენტთა გამოსახულებას 6.11 არ აქვს ნათელი, პირდაპირი, გარკვეულად გამოხატული, არაორაზროვანი და ზუსტი — ერთი სიტყვით, *ექსპლიციტური* — ამონახსნი. მის გადასაწყვეტად შეიძლება იყოს გამოყენებული მიმართული დაშვების ალგორითმი (იხ. ქვედანაყოფი 13.3.3). გარდა ამისა, ეს ამოცანა ყოველთვის მკაცრად ამოზნექილი არ არის (კერძოდ მაშინ, როცა $p > n$) და ამიტომ აუცილებელიც არ არის ჰქონდეს ერთადერთი ამონახსნი. პრაქტიკაში ეს განსაკუთრებით პრობლემატურია, როცა შეუძლებელია ცვლადების განხილვა ალბათობათა უწყვეტი განაწილებების რეალიზაციებად. მიუხედავად ამისა, შეიძლება იმის დამტკიცება, რომ არანულოვან კოეფიციენტებს ორ ამონახსნში აუცილებლად ერთნაირი ნიშნები აქვს. ამრიგად, ცვლადის გავლენა ერთნაირი მიმართულებისაა ყველა განსახილველ ამონახსნში, რაც აადვილებს LASSO-ს საშუალებით აგებული მოდელის ინტერპრეტაციას.

3.4 ალბათური მიდგომა

უპირველეს ყოვლისა, დავუშვათ, რომ შემთხვევით ცვლადს აქვს გაუსის (ნორმალური) ალბათობათა განაწილების $\mathcal{N}\left(l,\sigma^2\right)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{x-l}{\sigma}\right)^2\right)$ სიმკვრივე, სადაც l *ადგილმდებარეობის* (*location*) ნამდვილი პარამეტრია ($l\in\mathbb{R}$), ხოლო $\sigma^2\in\mathbb{R}_{>0}$ საშუალო კვადრატული გადახრის კვადრატი *მასშტაბის* (*scale*) ნამდვილ პარამეტრს წარმოადგენს. განმარტებათა საფუძველზე ადვილად მტკიცდება, რომ ასეთნაირად განაწილებული შემთხვევითი ცვლადის საშუალო მნიშვნელობა (მათემატიკური ლოდინი) არის l, ხოლო დისპერსია σ^2-ს შეადგენს. ამიტომ, თუ მოვითხოვთ, რომ მათემატიკური ლოდინი უნდა უდრიდეს ნულს, მაშინ $l=0$ ტოლობა გვექნება. თუ ამასთან ერთად მოვითხოვთ, რომ დისპერსია უნდა იყოს $1-$ის ტოლი, მაშინ $\sigma^2=1$ ტოლობას ექნება ადგილი. ასე რომ, გაუსის ალბათობათა განაწილების სიმკვრივე ნულოვანი მათემატიკური ლოდინითა და ერთეულოვანი დისპერსიით (ე.წ. სტანდარტული ნორმალური განაწილების ალბათობათა სიმკვრივის ფუნქცია) შეიძლება $\mathcal{N}\left(l,\sigma^2\right)=\mathcal{N}\left(0,1\right)$ ნოტაციით იყოს ჩაწერილი.

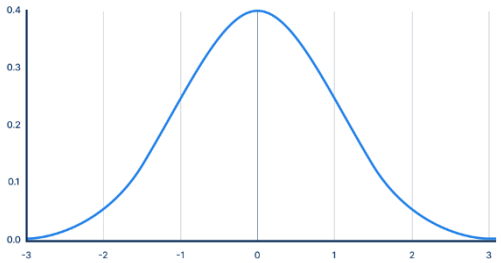
ახლა დავუშვათ, რომ შემთხვევით ცვლადს აქვს ლაპლასის ალბათობათა განაწილების $\mathrm{Lap}\left(l,s\right)=\frac{1}{2s}\exp\left(-\frac{\left|x-l\right|}{s}\right)$ სიმკვრივე, სადაც l *ადგილმდებარეობის* (location) ნამდვილი პარამეტრია, ხოლო $s>0$ *მასშტაბის* (scale) ასევე ნამდვილ პარამეტრს წარმოადგენს. განმარტებათა საფუძველზე ადვილად მტკიცდება, რომ ასეთნაირად განაწილებული შემთხვევითი ცვლადის საშუალო მნიშვნელობა (მათემატიკური ლოდინი) $l-$ის ტოლია, ხოლო მისი დისპერსია $2s^2-$ს შეადგენს. თუ მოვითხოვთ, რომ მათემატიკური ლოდინი უნდა უდრიდეს ნულს, მაშინ $l=0$ ტოლობა გვექნება. თუ ამასთან ერთად მოვითხოვთ, რომ დისპერსია უნდა იყოს $1-$ის ტოლი, მაშინ $2s^2=1$ განტოლებიდან $s=1/\sqrt{2}\approx 0.71$ შედეგი მიიღება. ასე რომ,

ლაპლასის ალბათობათა განაწილების სიმკვრივე ნულოვანი მათემატიკური ლოდინითა და ერთეულოვანი დისპერსიით შეიძლება $\mathrm{Lap}(l,s)=\mathrm{Lap}\left(0,1/\sqrt{2}\right)$ ნოტაციით იყოს ჩაწერილი.

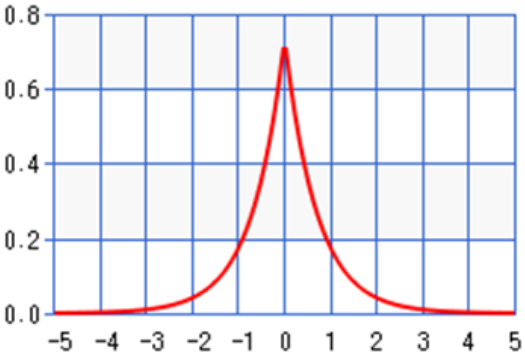
ℓ_1 რეგულარიზაცია შეესაბამება შეფასებას მაქსიმალური აპოსტერიორული მეთოდის გამოყენებით იმ ჰიპოთეზის პირობებში, რომ $\vec{\beta}^*$ -ის ამოღება ხდება B შემთხვევითი ვექტორიდან, რომელიც ალბათობათა გაუსის განაწილებას არ ემორჩილება, როგორც ℓ_2 რეგულარიზაციის დროს, არამედ აქვს ლაპლასის განაწილება ცენტრით 0 წერტილზე და ალბათობათა განაწილების სიმკვრივის შემდეგი ფუნქციით :

$$p_B\left(\vec{b}\right)=\prod_{j=1}^{p+1}\frac{1}{2\sigma}\exp\left(-\frac{1}{\sigma}\left|b_j\right|\right).$$

როგორც ნაჩვენებია ნახატზე 6.3, ლაპლასის კანონი ანიჭებს მეტ წონას თავის მათემატიკურ ლოდინს (ამ შემთხვევაში ნულს), ვიდრე გაუსის განაწილება ასეთივე დისპერსიით, რაც ამოწახსნის მიღების პარსიმონიის (კომპაქტურობის) ახსნასაც იძლევა.



a) გაუსის ალბათობათა განაწილების $\mathcal{N}(0,1)$ სიმკვრივე ნულოვანი მათემატიკური ლოდინითა და ერთეულოვანი დისპერსიით.



b) ლაპლასის ალბათობათა განაწილების $\mathrm{Lap}\left(0,1/\sqrt{2}\right)\approx\mathrm{Lap}(0,0.7)$ სიმკვრივე ნულოვანი მათემატიკური ლოდინითა და ერთეულოვანი დისპერსიით.

ნახატი 6.3 - გაუსის ალბათობათა განაწილების და ლაპლასის ალბათობათა განაწილების სიმკვრივეები ნულოვანი მათემატიკური ლოდინითა და ერთეულოვანი დისპერსიით.

ასე რომ, ემპირიული რისკის აპოსტერიორული მაქსიმუმით და ემპირიული რისკის მინიმიზაციით მიღებულ შეფასებათა შორის ეკვივალენტობა ზუსტად ისევე მიიღწევა, როგორც ℓ_2 რეგულარიზაციისთვის.

3.5 გეომეტრიული ინტერპრეტაცია

წინა შემთხვევის მსგავსად, ამოცანა 6.11 შეიძლება იყოს ჩამოყალიბებული ხელახლა შეზღუდვებიანი კვადრატული ოპტიმიზაციის პრობლემის სახით :

თეორემა 6.2 *თუ მოცემულია* $\lambda \in \mathbb{R}_+$, $X \in \mathbb{R}^{n\times p}$ *და* $\vec{y} \in \mathbb{R}^n$, *მაშინ არსებობს ერთადერთი* $t \in \mathbb{R}_+$ *მნიშვნელობა, ისეთი, რომ ამოცანა 6.11 ეკვივალენტურია შემდეგი პრობლემის :*

$$\arg\min_{\vec{\beta}\in\mathbb{R}^{p+1}} \left\| \left(\vec{y} - X\vec{\beta}\right) \right\|_2^2 \quad \text{ისეთი რომ} \quad \left\|\vec{\beta}\right\|_1 \le t\,. \tag{6.12}$$

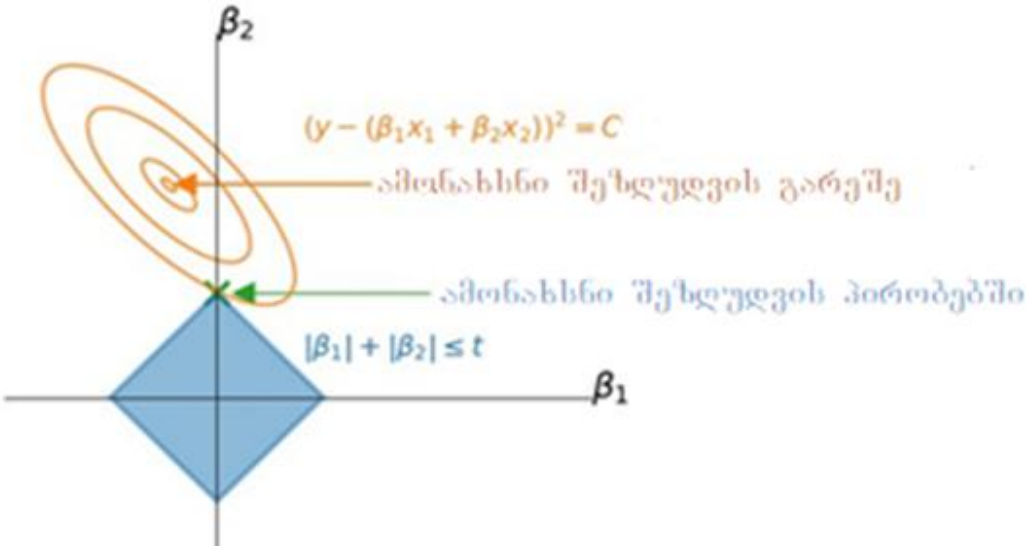
ახლა ამონახსნი უნდა იმყოფებოდეს t რადიუსის ℓ_1 სფეროს საზღვრებში. ვინაიდან ამ სფეროს აქვს «კუთხეები», ამიტომ კვადრატული ფორმის დონის წირები დიდი ალბათობით აღმოჩნდება მისი მხები წერტილზე, სადაც ერთი ან რამდენიმე კოორდინატი ნულის ტოლია (იხ. ნახატი 6.4).

3.6 რეგულარიზაციის გზა

LASSO-ს რეგულარიზაციის გზაზე (მაგალითად, ნახატი 6.5 იმავე მონაცემებზე, რომლებიც გამოყენებულია ნახატზე 6.1) ჩანს, რომ ცვლადები ერთიმეორის მიყოლებით ტოვებს მოდელს, ვიდრე ყველა კოეფიციენტი არ გახდება ნულის ტოლი. უნდა აღვნიშნოთ, რომ რეგულარიზაციის გზა ნებისმიერი ცვლადისათვის უბან-უბან წრფივია ; ეს LASSO-ს თვისებაა.

შენიშვნა

თუ ჭდის წინასწარმეტყველებაში წვლილი რამდენიმე კორელირებულ ცვლადს შეაქვს. მაშინ LASSO ეცდება მათ შორის მხოლოდ ერთი აირჩიოს, ყველა დანარჩენს ნულის ტოლი წონა მიანიჭოს და არ დაიწყოს წონების თანაბრად განაწილება, როგორც ეს თხემურ რეგრესიაში ხდება. ამრიგად, ძალზე პარსიმონიული (ეკონომიური, გაღარიბებული, გამარტივებული) მოდელები მიიღება. მაგრამ ამ ცვლადის არჩევა არის შემთხვევითი და შეიძლება იცვლებოდეს ოპტიმიზაციის პროცედურის განმეორებისას. ამიტომ LASSO გამორჩეულად მიდრეკილია არამდგრადობისკენ.



ნახატი 6.4 - შეზღუდვებიანი ოპტიმიზაციის (აქ ორ განზომილებაში) 6.12 ამოცანის ამონახსნი : განთავსებულია t რადიუსის ℓ_1 სფეროს მხებ უმცირეს კვადრატთა ჯამის დონის წირზე.

4 ელასტიკური ქსელი

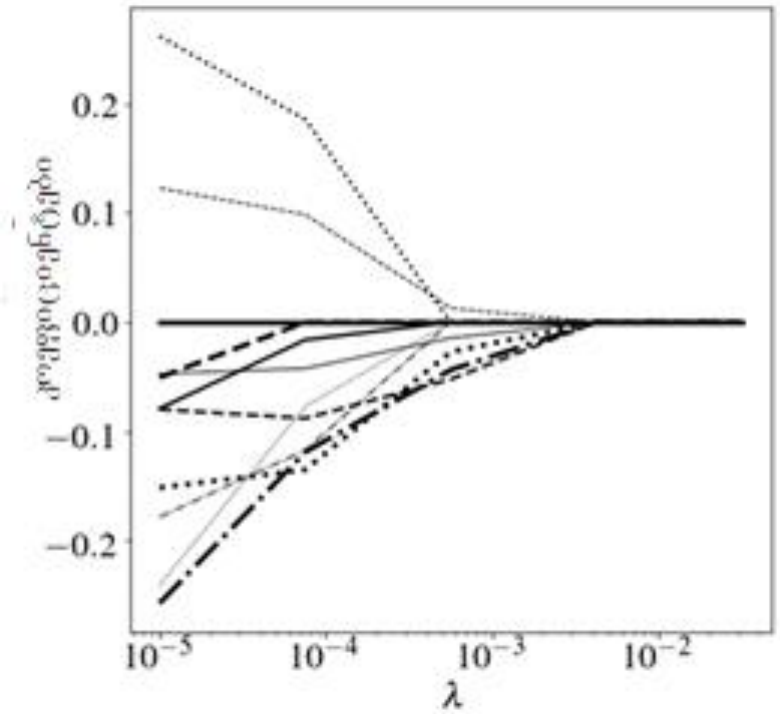
ℓ_1 ნორმით რეგულარიზაცია გამარტივებულ, პარსიმონიულ მოდელს იძლევა, რომლის ინტერპრეტირება გაადვილებულია, ხოლო ℓ_2 ნორმით რეგულარიზაცია თავიდან გვაცილებს ჭარბ სწავლებასა და კორელირებული ცვლადების დაჯგუფებას. ამასთან დაკავშირებით სავსებით ბუნებრივია ამ ორი ნორმის გაერთიანების სურვილი შემდეგ რეგულარიზატორში :

$$\Omega_{\text{enet}}\left(\vec{\beta}\right)=\left((1-\alpha)\left\|\vec{\beta}\right\|_1+\alpha\left\|\vec{\beta}\right\|_2^2\right). \tag{6.13}$$

ეს რეგულარიზატორი პარამეტრიზებულია $\alpha\in[0,1]$ სიდიდით. სახელდობრ, როცა $\alpha=0$, გამოიყენება ℓ_1 რეგულარიზაცია, ხოლო როცა $\alpha=1$, მაშინ — ℓ_2 რეგულარიზაცია.

განსაზღვრება 6.5 (ელასტიკური ქსელი) *ელასტიკური ქსელი* ეწოდება $f: x\mapsto\vec{\beta}^{\mathrm{T}}\vec{x}$ მოდელს, რომლის კოეფიციენტები მიღებულია ფორმულით

$$\arg\min_{\vec{\beta}\in\mathbb{R}^{p+1}}\left\|\left(\vec{y}-X\vec{\beta}\right)\right\|_2^2+\lambda\left((1-\alpha)\left\|\vec{\beta}\right\|_1+\alpha\left\|\vec{\beta}\right\|_2^2\right). \tag{6.14}$$



ნახატი 6.5 – LASSO-ს რეგულარიზაციის გზა მონაცემთა ნაკრებისათვის 12 ცვლადით. ყოველი წირი წარმოადგენს ერთერთი ამ ცვლადის რეგრესიის კოეფიციენტის ცვლილებას რეგულარიზაციის λ კოეფიციენტის ზრდისას : ცვლადების გამორიცხვა ერთიმეორის მიყოლებით ხდება.

ამონახსნი ელასტიკური ქსელისათვის მიღებულია მიმართული დაშვების ალგორითმით. ეს ამონახსნი ეკონომიურია, მაგრამ უფრო ნაკლებად, ვიდრე LASSO. მართლაც, რამდენიმე ძლიერ კორელირებული ცვლადის არსებობისას LASSO აირჩევს მხოლოდ ერთ-ერთს მათ შორის, მაშინ როცა ℓ_2 ნორმის ეფექტის გამო ელასტიკური ქსელი აირჩევს ყველა ცვლადს და მიანიჭებს მათ ერთნაირ კოეფიციენტებს.

5 საკვანძო მომენტები

— კოეფიციენტების $\vec{\beta}$ ვექტორის ფუნქციით წარმოდგენილი რეგულარიზაციის წევრის, დამატება წრფივი რეგრესიის ემპირიული რისკისადმი, ჭარბი სწავლების თავიდან აცილების საშუალებას იძლევა.

— თხემური რეგრესია რეგულარიზატორად $\vec{\beta}$ ვექტორის ℓ_2 ნორმას იყენებს ; იგი ყოველთვის ერთადერთ ანალიზურ ამონახსნს იძლევა და დამაჯგუფებელ გავლენას ახდენს კორელირებულ ცვლადებზე.

— LASSO $\vec{\beta}$ ვექტორის ℓ_1 ნორმას ხმარობს რეგულარიზატორად ; იგი ქმნის პარსიმონიულ, გამარტივებულ მოდელს და, ამრიგად, შეიძლება იყოს გამოყენებული განზომილების კონტროლირებადი შემცირებისათვის.

— ამოცანის სტრუქტურის შესაბამისად შესაძლებელია მრავალი სხვა რეგულარიზატორიც.

დამატებითი ინფორმაცია

- ℓ_1 და ℓ_2 ნორმებთან ერთად დამატებით შესაძლებელია $\Omega_{\ell_q}\left(\vec{\beta}\right)=\left\|\vec{\beta}\right\|_q^q$ სახის რეგულარიზატორების გამოყენებაც.

- რეგულარიზატორების ოჯახი, რომლებიც ცნობილია როგორც «სტრუქტურირებული» რეგულარიზატორები, შეიძლება იყოს გამოყენებული ისეთო ცვლადების ამოსარჩევად, რომლებიც ინარჩუნებს აპრიორი მოცემულ სტრუქტურას (გრაფის, ჯგუფის ან ხის). ეს მიდგომები გამოიყენება, კერძოდ, ბიოინფორმატიკის მიმართულებებში, მაგალითად მაშინ, როცა გვინდა მარტივი მოდელების აგება გენების ექსპრესიის (კოდირებული გენეტიკური ინფორმაციის გარდასახვის) საფუძველზე იმ დაშვებებით, რომ მეტაბოლური გზების (გენების ჯგუფთა) მხოლოდ მცირე რაოდენობაა მნიშვნელოვანი. ამ საკითხთან დაკავშირებით უფრო დაწვრილებითი ინფორმაციის მიღება შეიძლება სტატიიდან Huang et al. (2011).

- რაც შეეხება LASSO-ს მიერ გამოტანილი გადაწყვეტილების ერთადერთობის საკითხს, ამას შეიძლება გავეცნოთ რობ ტიბშირანის შვილის ნაშრომში Ryan J. Tibshirani (2013).

- წიგნი Hastie et al (2015) მთლიანად ეძღვნება LASSO-ს და მის ყველა განზოგადებას. აქ მოცემული ინფორმაცია პრაქტიკულად ამომწურავია და უდიდეს ინტერესს წარმოადგენს.

6 ბიბლიოგრაფია

1. Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical Learning with Sparsity : The Lasso and Generalizations*. CRC Press. http://web.stanford.edu/~hastie/StatLearnSparsity/ .

2. Huang, J., Zhang, T., and Metaxas, D. (2011). Learning with structured sparsity. *Journal of Machine Learning Research*, 12 : 3371–3412.

3. Tibshirani, R. J. (2013). The lasso problem and uniqueness. *Electronic Journal of Statistics*, 7 :1456–1490.

7 სავარჯიშოები

6.1 როდის არის რეგულარიზებული რეგრესიის მოდელი უფრო მგრძნობიარე, მიდრეკილი გადამეტებული სწავლებისადმი, როცა რეგულარიზაციის პარამეტრი დიდია, თუ მცირე ?

6.2 როდის არის თხემურ რეგრესიაში (ინგლ. ridge regression) წონები უფრო მნიშვნელოვანი, როცა რეგულარიზაციის პარამეტრი დიდია, თუ მცირე ?

6.3 ანიკას აქვს მონაცემთა ნაკრები, რომელიც შეიცავს n დაკვირვებას $\mathbb{R}^p$ – ში. ეს მონაცემები წარმოდგენილია $X\in\mathbb{R}^{n\times p}$ – ით და მარკირებულია (მონიშნულია) $\vec{y}\in\mathbb{R}^n$ ვექტორით. ყოველი ცვლადი მისი p მდგენელიდან შეესაბამება გარკვეულ ქალაქს. ამ ცვლადებზე ანიკას მიერ შექმნილია გრაფი : თითოეული კვანძი შეესაბამება ერთ ქალაქს და არსებობს ერთი წიბო ორ კვანძს შორის, თუ მათი შესაბამისი ორი ქალაქი უახლოესია. ანიკა აღნიშნავს ε სიმბოლოთი

თავისი გრაფის წიბოების სიმრავლეს. დაბოლოს, იგი შემდეგი მოდელის წვრთნას (სწავლებას) ახორციელებს : $\arg\min_{\vec{\beta}\in\mathbb{R}^{p+1}} \left\|\vec{y} - X\vec{\beta}\right\|_2^2 + \lambda \sum_{(u,v)\in\varepsilon} \left(\beta_u - \beta_v\right)^2$. რატომ ?

6.4 ბასამ განახორციელა წრფივი რეგულარიზებული რეგრესიის სწავლება (წვრთნა) თავის მონაცემებზე. მან შენიშნა, რომ პროგნოზის (წინასწარმეტყველების) შეცდომა დიდია როგორც სწავლების (წვრთნის) ანარჩევზე, ასევე შემოწმების ანარჩევზეც. რა უნდა გაკეთდეს ?

6.5 თუ არის რეკომენდებული რეგულარიზაციის გამოყენება, როცა ზოგიერთი ცვლადი კორელაციაშია ერთმანეთთან ?

6.6 კატომ გაწვრთნა წრფივი რეგრესია თავის მონაცემებზე და კმაყოფილი არ არის მუშაობის იმ მახასიათებლებით, რომლებსაც გამოყენებული საწვრთნელი სიმრავლით მიაღწია. მან შენიშნა, რომ ზოგიერთი მისი დაკვირვება ჭარბია (ზედმეტია). თუ მიეცემა კატოს საშუალება გააუმჯობესოს ეს მახასიათებლები რეგულარიზაციის გამოყენებით ?

6.7 თუ შეიძლება ℓ_1 რეგულარიზაციის გამოყენება პოლინომური რეგრესიის ამოცანაში ?

6.8 რისთვის ხდება Elastic net რეგრესიის გამოყენება Lasso-ს ნაცვლად ?

სავარჯიშოთა ამონახსნები

6.1 როცა λ მცირეა, ემპირიული რისკი დომინირებს, ჭარბობს და მოდელი, უფრო მეტად მოსალოდნელია, განაგრძობს ზედმეტად სწავლებას (წვრთნას).

6.2 როცა λ დიდია, მაშინ სწორედ რეგულარიზაციის წევრია ის ფაქტორი, რომელიც დომინირებს და წონები იძულებულია იყოს მცირე სიდიდის.

6.3 გრაფზე ერთმანეთთან დაკავშირებულ ცვლადებს ექნება ნაკლებად განსხვავებული კოეფიციენტები : ანიკა თვლის, რომ ორ ახლომდებარე ქალაქს შეაქვს თითქმის ერთნაირი წვლილი ჭდეში, რომლის მოდელირებას იგი ცდილობს.

6.4 საქმე გვაქვს არასაკმარისად ნასწავლ (არასაკმარისად გაწვრთნილ) მოდელთან : იგი ზედმეტად რეგულარიზებულია, აუცილებელია λ პარამეტრის მნიშვნელობის შემცირება.

6.5 დიახ : თხემური რეგრესია ამოხსნის სტაბილიზაციის საშუალებას შექმნის.

6.6 ნაკლებად მოსალოდნელია : არასაკმარისად ნასწავლი მოდელი და ჭარბი (ზედმეტი) დაკვირვებები გავლენას არ ახდენს ამონახსნზე.

6.7 დიახ (და ნებისმიერ რეგრესიაში) : საქმე მხოლოდ ცვლადების დამატებას ეხება.

6.8 უფრო სტაბილური ამონახსნის მისაღწევად.

ლექცია 7 ხელოვნური ნეირონული ქსელები

შინაარსი

გამოსახულებათა ავტომატური ანოტირებიდან ხმის ამოცნობამდე, *გოს* თამაშში ჩემპიონთა დამარცხების შემძლე კომპიუტერებიდან ავტონომიურ სატრანსპორტო საშუალებებამდე და საერთოდ ხელოვნური ინტელექტის სფეროში მრავალი უკანასკნელი წარმატება ეფუძნება ღრმა ნეირონულ ქსელებს, ხოლო *სიღრმისეული სწავლება* (ინგლ. *deep learning*) მთავარ თემად არის დღეს ქცეული საუბრებში (თუ პოლიტიკურ პაექრობას არ მივიღებთ მხედველობაში).

ხელოვნური ნეირონული ქსელები — ეს, ფაქტობრივად, სხვა არაფერია თუ არა პარამეტრული, პოტენციურად რთული, მოდელები : წრფივი რეგრესიისგან განსხვავებით ისინი ძალიან მოქნილი მოდელების ადვილად აგების საშუალებას იძლევა.

ამ ლექციაში ჩვენ განვიხილავთ ხელოვნური ნეირონული ქსელების კლასიდან ერთ-ერთის — მრავალშრიანი პერსეპტრონების — მუშაობის ძირითად პრინციპებს და მათ სწავლებას, წვრთნას. მოკლედ შევეხებით ღრმა (სიღრმისეული) ნეირონული ქსელების თემას, რომელიც, ჩვენი შეხედვით, უფრო რთულია და მოცემულ კურსში არ შედის. როგორც წესი, უნივერსიტეტებში იგი იკითხება როგორც ღრმა, ანუ სიღრმისეული სწავლების ცალკე კურსი. სწორედ ასეა მიღებული, მაგალითად, საქართველოს ტექნიკურ უნივერსიტეტშიც.

მიზნები

— ერთშრიანი ან მრავალშრიანი პერსეპტრონის შესაბამისი გადაწყვეტილების მიღების ფუნქციის ჩაწერა.

— პერსეპტრონის სწავლების პროცედურის რეალიზება.

— უკუგავრცელების პროცედურაში მრავალშრიანი პერსეპტრონის განახლების ახსნა.

— მრავალშრიანი პერსეპტრონის აგებისა და სწავლების ზოგიერთი ძირითადი პრინციპის გაგება.

— ღრმა სწავლების ამოცანათა გაგება.

1 პერსეპტრონი

ხელოვნური ნეირონული ქსელების ისტორია სათავეს იღებს 1950-ანი წლების დასაწყისში და ფსიქოლოგთა მცდელობებში — ისეთის, როგორიცაა ამერიკელი მეცნიერი ფრენკ როზენბლატი (Frank Rosenblatt, 1928-1971) — გაეგოთ ადამიანის ტვინის ფუნქციონირება. თავდაპირველად ისინი იქმნებოდა ძუძუმწოვრების თავის ტვინის ქერქში განთავსებული ბიოლოგიური ნეირონული ქსელებით ინფორმაციის დამუშავების მათემატიკური მოდელირებისათვის.

დღეს მათ სიახლოვეს ბიოლოგიურ რეალობასთან დიდი მნიშვნელობა არ აქვს და სწორედ მათი ეფექტურობა რთული არაწრფივი კავშირების მოდელირებაში გახდა ხელოვნური ნეირონული ქსელების წარმატების მიზეზი.

პირველი ხელოვნური ნეირონული ქსელი იყო ფრენკ როზენბლატის *პერსეპტრონი* (Rosenblatt, 1957).

ეს ქსელი არ იყო ღრმა, ჰქონდა ერთი შრე და მოდელირების შეზღუდული შესაძლებლობები.

1.1 მოდელი

პერსეპტრონი (ნახატი. 7.1) შედგენილია ერთი შემავალი შრით, სადაც p ნეირონი, ანუ *მოდული* (დასაშვებია ვიხმაროთ *ბლოკი*) არის წარმოდგენილი და თითოეული მათ შორის შეესაბამება ერთ შემავალ ცვლადს.

ყოველი ნეირონი გადასცემს თავისი შესასვლელის მნიშვნელობას მომდევნო შრეს. ამ p ნეირონს, ჩვეულებრივ, ემატება წანაცვლების მოდული (ბლოკი), რომელიც ყოველთვის გადასცემს 1-ის ტოლ მნიშვნელობას.

ეს მოდული შეესაბამება 1-ის იმ სვეტს, რომელიც ჩვენ დავამატეთ მონაცემებს წრფივ მოდელებში (განტოლება 5.5).მომავალში ნებისმიერი $\vec{x} = \left(x_1, x_2, \ldots, x_p\right)$ ვექტორი ჩანაცვლებული იქნება 1-ით გაზრდილი თავისივე ვერსიით :

$$\vec{x} = \left(1, x_1, x_2, \ldots, x_p\right).$$

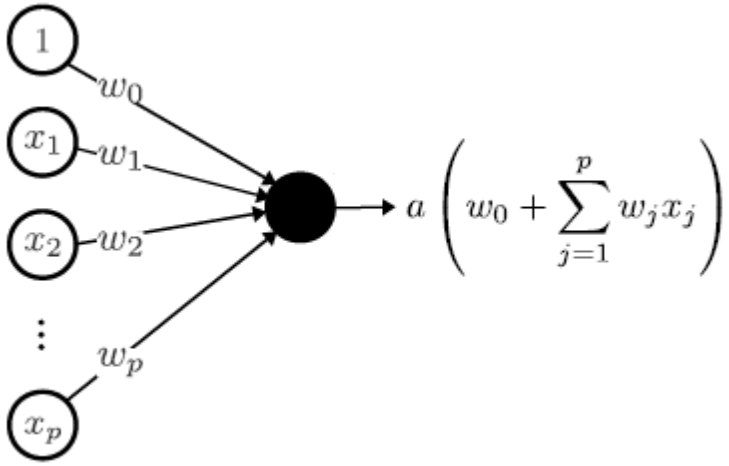
პერსეპტრონის პირველი და ერთადერთი შრე (შემავალი შრის შემდეგ) შეიცავს ერთ ნეირონს, რომელთანაც მიერთებულია შემავალი ფენის ყველა მოდული (ბლოკი).

ეს ნეირონი ითვლის $x_1, x_2, \ldots, x_p$ სიგნალების წრფივ $o\left(\vec{x}\right) = w_0 + \sum_{j=1}^{p} w_j x_j$ კომბინაციას. ამ სიგნალებს ნეირონი იღებს შესასვლელზე და მათ მიმართ a *აქტივაციის ფუნქციას* იყენებს. შედეგი მის მიერ გადაეცემა გამოსასვლელს. ეს გამოსასვლელი კი ახორციელებს პერსეპტრონის გადაწყვეტილების მიღების ფუნქციას.

ამრიგად, თუ w_j სიდიდეს ვუწოდებთ შემავალ j ბლოკსა და გამომავალ ნეირონს შორის კავშირის (ბმის) *წონას*, მაშინ ეს ნეირონი შემდეგ გამოსახულებას ანგარიშობს :

$$f\left(\vec{x}\right) = a\left(o\left(\vec{x}\right)\right) = a\left(w_0 + \sum_{j=1}^{p} w_j x_j\right) = a\left(\left\langle \vec{w}, \vec{x} \right\rangle\right). \tag{7.1}$$

მაშასადამე, საქმე ეხება *პარამეტრულ მოდელს.*



ნახატი 7.1 - პერსეპტრონის არქიტექტურა.

აქტივაციის ფუნქციები

რეგრესიის ამოცანის შემთხვევაში აქტივაციის ფუნქციად საკმარისია *იდენტურობის* (*იგივეობის*) *ფუნქციის* გამოყენება.

ბინარული კლასიფიკაციის შემთხვევაში შესაძლებელი იქნება სხვა მიდგომათა გამოყენება, სახელდობრ :

— ბინარული ჭდის პირდაპირი წინასწარმეტყველებისათვის გამოიყენება ზღურბლური ფუნქცია :

$$f:\vec{x}\mapsto\begin{cases}0 & \text{თუ } o(\vec{x})\le 0\\ 1 & \text{წინააღმდეგ შემთხვევაში}\end{cases}. \tag{7.2}$$

— დადებითი კლასისადმი კუთვნილების ალბათობის წინასწარმეტყველებისათვის, ისევე, როგორც ლოგისტიკური რეგრესიის შემთხვევაში (იხ. ქვედანაყოფი 5.3), ლოგისტიკური ფუნქცია გამოიყენება :

$$f:\vec{x}\mapsto\frac{1}{1+e^{o(\vec{x})}}=\frac{1}{1+\exp\left(\langle\vec{w},\vec{x}\rangle\right)}. \tag{7.3}$$

მრავალკლასიანი კლასიფიკაცია

მრავალკლასიანი კლასიფიკაციის ამოცანის შემთხვევაში ჩვენ ვცვლით პერსეპტრონის არქიტექტურას ისეთნაირად, რომ გამომავალ შრეში იყოს არა ერთი (1) ნეირონი, არამედ C ნეირონი, სადაც C — კლასების რაოდენობაა. მაშინ გამომავალი შრის $p+1$ ნეირონი დაკავშირებული იქნება თითოეულთან ამ გამომავალი ნეირონებიდან (შესაბამისად, გვექნება ბმის, კავშირის $(p+1)C$ წონა w_j^C აღნიშვნით).

ეს არქიტექტურა ნაჩვენებია ნახატზე. 7.2.

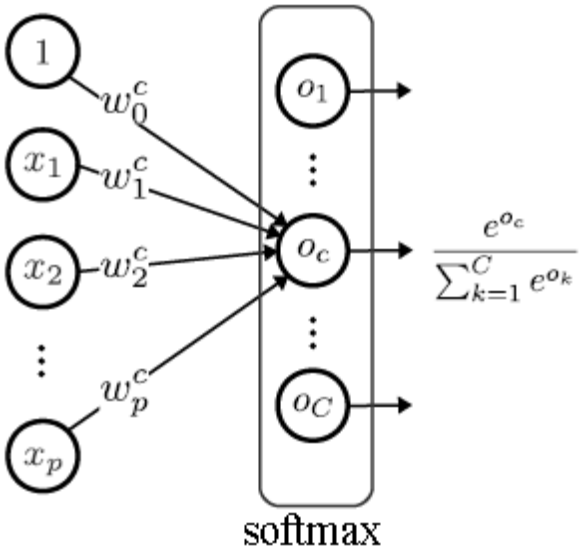
ამის შემდეგ აქტივაციის ფუნქციად გამოიყენება softmax ფუნქცია.

განსაზღვრება 7.1 (softmax ფუნქცია) softmax ფუნქცია, ანუ ნორმირებული ექსპონენციალური ფუნქცია, ეწოდება $\sigma : \mathbb{R}^C \to [0,1]^C$ ფუნქციას, რომელიც განისაზვრება შემდეგი გამოსახულებით :

$$\sigma(o_1, o_2, \ldots, o_C) = \frac{e^{o_C}}{\sum_{k=1}^{C} e^{o_k}}.$$

კლასიკური პერსეპტრონის შემთხვევაში o_k — ეს $k-$ურ გამომავალ ნეირონზე გამოთვლილი წრფივი კომბინაციაა : $o_k = \langle \vec{w}^k, \vec{x} \rangle$.

softmax ფუნქცია განაზოგადებს ლოგისტიკურ ფუნქციას, რომელიც ბინარულ შემთხვევაში გამოიყენება. იგი იძლევა C დადებით რეალურ რიცხვს 1-ის ტოლი ჯამით და შეიძლება იყოს განხილული როგორც arg max ფუნქციის ჩვეულებრივი ვერსია : როცა $o_C > o_{k \neq C}$, მაშინ $\sigma(\vec{o})_C \approx 1$ და $\sigma(\vec{o})_{k \neq C} \approx 0$.



ნახატი 7.2 - მრავალკლასიანი პერსეპტრონის არქიტექტურა.

1.2 სწავლება

პერსეპტრონის გასაწვრთნელად, როგორც ამას, ალბათ, პარამეტრული რეგრესიის შემთხვევაშიც აკეთებენ, ჩვენ უნდა ვეცადოთ ემპირიული რისკის მინიმიზაციის განხორციელება.

მაგრამ იმავდროულად დავუშვათ, რომ $(\vec{x}^i, y^i)$ დაკვირვებებისადმი წვდომა შესაძლებელია არა ერთდროულად, არამედ თანამიმდევრობით.

ეს დაშვება განპირობებულია ბიოლოგიური ნეირონული ქსელების პლასტიურობით : ისინი მუდმივ ადაპტირებას ახორციელებს მიღებული სიგნალების შესაბამისად. ამიტომ ჩვენ გამოვიყენებთ სწავლების *ინკრემენტულ* ალგორითმს, რომელიც ადაპტირებულია ერთიმეორის

მიყოლებით შემომავალ დაკვირვებებთან.

განსაზღვრება 7.2 (ინკრემენტული და ავტონომიური სწავლება) სწავლების ალგორითმს, რომელიც n დაკვირვებათა ერთ ნაკრებზე მუშაობს, *ავტონომიური ეწოდება*. გავრცელებულია ასევე *ოფლაინური სწავლების* დასახელებაც. ინგლისურ ენაში იგი ცნობილია როგორც *batch learning — პაკეტური სწავლება*.

ამისგან განსხვავებით, სწავლების ალგორითმს, რომელიც ერთ ან რამდენიმე ოპერაციას ყოველ ახალ მიღებულ დაკვირვებაზე ასრულებს, *ინკრემენტული* ეწოდება. ასევე გვხვდება დასახელება *ონლაინური სწავლება*. ინგლისურ ენაში იგი, ძირითადად, ცნობილია *online learning* ფორმით.

ემპირიული რისკის მინიმიზაციისათვის იტერაციული მიდგომით ვისარგებლოთ გრადიენტული ალგორითმით (იხ. ქვედანაყოფი 13.3.3). ალგორითმი იწყება $w_0^{(0)}, w_1^{(0)}, \ldots, w_p^{(0)}$ კავშირების (ბმების) წონათა ვექტორის შემთხვევითი ინიციალიზაციით, მაგალითად, $\vec{w} = \vec{0}$.

შემდეგ ყოველი დაკვირვებისათვის ამ ვექტორის კორექტირება ხდება ემპირიული რისკის გრადიენტის საწინააღმდეგო მიმართულებით. ეს გრადიენტი მიუთითებს ემპირიული რისკის უდიდესი დახრილობის მიმართულებას ; ქვევით დაშვებისას ვუახლოვდებით იმ წერტილს, სადაც ეს რისკი მინიმალურია. ფორმალურად ალგორითმის ყოველ იტერაციაზე ხდება ახალი $\left(\vec{x}^i, y^i\right)$ დაკვირვების აღება და ყოველი j-თვის კავშირების (ბმათა) წონების შემდეგნაირად განახლება :

$$w_j \leftarrow w_j - \eta \frac{\partial L\left(f\left(\vec{x}^i\right), y^i\right)}{\partial w_j}. \tag{7.4}$$

შესაძლებელია (და რეკომენდებულიც, თუ მონაცემები ძალიან დიდი მოცულობის არ არის) მონაცემების მთელ ანარჩევზე იტერაციის რამდენჯერმე განმეორება. როგორც წესი, იტერაცია მანამ მეორდება, ვიდრე ალგორითმის კრებადობა მიღწეული არ აღმოჩნდება $\mathcal{E}$-მდე სიზუსტით.

სწავლების სისწრაფე

ამ ალგორითმს აქვს $\eta > 0$ ჰიპერპარამეტრი, რომელიც წარმოადგენს გრადიენტული ალგორითმის ბიჯს და ხელოვნური ნეირონული ქსელების კონტექსტში მას *სწავლების სისწრაფე* (ინგლ. *learning rate*) ეწოდება. ეს ჰიპერპარამეტრი მნიშვნელოვან როლს თამაშობს : თუ იგი ძალიან დიდია, არის იმის რისკი, რომ ალგორითმი დაიწყებს ოსცილირებას ოპტიმალური გადაწყვეტილების მიდამოში ან გადაიხაროს კიდეც მისგან. და, პირიქითაც, თუ იგი ძალიან მცირეა ალგორითმის კრებადობა ძალიან შენელებული იქნება. ამიტომ უკიდურესად მნიშვნელოვანია სწავლების სისწრაფის სწორად არჩევა.

პრაქტიკაში ხშირად გამოიყენება სწავლების ადაპტირებადი სისწრაფე : შედარებით მაღალი დასაწყისში, შემდეგ სულ უფრო დაბალი გადაწყვეტილებასთან მიახლოებისას. ასეთი მიდგომა შეიძლება შევადაროთ ანალოგიურ ალგორითმებს, რომლებიც შექმნილია გრადიენტული

ალგორითმის ზოგადი შემთხვევისათვის (მაგალითად, წრფივი ძებნა *უკან დაბრუნებით — backtracking*, იხ. ქვედანაყოფი 13.3.4).

ბინარული (ორობითი) კლასიფიკაცია

ისტორიულად მოხდა ისე, რომ პირველად განხილული იყო ბინარული კლასიფიკაციის შემთხვევა ზღურბლის გამოყენებით აქტივაციის ფუნქციის როლში. ღირებულების (დანახარჯის) მაშინ არჩეული ფუნქცია დღეს ცნობილია როგორც *პერსეპტრონის კრიტერიუმი* :

$$L\left(f\left(\vec{x}^i\right), y^i\right) = \max\left(0, -y^i o\left(\vec{x}^i\right)\right) = \max\left(0, -y^i \langle \vec{w}, \vec{x} \rangle\right). \tag{7.5}$$

ეს კრიტერიუმი ახლოს დგას *სახსრის შეცდომის ფუნქციასთან* (ინგლ. *hinge error function*). როცა შესასვლელთა წრფივ კომბინაციას სწორი ნიშანი აქვს, პერსეპტრონის კრიტერიუმი უდრის ნულს. თუკი მას არასწორი ნიშანი აქვს, მაშინ პერსეპტრონის კრიტერიუმი მით მეტია, რაც უფრო დაცილებულია წრფივი კომბინაცია ნულოვან მნიშვნელობას.

ამ კრიტერიუმის გამოყენებით, წონათა განახლების წესი 7.4 შემდეგ სახეს იძენს :

$$w_j \leftarrow \begin{cases} 0 & \text{თუ } y^i o\left(\vec{x}^i\right) > 0 \\ -y^i x_j^i & \text{წინააღმდეგ შემთხვევაში} \end{cases}.$$

ამიტომ, როცა პერსეპტრონი უშვებს წინასწარმეტყველების შეცდომას, იგი გადაადგილებს გადაწყვეტილების მიღების საზღვარს ამ შეცდომის გამოსასწორებლად.

თეორემა 7.1 (თეორემა პერსეპტრონის კრებადობის შესახებ) *დავუშვათ, რომ მოცემულია n მონიშნული* $\mathcal{D} = \left\{\left(\vec{x}^i, y^i\right)_{i=1,\ldots,n}\right\}$ *დაკვირვება და* $\mathcal{D}, \gamma \in \mathbb{R}_+^*$.

თუ :

- $\forall i = 1, \ldots, n, \quad \left\|\vec{x}^i\right\|_2 \le \mathcal{D}$,
- *არსებობს* $\vec{u} \in \mathbb{R}^{p+1}$ *ისეთი, რომ*
  - $\left\|\vec{u}^i\right\|_2 = 1$ *და*
  - $\forall i = 1, \ldots, n, \quad y^i \langle \vec{u}, \vec{x} \rangle \ge \gamma$,

მაშინ პერსეპტრონის ალგორითმი იკრიბება და ეს მოითხოვს არაუმეტეს $\left(\frac{\mathcal{D}}{\gamma}\right)^2$ *ბიჯისა.*

ამერიკელი მათემატიკოსის ალბერტ ნოვიკოვის (Albert Boris J. Novikoff, 1928-) ეს ღრმაშინაარსიანი თეორემა მოცემულია ნაშრომში Novikoff, A. B. J. (1962).

რეგრესია

რეგრესიის შემთხვევაში ემპირიული ღირებულებისათვის გამოიყენება დანახარჯის კვადრატული ფუნქცია :

$$L\left(f\left(\vec{x}^i\right), y^i\right) = \frac{1}{2}\left(y^i - f\left(\vec{x}^i\right)\right)^2 = \frac{1}{2}\left(y^i - \left\langle \vec{w}, \vec{x} \right\rangle\right)^2 . \tag{7.6}$$

წონების განახლების წესი 7.4 შემდეგ სახეს იძენს :

$$w_j \leftarrow w_j - \eta\left(f\left(\vec{x}^i\right) - y^i\right)x_j^i . \tag{7.7}$$

ალბათური კლასიფიკაცია

ბინარული ან მრავალკლასიანი კლასიფიკაციის შემთხვევაში ღირებულების (დანახარჯის, დანაკარგის) ფუნქციად კროს-ენტროპია (ჯვარედინი ენტროპია) გამოიყენება :

$$L\left(f\left(\vec{x}^i\right), y^i\right) = -\sum_{c=1}^{C}\delta\left(y^i, c\right)\log f_c\left(\vec{x}^i\right) = -\sum\nolimits_{c=1}^{C}\delta\left(y^i, c\right)\log\frac{\exp\left(\left\langle \vec{w}^c, \vec{x} \right\rangle\right)}{\sum_{k=1}^{C}\exp\left(\left\langle \vec{w}^k, \vec{x} \right\rangle\right)} .$$

გამოთვლების სულ რამდენიმე სტრიქონი გვარწმუნებს, რომ წონათა განახლების წესი ისეთივეა, როგორც რეგრესიისათვის (7.7) :

$$w_j^k \leftarrow w_j^k - \eta\left(y_k^i - f\left(\vec{x}^i\right)\right)x_j^i .$$

ეს გამოთვლები სამართლიანია ბინარული (ორობითი) კლასიფიკაციისათვისაც, თანაც ფორმულით 7.7 მოცემული წესი ძალაში რჩება.

1.3 ბულის ფუნქციათა მოდელირება

ვინაიდან პერსეპტრონს წრფივი ფუნქციების გაწვრთნის უნარი აქვს, მას ასევე შეუძლია *ზოგიერთი* ლოგიკური ფუნქციის შესრულებაც ორობით შემავალ ვექტორებზე.

მაგალითად, პერსეპტრონი, რომლის წონებია $w_0 = -1.5, w_1 = 1$ და $w_2 = 1$, ანგარიშობს «AND» («და») ლოგიკური გამრავლების $\wedge$ ოპერატორს x_1 და x_2 არგუმენტთა შორის ნახატზე 7.3 მოცემული *ჭეშმარიტობის ცხრილის* (ინგლ. *truth table*) შესაბამისად.

x_1	x_2	$x_1 \wedge x_2$
0	0	0
0	1	0
1	0	0
1	1	1

ნახატი 7.3 - ჭეშმარიტობის ცხრილი «და»/«AND» $(\wedge)$ ლოგიკური გამრავლებისათვის.

ამის საპირისპიროდ, შეუძლებელია პერსეპტრონის გაწვრთნა ე.წ. XOR («eXclusif OR» — «ანგამორიცხვით») ლოგიკური ფუნქციის შესასრულებლად.

სხვანაირად «*ორის ტოლი მოდულით შეკრების*» სახელწოდებითაც ცნობილი ამ ლოგიკური

ფუნქციის ჭეშმარიტობის ცხრილი მოცემულია ნახატზე. 7.4 :

მართლაც, არცერთ წრფეს არ შეუძლია $(0,0)$ და $(1,1)$ წერტილების განცალკევება ერთი მხრივ, ხოლო $(0,1)$ და $(1,0)$ წერტილების — მეორე მხრივ.

x_1	x_2	$x_1 \oplus x_2$
0	0	0
0	1	1
1	0	1
1	1	0

ნახატი 7.4 - ჭეშმარიტობის ცხრილი ლოგიკური ფუნქციისათვის «ან-გამორიცხვით», ანუ «შეკრება ორის ტოლი მოდულით» — «eXclusif OR», ანუ «XOR» $(\oplus)$

2 მრავალშრიანი პერსეპტრონი

პერსეპტრონის მამოდელირებელი შესაძლებლობები შეზღუდულია, ვინაიდან იგი წრფივი მოდელია. პირველი ნეირონული მოდელებით გამოწვეული ენთუზიაზმის შემდეგ, სწორედ ეს წამოწყება გახდა გარკვეული იმედგაცრუების მიზეზი 1970-ანი წლების დასაწყისში, როდესაც ვერ მიიღეს, რასაც ელოდნენ.

სეიმურ პაპერტი (Seymour Aubrey Papert, 1928-2016, მათემატიკოსი და პედაგოგი) და მარვინ მინსკი (Marvin Lee Minsky, 1927–2016, კოგნიტოლოგი) იხსენებენ კიდეც თავიანთ ინტუიციას პერსეპტრონის იდეის გადატანის უსარგებლობის შესახებ მრავალშრიან არქიტექტურებზე : «*პერსეპტრონის შესწავლა ძალზე საინტერესო აღმოჩნდა, მისი სერიოზული შეზღუდვების მიუხედავად და, შესაძლოა, მათი წყალობითაც. ყურადღებას იპყრობს ზოგი მისი თავისებურება : წრფივობა, სწავლების მაინტრიგებელი თეორემა, პარალელური გამოთვლების პარადიგმის სიმარტივე. საფუძველი არ არის ვივარაუდოთ, რომ ყველა მისი ღირსება გადაიტანება მრავალშრიან ვერსიაზე. მიუხედავად ამისა, მნიშნელოვან კვლევით ამოცანად მიგვაჩნია მრავალშრიანი გაფართოების სტერილურობის შესახებ ჩვენი ინტუიციური აზრის დაზუსტება (ან უარყოფა)* ». (Minsky and Papert, 1972).

ისტორიამ დაამტკიცა, რომ ინტუიციამ უღალატა მათ.

2.1 არქიტექტურა

მრავალშრიანი პერსეპტრონი (ინგლ. MLP, Multi-Layer Perceptron) — ეს ნეირონული ქსელია, რომელიც აგებულია *შუალედური შრეების* ჩასმით შემავალ და გამომავალ შრეებს შორის. ზოგჯერ მათ *ფარულ შრეებს* (ინგლ. *hidden layers*) უწოდებენ. თითოეული ნეირონი შუალედურ ფენაში ან გამომავალ ფენაში იღებს შემავალი სიგნალის სახით წინა შრის ნეირონების გამოსასვლელებს. უკუკავშირი ერთი შრიდან წინა შრისკენ არ არსებობს ; ასეთ ნეირონულ ქსელს ეწოდება ნეირონული ქსელი მონაცემთა *პირდაპირი გადაცემით* (ინგლ. *feed-forward*) ან *ნეირონული ქსელი პირდაპირი გავრცელებით*.

აქტივაციის არაწრფივი ფუნქციების გამოყენებისას, როგორიცაა, მაგალითად, ლოგისტიკური ფუნქცია ან ჰიპერბოლური ტანგენსის ფუნქცია, იქმნება ძლიერ არაწრფივი პარამეტრული მოდელები.

მაგალითი

მაგალითის სახით განვიხილოთ ნახატზე 7.5 მოცემული პერსეპტრონი ორი ფარული შრით. დავუშვათ, რომ w_{jq}^{h} არის $h-1$ ფენის j ნეირონის h ფენის q ნეირონთან კავშირის წონა, a_h — აქტივაციის ფუნქცია, რომელიც გამოიყენება h ფენის გამოსასვლელზე, ხოლო p_h — ნეირონების რაოდენობა h ფენაში.

პირველი ფარული შრის რომელიღაც $q-$ური ნეირონის z_q^1 გამოსასვლელის მნიშვნელობა განისაზღვრება ფორმულით :

$$z_q^1 = a_1\left(\sum_{j=0}^{p} w_{jq}^1 x_j\right).$$

მეორე ფარული შრის რომელიღაც $q-$ური ნეირონის z_q^2 გამოსასვლელის მნიშვნელობა განისაზღვრება ფორმულით :

$$z_q^2 = a_2\left(\sum_{j=1}^{p_1} w_{jq}^2\, z_j^1\right).$$

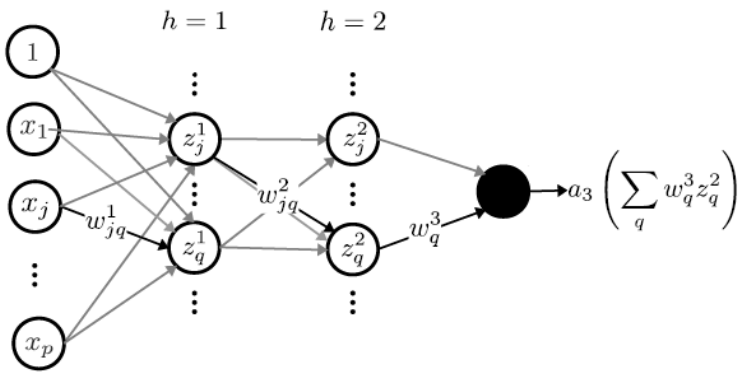
დაბოლოს, პერსეპტრონის გამოსასვლელისათვის გვაქვს :

$$f\left(\vec{x}\right) = a_3\left(\sum_{j=1}^{p_2} w_j^3\, z_j^2\right).$$

ამრიგად, თუ დავუშვებთ, რომ ფარული ფენების ყველა ნეირონისათვის ლოგისტიკური ფუნქცია გამოიყურება, მაშინ პერსეპტრონის გამოსასვლელი შემდეგ სახეს იძენს :

$$f\left(\vec{x}\right) = a_3\left(\sum_{j=0}^{p_2} w_{jq}^3 \frac{1}{1+\exp\left(-\sum_{j=0}^{p_1} w_{jq}^2 \frac{1}{1+\exp\left(-\sum_{j=0}^{p} w_{jq}^1 x_j\right)}\right)}\right),$$

რამაც უნდა დაგვარწმუნოს მრავალშრიანი პერსეპტრონის უნარში განახორციელოს არაწრფივი ფუნქციების მოდელირება.



ნახატი 7.5 - მრავალშრიანი პერსეპტრონის არქიტექტურა.

ყურადღება

მრავალშრიანი პერსეპტრონი — ეს პარამეტრული მოდელია, რომლის პარამეტრებია კავშირების w_{jq}^h წონები. (შრეთა რაოდენობა და ამ შრეებში ნეირონების რიცხვი ჰიპერპარამეტრების ნაწილს წარმოადგენს : ვუშვებთ, რომ ისინი ფიქსირებულია და მათი შესწავლა არ ხდება). ამრიგად, პარამეტრების (ე.ი. კავშირების წონათა) რაოდენობა ამ მოდელს მით უფრო მეტი აქვს, რაც უფრო მეტია შუალედური შრე და ამ შრეებში ნეირონების რიცხვი. სწორედ იმისათვის, რომ ზედმეტად სწავლების თავიდან აცილება მოხდეს, ღრმა ნეირონული ქსელები მოითხოვს მონაცემების უზარმაზარ მასივებს კარგად გაწვრთნილი მოდელების მისაღებად.

2.2 უნივერსალური აპროქსიმაცია

თეორემა 7.2 (უნივერსალური აპროქსიმაცია) *დავუშვათ, რომ $a:\mathbb{R}\to\mathbb{R}$ — შეზღუდული, უწყვეტი, ზრდადი ფუნქციაა და იგი არ წარმოადგენს კონსტანტას (მუდმივას), ხოლო K — $\mathbb{R}^p$ – ის კომპაქტური ქვესიმრავლეა. თუ მოცემულია $\mathcal{E}>0$ და უწყვეტი f ფუნქცია K-ზე, მაშინ არსებობს მთელი m რიცხვი, m რაოდენობის $\{d_i\}_{i=1,\ldots,m}$ სკალარი, m რაოდენობის $\{b_i\}_{i=1,\ldots,m}$ სკალარი და m რაოდენობის $\{\vec{w}_i\}_{i=1,\ldots,m}$ ვექტორი $\mathbb{R}^p$ – დან, ისეთი, რომ ყველა $\vec{x}\in K$ -თვის :*

$$\left| f\left(\vec{x}\right) - \sum_{i=1}^{m} d_i\, a\left(\left\langle \vec{w}_i, \vec{x}\right\rangle + b_i\right)\right| \le \mathcal{E}\,.$$

სხვა სიტყვებით რომ ვთქვათ, ნებისმიერი უწყვეტი ფუნქცია $\mathbb{R}^p$ – ის კომპაქტურ ქვესიმრავლეზე შეიძლება იყოს აპროქსიმირებული სიზუსტის ნებისმიერი ხარისხით მრავალშრიანი პერსეპტრონით, რომლის შუალედური შრე შეიცავს ნეირონების სასრულ რიცხვს.

ამერიკელი მათემატიკოსის ჯორჯ ციბენკოს (George Cybenko, 1989) მიერ მიღებული და ავსტრიელი მათემატიკოსის კურტ ჰორნიკის (Kurt Hornik, 1991) მიერ დაზუსტებული ეს თეორემა, მკვეთრად კონტრასტირებს, უპირისპირდება პაპერტისა და მინსკის ინტუიციას : ამჟამად შესაძლებელია ნებისმიერი ფუნქციების წვრთნა, სწავლება. მაგრამ აქვე უნდა აღვნიშნოთ,

რომ ეს შედეგი არ გვაძლევს არც იმ ნეირონების რიცხვს, რომლებმაც უნდა შეადგინოს ეს შუალედური შრე, არც ბმათა იმ წონებს, რომლებიც უნდა გამოვიყენოთ. ნეირონული ქსელები ერთი ფარული შრით, როგორც წესი, ნაკლებად ეფექტურია და პრაქტიკაში უკეთესი შედეგები ხშირად მიიღება შრეთა მეტი რაოდენობის გამოყენებისას.

2.3 XOR-ის მოდელირება მრავალშრიანი პერსეპტრონით

XOR ფუნქცია, რომლის ჭეშმარიტობის ცხრილი მოცემულია ნახატზე 7.4, ადვილად შეიძლება იყოს მოდელირებული ერთშრიანი პერსეპტრონით, რომელიც აქტივაციის ფუნქციებად მხოლოდ ზღურბლურ ფუნქციებს იყენებს. მაგალითად, შეიძლება შუალედური შრის გამოყენება სამი ნეირონით, ერთი წანაცვლებით და შემდეგი წონებით :

– $w_{01}^1 = -0.5, w_{11}^1 = 1, w_{21}^1 = -1;$ და ამრიგად $o_1^1 = -0,5 + x_1 - x_2$, რაც $(1,0)-$ის გამოყოფის საშუალებას იძლევა სხვა შესაძლებლობიდან ;

– $w_{02}^1 = -0.5, w_{12}^1 = -1, w_{22}^1 = +1;$ და ამრიგად $o_2^1 = -0,5 - x_1 + x_2$, რაც $(0,1)-$ის გამოყოფის საშუალებას იძლევა სხვა შესაძლებლობიდან ;

– $w_0^2 = -0.5, w_1^2 = 1, w_2^2 = 1;$ ასე რომ დასასრულ $f\left((x_1, x_2)\right) = -0.5 + \text{sign}\left(o_1^1\right) + \text{sign}\left(z_2^1\right)$.

2.4 სწავლება უკუგავრცელებით

გრადიენტული განპირობებულობა და არასტაბილურობა

მნიშვნელოვანია ხაზი გაესვას იმ გარემოებას, რომ ემპირიული რისკის მინიმიზაცია მრავალშრიანი პერსეპტრონისთვის არ არის ამოზნექილი ოპტიმიზაციის ამოცანა. ეს გვაიძულებს გამოვიყენოთ მიმართული დაშვების ალგორითმი (იხ. ქვედანაყოფი 13.3.3), რომელიც გლობალურ მინიმუმზე კრებადობის გარანტიას არც კი იძლევა.

გავიხსენოთ მათემატიკიდან, რომ *ფუნქციის განპირობებულობის რიცხვი* არგუმენტის მიმართ ზომავს ფუნქციის მნიშვნელობის ცვლილების სიდიდეს არგუმენტის მცირე ცვლილებისას. მართლაც, ღირებულების (ხარჯის) ფუნქციის გრადიენტი კავშირის (ბმის) წონით, როგორც წესი, განპირობებულია ცუდად, რაც იმას ნიშნავს, რომ ალგორითმის მუშაობის საწყისი პირობების მცირე შეშფოთებაც კი მიგვიყვანს სრულიად სხვა და მოულოდნელ შედეგამდე. გარდა ამისა, გრადიენტი წონის მიხედვით ერთ-ერთ პირველ შუალედურ შრეში ხშირად არასტაბილურია. იგი ან ძალიან მცირეა, რაც ანელებს სწავლების პროცესს და ეს ცნობილია როგორც *გაქრობადი გრადიენტის* (ინგლ. *vanishing gradient*) პრობლემა, ან ყოველ იტერაციასთან ერთად იძენს სულ უფრო ზრდად მნიშვნელობებს და ეს იწვევს *ფეთქებადი გრადიენტის* (ინგლ. *exploding gradient*) პრობლემას.

კავშირების (ბმების) წონათა ინიციალიზაცია, ცვლადების სტანდარტიზაცია, სწავლების სისწრაფისა და აქტივაციის ფუნქციის არჩევა გავლენას ახდენს მრავალშრიანი პერსეპტრონის კარგი გადაწყვეტილებისადმი კრებადობის უნარზე.

სწორედ ამიტომ წარმოადგენს მრავალშრიანი პერსეპტრონის სწავლება (წვრთნა) რთულ ამოცანას. და მხოლოდ 2006 წელს ბრიტანელ-კანადელი კოგნიტიური ფსიქოლოგის და კომპიუტერული მეცნიერის, განათლებულ სამყაროსაგან «ხელოვნური ინტელექტის ნათლიად»

(«Godfather of AI») შერაცხული ჯეფ ჰინტონის (Geoffrey Everest Hinton, 1947-) და სხვა სპეციალისტთა ნაშრომებმა, ასევე კომპიუტერების სიმძლავრის გაზრდამ ამ სიძნელეთა დაძლევის საშუალება უზრუნველყო და დააბრუნა ნეირონული ქსელების კვლევა წინა პლანზე.

გაჯერებულობა (სატურაცია)

კიდევ ერთი პრობლემა, რომელიც მრავალშრიან პერსეპტრონთან არის დაკავშირებული, იმ ვითარებაში ჩნდება, როცა ლოგისტიკური მოდულები (ბლოკები) ან 0-თან, ან 1-თან ძალიან ახლო მნიშვნელობებს გვიბრუნებს.

ეს მაშინ ხდება, როცა იმ სიგნალების შეწონილ o_j^h ჯამს, რომლებსაც ისინი იღებს შესასვლელზე, ძალიან დიდი ამპლიტუდა აქვს, სხვა სიტყვებით, როცა ბმათა (კავშირების) წონებს თავად აქვს დიდი ამპლიტუდა. ასეთ პირობებში ამ წონათა განახლება მხოლოდ უმნიშვნელო გავლენას ახდენს წინასწარმეტყველებაზე და ამბობენ, რომ ქსელი *გაჯერებულია*, ანუ *სატურირებულია*.

ასეთი სიტუაციის თავიდან ასაცილებლად შეიძლება ℓ_2 რეგულარიზაციას მივმართოთ (იხ. ქვედანაყოფი 6.2), რომელიც ხელოვნური ნეირონების კონტექსტში ცნობილია როგორც *შეწონვის დეგრადაცია* (ინგლ. weight decay — წონაში დაკლება), ანუ *შეწონვის მაჩვენებლების გაუარესება*. საქმე ეხება $L\left(f\left(\vec{x}^i\right), y^i\right)$ ემპირიული რისკისადმი კავშირის (ბმის) წონათა ვექტორის ევკლიდეს ნორმის დამატებას.

უკუგავრცელება (შექცეული გავრცელება)

მიუხედავად ამისა, მრავალშრიანი პერსეპტრონის სწავლების (წვრთნის) მთავარი, ძირითადი პრინციპი, რომელიც ცნობილია როგორც *შექცეული გავრცელება,* ანუ *უკუგავრცელება* (ინგლ. *backpropagation*, რაც ხშირად, შემოკლებისას, *backprop*-მდეც კი დაიყვანება), უკვე რამდენიმე ათწლეულია მისაწვდომი. ისევე, როგორც პერსეპტრონის შემთხვევაში შუალედური შრის გარეშე, იგი ეფუძნება გრადიენტული ალგორითმის გამოყენებას $L\left(f\left(\vec{x}^i\right), y^i\right)$ რისკის მინიმიზაციისათვის ყოველი ახალი დაკვირვებისათვის.

იმისათვის, რომ განახლდეს $h-1$ შრის j ნეირონიდან h შრის q ნეირონამდე კავშირის (ბმის) w_{jq}^h წონა, აუცილებელია $\frac{\partial L\left(f\left(\vec{x}^i\right), y^i\right)}{\partial w_{jq}^h}$.წარმოებულის გამოთვლა. ამისათვის შეიძლება გამოვიყენოთ *თეორემა რთული ფუნქციის წარმოებულის შესახებ* (ინგლ. *chain rule*— *ჯაჭვის წესი*). შემოვიღოთ o_j^h აღნიშვნა h შრის $j-$ური ნეირონის შესასვლელების წრფივი კომბინაციისათვის, მაშინ, ცხადია, $z_j^h = a_h\left(o_j^h\right)$. პირობითად მივიჩნიოთ, რომ $z_j^0 = x_j$. ასე რომ,

$$\left.\begin{aligned}\frac{\partial L\left(f\left(\vec{x}^{\,i}\right),y^i\right)}{\partial w_{jq}^h}&=\frac{\partial L\left(f\left(\vec{x}^{\,i}\right),y^i\right)}{\partial o_q^h}\frac{\partial o_q^h}{\partial w_{jq}^h}=\frac{\partial L\left(f\left(\vec{x}^{\,i}\right),y^i\right)}{\partial z_q^h}\frac{\partial z_q^h}{\partial o_q^h}\frac{\partial o_q^h}{\partial w_{jq}^h}\\&=\left(\sum_{r=1}^{p_{h+1}}\frac{\partial L\left(f\left(\vec{x}^{\,i}\right),y^i\right)}{\partial o_r^{h+1}}\frac{\partial o_r^{h+1}}{\partial z_q^h}\right)\frac{\partial z_q^h}{\partial o_q^h}\frac{\partial o_q^h}{\partial w_{jq}^h}\\&=\left(\sum_{r=1}^{p_{h+1}}\frac{\partial L\left(f\left(\vec{x}^{\,i}\right),y^i\right)}{\partial o_r^{h+1}}w_{qr}^{h+1}\right)a_h'\left(o_q^h\right)z_j^{h-1}\end{aligned}\right\}. \qquad (7.8)$$

ამრიგად, h შრის წონების განახლებისათვის აუცილებელი გრადიენტი, გამოითვლება როგორც იმ $\frac{\partial L\left(f\left(\vec{x}^{\,i}\right),y^i\right)}{\partial o_r^{h+1}}$ გრადიენტების ფუნქცია, რომლებიც საჭიროა $(h+1)$ შრის წონების განახლებისათვის.

ეს საშუალებას მოგვცემს გავამარტივოთ გამოთვლები *მემოიზაციის* (ინგლ. *memoization*) მეთოდის გამოყენების ხარჯზე, ე.ი. თავიდან ავიცილოთ იმ თანაფარდობების ხელახალი გამოთვლა, რომლებიც ჩვენს პროცედურაში რამდენჯერმე გვხვდება.

უფრო ზუსტად, მრავალშრიანი პერსეპტრონის სწავლება უკუგავრცელების მეთოდით იმაში მდგომარეობს, რომ ყოველი დასამუშავებელი $\left(\vec{x}^{\,i},y^i\right)$ დაკვირვებისათვის ერთმანეთს ენაცვლება პირდაპირი გავრცელების ეტაპი, რომელზეც თითოეული ნეირონის გამოსასვლელები გამოითვლება და შეცდომის უკუგავრცელების ფაზა, რომელშიც წონათა განახლება ხდება : 1). განახლების ეს პროცესი იწყება იმ წონებით, რომლებიც უკანასკნელი შუალედური შრიდან გამომავალი მოდულისკენ (ბლოკისკენ) მიდის და 2). გრძელდება ქსელში «ზევით (მაღლა) ასვლით» შესასვლელის წონებისკენ, რომლებითაც იგი პირველ შუალედურ შრეს უკავშირდება.

მაგალითი

ავიღოთ ნახატზე 7.5 წარმოდგენილი ქსელი ორი შუალედური შრით და გამოვიყენოთ : *იდენტურობა* აქტივაციის უკანასკნელ a_3 ფუნქციად, კვადრატული შეცდომის ფუნქცია და კიდევ ლოგისტიკური აქტივაციები a_1 და a_2 ფუნქციებისათვის. ამასთან ერთად გავიხსენოთ, რომ ლოგისტიკური ფუნქციის წარმოებული შეიძლება შემდეგი სახით ჩაიწეროს :

$$\sigma'(u)=u'\sigma(u)\left(1-\sigma(u)\right).$$

პირდაპირი გავრცელებისათვის ჩავატაროთ შემდეგი გამოთვლები :

$$\left.\begin{aligned}o_q^1&=\sum_{j=0}^{p}w_{jq}^1x_j;\ z_q^1=\sigma\left(o_q^1\right)\\o_q^2&=\sum_{j=0}^{p_1}w_{jq}^2z_j^1;\ z_q^2=\sigma\left(o_q^2\right)\\o^3&=\sum_{j=0}^{p_2}w_j^3z_j^2;\ f\left(\vec{x}^{\,i}\right)=z^3=o^3\end{aligned}\right\}.$$

უკუგავრცელებისას კი ჯერ გამოითვლება

$$\frac{\partial L\left(f\left(\vec{x}^i\right),y^i\right)}{\partial w_j^3}=\left(f\left(\vec{x}^i\right)-y^i\right)\frac{\partial f\left(\vec{x}^i\right)}{\partial w_j^3}=\left(f\left(\vec{x}^i\right)-y^i\right)z_j^2$$

გამოსახულება $f\left(\vec{x}^i\right)$ და z_j^2 მნიშვნელობათა გამოყენებით, რომლებიც ამას წინათ იყო შენახული გავრცელების პროცესში. ამრიგად,

$$w_j^3\leftarrow w_j^3-\eta\left(f\left(\vec{x}^i\right)-y^i\right)z_j^2\,.$$

ამის შემდეგ შეგვიძლია გამოვიყენოთ ფორმულა 7.8 და გამოვთვალოთ კერძო წარმოებული :

$$\frac{\partial L\left(f\left(\vec{x}^i\right),y^i\right)}{\partial w_{jq}^2}=\frac{\partial L\left(f\left(\vec{x}^i\right),y^i\right)}{\partial o_q^2}\frac{\partial o_q^2}{\partial w_{jq}^2},$$

სადაც

$$\frac{\partial L\left(f\left(\vec{x}^i\right),y^i\right)}{\partial o_q^2}=\frac{\partial L\left(f\left(\vec{x}^i\right),y^i\right)}{\partial f\left(\vec{x}^i\right)}w_q^3\sigma'\left(o_q^2\right)=\left(f\left(\vec{x}^i\right)-y^i\right)w_q^3z_q^2\left(1-z_q^2\right)\qquad(7.9)$$

და

$$\frac{\partial o_q^2}{\partial w_{jq}^2}=z_j^1\,.$$

ამიტომ შეგვიძლია გამოვიყენოთ $f\left(\vec{x}^i\right)$, z_q^2 და z_j^1 მნიშვნელობები, რომლებიც შენახული იყო ახლახან პირდაპირი გავრცელების დროს და w_q^3, რომელიც ასევე ამ ცოტა ხნის წინ განახლდა, w_{jq}^2 მნიშვნელობის შემდეგნაირად განახლების მიზნით :

$$w_{jq}^2\leftarrow w_{jq}^2-\eta\left(f\left(\vec{x}^i\right)-y^i\right)w_q^3z_q^2\left(1-z_q^2\right)z_j^1\,.$$

დაბოლოს, შეგვიძლია ხელახლა გამოვიყენოთ განტოლება 7.8 და გამოვთვალოთ კერძო წარმოებული :

$$\frac{\partial L\left(f\left(\vec{x}^i\right),y^i\right)}{\partial w_{jq}^1}=\left(\sum_{r=1}^{p_2}\frac{\partial L\left(f\left(\vec{x}^i\right),y^i\right)}{\partial o_r^2}w_{qr}^2\right)z_q^1\left(1-z_q^1\right)x_j\,.$$

გამოთვლათა დასასრულებლად ყველაფერი გვაქვს : z_q^1 გამოანგარიშებული იყო პირდაპირი გავრცელების პროცესში, w_{qr}^2 წონები განახლებული გვქონდა წინა ბიჯზე, ხოლო კერძო წარმოებულები $\frac{\partial L\left(f\left(\vec{x}^i\right),y^i\right)}{\partial o_q^2}$ ასევე წინა ბიჯზე იყო დადგენილი (7.9) თანაფარდობის სახით. ამიტომ შეგვიძლია ადვილად განვახორციელოთ უკუგავრცელების უკანასკნელი ბიჯიც :

$$w_{jq}^1 \leftarrow w_{jq}^1 - \eta \left(\sum_{r=1}^{p_2} \frac{\partial L\left(f\left(\vec{x}^i\right), y^i \right)}{\partial o_r^2} w_{qr}^2 \right) z_q^1 (1 - z_q^1) x_j \ .$$

შენიშვნა

ცხადია, ყოველ შუალედურ შრეში შეიძლება წანაცვლების ბლოკის (მოდულის) დამატება ; მაშინაც წარმოებულების გამოანგარიშება აქ გამოყენებულ პრინციპზე დაყრდნობით იქნება განსახორციელებელი.

2.5 ღრმა სწავლების კავშირი ნეირონულ ქსელებთან

გამოსახულებათა ანალიზიდან ავტონომიურ (უპილოტო) ავტომობილებამდე და მეტყველების ამოცნობამდე — საერთოდ, მანქანური სწავლების სფეროში უკანასკნელი მიღწევების უმრავლესობა, რომლებიც პრესაში შუქდებოდა, — ეფუძნება ღრმა ნეირონულ ქსელებს (ინგლ. *deep neural nets*).

ამ სფეროს დაწვრილებითი განხილვა სცილდება ლექციათა ჩვენი კურსის ფარგლებს და ამიტომ აქ ჩვენ მხოლოდ ზოგად ნიშნებს შევეხებით.

ღრმა სწავლების სფერო, ძირითადად, ეფუძნება იმ პრინციპებს, რომლებიც ჩვენ ამ ცოტა ხნის წინ გავაშუქეთ.

ფაქტობრივად, მრავალშრიანი პერსეპტრონი ღრმა, სიღრმისეული ხდება, როგორც კი მასში ჩნდება შრეთა საკმარისი რაოდენობა — განსაზღვრება «*საკმარისი*» დამოკიდებულია ავტორებზე.

ღრმა სწავლების სფერო ინტერესდება ასევე მრავალი სხვა არქიტექტურით, მაგალითად, რეკურენტული ნეირონული ქსელებით (ინგლ. *RNN, Recursive Neural Nets*), სახელდობრ, ქსელებით ხანგრძლივი მოკლევადიანი მეხსიერებით (ინგლ. *LSTM, Long Short-Term Memory*) მიმდევრობითი (მაგალითად, ტექსტური ან დროითი) მონაცემების მოდელირებისათვის და კონვოლუციური (ნახვევის ტიპის) ნეირონული ქსელებით (ინგლ. *CNN, Convolutional Neural Nets*) ან კაფსულური ქსელებით (ინგლ. *CapsNet, Capsule Neural Network*) გამოსახულებათა დასამუშავებლად. ყველა შემთხვევაში საქმე ეხება, ფაქტობრივად, იმას, რომ :

— გამოვიყენოთ ეს არქიტექტურები (პოტენციურად ძალიან რთული) პარამეტრული მოდელების შესაქმნელად ;

— გავწვრთნათ, დავასწავლოთ წონითი კოეფიციენტები მიმართული დაშვების ალგორითმის საშუალებით.

ღრმა სწავლების სირთულეები

მაგრამ ღრმა ნეირონული ქსელების წონათა სწავლება დაკავშირებულია ტექნიკურ სირთულეებთან : ოპტიმიზაციის გადასაწყვეტი ამოცანა არ არის ამოზნექილი და კრებადობა «კარგ» ლოკალურ მინიმუმზე — ადვილ ამოცანათა რიცხვს არ მიეკუთვნება. რაც უფრო რთულია ქსელი, მით უფრო რთულია ამოცანა და პროგრესი ამ სფეროში შესაძლებელია მხოლოდ მისი გადაწყვეტის შემამსუბუქებელი მეთოდების შექმნას.

გარდა ამისა, ღრმა ნეირონულ ქსელებს უამრავი პარამეტრი აქვს და ამიტომ ეს ქსელები მოითხოვს მონაცემთა დიდ მოცულობებს ზედმეტი, გადაჭარბებული სწავლების თავიდან ასაცილებლად.

აქედან გამომდინარე, როგორც წესი, აუცილებელია, რომ მათი განთავსება განაწილებულ არქიტექტურებზე ხდებოდეს.

ასე რომ, მიუხედავად წარმატებებისა გამოყენების ზოგ სფეროში, ღრმა სწავლება რთულია განსახორციელებლად და ყოველთვის არ წარმოადგენს ოპტიმალურ მიგნებას მანქანური სწავლების ამოცანათა გადასაწყვეტად, განსაკუთრებით მაშინ, როცა საქმე გვაქვს მონაცემთა უმნიშვნელო კრებულთან.

წარმოდგენათა დასწავლა

ღრმა ნეირონული ქსელები ხშირად უკავშირდება *შესამეცნებელ საკითხზე წარმოდგენების* ან *შესამეცნებელ მოვლენაზე* (*ობიექტზე*) *ნიშნების დასწავლის* ცნებას. მართლაც, შეიძლება ჩაითვალოს, რომ ყოველი მომდევნო შუალედური შრე $\left(z_1^h, z_2^h, \ldots, z_{p_h}^h\right)$ მონაცემების ახლებურად წარმოდგენას სწავლობს წინა შრის წარმოდგენის საფუძველზე, ვიდრე შესაძლებელი არ გახდება წრფივი ალგორითმის გამოყენება უკანასკნელ შუალედურ შრესა და ქსელის გამოსასვლელს შორის. ეს ასპექტი გამოიყენება, კერძოდ, ავტოენკოდერებში (ინგლ. *autoencoder*), რომლებიც დაწვრილებით იქნება განხილული მე-11 ლექციაში.

3 საკვანძო მომენტები

— პერსეპტრონი გამოიყენება წრფივი პარამეტრული მოდელების სწავლებისათვის და იწვრთნება თავისი წონების იტერაციული განახლების გზით მიმართული დაშვების ალგორითმის საფუძველზე.

— მრავალშრიანი პერსეპტრონი შეიძლება იყოს გამოყენებული არაწრფივი პარამეტრული მოდელების სწავლებისათვის და უზრუნველყოფს მოდელირების დიდ მოქნილობას.

— მრავალშრიანი პერსეპტრონი იწვრთნება *უკუგავრცელების* (*შექცეული გავრცელების*) მეთოდით, რომელშიც თეორემა რთულ ფუნქციათა წარმოებულზე შერწყმულია *მემოიზაციის* ანუ *შენახვის* (ინგლ. *memoization*) პროცედურასთან, რომლის არსია ფუნქციის წინა შედეგების *შენახვა* და მათი *მომდევნო გამოყენება*, ამ ფუნქციის ხელმეორედ გამოძახების ნაცვლად და ამით მაღალი გამოთვლითი ეფექტურობის უზრუნველსაყოფად. არ უნდა აგვერიოს სიტყვასთან *მემორიზაცია*, რაც მხოლოდ *დამახსოვრებას* ნიშნავს (ამ ოპერაციის დანიშნულების დაუზუსტებლად).

— ოპტიმიზაციის ამოცანა მრავალშრიანი პერსეპტრონისათვის და, უფრო ზოგად შემთხვევაში, ნებისმიერო ღრმა ხელოვნური ნეირონული ქსელისათვის არ არის ამოზნექილი და კარგი მინიმუმის მიღება არც ისე ადვილია.

დამატებითი ინფორმაცია

• ამ ლექციაში ჩვენ მხოლოდ ზოგად ასპექტში განვიხილეთ ღრმა სწავლების ძირითადი სტრუქტურული ელემენტები. პრობლემის უფრო დაწვრილებით გაცნობა შეიძლება წიგნებიდან Goodfellow et al (2016) და ფრანგიშვილი et al (2020).

· http://playground.tensorflow.org/ — «მუშაობის», «კვლევის», «თამაშისა» და «ექსპერიმენტირების» საშუალებას იძლევა ღრმა სწავლების ქსელის არქიტექტურასთან და სწავლებასთან (წვრთნასთან) ბრაუზერში.

4 ბიბლიოგრაფია

1. Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4) :303–314.

2. Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT Press, Cambridge, MA. https://www.deeplearningbook.org/

3. Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2) :251–257.

4. Minsky, M. and Papert, S. (1972). *Perceptrons: an Introduction to Computational Geometry*. MIT Press, Cambridge, MA.

5. Novikoff, A. B. J. (1962). On convergence proofs on perceptrons. In *Proceedings of the Symposium on the Mathematical Theory of Automata*, Polytechnic Institute of Brooklyn, Vol. XII, pages 615–622: https://cs.uwaterloo.ca/~y328yu/classics/novikoff.pdf .

6. Rosenblatt, F. (1957). The perceptron. A perceiving and recognizing automaton. Technical Report 85-460-1, Cornell Aeronautical Laboratory, INC., Buffalo, N. Y., Project PARA, Prepared by: Frank Rosenblatt, Project Engineer : https://blogs.umass.edu/brain-wars/files/2016/03/rosenblatt-1957.pdf .

7. ფრანგიშვილი, ა. ი., ნამიჩეიშვილი, ო. მ., გოგიაშვილი ჟ.გ. (2020). ღრმა სწავლება, თბილისი : საგამომცემლო სახლი „ტექნიკური უნივერსიტეტი“, 155 გვ.

5 სავარჯიშოები

7.1 ააგეთ პერსეპტრონი, რომელიც ახორციელებს «ან» ლოგიკურ ფუნქციას სამ ცვლადს შორის.

7.2 ანიკამ ჩაატარა რამდენიმე წრფივი ალგორითმის სწავლება (წვრთნა) ათასი დაკვირვების შემცველ თავის მონაცემებზე და კმაყოფილი არ არის მათი შედეგებით. იგი ვარაუდობს, რომ რთული არაწრფივი ფუნქცია უფრო შესაფერისი ქნება. ანიკა მერყეობს ორ გადაწყვეტილებას შორის. პირველია ნეირონული ქსელის გამოყენება, რომელშიც 3 შუალედური შრე იქნება 100 ნეირონით თითოეულში, ხოლო მეორე გადაწყვეტილება გულისხმობს ნაკლებად მძლავრი ქსელის გამოყენებას ერთი შუალედური შრით, სადაც დაახლოებით 10 ნეირონი იქნება წარმოდგენილი. რომელი გადაწყვეტილება უნდა მიიღოს ანიკამ ?

7.3 თუ არის დამოკიდებული ნეირონული ქსელის მიერ შესასწავლი მოდელის ხარისხი : აქტივაციის ფუნქციების არჩევაზე ? შუალედურ შრეთა რაოდენობაზე ? ნეირონების რიცხვზე თითოეულ შრეში ? გამოყენებულ მონაცემთა რაოდენობაზე ?

7.4 ბაჩო წვრთნის ნეირონულ ქსელს და ამჩნევს, რომ რამდენიმე იტერაციის შემდეგ ღირებულების (დანახარჯების) ფუნქციის შემცირება წყდება. რაში შეიძლება მდგომარეობდეს ამის მიზეზი ?

7.5 ნეირონული ქსელის მამოდელირებელი უნარი (მოდელირების შესაძლებლობები) რომელი ფაქტორის გადიდებისას იზრდება : ფარულ ფენათა რაოდენობის ? სწავლების სიჩქარის ? წონათა გაუარესების მარეგულირებელი პარამეტრის ?

7.6 დაუმატეთ წანაცვლების ბლოკი ამ ლექციის მე-3 ნახატზე მოცემული მრავალშრიანი პერსეპტრონის არქიტექტურის თითოეულ შუალედურ შრეში. ჩაწერეთ უკუგავრცელების (შექცეული გავრცელების) წესები ლოგისტიკური ფუნქციის გამოყენებით აქტივაციის ფუნქციად ყოველ შრეში და კროს-ენტროპიის გამოყენებით დანახარჯების (ფასის) ფუნქციად (განიხილება, ცხადია, ბინარული კლასიფიკაცია).

7.7 განვიხილოთ პირდაპირი გავრცელების მრავალკლასიანი ნეირონული ქსელი შუალედური შრით, რომელიც შექმნილია $p+1$ შემავალი ბლოკით (ერთი ცალით ყოველ ცვლადზე და კიდევ ერთით წანაცვლებაზე), $H+1$ შუალედური ლოგისტიკური ბლოკით (წანაცვლების ჩათვლით) და softmax ტიპის K გამომავალი ბლოკით (ერთი ცალით თითოეულ კლასზე). უნდა ჩაიწეროს უკუგავრცელების (შექცეული გავრცელების) წესები.

სავარჯიშოთა ამონახსნები

7.1 $f\left(x_1, x_2, x_3\right) = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3$ და ერთადერთი მნიშვნელობა, რომლისთვისაც f უნდა იყოს უარყოფითი, არის $\left(0,0,0\right)$. სწორედ ამიტომ, $w_0 < 0$ და ყველა $j = 1,2,3$ მნიშვნელობისთვის, $w_0 + w_j > 0$. ასე რომ შესაძლებელი $w_0 = -1$ და $w_1 + w_2 + w_3 = 2$ არჩევანის გაკეთება.

7.2 სამი შუალედური შრის შემცველ ქსელს 400-მდე პარამეტრი აქვს, რაც ძალიან ბევრია განსაზღვრისათვის მხოლოდ ერთი ათასი დაკვირვებიდან. გაცილებით უკეთესია დაწყება უფრო მცირე ქსელით, რომელიც ნაკლები ალბათობით, ნაკლები რისკით იქნება ზედმეტად გაწვრთნილი იმისათვის, რომ აღმოჩნდეს ბლოკირებული ერთ-ერთ ლოკალურ მინიმუმზე.

7.3 მოდელის ხარისხი დამოკიდებულია ყველა ჩამოთვლილ ფაქტორზე.

7.4 ალბათ, ქსელმა მიაღწია ლოკალურ მინიმუმს. შესაძლოა, ეს მოხდა სწავლების ძალზე დაბალი სიჩქარის ან ქსელის გაჯერების გამო. რჩევაა : შეიცვალოს ქსელის ინიციალიზაცია, მოხდეს სწავლების ადაპტირებადი სიჩქარის გამოყენება, განხორციელდეს დასახმარებლად რეგულარიზაცია (წონათა შემცირების სახით).

7.5 მხოლოდ და მხოლოდ ფარულ ფენათა რაოდენობის გადიდებით.

7.6 შემავალი შრის წანაცვლების ბლოკის w_{0j}^1 წონები ემატება პირველი შუალედური შრის ნეირონებს, პირველი შუალედური შრის წანაცვლების ბლოკის w_{0j}^2 წონები ემატება მეორე შრის ნეირონებს, ხოლო მეორე შუალედური შრის ბლოკის w_0^3 წონა ემატება გამომავალ ნეირონს.

$$L\left(f\left(\vec{x}\right), y\right) = -y \ln f\left(\vec{x}\right) - \left(1-y\right)\ln\left(1 - f\left(\vec{x}\right)\right)$$

და

$$\sigma'\left(u\right) = u'\sigma\left(u\right)\left(1-\sigma\left(u\right)\right).$$

ამრიგად:

$$\frac{\partial L\left(f\left(\vec{x}^i\right),y^i\right)}{\partial w_j^3}=\frac{\partial f\left(\vec{x}^i\right)}{\partial w_j^3}\frac{f\left(\vec{x}^i\right)-y}{f\left(\vec{x}^i\right)\left(1-f\left(\vec{x}^i\right)\right)}=\left(f\left(\vec{x}^i\right)-y^i\right)z_j^2,$$

ე.ი. $w_0^3\leftarrow w_0^3-\eta\left(f\left(\vec{x}^i-y^i\right)\right)$ და, $j>0$ მნიშვნელობებისთვის, $w_j^3\leftarrow w_j^3-\eta\left(f\left(\vec{x}^i-y^i\right)\right)z_j^2$.

$$\frac{\partial L\left(f\left(\vec{x}^i\right),y^i\right)}{\partial w_{jq}^2}=\frac{\partial L\left(f\left(\vec{x}^i\right),y^i\right)}{\partial o_q^2}\frac{\partial o_q^2}{\partial w_{jq}^2}=\left(f\left(\vec{x}^i\right)-y^i\right)w_q^3z_q^2\left(1-z_q^2\right)z_j^1.$$

ასე რომ :

$$w_{0q}^2\leftarrow w_{0q}^2-\eta\left(f\left(\vec{x}^i\right)-y^i\right)w_q^3z_q^2\left(1-z_q^2\right)$$

და $j>0$ მნიშვნელობებისთვის

$$w_{jq}^2\leftarrow w_{jq}^2-\eta\left(f\left(\vec{x}^i\right)-y^i\right)w_{jq}^3z_q^2\left(1-z_q^2\right)z_j^1.$$

დაბოლოს :

$$w_{0q}^1\leftarrow w_{0q}^1-\eta\left(\sum_{r=1}^{p_2}\frac{\partial L\left(f\left(\vec{x}^i\right),y^i\right)}{\partial o_r^2}w_{qr}^2\right)z_q^1\left(1-z_q^1\right).$$

ამრიგად, ყველა $j>0$ მნიშვნელობისთვის გვაქვს :

$$w_{jq}^1\leftarrow w_{jq}^1-\eta\left(\sum_{r=1}^{p_2}\frac{\partial L\left(f\left(\vec{x}^i\right),y^i\right)}{\partial o_r^2}w_{qr}^2\right)z_q^1\left(1-z_q^1\right)x_j.$$

7.7 აღვნიშნოთ w_{jh} სიმბოლოთი შემავალი j ბლოკის შუალედურ h ბლოკთან კავშირის (შეერთების, ბმის) წონა, v_{hk} სიმბოლოთი — შუალედური h ბლოკის გამომავალ k ბლოკთან კავშირის (შეერთების, ბმის) წონა, z_h^i სიმბოლოთი — შუალედური h ბლოკის გამოსასვლელი $\vec{x}^i$ შესასვლელისთვის და f_k სიმბოლოთი — ფუნქცია, რომელიც გამოთვლილია გამოსასვლელის k -ურ ნეირონში. მაშინ :

$$\frac{\partial f_l\left(\vec{x}^i\right)}{\partial v_{hk}}=\begin{cases}z_h^if_k\left(\vec{x}^i\right)\left(1-f_k\left(\vec{x}^i\right)\right), & l=k\\ -z_h^if_k\left(\vec{x}^i\right)f_l\left(\vec{x}^i\right), & l\neq k\end{cases}$$

და

$$\frac{\partial L\left(y^i,f_l\left(\vec{x}^i\right)\right)}{\partial v_{hk}}=\left(\sum_{l\neq k}y_l^i\right)z_h^if_k\left(\vec{x}^i\right)-y_k^iz_h^if_k\left(\vec{x}^i\right)=z_h^i\left(f_k\left(\vec{x}^i\right)-y_k^i\right),$$

ვინაიდან, თუ $y_k^i = 1$, მაშინ $\sum_{l \neq k} y_l^i = 0$ და, პირიქით.

ასე რომ, მიიღწევა შემდეგი თანაფარდობები :

$$v_{hk} \leftarrow v_{hk} - \eta z_h^i \left(f_k\left(\vec{x}^i\right) - y_k^i \right)$$

და

$$w_{jh} \leftarrow w_{jh} - \eta \sum_{k=1}^{K} v_{hk} \left(f_k\left(\vec{x}^i\right) - y_k^i \right) z_h^i \left(1 - z_h^i\right) x_j^i \ .$$

ლექცია 8 უახლოესი მეზობლების მეთოდი

შინაარსი

მოცემულ კურსში წარმოდგენილია კონცეპტუალურად ძალიან მარტივი, მაგრამ მძლავრი არაპარამეტრული პროგნოზირების ალგორითმი, რომელიც შეიძლება იყოს გამოყენებული გადაწყვეტილების მიღების რთული საზღვრების ასაგებად მონაცემთა განაწილებების შესახებ ყოველგვარი დაშვების გარეშე. «უახლოესი მეზობლების» სახელწოდებით ცნობილი ეს ალგორითმი ეფუძნება «*ერთნაირი ბუმბულით შემოსილი ფრინველები ერთად იყრის თავს*» (ინგლ. «*Birds of a feather flock together*») პრინციპს და იყენებს უახლოესი მაგალითების ჭდეებს გადაწყვეტილების მისაღებად.

ამ ალგორითმის ძალა მდგომარეობს დაკვირვებებს შორის მანძილებისა და მსგავსებათა ფუნქციის არჩევაში. ამ თავში ჩვენ ასევე განვიხილავთ მანძილებისა და მსგავსებათა განსაზღვრას მონაცემთა სხვადასხვა წარმოდგენისათვის. აქვე შევისწავლით შემთხვევებს, როცა p განზომილების ნამდვილი ვექტორით წარმოდგენა არ არის აუცილებლად ყველაზე შესაფერისი.

დაბოლოს, ვნახავთ, როგორ შეიძლება იყოს გამოყენებული უახლოესი მეზობლები პრაქტიკაში ერთობლივი (თანამშრომლობითი) ფილტრაციისათვის (ინგლ. *collaborative filtering*).

მიზნები

— k უახლოეს მეზობელთა ალგორითმის რეალიზაცია.

— მანძილების და მსგავსებათა გამოთვლა მონაცემების სხვადასხვა ტიპის წარმოდგენისათვის.

— გადაწყვეტილების მიღების საზღვრის განსაზღვრა უახლოეს მეზობელთა ალგორითმისათვის.

– k უახლოესი მეზობლის ალგორითმის მაღალი განზომილების პირობებში მუშაობის შეუძლებლობის მიზეზთა ახსნა.

1 უახლოესი მეზობლის მეთოდი

1.1 მეთოდი

განსაზღვრება 8.1 (უახლოესი მეზობლის ალგორითმი) დავუშვათ, რომ მოცემულია n მონიშნული დაკვირვების $\mathcal{D} = \left\{ \left(\vec{x}^i, y^i \right)_{i=1,\ldots,n} \right\}$ სიმრავლე და d მანძილი მათ შორის $\mathcal{X}$-ზე, მაშინ *უახლოესი მეზობლის ალგორითმი* ეწოდება ალგორითმს, რომელიც ახორციელებს ახალი $\vec{x}$ დაკვირვების მონიშვნას საწვრთნელი სიმრავლიდან აღებული უახლოესი წერტილის ჭდით :

$$f\left(\vec{x}\right) = y^{\arg\min\limits_{i=1,\ldots,n} d\left(\vec{x},\vec{x}^i\right)}.$$

⚙

ეს ალგორითმი შეიძლება წარმატებით იყოს გამოყენებული როგორც კლასიფიკაციის, ასევე რეგრესიის ამოცანის ამოსახსნელად.

1.2 ვორონოის დიაგრამა

უახლოესი მეზობლის ალგორითმი ახდენს $\mathcal{X}$ სივრცის დაყოფას n ნაწილად : $i-$ური ნაწილი ამ ნაწილებიდან — ეს $\mathcal{X}$ სივრცის იმ წერტილების სიმრავლეა, რომელთა უახლოეს მეზობელს $\mathcal{D}$-ში წარმოადგენს $\vec{x}^i$:

$$\mathcal{X} = \bigcup_{i=1}^{n} \left\{ \vec{x} \in \mathcal{X} : d\left(\vec{x},\vec{x}^i\right) \le d\left(\vec{x},\vec{x}^l\right) \; \forall \vec{x}^l \in \mathcal{D} \right\}.$$

ასეთ დაყოფას *ვორონოის დიაგრამა* (ინგლ. *Voronoi diagram*) ეწოდება და ტერმინი დაკავშირებული უკრაინელი (რუსეთის იმპერიიდან) მათემატიკოსის სახელთან — Georgy Theodosiyovych Voronoi (1868-1908).

განსაზღვრება 8.2 (ვორონოის დიაგრამა) დავუშვათ, რომ მოცემულია მეტრიკული $\mathcal{X}$ სივრცე $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ მანძილით და $\mathcal{X}$-ის ელემენტების S სასრული სიმრავლე.

თუ მოცემულია *წყაროდ, თესლად, მარცვლად* (ან *ჩანასახად,* ინგლ. germ) წოდებული $\vec{u} \in S$ ელემენტი, მაშინ *ვორონოის უჯრედი* წარმოადგენს სიმრავლეს, რომელიც შემდეგი გამოსახულებით განისაზღვრება :

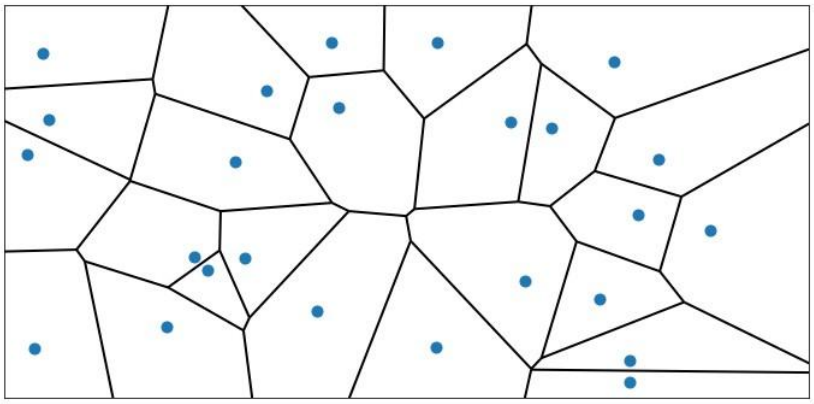
$$\mathrm{Vor}\left(\vec{u}\right) = \left\{ \vec{x} \in \mathcal{X} : \forall \vec{v} \in S, \; d\left(\vec{x},\vec{u}\right) \le d\left(\vec{x},\vec{v}\right) \right\}.$$

ვორონოის უჯრედი — ეს ამოზნექილი სიმრავლეა, ხოლო ვორონოის ყველა უჯრედის გაერთიანება აყალიბებს $\mathcal{X}$ მეტრიკული სივრცის ვორონოის დაყოფად წოდებულ *ვორონოის დიაგრამას* :

$$\mathcal{X} = \bigcup_{\vec{u} \in S} \mathrm{Vor}\left(\vec{u}\right).$$

⚙

ნახატზე 8.1 მოცემულია სიბრტყეზე ევკლიდეს მანძილის გამოყენებით გაანგარიშებული ვორონოის დიაგრამის მაგალითი. უჯრედები წარმოადგენს ამოზნექილ მრავალკუთხედებს. ყველა წერტილს ერთისა და იმავე უჯრედის შიგნით აქვს იგივე ჭდე, რაც მათ მარცვალს.

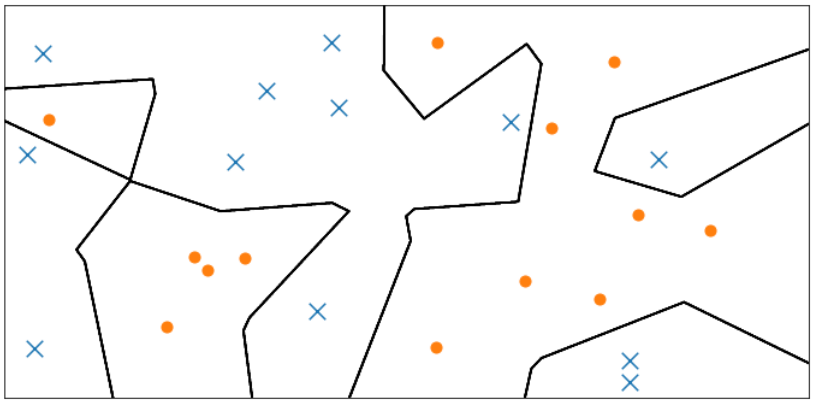


ნახატი 8.1 - ვორონოის დიაგრამა 25 წერტილისათვის ორგანზომილებიან სივრცეში ევკლიდეს მანძილის გამოყენებით.

უახლოესი მეზობლის ალგორითმის წარმოდგენის სხვა ხერხი — ეს $\mathcal{X}$ მეტრიკული სივრცის დაყოფაა იმდენ ნაწილად, რამდენიც სხვადასხვა ჭდე არსებობს საწვრთნელ სიმრავლეში. y ჭდის შესაბამისი ნაწილი, წარმოადგენს ვორონოის იმ უჯრედების გაერთიანებას, რომლებიც შეესატყვისება y ჭდით მონიშნულ დაკვირვებას $\mathcal{D}$ – დან :

$$\mathcal{X} = \bigcup_{y} \left\{ \vec{x} \in \mathcal{X} : y^{\arg\min_{i=1,\ldots,n} d\left(\vec{x},\vec{x}^{i}\right)} = y \right\}.$$

ასე რომ, ეს ძალიან მარტივი ალგორითმი შეიძლება იყოს გამოყენებული გადაწყვეტილების მიღების გაცილებით უფრო რთული საზღვრის ასაგებად, ვიდრე წრფივი პარამეტრული ალგორითმი აკეთებს ამას, როგორც ნაჩვენებია ნახატზე 8.2.



ნახატი 8.2 - უახლოესი მეზობლის ალგორითმის გადაწყვეტილების მიღების საზღვარი ნახ.8.1-ის უკვე მონიშნული მონაცემებისათვის

2 უახლოესი მეზობლების მეთოდი

უახლოესი მეზობლის მეთოდს ის ნაკლოვანება აქვს, რომ იგი ძალიან მგრძნობიარეა ხმაურის მიმართ : თუ დაკვირვება მონიშნულია შეცდომით (რაც სავსებით შესაძლებელია, ნაწილობრივ

გაზომვათა ხმაურის გამო და ნაწილობრივ გაბატონებულ წარმოდგენათა არასრულყოფილების მიზეზით) ან ცუდად არის განსაზღვრული, მაშინ მისი ყველა წერტილი ვორონოის სათანადო უჯრედში არასწორად იქნება მონიშნული. ამ მეთოდის საიმედოობის ასამაღლებლად მიმართავენ ჭდის მისანიჭებელი დაკვირვების *რამდენიმე* მეზობლის «აზრის» გაერთიანებას.

2.1 k უახლოესი მეზობლის მეთოდი

განსაზღვრება 8.3 (k უახლოესი მეზობლის ალგორითმი) თუ მოცემულია n მონიშნული დაკვირვების $\mathcal{D} = \left\{ \left(\vec{x}^i, y^i \right)_{i=1,\ldots,n} \right\}$ სიმრავლე, d მანძილი მეტრიკულ $\mathcal{X}$ სივრცეზე და $k \in \mathbb{N}^*$ ჰიპერპარამეტრი, მაშინ ალგორითმს, რომელიც მდგომარეობს ახალი $\vec{x}$ დაკვირვების მონიშვნაში საწვრთნელი სიმრავლის იმ k წერტილის ჭდეთა შესაბამისად, ვისთანაც იგი ყველაზე უფრო ახლოსაა, k *უახლოესი მეზობლის ალგორითმი* ეწოდება და ხშირად «k *Nearest Neighbors*» ინგლისური ფრაზის შესატყვისი *kNN* აკრონიმით აღინიშნება. თუ $\mathcal{D}$-ში $\vec{x}$ დაკვირვების k უახლოესი მეზობლის სიმრავლისათვის $\mathcal{N}_k\left(\vec{x}\right)$ ნოტაციას შემოვიღებთ, მაშინ k უახლოესი მეზობლის ალგორითმი შეიძლება შემდეგნაირად იყოს ასახული :

— კლასიფიკაციის ამოცანის ამოსახსნელად *მაჟორიტარული კენჭისყრა* გამოიყენება, და ამ დროს $\vec{x}$ დაკვირვება იძენს *უმრავლესობის მიერ შემოთავაზებულ* ჭდეს თავისი k უახლოესი მეზობლის ჭდეთა რიცხვიდან :

$$f\left(\vec{x}\right) = \arg\max_{c} \sum_{i:\vec{x}^i \in \mathcal{N}_k(\vec{x})} \delta\left(y^i, c\right).$$

— რეგრესიის ამოცანისათვის $\vec{x}$ დაკვირვება იძენს ჭდის სახით თავისი k უახლოესი მეზობლის ჭდეთა *საშუალო მნიშვნელობას* :

$$f\left(\vec{x}\right) = \frac{1}{k} \sum_{i:\vec{x}^i \in \mathcal{N}_k(\vec{x})} y^i \,.$$

იმ შემთხვევაში, როცა $k = 1$, უბრუნდებიან უახლოესი მეზობლის ალგორითმს.

k უახლოესი მეზობლის ალგორითმი წარმოადგენს *არაპარამეტრული სწავლების* მაგალითს : გადაწყვეტილების მიღების ფუნქცია გამოისახება როგორც დანაკვირვები მონაცემების ფუნქცია, და არა როგორც ცვლადების ანალიზური ფუნქციის სიმბოლური ჩანაწერი (ე.ი. ფორმულა).

ამ ალგორითმის მუშაობა შეიძლება შევადაროთ *მსჯელობებს, არგუმენტაციას, მოქმედებებს კონკრეტული შემთხვევების ან სიტუაციების საფუძველზე*, რაც მდგომარეობს იმაში, რომ ადამიანი მოქმედებს წინად ანალოგიურ სიტუაციებში გაკეთებული არჩევანის გახსენებით. სასამართლო პრაქტიკაშიც კი არსებობს პრეცედენტის დოქტრინა (ლათ. stare decisis) — როცა სასამართლო მიჰყვება თავისი ან ზემდგომი ორგანოების წინა გადაწყვეტილებების ლოგიკას. ან კიდევ, მაგალითად, ექიმი, რომელიც მკურნალობს თავის პაციენტს და იხსენებს როგორ ხდებოდა მსგავსი სიმპტომების მქონე პაციენტების გამოჯანმრთელება წარსულში.

k უახლოესი მეზობლის ალგორითმის გადაწყვეტილების მიღების საზღვარი უბან-უბან წრფივია ; k პარამეტრის გაზრდისას იგი «მარტივდება», ვინაიდან მაჟორიტარული კენჭისყრა იმ ცდომილებათა კომპენსირების საშუალებას იძლევა, რომლებიც იქმნება ცალკეული მაგალითებით, როგორც ეს ნაჩვენებია ნახატზე 8.3.

ნახატი 8.3 - ხუთი უახლოესი მეზობლის ალგორითმის გადაწყვეტილების საზღვარი 8.1 და 8.2 ნახატების მონაცემებისათვის.

2.2 ზარმაცი სწავლება

k უახლოესი მეზობლის ალგორითმს ზოგჯერ *ზარმაც სწავლებას* (ინგლ. *lazy learning*) იწოდებენ. ეს დაკავშირებულია იმასთან, რომ სწავლების პროცედურა მდგომარეობს მხოლოდ მონაცემების *შენახვაში* საწვრთნელი ნაკრებიდან და არავითარ გამოთვლებს არ გულისხმობს. მაგრამ ეს შეიძლება შემზღუდველ ფაქტორად იქცეს მეხსიერების დონეზე, თუ საწვრთნელი სიმრავლე ძალიან დიდია.

ამისგან განსხვავებით, წინასწარმეტყველების პროცედურა მოითხოვს მანძილის, ანუ დაშორების გამოთვლას დაკვირვებიდან და ეს მანძილი უნდა მოინიშნოს საწვრთნელი სიმრავლის *თითოეული* დაკვირვებისათვის. ამან მნიშვნელოვანი გამოთვლითი დანახარჯები შეიძლება მოითხოვოს, თუ საწვრთნელი სიმრავლე ძალიან დიდია. წინასწარმეტყველებისათვის ჯერ საჭიროა n მანძილის გამოთვლა, რაც p განზომილების პირობებში $\mathcal{O}(np)$ რიგის სირთულის პროცედურაა, ხოლო შემდეგ უნდა მოიძებნოს ყველა ამ მანძილიდან k უმცირესი მანძილი, რაც $\mathcal{O}(n \log k)$ სირთულის ოპერაციას წარმოადგენს.

2.3 უახლოესი მეზობლების რაოდენობა

როგორც ზემოთ იყო აღწერილი, მიზანშეწონილია $k > 1$ მნიშვნელობის გამოყენება იმისათვის, რათა k უახლოესი მეზობლის ალგორითმი გავხადოთ ხმაურის მიმართ უფრო მედეგი. მაგრამ, პირიქით, თუ $k = n$ მნიშვნელობის გამოყენება მოხდა, მაშინ კლასიფიკაციის ამოცანის შემთხვევაში ალგორითმი იწინასწარმეტყველებს $\mathcal{D}$ ნაკრების მაჟორიტარულ კლასს, ხოლო რეგრესიის ამოცანის შემთხვევაში — ამავე $\mathcal{D}$ ნაკრების ჭდეთა საშუალო მნიშვნელობას, რაც ასეთივე არადამაკმაყოფილებელ გადაწყვეტილებად უნდა იყოს მიჩნეული.

ამიტომ აუცილებელი იქნება შუალედური მნიშვნელობის არჩევა, რაც, ჩვეულებრივ, კროს-ვალიდაციით (ჯვარედინი შემოწმებით) კეთდება. ზოგჯერ ასევე $k \approx \sqrt{n}$ ევრისტიკაც გამოიყენება.

2.4 ვარიანტები

ε – მეზობლები

ნაცვლად იმისა, რომ უახლოესი მეზობლების ფიქსირებული რაოდენობის განხილვა ხდებოდეს, ჩვენ შეგვიძლია უპირატესობა მივცეთ ყველა იმ საწვრთნელ მაგალითს, რომლებიც საკმარისად ახლოსაა მოსანიშნავ დაკვირვებასთან : ეს საწვრთნელი სიმრავლის უფრო ეფექტურად გამოყენების საშუალებას იძლევა იმ არეებში, სადაც მისი სიმკვრივე მეტია. გარდა ამისა, არა ფარდობითი, არამედ აბსოლუტური გაგებით ახლო მაგალითებზე დაფუძნებული პროგნოზები ინტუიციურად უფრო საიმედოა, ვიდრე ერთმანეთისაგან დაშორებულ მაგალითებზე გაკეთებული პროგნოზები.

განსაზღვრება 8. 4 (ε – მეზობელთა ალგორითმი) დავუშვათ, რომ მოცემულია n მონიშნულ (მარკირებულ) დაკვირვებათა $\mathcal{D} = \left\{ \left(\vec{x}^i, y^i \right)_{i=1,\ldots,n} \right\}$ ნაკრები, d მანძილი $\mathcal{X}$ მეტრიკულ სივრცეზე და $\varepsilon \in \mathbb{R}_+$ ჰიპერპარამეტრი. მაშინ *ε – მეზობელთა* (ინგლ. *ε – ball neighbors*) *ალგორითმი* ეწოდება ალგორითმს, რომელიც მდგომარეობს ახალი $\vec{x}$ დაკვირვების მონიშვნაში საწვრთნელი სიმრავლის ყველა იმ წერტილის ჭდეთა შესაბამისად, რომელიც $\vec{x}$-დან ε-ზე ნაკლები მანძილით არის განთავსებული.

— კლასიფიკაციის ამოცანისათვის :

$$f\left(\vec{x}\right) = \arg\max_{c} \sum_{i:d\left(\vec{x},\vec{x}^i\right)\le\varepsilon} \delta\left(y^i, c\right).$$

— რეგრესიის ამოცანისათვის :

$$f\left(\vec{x}\right) = \arg\max_{c} \frac{1}{\left|\left\{i : d\left(\vec{x}, \vec{x}^i\right) \le \varepsilon\right\}\right|} \sum_{i:d\left(\vec{x},\vec{x}^i\right)\le\varepsilon} y^i \,.$$

მოცემული ალგორითმის სავსებით გასაგები შეზღუდვაა ალტერნატიული სტრატეგიის განსაზღვრის აუცილებლობა მაგალითების არარსებობის შემთხვევაში არჩეული ε რადიუსის სფეროში.

მეზობლების შეწონვა

ალტერნატივის სახით, რათა იმ საკმაოდ დამაჯერებელი წარმოდგენის გათვალისწინება მოხდეს, რომ ახლო მეზობელი დაშორებულ მეზობელზე, მართლაც, უფრო საიმედოა წინასწარმეტყველების გასაკეთებლად, თითოეული მეზობლის წვლილის *შეწონვა* შეიძლება განხორციელდეს მოსანიშნავი დაკვირვებიდან მისი დაშორების ფუნქციის სახით, როგორც წესი, შემდეგი ფორმულების საფუძველზე :

$$w_i = \frac{1}{d\left(\vec{x}, \vec{x}^i\right)}$$

ან

$$w_i = e^{-\left(\frac{1}{2} d\left(\vec{x}, \vec{x}^i\right)\right)}.$$

კლასიფიკაციის ამოცანის შემთხვევაში ნებისმიერი c კლასისათვის გამოითვლება ჯამი, სადაც თითოეული შესაკრები წარმოადგენს ამ კლასისადმი მიკუთვნებული რომელიმე ერთი მეზობლის წვლილს; შემდეგ, ყველა კლასისათვის ასეთნაირად მიღებული ჯამების შედარებით, ანალიზი ტარდება და შედეგი გამოისახება ფორმულით :

$$f\left(\vec{x}\right) = \arg\max_{c} \sum_{i:\vec{x}^i \in \mathcal{N}_k(\vec{x})} w_i \delta\left(y^i, c\right),$$

სადაც δ არის ერთის ტოლი, როცა y^i ჭდე მიეკუთვნება c კლასს და ნულის ტოლი — წინააღმდეგ შემთხვევაში.

რეგრესიის ამოცანის შემთხვევაში ეს უბრალოდ საშუალო მნიშვნელობის შეწონვის საკითხია :

$$f\left(\vec{x}\right) = \frac{1}{k} \sum_{\vec{x}^i \in \mathcal{N}_k(\vec{x})} w_i y^i .$$

3 მანძილები და მსგავსებები

k უახლოესი მეზობლის ალგორითმის არსებითი ელემენტია *მანძილი*. იგი საშუალებას იძლევა განისაზღვროს, რომელი დაკვირვებებია საწვრთნელ ნაკრებში უახლოესი დაკვირვებები მოსანიშნავ წერტილთან მიმართებაში.

ყურადღება

k უახლოესი მეზობლის ალგორითმი მგრძნობიარეა არარელევანტური (შეუსაბამო) ნიშნების მიმართ, რომლებიც გათვალისწინებული აღმოჩნდება მანძილის გამოთვლისას და შეიძლება დაამახინჯოს იგი.

უფრო მეტიც, დიდი განზომილების პირობებში ეს ალგორითმი *განზომილების წყევლით* (ინგლ. *Curse of dimensionality*) გამოწვეულ *უბედურებაში* აღმოჩნდება (იხ. ქვედანაყოფი 11.1.3) : ყველა მაგალითი ძალიან დაშორებული იქნება დაკვირვებისგან, რომლის მონიშვნას ვცდილობთ და ინტუიცია, რომლის გამოყენებით შეგვეძლო პროგნოზირება ახლომდებარე მაგალითების ჭდეთა საფუძველზე, უკვე არ იმუშავებს.

3.1 მანძილები

გავიხსენოთ აქ მანძილის განსაზღვრება :

განსაზღვრება 5 (მანძილი) მოცემული $\mathcal{X}$ სიმრავლის პირობებში ნებისმიერ $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ ფუნქციას, რომელიც შემდეგ სამ თვისებას აკმაყოფილებს, ეწოდება მანძილი $\mathcal{X}$ -ზე :

1. დაყოფა : $\forall \vec{u}, \vec{v} \in \mathcal{X} \times \mathcal{X}, \;\; d(\vec{u}, \vec{v}) = 0 \Leftrightarrow \vec{u} = \vec{v}$;

2. სიმეტრია : $\forall \vec{u},\vec{v} \in \mathcal{X}\times\mathcal{X},\ \ d(\vec{u},\vec{v}) = d(\vec{v},\vec{u})$;

3. სამკუთხა (სამკუთხედის) უტოლობა : $\forall \vec{t},\vec{u},\vec{v} \in \mathcal{X}^3,\ d(\vec{u},\ \vec{v}) \le d(\vec{u},\ \vec{t}) + d(\vec{t},\ \vec{v})$.

უკანასკნელ პირობას ხშირად მინკოვსკის (გერმ. Hermann Minkowski, 1864-1909) უტოლობასაც უწოდებენ.

იმ შემთხვევაში, როცა $\mathcal{X} = \mathbb{R}^p$, ყველაზე უფრო ხშირად მინკოვსკის მანძილი გამოიყენება :

განსაზღვრება 8.6 (მინკოვსკის მანძილი) როცა მოცემულია ფიქსირებული q , რომელიც $q \ge 1$ პირობას აკმაყოფილებს, მაშინ *მინკოვსკის მანძილი* ეწოდება შემდეგი სახით განსაზღვრულ მანძილს :

$$d_q : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$$

$$\vec{u},\vec{v} \mapsto \left\|\vec{u}-\vec{v}\right\|_q = \left(\sum_{j=1}^{p} \left|u_j - v_j\right|^q \right)^{1/q} .$$

ევკლიდეს მანძილი კერძო შემთხვევაა $q = 2$ მნიშვნელობისათვის :

$$d_2\left(\vec{u},\vec{v}\right) = \sqrt{\sum_{j=1}^{p}\left(u_j - v_j\right)^2} .$$

როცა $q = 1$, მაშინ საუბარი გვიწევს *მანჰეტენის მანძილზე*, ანუ *ტაქსის მძღოლის მანძილზე*, ვინაიდან სიბრტყეზე მანჰეტენის მანძილი პრაქტიკულად ნიშნავს $\vec{u}$ და $\vec{v}$ წერტილებს შორის გავლილი მანძილის გამოანგარიშებას მხოლოდ ღერძულა ხაზების პარალელურად მოძრაობისას, როგორც ეს მოხდებოდა ტაქსის შემთხვევაში მანჰეტენზე, სადაც ქუჩები ქმნის *თითქმის მართკუთხა ბადეს* (ზემოდან დაკვირვებისას) : $d_1\left(\vec{u},\vec{v}\right) = \sum_{j=1}^{p}\left|u_j - v_j\right|$.

დაბოლოს, ℓ_∞ *მანძილი*, რომელსაც ასეთი ნოტაცია შეიძლება მიეცეს ანალოგიით ℓ^∞ მიმდევრობათა სივრცესთან, მაგრამ ზემოთ მოცემულ აღნიშვნებთან ლოგიკური თავსებადობის შესანარჩუნებლად აქ აღინიშნება როგორც d_∞ და იგი წარმოადგენს მაქსიმალურ სხვაობას $\vec{u}$ და $\vec{v}$ მანძილებს შორის ერთ განზომილებაზე :

$$d_\infty\left(\vec{u},\vec{v}\right) = \max_{j=1,\ldots,p}\left|u_j - v_j\right|.$$

ეს მანძილი ასევე ცნობილია როგორც პაფნუტი ჩებიშევის (რუს. Пафнутий Львович Чебышёв, 1821-1894) მანძილი.

თეორემა 8.1 ℓ_∞ *მანძილი მინკოვსკის მანძილის ზღვარს წარმოადგენს, როცა* $q \to \infty$.

მტკიცებულება. დავუშვათ $k = \arg\max_{j=1,\ldots,p}\left|u_j - v_j\right|$. მაშინ :

– დავუშვათ, რომ $|u_k - v_k| = 0$ და, მაშასადამე, $|u_j - v_j| = 0 \ \forall j$, ასეთ შემთხვევაში მინკოვსკის მანძილი $\vec{u}$ და $\vec{v}$ წერტილებს შორის 0-ის ტოლია, q პარამეტრის ნებისმიერი მნიშვნელობისათვის და, ამრიგად $\lim_{q\to\infty} d_q(\vec{u},\vec{v})$;

– დავუშვათ, რომ $|u_k - v_k| > 0$ და მაშინ შესაძლებელი იქნება ჩავწეროთ :

$$d_q(\vec{u},\vec{v}) = |u_k - v_k| \left(\sum_{j=1}^{p} \left(\frac{|u_j - v_j|}{|u_k - v_k|} \right)^q \right)^{1/q} \le |u_j - v_j| \, p^{1/q},$$

ვინაიდან j-ის ყველა მნიშვნელობისათვის ადგილი აქვს $|u_j - v_j| < |u_k - v_k|$ თანაფარდობას ტოლობით არაუმეტეს p შემთხვევისა. მეორე მხრივ,

$$d_q(\vec{u},\vec{v}) = \left(|u_k - v_k|^q + \sum_{j\neq k} |u_j - v_j|^q \right)^{1/q} \ge |u_k - v_k|,$$

რადგან $|u_j - v_j|^q > 0$ ყველა j ინდექსისათვის. ასე რომ,

$$|u_k - v_k| \le d_q(\vec{u},\vec{v}) \le |u_j - v_j| \, p^{1/q}.$$

ვინაიდან $\lim_{q\to\infty} p^{1/q} = 1$, ამიტომ $\lim_{q\to\infty} d_q(\vec{u},\vec{v}) = |u_k - v_k|$.

□

3.2 მსგავსებები რეალურ ვექტორებს შორის

მაგრამ k უახლოესი მეზობლის ალგორითმის განსახორციელებლად სავალდებულო არ არის მანძილის გამოყენება : საკმარისია *მსგავსების* ცნება.

განსაზღვრება 8.7 (მსგავსება) ნებისმიერ $s : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}_+$ ფუნქციას, რომელიც მით უფრო დიდია, რაც უფრო მსგავსია მის მიერ $\mathcal{X}$-ის ერთმანეთთან შესადარებელი ორი ელემენტი, ეწოდება *მსგავსება* $\mathcal{X}$-ზე.

✿

მანძილისგან განსხვავებით, მსგავსებას განსაკუთრებული მათემატიკური თვისებები არ აქვს. ნებისმიერი მანძილი შეიძლება იყოს გარდასახული მსგავსებად, მაგალითად, შემდეგი გარდაქმნებით :

$$s(\vec{u},\vec{v}) = -d(\vec{u},\vec{v})$$

ან

$$s(\vec{u},\vec{v}) = \frac{1}{1+d(\vec{u},\vec{v})}.$$

მსგავსების ცნება, რომელიც ხშირად გამოიყენება, როცა $\mathcal{X} = \mathbb{R}^p$, — ეს *კორელაციის კოეფიციენტია.*

განსაზღვრება 8.8 (კორელაციის კოეფიციენტი) დავუშვათ, რომ მოცემულია ორი $\vec{u},\vec{v} \in \mathbb{R}^p$ წერტილი, მაშინ *კორელაციის კოეფიციენტი,* ანუ *პირსონის* (ინგლ. Karl Pearson, 1857-1936) *კოეფიციენტი* $\vec{u}$ და $\vec{v}$ წერტილებს შორის ეწოდება შემდეგ სიდიდეს :

$$\rho(\vec{u},\vec{v}) = \frac{\sum_{j=1}^{p}\left(u_j - \frac{1}{p}\sum_{k=1}^{p}u_k\right)\left(v_j - \frac{1}{p}\sum_{k=1}^{p}v_k\right)}{\sqrt{\sum_{j=1}^{p}\left(u_j - \frac{1}{p}\sum_{k=1}^{p}u_k\right)^2}\cdot\sqrt{\sum_{j=1}^{p}\left(v_j - \frac{1}{p}\sum_{k=1}^{p}v_k\right)^2}}.$$

ვუწოდებთ $\vec{u}'$ ვექტორს (შესაბამისად $\vec{v}'$ ვექტორს) $\vec{u}$ ვექტორის (შესაბამისად $\vec{v}$ ვექტორის) ცენტრირებული ვერსია, ე.ი. ისეთი ვექტორი, რომელიც მიიღება საშუალო მნიშვნელობის გამოკლებით მისი კოორდინატებიდან :

$$\vec{u}' = \vec{u} - \frac{1}{p}\sum_{j=1}^{p}u_j\,.$$

მაშინ ეს კოეფიციენტი მარტივდება და შემდეგ სახეს იძენს :

$$\rho(\vec{u},\vec{v}) = \frac{\sum_{j=1}^{p}u'_j v'_j}{\sqrt{\sum_{j=1}^{p}\left(u'_j\right)^2}\cdot\sqrt{\sum_{j=1}^{p}\left(v'_j\right)^2}} = \frac{\langle\vec{u}',\vec{v}'\rangle}{\|\vec{u}'\|_2\,\|\vec{v}'\|_2}$$

და, ამრიგად, წარმოადგენს $\vec{u}'$ და $\vec{v}'$ ცენტრირებულ ვექტორებს შორის კუთხის კოსინუსს. ამიტომ ρ კოეფიციენტს ზოგჯერ *კოსინუსის მსგავსებას* უწოდებენ.

თუ $\vec{u}$ და $\vec{v}$ ვექტორები არა მხოლოდ ცენტრირებულია, არამედ ნორმირებულიც, მაშინ კოსინუსური მსგავსება დაიყვანება $\vec{u}$ და $\vec{v}$ ვექტორების სკალარულ ნამრავლამდე. ამ სკალარული ნამრავლის აბსოლუტური მნიშვნელობა მით მეტია რაც უფრო გამოხატულია $\vec{u}$ და $\vec{v}$ ვექტორების კოლინეარაობა (სხვა სიტყვებით, პარალელურობა) და ნულის ტოლია, როცა ეს ვექტორები ორთოგონალურია.

მეათე ლექციაში განხილული იქნება სკალარული ნამრავლის ცნების $\mathbb{R}^p$-ზე გაფართოება *ბირთვის* ცნებამდე $\mathcal{X}$-ზე, რაც *მსგავსების* სხვა *ზომათა* ადვილად განსაზღვრის საშუალებას იძლევა.

შენიშვნა

ზოგად შემთხვევაში კოსინუსური მსგავსება და ევკლიდეს მანძილი არ განსაზღვრავს მეზობლობის ერთსა და იმავე მიდამოებს.

3.3 მსგავსებები სიმრავლეებს შორის

აქამდე ლექციათა ამ კურსში საქმე გვქონდა მხოლოდ ისეთ მონაცემებთან, რომლებიც შეგვეძლო წარმოგვედგინა ნამდვილი ვექტორის სახით. მაგრამ ხშირად ხდება, რომ ჩვენი დაკვირვებების წარმოდგენა უფრო მოხერხებულია ერთისა და იმავე ქვესიმრავლეთა სახით ელემენტების ერთისა და იმავე სასრული სიმრავლიდან. მაგალითად, სიმბოლოების სტრიქონი შეიძლება იყოს წარმოდგენილი მასში მოცემული ასოებით, ხოლო ტექსტური დოკუმენტი — მასში მოთავსებული სიტყვების ნაკრებით.

ეს წარმოდგენები შეიძლება იყოს გარდასახული ორობით (ბინარულ) ვექტორებად : $\mathcal{E}$ სიმრავლის $\mathcal{S}$ ქვესიმრავლე შეიძლება იყოს წარმოდგენილი $|\mathcal{E}|$ ზომის ორობითი ვექტორის სახით, რომლის თითოეული ბიტი შეესაბამება $\mathcal{E}$ სიმრავლის e ელემენტს და აქვს 1-ის ტოლი მნიშვნელობა, თუ $e \in \mathcal{S}$ და 0-ის ტოლი — წინააღმდეგ $e \notin \mathcal{S}$ შემთხვევაში. შემდეგ ამ ვექტორული წარმოდგენების მიმართ შეიძლება მანძილების ან მსგავსებათა განსაზღვრის წინა მეთოდების გამოყენება.

მაგრამ ჩვენ ასევე შეგვიძლია მანძილებისა და მსგავსებათა უშუალო, პირდაპირი განსაზღვრა ამ ნაკრებებზე ვექტორული წარმოდგენის გამოუყენებლად, რომლის განზომილება შეიძლება ძალიან დიდი იყოს (როგორც სიტყვათა რაოდენობა ლექსიკონში, თუ დოკუმენტის წარმოდგენა ამ ლექსიკონში არსებული სიტყვებით ხდება), თუმცა პოტენციურად მწირიც (ვინაიდან დოკუმენტების უმრავლესობაში გამოყენებული იქნება სიტყვათა უმნიშვნელო რაოდენობა ლექსიკონთან შედარებით).

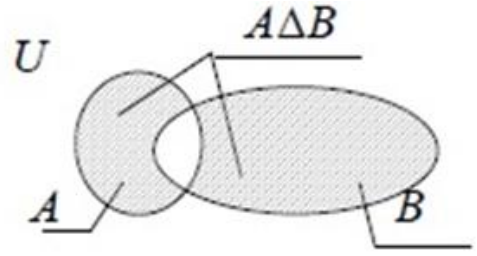
ორი სიმრავლის შედარება შეიძლება ხორციელდებოდეს იმ ელემენტთა რაოდენობის ფუნქციით, რომლებიც მხოლოდ ერთ-ერთ მათგანში გვხვდება.

განსაზღვრება 8.9 (ჰემინგის მანძილი) *ჰემინგის* (*Hamming*) *მანძილი* $\mathcal{E}$ სიმრავლის ორ — $\mathcal{S}$ და $\mathcal{T}$ — ქვესიმრავლეს შორის განისაზღვრება როგორც

$$H(\mathcal{S},\mathcal{T}) = |\mathcal{S} \Delta \mathcal{T}|,$$

სადაც $\mathcal{S} \Delta \mathcal{T} = \{e \in \mathcal{S} \setminus \mathcal{T}\} \cup \{e \in \mathcal{T} \setminus \mathcal{S}\}$ — $\mathcal{S}$ და $\mathcal{T}$ ქვესიმრავლეთა *სიმეტრიული სხვაობაა*, ე.ი. ზუსტად იმ ელემენტების შემცველი სიმრავლე, რომლებიც მიეკუთვნება მხოლოდ ერთ-ერთს ამ ორ სიმრავლიდან : ან $\mathcal{S}$-ს, ან $\mathcal{T}$-ს.

აქვე ქვემოთ, შეხსენების მიზნით, მოცემულია რომელიღაც აბსტრაქტული U სიმრავლის A და B ქვესიმრავლეთა $A \Delta B = (A \setminus B) \cup (B \setminus A)$ *სიმეტრიული სხვაობის* ლოგიკური ოპერაციის მაილუსტრირებელი *გეომეტრიული დიაგრამა* :



თუ გამოვიყენებთ $\mathcal{S}$ და $\mathcal{T}$ ვექტორების ბინარულ წარმოდგენებს, მაშინ ჰემინგის (ინგლ. Richard Hamming, 1915-1998) მანძილი ამ ორ ვექტორს შორის მათი მანჰეტენის მანძილის ეკვივალენტური და განსხვავებული ბიტების რაოდენობის ტოლი იქნება.

ითვლება, რომ ორ სიმრავლეს შორის მსგავსება მით უფრო დიდია, რაც უფრო მეტი საერთო ელემენტი აქვს მათ. მაგრამ, თუ ჩვენ ვადარებთ, სავარაუდოდ, ძალიან განსხვავებული ზომების სიმრავლეებს, მაშინ მხოლოდ იმ ელემენტების რიცხვთა შედარება უნდა მოხდეს, რომლებსაც *რაიმე საერთო შეიძლება გააჩნდეს* : ორ დიდ სიმრავლეს, ბუნებრივია, მეტი საერთო ელემენტი ექნება, ვიდრე ორ მცირე სიმრავლეს. სწორედ ამ პრობლემის დასაძლევად შვეიცარიელი ბოტანიკოსის პოლ ჟაკარის (Paul Jaccard, 1868-1944) მსგავსება გამოიყენება.

განსაზღვრება 8.10 (ჟაკარის მსგავსება) თუ მოცემულია ელემენტთა $\mathcal{E}$ სიმრავლე, მაშინ *ჟაკარის მსგავსება, ტანიმოტოს მსგავსება (Toshio Tanimoto)* ან კიდევ *ჟაკარის ინდექსი* ეწოდება შემდეგ ფუნქციას :

$$J : 2^{\mathcal{E}} \times 2^{\mathcal{E}} \to [0,1]$$
$$\mathcal{S}, \mathcal{T} \mapsto \frac{|\mathcal{S} \cap \mathcal{T}|}{|\mathcal{S} \cup \mathcal{T}|}.$$

აქ $2^{\mathcal{E}}$ აღნიშნავს $\mathcal{E}-$ის ნაწილების სიმრავლეს, სხვანაირად რომ ვთქვათ, თავის ქვესიმრავლეთა სიმრავლეს. ასეთი აღნიშვნა იმიტომ არის არჩეული, რომ, ყოველ შემთხვევაში, სასრული განზომილების შემთხვევაში, $\mathcal{S} \subseteq \mathcal{E}$ ქვესიმრავლე შეიძლება წარმოვადგინოთ როგორც $\mathcal{E}-$ის ბინარული ფუნქცია $\{0,1\}$ სიმრავლეზე, რომელიც აკავშირებს (1) ერთს $\mathcal{E}-$ის იმ ელემენტთან, რომელიც წარმოდგენილია $\mathcal{S}$-ში, ხოლო (0) ნულს — დანარჩენებთან.

თუ ყოველ «სიმრავლეში» ელემენტები შეიძლება რამდენჯერმე (რამდენიმეჯერ) გვხვდებოდეს (ამ შემთხვევაში საუბარია უკვე არა სიმრავლეთა, არამედ მულტისიმრავლეთა შესახებ), მაშინ ელემენტების ჯერადობის გათვალისწინება, მხედველობაში მიღება შესაძლებელია ე.წ. MinMax-მსგავსების საშუალებით :

განსაზღვრება 8.11 (MinMax-მსგავსება) დავუშვათ რომ მოცემულია ელემენტთა $\mathcal{E}$ სიმრავლის $\mathcal{S}$ მულტისიმრავლე და $e \in \mathcal{E}$ ელემენტი. აღვნიშნოთ $m_{\mathcal{S}}(e)$ სიმბოლოთი e ელემენტის ჯერადობა $\mathcal{S}$ მულტისიმრავლეში, ე.ი. $\mathcal{S}$ მულტისიმრავლეში მისი შესვლის რაოდენობა. თუ მოცემულია ელემენტების $\mathcal{E}$ სიმრავლე, მაშინ *MinMax-მსგავსება* $\mathcal{E}$ სიმრავლის ორ — $\mathcal{S}$ და $\mathcal{T}$ — მულტისიმრავლეს შორის ეწოდება ფუნქციას, რომელიც გვიბრუნებს შემდეგ შედეგს :

$$\mathrm{MinMax}(\mathcal{S}, \mathcal{T}) = \frac{\sum_{e \in \mathcal{S} \cap \mathcal{T}} \min\left(m_{\mathcal{S}}(e), m_{\mathcal{T}}(e)\right)}{\sum_{e \in \mathcal{S} \cup \mathcal{T}} \max\left(m_{\mathcal{S}}(e), m_{\mathcal{T}}(e)\right)}.$$

3.4 მსგავსებები კატეგორიალურ მონაცემებს შორის

ცვლადების კიდევ ერთი ტიპი, რომელიც ხშირად გვხვდება, — ეს *კატეგორიალური* ცვლადებია, ვთქვათ, ისეთი, როგორიცაა სქესი ან ადამიანის სოციალურ-პროფესიული კატეგორია, კვირის დღე რომელიმე მოვლენისათვის ან სიმსივნის სტადია. ამ ცვლადების წარმოდგენა რიცხვით კი არ ხდება, არამედ ელემენტით შესაძლებლობათა ჩამონათვალში.

მიუხედავად იმისა, რომ ყოველთვის შეგვიძლია ნებისმიერად მივაბათ თითოეულს ამ შესაძლებლობებიდან რიცხვი და წარმოვადგინოთ ცვლადი ამ რიცხვით, ეს, როგორც წესი, საუკეთესო გადაწყვეტილება არ არის.

უკანასკნელი დაკავშირებულია იმასთან, რომ მეზობელი რიცხვებით წარმოდგენილი ორი კატეგორია სწავლების ალგორითმით განხილული იქნება როგორც უფრო მსგავსი, ვიდრე ორი კატეგორია, როლებიც ასახულია ერთმანეთისაგან მეტად დაშორებული რიცხვებით.

ორ კატეგორიას შორის ყოველთვის როდი არსებობს მკაფიოდ გამოხატული დისტანცია : პროფესია «სახელმწიფო სამსახურის ხელმძღვანელი» უფრო ახლოს დგას პროფესიასთან «ხელოსანი» თუ პროფესიასთან «კომერსანტი» ?

უპირველეს ყოვლისა, აპრიორი შეიძლება ორი შემთხვევის გამოყოფა : ან ორივე დაკვირვება მიეკუთვნება ერთსა და იმავე კატეგორიას, ან ორ სხვადასხვა კატეგორიას.

იმისათვის, რომ შესაძლებელი იყოს უახლოეს მეზობელთა მეთოდის გამოყენება ℓ_q ნორმით განსაზღვრული მანძილით ან რომელიმე პარამეტრული მეთოდი რომელიმე კატეგორიალური ცვლადის მიმართ, ჩვენ ხშირად გამოვიყენებთ ე.წ. *სწრაფ კოდირებას* (ინგლ. *one-hot encoding*).

ხმარობენ ასევე *ერთწერტილოვანი, ერთჯერადი, ცხელი* კოდირების სახელწოდებებსაც, თუმცა ყველა შემთხვევაში იგულისხმება პროცესი, რომლის საშუალებითაც კატეგორიალურ ცვლადებს ეძლევა მანქანური სწავლების ალგორითმებისათვის მისაღები ფორმა.

განსაზღვრება 8.12 (სწრაფი კოდირება) თუ გვაქვს კატეგორიალური ცვლადი, რომელსაც m სხვადასხვა მნიშვნელობის მიღება შეუძლია, მაშინ *სწრაფი კოდირება* (ინგლ. *one-hot encoding*) ეწოდება ამ ცვლადის წარმოდგენას m განზომილების ბინარულ (ორობით) ვექტორად, რომელშიც ყოველი ბიტი შეესაბამება ერთ-ერთს ამ მნიშვნელობებიდან და ეს ბიტი დგება $1-$ ზე, თუ კატეგორიალური ცვლადი იძენს შესაბამის მნიშვნელობას, და $0-$ ზე — წინააღმდეგ შემთხვევაში.

მაგალითი

დავუშვათ, რომ ჩვენი დაკვირვებებია საგაზეთო სტატიები და გვაქვს კატეგორიალური ცვლადი, რომელიც განსაზღვრავს მათ თემატიკას როგორც ერთ-ერთს ჩამონათვალიდან :

«*პოლიტიკა*», «*საზოგადოება*», «*ეკონომიკა*», «*სპორტი*», «*მეცნიერება*», «*კულტურა*».

ამ ცვლადის სწრაფი კოდირება მდგომარეობს მის ჩანაცვლებაში ექვსი ბინარული ცვლადით, რომლებსაც სათანადო მნიშვნელობები ექნება :

$(1,0,0,0,0,0)$, თუ სტატია ეხება პოლიტიკას ; $(0,1,0,0,0,0)$, თუ სტატია ეხება საზოგადოებას ; $(0,0,1,0,0,0)$, თუ სტატია ეხება ეკონომიკას ; $(0,0,0,1,0,0)$, თუ სტატია ეხება სპორტს ; $(0,0,0,0,1,0)$, თუ სტატია ეხება მეცნიერებას ; $(0,0,0,0,0,1)$, თუ იგი ეხება კულტურას.

მას შემდეგ, რაც მონაცემები კოდირებული აღმოჩნდება ასეთი სახით, შესაძლებელი გახდება რეალური ვექტორებისათვის განსაზღვრული მანძილებისა და მსგავსებების გამოყენება.

შენიშვნა

კვირის დღეთა (ან წლის თვეთა) შემთხვევაში კოდირების პრობლემა რიცხვით $1-$ დან $7-$ მდე (ან $1-$ დან $12-$ მდე) მდგომარეობს ამ კატეგორიების ციკლურობაში (წრიულობაში) :

თებერვალი (2), რა თქმა უნდა, უფრო ახლოსაა — დროის მიხედვით — მაისთან $(5-2=3)$, ვიდრე ივლისთან $(7-2=5)$, მაგრამ იგი კიდევ უფრო ახლოსაა დეკემბერთან $(12-2=10)$.

დროითი ასეთი სიახლოვის შესანარჩუნებლად შესაძლებელია ამ სიდიდეთა თანაბარი მანძილით განლაგება $1-$ ის ტოლი რადიუსის წრეწირზე და თითოეულის წარმოდგენა ამ პოზიციის კოსინუსით და სინუსით.

ამრიგად, თებერვალი, მაისი და დეკემბერი აღმოჩნდება წარმოდგენილი, ამ თანამიმდევრობის შესაბამისად, შემდეგი გამოსახულებებით :

$$\left(\cos\left(\frac{2\pi}{6}\right),\sin\left(\frac{2\pi}{6}\right)\right)=\left(\frac{1}{2},\frac{\sqrt{3}}{2}\right),$$

$$\left(\cos\left(\frac{5\pi}{6}\right),\sin\left(\frac{5\pi}{6}\right)\right)=\left(-\frac{\sqrt{3}}{2},\frac{1}{2}\right),$$

$$\left(\cos(2\pi),\sin(2\pi)\right)=(1,0).$$

ევკლიდეს მანძილი თებერვლიდან მაისამდე არის $\sqrt{2}$, ხოლო იგივე პარამეტრი თებერვლიდან დეკემბრამდე 1-ის ტოლია, რაც ამ თვეთა დროში დისტანცირების ლოგიკას შეესაბამება.

4 ერთობლივი (თანამშრომლობითი) ფილტრაცია

ერთობლივი, ანუ *თანამშრომლობითი ფილტრაცია* (ინგლ. *collaborative filtering*) საფუძვლად უდევს ყველა სარეკომენდაციო სისტემას. ასეთია, მაგალითად, Netflix ან Amazon სისტემები, რომლებიც იყენებს ჯგუფის რეიტინგებს რეკომენდაციების შესადგენად ცალკეული მომხმარებლისათვის.

ამ პრობლემის გადაწყვეტის მრავალი ხერხი არსებობს. მაგრამ ერთ-ერთი მათ შორის ეყრდნობა ზემოთ განხილულ პრინციპს და, ბუნებრივია, სწორედ მას დაეთმობა ახლა ყურადღება.

უფრო ზუსტად, მოცემულია ობიექტების (ფილმების, წიგნების, საკანცელარიო საქონლის და

სხვა მსგავსი ობიექტის) $\mathcal{S}$ სიმრავლე და მომხმარებლების $\mathcal{X}$ სიმრავლე, იმავდროულად ვუშვებთ $r:\mathcal{X}\times\mathcal{S}\mapsto\mathbb{R}$ ფუნქციის არსებობას, ისეთის, რომ $r(u,a)$ — ეს u მომხმარებლის მიერ a ობიექტისათვის გამოტანილი რეიტინგია. ეს ფუნქცია ცნობილია ნაწილობრივ იმ აზრით, რომ ყველა მომხმარებელს არ შეუფასებია ყველა ობიექტი.

გარდა ამისა, კიდევ ორი სიდიდე შემოგვაქვს :

1. ობიექტების ნებისმიერი $(a,\ b)\in\mathcal{S}\times\mathcal{S}$ წყვილისათვის $\mathcal{X}$ სიმრავლის $\mathcal{U}(a,\ b)$ ქვესიმრავლე, რომელიც შეიცავს მხოლოდ იმ მომხმარებლებს, რომლებმაც შეაფასა a ობიექტიც და b ობიექტიც ;

2. ნებისმიერი $u\in\mathcal{X}$-თვის $\bar{r}(u)$ საშუალო მნიშვნელობა u-ს მიხედვით.

მაშინ შესაძლებელი ხდება ობიექტებს შორის *მსგავსების* განსაზღვრა *კოსინუსური მსგავსების* საფუძველზე :

$$s(a,b)=\frac{\sum_{u\in\mathcal{U}(a,\ b)}\left(r(u,a)-\bar{r}(u)\right)\left(r(u,b)-\bar{r}(u)\right)}{\sqrt{\sum_{u\in\mathcal{U}(a,\ b)}\left(r(u,a)-\bar{r}(u)\right)^2\cdot\sum_{u\in\mathcal{U}(a,\ b)}\left(r(u,a)-\bar{r}(u)\right)^2}}. \tag{8.1}$$

ახლა იმ ობიექტებს შორის, რომლებიც მონიშნულია u მომხმარებლის მიერ, გარკვეული a ობიექტის k უახლოესი მეზობლისთვის $\mathcal{N}_u^k(a)$ აღნიშვნა შემოვიღოთ.

მაშინ u მომხმარებელს შეიძლება ვურჩიოთ a ობიექტი შემდეგი თანაფარდობის საფუძველზე :

$$f(u,a)=\frac{\sum_{b\in\mathcal{N}_u^k(a)}s(a,b)r(u,b)}{\sum_{b\in\mathcal{N}_u^k(a)}\left|s(a,b)\right|}.$$

5 საკვანძო მომენტები

— k უახლოესი მეზობლის ალგორითმს აქვს ზარმაცი სწავლება : კომპენსაციისათვის წინასწარმეტყველების ალგორითმული დანახარჯი შეიძლება იყოს გაზრდილი, თუ საწვრთნელი ბაზა დიდია.

— წინასწარმეტყველებათა ხარისხი k უახლოესი მეზობლის ალგორითმით დამოკიდებულია, ძირითადად, კარგი მანძილის ან მსგავსების არჩევაზე.

— k უახლოესი მეზობლის ალგორითმი განსაკუთრებით კარგად მუშაობს *მცირე განზომილების* პირობებში :

- დავალების შესრულება მას უფრო სწრაფი აქვს ;
- იგი ნაკლებად არის მიდრეკილი არარელევანტური ცვლადებით გამოწვეული წანაცვლებისადმი (მიკერძოებულობისადმი).
- იგი ნაკლებად არის მიდრეკილი განზომილების ჭირით გამოწვეული ტანჯვისადმი.

დამატებითი ინფორმაცია

• სახეთა სტატისტიკური ამოცნობის პრობლემებისადმი მიძღვნილ Andrew R. Webb (1999) წიგნში მოცემულია რამდენიმე ძალზე საინტერესო მანძილი, რომლებიც შეიძლება იყოს გამოყენებული k უახლოესი მეზობლის ალგორითმის ფარგლებშიც.

• მონაცემთა ისეთი სტრუქტურები, როგორიცაა *kd ხეები* (Bentley, 1957) ან *ჰეშ-ცხრილები*, საწვრთნელი სიმრავლის ისეთნაირად შენახვის საშუალებას იძლევა, რომ უახლოესი მეზობლების პოვნა გაადვილებული იყოს. ამ სტრუქტურათა იდეა მდგომარეობს იმაში, რომ მოხდეს ერთმანეთთან ძალიან ახლოს მყოფი მაგალითების დაჯგუფება.

• ლიტერატურა სარეკომენდაციო სისტემების შესახებ ძალზე მდიდარია. კლასიკიდან მაგალითის სახით პირველ რიგში შეიძლება დავასახელოთ თუნდაც Ricci et al (2016) და Charu C. Aggarwal (2016).

• სტატიები k უახლოესი მეზობლის შესახებ თავმოყრილია ზარმაცი სწავლებისადმი მიძღვნილ სპეციალურ გამოშვებაში Aha (1997).

• სადისკუსიოდ გამოტანილ სტატიაში Klaus Hechenbichler და Klaus Schliep (2004) მოცემულია უახლოესი მეზობლების შეწონვის მეთოდების ღრმა და ძალზე საინტერესო მიმოხილვა.

6 ბიბლიოგრაფია

1. Aggarwal, C. C. (2016). *Recommender Systems: The Textbook*. Springer, 518 pages.

2. Aha, D. W. (1997). Special issue on lazy learning. *Artificial Intelligence Review*, 11(1–5) :7–423.

3. Bentley, J. L. (1957). Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9) :509–517.

4. Hechenbichler, K. and Schliep, K. (2004). Weighted k-nearest-neighbor techniques and ordinal classification. In *SFB* — Sonderforschungsbereich (Collaborative Research Centres, German research projects) 386, Paper 399 (2004), 17 pages : https://epub.ub.uni-muenchen.de/1769/1/paper_399.pdf

5. Ricci, F., Rokach, L., and Shapira, B. (2015). *Recommender Systems Handbook*. Springer, 2 edition, 1003 pages.

6. Webb, A. (1999). *Statistical Pattern Recognition*. Arnold, London, 454 pages.

7 სავარჯიშოები

8.1 ბინარული (ორობითი) კლასიფიკაციისათვის ანიკას აქვს მონაცემთა შემდეგი ნაკრები :

x_1	1	2	2	2	3	3
x_2	2	1	2	3	1	2
x_3	+	+	-	+	-	+

1. აჩვენეთ უახლოესი მეზობლის ალგორითმის გადაწყვეტილების მიღების საზღვარი.

2. აჩვენეთ 3 უახლოესი მეზობლის ალგორითმის გადაწყვეტილების მიღების საზღვარი.

3. რამდენ შეცდომას უშვებს ეს ორი კლასიფიკატორი საწვრთნელ სიმრავლეზე ?

4. როგორ აღნიშნავს (როგორ ჭდეს მიანიჭებს) ეს ორი კლასიფიკატორი $(4;0.5)$ წერტილს ?

8.2 როცა მოცემულია n დაკვირვება p განზომილებაში, როგორია მაშინ უახლოესი მეზობლის ალგორითმის წვრთნის შეცდომა (კლასიფიკაციისას) ?

8.3 თუ არის k უახლოესი მეზობლის ალგორითმის სწავლება (გაწვრთნა) უფრო შრომატევადი, ვიდრე მისი გამოყენება ჭდის მისანიჭებლად ახალი დაკვირვებისათვის ?

8.4 ბაჩო იხილავს k უახლოესი მეზობლის ალგორითმს. მას ეჭვი აქვს, რომ მისი მონაცემები ძალზე დაბინძურებულია, შეიცავს ხმაურს. რა არის საჭირო : რომ მან გაზარდოს k, თუ შეამციროს იგი ?

8.5 ვეკო ახასიათებს მელოდიებს. იგი განიხილავს ცვლადებად 7 ნოტს. ესენია :

do, re, mi, fa, sol, la და si .

სამი — A , B და C — მელოდია, რომელიც მას აინტერესებს, ასეთია :

$$\left.\begin{array}{l} \text{A = si si do re re do si la sol sol la si la sol sol} \\ \text{B = do do do re mi re do mi re re do} \\ \text{C = sol sol si la si la sol sol si si la si} \end{array}\right\}.$$

1. როგორია ჰემინგის მანძილი A-სა და B-ს შორის? A-სა და C-ს შორის? ასევე B-სა და C-ს შორის ?

2. როგორია მსგავსება შვეიცარიელი ბოტანიკოსის პოლ ჟაკარის (Paul Jaccard, 1868-1944) მიხედვით A-სა და B-ს შორის ? A-სა და C-ს შორის ? ასევე B-სა და C-ს შორის ?

3. როგორია MinMax სახის (მინიმაქსური) მსგავსება A-სა და B-ს შორის ? A-სა და C-ს შორის ? ასევე B-სა და C-ს შორის ?

8.6 დურმიშხანს სურს იწინასწარმეტყველოს (გაარკვიოს) ჩაია თუ ყავა სასმელი. მან შემდეგი მონაცემები შეაგროვა :

მოცულობა (mL)	250	100	125	250
კოფეინი (g)	0.025	0.010	0.050	0.100
სასმელი	ჩაი	ჩაი	ყავა	ყავა

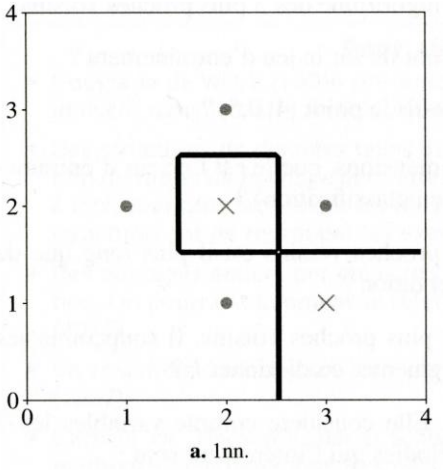
1. უახლოესი მეზობლის ალგორითმის გამოყენებისას ევკლიდეს მანძილით, როგორია ნაწინასწარმეტყველი ჭდე 125 mL მოცულობის სასმელისთვის, რომელიც შეიცავს კოფეინის 0.015 g რაოდენობას ?

2. დურმიშხანის აზრით, ეს კლასიფიკაცია კორექტული არ არის. რისი გაკეთება შეუძლია მას მდგომარეობის გამოსასწორებლად ?

სავარჯიშოთა ამონახსნები

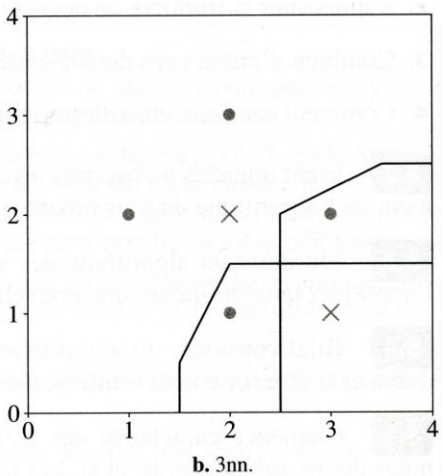
8.1

1.



a. 1nn.

2.



b. 3nn.

3. $k=1$: არავითარი შეცდომა.

$k=3$: ოთხი შეცდომა.

4. $(4, 0.5)$ წერტილის უახლოესია $(3,1)$ წერტილი და იგი უარყოფითია.

შემდეგი ორი უახლოესი წერტილია $(3, 2)$ და $(2,1)$, თანაც ისინი დადებითია ორივე.

$k = 1$ მნიშვნელობისათვის $\hat{y} = +$;

$k = 3$ მნიშვნელობისათვის $\hat{y} = -$.

8.2 0: საწვრთნელ სიმრავლეში დაკვირვების უახლოესი მეზობელი თავად ეს დაკვირვებაა.

8.3 არა, ეს ზარმაცი ალგორითმია და მას არ აქვს წვრთნის (სწავლების) დრო.

8.4 k სიდიდის გაზრდამ უნდა გახადოს ალგორითმი უფრო მედეგი ხმაურის მიმართ.

8.5

1. ჰემინგის მანძილები შვიდბიტიანი წარმოდგენებით (1 თითოეულ ნოტზე/ბგერაზე) ასეთია :

$$A = 1100111;\ B = 1110000;\ C = 0000111.$$

ამიტომ :

$$d(A, B) = 4;\ d(A, C) = 2;\ d(B, C) = 6.$$

2. ჟაკარის მსგავსებები ასეთია :

$$s(A, B) = \frac{2}{6} \approx 0.33;\ s(A, C) = \frac{3}{5} = 0.6;\ s(B, C) = \frac{0}{6} = 0.$$

3. მინიმაქსური მსგავსებები შვიდბაიტიანი (შვიდოქტეტიანი) წარმოდგენებით ასეთია :

$$A = (2, 2, 0, 0, 4, 3, 4);\ B(5, 4, 2, 0, 0, 0, 0);\ C(0, 0, 0, 0, 4, 3, 4).$$

მაშასადამე :

$$s(A, B) = \frac{2+2+0+0+0+0+0}{5+4+2+0+4+3+4} = \frac{4}{22} = \frac{2}{11} \approx 0.18;$$
$$s(A, C) = \frac{0+0+0+0+4+3+4}{2+3+0+0+4+3+4} = \frac{11}{16} \approx 0.69;$$
$$s(B, C) = \frac{0+0+0+0+0+0+0}{5+4+2+0+4+3+4} = 0.$$

8.6

1. დაკვირვება, რომელთანაც ჩვენი სასმელი უახლოესია, $(125, 0.050)$–ს შეადგენს. ამიტომ დურმიშხანი პროგნოზირებს, რომ ეს სასმელი ყავა უნდა იყოს.

2. ეს სასმელი სინამდვილეში უფრო ჰგავს ჩაის — მასში კოფეინის მცირე რაოდენობის გათვალისწინებით. მოცულობა დომინირებს კოფეინის შემცველობაზე. ამიტომ მონაცემების ნორმალიზება უნდა მოხდეს.

ლექცია 9 ხეები და ტყეები

შინაარსი

უახლოესი მეზობლის ალგორითმი გამოიყენება მეტრიკული არაპარამეტრული მოდელების ასაგებად, ე.ი. ისინი ეფუძნება შესაბამისი მანძილის ან მსგავსების განსაზღვრას დაკვირვებებს შორის. ეს ყოველთვის ადვილი როდია და ამიტომ *გადაწყვეტილების ხეები* უდგება ამ პრობლემას სხვანაირად. ამ *არამეტრიკულ*, *იერარქიულ* და *არაპარამეტრულ* ობიექტებს საინტერესო თვისებები ახასიათებს, განსაკუთრებით, როცა ლაპარაკია დისკრეტულ ნიშნებზე. მაგრამ, როგორც წესი, ისინი ცუდად იწვრთნება და მათ განზოგადების ძალიან სუსტად გამოხატული უნარი აქვს.

ამ ლექციაში, ხეების თვისებათა და მათი აგებულობის უფრო დაწვრილებითი შესწავლის შემდეგ, ჩვენ დავინახავთ, როგორ *გავაერთიანოთ* ისინი უფრო მძლავრი მოდელების შესაქმნელად. ეს მოდელები ცნობილია *ანსამბლურ*, ანუ *ჯგუფურ* მოდელებად, რომლებიც ბევრად მეტი მნიშვნელობისაა, ვიდრე მათი შემადგენელი ნაწილების უბრალო ჯამი.

მიზნები

- გადაწყვეტილების მიმღები ხის უპირატესობებისა და ნაკლოვანებების დაწვრილებითი აღწერა.
- გადაწყვეტილების მიმღების ხის აგება.
- სუსტი მოსწავლეების პარალელურად ან მიმდევრობით გაერთიანების შესახებ ცოდნის მიღება და ამ პროცედურისადმი არსებული ინტერესის მიზეზთა გაგება.

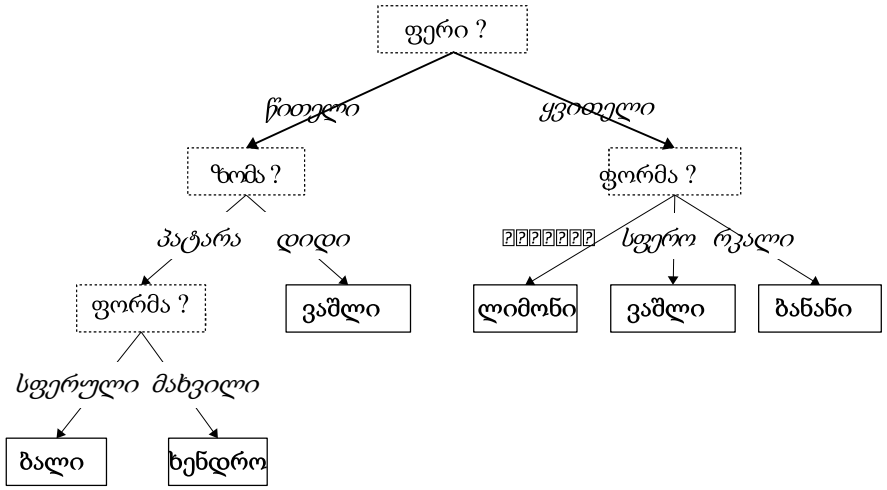
1 გადაწყვეტილების ხეები

1.1 იერარქიული სწავლება

აქამდის განხილულ მოდელებში $\vec{x}$ დაკვირვების მოსანიშნავად მხოლდ ერთ ოპერაციას მიმართავენ. ასეთი მოდელები იყენებს ცვლადების ერთსა და იმავე ნაკრებს ყველა კლასისთვის და ანიჭებს მათ ერთნაირ მნიშვნელოვნობას ყველა დაკვირვებისათვის. ეს მოდელები ცუდად მიესადაგება იმ შემთხვევებს, როცა კლასებს აქვს მულტიმოდალური

განაწილება, ე.ი. როცა კლასები განისაზღვრება განსხვავებული ცვლადებით სივრცის სხვადასხვა არეში ან მაშინ, როცა ცვლადები დისკრეტულია.

მათგან განსხვავებით, გადაწყვეტილების ხეები წარმოადგენს *იერარქიულ მოდელებს*, რომლებიც მოქმედებს როგორც პირობითი ტესტების მიმდევრობითი სერია, რომელშიც ყოველი ტესტი დამოკიდებულია თავის წინამორბედებზე. ისინი ფართოდ გამოიყენება მანქანური სწავლების სამყაროს მიღმაც, მაგალითად, დიაგნოზის დასმის ეტაპების აღსაწერად, ექიმის მიერ მკურნალობის სტრატეგიის ასარჩევად ან შესაძლო გზების საპოვნელად «წიგნში, რომლის გმირი თქვენ ხართ». ნახატზე 9.1 ნაჩვენებია გადაწყვეტილების მიღების ერთი ასეთი ხე.



ნახატი 9.1 - გადაწყვეტილების ხის მაგალითი ხილის მოსანიშნავად.

ნახატი 9.1 გადაწყვეტილების ხეთა სამი თვისების ილუსტრირების საშუალებას გვაძლევს :

— ისინი დისკრეტული ატრიბუტების (ვთქვათ, ფორმის, ზომის და ფერის) დამუშავების საშუალებას იძლევა ამ ატრიბუტებისათვის (ნიშნებისათვის) მსგავსების ან მოწესრიგებულობის კონცეფციის მოუთხოვნელად (ეს ცნობილია როგორც *არამეტრიკული სწავლება*) ;

— ისინი მრავალკლასიანი კლასიფიკაციის ამოცანის დამუშავების საშუალებას იძლევა ბინარული (ორობითი) კლასიფიკაციების გამოუყენებლად ;

— ისინი შეიძლება იყოს გამოყენებული სამუშაოდ მულტიმოდალურ კლასებთან (როგორც მოცემულ შემთხვევაში ჭდისათვის «ვაშლი», რომელიც მინიჭებული აქვს მოზრდილ წითელ ნაყოფს ან სფერულ ყვითელ ხილს).

განსაზღვრება 9.1 (გადაწყვეტილების ხე) პროგნოზირების (წინასწარმეტყველების) მოდელს, რომელიც შეიძლება იყოს წარმოდგენილი ხის სახით, გადაწყვეტილების (მიღების) ხე ეწოდება. ხის ყოველი კვანძი ამოწმებს ცვლადის პირობას, ხოლო მისი შვილობილი კვანძებიდან თითოეული შვილი შეესაბამება შესაძლო პასუხს ამ პირობაზე. ხის ფოთლები შეესაბამება ჭდეს.

რომელიმე დაკვირვების ჭდის წინასწარმეტყველებისათვის კვალდაკვალ «მიჰყვებიან» ხის ფესვიდან ტესტზე გაცემულ პასუხებს და აბრუნებენ მიღწეული ფოთლის ჭდეს.

1.2 სივრცის დაყოფა გადაწყვეტილების ხით

გადაწყვეტილების ხე ყოფს დაკვირვებათა $\mathcal{X}$ სივრცეს იმდენ არედ, რამდენიც მას ფოთოლი აქვს ; ერთსა და იმავე არეში ყველა დაკვირვება ერთსა და იმავე ჭდეს იძენს. კლასიფიკაციის ამოცანის შემთხვევაში ეს ჭდე წარმოადგენს უმრავლესობის ჭდეს ამ არეში.

თუ დავუშვებთ, რომ $\mathcal{X}$ სივრცეში გვაქვს n რაოდენობის $\vec{x}^1, \vec{x}^2, \ldots, \vec{x}^n$ დაკვირვება, რომლებიც მონიშნულია $y^1, y^2, \ldots, y^n$ სახით, და ასევე R რაოდენობის $R_1, R_2, \ldots, R_R$ არე, მაშინ შეგვიძლია ჩავწეროთ :

$$f(\vec{x}) = \sum_{r=1}^{R} \delta_{\vec{x} \in R_r} \arg\max_{c=1,\ldots C} \sum_{i:\vec{x}^i \in R_r} \delta(y^i, c). \tag{9.1}$$

მოცემულ კურსში ჩვენ არ ვაპირებთ ბინარულ და მრავალკლასიან შემთხვევათა ცალ-ცალკე განხილვას, რადგან პირველი გამომდინარეობს მეორედან, თუ აღვნიშნავთ ამ ორ კლასს ციფრებით 1 და 2 , და არა 0 -ით და 1-ით, როგორც ამას, ჩვეულებრივ, ვაკეთებთ.

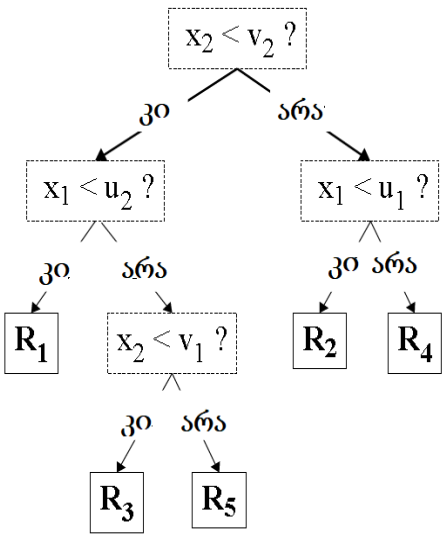
რეგრესიის ამოცანისათვის ეს ჭდე წარმოადგენს მოცემულ არეში დაკვირვებების საშუალო ჭდეს :

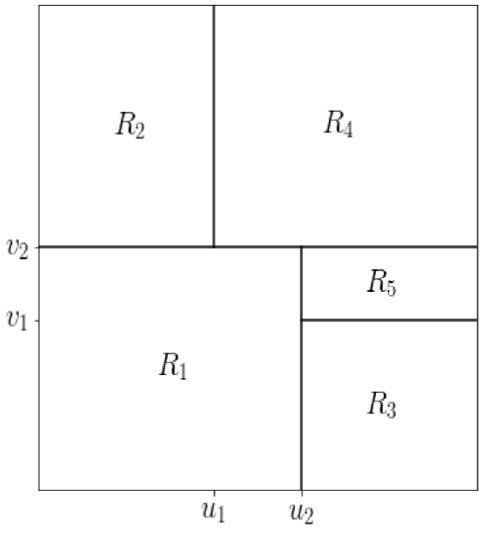
$$f(\vec{x}) = \sum_{r=1}^{R} \delta_{\vec{x} \in R_r} \frac{1}{|R_r|} \sum_{i:\vec{x}^i \in R_r} y^i . \tag{9.2}$$

ეს აღნიშვნები მოუხერხებელია.

სასურველია არ ვეცადოთ მოცემული ხისათვის *გადაწყვეტილების მიღების ფუნქციის* ჩაწერა ანალიზური ფორმით.

მაგრამ მნიშვნელოვანია გვახსოვდეს, რომ გადაწყვეტილების ხე რეკურსიულად ყოფს საწვრთნელ სიმრავლეს სულ უფრო და უფრო მცირე ქვესიმრავლეებად. ასეთი დაყოფა ნაჩვენებია ნახატზე 9.2 ნამდვილი ცვლადებისათვის.





ნახატი 9.2 - გადაწყვეტილების ხე (მარცხნივ) ყოფს $\mathbb{R}^2$ სივრცეს 5 ზონად (მარჯვნივ)

2. ხის აგების სტრატეგია

2.1 CART ალგორითმი გადაწყვეტილების ხის გასაწვრთნელად

გადაწყვეტილების ხის დასასწავლად გამოსაყენებელ ალგორითმს CART აკრონიმით აღნიშნავენ, რომელიც ნაწარმოებია ინგლისური ფრაზიდან *Classification And Regression Tree* — კლასიფიკაციისა და რეგრესიის ხე (Breiman et al., 1984).

ეს არის სივრცის დაყოფის ალგორითმი ხარბი, რეკურსიული და გამყოფი მიდგომის გამოყენებით.

CART ალგორითმი შეიძლება იყოს გამოყენებული *ბინარული გადაწყვეტილების ხის* გასაწვრთნელად — ე.ი. ისეთი ხის, რომელშიც ყოველ კვანძს აქვს ზუსტად ორი შვილობილი — ზოგადობის დაუკარგავად. მართლაც, ნებისმიერი ხე შეიძლება იყოს ხელახლა წარმოდგენილი ბინარული ხით, როგორც ეს ნაჩვენებია ნახატზე 9.3.

როგორც ნაჩვენებია ნახატზე 9.2, CART ალგორითმი ახორციელებს მონაცემების *ერთდროულ* დაყოფას *სათითაოდ,* რაც ქმნის გადაწყვეტილების მიღების საზღვარს, რომელიც ღერძების ორთოგონალურია.

შემდგომში ჩვენ ვვარაუდობთ, რომ მონაცემები განსაზღვრულია p განზომილების $\mathcal{X}$ სივრცეში.

განსაზღვრება 9.2 (მაცალკევებელი ცვლადი) CART ალგორითმით აგებული გადაწყვეტილების ხის ყოველ კვანძთან ასოცირებულია *მაცალკევებელი ცვლადი* (ინგლ. *splitting variable*) $j \in \{1,\ldots,p\}$, რომლის შესაბამისად მოხდება მონაცემების დაყოფა.

ეს ცვლადი მოცემული კვანძის *შვილობილთა* შესაბამის ორ არეს განსაზღვრავს.

იმ შემთხვევაში, როცა მაცალკევებელი ცვლადი *ბინარული* ცვლადია, მაშინ ხსენებული ორი არე ასეთია :

$$R_l(j,s) = \{\vec{x} : x_j = 0\};$$

$$R_r(j,s) = \{\vec{x} : x_j = 1\}.$$

თუ მაცალკევებელი ცვლადი *დისკრეტულია* ცვლადია, რომელსაც ორზე მეტი მნიშვნელობის (ან მოდალობის) მიღება შეუძლია, მაშინ მას თან ახლავს ამ მნიშვნელობათა

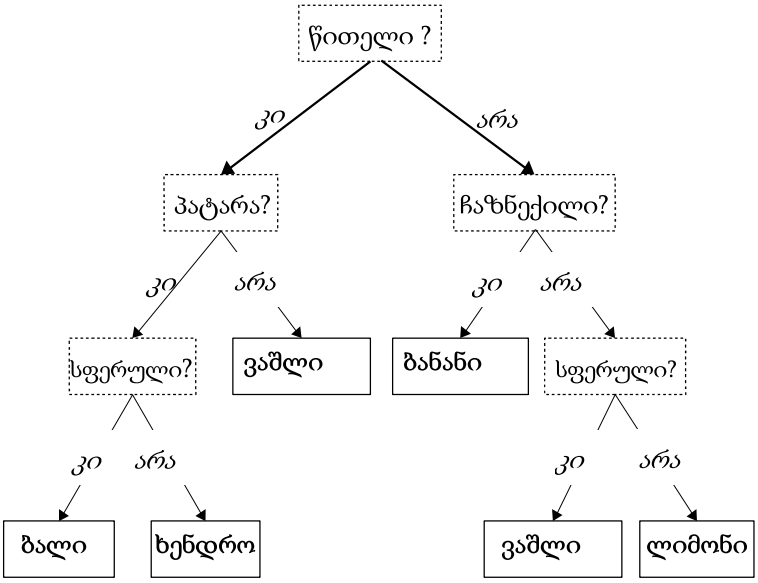
$$\mathcal{S} \subset \mathrm{dom}(x_j)$$

ქვესიმრავლე.

ამ შემთხვევაში ორი არე ასეთია :

$$R_l\left(j,s\right)=\left\{\vec{x}:x_j\in\mathcal{S}\right\};$$

$$R_r\left(j,s\right)=\left\{\vec{x}:x_j\notin\mathcal{S}\right\}.$$



ნახატი 9.3 - ნახატზე 9.1 მოცემული გადაწყვეტილების ხის ბინარული ვერსია.

დაბოლოს, თუ მაცალკევებელი ცვლადი *ნამდვილია*, მაშინ მას ახლავს s *მაცალკევებელი წერტილი* (ინგლ. *splitting point*), რომელიც წარმოადგენს ატრიბუტის — მთავარი დამახასიათებელი ნიშნის — მნიშვნელობას, რომლის მიხედვით იქნება მიღებული გადაწყვეტილება.

ამ შემთხვევაში ორი არე ასეთია :

$$R_l\left(j,s\right)=\left\{\vec{x}:x_j<s\right\};$$

$$R_r\left(j,s\right)=\left\{\vec{x}:x_j\geq s\right\}.$$

CART ალგორითმის თითოეულ იტერაციაზე ხდება j ცვლადის ყველა შესაძლო მნიშვნელობის გადარჩევა და, აუცილებლობის შემთხვევაში, s-ის ან $\mathcal{S}$-ის ყველა შესაძლო მნიშვნელობათა გადარჩევაც წინასწარ მოცემული *კრიტერიუმის* მინიმუმის უზრუნველმყოფი $\left(j,s\right)$ წყვილის დასადგენად.

შენიშვნა

რომელიღაც უწყვეტი x_j ცვლადის შემთხვევაში, თუ ჩვენ ვვარაუდობთ ამ ცვლადის მიერ $\mathcal{D}$ ნაკრებზე მისაღებ მნიშვნელობებს ანუ $x_j^1\leq x_j^2\leq\ldots,\leq x_j^n$ მონაცემებს, მაშინ s-ის შესაძლო მნიშვნელობებს წარმოადგენს $\frac{x_j^{i+1}-x_j^i}{2}$ სიდიდეები i ინდექსის ყველა ისეთი მნიშვნელობისათვის, სადაც $x_j^{i+1}\neq x_j^i$.

რეგრესიის ამოცანის შემთხვევაში, როცა კრიტერიუმს, რომლის მინიმიზირება არის აუცილებელი, ჩვეულებრივ, წარმოადგენს საშუალო კვადრატული შეცდომა, მთავარი საკითხი შემდეგში მდგომარეობს : ζ ნაწილად ($\zeta = j$, ან $\zeta = (j, \mathcal{S})$, ან $\zeta = (j, s)$ პირობებში) დაყოფისას რომელ წერტილზე შეიძლება ახლობელ ჭდეთა მქონე დაკვირვებების დაჯგუფება ? სწორედ ამიტომ *დაყოფის ცვლადისა* და *წერტილის* არჩევა ასეთნაირად ხდება :

$$\arg\min_{j,s}\left(\sum_{i:\vec{x}^i\in R_l(\zeta)}\left(y^i - y_l(\zeta)\right)^2 + \sum_{i:\vec{x}^i\in R_r(\zeta)}\left(y^i - y_r(\zeta)\right)^2\right),$$

სადაც $y_l(\zeta)$ (შესაბამისად $y_r(\zeta)$) — ეს $R_l(\zeta)$ (შესაბამისად $R_r(\zeta)$) არესთან ასოცირებული (დაკავშირებული) ჭდეა მოცემულ ეტაპზე, ე.ი ამ არეში წარმოდგენილ ჭდეთა საშუალო მნიშვნელობა.

კლასიფიკაციის ამოცანის შემთხვევაში იყენებენ არა საშუალო კვადრატულ შეცდომას, არამედ *მინარევის კრიტერიუმს* (ინგლ. *impurity criterion*). ასეთი დასახელება დაკავშირებულია იმასთან, რომ ეს კრიტერიუმი რაოდენობრივად განსაზღვრავს მოცემული არის იმ კლასთა ელემენტებით «*დაბინძურების*» ხარისხს, რომლებიც უმრავლესობას არ შეადგენს. თუ ხსენებულ კრიტერიუმს Imp სიმბოლოთი აღვნიშნავთ, (ინგლისურიდან : **Imp**urity — მინარევი, დაჭუჭყიანება, დაბინძურება, დანაგვიანება) შეიძლება *დაყოფის ცვლადისა და წერტილის* არჩევა შემდეგი სახით :

$$\arg\min_{\zeta}\left(\frac{|R_l(\zeta)|}{n}\operatorname{Imp}\left(R_l(\zeta)\right) + \frac{|R_r(\zeta)|}{n}\operatorname{Imp}\left(R_r(\zeta)\right)\right).$$

ეს კი კვლავ ხარბი ალგორითმია : გარანტია არ არის, რომ ამ სტრატეგიის გამოყენების შედეგად მიღებული იქნება გადაწყვეტილების ხე მინიმალური მინარევით ან საშუალო კვადრატული შეცდომით.

2.2 მინარევების კრიტერიუმები

მანქანურ სწავლებაში მინარევი განიხილება როგორც სხვა კლასთა არსებობა გადაწყვეტილებათა ხის სიმრავლეში. არსებობს მინარევების რამდენიმე კრიტერიუმი, რომლებსაც ჩვენ დაწვრილებით განვიხილავთ ამ ნაწილში : *კლასიფიკაციის შეცდომა, კროს-ენტროპია (ჯვარედინი ენტროპია)* და *ჯინის* (*Gini*) *მინარევი.*

უკანასკნელი კრიტერიუმი დაკავშირებულია ცნობილი იტალიელი სტატისტიკოსის, დემოგრაფის, ეთნოლოგის, სოციოლოგის და ფაშიზმის ერთ-ერთი წამყვანი თეორეტიკოსის, «*ფაშიზმის მეცნიერული საფუძვლები*»[2] (1927) ტრაქტატის ავტორის — კორადო ჯინის (Corrado Gini, 1884-1965) — სახელთან. ამ კრიტერიუმების განსაზღვრისათვის ჩვენ გამოვიყენებთ

[2] Corrado Gini The Scientific Basis of Fascism. Source: Political Science Quarterly, Vol. 42, No. 1 (Mar., 1927), pp. 99-115 (17 pages). Published by: Oxford University Press. https://www.jstor.org/stable/2142862

$p_c(R)$ ნოტაციას R არეში c კლასს მიკუთვნებული საწვრთნელი მაგალითების წილის აღსანიშნავად :

$$p_c(R) = \frac{1}{|R|} \sum_{i:\vec{x}^i \in R} \delta(y^i, c).$$

ჯერ ერთი, *კლასიფიკაციის შეცდომა* გამოიყენება R არის დაბინძურების შესაფასებლად ამ არიდან აღებული ისეთი მაგალითების *წილის* სახით, რომლებიც უმრავლესობის კლასს არ მიეკუთვნება :

$$\mathrm{Imp}(R) = 1 - \max_{c=1,\ldots,C} p_c(R). \tag{9.3}$$

თუ ყველა მაგალითი არეში მიეკუთვნება ერთსა და იმავე კლასს, მაშინ კლასიფიკაციის შეცდომა ამ არისათვის 0-ს უდრის ; და, პირიქით, თუ არე შეიცავს მაგალითების იმავე რაოდენობას, რაც თითოეულ კლასშია C კლასთა ერთობლიობიდან, მაშინ $p_c(R) \approx \frac{1}{C}$ ნებისმიერი c კლასისთვის, ხოლო კლასიფიკაციის შეცდომა $1 - \frac{1}{C}$ სიდიდეს უდრის ან $\frac{1}{2}$ -ის მნიშვნელობისაა ბინარული კლასიფიკაციის შემთხვევაში.

შემდეგ, რაც შეეხება *კროს-ენტროპიას*, ანუ *ჯვარედინ ენტროპიას*, იგი გამოიყენება R არის გაბინძურების ხარისხის დასადგენად ისეთი დაყოფის არჩევით, რომელიც ინფორმაციული მოგების მაქსიმიზაციას ახდენს : აგების მიზანია იმ დამატებითი ინფორმაციის რაოდენობის მინიმიზაცია, რომელიც აუცილებელია R-ის საწვრთნელი მაგალითების კორექტული მონიშვნისათვის :

$$\mathrm{Imp}(R) = -\sum_{c=1}^{C} p_c(R) \log_2 p_c(R). \tag{9.4}$$

თუ ყველა მაგალითი არეში მიეკუთვნება ერთსა და იმავე კლასს, მაშინ ამ არის კროს-ენტროპია 0-ს უდრის ; და, პირიქით, თუ არე შეიცავს თითოეული კლასის ერთისა და იმივე რაოდენობის მაგალითებს C კლასთა ნაკრებიდან, მაშინ კროს-ენტროპია $\log_2(C)$-ის ტოლია, ან 1-ის ტოლია ბინარული კლასიფიკაციის შემთხვევაში.

დაბოლოს, მინარევის ყველაზე უფრო ფართოდ გავრცელებული განსაზღვრებაა *ჯინის მინარევი*, რომელიც რაოდენობრივად აფასებს იმის ალბათობას, რომ მაგალითი საწვრთნელი სიმრავლიდან არაკორექტულად იქნება მონიშნული, თუ მისი მარკირება მოხდება შემთხვევითად R არეში განაწილების ფუნქციის შესაბამისად.

განსაზღვრება 9.3 (ჯინის მინარევი) გარკვეულ R არეში *ჯინის მინარევი* შემდეგი სახით განისაზღვრება :

$$\mathrm{Imp}(R) = \sum_{c=1}^{C} p_c(R)\left(1 - p_c(R)\right).$$

თუ ყველა მაგალითი არეში მიეკუთვნება ერთსა და იმავე კლასს, მაშინ ჯინის მინარევი ამ არისათვის 0 -ს უდრის ; და, პირიქით, თუ არე შეიცავს თითოეული კლასის ერთისა და იმივე რაოდენობის მაგალითებს C კლასთა ნაკრებიდან, მაშინ ჯინის მინარევი უდრის $1-\frac{1}{C}$ მნიშვნელობას, ან $\frac{1}{2}$ მნიშვნელობას ბინარული კლასიფიკაციის შემთხვევაში.

2.3 ხის გასხვლა

ჩვენ წარმოვადგინეთ ხის აგების სტრატეგია. ახლა უნდა შევძლოთ გადაწყვეტილების მიღება, როდის გავჩერდეთ. თუ ხე ღრმა არ არის, სავსებით მოსალოდნელია, იგი ვერ მოახერხებს პრობლემის კორექტულ ჩამოყალიბებას, მაშინ როცა ძალიან ღრმა ხე, სავარაუდოა, ზედმეტად ნასწავლი აღმოჩნდება.

კლასიკური მიდგომა ამ პრობლემის გადაჭრისადმი მდგომარეობს იმაში, რომ შეწყდეს არის დაყოფა, როცა იგი შეიცავს საწვრთნელი მაგალითების წინასწარ დადგენილ მინიმალურ რაოდენობას. ეს საშუალებას იძლევა თავიდან ავიცილოთ მხოლოდ ერთი წერტილის შემცველი ძალზე თავისებური ფოთლების აგება. ასევე შესაძლებელია ხის სიღრმის შეზღუდვა.

ასევე შეიძლება *რეგულარიზაციის* მეთოდის გამოყენება (იხ. ლექცია 6) ხის სირთულის სამართავად, რომელიც იზომება მის მიერ განსაზღვრულ არეთა რაოდენობით. ეს ცნობილია როგორც *გასხვლა სირთულეზე გაწეული ხარჯების კრიტერიუმით* (ინგლ. *cost-complexity pruning*) : T ხის სირთულეზე გაწეული ხარჯები მოიცემა შემდეგი თანაფარდობით :

$$C_{\lambda}(T)=\sum_{r=1}^{|T|} n_r \mathrm{Imp}(R_r)+\lambda|T|, \tag{9.5}$$

სადაც $|T|$ არის T ხით განსაზღვრულ არეთა რაოდენობა, R_r — $r-$ული არეა არეთა ამ რიცხვიდან, n_r — ამ არეში განთავსებული საწვრთნელი მაგალითების რიცხვი, $\lambda>0$ — ჰიპერპარამეტრია, რომელიც არეგულირებს შეცდომისა და სირთულის ზომის მნიშვნელოვნობას. მოცემული პროცედურა მოითხოვს ხის მთლიანად გაფართოებას, ხოლო შემდეგ მის გასხვლას გარკვეულ არეთა მიმდევრობითი დაჯგუფებით, სანამ კრიტერიუმი ოპტიმალური არ გახდება.

სამწუხაროდ, გადაწყვეტილების ხეები, როგორც წესი, ძალზე მარტივ მოდელებს ქმნის, რომელთა წინასწარმეტყველების ხარისხის მაჩვენებლები ოდნავაც არ აღემატება მონაცემებში ცვლილებების მიმართ არც თუ ისე მედეგი შემთხვევითი მოდელების მახასიათებლებს. ასეთ მოდელებს *სუსტ მოსწავლეებს* (ინგლ. weak learners) უწოდებენ. საბედნიეროდ, ამ ვითარების გამოსწორება ადვილად შეიძლება სწავლების *ანსამბლური* (*ჯგუფური*) მეთოდების დახმარებით. ანსამბლური სწავლება — მანქანური სწავლების მეთოდია, რომელიც იყენებს რამდენიმე დანასწავლ, გაწვრთნილ ალგორითმს წინასწარმეტყველების უკეთესი მაჩვენებლის მისაღწევად, ვიდრე ამის მიღების შესაძლებლობა გვექნებოდა თითოეული ალგორითმისგან ცალცალკე.

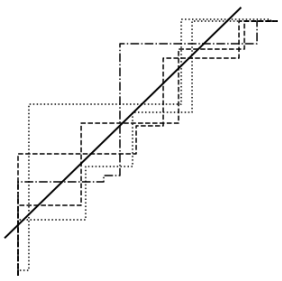
3 ანსამბლური (ჯგუფური) მეთოდები : ბრბოს სიბრძნე

ამ თემასთან დაკავშირებით იხილეთ: https://habr.com/ru/company/ods/blog/324402/

პრაქტიკაში ძალიან მძლავრი აღმოჩნდა ე.წ. ანსამბლური მეთოდები, რომლებიც ეყრდნობა იმ იდეას, რომ დიდი რაოდენობის სუსტი მოსწავლეების გაერთიანება საშუალებას იძლევა ისეთი შედეგის მიღებას, რომელიც მნიშვნელოვნად აღემატება მათ პერსონალურ შედეგს, რამდენადაც სუსტი მოსწავლეების შეცდომები აკომპენსირებს ერთმანეთს. ეს იდეა მოგვაგონებს *ბრძენი რბოს* (ინგლ. *wisdom of crowd*) კონცეფციას : თუ ჩვენს სტუდენტებს დავუსვამთ შეკითხვას ზვიად გამსახურდიას დაბადების წელზე, დიდი ალბათობით, მათი პასუხების საშუალო მნიშვნელობა საკმაოდ ახლოს იქნება სწორთან — 1939 წელთან ; მაგრამ, თუ ამ თარიღზე შეკითხვას შემთხვევითად არჩეულ რომელიღაც ერთ ადამიანს დავუსვამთ, ჩვენ არ გვექნება საშუალება აპრიორი გავიგოთ, იცის თუ არა მან ეს თარიღი ან შემთხვევითად პასუხობს.

მაგალითი

ამ კონცეფციის საილუსტრაციოდ, წარმოვიდგინოთ კლასიფიკაციის ორგანზომილებიანი ამოცანა, რომელშიც ორი კლასი გაყოფილია დიაგონალით, მაგრამ სწავლების ერთადერთ ალგორითმს, რომლის შესახებ ჩვენ ვსაუბრობდით, შეუძლია გადაწყვეტილების მიღების საფეხურების მხოლოდ შეზღუდული რაოდენობის მქონე კიბისებრ საზღვართან დაკავშირებით ჩაატაროს მოსწავლეთა წვრთნა. გადაწყვეტილების მიღების ათობით და ასობით ასეთი კიბისებრი საზღვრის გაერთიანებამ შეიძლება გაცილებით უკეთესი მიახლოება მოგვცეს გადაწყვეტილების მიღების ჭეშმარიტ საზღვართან. ეს მაგალითი ილუსტრირებულია ნახატზე. 9.4.



ნახატი 9.4 - გადაწყვეტილების მიღების ყოველი «კიბისებრი», «საფეხურებიანი» საზღვარი უწყვეტი დიაგონალით წარმოდგენილი ჭეშმარიტი საზღვრის ცუდი მიახლოებაა. მაგრამ ამ კიბეთა გაერთიანება უკეთეს მიახლოებას იძლევა დიაგონალთან.

ყურადღება

ანსამბლური მეთოდების თეორია ამტკიცებს, რომ, თუ გასაერთიანებელ მოდელებს სუსტი სწავლება აქვს ჩატარებული, ესე იგი მარტივი წვრთნებით და არასაკმარისი ეფექტურობით, მაშინ ეს მეთოდები ფუნქციონირების ხარისხის მაჩვენებლის გაუმჯობესების საშუალებას იძლევა ასეთივე მაჩვენებელთან შედარებით, რომელიც კი აქვს საუკეთესო ნიმუშს ხსენებულ მოდელებს შორის. პრაქტიკაში, თუ ცალკეული მოდელები უკვე კარგად მუშაობს და მდგრადია ხმაურის მიმართ, მაშინ ანსამბლური მოდელი შეიძლება არც კი აღმოჩნდეს რომელიმე ცალკეულ მოდელზე უკეთესი. და კიდევ ერთი იმ ფაქტის კონსტანტაციაა საჭირო,

რომ გადაწყვეტილებათა მიღების ხეები, უფრო ხშირად გამოიყენება მხოლოდ განმარტოებული (განმხოლოებული) მოდელებისათვის.

მონაცემთა ერთადერთი ნაკრების საფუძველზე *რამდენიმე* სუსტი მოდელის შექმნისადმი მიდგომათა ორი ძირითადი ოჯახი არსებობს. ორივე მიდგომა იყენებს *მონაცემთა ხელახალი ანარჩევის* (ინგლ. *data set resampling*) შექმნას, მაგრამ კონცეპტუალურად ძალიან განსხვავდება ერთმანეთისგან.

პირველი მიდგომა — *bagging* — წარმოადგენს *პარალელურ* მეთოდს, რომელშიც მოდელის სუსტი მოსწავლეები დამოუკიდებელია ერთმანეთისგან, ხოლო მეორე მიდგომა — *boosting* — წარმოადგენს მიმდევრობით მეთოდს, რომელშიც ყოველი ახალი მოსწავლე აიგება წინა მოსწავლის მუშაობის შედეგების საფუძველზე.

3.1 პარალელური მეთოდები : ბეგინგი

დავუშვათ, გვაქვს მონაცემების $\mathcal{D}$ ნაკრები, რომელიც შედგენილია n რაოდენობის $\vec{x}^1, \vec{x}^2, \ldots, \vec{x}^n$ დაკვირვებით $\mathcal{X}$ -დან და ისინი მონიშნულია $y^1, y^2, \ldots, y^n$ ჭდეებით. *ბეგინგი*, (ინგლ. *bagging* — **B**ootstrap **AGG**regat**ING** — საწყისი ჩატვირთვის აგრეგირება), რომელიც შემოთავაზებული იყო (Breiman, 1996) ნაშრომში გამოჩენილი ამერიკელი სტატისტიკოსის ლეო ბრეიმანის მიერ (Leo Breiman, 1928–2005), მდგომარეობს მონაცემთა $\mathcal{D}$ ნაკრების B სხვადასხვა ვერსიის ფორმირებაში ე.წ. *ბუტსტრეპ-ანარჩევის* (ინგლ. bootstrap sampling) გამოყენებით მიმდინარეობს (იხ. ქვედანაყოფი 3.1.4). ყოველი სუსტი მოდელის სწავლება (წვრთნა) ხორციელდება ერთ-ერთ ანარჩევზე, რაც შეიძლება პარალელურად მიმდინარეობდეს. შემდეგ B რაოდენობით მიღებული ცალკეული პროგნოზის გაერთიანება ხდება :

- ხმების უმრავლესობით კლასიფიკაციის ამოცანის შემთხვევაში ;
- საშუალო მნიშვნელობის აღების გზით რეგრესიის ამოცანის შემთხვევაში.

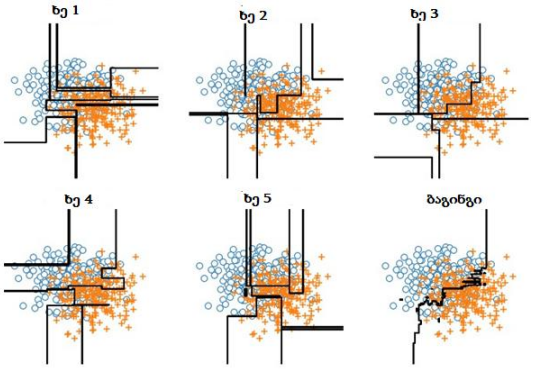
ბეგინგი ამცირებს ცალკეულ შემფასებელთა დისპერსიას. სწორედ ეს აძლევს მას საშუალებას მიაღწიოს მეტ მდგრადობას და უკეთეს პროგნოზს ამ შემფასებლებთან შედარებით.

ნახატი 9.5 ამ პრინციპის ილუსტრირებას იძლევა : პირველი 5 ხე, რომლებიც bagging მეთოდით გაწვრთნილ კლასიფიკატორს შეადგენს, მონაცემთა ნაკრების დაყოფას გაცილებით უფრო ცუდად ახორციელებს, ვიდრე bagging ალგორითმი, მაგრამ ისინი შეცდომას უშვებს სივრცის სხვადასხვა არეში და ამ არეთა გაერთიანებისას ეს შეცდომები აკომპენსირებს ერთმანეთს.

შემთხვევითი ტყეები

ანსამბლური (ჯგუფური) მეთოდების ძალა იმ შემთხვევაში ვლინდება, თუ სუსტი მოსწავლეები პირობითად დამოუკიდებელია მონაცემებისგან, ე.ი. მაქსიმალურად განსხვავდება ერთმანეთისგან, რომ მათი შეცდომების ურთიერთკომპენსირება მოხდეს. ამ მიზნის მისაღწევად *შემთხვევითი ტყეების* იდეა, რომელიც კვლავ ლეო ბრეიმანს ეკუთვნის, მდგომარეობს ცალკეული ხეების აგებაში არა მხოლოდ სხვადასხვა ანარჩევზე (როგორც ბეგინგის შემთხვევაში), არამედ სხვადასხვა *ცვლადების* გამოყენებითაც (Breiman, 2001).

უფრო ზუსტად რომ ვთქვათ, შემთხვევითი ტყის ფორმირებისათვის აგებული ხეები განსხვავდება CART ალგორითმის საშუალებით გაწვრთნილი ხეებისგან იმით, რომ თითოეულ კვანძში ჯერ შემთხვევით ირჩევენ $q < p$ ცვლადს, ხოლო შემდეგ *მათ შორის* გამოყოფენ მაცალკევებელ ცვლადს. კლასიფიკაციისას, ჩვეულებრივ, $q \approx \sqrt{p}$ სიდიდე გამოიყენება, რაც ასევე მნიშვნელოვნად ამცირებს გამოთვლების დროს, ვინაიდან თითოეულ კვანძში სულ რამდენიმე ცვლადი განიხილება (5 ამოცანისათვის 30 ცვლადით, 31 ამოცანისათვის 1000 ცვლადით). რეგრესიისათვის არაცხადად ნაგულისხმევი შეთანხმებით ხდება $q \approx \frac{p}{3}$ მნიშვნელობის არჩევა.



ნახატი 9.5 - დაყოფის ხარისხის ეფექტურობა იმ კლასიფიკატორის ტესტურ ნაკრებზე, რომელიც გაწვრთნილია bagging ალგორითმით (ქვემოთ მარჯვნივ) და ხუთი პირველი ხე, რომლებიც ამ კლასიფიკატორს შეადგენს

შენიშვნა

სხვადასხვა ხის პასუხების გასაშუალოებით, შემთხვევითი ტყეების გამოყენება წარმატებით შეიძლება როგორც რეგრესიის, ასევე კლასიფიკაციის ამოცანათა გადასაჭრელად.

პრაქტიკაში შემთხვევითი ხეები წარმოადგენს ერთ-ერთ ყველაზე ეფექტურ და განსახორციელებლად ადვილ ალგორითმს. მათი უპირატესობა იმაში მდგომარეობს, რომ ისინი ნაკლებად არის დამოკიდებული თავიანთ ჰიპერპარამეტრებზე, სხვა სიტყვებით : თითოეულ კვანძში გასათვალისწინებელი ცვლადების q რიცხვზე ; დაკვირვებათა რაოდენობაზე, რომლებიც ყოველი ხისათვის გამოიყენება (n რიცხვზე ჩვენ მიერ აღწერილ პროცედურაში, მაგრამ რომელიც შეიძლება შემცირებულიც იყოს) ; ხის ფოთლებში დაკვირვებათა მაქსიმალურ რიცხვზე (ჩვეულებრივ, განისაზღვრება 1-ით კლასიფიკაციაში და 5-ით — რეგრესიაში) ; დაბოლოს, ხეთა რიცხვზე, იმ პირობით, რომ იგი საკმარისად დიდია.

3.2 მიმდევრობითი მეთოდები : ბუსტინგი

მაშინ როცა ანსამბლის აგება ეფუძნება დაშვებას, რომ მოსწავლეებს კარგი მაჩვენებლებით დაკვირვებათა $\mathcal{X}$ სივრცის სხვადასხვა არეში ექნება არაკორელირებული შეცდომები და ამიტომ შეიძლება ვენდოთ იმას, რომ მათი უმრავლესობა მოცემულ წინასწარმეტყველებას უშეცდომოთ გააკეთებს, *მიმდევრობითი* მიდგომა ანსამბლის აგებისადმი ცდილობს შექმნას სუსტი მოდელები, რომლებშიც ძირითადი ყურადღება მოდელის ხარვეზებს და ცდომილებებს დაეთმობა.

ამას ეწოდება *ბუსტინგი* (ინგლ. boosting) : მიმდევრობითი იტერაციების მსვლელობისას სუსტი მოსწავლეები *აძლიერებს, ზრდის, ამაღლებს* («*booster*») თავიანთ ეფექტურობას (ხარისხობრივ მახასიათებლებს, მაჩვენებლებს) ფინალურ მოდელში, რომელიც მათ აერთიანებს.

პირველ წინადადებას ამ მიმართულებით წარმოადგენს AdaBoost ალგორითმი. მისი გამომგონებლებია ამერიკელი რობერტ ელიას შაპირი (Robert Elias Schapire, 1963-) და ებრაელი იოავ ფროინდი (Yoav Freund, 1961-). ამ გამორჩეული ნაშრომისათვის მანქანური სწავლების სფეროში 2003 წელს ორივემ გიოდელის პრემია (Kurt Friedrich Gödel prize) მიიღო.

ალგორითმთან დაკავშირებული ყველა მასალა ასახულია ნაშრომში Schapire et al (1997).

AdaBoost

AdaBoost ალგორითმი, თავისი ***Ada**ptive **Boost**ing* ფრაზიდან მიღებული დასახელებით, წარმოადგენს კლასიფიკატორის იტერაციული აგების მეთოდს, რომელიც აიძულებს სუსტ კლასიფიკატორს ორიენტაცია აიღოს მოდელების შეცდომებზე წონითი კოეფიციენტების სისტემის შემოღებით საწვრთნელი მაგალითებისათვის.

განსაზღვრება 9.4 (AdaBoost) დავუშვათ, რომ გვაქვს ბინარული კლასიფიკაციის მონაცემთა $\mathcal{D} = \left\{\left(\vec{x}^i, y^i\right)\right\}_{i=1,\ldots,n}$ ნაკრები, იტერაციათა M რიცხვი და სწავლების ალგორითმი. დავუშვათ ასევე, რომ მონაცემთა მოცემული $\mathcal{S}$ ნაკრებისთვის f_S ნოტაცია აღნიშნავს ამ ალგორითმით დასაბრუნებელი გადაწყვეტილების მიღების ფუნქციას. დაბოლოს, მივიჩნევთ, რომ $\mathcal{Y} = \{-1,1\}$ და f_S ფუნქციას აქვს მნიშვნელობა $\mathcal{Y}$ -ში.

მონაცემთა შეწონილ $\mathcal{D}' = \left\{\left(w_i, \vec{x}^i, y^i\right)\right\}_{i=1,\ldots,n} \in \mathbb{R}^n \times \mathcal{X}^n \times \mathcal{Y}^n$ ნაკრებად მიიჩნევენ $\left\{\left(\vec{x}^i, y^i\right)\right\}_{i=1,\ldots,n}$ მონაცემთა ნაკრებს, რომელშიც $i-$ურ დაკვირვებას მინიჭებული აქვს w_i წონა. აქ, ჩვეულებრივ, ვგულისხმობთ, რომ სწავლების ჩვენ მიერ გამოყენებულ ალგორითმს შეუძლია ამ წონების გათვალისწინება, მხედველობაში მიღება. გადაწყვეტილების ხეთა შემთხვევაში საწვრთნელი მაგალითების შეწონილობა აისახება დაბინძურების კრიტერიუმში მათი წონით ; გადაწყვეტილება ასევე მიიღება ხმათა შეწონილი უმრავლესობით.

AdaBoost ალგორითმი მოსწავლეთა (გასაწვრთნელთა) ანსამბლის (ჯგუფის) აგების შემდეგ პროცედურას ეწოდება :

1. $w_1^1, w_2^1, \ldots, w_n^1$ წონითი კოეფიციენტების ინიციალიზაცია $\frac{1}{n}$ მნიშვნელობ-ით(-ამდე)
2. $m = 1, 2, \ldots, M$ მნიშვნელობებისათვის :
   (a) გადაწყვეტილების ფუნქციის სწავლება (წვრთნა) მონაცემების შეწონილ ნაკრებზე $\mathcal{D}_m = \left\{\left(w_i^m, \vec{x}^i, y^i\right)\right\}_{i=1,\ldots,n}$:

$$f_m = f_{\mathcal{D}_m}$$

(b) ამ მოდელის შეწონილი შეცდომის გაანგარიშება :

$$\epsilon_m = \sum_{i=1}^{m} w_i^m \delta\left(f_m\left(\vec{x}^i\right), y^i\right). \tag{9.6}$$

(c) აქედან გამომდინარე, დასკვნა ნდობის დონეზე, რომელიც შეიძლება იყოს დაკავშირებული ამ შეცდომასთან :

$$\alpha_m = \frac{1}{2}\log\frac{1-\epsilon_m}{\epsilon_m}. \tag{9.7}$$

რაც უფრო ნაკლებია მოდელის გლობალური შეცდომა, მით მეტია α_m და, ამრიგად, შესაძლებელია მეტი ნდობის გამოცხადება ამ მოდელისათვის.

(d) წონების განახლება მეტი მნიშვნელოვნობის მისაცემად საწვრთნელი მაგალითისათვის, რომელშიც გადაწყვეტილების f_m ფუნქცია ცდება (ე.ი. უშვებს შეცდომას) :

$$w_i^{m+1} = \frac{1}{Z_m} w_i^m e^{-\alpha_m y^i f_m\left(\vec{x}^i\right)}, \text{ სადაც } Z_m = \sum_{l=1}^{nw} w_l^m e^{-\alpha_m y^l f_m\left(\vec{x}^l\right)}. \tag{9.8}$$

Z_m სიდიდის როლი მდგომარეობს იმაში, რომ w_i^{m+1} კოეფიციენტების ჯამი უდრიდეს 1-ს.

3. დასკვნითი გადაწყვეტილების მიღების ფუნქციის დაბრუნება :

$$f : \vec{x} \mapsto \sum_{m=1}^{M} \alpha_m f_m\left(\vec{x}\right).$$

AdaBoost ალგორითმისათვის სუსტ მოსწავლეებად (გასაწვრთნელებად), ტრადიციულად, მიიჩნევენ გადაწყვეტილების ხის ძალზე განსაკუთრებულ ტიპს : გადაწყვეტილების მიღების ხეს 1-ის ტოლი სიღრმით. ასეთი ხე ცნობილია *stump* [stʌmp] სახელწოდებით, რაც ინგლისურად ნიშნავს *კუნძს* — ხის მოჭრის შემდეგ მიწაში დარჩენილ ფესვებიან ნაწილს.

გრადიენტული ბუსტინგი

AdaBoost ალგორითმი ფაქტობრივად იმ მეთოდების ოჯახის კერძო შემთხვევაა, რომლებიც ცნობილია *გრადიენტული ბუსტინგის* სახელწოდებით (ინგლ. *gradient boosting* ან *GBOOST*). გრადიენტული ბუსტინგის თეორიული საფუძვლები მოცემული იყო ჯერ კიდევ 1999 წელს, ერთი მხრივ, ჯერომ ფრიდმანის (Jerome H. Friedman) მიერ, და მეორე მხრივ ლევ მეისონის (Llew Mason), ჯონატან ბაქსტერის (Jonathan Baxter), პიტერ ბარტლეტის (Peter Bartlett) და მარკუს ფრინის (Marcus Frean) მიერ : Friedman (2001) და Mason et al., (1999).

ასე რომ, გრადიენტული ბუსტინგი — მანქანური სწავლების ისეთი ტექნიკაა კლასიფიკაციისა და რეგრესიის ამოცანებისათვის, რომელიც წინასწარმეტყველების მოდელს აგებს სუსტი მაპროგნოზირებელი მოდელების (ჩვეულებრივ, გადაწყვეტილების ხეთა) ანსამბლის ფორმით.

სტრუქტურა, რომლის აღწერას ახლა ვიწყებთ, უპირველეს ყოვლისა, AdaBoost ალგორითმის მუშაობის პრინციპის უკეთ გაგების საშუალებას იძლევა, ვინაიდან მასში განიხილება *ემპირიული რისკის მინიმიზაცია* $\mathcal{D}$-ზე და *შეცდომის ექსპონენციალური ფუნქცია* ღირებულების

(დანახარჯთა) ფუნქციის როლში.

განსაზღვრება 9.5 (ექსპონენციალური შეცდომა) კლასიფიკაციის მოცემული ამოცანის პირობებში ღირებულების (დანახარჯთა) შემდეგ ფუნქციას ეწოდება *ექსპონენციალური შეცდომის* ფუნქცია, ანუ *ექსპონენციალური დანაკარგი*:

$$\left.\begin{array}{c}\{-1,1\}\times\mathbb{R}\to\mathbb{R}\\ y,\ f(\vec{x})\mapsto e^{-yf(\vec{x})}\end{array}\right\}.$$

ავიღოთ AdaBoost ალგორითმი და ვუწოდოთ F_m ფუნქციას *გადაწყვეტილების (მიღების) კუმულატიური ფუნქცია*:

$$F_m:\vec{x}\mapsto\sum_{k=1}^{m}\alpha_k f_k(\vec{x}).$$

შევნიშნავთ, რომ ზედსართავი სახელი «*კუმულატიური*» განსაზღვრავს ისეთ სუბსტანციას, რომელიც თანამიმდევრული ნაზრდებით ხასიათდება ან მატულობს რაოდენობას, ხარისხს, ძალას და ა.შ.

m-ურ ბიჯზე F_m-ის ექსპონენციალური შეცდომა საწვრთნელ სიმრავლეზე შემდეგნაირად აისახება:

$$\left.\begin{aligned}E_m&=\frac{1}{n}\sum_{i=1}^{n}\exp\left(-y^i\sum_{k=1}^{m-1}\alpha_k f_k(\vec{x}^i)\right)\exp\left(-\alpha_m y^i f_m(\vec{x}^i)\right)\\&=\frac{1}{n}\sum_{i=1}^{n}\exp\left(-y^i F_{m-1}(\vec{x}^i)\right)\exp\left(-\alpha_m y^i f_m(\vec{x}^i)\right)\end{aligned}\right\}.$$

ახლა განვსაზღვროთ $w_i^m=\exp\left(-y^i F_{m-1}(\vec{x}^i)\right)$, მაშინ:

$$E_m=\frac{1}{n}\sum_{i=1}^{n}w_i^m\exp\left(-\alpha_m y^i f_m(\vec{x}^i)\right).$$

რამდენადაც f_m არის სიდიდე $\{-1,1\}$-სიმრავლიდან, რომლის ელემენტებია -1 და 1, ამიტომ $y^i f_m(\vec{x}^i)$ ნამრავლი იქნება 1, თუ წინასწარმეტყველება სწორია, და -1 წინააღმდეგ შემთხვევაში, მაშასადამე

$$E_m=\frac{1}{n}\sum_{i=1}^{n}w_i^m e^{-\alpha_m}+\frac{1}{n}\sum_{i:\,f_m(\vec{x}^i)\neq y^i}w_i^m\left(e^{\alpha_m}-e^{-\alpha_m}\right).$$

ეს შეცდომა მინიმალურია, თუ α_m პარამეტრს აქვს (9.7)-განტოლებით მოცემული სახე. ამრიგად, AdaBoost ალგორითმი სუსტ მოსწავლეთ ისეთნაირად აერთიანებს, რომ მოხდეს გლობალური კლასიფიკატორის ექსპონენციალური შეცდომის მინიმიზაცია ყოველ ეტაპზე.

ექსპონენციალური შეცდომა შეიძლება იყოს ჩანაცვლებული ღირებულების (დანახარჯის) სხვა ფუნქციით, მაგალითად, კროს-ენტროპიით (ჯვარედინი ენტროპიით) ან კვადრატული

შეცდომით : სწორედ ეს არის ის, რასაც *გრადიენტულ ბუსტინგს* უწოდებენ. ამჟამად GBOOST ალგორითმი მანქანური სწავლების ერთ-ერთი ყველაზე პოპულარული ალგორითმია.

4 საკვანძო მომენტები

- გადაწყვეტილების ხეები — ეს ინტერპრეტირებადი მოდელებია, რომლებსაც, ბუნებრივია, შეუძლია მუშაობა რამდენიმე (ნამდვილი, დისკრეტული და ბინარული) სახის ცვლადთან და ეს მოდელები ადვილად ემორჩილება მულტიმოდალური განაწილებების მრავალკლასიან სწავლებას.

- გადაწყვეტილების ხეებს ის დიდი ნაკლოვანება აქვს, რომ ისინი სუსტი მოსწავლეებია და საერთოდ მათ მოდელირების ძალზე შეზღუდული შესაძლებლობები გააჩნია იმისათვის, რომ მათი ეფექტურობა მკვეთრად გამოირჩეოდეს პრაქტიკაში.

- ანსამბლური (ჯგუფური) მეთოდები გადაწყვეტილების ხეთა მსგავსი სუსტად დამსწავლელი მოდელების შეზღუდვათა აღმოფხვრის საშუალებას იძლევა ამ მოდელების ისეთნაირად გაერთიანების გზით, რომ მოხდეს მათი შესაბამისი შეცდომების დაკომპენსირება.

- ანსამბლური (ჯგუფური) პარალელური მეთოდები, როგორიცაა ბეგინგი (bagging) ან შემთხვევითი ტყეები (random forests), აგებს ცუდად (სუსტად) დასწავლილი (გაწვრთნილი) მოდელების დიდ რაოდენობას მონაცემთა ბუტსტრაპ-ანარჩევზე (bootstrap sampling), რაც გულისხმობს მონაცემების წყაროდან აღებული არჩევითი მონაცემების მრავალჯერად ჩატვირთვას პოპულაციის პარამეტრის შესაფასებლად.

- შემთხვევითი ტყეები თავიანთ ხეებს ისეთნაირად წვრთნის, რომ ისინი ერთი მეორისგან მონაცემების მიხედვით დამოუკიდებელი იყოს, რაც იმ ცვლადების შემთხვევითად შერჩევის გზით მიიღწევა, რომლებიც გათვალისწინებული იქნება ყოველი კვანძის შექმნისას.

- ბუსტინგი (boosting) ცუდად (სუსტად) გაწვრთნილი (დანასწავლი) მოდელების მიმდევრობით გაერთიანებას ახორციელებს, რათა მისცეს მეტი მნიშვნელობა იმ საწვრთნელ მაგალითებს, რომლებზეც წინასწარმეტყველებანი უფრო ცუდია.

დამატებითი ინფორმაცია

- შემთხვევითი ტყეები ასევე შეიძლება იყოს გამოყენებული თითოეული ცვლადის *მნიშვნელოვნობის ზომის (ხარისხის)* დასადგენად, რომელიც ეფუძნება მისი გამოყენების სიხშირეს მაცალკევებელ ცვლადად ტყის ყველა ხეთა შორის. უფრო კონკრეტულად, მნიშვნელოვნობა *გადაადგილების* (ინგლ. swap) მიხედვით ზომავს ეფექტურობის სხვაობას ტესტურ ნაკრებზე შეფასებულ ხესა და იმავე ტესტურ ნაკრებზე შეფასებულ ისეთ ხეს შორის, რომელშიც ელემენტები (შემავალი მონაცემები) $j-$ური ცვლადისათვის შემთხვევით იყო გადაადგილებული : თუ ეს ცვლადი მნიშვნელოვანია, მაშინ მეორე ხე უარესად უნდა ფუნქციონირებდეს. სხვა ზომა (მაჩვენებელი), რომელიც ზოგჯერ გამოიყენება — დაბინძურების კრიტერიუმის შემცირება — წარმოადგენს ჯინის კრიტერიუმის შემცირებათა შეწონილ ჯამს ხის ყველა *განსხლევისათვის* (გასხვლისას აჭრილი ტოტისათვის), რომლებიც რეალიზებულია მნიშვნელოვნობის გამზომი ცვლადით.

- GBOOST (Gradient BOOSTing) გრადიენტული ბუსტინგის განსაკუთრებით ცნობილ რეალიზაციას, რომელიც ფაქტობრივად მისი სინონიმიც კი გახდა, XGBOOST (eXtreme Gradient Boosting) რეალიზაცია წარმოადგენს (Chen and Guestrin, 2016). იგი ხელმისაწვდომია რესურსების უნიფიცირებულ ლოკატორზე (URL) https://xgboost.ai/ ან კიდევ https://github.com/dmlc/xgboost/.

5 ბიბლიოგრაფია

1. Breiman, L. (1996). Bagging predictors. Machine Learning, 26 :123–140.

2. Breiman, L. (2001). Random forests. Machine Learning, 45(1) :5–32.

3. Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). Classification and Regression Trees. Wadsworth International Group, Belmont, CA.

4. Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 785–794, New York, NY, USA. ACM.

5. Friedman, J. H. (2001). Greedy function approximation : a gradient boosting machine. The Annals of Statistics, 29(5) :1189–1232.

6. Mason, L., Baxter, J., Bartlett, P., and Frean, M. (1999). Boosting algorithms as gradient descent. In Proceedings of the 12th International Conference on Neural Information Processing Systems, pages 512–518, Cambridge, MA, USA. MIT Press.

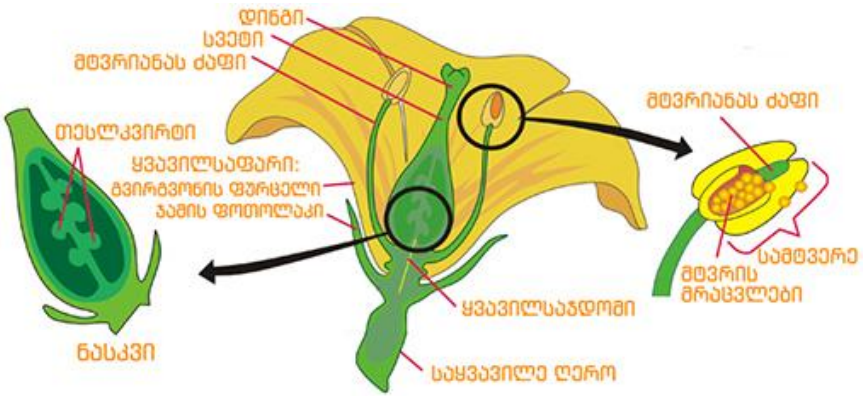
7. Schapire, R., Freund, Y., Bartlett, P., and Lee, W. S. (1997). Boosting the margin: a new explanation for the effectiveness of voting methods. The Annals of Statistics, 26 :322–330.

6 სავარჯიშოები

9.1 ანიკამ გაწვრთნა 50 პრედიქტორი თავის მონაცემებზე და ახლა იგი ფიქრობს მათი პროგნოზების გასაშუალოებას. ყველაზე მეტად როდის არის მოსალოდნელი, რომ ეს სტრატეგია იძლეოდეს უფრო მაღალ ეფექტურობას, ვიდრე ეს შეუძლია საუკეთესოს 50 მოდელიდან : როცა ცალკეული მოდელების მახასიათებლები კარგია თუ ცუდი ?

9.2 ბენომ გაწვრთნა გადაწყვეტილებათა ხე და შენიშნა, რომ მისი ეფექტურობა უფრო მაღალია საწვრთნელ სიმრავლეზე, ვიდრე სატესტოზე. უნდა გაზარდოს თუ უნდა შეამციროს ბენომ თავისი ხის სიღრმე ?

9.3 ქეთინომ შეადგინა მონაცემთა შემდეგი ნაკრები, რომელიც შეიცავდა 10 მცენარისთვის მათი ჯამების ფოთოლაკების (ინგლ. sepal, calyx) სიგრძესა და სიგანეს (იხ. მცენარის ყვავილის ელემენტები ნახატზე ქვემოთ).



მას სურს გამოყოს სხვებისგან ბალახოვანი მცენარეები (ლათ. plantae herbaceae), რომლებიც მიეკუთვნება Iris virginica (+) სახეს :

ჭდე	+	+	+	+	+	+	-	-	-	-
სიგრძე (cm)	6.7	6.7	6.3	6.5	6.2	5.9	6.1	6.4	6.6	6.8
სიგანე (cm)	3.3	3.0	2.5	3.0	3.4	3.0	2.8	2.9	3.0	2.8

1. გამოთვალეთ იტალიელი მათემატიკოსის კორადო ჯინის (Corrado Gini, 1884-1965) «დამაბინძურებელი» მინარევი სეპარაციის (გამოყოფის) ყველა შესაძლო წერტილისათვის განმაცალკევებელი ცვლადის როლში ჯამის ფოთოლაკის სიგრძის გამოყენებით.

2. გამოთვალეთ ჯინის «დამაბინძურებელი» მინარევი (ინგლ. Gini impurity) სეპარაციის (გამოყოფის) ყველა შესაძლო წერტილისათვის განმაცალკევებელი ცვლადის როლში ჯამის ფოთოლაკის სიგანის გამოყენებით.

3. როგორი იქნება გადაწყვეტილებათა ხის პირველი კვანძი, თუ ეს ხე გაწვრთნილია ქეთინოს მონაცემებზე ჯინის მინარევით ?

9.4 დიმას უნდოდა ეწინასწარმეტყველა, მოვა თუ არა დროზე მისი ავტობუსი. იგი აგროვებს 7 დღის მონაცემებს : ამინდი, ავტობუსში შესვლის მცდელობა არაპიკურ ან პიკურ დროს და მოცემულ დღეს დანიშნული მნიშვნელოვანი შეხვედრის არსებობა ან არარსებობა.

კარგი ამინდი	არა	არა	კი	კი	არა	კი	არა
არაპიკური საათი	კი	არა	არა	კი	კი	არა	არა
მნიშვნელოვანი შეხვედრა	კი	კი	არა	არა	არა	კი	არა
დროზე მოსვლა	+	-	+	+	+	-	-

ააგეთ შესაბამისი გადაწყვეტილებების ხე ჯინის (Gini) «დამაბინძურებელი» მინარევის გამოყენებით

9.5 ირინა იყენებს GBOOST უტილიტს (სერვისულ პროგრამას, რომელიც კომპიუტერის ეფექტური გამოყენების საშუალებას აძლევს მომხმარებელს) და იმ დასკვნამდე მიდის, რომ ხსენებული მომსახურე პროგრამა გადამეტებული სწავლების მდგომარეობაშია. რა პარამეტრი შეიძლება შეეცვალოს უტილიტს ?

9.6 ოთოს სურს გადაწყვეტილებათა ხეების საფუძველზე აგებული სწავლების მეთოდის გამოყენება, მაგრამ მას აკლია ზოგიერთი ცვლადის მნიშვნელობა გარკვეული დაკვირვებისათვის. რისი გაკეთება შეუძლია მას ?

9.7 დაჯის აქვს მონაცემთა ნაკრები, რომელიც შეიცავს p განზომილების n დაკვირვებას.

1. როგორია ამ მონაცემებზე გადაწყვეტილებათა ხის ფოთლების მაქსიმალური რიცხვი ?

2. როგორია ამ მონაცემებზე გადაწყვეტილებათა ხის მაქსიმალური სიღრმე ?

სავარჯიშოთა ამონახსნები

9.1 როცა ცალკეული მოდელების მახასიათებლები ცუდია (სუსტად არის გაწვრთნილი) .

9.2 ხის სიღრმის შემცირება გამოიწვევს იმას, რომ მოდელი ნაკლებ მიდრეკილებას გამოიჩენს გადამეტებული სწავლებისადმი.

9.3

1.

s	<5.9	5.9	6.1	6.2	6.3	6.4	6.5	6.6	6.7	6.8
GI	-	**0.444**	**0.475**	**0.476**	**0.450**	**0.480**	**0.467**	**0.476**	**0.400**	-

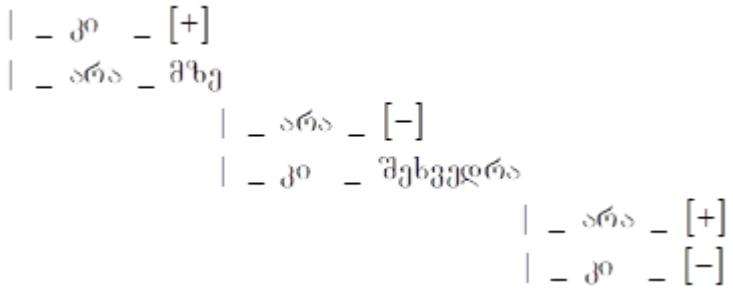
2.

s	<2.5	2.5	2.8	2.9	3.0	3.3	3.4
GI	-	**0.444**	**0.419**	**0.317**	**0.400**	**0.444**	-

3. ჯინის «დამაბინძურებელი» მინარევის უმცირესი მნიშვნელობა არის 0.317, რომელიც მიღებულია სიგრძით გამოყოფისას 2.9 სანტიმეტრის ზღურბლზე. ამრიგად, გადაწყვეტილებათა ხის პირველი კვანძი ადარებს ჯამთა ფოთოლაკების სიგრძეს 2.9 მნიშვნელობას. მარცხნივ — 4 მცენარეა 1 *Iris virjinica* ბალახის ჩათვლით ; მარჯვნივ — 6 მცენარეა 5 *Iris virjinica* ბალახის ჩათვლით.

9.4

არაპიკური საათი



9.5 პრედიქტორების რაოდენობა (რომელიც ალბათ, ძალიან დიდია) და გრადიენტული ალგორითმის სწავლების სიჩქარე (რომელიც, შესაძლოა, ოდნავ მომატებულია ასევე).

9.6 ოთოს შეუძლია გადაწყვეტილებათა ხის აგება იმ ცვლადების იგნორირებითაც, რომლებიც მას აკლია. სწავლებისას (წვრთნისას) მინარევის გამოსათვლელად გამოირიცხება დაკვირვებები, რომლებისთვისაც განმაცალკევებელი ცვლადის მნიშვნელობა მიუწვდომელია ან არ არსებობს. პროგნოზირებისას, თუ აუცილებელია გამოყოფა (განცალკევება) ცვლადით, რომლის მნიშვნელობა უცნობია, სვლას აგრძელებენ ორი შტოს გასწვრივ (რისთვისაც პოტენციურად შეიწონება თითოეული შტო მასში მოპოვებული საწვრთნელი ნაკრების დაკვირვებების წილით).

9.7

1. ყოველი ფოთოლი შეიცავს თუნდაც ერთ დაკვირვებას. ამიტომ ამ მოცემებზე გადაწყვეტილებათა ხის *ფოთლების მაქსიმალური რაოდენობა* არის n.

2. ყოველი ფოთოლი შეიცავს თუნდაც ერთ დაკვირვებას. ამიტომ ამ მოცემებზე გადაწყვეტილებათა ხის *მაქსიმალური სიღრმე* არის n (თუ ყოველი კვანძი გზაზე ფესვსა და ფოთოლს შორის შეიცავს დაკვირვებას).

ლექცია 10 საყრდენი ვექტორების მანქანები და ბირთვული მეთოდები

შინაარსი

საყრდენი ვექტორების მანქანები (ინგლ. SVM, Support Vector Machines) წარმოადგენს მანქანური სწავლების მძლავრ ალგორითმებს. ისინი ეფუძნება ვლადიმირ ვაპნიკის (Vladimir Vapnik) და ალექსანდრ ლერნერის (Aleksandr Lerner) მიერ 1963 წელს (Vapnik and Lerner, 1963) შემოთავაზებულ წრფივ ალგორითმს, მაგრამ წრფივ მოდელებზე გაცილებით მეტის დასწავლა არის შესაძლებელი. მართლაც, 1990-ანი წლების დასაწყისში ვლადიმირ ვაპნიკმა (Vladimir Vapnik), ბერნჰარდ ბოზერმა (Bernhard Boser), იზაბელ გიონმა (Isabelle Guyon) და კორინა კორტესმა (Corinna Cortes) (Boser et al., 1992; Cortes and Vapnik, 1995) იპოვეს მათი ეფექტურად გაფართოების ხერხი არაწრფივი მოდელების დასწავლამდე (გაწრთვნამდე) *ბირთვული ტრიუკის* (*ხრიკის*, *ოინის*) გამოყენებით (ინგლ. kernel trick, kernel function). ამ თავში წარმოდგენილია ამ მიდგომის სხვადასხვა ვერსია კლასიფიკაციის ამოცანისათვის და, ამრიგად, წარმოდგენილია *ბირთვული მეთოდების* ოჯახი.

მიზნები

— ფართო ნაპრალიანი კლასიფიკატორის განსაზღვრა სეპარაბელურ შემთხვევაში ;

— პრიმალური (პირველადი) და დუალური ოპტიმიზაციის შესაბამისი ამოცანების ჩაწერა ;

— ამ ამოცანათა ხელმეორედ ჩაწერა არასეპარაბელურ შემთხვევაში ;

— ბირთვული ტრიუკის ამოქმედება საყრდენი ვექტორების მანქანის გამოსაყენებლად მოქნილი ნაპრალით არაწრფივ შემთხვევაში ;

— ბირთვების დადგენა (განსაზღვრა) ნამდვილი რიცხვებით ასახული მონაცემებისათვის

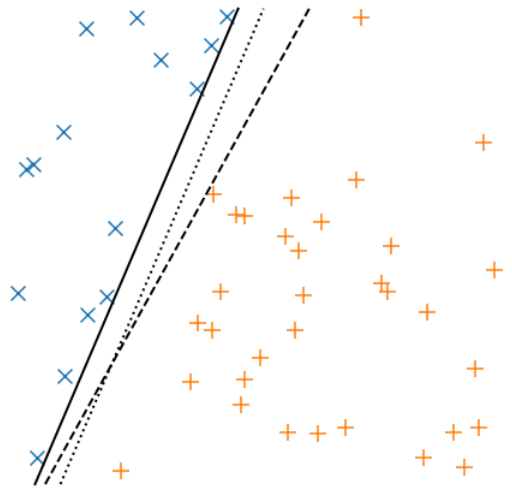
ან სიმბოლური სტრიქონებისათვის.

1 წრფივად სეპარაბელური შემთხვევა : საყრდენი ვექტორების მანქანა ხისტი ნაპრალით

ამ თავში ჩვენ განვიხილავთ ბინარული (ორობითი) კლასიფიკაციის ამოცანას. მოცემულ ქვედანაყოფში კი ჩვენ დავუშვებთ, რომ შეიძლება ისეთი წრფივი მოდელის პოვნა, რომელიც არ დაუშვებს შეცდომებს ჩვენს მონაცემებზე : ეს სწორედ რომ ის არის, რასაც *წრფივად სეპარაბელური* სცენარი, ანუ მოქმედების მონახაზი ეწოდება.

განსაზღვრება 10.1 (წრფივი სეპარაბელურობა) დავუშვათ, რომ $\mathcal{D}=\left\{\left(\vec{x}^i, y^i\right)\right\}_{i=1,\ldots,n}$ არის n დაკვირვების შემცველ მონაცემთა ნაკრები. ასევე დავუშვათ, რომ $\vec{x}^i \in \mathbb{R}^p$ და $y^i \in \{-1,1\}$. ჩვენ ვამბობთ, რომ $\mathcal{D}$ ნაკრები *წრფივად სეპარაბელურია*, თუ $\mathbb{R}^p$ – ში არსებობს თუნდაც ერთი ისეთი ჰიპერსიბრტყე, რომ $(+1)$ – ით აღნიშნული ყველა დადებითი წერტილი განთავსებულია ამ ჰიპერსიბრტყის ერთი მხრიდან, ხოლო (-1) – ით აღნიშნული ყველა უარყოფითი წერტილი — მეორე მხრიდან.

ამ შემთხვევაში ფაქტობრივად არსებობს *მაცალკევებელ ჰიპერსიბრტყეთა* უსასრულო რიცხვი, რომლებიც გამორიცხავს კლასიფიკაციის შეცდომათა დაშვებას (იხ. ნახატი 10.1). ეს ჰიპერსიბრტყეები ემპირიული რისკის მინიმიზაციის თვალსაზრისით არის ეკვივალენტური მოდელები.



ნახატი 10.1 - ჰიპერსიბრტყეთა (ორგანზომილებიან სივრცეში — წრფეთა) უსასრულო რიცხვი აცალკევებს (×) უარყოფით წერტილებს (+) დადებით წერტილებისგან.

1.1 მაცალკევებელი ჰიპერსიბრტყის ნაპრალი

დამატებითი ინფორმაციის არარსებობისას პუნქტირულ მაცალკევებელ ჰიპერსიბრტყეს ნახატზე 10.1 თითქოსდა მეტი უპირატესობა უნდა მიეცეს. ეს აიხსნება იმით, რომ, ვინაიდან იგი თანაბრად დაშორებულია უახლოესი დადებითი და უახლოესი უარყოფითი დაკვირვებებისგან, შეიძლება ითქვას, იგი ყოფს, ჭრის ორად მსხლისებრ არეს დადებით და

უარყოფით წერტილებს «შორის». მაცალკევებელი ჰიპერსიბრტყის *ნაპრალი* (ინგლ. margin - მარჟა) ამ ხედვის ფორმალიზების საშუალებას გვაძლევს.

განსაზღვრება 10.2 (ნაპრალი) მაცალკევებელი ჰიპერსიბრტყის γ *ნაპრალი* არის მანძილი ამ ჰიპერსიბრტყიდან საწვრთნელი (დამსწავლელი) სიმრავლის (ნაკრების) უახლოეს დაკვირვებამდე.

ამრიგად, სასურველი მაცალკევებელი ჰიპერსიბრტყე — ეს ის ჰიპერსიბრტყეა, რომელიც *მაქსიმუმს ანიჭებს ნაპრალს*. ამ შემთხვევაში არსებობს თუნდაც ერთი უარყოფითი და ერთი დადებითი დაკვირვება, რომლებიც იმყოფება γ მანძილზე მაცალკევებელი ჰიპერსიბრტყიდან : წინააღმდეგ შემთხვევაში, თუ, მაგალითად, ყველა უარყოფითი დაკვირვება იქნებოდა უფრო დიდ მანძილზე, ვიდრე γ მნიშვნელობაა, მაცალკევებელი ჰიპერსიბრტყიდან, ჩვენ შევძლებდით ამ ჰიპერსიბრტყის მიახლოებას უარყოფით დაკვირვებებთან და ნაპრალის გაზრდას.

ასეთ ვითარებაში ჩვენ შეგვიძლია განვსაზღვროთ, H მაცალკევებელი ჰიპერსიბრტყის გარდა, მის მიმართ პარალელური H_+ და H_- ჰიპერსიბრტყეები, რომლებიც განლაგებულია γ მანძილზე ორივე მხრიდან. H_+ შეიცავს თუნდაც ერთ დადებით დაკვირვებას, ხოლო H_- — თუნდაც ერთ უარყოფითს.

განსაზღვრება 10.3 (საყრდენი ვექტორები) დაკვირვებებს დამსწავლელ (საწვრთნელ) სიმრავლეზე, რომლებიც მაცალკევებელი ჰიპერსიბრტყიდან დაშორებულია γ მანძილით, *საყრდენ ვექტორებს* უწოდებენ. ამ სახელწოდებას განაპირობებს ის გარემოება, რომ სწორედ მათ «ეყრდნობა» H_+ და H_- ჰიპერსიბრტყეები.

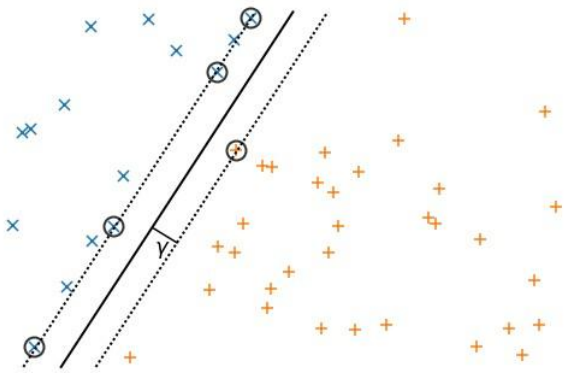
სწორედ აქედან არის ნაწარმოები *საყრდენ ვექტორთა მანქანის* ინგლისური დასახელება *Support Vector Machine*, ანუ შემოკლებით *SVM*. მას ზოგჯერ ფრანგები ასევე უწოდებენ «*სეპარატორს ვეებერთელა მარჟით*», რადგან ფრანგულ ენაზე ეს ფრაზა «Séparatrice à Vaste Marge» ზუსტად გადმოიცემა ინგლისური SVM აბრევიატურით.

თუ დაკვირვებას, რომელიც საყრდენი ვექტორია, ოდნავ დავძრავთ, ეს გამოიწვევს გაურკვევლობის ზონის გადაადგილებას და მაცალკევებელი ჰიპერსიბრტყის შეცვლას. და, პირიქით, თუ ოდნავ დავძრავთ დაკვირვებას, რომელიც საყრდენი ვექტორი არ არის, მაშინ H არ შეიცვლება : საყრდენი ვექტორები — ეს ის დაკვირვებებია, რომლებიც ამაგრებს გადაწყვეტილების საძირკველს.

ყველა დადებითი დაკვირვება განთავსებულია H_+-ის გარეთ, ხოლო ყველა უარყოფითი — H_--ის გარეთ.

განსაზღვრება 10.4 (გაურკვევლობის ზონა) არეს H_- ჰიპერსიბრტყესა და H_+ ჰიპერსიბრტყეს შორის ეწოდება *გაურკვევლობის ზონა* (ინგლ. *indecision zone*). ეს ზონა არ შეიცავს დაკვირვებებს.

ნახატი 10.2 ამ ცნებათა (კონცეფციათა) ილუსტრირებას იძლევა.



ნახატი 10.2 – უწყვეტი წირით ნაჩვენები მაცალკევებელი ჰიპერსიბრტყის γ ნაპრალი არის ამ ჰიპერსიბრტყის მანძილი უახლოეს დაკვირვებამდე.

როცა ეს ნაპრალი მაქსიმალურია, თუნდაც ერთი უარყოფითი და ერთი დადებითი დაკვირვება იმყოფება γ მანძილზე მაცალკევებელი ჰიპერსიბრტყიდან. ჰიპერსიბრტყეები (აქ პუნქტირული წირები), რომლებიც მაცალკევებელი ჰიპერსიბრტყის პარალელურია და ხსენებულ დაკვირვებათა გავლით გადის, განსაზღვრავს გაურკვევლობის ზონას. ამ ჰიპერსიბრტყეებზე განთავსებული (და წრეწირებით შემოვლებული) დაკვირვებები საყრდენ ვექტორებს წარმოადგენს.

1.2 საყრდენი ვექტორების მანქანა ხისტი ნაპრალით

სასურველი მაცალკევებელი H ჰიპერსიბრტყის განტოლებას აქვს $\langle \vec{w}, \vec{x} \rangle + b = 0$ სახე, სადაც $\langle , \rangle$ — სკალარული ნამრავლია $\mathbb{R}^p$-ზე. H_+ ჰიპერსიბრტყე მისი პარალელურია, ამიტომ $\langle \vec{w}, \vec{x} \rangle + b = \text{constante}$ არის H_+-ის განტოლება. ჩვენ შეგვიძლია ეს კონსტანტა გავუტოლოთ ერთს ზოგადობის დაუკარგავად $\left(\langle \vec{w}, \vec{x} \rangle + b = \text{constante} = 1\right)$: მართლაც, თუ ჩვენ გავამრავლებთ $\vec{w}$ და b სიდიდეებს ნებისმიერ კონსტანტაზე, ეს გავლენას არ მოახდენს H-ის განტოლებაზე. ვინაიდან H_- სიმეტრიულია H-ის და ეს უკანასკნელი ასეთივე მიმართებაშია H_+-თან, ამიტომ H_--ის განტოლება იქნება $\langle \vec{w}, \vec{x} \rangle + b = -1$. დაბოლოს, ნაპრალი, რომელიც წარმოადგენს მანძილს H-დან H_+-მდე, უდრის $\gamma = \frac{1}{\|\vec{w}\|_2}$ სიდიდეს.

ამრიგად, ყველა დადებითი დაკვირვება აკმაყოფილებს $\langle \vec{w}, \vec{x} \rangle + b \geq 1$ პირობას. მსგავსად ამისა, ყველა უარყოფითი წერტილი აკმაყოფილებს $\langle \vec{w}, \vec{x} \rangle + b \leq 1$ პირობას. ასე რომ, თითოეული წერტილისათვის ჩვენი დამსწავლელი (საწვრთნელი) სიმრავლიდან სრულდება $y^i \left(\langle \vec{w}, \vec{x}^i \rangle + b \right) \geq 1$ პირობა. ტოლობა სრულდება საყრდენი ვექტორებისათვის.

ამრიგად, ჩვენ ვცდილობთ აქ $\frac{1}{\|\vec{w}\|_2}$ გამოსახულების მაქსიმიზაციას n რაოდენობის $y^i\left(\left\langle \vec{w}, \vec{x}^i \right\rangle + b\right) \geq 1$ შეზღუდვათა პირობებში. ასე რომ, ჩვენი ამოცანა შეიძლება იყოს ფორმალიზებული შემდეგი სახით :

განსაზღვრება 10.5 (საყრდენი ვექტორების მანქანა ხისტი ნაპრალით — პრიმალური ფორმულირება) ოპტიმიზაციის შემდეგ ამოცანას ეწოდება *საყრდენი ვექტორების მანქანა ხისტი ნაპრალით* :

$$\arg \min_{\vec{w}\in\mathbb{R}^p, b\in\mathbb{R}} \frac{1}{2}\|\vec{w}\|_2^2 \text{ ისეთი, რომ } | \, y^i\left(\left\langle \vec{w}, \vec{x}^i \right\rangle + b\right) \geq 1, i = 1, \ldots, n \ . \tag{10.1}$$

დავუშვათ, რომ $\vec{w}^*, b^*$ — ეს 10.1 განტოლების ამონახსნია. მაშინ გადაწყვეტილების მიღების ფუნქცია შემდეგი სახით იქნება მოცემული :

$$f\left(\vec{x}\right) = \left\langle \vec{w}^*, \vec{x} \right\rangle + b^* \ . \tag{10.2}$$

1.3 დუალური ფორმულირება

განტოლებით 10.1 განსაზღვრული ამოცანა არის ამოზნექილი ოპტიმიზაციის ამოცანა n შეზღუდვით, რომელთა შორის თითოეული შეესაბამება დამსწავლელი (საწვრთნელი) სიმრავლის ერთ-ერთ წერტილს. უფრო მეტიც, ყველა ეს შეზღუდვა არის აფინური. ამრიგად, სლეიტერის (Morton Lincoln Slater, 1921-2002) პირობები იმის გარანტიას იძლევა, რომ ხისტი ნაპრალიანი საყრდენი ვექტორების მანქანის მიზნობრივი ფუნქცია (რომელიც 10.1 განტოლებაშია მოცემული) მინიმუმს იძენს თავისი დუალური ამოცანის იმავე წერტილებში (იხ. ქვედანაყოფი 13.4.3). ამრიგად, ჩვენ შეგვიძლია მივიღოთ ამოცანის მეორე ეკვივალენტური ფორმულირება :

თეორემა 10.1 (საყრდენი ვექტორების მანქანა ხისტი ნაპრალით — დუალური ფორმულირება) *განტოლებით 10.1 განსაზღვრული ამოცანა ეკვივალენტურია შემდეგი პრობლემის :*

$$\left.\begin{array}{l} \max\limits_{\vec{\alpha}\in\mathbb{R}^n} \sum\limits_{i=1}^{n} \alpha_i - \frac{1}{2}\sum\limits_{i=1}^{n}\sum\limits_{l=1}^{m} \alpha_i \alpha_l y^i y^l \left\langle \vec{x}^i \vec{x}^l \right\rangle \\ \text{ისეთი, რომ } \sum\limits_{i=1}^{n} \alpha_i y^i = 0; \alpha_i \geq 0, i = 1, \ldots, n \end{array}\right\}. \tag{10.3}$$

მტკიცებულება. ეს შეზღუდვებიანი ამოზნექილი ამოცანის ფორმულირებაა (იხ. ქვედანაყოფი 13.4). შემოგვაქვს ლაგრანჟის n მამრავლი $\{\alpha_i\}_{i=1,\ldots,n}$ — თითო-თითო ყოველ შეზღუდვაზე. მაშასადამე, *ლაგრანჟიანი* წარმოადგენს შემდეგი სახის ფუნქციას :

$$\left.\begin{array}{l} L:\mathbb{R}^p\times\mathbb{R}\times\mathbb{R}_+^n\to\mathbb{R} \\ \vec{w},b,\vec{\alpha}\mapsto\frac{1}{2}\|\vec{w}\|_2^2-\sum_{i=1}^n\alpha_i\left(y^i\left(\left\langle\vec{w},\vec{x}^i\right\rangle+b\right)-1\right) \end{array}\right\}. \tag{10.4}$$

ამრიგად, *ლაგრანჟის დუალური ფუნქცია* — ეს შემდეგი სახის ფუნქციაა :

$$\left.\begin{array}{l} q:\mathbb{R}_+^n\to\mathbb{R} \\ \vec{\alpha}\mapsto\inf\limits_{\vec{w}\in\mathbb{R}^p,b\in\mathbb{R}}L\left(\vec{w},b,\vec{\alpha}\right) \end{array}\right\}. \tag{10.5}$$

დაბოლოს, განტოლებით 10.1 წარმოდგენილი პრობლემის დუალურ ამოცანას ასეთი ფორმა აქვს :

$$\max_{\vec{\alpha}\in\mathbb{R}_+^n}\inf_{\vec{w}\in\mathbb{R}^p,b\in\mathbb{R}}\frac{1}{2}\|\vec{w}\|_2^2-\sum_{i=1}^n\alpha_i\left(y^i\left(\left\langle\vec{w},\vec{x}^i\right\rangle+b\right)-1\right). \tag{10.6}$$

ლაგრანჟიანი არის ამოზნექილი $\vec{w}$–ზე და, ამრიგად, მინიმალური, როცა მისი გრადიენტი $\vec{w}$ ცვლადით უდრის ნულს, ე.ი. როცა

$$\vec{w}=\sum_{i=1}^n\alpha_i y^i\vec{x}^i\,. \tag{10.7}$$

უფრო მეტიც, იგი აფინურია b წერტილზე. ამიტომ მისი ინფიმუმი არის $-\infty$ მხოლოდ იმ შემთხვევაში, თუ მისი გრადიენტი b წერტილზე არ უდრის ნულს (ამ ვითარებაში აფინური ფუნქცია არის «ბრტყელი»), ე.ი. როცა

$$\sum_{i=1}^n\alpha_i y^i=0\,. \tag{10.8}$$

ამიტომ მეორე შემთხვევაში q დუალური ფუნქცია არის მაქსიმიზებული.

თუ ჩავანაცვლებთ $\vec{w}$ სიდიდეს მისი გამოსახულებით (განტოლება 10.7) დუალური ფუნქციის ჩანაწერში, განტოლება 10.6 შეიძლება, ამრიგად, ხელმეორედ ჩამოვაყალიბოთ შემდეგი სახით :

$$\left.\begin{array}{l} \max\limits_{\vec{\alpha}\in\mathbb{R}^n}\frac{1}{2}\sum_{i=1}^n\sum_{l=1}^m\alpha_i\alpha_l y^i y^l\left\langle\vec{x}^i,\vec{x}^l\right\rangle-\sum_{i=1}^n\alpha_i y^i\sum_{l=1}^n\alpha_l y^l\left\langle\vec{x}^l,\vec{x}^i\right\rangle-\sum_{i=1}^n\alpha_i y^i b+\sum_{i=1}^n\alpha_i \\ \text{ისეთი, რომ }\sum_{i=1}^n\alpha_i y^i=0;\alpha_i\geq 0,i=1,\ldots,n \end{array}\right\}.$$

სასურველი შედეგი მიიღება განტოლების 10.8 გამოყენებით.

ამიტომ, პრიმალური ამოცანა რომ ამოვხსნათ (განტოლება 10.1), შეიძლება დავიწყოთ დუალური ამოცანის ამოხსნით (განტოლება 10.3). ამოცანის 10.3 ამონახსნად ავიღოთ $\vec{\alpha}^*$. განტოლება 10.7 გვაძლევს 10.1 განტოლების ამონახსნს $\vec{w}$ ცვლადისთვის : $\vec{w}^*=\sum_{i=1}^n\vec{\alpha}^* y^i\vec{x}^i$.

b^* სიდიდის დასადგენად შეიძლება დავუბრუნდეთ საყრდენი ვექტორების მანქანის საწყის ფორმულირებას : დადებითი საყრდენი ვექტორები განთავსებულია H_+ ჰიპერსიბრტყეზე და

აკმაყოფილებს $\left\langle \vec{w}^*, \vec{x}^i \right\rangle + b^* = 1$ პირობას.

გარდა ამისა, რადგან დადებითი დაკვირვებები ყველაზე უფრო ახლოსაა განთავსებული H მაცალკევებელ ჰიპერსიბრტყესთან, ამიტომ დადებითი საყრდენი ვექტორები $\left\langle \vec{w}^*, \vec{x}^i \right\rangle$-ის მინიმიზაციას ახდენს. მაშასადამე,

$$b^* = 1 - \min_{i:y^i=+1} \left\langle \vec{w}^*, \vec{x}^i \right\rangle .$$

დაბოლოს, *გადაწყვეტილების მიღების ფუნქცია* მოიცემა თანაფარდობით :

$$f\left(\vec{x}\right) = \sum_{i=1}^{n} \alpha_i^* y^i \left\langle \vec{x}^i, \vec{x} \right\rangle + b^* . \tag{10.9}$$

ჩანართი

ალგორითმული სირთულე

საყრდენ ვექტორთა მანქანის პრიმალური ფორმულირება წარმოადგენს $p+1$ განზომილების ოპტიმიზაციის ამოცანას, მაშინ როცა დუალური ფორმულირება — n განზომილების ოპტიმიზაციის ამოცანას. თუ ჩვენ ცოტა მონაცემი და ბევრი ცვლადი გვაქვს, უპირატესობა დუალურ ფორმულირებას უნდა მიეცეს, წინააღმდეგ შემთხვევაში უპირატესობა პრიმალური ამოცანის ამოხსნას ეძლევა.

1.4 გეომეტრიული ინტერპრეტაცია

კარუშ-კუნ-ტაკერის (Karush-Kuhn-Tucker, KKT) (იხ. ქვედანაყოფი 13.4.4) *პირობები* $\vec{\alpha}^*$-სა და $\left(\vec{w}^*, b^*\right)$-ს შორის კავშირის უფრო ზუსტად დახასიათების საშუალებას იძლევა. აღვნიშნოთ ϕ სიმბოლოთი ფუნქცია, რომელიც $\vec{w}$-თან აკავშირებს $\frac{1}{2}\|\vec{w}\|_2^2$ სიდიდეს, ხოლო g_i სიმბოლოთი — ფუნქცია, რომელიც $\vec{w}, b$-თან ამყარებს $g_i\left(\vec{w}, b\right) = y^i\left(\left\langle \vec{w}, \vec{x}^i \right\rangle + b\right) - 1$ სახის კავშირს. ნებისმიერი $1 \le i \le n$ მნიშვნელობისათვის *დამატებითი გადახრის* (ინგლ. complementary deviation, ფრანგ. écart complémentaire) პირობა ნიშნავს, რომ $\alpha_i^* g_i\left(\vec{w}^*, b^*\right) = 0$.

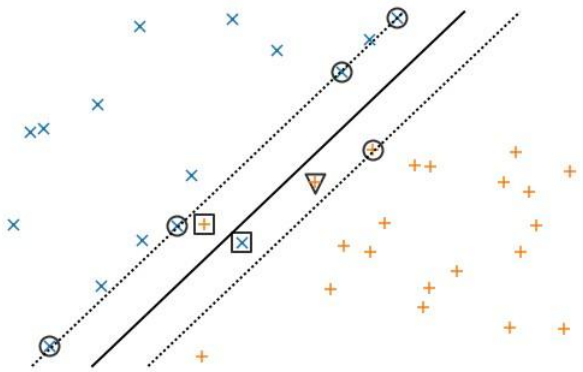
უფრო ზუსტად რომ ვთქვათ, პირობათა ნებისმიერი i მნიშვნელობისათვის ორი შემთხვევა არის შესაძლებელი :

- $\alpha_i^* = 0$: ϕ-ის მინიმიზატორი აკმაყოფილებს შეზღუდვას და $g_i\left(\vec{w}^*, b^*\right) > 0$, სხვანაირად რომ ვთქვათ, $\vec{x}^i$ წერტილი იმყოფება H_+ ან H_- ჰიპერსიბრტყეთა გარეთ ;
- $\alpha_i^* > 0$: შეზღუდვა მოწმდება შესრულებადობის (განხორციელებადობის, რეალიზებადობის) ზონის საზღვარზე, ე.ი. როცა სრულდება $g_i\left(\vec{w}^*, b^*\right) = 0$ ტოლობა და $\vec{x}^i$ არის საყრდენი ვექტორი.

ასე რომ, *საყრდენი ვექტორები — დაკვირვებებია მონაცემთა ნაკრებიდან, რომელიც შეესაბამება ლაგრანჟის არანულოვან* α_i^* *მამრავლს.*

2 წრფივად არასეპარაბელური შემთხვევა : საყრდენი ვექტორების მანქანა მოქნილი (რბილი) ნაპრალით

ჩვენი მონაცემები, როგორც წესი, არ არის წრფივად განცალკევებადი. ასეთ შემთხვევაში, ნებისმიერი მაცალკევებელი ჰიპერსიბრტყის არჩევის პირობებშიც კი, ზოგიერთი წერტილი მცდარად იქნება კლასიფიცირებული, სხვები — სწორად, მაგრამ მხოლოდ გაურკვევლობის ზონაში (არეში). ეს ცნებები ილუსტრირებულია ნახატზე 10.3.



ნახატი 10.3 - არცერთ წრფივ კლასიფიკატორს არ შეუძლია იდეალურად ამ მონაცემების განცალკევება. კვადრატით მონიშნული დაკვირვებები კლასიფიცირებულია არასწორად. სამკუთხედით მონიშნული დაკვირვება კლასიფიცირებულია სწორად, მაგრამ გაურკვევლობის არეში. ეს დაკვირვება რომ ხსენებული ზონის საზღვარზე იმყოფებოდეს, სხვა სიტყვებით, იგი რომ საყრდენი ვექტორი იყოს, მაშინ ნაპრალი გაცილებით უფრო მცირე იქნებოდა.

2.1 საყრდენი ვექტორების მანქანა მოქნილი (რბილი) ნაპრალით

ახლა ჩვენი მიზანია კომპრომისის პოვნა კლასიფიკაციის შეცდომებსა და ნაპრალის ზომას შორის. როგორც საყრდენი ვექტორების მანქანაში ხისტი ნაპრალით (განტოლება 10.1), აქაც ვცდილობთ $\frac{1}{2}\|\vec{w}\|_2^2$ ნაპრალის კვადრატის შებრუნებული სიდიდის, რომელსაც $C\times\sum_{i=1}^{n} l\left(f\left(\vec{x}^i\right),y^i\right)$ ცდომილების წევრი ემატება, მინიმუმამდე დაყვანას. აქ $C\in\mathbb{R}^+$ - საყრდენი ვექტორების მანქანის ჰიპერპარამეტრია, ხოლო L — ღირებულების ფუნქცია :

$$\arg\min_{\vec{w}\in\mathbb{R}^p,b\in\mathbb{R}} \frac{1}{2}\|\vec{w}\|_2^2 + C\sum_{i=1}^{n} L\left(\left\langle \vec{w},\vec{x}^i\right\rangle + b, y^i\right). \tag{10.10}$$

ღირებულების (დანახარჯების) C ჰიპერპარამეტრი გამოიყენება ნაპრალურ და მოდელურ ცდომილებათა ფარდობითი მნიშვნელოვნობის სამართავად დამსწავლელ (საწვრთნელ) ნაკრებზე (სიმრავლეზე). ამრიგად, იგი ანიჭებს ნაპრალს გარკვეულ *მოქნილობას* და ასეთ შემთხვევაში ლაპარაკობენ საყრდენი ვექტორების მანქანაზე *მოქნილი, რბილი, რეგულირებადი ნაპრალით* (ინგლ. SVM with *flexible margin*).

შეძლებისდაგვარად ცდილობენ, რომ ყოველი $\vec{x}$ დაკვირვება y ჭდით იმყოფებოდეს გაურკვევლობის ზონის გარეთ, ანუ აკმაყოფილებდეს $y\,f\left(\vec{x}\right)\geq 1$ პირობას. ამიტომ ჩვენ გამოვიყენებთ *სახსრულ შეცდომას* (იხ. ქვედანაყოფი 2.12) როგორც ღირებულების (დანახარჯის, დანაკარგის) ფუნქციას. მკაცრი მათემატიკური ენით რომ ვისარგებლოთ, გამოყენებული იქნება

დანაკარგთა უბან-უბან წრფივი (ინგლ. hinge loss) ფუნქცია. ამრიგად, განტოლებას 10.10 შეიძლება სხვა სახეც მიეცეს.

განსაზღვრება 10.6 (საყრდენი ვექტორების მანქანა მოქნილი/რბილი ნაპრალით) შემდეგი ოპტიმიზაციური ამოცანის ამონახსნს ეწოდება *საყრდენი ვექტორების მანქანა მოქნილი (რბილი) ნაპრალით :*

$$\arg \min_{\vec{w}\in\mathbb{R}^p, b\in\mathbb{R}} \frac{1}{2}\|\vec{w}\|_2^2 + C\sum_{i=1}^{n}\left[1-y^i f\left(\vec{x}^i\right)\right]_+ . \tag{10.11}$$

შენიშვნა

ეს ფორმულირება წარმოადგენს ემპირიული რისკის მინიმიზაციის რეგულარიზაციას ნორმის ℓ_2 წევრის საშუალებით. იგი ჰგავს *თხემურ რეგრესიას* (ინგლ. *ridge regression*), რომლის შესახებ ვსაუბრობდით მე-6 ლექციაში, C უკუპროპორციულია λ პარამეტრის, მაგრამ შეცდომის ფუნქცია განსხავებულია.

განსაზღვრება 10.7 (საყრდენი ვექტორების მანქანა მოქნილი/რბილი ნაპრალით — პრიმალური ფორმულირება) თუ შემოვიტანთ განსახილველად $\xi_i = \left[1-y^i f\left(\vec{x}^i\right)\right]_+$ *მაკორექტირებელ*, ანუ *მარეგულირებელ ცვლადს* (ინგლ. slack variable) დამსწავლელი (საწვრთნელი) სიმრავლის თითოეული დაკვირვებისათვის, ოპტიმიზაციის ამოცანა 10.11 შემდეგი გამოთვლების ეკვივალენტურია :

$$\left.\begin{array}{l} \arg \min_{\vec{w}\in\mathbb{R}^p, b\in\mathbb{R}, \vec{\xi}\in\mathbb{R}^n} \frac{1}{2}\|\vec{w}\|_2^2 + C\sum_{i=1}^{n}\xi_i \\ \text{ისეთი, რომ } \left(\left\langle \vec{w}, \vec{x}^i \right\rangle + b\right) \geq 1-\xi_i, i=1,\ldots,n \\ \xi_i \geq 0, i=1,\ldots,n \end{array}\right\}. \tag{10.12}$$

2.2 დუალური ფორმულირება

როგორც საყრდენ ვექტორთა მანქანის შემთხვევაში ხისტი ნაპრალით, ეს ოპტიმიზაციის ამოზნექილი ამოცანაა, ამჯერად $2n$ შეზღუდვით, თანაც ყველა შეზღუდვა აფინურია და მორტონ სლეიტერის (Morton Lincoln Slater, 1921-2002) პირობები ძალაშია.

თეორემა 10.2 (საყრდენი ვექტორების მანქანა მოქნილი/რბილი ნაპრალით — დუალური ფორმულირება) *ამოცანა 10.12 შემდეგი გამოთვლების ეკვივალენტურია :*

$$\left.\begin{array}{l} \max_{\vec{\alpha}\in\mathbb{R}} \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{l=1}^{n}\alpha_i\alpha_l y^i y^l \left\langle \vec{x}^i, \vec{x}^l \right\rangle \\ \text{ისეთი, რომ } \sum_{i=1}^{n}\alpha_i y^i = 0; 0 \leq \alpha_i \leq C, i=1,\ldots,n \end{array}\right\}. \tag{10.13}$$

მტკიცებულება. შემოვიტანოთ განსახილველად $\{\alpha_i, \beta_i\}_{i=1,\ldots,n}$ ლაგრანჟის $2n$ მამრავლი და ჩავწეროთ ლაგრანჟიანი :

$$\left.\begin{array}{l} L : \mathbb{R}^p \times \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}_+^n \times \mathbb{R}_+^n \to \mathbb{R} \\ \qquad \vec{w}, b, \vec{\xi}, \vec{\alpha}, \vec{\beta} \mapsto \frac{1}{2}\|\vec{w}\|_2^2 + C\sum_{i=1}^n \xi_i - \sum_{i=1}^n \alpha_i \left(\left(y^i \left\langle \vec{w}, \vec{x}^i \right\rangle + b \right) - 1 + \xi_i \right) - \sum_{i=1}^n \beta_i \xi_i \end{array}\right\}. \quad (10.14)$$

ამრიგად, *ლაგრანჟის დუალური ფუნქცია* — ეს შემდეგი სახის ფუნქციაა :

$$\left.\begin{array}{l} q : \mathbb{R}_+^n \times \mathbb{R}_+^n \to \mathbb{R} \\ \qquad \vec{\alpha}, \vec{\beta} \mapsto \inf\limits_{\vec{w} \in \mathbb{R}^p, b \in \mathbb{R}, \vec{\xi} \in \mathbb{R}^n} L\left(\vec{w}, b, \vec{\xi}, \vec{\alpha}, \vec{\beta} \right) \end{array}\right\}. \quad (10.15)$$

ასე რომ, 10.12 განტოლებით წარმოდგენილ დუალურ ამოცანა ამ ფორმით ჩაიწერება :

$$\max_{\vec{\alpha} \in \mathbb{R}_+^n, \vec{\beta} \in \mathbb{R}^p} \inf_{\vec{w} \in \mathbb{R}^p, b \in \mathbb{R}, \vec{\xi} \in \mathbb{R}^n} \frac{1}{2}\|\vec{w}\|_2^2 + C\sum_{i=1}^n \xi_i - \sum_{i=1}^n \alpha_i \left(\left(y^i \left\langle \vec{w}, \vec{x}^i \right\rangle + b \right) - 1 + \xi_i \right) - \sum_{i=1}^n \beta_i \xi_i \,. \quad (10.16)$$

მსგავსად იმისა, რაც ვქონდა საყრდენ ვექტორთა მანქანის შემთხვევაში ხისტი ნაპრალით, ლაგრანჟიანი მინიმალურია, როცა მისი გრადიენტი $\vec{w}$ ცვლადით ნულის ტოლია, ე.ი. როცა

$$\vec{w} = \sum_{i=1}^n \alpha_i y^i \vec{x}^i \,. \quad (10.17)$$

როგორც ადრე, იგი აფინურია b -ში და მისი ინფიმუმი, მაშასადამე, არის $-\infty$, მაგრამ მხოლოდ მაშინ, თუ მისი გრადიენტი b -ში უდრის ნულს, ე.ი. როცა

$$\sum_{i=1}^n \alpha_i y^i = 0 \,. \quad (10.18)$$

უფრო მეტიც, იგი აფინურია $\vec{\xi}$ -ში და მისი ინფიმუმი, მაშასადამე, არის $-\infty$, მაგრამ მხოლოდ მაშინ, თუ მისი გრადიენტი $\vec{\xi}$ -ში უდრის ნულს, ე.ი. როცა

$$\beta_i = C - \alpha_i, i = 1, \ldots, n \,. \quad (10.19)$$

ამრიგად, დუალური q ფუნქცია 10.18 და 10.19 განტოლებათა შესრულებისას მაქსიმუმს აღწევს.

დუალურ ფუნქციაში $\vec{w}$ სიდიდის ჩანაცვლებისას მისივე გამოსახულებით განტოლებიდან 10.17, პრობლემა 10.16 შეიძლება იქნეს ახლებურად ჩამოყალიბებული შემდეგი ფორმით :

$$\left.\begin{array}{l} \max\limits_{\vec{\alpha} \in \mathbb{R}^n} -\frac{1}{2}\sum_{i=1}^n \sum_{l=1}^n \alpha_i \alpha_l y^i y^l \left\langle \vec{x}^i, \vec{x}^l \right\rangle + \sum_{i=1}^n \alpha_i + C\sum_{i=1}^n \xi_i - \sum_{i=1}^n (C - \alpha_i)\xi_i - \sum_{i=1}^n \alpha_i \xi_i \\ \text{ისეთი, რომ } \sum_{i=1}^n \alpha_i y^i = 0; \alpha_i \geq 0, i = 1, \ldots, n; C - \alpha_i \geq 0, i = 1, \ldots, n; \end{array}\right\}.$$

□

შენიშვნა

ხისტი ნაპრალიანი საყრდენ ვექტორთა მანქანის ერთადერთი განსხვავება დუალური ფორმულირებისგან (განტოლება 10.3) მდგომარეობს $\alpha_i \le C,\ i=1,\ldots,n$ შეზღუდვაში.

2.3 გეომეტრიული ინტერპრეტაცია

წინა შემთხვევის მსგავსად, კარუშ-კუნ-ტაკერის პირობები $\vec{\alpha}^*$ და $\left(\vec{w}^*, b^*\right)$ სიდიდეთა შორის კავშირის უფრო ზუსტად დახასიათების საშუალებას გვაძლევს.

ახლა ყოველი i -ური დაკვირვებისათვის გვაქვს ორი *დამატებითი გადახრის* პირობა :

- $\alpha_i^* g_i\left(\vec{w}^*, b^*\right) = 0$ და
- $\beta_i^* h_i\left(\vec{w}^*, b^*\right) = 0$, ან კიდევ $\left(C - \alpha_i^*\right) h_i\left(\vec{w}^*, b^*\right) = 0$, სადაც
  $h_i\left(\vec{w}^*, b^*\right) = \left[1 - y^i\left(\left\langle \vec{w}^*, \vec{x}^i \right\rangle + b\right)\right]_+$.

ამრიგად, ყოველი i -ური დაკვირვებისათვის, სამი შესაძლებლობა გვაქვს :

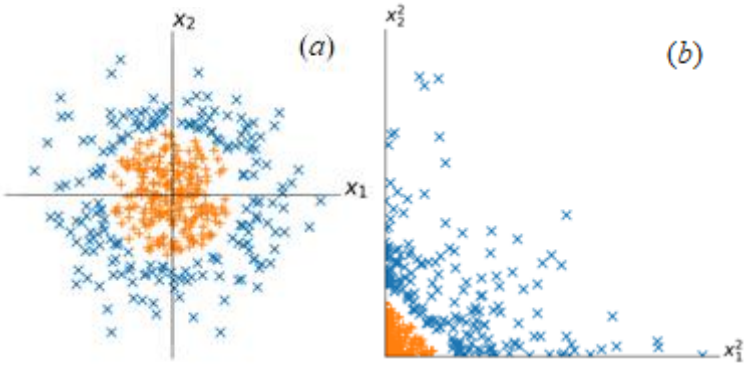
- $\alpha_i^* = 0$: $\frac{1}{2}\|\vec{w}\|_2^2$-ის მინიმიზატორი აკმაყოფილებს შეზღუდვას და $y^i\left(\left\langle \vec{w}^*, \vec{x}^i \right\rangle + b\right) > 1$ უტოლობას. $\vec{x}^i$ დაკვირვება კვლავ გაურკვევლობის ზონის *გარეთ* იმყოფება.
- $0 < \alpha_i^* < C = 0$: როგორც ადრე, $\vec{x}^i$ არის საყრდენი ვექტორი, რომელიც განთავსებულია გაურკვევლობის ზონის საზღვარზე.
- $\beta_i^* = 0$: $\alpha_i^* = C$ და ასეთ ვითარებაში $\left[1 - y^i\left(\left\langle \vec{w}^*, \vec{x}^i \right\rangle + b\right)\right]_+ > 0$. ამ შემთხვევაში გაურკვევლობის ზონის საზღვრის მიმართ $\vec{x}^i$ დაკვირვება *არასათანადო მხარეზეა*.

შენიშვნა

საყრდენი ვექტორების მანქანა შეიძლება გამოვიყენოთ მრავალკლასიანი კლასიფიკატორის ასაგებადაც «*ერთი ყველას წინააღმდეგ*» (ინგლ. «one-versus-all») ან «ერთი ერთის წინააღმდეგ» (ინგლ. «one-versus-one») მიდგომათა გამოყენებით (იხ. ქვედანაყოფი 2.1.2).

3 არაწრფივი შემთხვევა : ბირთვული საყრდენი ვექტორების მანქანა

ხშირად წრფივი ფუნქცია არ გამოდგება ჩვენი მონაცემების განსაცალკევებლად (იხ., მაგალითად, ნახატი 10.4(a)). რა უნდა გაკეთდეს/მოვიმოქმედოთ ამ შემთხვევაში ?



(*a*) წრეწირი, ალბათ, გაცილებით უკეთ აცალკევებს ამ მონაცემებს, ვიდრე წრფე.
(*b*) გარდასახვის შემდეგ $\phi:(x_1,\ x_2)\mapsto(x_1^2,\ x_2^2)$ ფუნქციით მონაცემები წრფივად სეპარაბელურია ხელახალი აღწერის სივრცეში.
ნახატი 10.4 - მონაცემთა გარდასახვა მათი გაცალკევების საშუალებას იძლევა ხელახალი აღწერის სივრცეში.

3.1 ხელახალი აღწერის სივრცე

ნახატზე 10.4(*a*) წარმოდგენილი მონაცემების შემთხვევაში ორი კლასის განსაცალკევებლად $x_1^2+x_2^2=R^2$ განტოლებით მოცემული წრეწირი უფრო მისაღები ჩანს, ვიდრე წრფე.

ახლა, მართალია, $f:\mathbb{R}^2\to\mathbb{R},\vec{x}\mapsto x_1^2+x_2^2-R^2$ ფუნქცია არაწრფივია $\vec{x}=\left(x_1,x_2\right)$ სივრცეზე, მაგრამ იგი წრფივია $\left(x_1^2,x_2^2\right)$ სივრცეზე.

ამიტომ განვსაზღვროთ $\phi:\mathbb{R}^2\to\mathbb{R}^2,\left(x_1,x_2\right)\mapsto\left(x_1^2,x_2^2\right)$ ასახვა. გადაწყვეტილების მიღების f ფუნქცია წრფივია $\phi\left(\vec{x}\right)$: $f\left(\vec{x}\right)=\phi\left(\vec{x}\right)_1+\phi\left(\vec{x}\right)_2-R^2$ ასახვაზე. მაშასადამე, შეგვიძლია დავასწავლოთ იგი საყრდენი ვექტორების მანქანით *მონაცემთა სახეებზე*, რომლებიც ϕ ასახვას იყენებს.

უფრო ზოგად შემთხვევაში ჩვენ ახლა ჩავთვლით, რომ დაკვირვებები განსაზღვრულია ნებისმიერ $\mathcal{X}$ სივრცეზე, რომელიც შეიძლება იყოს წარმოდგენილი არა მხოლოდ $\mathbb{R}^p$ კლასით, არამედ, მაგალითად, სიმბოლოთა სტრიქონების სიმრავლითაც მოცემულ ალფაბეტზე, ყველა გრაფის სივრცით ან ფუნქციათა სივრცით.

განსაზღვრება 10.8 (ხელახალი აღწერის სივრცე) ჰილბერტის $\mathcal{H}$ სივრცე, რომელშიც სასურველია მონაცემების ხელახალი აღწერა $\mathcal{X}\to\mathcal{H}$ ასახვის გამოყენებით საყრდენი ვექტორების მანქანის გასაწვრთნელად დაკვირვებათა სახეებზე დამსწავლელ სიმრავლეზე, *ხელახალი აღწერის* სივრცე ეწოდება.

მონაცემთა ხელახალი აღწერა ჰილბერტის სივრცეში წრფივი ალგორითმის გამოყენების საშუალებას იძლევა, მაგალითად, საყრდენი ვექტორების მანქანის რბილი ნაპრალით, არაწრფივი ამოცანის ამოსახსნელად.

3.2 საყრდენი ვექტორების მანქანა ხელახალი აღწერის სივრცეში

ამრიგად, საყრდენი ვექტორების მანქანის დასწავლისათვის (გასაწვრთნელად) ჩვენი დაკვირ-

ვებების სახეებზე ხელახალი აღწერის $\mathcal{H}$ სივრცეში, აუცილებელია ამოვხსნათ, 10.13 განტოლების თანახმად, შემდეგი ამოცანა :

$$\left.\begin{array}{l}\max\limits_{\vec{\alpha}\in\mathbb{R}} \sum\limits_{i=1}^{n}\alpha_i - \frac{1}{2}\sum\limits_{i=1}^{n}\sum\limits_{l=1}^{n}\alpha_i\alpha_l y^i y^l \left\langle \phi\left(\vec{x}^{\,i}\right), \phi\left(\vec{x}^{\,l}\right)\right\rangle_{\mathcal{H}} \\ \text{ისეთი, რომ } \sum\limits_{i=1}^{n}\alpha_i y^i = 0; 0 \le \alpha_i \le C, i = 1,\ldots,n\end{array}\right\}. \tag{10.20}$$

მაშინ გადაწყვეტილების მიღების ფუნქცია მოცემული აღმოჩნდება შემდეგი გამოსახულებით (იხ. განტოლება 10.9) :

$$f\left(\vec{x}\right) = \sum_{i=1}^{n}\alpha_i^* y^i \left\langle \phi\left(\vec{x}^{\,i}\right), \phi\left(\vec{x}\right)\right\rangle_{\mathcal{H}} + b^* . \tag{10.21}$$

3.3 ბირთვული ტრიუკი (ხრიკი, ოინი, ფანდი)

განტოლებებში 10.20 და 10.21 დაკვირვებათა სახეები $\mathcal{H}$ სივრცეში ფიგურირებს მხოლოდ სკალარულ ნამრავლებში $\mathcal{H}$-ზე. ამიტომ $\phi: \mathcal{X} \to \mathcal{H}$ ფუნქციის ნაცვლად შეიძლება *ბირთვად* წოდებული შემდეგი ფუნქციის გამოყენება :

$$\left.\begin{array}{l} k : \mathcal{X}\times\mathcal{X} \to \mathbb{R} \\ \qquad \vec{x}, \vec{x}' \mapsto \left\langle \phi\left(\vec{x}\right), \phi\left(\vec{x}'\right)\right\rangle_{\mathcal{H}}\end{array}\right\}.$$

ახლა ჩვენ შეგვიძლია განვსაზღვროთ *ბირთვული საყრდენი ვექტორების მანქანა* :

განსაზღვრება 10.9 (ბირთვული საყრდენი ვექტორების მანქანა) ოპტიმიზაციის შემდეგი ამოცანის ამონახსნს ეწოდება *ბირთვული საყრდენი ვექტორების მანქანა* (ინგლ. kernel SVM) :

$$\left.\begin{array}{l}\max\limits_{\vec{\alpha}\in\mathbb{R}} \sum\limits_{i=1}^{n}\alpha_i - \frac{1}{2}\sum\limits_{i=1}^{n}\sum\limits_{l=1}^{n}\alpha_i\alpha_l y^i y^l k\left(\vec{x}^{\,i}, \vec{x}^{\,l}\right) \\ \text{ისეთი, რომ } \sum\limits_{i=1}^{n}\alpha_i y^i = 0; 0 \le \alpha_i \le C, i = 1,\ldots,n\end{array}\right\}. \tag{10.22}$$

მაშინ *გადაწყვეტილების მიღების ფუნქცია* მოცემული აღმოჩნდება გამოსახულებით (იხ. განტოლება 10.9) :

$$f\left(\vec{x}\right) = \sum_{i=1}^{n}\alpha_i^* y^i k\left(\vec{x}^{\,i}, \vec{x}\right) + b^* . \tag{10.23}$$

დამოუკიდებლად იმისა, გვინდა საყრდენ ვექტორთა მანქანის გაწვრთნა ან მისი გამოყენება, ჩვენ არ გვჭირდება ϕ-ის ცოდნა ცხადი სახით, ჩვენთვის საკმარისია k ბირთვის ცოდნა. ეს ნიშნავს, რომ ჩვენ არ გვჭირდება რაიმე გამოთვლების განხორციელება $\mathcal{H}$-ზე, რომელსაც, როგორც წესი, ძალიან მაღალი განზომილება აქვს : ეს ცნობილია *ბირთვული ტრიუკის*

(*ხრიკის, ოინის, ფანდის*) სახელწოდებით. ზოგადი სახით ბირთვული ტრიუკი წრფივი სწავლების სხვა ალგორითმის მიმართაც შეიძლება იყოს გამოყენებული, როგორიცაა, მაგალითად, თხემური რეგრესია (იხ. ქვედანაყოფი 6.2), Principal Component Analysis (PCM) - მთავარ კომპონენტთა მეთოდი (იხ. ქვედანაყოფი 11.3.1) ან k – საშუალოთა მეთოდი (იხ. ქვედანაყოფი 12.4).

3.4 ბირთვები

მათემატიკური დახასიათება

განსაზღვრება 10.10 (ბირთვი) *ბირთვი* ეწოდება ორი ცვლადის ნებისმიერ k ფუნქციას, რომელიც წარმოდგენილია ასახვათა სკალარული ნამრავლით თავისი ცვლადების ჰილბერტის სივრცეში. ასე რომ, ბირთვი — ეს უწყვეტი, სიმეტრიული და ნახევრად განსაზღვრული დადებითი ფუნქციაა

$$\forall N \in \mathbb{N}, \forall \left(\vec{x}^1, \vec{x}^2, \ldots, \vec{x}^N\right) \in \mathcal{X}^N \text{ და } \left(\alpha_1, \alpha_2, \ldots, \alpha_N\right) \in \mathbb{R}^N, \sum_{i=1}^{N}\sum_{l=1}^{N} \alpha_i \alpha_l k\left(\vec{x}^i, \vec{x}^l\right) \geq 0 .$$

⚙

განსაზღვრება 10.11 (გრამის მატრიცა) თუ მოცემულია n რაოდენობის $\vec{x}^1, \vec{x}^2, \ldots \vec{x}^n \in \mathcal{X}^n$ დაკვირვება და k ბირთვი $\mathcal{X}$-ზე, მაშინ ამ დაკვირვებათა *გრამის მატრიცა, ანუ გრამიანი* (ინგლ. *Gram matrix*, *Gramian matrix* ან უბრალოდ *Gramian*) ეწოდება $K \in \mathbb{R}^{n \times n}$ მატრიცას, ისეთს, რომ $K_{il} = k\left(\vec{x}^i, \vec{x}^l\right)$. ეს მატრიცა დადებითად ნახევრადგანსაზღვრულია.

მატრიცის სახელწოდება დაკავშირებულია გამოჩენილი დანიელი მათემატიკოსის იორგენ გრამის (Jørgen Pedersen Gram, 1850 – 1916) სახელთან.

⚙

თეორემა 10.3 (მურ-არონშაინის) *ნებისმიერი დადებითი ნახევრადგანსაზღვრული სიმეტრიული* $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ *ფუნქციისათვის, არსებობს ჰილბერტის* $\mathcal{F}$ *სივრცე და* $\psi : \mathcal{X} \to \mathcal{F}$ *ასახვა, ისეთი რომ ყველა* $\vec{x}, \vec{x}' \in \mathcal{X} \times \mathcal{X}$ *წყვილისათვის ადგილი აქვს* $k(\vec{x}, \vec{x}') = \left\langle \psi\left(\vec{x}\right), \psi\left(\vec{x}'\right) \right\rangle_{\mathcal{F}}$ *თანაფარდობას.*

⚙

მტკიცებულება. ამ შედეგის მტკიცებულება შეიძლება ვიპოვოთ პოლონური ფესვების ამერიკელი მათემატიკოსის ნახმან არონშაინის (Nachman Aronszajn, 1907 – 1980) ორიგინალურ სტატიაში (Aronszajn, 1950), სადაც იგი ამ თეორემის ფორმულირებას მიაწერს ამერიკელ მათემატიკოსს ელაიაკიმ ჰეისტინგ მურს (Eliakim Hastings Moore. 1862 – 1932).

□

ინტუიციურად ბირთვი შეიძლება იყოს ინტერპრეტირებული სკალარულ ნამრავლად ჰილბერტის სივრცეზე, სხვა სიტყვებით, ფუნქციად, რომელიც ზომავს *მსგავსებას* $\mathcal{X}$-ის ორ ობიექტს შორის. ამრიგად, ბირთვები შეიძლება განისაზღვროს ობიექტებს შორის მსგავსების აგებით და შემდეგ იმის შემოწმებით, რომ ეს მსგავსება დადებითად ნახევრადგანსაზღვრულია

ბირთვები ნამდვილი ვექტორებისათვის

თუ $\mathcal{X}=\mathbb{R}^p$, მაშინ მურ-არონშაინის თეორემა შემდეგი ბირთვების განსაზღვრის საშუალებას იძლევა.

განსაზღვრება 10.12 (კვადრატული ბირთვი) ბირთვს, რომელიც განისაზღვრება $k\left(\vec{x},\vec{x}'\right)=\left(\left\langle \vec{x},\vec{x}'\right\rangle +c\right)^2$ თანაფარდობით, სადაც $c\in\mathbb{R}^+$, *კვადრატული ბირთვი* ეწოდება.

ამ ბირთვის შესაბამისი ასახვა ასეთია :

$$\phi:\vec{x}\mapsto\left(x_1^2,\ldots,x_p^2,\sqrt{2}x_1x_2,\ldots,\sqrt{2}x_1x_p,\ldots,\sqrt{2}x_{p-1}x_p,\ldots,\sqrt{2c}x_1,\ldots,\sqrt{2c}x_p,c\right).$$

ამრიგად, ϕ ასახვას აქვს მნიშვნელობა $2p+\frac{p\left(p-1\right)}{2}+1$ განზომილების სივრცეზე : k ბირთვით და ბირთვული ტრიუკით (ხრიკით, ოინით) სარგებლობა იქნება უფრო ეფექტური, ვიდრე დაკვირვებათა სახეების გამოთვლა $\phi-$ით საყრდენ ვექტორთა მანქანის ალგორითმის გამოყენებამდე მათ მიმართ.

განსაზღვრება 10.13 (პოლინომური ბირთვი) $d\in\mathbb{N}$ რიგის *პოლინომური ბირთვი* ეწოდება შემდეგი სახით განსაზღვრულ ბირთვს :

$$k\left(\vec{x},\vec{x}'\right)=\left(\left\langle \vec{x},\vec{x}'\right\rangle +c\right)^d.$$

კვადრატული ბირთვის მსგავსად, $c\in\mathbb{R}^+$ წარმოადგენს პარამეტრს, რომელიც d-ზე ნაკლები ხარისხის წევრთა ჩართვის საშუალებას იძლევა. ამ ბირთვს შეესაბამება ხელახალი აღწერის სივრცე განზომილებათა იმდენი რიცხვით, რამდენიც არსებობს p ცვლადთა მონომი (ერთწევრი) d-ზე ნაკლები ან მისი ტოლი ხარისხით, ე.ი. $\left(\begin{array}{c}p+d\\d\end{array}\right)$ რიცხვით.

განსაზღვრება 10.14 (გაუსის რადიალური ბირთვი) *გაუსის რადიალური ბირთვი,* ანუ RBF ბირთვი (ინგლ. Radial Basis Function - რადიალური საბაზისო ფუნქცია), გატარების $\sigma>0$ ზოლით, — ეს ბირთვია, რომელიც $k\left(\vec{x},\vec{x}'\right)=\exp\left(-\frac{\left\|\vec{x}-\vec{x}'\right\|^2}{2\sigma^2}\right)$ ფორმულით განისაზღვრება.

ეს ბირთვი შეესაბამება *უსასრულო* განზომილების ხელახალი აღწერის სივრცეს. მართლაც, თუ გამოვიყენებთ ექსპონენციალური ფუნქციის დაშლას მთელრიცხვა მწკრივად, შეიძლება ამ ბირთვის წარმოდგენა შემდეგი სახით :

$$\left.\begin{aligned} k\left(\vec{x},\vec{x}'\right) &= \exp\left(-\frac{\left\|\vec{x}\right\|^2}{2\sigma^2}\right)\exp\left(-\frac{\left\langle\vec{x},\vec{x}'\right\rangle}{2\sigma^2}\right)\exp\left(-\frac{\left\|\vec{x}'\right\|^2}{2\sigma^2}\right) \\ &= \psi\left(\vec{x}\right)\sum_{r=0}^{+\infty}\left(-\frac{\left\langle\vec{x},\vec{x}'\right\rangle^r}{\sigma^{2r}r!}\right)\psi\left(\vec{x}'\right) = \sum_{r=0}^{+\infty}\left(-\frac{\left\langle\psi\left(\vec{x}\right)^{1/r}\vec{x},\psi\left(\vec{x}'\right)^{1/r}\vec{x}'\right\rangle^r}{\sigma^{2r}r!}\right) \end{aligned}\right\},$$

სადაც $\psi:\mathbb{R}^p\to\mathbb{R},\vec{x}\mapsto\exp\left(-\frac{\left\|\vec{x}\right\|^2}{2\sigma^2}\right)$.

ამ სამი ბირთვისათვის უფრო მარტივია და ეფექტურიც ბირთვის პირდაპირი და უშუალო გაანგარიშება $\mathbb{R}^p$ სივრცის ორ ვექტორს შორის, ვიდრე მათი სახეების სკალარული ნამრავლის გამოთვლა ხელახალი აღწერის შესაბამის სივრცეში, რომელიც, შეიძლება, უსასრულო განზომილების აღმოჩნდეს კიდეც.

ბირთვები სიმბოლური სტრიქონებისათვის

ბირთვული ტრიუკი ასევე იძლევა რთულ მონაცემებთან მუშაობის საშუალებას ისეთნაირად, რომ ეს არ მოითხოვდეს ამ მონაცემების წინასწარ გადაყვანას ფიქსირებული სიგრძის ვექტორულ წარმოდგენაში. ეს განსაკუთრებით აქტუალურია *სიმბოლური სტრიქონების* სახით მოცემული მონაცემებისათვის, როგორიცაა ტექსტი ან ბიოლოგიური ჯაჭვები, მაგალითად, ნუკლეინური ოთხი ფუძის შესაბამისი ალფაბეტის ოთხი ასოთი განსაზღვრული დეზოქსირიბონუკლეინის მჟავა (დნმ) ან ოცდაერთი ამინომჟავის ალფაბეტით განსაზღვრული ცილები.

$\mathcal{A}$ ალფაბეტის მოცემისას ერთდროულად შეიძლება $\mathcal{X}=\mathcal{A}^*$ აღნიშვნის გამოყენება $\mathcal{A}$-ზე განსაზღვრულ სიმბოლოთა სტრიქონების სიმრავლისათვის. ბირთვების უმრავლესობა $\mathcal{X}$-ზე იმ გარემოებიდან განისაზღვრება, რომ, რაც უფრო მეტი საერთო ქვესტრიქონი აქვს ორ, x და x' სტრიქონს, მით უფრო მსგავსია ისინი ერთმანეთის. შემოვიღოთ ქვესტრიქონების სიგრძისათვის $k\in\mathbb{N}$ აღნიშვნა და გარდავსახოთ x სტრიქონი $\left|\mathcal{A}\right|^k$ სიგრძის ვექტორად $\phi:x\mapsto\left(\psi_u\left(x\right)\right)_{u\in\mathcal{A}^k}$ ასახვის საშუალებით, სადაც $\psi_u\left(x\right)$ ნოტაცია გვიჩვენებს u სიმბოლოს რაოდენობას x სტრიქონში. ψ შეიძლება იყოს ისეთნაირად მოდიფიცირებული, რომ დასაშვები გახდეს რეგულირების გარკვეული უზუსტობანი ან ხარვეზები, «ხვრელები» (ინგლ. *gaps*) ამისათვის ჯარიმის ან სხვა სასჯელის შემოღებით.

მაშინ ჩვენ შეგვიძლია შემდეგი სახით განვსაზღვროთ *ბირთვი სიმბოლური სტრიქონებისთვის*:

$$\left.\begin{aligned} k:\mathcal{A}^*\times\mathcal{A}^* &\to\mathbb{R} \\ x,x' &\mapsto\sum_{u\in\mathcal{A}^k}\psi_u\left(x\right)\psi_u\left(x'\right) \end{aligned}\right\}.$$

ფორმალურად ეს ბირთვი მოთხოვს ჯამის გამოანგარიშებას $\left|\mathcal{A}^k\right|=\left|\mathcal{A}\right|^k$ წევრით. მაგრამ მისი გამოთვლა გაცილებით უფრო ეფექტურად შეიძლება მხოლოდ იტერაციის განხორციელებით

x-ში წარმოდგენილი k სიგრძის $\left(|x|+1-k\right)$ რაოდენობის სტრიქონზე, ამ დროს ჯამის დანარჩენი წევრები აუცილებლად უნდა იყოს 0-ის ტოლი. ასე რომ, საქმე ეხება $\mathcal{O}\left(|x|+|x'|\right)$ სირთულის გამოთვლას.

მაგალითი

ადამიანის ცილების შემთხვევაში, თუ ჩვენ ვირჩევთ $k=8$ სიდიდეს, ამით ვცლით გამოთვლას 37 მილიარდზე $\left(21^8\right)$ მეტი განზომილების ხელახალი აღწერის სივრცეში 500-ზე ნაკლები შესაკრების შემცველი ჯამით (ადამიანის ცილის საშუალო სიგრძე 485 ამინომჟავას შეიცავს)

ეს იდეები შეიძლება იყოს გამოყენებული ასევე *გრაფთა ბირთვების* განსაზღვრისათვის, თუ შევცვლით სიმბოლოთა ქვესტრიქონებს ქვეგრაფებით. მაგრამ ქვეგრაფების იზომორფიზმის პრობლემა ზოგად შემთხვევაში არის NP-სრული, რაც ზღუდავს განსახილველად შესაძლებელი ქვეგრაფების ტიპებს.

გრაფთა იზომორფიზმის ამოცანა თავისთავად მარტივად გამოიყურება : საჭიროა განისაზღვროს ორი გრაფის იზომორფულობა, ანუ სხვანაირად, დადგინდეს წვეროების უბრალო გადაადგილებით ერთი გრაფის ტრანსფორმირების შესაძლებლობა მეორე გრაფად კავშირების შენარჩუნებით წვეროებს შორის. ესაა და ეს. მიუხედავად ესოდენ მოჩვენებითი სიმარტივისა, ამ ამოცანის გადაჭრა ძნელია, ვინაიდან მცირე ზომის გრაფებსაც კი შეუძლია უამრავი განსხვავებული ფორმის მიღება.

აღსანიშნავია, რომ 2015-2017 წლებში გამოჩნდა «Graph Isomorphism in Quasipolynomial Time» თემატიკის ნაშრომები უნგრული წარმომავლობის ამერიკელი მათემატიკოსის ლასლო ბაბაის (László Babai, 1950-) ჩიკაგოს უნივერსიტეტის კომპიუტერული მეცნიერებისა და მათემატიკის დეპარტამენტიდან. ამ ნაშრომებში მან წარმოადგინა სწრაფი ახალი ალგორითმი გრაფების იზომორფიზმის ამოცანის გადასაჭრელად — გამოთვლითი სირთულის თეორიის ერთ-ერთი ფუნდამენტური პრობლემის. ზოგიერთი ექსპერტის აზრით, ეს არის ერთ-ერთი ყველაზე მნიშვნელოვანი შედეგი თეორიულ კომპიუტერულ მეცნიერებაში ბოლო რამდენიმე ათწლეულის მონაკვეთზე. მიჩნეულია, რომ მისი ალგორითმით გამოთვლა კვაზი-პოლინომურ $\exp(P((\log n)))$ დროს მოითხოვს, სადაც n წვეროთა რიცხვია, $P\left(\log\left(n\right)\right)$ კი $\log\left(n\right)$ არგუმენტის პოლინომია.

ბირთვები სიმრავლეებისათვის

ჩვენ მიერ მე-8 ლექციის 3.3 სექციაში განსაზღვრული მსგავსებებიდან სიმრავლეთა შორის, ზოგიერთი სინამდვილეში ბირთვს წარმოადგენს.

თეორემა 10.4 *მსგავსება შვეიცარიელი ბოტანიკოსის პოლ ჟაკარის (Paul Jaccard, 1868–1944) მიხედვით და მსგავსება მინიმაქსის (MinMax) პრინციპის მიხედვით ბირთვებია.*

მტკიცებულება. ამ თეორემის მტკიცებულება ეფუძნება ჯონ გაუერის (John Clifford Gower, 1930–2019) შედეგს : Gower (1971).

□

4 თხემური რეგრესია ბირთვული ფუნქციის გამოყენებით

თეორემა 10.5 *ბირთვული ტრიუკი (ხრიკი, ოინი) წრფივ თხემურ რეგრესიაშიც გამოიყენება.*

✿

მტკიცებულება. გავიხსენოთ, რომ თხემური რეგრესიის მაპროგნოზირებელი ფუნქცია შემდეგი სახისაა :

$$f\left(\vec{x}\right)=\left(\vec{x},\vec{\beta}^{*}\right), \tag{10.24}$$

სადაც $\vec{\beta}^{*}$ მოცემულია განტოლებით 6.7 :

$$\vec{\beta}^{*}=\left(\lambda I_{p}+X^{\mathrm{T}}X\right)^{-1}X^{\mathrm{T}}\vec{y}\,. \tag{10.25}$$

აქ $X\in\mathbb{R}^{n\times p}$ — საწყისი (საანგარიშო) მატრიცაა, რომელიც — (5.5) გარდაქმნის შესაბამისად — ფართოვდება $1-$ანების შემცველი სვეტით. გამოსახულება (10.25), მარცხნიდან $\left(\lambda I_{p}+X^{T}X\right)$ ჯამზე გამრავლებით, შეიძლება შემდეგი სახით ჩაიწეროს :

$$\vec{\beta}^{*}=\mathrm{X}^{\mathrm{T}}\vec{\alpha}\text{ , სადაც }\vec{\alpha}=\frac{1}{\lambda}\left(y-X\vec{\beta}^{*}\right).$$

ასე რომ, $\lambda\vec{\alpha}=y-XX^{\mathrm{T}}\vec{\alpha}$ და, ამრიგად,

$$\vec{\alpha}=\left(\lambda I_{n}+XX^{\mathrm{T}}\right)^{-1}y.$$

განტოლება 10.24 ახლა შემდეგ ფორმას იძენს :

$$f\left(\vec{x}\right)=\vec{x}X^{\mathrm{T}}\vec{\alpha}=\vec{x}X^{\mathrm{T}}\left(XI_{n}+XX^{\mathrm{T}}\right)^{-1}y\,. \tag{10.26}$$

ახლა დავუშვათ, რომ $k:\mathcal{X}\times\mathcal{X}\mapsto\mathbb{R}$ — ბირთვია. არსებობს ჰილბერტის $\mathcal{H}$ სივრცე და $\phi:\mathcal{X}\to\mathcal{H}$ ასახვა, ისეთი რომ ყველა $\vec{x},\vec{x}'\in\mathcal{X}\times\mathcal{X}$-თვის k ბირთვი $k\left(\vec{x},\vec{x}'\right)=\left\langle\vec{x},\vec{x}'\right\rangle_{\mathcal{H}}$ ფორმულით აისახება.

თხემური რეგრესიის გამოყენება დაკვირვებათა სახეების მიმართ $\mathcal{H}$ სივრცეზე შემდეგი თანაფარდობის გამოთვლამდე დაიყვანება :

$$f_{\phi}\left(\vec{x}\right)=\phi\left(\vec{x}\right)\Phi^{T}\left(\lambda I_{n}+\Phi\,\Phi^{T}\right)^{-1}y\,,$$

სადაც $\Phi\in\mathbb{R}^{n\times d}$ — მატრიცაა, რომლის $i-$ური სტრიქონი, რომელიც წარმოადგენს ϕ ასახვით $\vec{x}^{i}$ მონაცემების d განზომილების $\mathcal{H}$ სივცეზე მოცემულ სახეს. ეს ფუნქცია შეიძლება ხელახლა იყოს გადაწერილი ასეთი სახით :

$$f_{\phi}\left(\vec{x}\right)=k\left(\lambda I_{n}+K\right)^{-1}y, \tag{10.27}$$

სადაც $k \in \mathbb{R}^n$ — ვექტორია, რომლის i – ური შემავალი ელემენტი ასახულია

$$k_i = \left\langle \phi\left(\vec{x}\right), \phi\left(\vec{x}^i\right) \right\rangle_{\mathcal{H}} = k\left(\vec{x}, \vec{x}^i\right)$$

ფორმულით და $K \in \mathbb{R}^{n \times n}$ — მატრიცაა, რომლის შემავალი K_{il} ელემენტი ასე ჩაიწერება :

$$K_{il} = \left\langle \phi\left(\vec{x}^i\right), \phi\left(\vec{x}^l\right) \right\rangle_{\mathcal{H}} = k\left(\vec{x}^i, \vec{x}^l\right).$$

ამიტომ არ არის ϕ-ის და $\mathcal{H}$-ის ცხადი ფორმით გამოყენების აუცილებლობა თხემური რეგრესიის გასაწვრთნელად ხელახალი აღწერების $\mathcal{H}$ სივრცეზე, რომელიც განსაზღვრულია ბირთვით.

□

5 საკვანძო მომენტები

— საყრდენი ვექტორების მანქანა რბილი (მოქნილი) ნაპრალით — ეს ბინარული კლასიფიკაციის წრფივი ალგორითმია, რომელიც ცდილობს კლასებს შორის ნაპრალის ცნებით ფორმალიზებული განცალკევების მაქსიმიზაციას და იმავდროულად დამსწავლელ (საწვრთნელ) სიმრავლეზე კლასიფიკაციის შეცდომათა რაოდენობის გაკონტროლებას.

— ბირთვული ტრიუკი (ხრიკი) ამ ალგორითმის ეფექტურად გამოყენების საშუალებას იძლევა ისევე, როგორც სხვა წრფივი ალგორითმის (მაგალითად, თხემური რეგრესიის) ხელახალი აღწერის სივრცეში და თანაც დაკვირვებათა სახეების გამოუთვლელად ცხადი ფორმით ამ სივრცეში.

— კვადრატული, პოლინომური და რადიალური გაუსის ბირთვები სულ უფრო და უფრო დიდი განზომილების ხელახალი აღწერის სივრცეთა გამოყენების საშუალებას იძლევა.

დამატებითი ინფორმაცია

- http://www.kernel-machines.org/ საიტზე წარმოდგენილია ბირთვულ მეთოდებთან დაკავშირებული ბიბლიოგრაფიული და პროგრამული რესურსების დიდი რაოდენობა.

- საყრდენი ვექტორების მანქანას მრავალი ნაშრომი ეძღვნება ; მათ შორისაა Schölkopf და Smola (2002), ასევე ცალკეული თავები საყრდენი ვექტორების მანქანის შესახებ წიგნებში Vladimir Cherkassky და Filip Mulier (1998), Vert et al. (2004) და Burges (1998).

- საყრდენი ვექტორების მანქანათა ალგორითმები სწრაფად აღმოჩნდა გადატანილი რეგრესიათა ამოცანებზე. ეს ეხება, მაგალითად, ℓ_2 ნორმით რეგულარიზებული წრფივი რეგრესიის ამოცანას, მაგრამ ϵ – უგრძნობელი დანაკარგებით (იხ. ქვედანაყოფი 2.4.3) დანახარჯების ფუნქციის როლში. ამჟამად მსგავს ალგორითმებს ზოგჯერ *SVR ალგორითმებს* უწოდებენ (*SVR* აკრონიმი ნაწარმოებია ინგლისური ფრაზიდან *Support Vector Regression*, რომელიც ქართულად ითარგმნება როგორც *საყრდენი ვექტორის რეგრესია*). უფრო დაწვრილებითი ინფორმაციის მოპოვება შეიძლება დამხმარე სახელმძღვანელოში Smola and Schölkopf (1998).

- ტიპური კვადრატული ოპტიმიზატორების ალტერნატივას *SMO ალგორითმი* (ინგლ. *Sequential Minimal Optimisation* — მიმდევრობითი მინიმალური ოპტიმიზაცია) წარმოადგენს. იგი პირველად გასაჯაროვდა ტექნიკურ ანგარიშში John Platt (1998). ამ ალგორითმმა ხელი შეუწყო საყრდენი ვექტორების მანქანის SVM ალგორითმების გავრცელებას და დააჩქარა მათი ოპტიმიზაცია. სწორედ ეს ალგორითმია რეალიზებული ბიბლიოთეკაში LibSVM (Chang and Lin, 2008), რომელიც დღემდე რჩება ერთ-ერთ ყველაზე გავრცელებულ ბიბლიოთეკად საყრდენ ვექტორთა მანქანის გასაწვრთნელად ; უფრო მეტიც, scikit-learn (მანქანური სწავლების ბიბლიოთეკა Python ენაზე) და Shogun (უფასო პროგრამული ბიბლიოთეკა ღია საწყისი კოდით მანქანური სწავლებისათვის C++ ენაზე), მაგალითად, ეფექტურად იყენებს მას უკვე ხელმეორედ თავისი მიზნებისათვის.

- ცნობილია, რომ SVM საყრდენი ვექტორების მანქანათა გადაწყვეტილების მიღების ფუნქცია ალბათობის მოდელირებას ვერ ახდენს. ამ ნაკლოვანების აღმოსაფხვრელად ნაშრომში Platt (1999) შემოთავაზებულია სიგმოიდური ფუნქციის გაწვრთნა, რომელიც წინასწარმეტყველებათა კონვერტირებას, გარდასახვას განახორციელებს ალბათობებად.

- საყრდენი ვექტორების მანქანა შეიძლება გაფართოვდეს ისეთ დაკვირვებათა მოდელირებისათვის, რომლებიც ერთსა და იმავე კლასს მიეკუთვნება (უარყოფითი მაგალითების გარეშე). შემდეგ ეს მეთოდი ახალ დაკვირვებათა კლასიფიცირების საშუალებას იძლევა იმისდა მიხედვით, მიეკუთვნება ისინი თუ არა კლასს და, ამრიგად, შეიძლება იყოს გამოყენებული ანომალიების გამოსავლენად, ე.ი. დაკვირვებების, რომლებიც არ მიეკუთვნება კლასს. არსებობს ორი ვარიანტი : *ერთკლასიანი საყრდენი ვექტორების მანქანა — One-class SVM* (Schölkopf et al., 2000), რომელიც აცალკევებს, გამოყოფს დაკვირვებებს წყაროსგან (სათავისგან) ; და *საყრდენი ვექტორით მონაცემთა აღწერის* მეთოდი — *Support Vector Data Description* (Tax and Duin, 2004), რომელიც იმ სფერული საზღვრის პოვნის საშუალებას იძლევა (ხელახალი აღწერის სივრცეში, ბირთვული ვერსიის გამოყენებისას), რომლითაც მთლიანად გარშემორტყმულია მონაცემები.

- უფრო დაწვრილებით ბირთვებზე სიმბოლური სტრიქონებისათვის და ამ ბირთვების გამოყენების შესახებ ბიოინფორმატიკაში შეიძლება მივმართოთ «Perspectives of Neural-Symbolic Integration» კრებულის (Editors: Barbara Hammer, Pascal Hitzler) თავს «Kernels for Strings and Graphs», რომლის ავტორებია Craig Saunders & Anthony Demco (2007). რაც შეეხება ბირთვებს გრაფებისათვის, ამ სფეროს კარგ მიმოხილვას იძლევა M. Benoît Gaüzère (2013) დისერტაციის მეორე თავი.

6 ბიბლიოგრაფია

1. Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3) :337–404.

2. Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth annual Workshop on Computational Learning Theory*, pages 144–152, Pittsburgh, Pennsylvania, United States. ACM.

3. Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2 :121–167.

4. Chang, C.-C. and Lin, C.-J. (2008). LibSVM : A library for support vector machines. http://www.kernel-machines.org/ .

5. Cherkassky, V. and Mulier, F. (1998). *Learning from Data : Concepts, Theory, and Methods*. John Wiley and Sons, New York, 538 pages.

6. Cortes, C. et Vapnik, V. (1995). Support vector networks. *Machine Learning*, 20 :273–297.

7. Gaüzère, B. (2013). *Application des méthodes à noyaux sur graphes pour la prédiction des propriétés des molécules.* Thèse de doctorat, Spécialité : Traitement du signal et des images, Université de Caen Basse-Normandie, Directeur de thèse : Pr. Luc Brun, Co-directeur de thèse : Pr. Didier Villemin, 198 pages.

8. Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, pages 857–871.

9. Platt, J. C. (1998). Sequential minimal optimization : a fast algorithm for training support vector machines.Technical Report MSR-TR-98-14, Microsoft Research.

10. Platt, J. C. (1999). Probabilities for support vector machines. In *Advances in Large Margin Classifiers*, pages 61–74. MIT Press, Cambridge, MA.

11. Saunders, C. and Demco, A. (2007). Kernels for strings and graphs. In *Perspectives of Neural-Symbolic Integration*, Studies in Computational Intelligence, pages 7–22. Springer, Berlin, Heidelberg.

12. Schölkopf, B., Williamson, R. C., Smola, A. J., Shawe-Taylor, J., and Platt, J. C. (2000). Support vector method for novelty detection. In *Advances in neural information processing systems*, pages 582–588.

13. Schölkopf, B. and Smola, A. J. (2002). *Learning with Kernels : Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA.

14. Smola, A. J. and Schölkopf, B. (1998). A tutorial on support vector regression. *NeuroCOLT*, TR-1998-030.

15. Tax, D. M. and Duin, R. P. (2004). Support vector data description. *Machine learning*, 54(1) :45–66.

16. Vapnik, V. and Lerner, A. (1963). Pattern recognition using generalized portrait method. *Automation and Remote Control*, 24.

17. Vert, J.-P., Tsuda, K., and Schölkopf, B. (2004). A primer on kernel methods. In *Kernel Methods in Computational Biology*, pages 35–70. MIT Press, Cambridge, MA.

7 სავარჯიშოები

10.1 ბირთვები. შემდეგ ფუნქციებს შორის, რომელი არის ბირთვი ?

1.

$$k:[0,1]\times[0,1]\to\mathbb{R}$$
$$x,x'\mapsto\min(x,x').$$

2.

$$k:\mathbb{R}\times\mathbb{R}\to\mathbb{R}$$
$$x,x'\mapsto\min(x,x').$$

3.

$$k:\mathbb{R}^p\times\mathbb{R}^p\to\mathbb{R}$$
$$x,x'\mapsto\sqrt{\|\vec{x}-\vec{x}'\|_2^2+1}.$$

4.

$$k:\mathbb{R}^p\times\mathbb{R}^p\to\mathbb{R}$$
$$x,x'\mapsto\prod_{j=1}^{p}h\left(\frac{x_j-a}{b}\right)h\left(\frac{x'_j-a}{b}\right),$$

სადაც $h:u\mapsto\exp\left(-u^2\right)$ და $(a,b)\in\mathbb{R}\times\mathbb{R}^*$.

10.2 $\mathbb{R}^p$ სივრცეში n რაოდენობის $\vec{x}^1,\vec{x}^2,\ldots,\vec{x}^n$ დაკვირვებათა და მათი ბინარული $y^1,y^2,\ldots,y^n$ ჭდეების გათვალისწინებით, ანიკას სურს ჩაუტაროს სწავლება საყრდენი ვექტორების მანქანას (ინგლ. SVM, Support Vector Machine), რომელსაც $\sigma>0$ სიგანის რადიალური გაუსის ბირთვის გატარების ზოლი აქვს.

თავდაპირველად ანიკა ამ ბირთვის იორგენ გრამის (დანიელი მათემატიკოსი Jørgen Pedersen Gram, 1850-1916) მატრიცას განიხილავს საწვრთნელ მონაცემებზე.

ამის შედეგად იგი იმ დასკვნამდე მიდის, რომ მნიშვნელობები დიაგონალზე რამდენიმე რიგით აღემატება არადიაგონალურ მნიშვნელობებს.

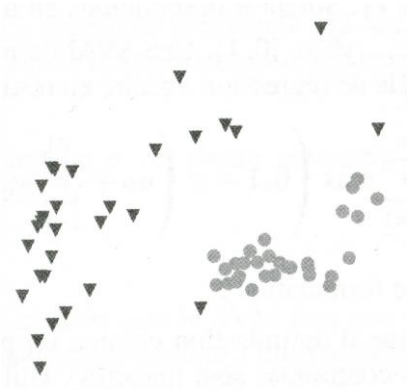
1. შეძლებს საყრდენი ვექტორების მანქანა კარგად დასწავლას ?

2. შეძლებს ანიკა თავისი საყრდენი ვექტორების მანქანის გაუმჯობესებას σ გატარების ზოლის მნიშვნელობის შეცვლით ?

როგორ უნდა განახორციელოს ანიკამ ეს ?

10.3 კვადრატული საყრდენი ვექტორების მანქანა (ინგლ. Support Vector Machine Quadratic). ბაადურს აქვს ორგანზომილებიანი ბინარული კლასიფიკაციის ამოცანა ქვემოთ ნახატზე წარმოდგენილი მონაცემებით.

იგი იყენებს საყრდენი ვექტორების მანქანას კვადრატული ბირთვით და ღირებულების C პარამეტრით.



1. როგორია გადაწყვეტილების მიღების საზღვარი C პარამეტრის «ძალიან მაღალი» მნიშვნელობისათვის ?

2. როგორია გადაწყვეტილების მიღების საზღვარი C პარამეტრის «ძალიან დაბალი» მნიშვნელობისათვის ?

3. ორი ვერსიიდან (C პარამეტრის მაღალი და დაბალი მნიშვნელობებით) რომელი განახორციელებს განზოგადებას უკეთ ?

4. C პარამეტრის *დიდი* მნიშვნელობისათვის დახატეთ «სამკუთხედი» კლასის დამატებითი დაკვირვება, რომელიც გავლენას არ მოახდენს გადაწყვეტილების მიღების საზღვარზე.

5. C პარამეტრის *მცირე* მნიშვნელობისათვის დახატეთ იმავე «სამკუთხედი» კლასის დამატებითი დაკვირვება, რომელიც გავლენას არ მოახდენს გადაწყვეტილების მიღების საზღვარზე.

10.4 როგორც ცნობილია, Lasso (ინგლ. Least absolute shrinkage and selection operator — უმცირესი აბსოლუტური შეკვეცა და არჩევის ოპერატორი) წრფივი რეგრესიული მოდელის კოეფიციენტთა შეფასების მეთოდია. შესაძლებელია *ბირთვული ტრიუკის* (*იინის, ფანდის, ხრიკის*) გამოყენება ამ მეთოდის მიმართ ისევე, როგორც *თხემური რეგრესიისადმი* (ინგლ. Ridge Regression) ?

10.5 კატოს აქვს მრავალი მილიონი დაკვირვების და მრავალი ასეული ცვლადის შემცველი მონაცემების კრებული. საყრდენ ვექტორთა წრფივი მანქანის (ინგლ. SVM, Support Vector Machine) წვრთნის ჩასატარებლად რას ურჩევდით კატოს : ოპტიმიზაციის ამოცანის პრიმალური (პირველადი, მთავარი, ძირითადი) სახის თუ მისი დუალური (მეორეული) ფორმის გამოყენებას ?

10.6 დიმიტრიმ გაწვრთნა საყრდენი ვექტორების მანქანა (ინგლ. SVM, Support Vector Machine) პოლინომური ბირთვით თავის მონაცემებზე და იმ დასკვნამდე მივიდა, რომ მისი საყრდენი ვექტორების მანქანა (ინგლ. SVM, Support Vector Machine), ეტყობა, საკმარისად ნასწავლი არ არის. რომელი ჰიპერპარამეტრ(ებ)ი უნდა შეიცვალოს და როგორ ?

10.7 რეგრესია საყრდენი ვექტორებით (ინგლ. **SVR, Support Vector Regression).** მანქანები საყრდენი ვექტორებით (ინგლ. SVM, Support Vector Machine) კარგად არის ცნობილი კლასიფიკაციის ამოცანებში. მაგრამ SVM-ის გამოყენება რეგრესიაში არ არის დამაჯერებლად

დასაბუთებული. მოდელების ეს ტიპები ცნობილია როგორც *რეგრესია საყრდენი ვექტორებით* (ინგლ. SVR, Support Vector Regression).

დავუშვათ, რომ გვაქვს $\mathcal{D}=\left(\vec{x}^i,y^i\right)_{i=1,..,n}$ საწვრთნელი სიმრავლე, სადაც $\vec{x}^i\in\mathbb{R}^p$ და $y^i\in\mathbb{R}$. დავუშვათ, არსებობს ისეთი $\varepsilon>0,\vec{w}\in\mathbb{R}^p,\beta\in\mathbb{R}$, რომ ყველა $i=1,\ldots,n$ ინდექსისათვის, y^i სიდიდე შეიძლება იყოს აპროქსიმირებული $f\left(\vec{x}^i\right)=\left(\vec{w},\vec{x}^i\right)+b$ გამოსახულებით გარკვეული ε სიზუსტით :

$$\left|\left\langle\vec{w},\vec{x}^i\right\rangle+b-y^i\right|\leq\varepsilon.$$

SVR-ის (საყრდენი ვექტორების რეგრესიის) მიზანია $\vec{w}$ და b სიდიდეთა პოვნა $\min_{\vec{w},b}\frac{1}{2}\left\|\vec{w}\right\|_2^2$ ამოცანის ამოხსნით

$$\left\langle\vec{w},\vec{x}^i\right\rangle+b-y^i\leq\varepsilon$$

და

$$-\left(\left\langle\vec{w},\vec{x}^i\right\rangle+b\right)+y^i\leq\varepsilon$$

შეზღუდვათა პირობებში ყველა $i=1,\ldots,n$ ინდექსისათვის.

1. ჩაწერეთ ოპტიმიზაციის ზემოთ მოცემული ამოცანის დუალური ფორმა.

2. ჩაწერეთ f ოპერაცია დუალური ცვლადების და არა $\vec{w}$ ვექტორის ფუნქციის სახით.

3. როგორი დაკვირვებებისათვის არის ლაგრანჟის მამრავლები ნულის ტოლი ?

4. თუ არის შესაძლებელი ბირთვის ტრიუკის (ოინის, ფანდის, ხრიკის) გამოყენება SVR-ის (Support Vector Regression — რეგრესია საყრდენი ვექტორებით) შემთხვევაში ?

10.8 ℓ_1 ნორმის საყრდენი ვექტორების მანქანა (SVM). დავუშვათ, რომ გვაქვს p განზომილების n რაოდენობის $\vec{x}^1,\vec{x}^2,\ldots,\vec{x}^n$ დაკვირვება და მათი $y^1,y^2,\ldots,y^n\in\{0,1\}$ ჭდეები. ℓ_1 ნორმის საყრდენი ვექტორების მანქანა წვრთნის (და ადგენს) წრფივი რეგრესიის მოდელის $w_0,w_1,\ldots,w_p$ წონებს შემდეგი ამოცანის გადაჭრის გზით :

$$\vec{w}^*\in\arg\min_{\vec{w}\in\mathbb{R}^{p+1}}\sum_{i=1}^{n}\max\left(0.1-y^i\left(w_0+\sum_{j=1}^{p}w_ix_j^i\right)\right)+\lambda\left\|\vec{w}\right\|_1.$$

1. როგორია ამ ფორმულირების ეფექტი ?

2. ხელახლა ჩამოაყალიბეთ ოპტიმიზაციის ეს ამოცანა წრფივი ამოცანის სახით, რომელშიც სამინიმიზაციო ფუნქციაც და შეზღუდვებიც წრფივია. ამისათვის გამოიყენეთ

$$\xi_i=\max\left(1-y^i\left(w_0+\sum\nolimits_{j=1}^{p}w_ix_j^i\right)\right)$$

გადახრის ან, რაც იგივეა, გაბნევის ცვლადები და დაშალეთ თითოეული w_j ცვლადი დადებითი და უარყოფითი ნაწილების $w_j = v_j^+ - v_j^-$ ჯამად.

3. ეფექტურობის რომელი მაჩვენებლის გამოყენებაა საჭირო ალგორითმის შესაფასებლად მონაცემებზე ?

4. როგორ ავირჩიოთ λ და შევაფასოთ მოდელის უნარი განახორციელოს განზოგადება ?

5. თუ ახდენს ცვლადების ნორმალიზაცია გავლენას მოცემულ ალგორითმზე ?

სავარჯიშოთა ამონახსნები

10.1

1. კი. ქვემოთ ჩამოთვლილი ყველა სიდიდისათვის

$$n \in \mathbb{N}, \left(a_1, a_2, \ldots, a_n\right) \in \mathbb{R}^n, x^1, x^2, \ldots, x^n \in [0,1]:$$

$$\sum_{i=1}^n \sum_{j=1}^n a_i a_j k\left(x^i, x^j\right) = \sum_{i=1}^n a_i a_j \int_0^1 \mathbb{I}_{t \le x^i} \int_0^1 \mathbb{I}_{t \le x^j} dt$$

$$= \int_0^1 \left(\sum_{i=1}^n a_i \mathbb{I}_{t \le x^i} \right) \left(\sum_{j=1}^n a_j \mathbb{I}_{t \le x^j} \right) dt,$$

რაც არის, ამრიგად, $x \mapsto \sum_{i=1}^n a_j \mathbb{I}_{t \le x^i}$ ფუნქციის კვადრატის ინტეგრალი და, მაშასადამე, იგი დადებითია.

2. არა. ავიღოთ $x^1 = 0, x^2 = 1$ მნიშვნელობები.

$$K = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}$$

მატრიცა არ არის განსაზღვრული, რადგან

$$\left(a_1, a_2\right) K \left(a_1, a_2\right)^{\mathrm{T}}$$

ნამრავლი შეიძლება იყოს დადებითი ან უარყოფითი.

3. კი, ვინაიდან

$$\vec{x}, \vec{x}' \mapsto \left\| \vec{x} - \vec{x}' \right\|_2^2$$

არის ბირთვი.

4. კი, რადგან

$$\prod_{j=1}^p h\left(\frac{x_j - a}{b} \right) h\left(\frac{x_j' - a}{b} \right) = g\left(\vec{x}\right) g\left(\vec{x}'\right),$$

სადაც

$$g : \vec{x} \mapsto \prod_{j=1}^{p} h\left(\frac{x_j - a}{b}\right).$$

10.2

1. გრამის მატრიცას მადომინირებელი დიაგონალი აქვს : $\vec{x}^i$-ის მსგავს ერთადერთ დაკვირვებას თავად ეს დაკვირვება წარმოადგენს, დანარჩენი დაკვირვებები კი დაშორებულია თანაბრად. ასე რომ, საყრდენი ვექტორების მანქანის (ინგლ. SVM, Support Vector Machine) კარგად გაწვრთნა არ ხერხდება.

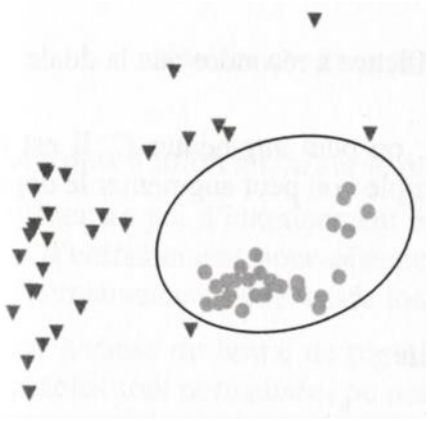
2. ბირთვის σ გამტარუნარიანობის გაზრდა უფრო გონივრულ მნიშვნელობებამდე მიგვიყვანს.

10.3

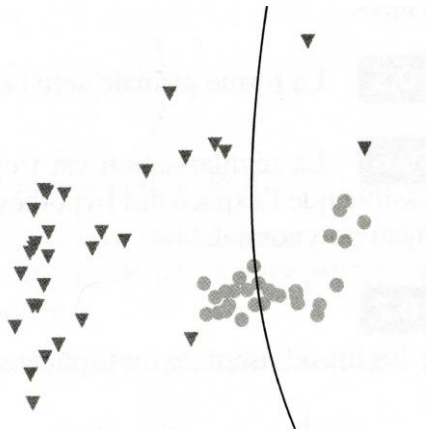
1. C სიდიდის მაღალი მნიშვნელობა ნიშნავს, რომ საყრდენი ვექტორების მანქანა (ინგლ. SVM, Support Vector Machine) შეცდომათა მცირე რაოდენობას დაუშვებს საწვრთნელ სიმრავლეზე.

კვადრატული ბირთვი განაპირობებს გადაწყვეტილების მიღების ელიფსოიდურ საზღვარს (იხ. ნახატი 10.1 *a.*).

2. C სიდიდის დაბალი, მცირე მნიშვნელობა ნიშნავს, რომ საყრდენი ვექტორების მანქანას (ინგლ. SVM, Support Vector Machine) ექნება კარგად ფუნქციონირების დიდი მარაგი, პოტენციალი, საწვრთნელ სიმრავლეზე შეცდომებსაც რომ უშვებდეს (იხ. ნახატი 10.1 *b.*).



a. როცა C დიდია.



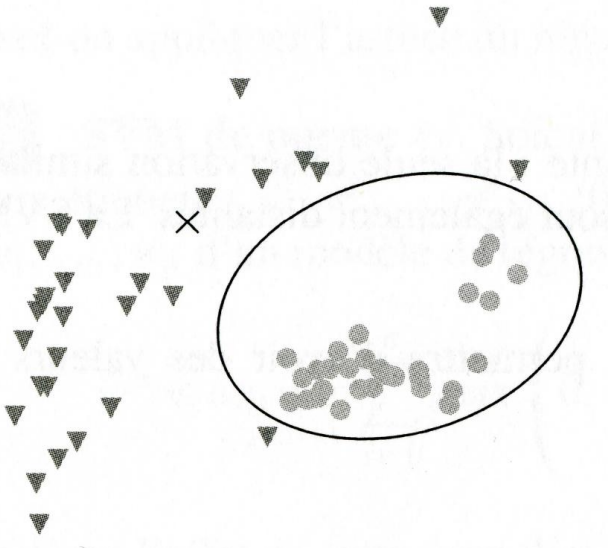
b. როცა C მცირეა.

ნახატი 10.5 - სავარჯიშო 10.3, შეკითხვები 1 და 2.

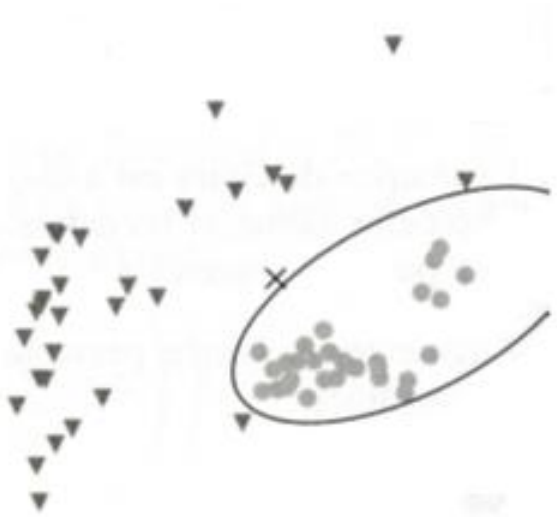
3. C სიდიდის დაბალი, მცირე მნიშვნელობა გადაჭარბებული სწავლების თავიდან აცილების საშუალებას იძლევა. წრეებთან ახლოს განლაგებული ორი სამკუთხედი აშკარად ანომალურია.

4. სწორად კლასიფიცირებული და გადაწყვეტილების მიღების საზღვრიდან დაშორებული დაკვირვებები გავლენას არ ახდენს ამ უკანასკნელზე (იხ. ნახატი 10.2 a.).

5. შეცდომით კლასიფიცირებული დაკვირვებები ან გადაწყვეტილების მიღების საზღვართან ახლოს განთავსებული დაკვირვებები გავლენას ახდენს ამ საზღვარზე (იხ. ნახატი 10.2b.).



a. «სამკუთხედი» კლასის ახალი $(\times)$ დაკვირვება გავლენას არ ახდენს გადაწყვეტილების მიღების საზღვარზე.



b. «სამკუთხედი» კლასის ახალი $(\times)$ დაკვირვება გავლენას ახდენს გადაწყვეტილების მიღების საზღვარზე.

ნახატი 10.6 - სავარჯიშო 10.3, შეკითხვები 4 და 5.

10.4 არა : ℓ_1 ნორმა არ ჩაიწერება როგორც სკალარული ნამრავლი სახეთა შორის.

10.5 პირველადი (პრიმალური, მთავარი, ძირითადი) ფორმა გაცილებით უფრო ეფექტურად იქნება გადაჭრილი, ვიდრე დუალური.

10.6 რეგულარიზაცია ძალზე დიდია :

შესაძლებელია C პარამეტრის მნიშვნელობის გაზრდა.

ასევე შესაძლებელია, რომ ჰიპოთეზათა სივრცე ზედმეტად მარტივია :

უნდა მოხდეს პოლინომური ბირთვის ხარისხის გაზრდა.

10.7

1. შემოვიტანოთ ლაგრანჟის $2n$ მამრავლი :

$$L = \frac{1}{2}\|\vec{w}\|_2^2 - \sum_{i=1}^{n} \alpha_i \left(\varepsilon - y^i + \left\langle \vec{w}, \vec{x}^i \right\rangle + b\right) + \\ -\sum_{i=1}^{n} \alpha_i^* \left(\varepsilon + y^i - \left\langle \vec{w}, \vec{x}^i \right\rangle - b\right)$$

,

სადაც $\alpha_i > 0, \alpha_i^* > 0$.

ლაგრანჟის გრადიენტი b ცვლადით გავუტოლოთ ნულს. მაშინ მივიღებთ :

$$\sum_{i=1}^{n} \left(\alpha_i - \alpha_i^*\right) = 0 .$$

ლაგრანჟის გრადიენტი $\vec{w}$ ცვლადით გავუტოლოთ ნულს. მაშინ გვექნება :

$$\vec{w} - \sum_{i=1}^{n} \left(\alpha_i - \alpha_i^*\right) \vec{x}^i = 0 .$$

ამრიგად, დუალურ ფორმას შემდეგი სახე აქვს :

$$\arg\max -\frac{1}{2}\sum_{i,j=1}^{n}\left(\alpha_i-\alpha_i^*\right)\left(\alpha_j-\alpha_j^*\right)\left\langle \vec{x}_i,\vec{x}_j\right\rangle + \\ -\varepsilon\sum_{i=1}^{n}\left(\alpha_i-\alpha_i^*\right)+\sum_{i=1}^{n}y^i\left(\alpha_i-\alpha_i^*\right)$$

სადაც შეზღუდვა შემდეგი ტოლობით ჩაიწერება :

$$\sum_{i=1}^{n}\left(\alpha_i-\alpha_i^*\right)=0\,.$$

2. $f\left(\vec{x}\right)=\sum_{i=1}^{n}\left(\alpha_i-\alpha_i^*\right)\left\langle \vec{x}^i,\vec{x}\right\rangle+b\,.$

3. კარუშ-კუნ-ტაკერის (Karush-Kuhn-Tucker) პირობები გვაძლევს :

$$\alpha_i\left(\varepsilon-y^i+\left\langle \vec{w},\vec{x}^i\right\rangle+b\right)=0$$

და

$$\alpha_i^*\left(\varepsilon+y^i-\left\langle \vec{w},\vec{x}^i\right\rangle-b\right)=0\,.$$

ასე რომ, $\alpha_i\neq 0$, თუ და მხოლოდ თუ

$$y^i-\left\langle \vec{w},\vec{x}^i\right\rangle+b=\varepsilon\,.$$

მსგავსად ამისა, $\alpha_i^*\neq 0$, თუ და მხოლოდ თუ

$$y^i-\left\langle \vec{w},\vec{x}^i\right\rangle+b=-\varepsilon.$$

ყველა დაკვირვება, რომლის ჭდე არის ε მანძილით დაშორებული ნაწინასწარმეტყველი სიდიდისგან, წარმოადგენს საყრდენ ვექტორს.

ხსენებული ვექტორები განსაზღვრავს $\vec{w}$ ვექტორს და, მაშასადამე, f ფუნქციასაც.

4. კი. როცა მოცემულია $k:\mathcal{X}\times\mathcal{X}\to\mathbb{R}$ ბირთვი, დუალური ფორმის გადაწერა შეიძლება ხორციელდებოდეს როგორც ხელახალი აღწერის სივრცის, ასევე შესაბამისი აპარატის (აპლიკაციის) ცხადი გამოყენების გარეშე :

$$\arg\max -\frac{1}{2}\sum_{i,j=1}^{n}\left(\alpha_i-\alpha_i^*\right)\left(\alpha_j-\alpha_j^*\right)k\left(\vec{x}_i,\vec{x}_j\right)+ \\ -\varepsilon\sum_{i=1}^{n}\left(\alpha_i-\alpha_i^*\right)+\sum_{i=1}^{n}y^i\left(\alpha_i-\alpha_i^*\right)$$

სადაც შეზღუდვა მოცემულია $\sum_{i=1}^{n}\left(\alpha_i-\alpha_i^*\right)=0$ ტოლობით.

მსგავსად, $f-$ის გამოსახულებისათვის :

$$f:\ f\left(\vec{x}\right)=\sum_{i=1}^{n}\left(\alpha_i-\alpha_i^*\right)k\left(\vec{x}^i,\vec{x}\right)+b\,.$$

ამიტომ პრობლემა შეიძლება იყოს დასმული და გადაწყვეტილი მიმართვის გარეშე ხელახალი აღწერის სივრცისადმი, მხოლოდ და მხოლოდ ბირთვების გამოყენებით.

10.8

1. ℓ_1 ნორმა მოდელის გამარტივებას უწყობს ხელს.

2. $\min_{\vec{v}^+>0,\vec{v}^-_j>0,\xi_i>0} \sum_{i=1}^{n} \xi_i + \lambda \sum_{j=0}^{p} \left(\vec{v}_j^+ + \vec{v}_j^- \right)$ შემდეგი სახის შეზღუდვისას :

$$y^i \left(w_0 + \sum_{j=1}^{p} \left(v_j^+ + v_j^- \right) x_j^i \right) \geq 1 - \xi_i \, .$$

3. ყოველგვარი დამატებითი ინფორმაციის გარეშე პრობლემის შესახებ, ფართობი ROC (ინგლ. Receiver Operating Characteristic — მიმღების ოპერაციული მახასიათებელი) მრუდის ქვეშ.

4. საწვრთნელი სიმრავლის და სატესტო სიმრავლის, ასევე კროს-ვალიდაციის (ჯვარედინი შემოწმების) გამოყენება საწვრთნელ სიმრავლეზე λ სიდიდის ასარჩევად წინასწარ განსაზღვრული (მოცემული) მნიშვნელობების მესერზე (ბადეზე).

ალტერნატივის სახით : ჩაშენებული კროს-ვალიდაცია, სხვანაირად, ჩაშენებული(ჩადებული) ჯვარედინი შემოწმება.

5. კი, რეგულარიზაციათა წევრის გამო მოდელი იქნება განსხვავებული შესაბამისად იმისა, ნორმალიზებულია ცვლადები თუ არა.

ლექცია 11 განზომილების რედუქცია (შემცირება)

შინაარსი

ზოგიერთ ამოცანაში მონაცემთა წარმოსადგენად გამოყენებული ცვლადების p რაოდენობა ძალიან დიდია.

ეს ხდება, მაგალითად, მაღალი გარჩევითობის გამოსახულებათა დამუშავებისას, სადაც ყოველი პიქსელი შეიძლება იყოს წარმოდგენილი რამდენიმე ცვლადით, ან გენომური მონაცემების ანალიზისას, როცა გენომის ასობით ათასი პოზიცია შეიძლება იყოს დასახასიათებელი

თუმცა ცვლადების მეტი რაოდენობის შემცველი მონაცემების წარმოდგენა, ინტუიციიდან გამომდინარე, უფრო მდიდარია, ასეთ ვითარებაში უფრო რთულია მაღალი წარმადობის მოდელის სწავლების განხორციელება.

ამ კურსში ჩვენ დაწვრილებით განვიხილავთ ამის მიზეზებს, ხოლო შემდეგ ცვლადების რიცხვის შესამცირებლად უფრო კომპაქტური წარმოდგენებისა და გაცილებით საიმედო მოდელების მისაღებად შევისწავლით რიგ, როგორც კონტროლირებად, ასევე არაკონტროლირებად, მეთოდს.

მიზნები

— მონაცემთა ნაკრების ზომის შემცირებისადმი არსებული ინტერესის ახსნა ;

— ცვლადების სელექციასა (შერჩევასა) და ცვლადების ექსტრაქციას (ამოღებას) შორის განსხვავების დასაბუთება ;

— ამ ორი მიდგომის ძირითადი მეთოდების გაცნობა ;

— მთავარი კომპონენტების ანალიზის გაგება დისპერსიის მაქსიმიზაციის თვალსაზრისით და ასევე როგორც მატრიცული ფაქტორიზაციის მეთოდის.

1 მოტივაცია

განზომილების რედუქციის (შემცირების, შეკვეცის) მიზანია — მონაცემთა $X \in \mathbb{R}^{n \times p}$ წარმოდგენის გარდასახვა $X^* \in \mathbb{R}^{n \times m}$ წარმოდგენად, სადაც $m \square p$. ამისათვის მრავალი მიზეზი არსებობს და ისინი დაწვრილებით არის აღწერილი ამ განყოფილებაში.

1.1 მონაცემთა ვიზუალიზაცია

თუმცა მანქანური სწავლების მიზანს წარმოადგენს მონაცემთა დამუშავების ავტომატიზაცია, ამ კურსში წარმოდგენილი მეთოდების სწორად გამოყენებისათვის აუცილებელია კარგად გვესმოდეს კონკრეტული პრობლემა, რომლის გადაჭრას ჩვენ ვცდილობთ, ისევე, როგორც ჩვენ ხელთ არსებული მონაცემები. ეს აადვილებს გამოსაყენებელი ცვლადების განსაზღვრას, ნებისმიერი ანომალიური გადახრების თავიდან აცილებას და ალგორითმების არჩევას. ამისათვის ძალიან სასარგებლოა გაგვაჩნდეს მონაცემთა გადახედვის (ვიზუალიზაციის) შესაძლებლობა, მაგრამ ეს ადვილი საქმე არ არის ცვლადების უზარმაზარი p რაოდენობის არსებობისას. ცვლადების შეზღუდვა განზომილებათა მცირე რიცხვით აადვილებს მონაცემების ვიზუალიზაციას, ეს გარკვეული ინფორმაციის დაკარგვასაც რომ ნიშნავდეს გარდასახვის პროცესში.

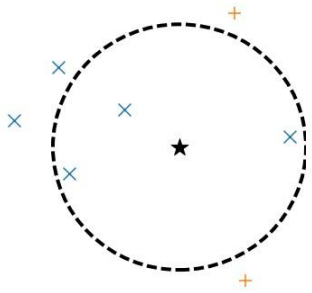
1.2 ალგორითმული დანახარჯების შემცირება

ეს ასპექტი წმინდა გამოთვლითი ხასიათისაა : მონაცემთა განზომილების შეკვეცა ამცირებს როგორც მათ მიერ მეხსიერებაში დაკავებულ ადგილს, ასევე გამოთვლათა დროსაც. გარდა ამისა, თუ ზოგიერთი ცვლადი არის უსარგებლო ან ჭარბი, მაშინ არ არსებობს მათი მიღების აუცილებლობა ახალი დაკვირვებებისათვის. ეს კი ამცირებს მონაცემთა შეგროვების ღირებულებას.

1.3 მოდელის ხარისხის გაუმჯობესება

ჩვენი ძირითადი მოტივაცია მონაცემთა მოცულობის შესამცირებლად მდგომარეობს იმაში, რომ გაგვიჩნდეს შესაძლებლობა კონტროლირებადი სწავლების გაცილებით უფრო ხარისხიანი მოდელების ასაგებად. მაგალითად, ცვლადების ნაკლებ რაოდენობაზე აგებული პარამეტრული მოდელი გაცილებით მცირე ალბათობით აღმოჩნდება ზედმეტად, ანუ გადაჭარბებულად დასწავლილი, როგორც ამაში დავრწმუნდით ლასოს მაგალითზე მეექვსე ლექციაში.

უფრო მეტიც, თუ გასაზომ ცვლადებიდან ზოგიერთს კავშირი არ აქვს სწავლების ჩვენთვის საინტერესო პრობლემასთან, მათ შეუძლია სწავლების ალგორითმი შეცდომამდე მიიყვანოს. მაგალითის სახით ავიღოთ უახლოესი მეზობლების ალგორითმი (იხ. ლექცია 8), რომელიც აფუძნებს თავის წინასწარმეტყველებებს მონიშნული დაკვირვებისადმი უახლოესი საწვრთნელი მაგალითების ჭდეთა გამოყენებით. დავუშვათ, რომ ვიყენებთ ჰერმან მინკოვსკის (გერმ. Hermann Minkowski, 1864-1909) მანძილს (მეტრიკას) ; მაშინ ყველა ცვლადს აქვს ერთნაირი ღირებულება უახლოესი მეზობლების გამოანგარიშებისას. ამიტომ ცვლადებს, რომლებსაც არ აქვს მნიშვნელობა, შეუძლია დაამახინჯოს უახლოესი მეზობლის დადგენა და შემოიტანოს ხმაური მოდელში, როგორც ეს ნაჩვენებია ნახატზე 11.1.



ნახატი 11.1 - ორივე განზომილების გამოყენებისას ვარსკვლავის სამი უახლოესი მეზობელია, უმთავრესად, $(\times)$ წერტილები. თუ გამოიყენება მხოლოდ ერთი ცვლადი აბსცისათა ღერძზე, მაშინ ვარსკვლავის სამი მეზობლის უმრავლესობას $(+)$ წერტილები შეადგენს. ამრიგად, როცა ცვლადს ორდინატთა ღერძზე არავითარი კავშირი აქვს საქმესთან, მაშინ იგი დაამახინჯებს უახლოესი სამი მეზობლის ალგორითმის შედეგს.

დაბოლოს, მაღალი განზომილების პირობებში ჩვენ ვაწყდებით მოვლენას, რომელიც ცნობილია როგორც «განზომილების ჭირი, განზომილების უბედურება» (ინგლ. *curse of dimensionality*).

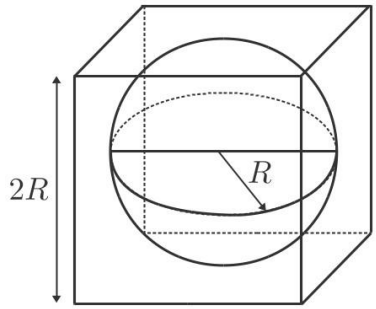
ეს ტერმინი ახასიათებს იმ ფაქტს, რომ დაბალი განზომილების პირობებში განვითარებული ინტუიცია, მიგნების უნარი, მაღალი განზომილების პირობებში ყოველთვის როდი მართლდება.

ამრიგად, მაღალი განზომილების პირობებში მანქანური სწავლების ყველა მაგალითს, როგორც წესი, აქვს ერთმანეთისგან დაშორების ტენდენცია. ამ აზრის გასაგებად მოვათავსოთ ჩვენი საკუთარი თავი p განზომილებაში და განვიხილოთ $\vec{x}$ დაკვირვებაზე ცენტრირებული $R \in \mathbb{R}_+^*$ რადიუსის $S(\vec{x}, R)$ ჰიპერსფერო და ასევე $C(\vec{x}, R)$ ჰიპერკუბი, რომელშიც ჩაწერილია ეს ჰიპერსფერო.

$S(\vec{x}, R)$ ჰიპერსფეროს მოცულობა არის $\frac{2R^p \pi^{p/2}}{p\Gamma(p/2)}$, ხოლო $2R$ სიგრძის გვერდიანი $C(\vec{x}, R)$ ჰიპერკუბის მოცულობა — $2^p R^p$. ამიტომ

$$\lim_{p \to \infty} \frac{\mathrm{Vol}(C(\vec{x}, R))}{\mathrm{Vol}(S(\vec{x}, R))} = 0.$$

ამ შედეგის არსი ასეთია : ალბათობა იმისა, რომ $C(\vec{x}, R)$-ში განთავსებული მაგალითი მიეკუთვნება $S(\vec{x}, R)$-ს, ტოლია $\frac{\pi}{4} \approx 0.79$-ის, როცა $p = 2$ და ტოლია $\frac{\pi}{6} \approx 0.52$-ის, როცა $p = 3$ (იხ. ნახატი. 11.2), დაბოლოს, ხდება ძალზე მცირე, როცა p დიდია — მონაცემები, როგორც წესი, შესამჩნევად დაშორებულია ერთმანეთისგან. ეს ნიშნავს, რომ ალგორითმები, რომლებიც აგებულია *მსგავსების ცნების* გამოყენებით (ვთქვათ, ისეთის, როგორიცაა უახლოესი მეზობლების, გადაწყვეტილებათა ხის ან მხარდამჭერი/საყრდენი ვექტორების მანქანების — Support Vector Machine) აუცილებლად როდი მუშაობს მაღალი განზომილების პირობებში. ამიტომ კარგი მოდელების ასაგებად შეიძლება საჭირო გახდეს განზომილების შემცირება.



ნახატი 11.2 - აქ 3-ის ტოლ განზომილებაში განიხილება $a = 2R$ სიგრძის გვერდიანი კუბის მოცულობის წილი R რადიუსის იმ სფეროს შიგნით, რომელიც ჩაწერილია ამ კუბში.

არსებობს ჩვენი მონაცემების ზომის შემცირების ორი ხერხი :

1. *ცვლადების შერჩევა*, რომელიც მდგომარეობს გარკვეული $p - m$ რაოდენობის *ცვლადის გამორიცხვაში* მონაცემებიდან ;
2. *ცვლადების ამოღება*, რომელიც მდგომარეობს m *ახალი ცვლადის შექმნაში* იმ p ცვლადიდან, რომლებიც თავიდანვე გვაქვს.

კურსის დარჩენილი ნაწილი დაწვრილებით აღწერს ამ ორ მიდგომას.

2 ცვლადების შერჩევა

მიდგომები *ცვლადების შერჩევისადმი* გულისხმობს m ცვლადის შერჩევას შესანახად არსებული p ცვლადიდან და დანარჩენი $p - m$ ცვლადის იგნორირებას. ამ მიდგომების დაყოფა შეიძლება სამ კატეგორიად : *ფილტრაციის* მეთოდები, *კონტეინერული (კონტეინერიზაციის)* მეთოდები და *ჩაშენებული* მეთოდები. მეთოდები, რომლებსაც ჩვენ აქ განვიხილავთ, არის კონტროლირებადი : ვგულისხმობთ, რომ გვაქვს მონაცემების $\mathcal{D} = \left\{\left(\vec{x}^i, y^i\right)\right\}_{i=1,\ldots,n}$ ნაკრები ჭდეებით, სადაც $\vec{x}^i \in \mathbb{R}^p$.

2.1 ფილტრაციის მეთოდები

ცვლადების შერჩევა *ფილტრაციის* გზით გულისხმობს *შერჩევის კრიტერიუმის* დამოუკიდებლად გამოყენებას თითოეულისადმი p ცვლადიდან ; მიზანია $p-$ური ცვლადის *მნიშვნელოვნობის დონის* (ინგლ. *significance level*) რაოდენობრივი შეფასება მონაცემების ნაკრებში y ჭდის მიმართ.

ასეთი კრიტერიუმების რამდენიმე მაგალითია : კორელაცია ჭდით, ვთქვათ, $\mathcal{X}^2$-ის სტატისტიკური ტესტი კლასიფიკაციის პრობლემის შემთხვევაში, ან *ურთიერთინფორმირებულობა* (ინგლ. *mutual information*) — ორი შემთხვევითი სიდიდის სტატისტიკური ფუნქცია, რომელიც აღწერს ერთ შემთხვევით სიდიდეში ინფორმაციის რაოდენობის შემცველობას მეორე შემთხვევითი სიდიდის შესახებ.

კორელაცია j ცვლადსა და y ჭდეს შორის გამოითვლება კორელაციის მსგავსად ნაწინასწარმეტყველებ ჭდესა და ფაქტობრივ ჭდეს შორის (იხ. განტოლება 3.2) :

$$R_j = \frac{\sum_{i=1}^{n}\left(y^i - \frac{1}{n}\sum_{i=1}^{n} y^i\right)\sum_{i=1}^{n}\left(x_j^i - \frac{1}{n}\sum_{i=1}^{n}\vec{x}^i\right)}{\sqrt{\sum_{i=1}^{n}\left(y^i - \frac{1}{n}\sum_{i=1}^{n} y^i\right)^2}\sqrt{\sum_{i=1}^{n}\left(x_j^i - \frac{1}{n}\sum_{l=1}^{n} x_j^l\right)^2}}. \qquad (11.1)$$

განსაზღვრება 11.1 (ურთიერთინფორმირებულობა) *ურთიერთინფორმირებულობა* ორ — X_j და Y — შემთხვევით სიდიდეს შორის ზომავს მათ დამოკიდებულებას ალბათური თვალსაზრისით ; იგი უდრის ნულს მაშინ და მხოლოდ მაშინ, როცა სიდიდეები დამოუკიდებელია და იზრდება მათი დამოკიდებულების ხარისხის ზრდისას. დისკრეტულ შემთხვევაში იგი განისაზღვრება თანაფარდობით

$$I\left(X_j, Y\right) = \sum_{x_j, y} \mathbb{P}\left(X_j = x_j, Y = y\right) \log \frac{\mathbb{P}\left(X_j = x_j, Y = y\right)}{\mathbb{P}\left(X_j = x_j\right)\mathbb{P}\left(Y = y\right)},$$

ხოლო უწყვეტ შემთხვევაში — ფორმულით

$$I\left(X_j, Y\right) = \int_{x_j}\int_{y} p\left(x_j, y\right) \log \frac{p\left(x_j, y\right)}{p\left(x_j\right)p\left(y\right)} dx_j dy.$$

კოზაჩენკო-ლეონენკოს შემფასებელი არის ერთ-ერთი მათ შორის, რომლებიც ყველაზე უფრო ხშირად გამოყენება ურთიერთინფორმირებულობის შესაფასებლად (Л. Ф. Козаченко და Н. Н. Леоненко, 1987).

ფილტრაციის მეთოდები გულისხმობს ინდივიდუალური მიდგომის გამოყენებას ცვლადებისადმი : ამ მეთოდებს არ შეუძლია მათი ერთობლივი მოქმედების გათვალისწინება. ხსენებული პრობლემის მაილუსტრირებელი კლასიკური მაგალითია ე.წ. *გამომრიცხავი „ან“* $\left(\mathrm{XOR}\right)$ ლოგიკური ფუნქცია (eXclusive-OR) : ცალკე აღებული, x_1 ცვლადი (შესაბამისად, x_2 ცვლადი) კორელაციაში არ არის $y = x_1 \mathrm{XOR}\, x_2$ ცვლადთან, მაშინ როცა ერთად ეს ორი ცვლადი იდეალურ ახსნას აძლევს y ჭდეს.

2.2 კონტეინერული მეთოდები (კონტეინერიზაციის მეთოდები)

კონტეინერიზაციის, ანუ კონტეინერული მეთოდები (ინგლ. ***wrapper methods,*** შეფუთვის მეთოდები) გულისხმობს მცდელობას განისაზღვროს ცვლადების საუკეთესო ქვესიმრავლე *სწავლების მოცემული ალგორითმისათვის.* ხშირად ამას უწოდებენ *ქვესიმრავლის არჩევას* (ინგლ. ***subset selection***), და არა ცვლადების არჩევას.

ხსენებული მეთოდები, ხშირად, ევრისტიკულია, ე.ი. დაფუძნებულია ინტუიციაზე, საზრიანობაზე, ანალოგიებზე, გამოცდილებაზე, გამომგონებლობაზე, რაც უნდა ეყრდნობოდეს ადამიანის ტვინის განსაკუთრებულ თვისებებსა და ადამიანის უნარს გადაწყვიტოს ამოცანები, რომლებისთვისაც ფორმალური მათემატიკური ალგორითმი, გადაწყვეტის წესი უცნობია.

ეს დაკავშირებულია იმასთან, რომ, ჩვეულებრივ, შეუძლებელია ალგორითმის მწარმოებლურობის განსაზღვრა მონაცემთა ნაკრებზე მისი სწავლების გარეშე. ამომწურავი სტრატეგიის გამოყენება მიზანშეუწონელია, გარდა იმ შემთხვევისა, როცა ცვლადების რიცხვი მცირეა, ვინაიდან p ცვლადის ქვესიმრავლეთა რიცხვი არის $2^p - 1$ ($\emptyset$-ის გამოკლებით). ამიტომ გამოიყენებენ ხარბ (გაუმაძღარ) მიდგომას (ინგლ. *greedy approach*).

ასეთი მიდგომების მაგალითებია *ძებნა ქვევიდან ზევით (აღმავალი ძებნა)*, *ძებნა ზევიდან ქვევით (დაღმავალი ძებნა)* და *მცურავი ძებნა*, რომლებიც აღწერილი იქნება ქვემოთ.

დავუშვათ, რომ მოცემულია მონაცემთა $\mathcal{D} = (X, \vec{y})$ ნაკრები, სადაც $X \in \mathbb{R}^{n \times p}$, ცვლადების $\mathcal{E} \subset \{1, 2, \ldots, p\}$ ქვესიმრავლე და სწავლების ალგორითმი. ასევე დავუშვათ, რომ $X_\mathcal{E} \in \mathbb{R}^{n \times |\mathcal{E}|}$ არის $\mathcal{E}$-ში წარმოდგენილი ცვლადებით შეზღუდული X მატრიცა. მივიჩნიოთ, რომ $E_\mathcal{D}(\mathcal{E})$ არის $(X_\mathcal{E}, \vec{y})$ სიმრავლეზე გაწვრთნილი სწავლების ამ ალგორითმის განზოგადების შეცდომის შეფასება. ჩვეულებრივ, ეს შეფასება მიიღება სატესტო ნაკრებიდან ან ჯვარედინი ვალიდაციის (შემოწმების) გზით.

კვლევა «ქვემოდან ზევით» გულისხმობს დაწყებას ცვლადების ცარიელი ნაკრებიდან და ამ ნაკრებში ყოველი იმ ცვლადის მიყოლებით შეტანას (დამატებას), რომელიც ყველაზე უკეთ აუმჯობესებს ალგორითმის ეფექტურობას (მწარმოებლურობას, სამომხმარებლო მახასიათებელს), ვიდრე ამ მაჩვენებლის გაუმჯობესება შეუძლებელი გახდება.

განსაზღვრება 11.2 (ძებნა ქვევიდან ზევით, ანუ აღმავალი ძებნა) ძებნა ქვევიდან ზევით (ინგლ. *forward search — პირდაპირი ძებნა, ძებნა წინსვლით, აღმავალი ძებნა*) ეწოდება ცვლადების შერჩევის შემდეგ ხარბ (გაუმაძღარ) პროცედურას :

1. ინიციალიზაციის განხორციელება $\mathcal{F} = \emptyset$.
2. $\mathcal{F}$ -ში დასამატებლად (შესატანად) საუკეთესო ცვლადის მოძიება :

   $$j^* = \arg \min_{j \in \{1, \ldots, p\} \setminus \mathcal{F}} E_\mathcal{D}(\mathcal{F} \cup \{j\}).$$

3. თუ $E_\mathcal{D}(\mathcal{F} \cup \{j\}) > E_\mathcal{D}(\mathcal{F})$: გაჩერება,

   წინააღმდეგ შემთხვევაში : $\mathcal{F} \leftarrow \mathcal{F} \cup \{j\}$; ხელახლა დაწყება პუნქტებით $2-3$.

უარეს შემთხვევაში (მაშინ, როცა უნდა განხორციელდეს იტერირება $\mathcal{F} = \{1, 2, \ldots, p\}$-მდე), ეს ალგორითმი მოითხოვს მონაცემთა ნაკრებზე სწავლების ალგორითმის შეფასებათა სირთულის $\mathcal{O}(p^2)$ რიგს, რაც შეიძლება იყოს საკმაოდ შრომატევადი და ძვირადღირებული, მაგრამ გაცილებით უფრო ეფექტური, ვიდრე $\mathcal{O}(2^p)$ მნიშვნელობა, რომელსაც ამომწურავი მიდგომა მოითხოვს.

და, პირიქით, დაღმავალი ძებნა მდგომარეობს იმაში, რომ ალგორითმის დაწყება ხდება ყველა

ცვლადის ნაკრებიდან და გრძელდება მიყოლებით თითოეული იმ ცვლადის ამოღებით, რომელიც ყველაზე უკეთ აუმჯობესებს ალგორითმის ხარისხის მაჩვენებელს, ვიდრე ეს გაუმჯობესება არ შეწყდება.

განსაზღვრება 11.3 (დაღმავალი ძებნა) *დაღმავალი ძებნა* (ან ინგლისურად : *backward search* — შექცეული ძებნა, ძებნა უკუსვლით) ეწოდება ცვლადების არჩევის (შერჩევის) შემდეგ ხარბ (გაუმაძღარ) პროცედურას :

1. ინიციალიზაციის განხორციელება $\mathcal{F} = \{1, 2, \ldots, p\}$.
2. $\mathcal{F}$ -დან ამოსაღებად საუკეთესო ცვლადის მოძიება : $j^* = \arg\min_{j \in \mathcal{F}} E_{\mathcal{D}}\left(\mathcal{F} \setminus \{j\}\right)$.
3. თუ $E_{\mathcal{D}}\left(\mathcal{F} \setminus \{j\}\right) > E_{\mathcal{D}}\left(\mathcal{F}\right)$: გაჩერება,

   წინააღმდეგ შემთხვევაში : $\mathcal{F} \leftarrow \mathcal{F} \setminus \{j\}$; ხელახლა დაწყება პუნქტებით $2-3$.

დაღმავალი მიდგომის უპირატესობა აღმავალთან შედარებით მდგომარეობს იმაში, რომ იგი აუცილებლად იძლევა ცვლადების ქვესიმრავლეს, რომელიც ყველა ცვლადის ერთობლიობაზე უკეთესია. მართლაც, თუ ჩვენ ვერ ვპოულობთ ცვლადს $\mathcal{F}$ -ში დასამატებლად აღმავალი მეთოდის ერთ-ერთ ეტაპზე, ეს ჯერ კიდევ არ ნიშნავს იმას, რომ ალგორითმის ეფექტურობა $\left(X_{\mathcal{F}}, \vec{y}\right)$-ზე უკეთესია, ვიდრე $\left(X, \vec{y}\right)$-ზე.

მცურავი ძებნა შესაძლებლობათა სივრცის გამოკვლევის საშუალებას ამ ორი მიდგომის კომბინაციის გზით რამდენადმე სხვანაირად იძლევა.

განსაზღვრება 11.4 (მცურავი ძებნა) თუ მოცემულია ორი მკაცრად დადებითი მთელრიცხვა q და r პარამეტრი, მაშინ *მცურავი ძებნა* (ინგლ. *floating search*) ეწოდება ცვლადების შერჩევის შემდეგ ხარბ (გაუმაძღარ) პროცედურას :

1. ინიციალიზაციის განხორციელება $\mathcal{F} = \emptyset$.
2. $\mathcal{F}$ -ში დასამატებლად (შესატანად) q რაოდენობის საუკეთესო ცვლადთა მოძიება :

   $S^* = \arg\min_{S \subseteq \{1,\ldots,p\} \setminus \mathcal{F}, |S| = q} E_{\mathcal{D}}\left(\mathcal{F} \cup S\right)$.
3. თუ $E_{\mathcal{D}}\left(\mathcal{F} \cup S\right) < E_{\mathcal{D}}\left(\mathcal{F}\right)$: $\mathcal{F} \leftarrow \mathcal{F} \cup S$.
4. $\mathcal{F}$ -დან ამოსაღებად r რაოდენობის საუკეთესო ცვლადთა მოძიება :

   $S^* = \arg\min_{S \subseteq \{1,\ldots,p\} \setminus \mathcal{F}, |S| = r} E_{\mathcal{D}}\left(\mathcal{F} \cup S\right)$.
5. თუ $E_{\mathcal{D}}\left(\mathcal{F} \setminus S\right) > E_{\mathcal{D}}\left(\mathcal{F}\right)$: გაჩერება,

   წინააღმდეგ შემთხვევაში : $\mathcal{F} \leftarrow \mathcal{F} \setminus S$; ხელახლა დაწყება პუნქტებით $2-5$.

2.3 ჩაშენებული მეთოდები (მიდგომები)

დაბოლოს, *ჩაშენებული მეთოდები, მიდგომები* (ინგლ. *embedded approaches*), რომლებიც მოდელთან ერთად იმავდროულად იმასაც სწავლობს, რომელი ცვლადების ჩართვა უნდა მოხდეს.

როგორც წესი, ეს მარტივი, ეკონომიური (მომჭირნე, ყაირათიანი) პარამეტრული მოდელებია, ე.ი. ისეთი, რომლებშიც ზოგიერთი ცვლადისათვის მინიჭებული კოეფიციენტები ნულს უდრის.

შემდეგ ეს ცვლადები მოდელიდან აღმოიფხვრება, ასეთი მოდელების ყველაზე თვალსაჩინო მაგალითია LASSO (Least Absolute Shrinkage and Selection Operator — უმცირესი აბსოლუტური შეკვეცის და შერჩევის ოპერატორი), რომელიც ადრე გვაქვს განხილული (იხ. მეექვსე ლექცია).

ეს ნიშანთა სივრცის განზომილების შემცირების მეთოდია, რომელიც შემოთავაზებული იყო კანადელი სტატისტიკოსის ტორონტოს უნივერსიტეტიდან რობერტ ტიბშირანის (Robert Tibshirani, 1956-) მიერ 1994 წლის იანვარში[3].

ასეთივე გზით ცვლადთა მნიშვნელოვნობის ზომა შემთხვევით ტყეებში (იხ. მეცხრე ლექცია) შეიძლება იყოს გამოყენებული გადაწყვეტილების მისაღებად აღმოსაფხვრელი ცვლადების დასადგენად.

3 ცვლადების ამოღება

3.1 მთავარი კომპონენტების ანალიზი

ყველაზე კლასიკური მეთოდი მონაცემთა მასივის (ნაკრების) განზომილების შესამცირებლად ცვლადების ამოღების გზით — ეს *მთავარ კომპონენტთა ანალიზია* (ინგლ. Principal Component Analysis, PCA). იგი მონაცემთა განზომილების შემცირების ერთ-ერთი ძირითადი ხერხია ინფორმაციის რაც შეიძლება ნაკლები რაოდენობის დანაკარგის უზრუნველყოფით. გამოგონებულია 1901 წელს კარლ პირსონის (Karl Pearson, 1857—1936) — ინგლისელი მათემატიკოსის, სტატისტიკოსის, ბიოლოგისა და ფილოსოფოსის, მათემატიკური სტატისტიკის დამაარსებლისა და ბიომეტრიკის ერთ-ერთი ფუძემდებლის — მიერ.

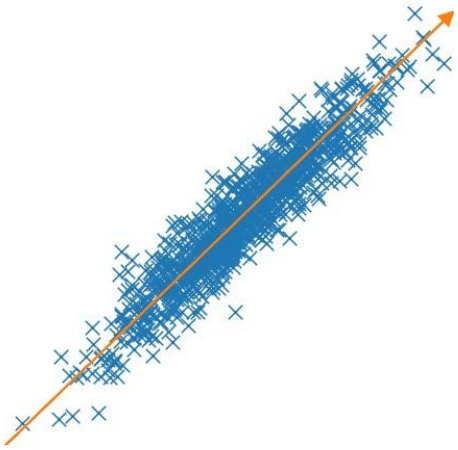
მთავარ კომპონენტთა ანალიზი გამოიყენება მრავალ სფეროში, მათ შორის ეკონომომეტრიკაში, ბიოინფორმატიკაში, გამოსახულებათა დასამუშავებლად, ასევე მონაცემების შეკუმშვის მიზნით და საზოგადოებრივ მეცნიერებებში.

დისპერსიის მაქსიმიზაცია

მთავარ კომპონენტთა ანალიზის ძირითადი იდეა მდგომარეობს მონაცემთა ისეთ წარმოდგენაში, რომ მოხდეს მათი დისპერსიის მაქსიმიზაცია ახალი განზომილებებით და შეიძლებოდეს

[3] იხ. მისი სტატია : Robert Tibshirani, «Regression Shrinkage and Selection via the Lasso», Journal of the Royal Statistical Society, Series B (Statistical Methodology), Vol. 58, No. 1 (1996), pp. 267–288 (22 pages), Published By Oxford University Press, https://www.jstor.org/stable/2346178)

მაგალითების ერთმანეთისგან გარჩევის შენარჩუნება ახალ წარმოდგენაში (იხ. ნახატი. 11.3).



ნახატი 11.3 – მონაცემების დისპერსია მაქსიმალურია ისრით ნაჩვენები ღერძის გასწვრივ.

ფორმალურად, ახალი $\mathcal{X}$ წარმოდგენა განისაზღვრება ორთონორმირებული ბაზისით, რომელზეც ხდება მონაცემთა X მატრიცის პროეცირება.

განსაზღვრება 11.5 (მთავარ კომპონენტთა ანალიზი) $X \in \mathbb{R}^{n \times p}$ მატრიცის *მთავარ კომპონენტთა ანალიზი* — ეს ორთოგონალური წრფივი გარდასახვაა, რომელიც X-ის გამოსახვის საშუალებას იძლევა ახალ ორთონორმირებულ ბაზისში ისე, რომ X-ის უდიდესი დისპერსია პროექციის მიხედვით თანაბრდება ამ ახალი ბაზისის პირველ ღერძთან, მეორე უდიდესი დისპერსია — მეორე ღერძთან, და ასე შემდეგ.

მონაცემთა ამ ახალი ბაზის ღერძებს *მთავარ კომპონენტებს* უწოდებენ, შემოკლებით PC — ინგლისური ფრაზიდან *Principal Components*.

ყურადღება

ლექციის ამ ქვედანაყოფის დარჩენილ ნაწილში ნაგულისხმევია, რომ ცვლადები სტანდარტიზებულია ისეთნაირად, რომ ყველას გააჩნდეს 0-ის ტოლი საშუალო მნიშვნელობა და 1-ის ტოლი დისპერსია. ამით ისეთი ვითარების აცილება მოხდება, როცა ცვლადები ამ პარამეტრთა მეტი მნიშვნელობებით უფრო მნიშვნელოვანი აღმოჩნდება, ვიდრე ცვლადები მათი მცირე მნიშვნელობებით. ეს არის მთავარ კომპონენტთა ანალიზის (PCA-ის) გამოყენებისათვის წინასწარი აუცილებელი პირობა. სტანდარტიზაცია მიიღწევა თითოეული ცვლადის საშუალო მნიშვნელობის ცენტრირებისა და დისპერსიის შემცირების გზით :

$$x_j^i \leftarrow \frac{x_j^i - \overline{x}_j}{\sqrt{\frac{1}{n}\sum_{l=1}^{n}\left(x_j^l - \overline{x}_j\right)^2}}, \tag{11.2}$$

სადაც $\overline{x}_j = \frac{1}{n}\sum_{l=1}^{n} x_j^l$. სწორედ ამ შემთხვევაში იტყვიან, რომ X მატრიცა *ცენტრირებულია* : ყოველ მის სვეტს აქვს 0 – ის ტოლი საშუალო მნიშვნელობა.

თეორემა 11.1 *დავუშვათ, რომ* $X \in \mathbb{R}^{n\times p}$ *— ცენტრირებული მატრიცაა* $\Sigma = \frac{1}{n} X^{\mathrm{T}} X$ *ემპირიული კოვარიაციით.* X *მატრიცის მთავარი კომპონენტები — ეს* $X-$*ის საკუთარი ვექტორებია, რომლებიც მოწესრიგებულია საკუთარი მნიშვნელობის კლებადობით.*

მტკიცებულება. დავიწყოთ იმის მტკიცებულებით, რომ ნებისმიერი $\vec{w} \in \mathbb{R}^p$ ვექტორის შემთხვევაში X ვექტორის $\vec{w}$ ვექტორზე პროექციის დისპერსია უდრის $w^{\mathrm{T}}\Sigma w$ სიდიდეს.

$X \in \mathbb{R}^{n\times p}-$ის პროექცია $\vec{w} \in \mathbb{R}^p-$ზე — ეს $\vec{z} = X\vec{w}$ ვექტორია. ვინაიდან X ცენტრირებულია, $\vec{z}-$ის საშუალო მნიშვნელობა შემდეგი მნიშვნელობისაა :

$$\frac{1}{n}\sum_{i=1}^{n} z_i = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p} x_j^i w_j = \frac{1}{n}\sum_{j=1}^{p} w_j \sum_{i=1}^{n} x_j^i = 0\,.$$

მისი დისპერსია კი შეადგენს :

$$\mathrm{Var}\left[\vec{z}\right] = \frac{1}{n}\sum_{i=1}^{n} zz_i^2 = \frac{1}{n}\vec{w}^{\mathrm{T}} X^{\mathrm{T}} X\vec{w} = \vec{w}^{\mathrm{T}}\Sigma\vec{w}\,.$$

ახლა ვუწოდოთ $\vec{w}_1 \in \mathbb{R}^p$ ვექტორს პირველი მთავარი კომპონენტი. $\vec{w}_1$ ორთონორმირებულია და იმავდროულად $X\vec{w}_1-$ის დისპერსია მაქსიმალურია :

$$\left.\begin{array}{l} \vec{w}_1 = \arg\max\limits_{\vec{w}\in\mathbb{R}^p} \vec{w}^{\mathrm{T}}\Sigma\vec{w} \\ \text{და ამასთან ერთად } \left\|\vec{w}_1\right\|_2 = 1 \end{array}\right\}. \qquad (11.3)$$

ეს კვადრატული ოპტიმიზაციის ამოცანაა ტოლობათა სახის შეზღუდვებით (იხ. ქვედანაყოფი 13.4), რომელიც შეიძლება ამოვხსნათ $\alpha_1 > 0$ ლაგრანჟის მამრავლის შემოტანით და შემდეგი სახის ლაგრანჟიანის დაწერით :

$$L\left(\alpha_1, \vec{w}\right) = \vec{w}^{\mathrm{T}}\Sigma\vec{w} - \alpha_1\left(\left\|\vec{w}\right\|_2 - 1\right).$$

ძლიერი დუალობის გამო, $\vec{w}^{\mathrm{T}}\Sigma\vec{w}-$ის მაქსიმუმი $\left\|\vec{w}_1\right\|_2 = 1$ შეზღუდვისას $\min\limits_{\alpha_1}\sup\limits_{\vec{w}\in\mathbb{R}^p} L\left(\alpha_1, \vec{w}\right)$ სიდიდეს უდრის. ლაგრანჟიანის სუპრემუმი — რომლის ცნება ცნობილია სიმრავლეთა თეორიიდან — მიიღება წერტილზე, სადაც მისი გრადიენტის განულება ხდება, რაც იმას ნიშნავს, რომ :

$$2\Sigma\vec{w} - 2\alpha_1\vec{w} = 0\,.$$

ამრიგად, $\Sigma\vec{w}_1 = \alpha_1\vec{w}_1$ და $\left(\alpha_1, \vec{w}_1\right)$ ქმნის $\Sigma-$ის წყვილს (საკუთარი მნიშვნელობა, საკუთარი ვექტორი).

Σ მატრიცის ყველა საკუთარ ვექტორს შორის, $\vec{w}_1$ ვექტორი წარმოადგენს იმ ვექტორს, რომელიც $\vec{w}_1{}^{\mathrm{T}}\Sigma\vec{w}_1 = \alpha_1\left\|\vec{w}_1\right\|_2 = \alpha_1$ დისპერსიას მაქსიმუმს ანიჭებს.

ამრიგად, α_1 არის Σ მატრიცის უდიდესი საკუთარი მნიშვნელობა (საკმარისია გავიხსენოთ, რომ $X^{\mathrm{T}}X$ გამოსახულებით განსაზღვრული Σ მატრიცა არის დადებითი ნახევრადგანსაზღვრული მატრიცა და ყველა მისი საკუთარი მნიშვნელობა დადებითია).

X – ის მეორე მთავარი კომპონენტი აკმაყოფილებს შემდეგ პირობებს :

$$\left.\begin{array}{l} \vec{w}_2 = \arg\max\limits_{\vec{w}\in\mathbb{R}^p} \vec{w}^{\mathrm{T}}\Sigma\vec{w} \\ \text{ისეთი რომ } \left\|\vec{w}_2\right\|_2 = 1 \text{ და } \vec{w}^{\mathrm{T}}\vec{w}_1 = 0 \end{array}\right\}. \tag{11.4}$$

ეს უკანასკნელი შეზღუდვა იმის გარანტირების საშუალებას იძლევა, რომ მთავარ კომპონენტთა ბაზისი ორთონორმირებულია.

ახლა შემოვიტანოთ ლაგრანჟის ორი — $\alpha_2 > 0$ და $\beta_2 > 0$ — მამრავლი და მივიღოთ შემდეგი ლაგრანჟიანი :

$$L\left(\alpha_2, \beta_w, \vec{w}\right) = \vec{w}^{\mathrm{T}}\Sigma\vec{w} - \alpha_2\left(\left\|\vec{w}\right\|_2^2 - 1\right) - \beta_2\vec{w}^{\mathrm{T}}\vec{w}_1 .$$

წინა შემთხვევის მსგავსად, ამ ლაგრანჟიანის სუპრემუმი $\vec{w}$-ზე მიიღწევა იქ, სადაც მისი გრადიენტის განულება ხდება :

$$2\Sigma\vec{w}_2 - 2\alpha_2\vec{w}_2 - \beta_2\vec{w}_1 = 0 .$$

მარცხენა ნაწილის გამრავლების შედეგად $\vec{w}_1^{\mathrm{T}}$-ზე, მიიღება :

$$2\vec{w}_1^{\mathrm{T}}\Sigma\vec{w}_2 - 2\alpha_2\vec{w}_1^{\mathrm{T}}\vec{w}_2 - \beta_2\vec{w}_1^{\mathrm{T}}\vec{w}_1 = 0 ,$$

საიდანაც ვასკვნით, რომ $\beta_2 = 0$ და, ამ მნიშვნელობის შეტანით წინა განტოლებაში, გვექნება $2\Sigma\vec{w}_2 - 2\alpha_1\vec{w}_2 = 0$ თანაფარდობა, რომლის მსგავსი უკვე მიღებული იყო $\vec{w}_1$-ის შემთხვევაში. ამრიგად, $\left(\alpha_2, \vec{w}_2\right)$ ქმნის Σ მატრიცის წყვილს (საკუთარი მნიშვნელობა, საკუთარი ვექტორი) და α_2 მაქსიმალურია : ამიტომ იგი აუცილებლად უნდა იყოს Σ – ის მეორე საკუთარი მნიშვნელობა.

მსგავსი მსჯელობები გრძელდება შემდეგი მთავარი კომპონენტებისათვის.

□

შენიშვნა

ამრიგად, 11.1 თეორემის დამტკიცება სხვანაირადაც შეიძლება, თუ ვისარგებლებთ იმ ფაქტით, რომ Σ მატრიცა თავისი აგებულებით დადებითად განსაზღვრულია და შესაძლებელია მისი დიაგონალიზაცია ორთონორმირებული ბაზისის ცვლილებით : $\Sigma = Q^{\mathrm{T}}\Lambda Q$, სადაც $\Lambda \in \mathbb{R}^{p\times p}$ — დიაგონალური მატრიცაა, რომლის დიაგონალური მნიშვნელობები Σ მატრიცის საკუთარი მნიშვნელობებია. ამრიგად,

$$\vec{w}_1^{\mathrm{T}}\Sigma\vec{w}_1 = \vec{w}_1^{\mathrm{T}}Q^{\mathrm{T}}\Lambda Q\vec{w}_1 = \left(Q\vec{w}_1\right)^{\mathrm{T}}\Lambda\left(Q\vec{w}_1\right).$$

თუ $\vec{w}_1^{\mathrm{T}}\Sigma\vec{w}_1$ სიდიდის მაქსიმიზაციისათვის შემოვიტანთ $\vec{v}=Q\vec{w}_1$ აღნიშვნას, სადაც Q თანამამრავლი ორთონორმირებულია, ხოლო $\vec{w}_1$ ვექტორის ნორმა 1-ს უდრის, მაშინ დასადგენი ხდება 1-ის ტოლი ნორმის მქონე $\vec{v}$ ვექტორი, რომელიც მაქსიმუმს ანიჭებს $\sum_{j=1}^{p} v_j^2\lambda_j$ ჯამს. ვინაიდან Σ მატრიცა განსაზღვრულია დადებითად, $\lambda_j \geq 0 \ \forall j=1,\ldots,p.$ უფრო მეტიც, $\left\|\vec{v}\right\|_2 = 1$ ტოლობიდან გამომდინარეობს, რომ $0 \leq v_j^2 \leq 1 \ \forall j=1,\ldots,p$.

ასე რომ, $\sum_{j=1}^{p} v_j^2\lambda_j \leq \max\limits_{j=1,\ldots,p}\lambda_j \sum_{j=1}^{p} v_j^2 \leq \max\limits_{j=1,\ldots,p}\lambda_j$ და ეს მაქსიმუმი მაშინ მიიღწევა, როცა $v_j=1$ და $v_k=0 \ \forall k\neq j$. მაშასადამე, დგინდება, რომ $\vec{w}_1$ — ეს საკუთარი ვექტორია, რომელიც შეესაბამება Σ მატრიცის იდიდეს საკუთარ მნიშვნელობას, და ასე შემდეგ.

სინგულარულ მნიშვნელობებად დეკომპოზიცია (დაშლა)

თეორემა 11.2 *დავუშვათ, რომ $X\in\mathbb{R}^{n\times p}$ — ცენტრირებული მატრიცაა. X-ის მთავარ კომპონენტებს წარმოადგენს მისი მარჯვენა სინგულარული ვექტორები, რომლებიც, მოწესრიგებულია სინგულარული მნიშვნელობების კლებადობით.*

მტკიცებულება თუ X მატრიცას ჩავწერთ UDV^{T} ფორმით, სადაც $U\in\mathbb{R}^{n\times n}$ და $V\in\mathbb{R}^{p\times p}$ მატრიცები ორთოგონალურია, ხოლო $D\in\mathbb{R}^{n\times p}$ წარმოადგენს დიაგონალს, მაშინ

$$\Sigma = X^{\mathrm{T}}X = VDU^{\mathrm{T}}UDV^{\mathrm{T}} = VD^2V^{\mathrm{T}}$$

და X-ის სინგულარული მნიშვნელობები (D დიაგონალის ელემენტები) არის კვადრატული ფესვები Σ მატრიცის საკუთარი მნიშვნელობებიდან, მაშინ როცა X-ის მარჯვენა სინგულარული ვექტორები (V მატრიცის სვეტები) წარმოადგენს Σ მატრიცის საკუთარ ვექტორებს.

პრაქტიკულად, სინგულარულ მნიშვნელობებად დაშლის (ინგლ. SVD, Singular Value Decomposition) რეალიზაციები რიცხობრივად უფრო მდგრადია, ვიდრე სპექტრული დაშლით მიღებული რეალიზაციები. ამიტომ მეტი უპირატესობა ეძლევა $X-$ის მთავარი კომპონენტების გამოთვლას $X-$ის სინგულარულ მნიშვნელობებად დაშლის გამოყენებით, ვიდრე $X^{\mathrm{T}}X$ მატრიცის სპექტრული დაშლის გზით.

მთავარ კომპონენტთა რაოდენობის არჩევა

მონაცემთა განზომილების შემცირება მთავარ კომპონენტთა ანალიზის დახმარებით გულისხმობს შესანახ მთავარ კომპონენტთა რიცხვის *არჩევას.*

ამისათვის ჩვენ ვიყენებთ ამ კომპონენტების მიერ *ასახსნელი დისპერსიის* (ინგლ. Explained Variance) წილს : X მატრიცის დისპერსია გამოისახება როგორც Σ *მატრიცის კვალი* (ინგლ. matrix trace), რომელიც თავად არის საკუთარი მნიშვნელობების ჯამი — დიაგონალური ელემენტების ჯამი

ამრიგად, თუ მივიღებთ გადაწყვეტილებას შევინახოთ X მატრიცის პირველი m მთავარი კომპონენტი, მაშინ იმ დისპერსიის წილი, რომლის ახსნას ისინი ახერხებს, შემდეგი გამოსახულების ტოლია :

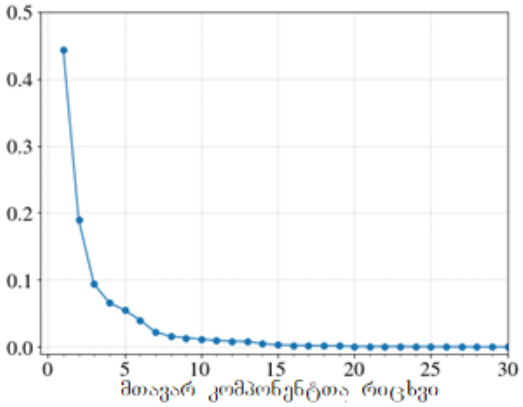
$$\frac{\alpha_1 + \alpha_2 + \cdots + \alpha_m}{\mathrm{Tr}(\Sigma)},$$

სადაც $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_p$ — Σ მატრიცის საკუთარი მნიშვნელობებია, დალაგებული კლების რიგით, ხოლო $\mathrm{Tr}(\Sigma)$ – ამავე მატრიცის ე.წ. *კვალია*. მატრიცის კვალი — ეს ოპერაციაა, რომელიც ასახავს კვადრატულ მატრიცათა სივრცეს ველზე. სადაც განსაზღვრულია მატრიცა (ნამდვილი მატრიცებისათვის — ნამდვილ რიცხვთა ველზე, კომპლექსური მატრიცებისათვის — კომპლექსურ რიცხვთა ველზე). მატრიცის კვალი — ეს მატრიცის მთავარი დიაგონალის ელემენტების ჯამია, მაგალითად, თუ a_{ij} რომელიღაც A მატრიცის ელემენტებია, მაშინ მისი კვალი $\mathrm{Tr}A = \sum_i a_{ii}$. მატრიცას ნულოვანი კვალით *უკვალო* ეწოდება (ინგლ. *traceless* ან *tracefree*). მათემატიკურ ტექსტებში კვალისათვის გვხვდება ორი აღნიშვნა : $\mathrm{Tr}\,A$ (ინგლისური სიტყვიდან *trace* — *კვალი*) და $\mathrm{Sp}A$ (გერმანული სიტყვიდან «*შპურ*» ჟღერადობით: Spur — კვალი).

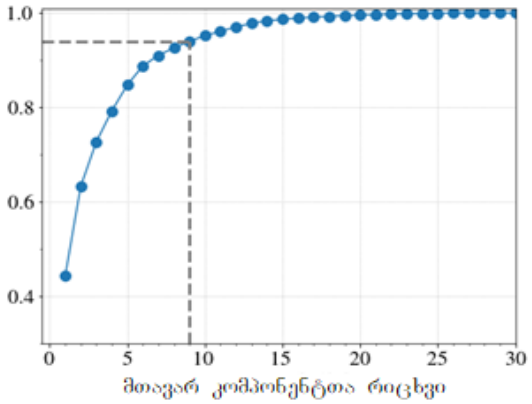
მთავარ კომპონენტთა რიცხვის ზრდასთან დაკავშირებით ტრადიციულია ორი პარამეტრის ევოლუციის განხილვა : ან დისპერსიის იმ წილის, რომელიც აიხსნება თითოეულის მიერ ამ კომპონენტთა შორის, ან მათი ერთობლივი, კუმულაციური ხასიათის ასეთი წილის, რომელიც შეიძლება ვიზუალურად იყოს წარმოდგენილი *საკუთარი მნიშვნელობების გრაფიკზე* (ინგლ. a scree plot, რაც მართლაც გვაგონებს *ქვიან ჩამონაშალს, შვავს* ბუნებაში). ხსენებული გრაფიკი ნაჩვენებია ნახატზე 11.4*a* . *Scree plot* სახელწოდებით ტერმინი შემოღებულია 1966 წელს ბრიტანული წარმომავლობის ამერიკელი ფსიქოლოგის რაიმონდ ბერნარდ კეტელის მიერ — Raymond Bernard Cattell, 1905-1998. ამ გრაფიკს იმისათვის იყენებენ, რომ დადგინდეს :

1. ან მთავარი კომპონენტების რაოდენობა, რომლებიც თავდაპირველად მოცემული დისპერსიიდან პროცენტის ახსნას იძლევა (მაგალითად, ნახატზე 11.4(b) ეს არის 95%) ;

2. ან გრაფიკის «კლაკნის» შესაბამისი მთავარი კომპონენტების რაოდენობა, როცა ახალი მთავარი კომპონენტის დამატებას ინფორმაცია, სავარაუდოდ, უკვე არ მოაქვს.



(*a*) მთავარი კომპონენტებიდან თითოეულის მიერ ახსნილი დისპერსიის პროცენტი. მთავარი კომპონენტების გამოყენებისას ახალი კომპონენტების დამატება უკვე ნამდვილად არ არის ინფორმაციის მომტანი.



(*b*) მთავარი კომპონენტებიდან თითოეულის მიერ ახსნილი დისპერსიის კუმულაციური (ერთობლივი) პროცენტი. თუ მივიჩნევთ, რომ ახსნილი დისპერსიის 95-პროცენტიანი წილი ჩვენთვის მისაღებია, მაშინ სავსებით დაკმაყოფილებული აღმოვჩნდებით ათი მთავარი კომპონენტით.

ნახატი 11.4 – მთავარ კომპონენტთა რიცხვის ასარჩევად ახსნილი დისპერსიების რაოდენობა გამოიყენება.

3.2 მონაცემთა მატრიცის ფაქტორიზაცია

აღვნიშნოთ m-ით მთავარი კომპონენტების რაოდენობა, რომლებიც გამოთვლილია მთავარ კომპონენტთა ანალიზის (PCA) მეთოდით და წარმოდგენილია $W \in \mathbb{R}^{p\times m}$ მატრიცით. n დაკვირვებათა რედუცირებული (შეკვეცილი) რაოდენობა ახალი განზომილების სივრცეში მიიღება X მატრიცის პროეცირებით W მატრიცის სვეტებზე, სხვა სიტყვებით რომ ვთქვათ, შემდეგი გამოსახულების გამოთვლით :

$$H = W^{\mathrm{T}} X \,. \tag{11.5}$$

$H \in \mathbb{R}^{m\times n}$ მატრიცა შეიძლება იყოს ინტერპრეტირებული როგორც მონაცემების ლატენტური, ანუ ფარული (ინგლ. *hidden*, საიდანაც მიღებულია H ნოტაცია) წარმოდგენა. სწორედ ამ წარმოდგენის გამოვლენას ვცდილობდით ჩვენ მთავარ კომპონენტთა ანალიზის (PCA, Principal Component Analysis) მეთოდით.

ვინაიდან W მატრიცის სვეტები წარმოადგენს ორთონორმირებულ ვექტორებს (საქმე ეხება XX^{T} მატრიცის საკუთარ ვექტორებს), შეგვიძლია 11.5 განტოლების გამრავლება მარცხნიდან W მატრიცაზე, იმისათვის, რომ მივიღოთ X მატრიცის *ფაქტორიზაცია* :

$$X = WH \,. \tag{11.6}$$

ამიტომ H მატრიცის სტრიქონებს X მატრიცის *ფაქტორებს* უწოდებენ.

საერთოდ კი მათემატიკაში *ფაქტორიზაცია* არის ობიექტის დეკომპოზიცია, დაშლა სხვა ობიექტთა, ანუ *ფაქტორთა* ნამრავლად, რომლებიც გადამრავლების შედეგად საწყის ობიექტს იძლევა.

ფაქტორული ანალიზი (ფაქტორთა ანალიზი)

ზემოთ განხილული ტიპის ფაქტორიზაცია არის ე.წ. *ფაქტორული ანალიზის* გაცილებით უფრო ზოგადი სტრუქტურის ნაწილი.

დავუშვათ, რომ $\left\{\vec{x}^1, \vec{x}^2, \ldots, \vec{x}^n\right\}$ დაკვირვებები არის $p-$განზომილებიანი ისეთი $\vec{x}$ შემთხვევითი სიდიდის რეალიზაციები, რომელიც გენერირებულია (გაჩენილია) შემდეგი მოდელით :

$$\vec{x} = W\vec{h} + \epsilon, \tag{11.7}$$

სადაც $\vec{h}$ არის $m-$განზომილებიანი $\vec{x}$ შემთხვევითი ცვლადის ლატენტური (ფარული) წარმოდგენა, ხოლო ϵ წარმოადგენს გაუსის (ნორმალურად განაწილებულ) ხმაურს:

$$\epsilon \sim \mathcal{N}\left(0, \Psi\right)$$

$\Psi \in \mathbb{R}^{p\times p}$ მატრიცით. დავუშვათ, რომ მონაცემები ცენტრირებულია 0 მათემატიკური ლოდინის მიმართ და $\vec{h}^1, \vec{h}^2, \ldots, \vec{h}^n$ ლატენტური ცვლადები, რომლებიც $\vec{h}$ – ის რეალიზაციებს წარმოადგენს, არის დამოუკიდებელი და ნორმალურად (გაუსის კანონით) განაწილებული ერთის ტოლი დისპერსიით, ე.ი. $\vec{h} \sim \mathcal{N}\left(0, I_m\right)$, სადაც I_m — იგივეობის მატრიცაა $m\times m$ განზომილებით. მაშინ $W\vec{h}$ ცენტრირებულია $0-$ის მიმართ და მისი კოვარიაცია WW^{T} სახით მოიცემა. ასე რომ,

$$\vec{x} \sim \mathcal{N}\left(0, WW^{\mathrm{T}} + \Psi\right).$$

თუ ჩვენ ასევე ჩავთვლით, რომ ϵ ხმაური *იზოტროპულია* და, ამრიგად, $\Psi = \sigma^2 I_p$, მაშინ $\vec{x} \sim \mathcal{N}\left(0, WW^{\mathrm{T}} + \sigma^2 I_p\right)$, სადაც W და σ პარამეტრების განსაზღვრა დამაჯერებლობის მაქსიმუმის მეთოდით შეიძლება. ასე მივიღებთ იმას, რაც ცნობილია როგორც *ალბათური მთავარ კომპონენტთა ანალიზი — Probabilistic principal components analysis* (Tipping and Bishop, 1999).

კლასიკური მთავარ კომპონენტთა ანალიზი (PCA, Principal Components Analysis) არის ალბათური მთავარ კომპონენტთა ანალიზის (Probabilistic PCA) ზღვრული შემთხვევა, რომელიც მაშინ მიიღება, როცა ხმაურის კოვარიაცია უსასრულო მცირე ხდება $\left(\sigma^2 \to 0\right)$. აქ შეიძლება გავიხსენოთ, რომ ϵ შემთხვევითი სიდიდის კოვარიაცია საკუთარ თავთან ამ შემთხვევითი სიდიდის $\mathrm{Var}\left(\epsilon\right)$ დისპერსიას (σ_ϵ სტანდარტული გადახრის კვადრატს) უდრის :

$$\mathrm{Cov}\left(\epsilon, \epsilon\right) = \mathrm{E}\left(\epsilon^2\right) - \left(\mathrm{E}\left(\epsilon\right)\right)^2 = \mathrm{Var}\left(\epsilon\right) = \sigma_\epsilon^2,$$

სადაც E არის მათემატიკური ლოდინისათვის (ინგლ. Expectation ან Expected value) მიღებული ნოტაცია (აღნიშვნა).

უფრო ზოგად შემთხვევაში შეიძლება დავუშვათ, რომ დანაკვირვები $x_1, x_2, \ldots, x_p$ ცვლადები პირობითად დამოუკიდებელია $h_1, h_2, \ldots, h_m$ ლატენტური ცვლადებისგან. ამ შემთხვევაში Ψ

არის $\Psi = \text{diag}\left(\psi_1, \psi_2, \ldots, \psi_p\right)$ დიაგონალური მატრიცა, სადაც ψ_j აღწერს x_j ცვლადისათვის დამახასიათებელ დისპერსიას . W, σ და $\psi_1, \psi_2, \ldots, \psi_p$ მნიშვნელობები კვლავ შეიძლება იყოს მიღებული მაქსიმალური დამაჯერებლობის მეთოდით. სწორედ რომ ეს არის ცნობილი *ფაქტორული ანალიზის* სახელწოდებით.

ფაქტორულ ანალიზში ჩვენ უკვე არ მივმართავთ დაშვებას ახალ განზომილებათა ორთოგონალურობის შესახებ. კერძოდ, შესაძლებელია გადაგვარებულ განზომილებათა მიღება, სხვა სიტყვებით რომ ვთქვათ, შესაძლებელია W მატრიცის იმ სვეტების მიღება, რომელთა ყველა კოორდინატი 0-ს უდრის.

მატრიცების დადებითი ფაქტორიზაცია

იმ შემთხვევაში, როცა X-ის ყველა მნიშვნელობა დადებითია, უფრო გასაგები შეიძლება იყოს 11.6 განტოლების მსგავსი ფაქტორიზაციის ძებნა, მაგრამ ისეთის, სადაც W და H მატრიცები ასევე დადებითია. ასეთ შემთხვევაში ლაპარაკობენ დადებითი მატრიცული ფაქტორიზაციის (ინგლ. *NMF*, *Non-negative Matrix Factorization*) შესახებ. ამოცანის ფორმალიზებულად ჩამოყალიბება კი ასეთი სახით ხდება :

$$\arg \min_{W \in \mathbb{R}^{p \times m}, H \in \mathbb{R}^{m \times n}} \left\| X - WH \right\|_F^2 , \tag{11.8}$$

სადაც $\left\| \cdot \right\|_F$ აღნიშნავს მატრიცის *ფრობენიუსის* (გერმ. Ferdinand Georg Frobenius, 1849-1917) *ნორმას* (ზოგჯერ ხმარობენ *ევკლიდეს ნორმას*), სხვა სიტყვებით რომ ვთქვათ, კვადრატულ ფესვს მისი შემავალი ელემენტების მოდულთა კვადრატების ჯამიდან. ამრიგად, $\left\| X - WH \right\|_F^2$ ადარებს X და WH მატრიცებს მათი ელემენტების ერთი მეორის მიყოლებით.

ამოცანა 11.8 შეიძლება იყოს გადაწყვეტილი მიმართული დაშვების ალგორითმით (იხ. ქვედანაყოფი 13.3.3).

მაგალითი

ეს ტექნიკა შეიძლება იყოს გამოყენებული, მაგალითად, მაყურებლების მიერ სხვადასხვა ფილმისათვის მიცემული შეფასებების ანალიზის ჩასატარებლად. დავუშვათ, რომ გვყავს n მაყურებელი, რომელიც აძლევს შეფასებას ნულსა და ხუთს შორის p რაოდენობის ფილმს ; არაუარყოფითი ფაქტორიზაცია საშუალებას იძლევა რომელიმე $\left(x_j^i\right)$ მნიშვნელობა ამ შეფასებებიდან იყოს ინტერპრეტირებული როგორც ორი ფაქტორის შეთავსება, შეხამება და შეწყობა. ესენია : i-ური მაყურებლისთვის $\vec{h}^i \in \mathbb{R}_+^m$-ით მოცემული ფილმების m *ასპექტის ფასეულობა* და ფილმში — W მატრიცის $j-$ური სტრიქონით ასახული — ნებისმიერი ამ *ასპექტის არსებობა.*

შენიშვნა

ეს ტექნიკა შეიძლება ასევე იყოს გამოყენებული *მატრიცის შესავსებად, დასასრულებლად, დასამთავრებლად* (ინგლ. matrix completion), ე.ი. იმ შემთხვევაში, როცა X მატრიცას აქვს დაუმთავრებელი, დაუსრულებელი მნიშვნელობები (ყველა მაყურებელს არ შეუფასებია

ყველა ფილმი). ნაცვლად იმისა, რომ ვეცადოთ $\|X - WH\|_F^2$ თანაფარდობის მინიმიზაცია, ჩვენ ვზღუდავთ ამ ჯამს X მატრიცის ცნობილი ელემენტებით (შემავალი მონაცემებით) : მაშინ 11.8 ამოცანა შემდეგ სახეს იძენს :

$$\arg \min_{W \in \mathbb{R}_+^{p \times m},\, H \in \mathbb{R}_+^{m \times n}} \sum_{i,j \text{ გამოუტოვებლად}} \left(X_{ij} - (WH)_{ij}\right)^2 . \tag{11.9}$$

იგივე ამოცანა ადვილი გადასაწყვეტია მიმართული დაშვების ალგორითმით (იხ. ქვედანაყოფი 13.3.3), და მაშინ X შეიძლება იყოს აპროქსიმირებული WH ნამრავლით. ეს X მატრიცის გამოტოვებული შემავალი მონაცემების წინასწარმეტყველების საშუალებას იძლევა WH მატრიცის შესაბამისი ელემენტებით.

ეს ტექნიკა გამოიყენება ე.წ. *კოლაბორაციული (ერთობლივი) ფილტრაციისათვის* (ინგლ. collaborative filtering ; იხ. ქვედანაყოფი 8.4), სადაც X მატრიცის სვეტები შეესაბამება მომხმარებლებს, ხოლო სტრიქონები — საქონელს.

3.3 ავტოენკოდერები

ქვედანაყოფში 7.2.5 დავრწმუნდით, რომ ხელოვნური ნეირონული ქსელები შეიძლება იყოს ინტერპრეტირებული როგორც ინსტრუმენტი მონაცემთა ისეთი მოდელის სწავლებისათვის (დაწყებული უკანასკნელი შუალედური შრიდან), რომელიც ადვილად ემორჩილება წრფივ კონტროლირებად სწავლებას. სწორედ ამ იდეას მივყავართ *ავტოენკოდერის* ცნებამდე, რომელიც ნეირონული ქსელების არაკონტროლირებადი სწავლებისათვის განზომილების რედუქციის (შეკვეცის) გამოყენების საშუალებას იძლევა.

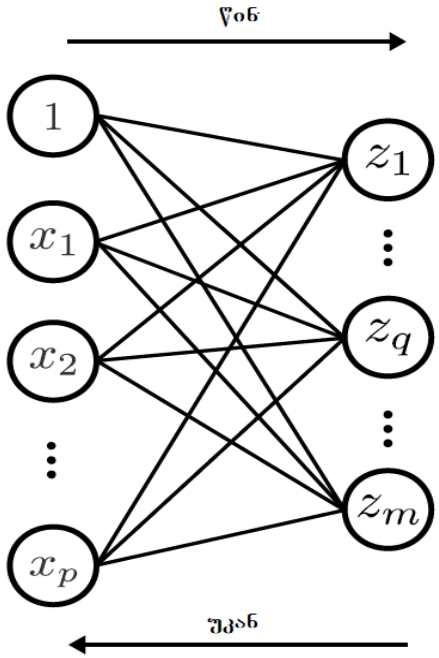
განსაზღვრება 11.6 (ავტოენკოდერი) *ავტოენკოდერი* — ეს ნეირონული ქსელია, რომლის გამომავალი ფენა შემავალი ფენის იდენტურია, ხოლო შუალედური შრე მათ ზომაზე ნაკლებია. ამ შუალედური შრის გამოსასვლელზე მიიღება მონაცემთა ახალი, უფრო კომპაქტური წარმოდგენა.

მიუხედავად იმისა, რომ წრფივი ავტოენკოდერები არსებობს, მათთან შედარებისას, ჩვეულებრივ, უპირატესობას *ბოლცმანის შეზღუდულ მანქანებს* (ინგლ. RBM, Restricted Boltzmann Machines) ანიჭებენ. ისინი შემოთავაზებული იყო 1986 წელს ამერიკელი მეცნიერ-კოგნიტივისტის პოლ სმოლენსკის (Paul Smolensky, 1955-) მიერ.

ბოლცმანის შეზღუდული მანქანა ნეირონების ორ შრეს შეიცავს, როგორც ეს ნაჩვენებია ნახატზე 11.5. პირველი შრე შეიცავს p ნეირონს, თითო-თითოს მონაცემთა ამსახველ x_j ცვლადზე ; მეორე ფენა შეიცავს m ნეირონს, სადაც $m < p$. ამ შრის გამოსასვლელზე $\vec{z} = (z_1, z_2, \ldots, z_m)$ ვექტორი არის $\vec{x}$-ის რედუცირებული (შეკვეცილი) წარმოდგენა. ჩავთვალოთ, რომ ეს ცვლადები ბინარულია მნიშვნელობებით $\{0,1\}$-ზე : ისინი შეიძლება შეესაბამოდეს პიქსელის ფერს შავ-თეთრ გამოსახულებაში.

ეს არქიტექტურა არის *შეზღუდული*, ვინაიდან ერთისა და იმავე შრის ნეირონებს შორის კავშირები არ არსებობს, ბოლცმანის მიერ ადრე შემოთავაზებული მანქანებისგან განსხვავებით,

რომლებიც სტატისტიკურ ფიზიკაში გამოიყენება. მეორე მხრივ, კავშირები ორ შრეს შორის ორივე მიმართულებით ვრცელდება : შემავალი შრიდან შუალედურისკენ და ასევე შუალედური შრიდან შემავალი შრისაკენ. ბმათა წონების არჩევა ისეთნაირად ხდება, რომ მოხდეს შესასვლელსა და გამოსასვლელს შორის იმ სხვაობის მინიმიზაცია, რომელიც მიიღება ქსელის გავლისას ჯერ წინ და შემდეგ უკან.



ნახატი 11.5 – არქიტექტურა ბოლცმანის შეზღუდული მანქანის, რომელიც აგებს p – განზომილებიანი მონაცემების m – განზომილებიან ასახვას.

ბოლცმანის შეზღუდული მანქანების თავისებურება ისაა, რომ ისინი იყენებს აქტივაციის სტოქასტიკურ ფუნქციებს :

$$\mathbb{P}\left(z_q = 1 | \vec{x}\right) = \sigma\left(b_q + \sum_{j=1}^{p} w_{jq} x_j\right), \tag{11.10}$$

სადაც b_q არის წანაცვლება $q-$ური შუალედური ნეირონის და

$$\mathbb{P}\left(x_j = 1 | \vec{z}\right) = \sigma\left(a_j + \sum_{q=1}^{m} w_{jq} z_q\right), \tag{11.11}$$

სადაც a_j არის $j-$ური ხილული ნეირონის წანაცვლება.

სტატისტიკურ ფიზიკაზე დაყრდნობით, ჩვენ შეგვიძლია ამ მოდელის ენერგიის განსაზღვრა შემდეგი სახით :

$$E\left(\vec{x}, \vec{z}\right) = -\sum_{j=1}^{p} a_j x_j + \sum_{q=1}^{m} b_q z_q - \sum_{j=1}^{p}\sum_{q=1}^{m} x_j w_{jq} z_q \,. \tag{11.12}$$

ამ მოდელის თანახმად, ალბათობის განაწილების ერთობლივი კანონი $\vec{x}$ და $\vec{z}$ შემთხვევითი

სიდიდეებისათვის შეიძლება იყოს ჩაწერილი როგორც ხსენებული ენერგიის ფუნქციად ასახული ბოლცმანის განაწილების კანონი :

$$\mathbb{P}\left(\vec{x},\vec{z}\right)=\frac{1}{Z}e^{-E(\vec{x},\vec{z})}\,, \tag{11.13}$$

სადაც Z ნორმირების კონსტანტა შეიძლება იყოს განხილული დაყოფის ფუნქციად (ინგლ. *Partition function*) და ჩაიწეროს ყველა შესაძლო მდგომარეობის შესაბამისი $\mathbb{P}\left(\vec{x},\vec{z}\right)$ ალბათობის შეკრებით :

$$Z=\sum_{\vec{u}\in\{0,1\}^p,\vec{v}\in\{0,1\}^m}e^{-E(\vec{u},\vec{v})}\,. \tag{11.14}$$

ამრიგად, შეგვიძლია გავიგოთ $\left\{a_j\right\}_{j=1,\ldots,p}$, $\left\{b_q\right\}_{q=1,\ldots,m}$ და $\left\{w_{jq}\right\}_{j=1,\ldots,p,q=1,\ldots,m}$ კოეფიციენტები მაქსიმალური დამაჯერებლობის მეთოდით.

მოცემული $\vec{x}^i$ დაკვირვებისათვის შექცეული ლოგარითმული დამაჯერებლობა უდრის შემდეგ გამოსახულებას :

$$-\log P\left(\vec{x}^i\right)=-\log\sum_{\vec{v}}P\left(\vec{x}^i,\vec{v}\right)=\log\sum_{\vec{u},\vec{v}}e^{-E(\vec{u},\vec{v})}-\log\sum_{\vec{v}}e^{-E\left(\vec{x}^i,\vec{v}\right)}\,. \tag{11.15}$$

ახლა ჩვენ გვჭირდება ამ თანაფარდობის მიღება ყოველი იმ პარამეტრის მიმართ, რომლის დადგენასაც ვცდილობთ. თუ ერთ-ერთ ასეთ პარამეტრს აღვნიშნავთ θ სიმბოლოთი (ასე რომ θ შეიძლება ერთნაირად ასახავდეს a_j, b_q ან w_{jq} სიდიდეებიდან ნებისმიერს), მაშინ $\left(-\log P\left(\vec{x}^i\right)\right)$ სიდიდის კერძო წარმოებული θ ცვლადით იქნება :

$$\left.\begin{aligned}\frac{\partial\left(-\log P\left(\vec{x}^i\right)\right)}{\partial\theta}&=-\frac{1}{\sum_{\vec{u},\vec{v}}e^{-E(\vec{u},\vec{v})}}\sum_{\vec{u},\vec{v}}\frac{\partial E\left(\vec{u},\vec{v}\right)}{\partial\theta}e^{-E(\vec{u},\vec{v})}+\frac{1}{\sum_{\vec{v}}e^{-E\left(\vec{x}^i,\vec{v}\right)}}\sum_{\vec{v}}\frac{\partial E\left(\vec{x}^i,\vec{v}\right)}{\partial\theta}e^{-E\left(\vec{x}^i,\vec{v}\right)}\\&=-\sum_{\vec{u},\vec{v}}\mathbb{P}\left(\vec{u},\vec{v}\right)\frac{\partial E\left(\vec{u},\vec{v}\right)}{\partial\theta}+\sum_{\vec{v}}\mathbb{P}\left(\vec{v}\mid\vec{x}^i\right)\frac{\partial E\left(\vec{x}^i,\vec{v}\right)}{\partial\theta}\end{aligned}\right\}.$$

ასე რომ, ეს გრადიენტი დაიყვანება ერთ *უარყოფით გრადიენტამდე* :

$$-\sum_{\vec{u},\vec{v}}\mathbb{P}\left(\vec{u},\vec{v}\right)\frac{\partial E\left(\vec{u},\vec{v}\right)}{\partial\theta}=-\mathbb{E}\left[\frac{\partial E\left(\vec{u},\vec{v}\right)}{\partial\theta}\right] \tag{11.16}$$

და კიდევ ერთ *დადებით გრადიენტამდე* :

$$\sum_{\vec{v}}\mathbb{P}\left(\vec{v}\mid\vec{x}^i\right)\frac{\partial E\left(\vec{x}^i,\vec{v}\right)}{\partial\theta}=\mathbb{E}\left[\frac{\partial E\left(\vec{x}^i,\vec{v}\right)}{\partial\theta}\right]. \tag{11.17}$$

ამ მათემატიკურ ლოდინთა სააპროქსიმაციოდ გამოვიყენოთ *გიბსის (Gibbs) ამოკრების (ამორჩევის) პროცედურა*, რომელიც მდგომარეობს $\vec{x}^i$ დაკვირვების იტერაციულ ამოღებაში,

პირდაპირი სვლის შესრულებაში $\vec{z}^i$ –ის მისაღებად, შექცეული სვლის (უკუსვლის) შესრულებაში $\vec{x}'^i$ გამოსასვლელის მისაღებად, შემდეგ კი უკანასკნელი პირდაპირი სვლის განხორციელებაში $\vec{z}'^i$ –ის მისაღებად და ჩვენი გრადიენტის აპროქსიმაციისათვის შემდეგი თანაფარდობის დახმარებით :

$$\frac{\partial E\left(\vec{x}^i, \vec{z}^i\right)}{\partial \theta} - \frac{\partial E\left(\vec{x}'^i, \vec{z}'^i\right)}{\partial \theta}. \tag{11.18}$$

11.12 განტოლებით აღწერილი E ენერგიის კერძო წარმოებულები ადვილად განისაზღვრება.

ეს გვაძლევს ეგრეთ წოდებული *კონტრასტული დივერგენციის* (ინგლ. contrastive divergence) ალგორითმს ბოლცმანის შეზღუდული მანქანის სწავლებისათვის. ხსენებული ალგორითმი პირველად შემოთავაზებული იყო ჯეფ ჰინტონის (Geoff Hinton, 2002) მიერ.

განსაზღვრება 11.7 (კონტრასტული დივერგენცია) დავუშვათ, რომ მოცემულია ბოლცმანის შეზღუდული მანქანა, რომლის ენერგია ასახულია განტოლებით 11.12 და სწავლების სიჩქარე η სიდიდით. მაშინ *კონტრასტული დივერგენციის* პროცედურა მდგომარეობას a_j, b_q და w_{jq} წონების სწავლების განხორციელებაში დაკვირვებათა გადარჩევით ამ წონების ინიციალიზაციის შემდეგ შემთხვევითი მნიშვნელობით. $\vec{x}^i$ დაკვირვებისათვის სრულდება შემდეგი ბიჯები :

1. შუალედური $\vec{z}^i$ ვექტორის ამოღება $\mathbb{P}\left(\vec{z} \mid \vec{x}^i\right)$ კანონის შესაბამისად ;
2. დადებითი გრადიენტის გამოთვლა, რომელიც იქნება x_j^i, თუ $\theta = a_j$; z_q^i, თუ $\theta = b_q$; დაბოლოს, $x_j^i z_q^i$, თუ $\theta = w_{jq}$;
3. $\vec{x}^i$ დაკვირვების $\vec{x}'^i$ რეკონსტრუქციის (აღდგენის) ამოღება $\mathbb{P}\left(\vec{x} \mid \vec{z}^i\right)$ კანონის შესაბამისად ;
4. შუალედური $\vec{z}'^i$ ვექტორის ამოღება $\mathbb{P}\left(\vec{z} \mid \vec{x}'^i\right)$ კანონის შესაბამისად ;
5. უარყოფითი გრადიენტის გამოთვლა, რომელიც იქნება $-x_j'^i$, თუ $\theta = a_j$; $-z_q'^i$, თუ $\theta = b_q$; დაბოლოს, $-x_j'^i z_q'^i$, თუ $\theta = w_{jq}$;
6. წონების განახლება შემდეგი თანაფარდობების შესაბამისად :

$$w_{jq} \leftarrow w_{jq} + \eta\left(x_j^i z_q^i - x_j'^i z_q'^i\right);\ a_j \leftarrow a_j + \eta\left(x_j^i - x_j'^i\right);\ b_q \leftarrow b_q + \eta\left(z_q^i - z_q'^i\right).$$

⚙

შენიშვნა

ბოლცმანის შეზღუდული მანქანები (ინგლ. RBM, Restricted Boltzmann Machines) შეიძლება იყოს გაერთიანებული ღრმა არქიტექტურის ფორმირებისათვის, რომელიც ცნობილია როგორც *ნდობის ღრმა ქსელი* (ინგლ. DBN, *Deep Belief Network*). ეს არქიტექტურა გამოიყენება ზომების არაკონტროლირებადი რედუქციისთვის (შესამცირებლად), თუმცა შეიძლება იყოს

გამოყენებული ზომების კონტროლირებადი რედუქციის მიზნითაც. ამ შემთხვევაში ნდობის ღრმა ქსელის უკანასკნელი შრის მიერთება ხდება გამომავალ შრესთან, რაც იძლევა კონტროლირებად ნეირონულ ქსელს. ჯეფ ჰინტონისა და რუსლან სალახუტდინოვის მიერ რეალიზებული ნდობის ღრმა ქსელი იყო ერთ-ერთი პირველი მოქმედი სტრუქტურა ღრმა ნეირონული ქსელების კლასში (Geoff Hinton and Ruslan Salakhutdinov, 2006).

მოკლედ რომ ვთქვათ, ნდობის ღრმა ქსელი (ინგლ. Deep Belief Network, DBN) — ეს წარმომქმნელი (წარმომშობი, შემქმნელი) გრაფიკული მოდელია, ანუ, სხვა სიტყვებით რომ ვთქვათ, სიღრმისეული ნეირონული ქსელების ერთ-ერთი ტიპი რამდენიმე ფარული შრით, რომლებშიც ერთისა და იმავე შრის ნეირონები კავშირში არ არის ერთმანეთთან და ბმები მხოლოდ მეზობელი შრის ნეირონებთან არსებობს.

3.4 სხვა არაწრფივი მიდგომა

მონაცემთა მოცულობის არაწრფივი შემცირებისათვის მრავალი მიდგომა იყო შემოთავაზებული. აქ ჩვენ განვიხილავთ რამდენიმე ყველაზე პოპულარულს მათ შორის ; მათი დაწვრილებითი ახსნა ცდება მანქანური სწავლების საწყისთა ამ კურსს.

ბირთვული მთავარი კომპონენტების ანალიზი

დავიწყოთ იმით, რომ მთავარი კომპონენტების ანალიზი ბირთვის გამოყენების საშუალებასაც იძლევა (იხ. ქვედანაყოფი 10.3.3). მიღებულ მეთოდს ეწოდება *ბირთვული მთავარი კომპონენტების ანალიზი* (ინგლ. *kPCA, Kernel Principal Component Analysis*).

იგი გამოიყენება მანქანურ სწავლებაში განზომილების არაწრფივი შემცირების მიზნით და წარმოადგენს მთავარი კომპონენტების ანალიზის კლასიკური ალგორითმის (ინგლ. PCA, Principal Component Analysis) გაფართოებას. უკანასკნელი წრფივი მეთოდია, რომელიც განსაზღვრავს მონაცემთა სიმრავლის განსაკუთრებულად მნიშვნელოვან ნიშნებს ან კომპონენტებს.

მრავალგანზომილებიანი პოზიციონირება

მრავალგანზომილებიანი პოზიციონირება, ანუ *მრავალგანზომილებიანი სკალირება* (ინგლ. MDS, MultiDimensional Scaling) შემოტანილია ხმარებაში მაიკლ კოკსისა და ტრევორ კოკსის მიერ (Michael A. A. Cox and Trevor F. Cox, 1994) და ეფუძნება დაკვირვებათა შორის $D \in \mathbb{R}^{n \times n}$ *განსხვავებულობის* მატრიცას (ინგლ. *dissimilarity matrix*). ხსენებული განსხვავებულობა შეიძლება იყოს მოცემული მეტრიკული მანძილით, მაგრამ ეს სავალდებულო არ არის. ალგორითმის მიზანია — ნაპოვნი იქნეს მონაცემთა ისეთი წარმოდგენა, რომელიც იცავს, ინარჩუნებს ამ განსხვავებულობას :

$$X^* = \arg\min_{Z \in \mathbb{R}^{n \times m}} \sum_{i=1}^{n} \sum_{l=i+1}^{n} \left(\left\| \vec{z}^{\,i} - \vec{z}^{\,l} \right\|_2 - D_{il} \right)^2 . \tag{11.19}$$

თუ ვიყენებთ ევკლიდეს მანძილს განსხვავებულობის, არამსგავსების ზომად, მაშინ მრავალგანზომილებიანი სკალირება (MDS, MultiDimensional Scaling) მთავარი კომპონენტების ანალიზის (PCA, Principal Component Analysis) ეკვივალენტურია.

მრავალგანზომილებიანი პოზიციონირება შეიძლება ასევე გამოიყენებოდეს ისეთი წერტილე-

ბის პოზიციონირებისათვის m – განზომილებიან სივრცეში, რომელთა კოორდინატები უცნობია. მაგალითად, მისი გამოყენება შეიძლება რუკაზე ქალაქების პოზიციათა შესაცვლელად მათ შორის არსებული მხოლოდ მანძილების გამოყენებით.

მრავალგანზომილებიანი სკალირების (MDS, MultiDimensional Scaling) ნაკლოვანებათა შორის ერთ-ერთი შეზღუდვა ისაა, რომ იგი ცდილობს მანძილის შენარჩუნებას დაკვირვებებს შორის მხოლოდ გლობალურ მასშტაბში. მრავალგანზომილებიანი სკალირების განსხვავებულობის მატრიცის აგების ეფექტურ ხერხს მონაცემთა ლოკალური სტრუქტურის შესანარჩუნებლად წარმოადგენს *IsoMap* (Joshua B. Tenenbaum et al., 2000) ალგორითმი, რომელიც გულისხმობს დაკვირვებებს შორის მეზობლობის გრაფის აგებას თითოეული მათგანის დაკავშირებით k უახლოეს მეზობელ დაკვირვებასთან.

ეს წიბოები შეიძლება შეიწონოს მანძილით დაკვირვებათა შორის, რომლებსაც ისინი აკავშირებს. შემდეგ მეზობლობის ამ გრაფზე შეიძლება გაანგარიშდეს განსხვავებულობა დაკვირვებებს შორის, ვთქვათ, ორ წერტილს შორის უმოკლესი გზის სიგრძის გამოყენებით.

t-SNE

დაბოლოს, *t-SNE* (ინგლ. *t-Student Neighborhood Embedding* ან *t-Distributed Stochastic Neighbor Embedding*) ალგორითმი (*მეზობლების სტოქასტიკური ჩალაგება t-განაწილებით*), რომელიც აღმოჩენილია 2008 წელს ლორენს ვან დერ მაატენისა (Laurens van der Maaten) და ჟეფ ჰინტონის (Geoff Hinton) მიერ, გვთავაზობს დაკვირვებებს შორის მანძილების განაწილების აპროქსიმაციას ე.წ. «*სტუდენტის*» განაწილებით (van der Maaten and Hinton, 2008).

ხსენებული განაწილების კანონი ეკუთვნის ინგლისელ სტატისტიკოსს უილიამ სილი გოსეტს (William Sealy Gosset, 1876-1937), რომელიც ცნობილია «Student»-ის ფსევდონიმით. მისი დამსაქმებლის ამხედრების გამო ამ ნიჭიერი თანამშრომლის მიერ ნებისმიერი პუბლიკაციის გამოქვეყნების წინააღმდეგ, სწორედ «Student»-ის ფსევდონიმით დაიბეჭდა 1908 წელს გოსეტის სტატია ე.წ. t– განაწილების ფუნქციის შესახებ.

ყოველი $\vec{x}^i$ დაკვირვებისათვის განისაზღვრება P_i სიდიდე, როგორც შემდეგი სახით მოცემული «სტუდენტის» t– განაწილების — ანუ უბრალოდ t– განაწილების — ალბათობა :

$$P_i\left(\vec{x}\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left\| \vec{x} - \vec{x}^i \right\|_2^2}{2\sigma^2} \right). \tag{11.20}$$

t– SNE გულისხმობს შემდეგი პრობლემის გადაჭრას :

$$\arg\min_Q \sum_{i=1}^{n} KL\left(P_i \,||\, Q_i\right), \tag{11.21}$$

სადაც KL აღნიშნავს კულბეკ-ლეიბლერის *დივერგენციას*, ანუ *შეუთანხმებლობას* — Kullback–Leibler divergence (იხ. ქვედანაყოფი 2.4.1), ხოლო Q აირჩევა p – ზე ნაკლები განზომილების «სტუდენტის» t– განაწილებებიდან.

კულბეკ-ლეიბლერის დივერგენცია, ანუ შეუთანხმებლობა შემოღებული იყო ამერიკელი კრიპტოანალიტიკოსებისა და მათემატიკოსების სოლომონ კულბეკისა (Solomon Kullback,

1907–1994) და რიჩარდ ლეიბლერის (Richard A. Leibler, 1914-2003) მიერ 1951 წელს.

ყურადღება მიაქციეთ იმ გარემოებას, რომ ეს ალგორითმი პოულობს ლოკალურ და არა გლობალურ მინიმუმს, ამიტომ გამოყენებული ინიციალიზაციის შესაბამისად შეიძლება სულ სხვადასხვა შედეგი იყოს მიღებული. გარდა ამისა.

ამ ალგორითმის სირთულე კვადრატულად არის დამოკიდებული დაკვირვებათა რაოდენობაზე.

4. საკვანძო მომენტები

- მონაცემთა განზომილების შემცირება კონტროლირებადი სწავლების ალგორითმის გამოყენების წინ გვეხმარება ამ ალგორითმის დროითი და სივრცითი საჭიროებები შემცირებაში, ასევე მისი ფუნქციური მახასიათებლების გაუმჯობესებაში.
- განასხვავებენ *ცვლადების შერჩევას*, რომელიც მდგომარეობს ჭარბი ან არაინფორმაციული ცვლადების გამორიცხვაში და *ცვლადების ამოღებას*, რომელიც მდგომარეობს მონაცემთა ახალი წარმოდგენის შექმნაში.
- შესაძლებელია მონაცემების ვიზუალიზაცია მათი პროეცირებით ორგანზომილებიან სივრცეზე, მაგალითად, მთავარ კომპონენტთა ანალიზის (PCA, Principal Component Analysis) საშუალებით ან მეზობლების სტოქასტიკური $t-$განაწილებით ჩალაგების ($t-$SNE, $t-$Student Neighborhood Embedding) მიდგომის გამოყენებით.
- ცვლადთა განზომილების შესამცირებლად მრავალი მეთოდის გამოყენება შეიძლება.

დამატებითი ინფორმაცია

- მთავარ კომპონენტთა ანალიზის დაწვრილებითი ანალიზი მოცემულია დამხმარე სახელმძღვანელოში Jonathon Shlens (2014).
- ცვლადების შერჩევის მეთოდების შესახებ დაწვრილებითი მიმოხილვის გასაცნობად შეიძლება მივმართოთ ნაშრომს Isabelle Guyon და André Elisseeff (2003).
- უფრო დაწვრილებით არაუარყოფითი მატრიცების ფაქტორიზაციაზე (NMF, Non-negative Matrix Factorization) დამატებითი ინფორმაციის მიღება შეიძლება, მაგალითად, სტატიიდან Daniel D. Lee & H. Sebastian Seung (1999) .
- უფრო დაწვრილებითი ინფორმაციის მისაღებად მომვლების (შეხვევის, გახვევის) მეთოდების შესახებ (ინგლ. wrapper methods) ცვლადების სიმრავლეთა სელექციისთვის (შერჩევისთვის) შეიძლება მივმართოთ ალან მილერის წიგნს Miller (1990).
- განზომილების შემცირების არაწრფივ Isomap მეთოდს ეძღვნება შემდეგი ვებგვერდი : http://web.mit.edu/cocosci/isomap/isomap.html .
- უფრო დაწვრილებითი ინფორმაციის მისაღებად $t-$SNE ალგორითმის გამოყენების შესახებ შესაძლებელია მივმართოთ https://lvdmaaten.github.io/tsne/ ვებგვერდს ან ინტერაქტიურ პუბლიკაციას — Wattenberg et al. (2016).

5. ბიბლიოგრაფია

1. Cox, T. F. and Cox, M. A. A. (1994). *Multidimensional Scaling*. Chapman and Hall., London.

2. Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3 :1157–1182.

3. Hinton, G. E. (2002). Training product of experts by minimizing contrastive divergence. *Neural Computation*, 14 :1771–1800.

4. Hinton, G. E. and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, 313 :504–507.

5. Kozachenko, L. F. and Leonenko, N. N. (1987). A statistical estimate for the entropy of a random vector. *Problemy Peredachi Informatsii (Problems Inform. Transmission)*, 23 :9–16.

6. Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755) :788–791.

7. Miller, A. J. (1990). *Subset Selection in Regression*. Chapman and Hall., London.

8. Shlens, J. (2014). A Tutorial on Principal Component Analysis. *arXiv [cs, stat]*. arXiv : 1404.1100.

9. Smolensky, P. (1986). Information processing in dynamical systems : foundations of harmony theory. In *Parallel Distributed Processing : Explorations in the Microstructure of Cognition*, volume 1 : Foundations, chapter 6, pages 194–281. MIT Press, Cambridge, MA.

10. Tenenbaum, J. B., de Silva, V., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500) :2319–2323.

11. Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal components analysis. *Journal of the Royal Statistical Society Series B*, 61 :611–622.

12. van der Maaten, L. and Hinton, G. (2008). Visualizing data using $t-$SNE. *Journal of Machine Learning Research*, 9 :2579–2605.

13. Wattenberg, M., Viégas, F., and Johnson, I. (2016). How to use $t-$SNE effectively. *Distill*. http://distill.pub/2016/misread-tsne .

6 სავარჯიშოები

11.1 ანუკას აქვს ორგანზომილებიანი მონაცემები, რომლებიც განლაგებულია $x_2 = x_1$ განტოლების დიაგონალზე.

1. როგორია იმ ორი მთავარი კომპონენტის კოორდინატები, რომლებიც მიღებულია მთავარ კომპონენტთა ანალიზის შედეგად (ინგლ. PCA, Principal Component Analysis) ?

2. როგორია ფაქტორული ანალიზის შედეგად მიღებული ორი მთავარი კომპონენტის კოორდინატები ?

11.2 ბუხუტიმ შეამცირა თავისი მონაცემების განზომილება. მას სურს მეხსიერებაში ამ მონაცემების მხოლოდ შემცირებული ვერსიის შენახვა, მაგრამ, აუცილებლობის შემთხვევაში, უნდა დარჩეს ამ მონაცემებიდან საწყისი მონაცემების ამოღების შესაძლებლობაც. მიიღწევა ეს ?

11.3 კატოს აქვს 100 ცვლადით წარმოდგენილი მონაცემები. იგი იყენებს მთავარ კომპონენტთა ანალიზს (ინგლ. PCA, Principal Component Analysis) და სურს იმდენი კომპონენტის დატოვება, რამდენიც აუცილებელია მისი მონაცემების დისპერსიის 90%-ის ასახსნელად.

1. რამდენი კომპონენტი ექნება კატოს იმ შემთხვევაში, როცა დაკვირვებები თითქმის იდეალურად შეთანხმებული ?

2. რამდენი კომპონენტი ექნება მას იმ შემთხვევაში, როცა დაკვირვებები შემთხვევითია ?

11.4 ჯემალს მონაცემთა 6-განზომილებიანი $\mathcal{D} = \left\{\left(\vec{x}^i, y^i\right)\right\}_{i=1,\ldots,n}$ ნაკრები აქვს. იგი ახორციელებს ნეიროქსელის სწავლებას ასეთ მონაცემებზე

$$\left\{\left(\left(x_1^i + x_2^i, x_2^i + x_3^i - x_4^i, x_6^i - x_5^i\right), y^i\right)\right\}_{i=1,\ldots,n}.$$

თუ ეხება საქმე განზომილების რედუქციას, ანუ შემცირებას ?

11.5 ელენეს აქვს 12 ცვლადით წარმოდგენილი 200 000 დაკვირვების შემცველი მონაცემების კრებული. თავისი მონაცემების ანალიზის დასაწყებად მას სურს მათი ვიზუალიზაცია ორ განზომილებაში. რას ურჩევთ მას ? მთავარ კომპონენტთა ანალიზის (ინგლ. PCA, Principal Component Analysis) მეთოდის გამოყენებას თუ მეზობლების სტოქასტიკურ ჩაშენებას (ჩადებას) $\mathrm{t}-$განაწილებით (ინგლ. $\mathrm{t}-$SNE, $\mathrm{t}-$distributed Stochastic Neighbor Embedding ?

11.6 ფორეს აქვს 100 ცვლადით აღწერილი 100 000 დაკვირვება. სინამდვილეში ჭდეები იყო შექმნილი მხოლოდ პირველი 5 ცვლადის გამოყენებით. მომდევნო 5 ცვლადი არის ამ 5 პირველი ცვლადის ზუსტი ასლი. დაბოლოს, დარჩენილი 90 ცვლადი არის შემთხვევითი.

1. რამდენი ცვლადი იქნება შერჩეული ფილტრაციის მიდგომის გამოყენებისას ?

2. რამდენი ცვლადი იქნება შერჩეული ჩაშენების (ჩადების) მიდგომის (ინგლ. the embedded approach) გამოყენებისას ?

11.7 **kPCA**(**k**ernel **P**rincipal **C**omponent **A**nalysis — ბირთვის მთავარ კომპონენტთა ანალიზი). დავუშვათ, რომ რაღაც საგვარეულო $\mathcal{X}$ სივრცეში არის n რაოდენობის $\vec{x}^1, \vec{x}^2, \ldots, \vec{x}^n$ დაკვირვება. დავუშვათ ასევე, რომ $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ არის ბირთვი. არსებობს ჰილბერტის (David Hilbert, 1862-1943, გერმანელი მათემატიკოსი) $\mathcal{H}$ სივრცე და ისეთი $\phi : \mathcal{X} \to \mathcal{H}$ ოპერაცია, რომ $k\left(\vec{x}, \vec{x}'\right) = \left\langle \phi\left(\vec{x}\right), \phi\left(\vec{x}'\right)\right\rangle_{\mathcal{H}}$.

დავუშვათ, რომ $K \in \mathbb{R}^{n \times n}$ არის იორგენ გრამის (Jørgen Pedersen Gram,1850-1916, დანიელი მათემატიკოსი) მონაცემთა მატრიცა. დავუშვათ ასევე, რომ Σ არის $\mathcal{H}$ სივრცეში დაკვირვებათა სახეების კოვარიაციული მატრიცა : $\Sigma = \frac{1}{n}\sum_{i=1}^{n} \phi(\vec{x}^i)\phi(\vec{x}^i)^{\mathrm{T}}$.

1. დავუშვათ, რომ $\left(\lambda, \vec{v}\right)$ არის Σ მატრიცის წყვილი (საკუთარი მნიშვნელობა, საკუთარი ვექტორი). აჩვენეთ, რომ $\vec{v}$ შეიძლება იყოს ჩაწერილი $\phi(\vec{x}^i)$ სიდიდეთა წრფივი კომბინაციის სახით.

2. აჩვენეთ, რომ $n\lambda K\alpha = K^2\alpha$.

3. აჩვენეთ, რომ $\mathcal{H}$ სივრცეში დაკვირვებათა სახეების მთავარი კომპონენტები შეიძლება იყოს მიღებული გრამის K მატრიცის სპექტრული დაშლით

4. აჩვენეთ, რომ $\vec{x} \in \mathcal{X}$ დაკვირვების $\phi(\vec{x})$ სახის პროექცია ჰილბერტის $\mathcal{H}$ სივრცეში დაკვირვებათა სახეების მთავარ კომპონენტზე შეიძლება იყოს გამოთვლილი ϕ ფუნქციის ან $\mathcal{H}$ სივრცის გამოუყენებლად.

11.8 დავუშვათ, რომ $\mathbb{R}^p$ სივრცეში გვაქვს ორი $\mathcal{A}$ და $\mathcal{B}$ სფერო ცენტრებით კოორდინატთა სათავეზე. $\mathcal{A}$ სფეროს რადიუსი არის 1, ხოლო $\mathcal{B}$ სფეროს რადიუსი $1-\varepsilon$ მნიშვნელობის ტოლია.

1. როგორია $\mathcal{A}$ სფეროს მოცულობის ფარდობითი წილი, რომელიც არ არის $\mathcal{B}$ სფეროში ?

2. რა კავშირი აქვს ამას ე.წ. *განზომილების რისხვასთან* ?

განზომილებათა რისხვა მანქანური სწავლების ერთ-ერთი უმნიშვნელოვანესი პრობლემაა, რომელიც გვეუბნება : რაც უფრო მეტია განზომილება, მით უფრო გაიშვიათებულია მონაცემები. სხვა სიტყვებით რომ ვთქვათ, ნიშანთა რაოდენობის ზრდასთან ერთად ჩვენ მიერ განსაზოგადებელი მონაცემების მოცულობა ექსპონენციალურად იზრდება. ინგლისურად განზომილების რისხვა გადმოიცემა «*the curse of dimensionality*» ტერმინით, რომელიც შემოტანილია ხმარებაში ამერიკელი მათემატიკოსის რიჩარდ ბელმანის (Richard Ernest Bellman,1920–1984) მიერ.

სავარჯიშოთა ამონახსნები

11.1

1. პირველი მთავარი კომპონენტი — ეს *დიაგონალის მიმმართველი ვექტორია* (ინგლ. the direction vector of the diagonal), $1-$ის ტოლი ნორმით, ამიტომ ეს ვექტორი $\left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right)$ სახისაა.

მეორე მთავარი კომპონენტი პირველის ორთოგონალურია, ასევე $1-$ის ტოლი ნორმით, ამიტომ ეს ვექტორი $\left(\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2}\right)$ სახისაა.

2. პირველი მთავარი კომპონენტი — ეს მთავარ კომპონენტთა ანალიზის (ინგლ. PCA, Principal Component Analysis) საშუალებით მიღებული კომპონენტია. ვინაიდან ეს კომპონენტი ახსნას მთლიან დისპერსიას აძლევს, ხოლო მეორე ვალდებული არ არის იყოს ორთოგონალური პირველის მიმართ, მეორე კომპონენტს $(0,0)$ კოორდინატები ექნება.

11.2 ზოგად შემთხვევაში, არა! განზომილების შემცირება შეიძლება იყოს განხილული დანაკარგებიან კომპრესიად (შეკუმშვად).

11.3

1. მხოლოდ ერთი კომპონენტი.

2. დაახლოებით 90 კომპონენტი.

11.4 კი : მონაცემები, რომლებიც წარმოდგენილი იყო ექვს განზომილებაში, ახლა წარმოდგენილია სამში.

11.5 ვინაიდან მანქანური სწავლების მძლავრი $t-$SNE (ინგლ. $t-$distributed Stochastic Neighbor Embedding) ალგორითმი, რომლითაც მეზობლების ჩაშენება სტოქასტიკურად ხდება $t-$განაწილებით, კვადრატულია, დაკვირვებათა რაოდენობა ამ მონაცემებზე ძალზე დიდი იქნება, რაც ეწინააღმდეგება მთავარ კომპონენტთა ანალიზს (ინგლ. PCA, Principal Component Analysis) – მონაცემთა განზომილების შემცირების ერთ-ერთ ძირითად მეთოდს. მაგრამ თავდაპირველი გაცნობითი ხასიათის კვლევის ჩასატარებლად იგი მაინც მისაღებია.

11.6

1. $10:$ 5 ინფორმაციული ცვლადი + მათი 5 ასლი.

2. 5.

11.7

1. განმარტების თანახმად, $\lambda\vec{v}=\Sigma\cdot\vec{v}$, მაშასადამე, $n\lambda\vec{v}=\sum_{i=1}^{n}\phi\left(x^i\right)\phi\left(x^i\right)^{\mathrm{T}}\vec{v}$ და ამის შემდეგ $\alpha_i=\frac{1}{n\lambda}\phi\left(\vec{x}^i\right)^{\mathrm{T}}\vec{v}$ გამოსახულების მოცემით, მოსალოდნელი დეკომპოზიცია (დაშლა) მიიღება.

2. $\vec{v}$ სიდიდის ჩანაცვლება $\sum_{i=1}^{n}\alpha_i\phi\left(x^i\right)$ გამოსახულებით $\lambda\vec{v}=\Sigma\cdot\vec{v}$ თანაფარდობაში გვაძლევს : $\lambda\sum_{i=1}^{n}\alpha_i\phi\left(\vec{x}^i\right)=\frac{1}{n}\sum_{i=1}^{n}\phi\left(x^i\right)\phi\left(x^i\right)^{\mathrm{T}}\sum_{i=1}^{n}\alpha_k\phi\left(\vec{x}^k\right)$.

K მატრიცის განმარტებით, $K\alpha$ არის n განზომილების ვექტორი, რომლის $l-$ური კოორდინატია $\left(K\alpha\right)_l=\sum_{i=1}^{n}\alpha_i\phi\left(\vec{x}^l\right)^{\mathrm{T}}\phi\left(\vec{x}^i\right)$. მარცხნიდან $\phi\left(\vec{x}^l\right)^{\mathrm{T}}$ გამოსახულებაზე გამრავლებით, მიიღება : $\lambda\sum_{i=1}^{n}\alpha_i\phi\left(\vec{x}^l\right)^{\mathrm{T}}\phi\left(\vec{x}^i\right)=\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{n}\phi\left(\vec{x}^l\right)^{\mathrm{T}}\phi\left(\vec{x}^i\right)\phi\left(\vec{x}^i\right)^{\mathrm{T}}\phi\left(\vec{x}^k\right)\alpha_k=\frac{1}{n}\left(KK\alpha\right)_l$, საიდანაც საბოლოო შედეგი გამომდინარეობს.

3. ამიტომ K მატრიცის სპექტრული დაშლა Σ კოვარიაციული მატრიცის სპექტრული დაშლის ეკვივალენტურია და იგი (K მატრიცის სპექტრული დაშლა) ჩვენ მთავარ კომპონენტებს გვაძლევს ჰილბერტის $\mathcal{H}$ სივრცეში.

4. $\phi\left(\vec{x}\right)$-ის პროექცია მთავარ კომპონენტზე, ესე იგი Σ კოვარიაციული მატრიცის $\vec{v}$ საკუთარი ვექტორი, გამოითვლება $\phi\left(\vec{x}\right)^{\mathrm{T}}\vec{v}=\sum_{i=1}^{n}\alpha_i\phi\left(\vec{x}\right)^{\mathrm{T}}\phi\left(\vec{x}^i\right)=\sum_{i=1}^{n}\alpha_i k\left(\vec{x},\vec{x}^i\right)$ ფორმულით.

11.8

1. $\rho=\left(\frac{2\pi^{p/2}}{p\Gamma\left(p/2\right)}-\frac{2\left(1-\varepsilon\right)^p\pi^{p/2}}{p\Gamma\left(p/2\right)}\right)\Bigg/\left(\frac{2\pi^{p/2}}{p\Gamma\left(p/2\right)}\right)=1-\left(1-\varepsilon\right)^p$.

2. ეს პროპორციული წილი მიისწრაფვის 1-კენ, როცა p მიისწრაფვის უსასრულობისკენ. დიდი განზომილების პირობებში წერტილები სფეროს შიგნით თავმოყრილია მის საზღვარზე ; ამიტომ პრაქტიკულად არ მოიძებნება ისეთი წერტილები, რომლებიც უფრო ახლოსაა განთავსებული კოორდინატთა სათავესთან. ამის გამო ყველა წერტილი იმყოფება ძალიან შორს ერთმანეთისგან და ძნელია სიახლოვის ცნების გამოყენება დასკვნების გასაკეთებლად.

ლექცია 12 კლასტერიზაცია

შინაარსი

როგორ უნდა ხდებოდეს მოუნიშნავი (არამარკირებული, არაეტიკეტირებული, ჭდედაუსმელი) მონაცემების შესწავლა ? რედუქციის (შემცირების) ზომა მათი ვიზუალიზაციის საშუალებას გვაძლევს ; მაგრამ ეგრეთ წოდებული *მონაცემთა დაყოფის* ანუ *კლასტერიზაციის* მეთოდები გაცილებით უფრო შორს წასვლის საშუალებას გვაძლევს. ეს გულისხმობს მონაცემთა დაყოფას *კლასტერებად* წოდებულ ერთგვაროვან ქვეჯგუფებად, რომლებსაც საერთო მახასიათებლები აქვს. ამ ლექციაში ჩვენ დაწვრილებით გავაშუქებთ ინტერესს ამ მეთოდების მიმართ და განვიხილავთ მიდგომების სამ ტიპს კლასტერიზაციისადმი : იერარქიულ კლასტერიზაციას, კლასტერიზაციას ცენტროიდებით და კლასტერიზაციას სიმკვრივით.

მიზნები

- კლასტერიზაციის ალგორითმის დანიშნულების ახსნა.
- კლასტერიზაციის ალგორითმის შედეგის შეფასება.
- იერარქიული კლასტერიზაციის, k – საშუალოთა მეთოდით კლასტერიზაციის, სიმკვრივით კლასტერიზაციის, სპექტრული კლასტერიზაციის და გაუსის ნარევის მოდელით კლასტერიზაციის დანერგვა და გამოყენება.

1 რატომ არის საჭირო ამ მონაცემების დაყოფა

მონაცემთა დაყოფის ალგორითმები მოუნიშნავი მონაცემების კვლევითი ანალიზის ჩატარების საშუალებას იძლევა. მაგალითად, შესაძლებელია მათი გამოყენება თუნდაც შემდეგი ფაქტორების საიდენტიფიკაციოდ (გამოსავლენად) : მომხმარებლების მსგავსი სტილის ქცევით (ეგრეთ წოდებული *ბაზრის სეგმენტაცია*), საზოგადოებრივი გაერთიანებების სოციალურ ქსელში, განმეორებადი მოტივების საფინანსო ოპერაციებში, ერთისა და იმავე ობიექტის პიქსელების გამოსახლებაზე *(გამოსახულების სეგმენტაცია)* ან პაციენტების, რომელთა სენი შეიძლება იყოს ახსნილი ერთისა და იმავე გენეტიკური პროფილით.

ისინი ასევე მონაცემების ვიზუალიზაციის საშუალებასაც იძლევა თითოეული კლასტერის თუნდაც ერთი საჩვენებელი მაგალითის გაანალიზებით.

დაბოლოს, კლასტერიზაციის ალგორითმები საშუალებას იძლევა გავავრცოთ ერთისა და იმავე კლასტერის ყველა დაკვირვებაზე ის თვისებები, რომლებიც ნამდვილად ჭეშმარიტია ამ კლასტერის რომელიმე ერთი ელემენტისთვის.

ეს განსაკუთრებით სასარგებლოა იმ შემთხვევებში, როცა მონაცემთა მონიშვნა (ჭდის მინიჭება, მარკირება) არის მეტისმეტად გაძნელებული ან ძვირადღირებული.

მაგალითი

განვიხილოთ ტექსტურ დოკუმენტთა ნაკრების ანოტირების მაგალითი. თითოეული ამ დოკუმენტისთვის ანოტაციის გაკეთება თემის (თემების) მიხედვით, რომელსაც იგი ეხება, — ძალიან მომქანცველი ამოცანაა.

უფრო მეტიც, ადამიანები, რომლებიც ამით არიან დაკავებული, სავსებით მოსალოდნელია, უშვებენ წინასწარ გულუხვრახ შეცდომებს უყურადღებობის გამო. ამიტომ ნაკლები ხარჯის გაწევას მოითხოვს და, შესაძლოა, უფრო ეფექტურიც კი იყოს ამ დოკუმენტების ავტომატურად დაჯგუფებისათვის თემების მიხედვით კლასტერიზაციის ალგორითმის გამოყენება. ამის შემდეგ საკმარისი იქნება ვთხოვოთ ადამიანს ყოველი კლასტერის თემის დასადგენად ამ კლასტერში განთავსებული მასალიდან მხოლოდ ერთი ან ორი დოკუმენტის წაკითხვა.

2 კლასტერიზაციის ალგორითმის ხარისხის შეფასება

ვინაიდან კლასტერიზაცია არაკონტროლირებადი (მასწავლებლის გარეშე მიმდინარე) პროცესია, აუცილებელი ხდება შეფასების ისეთი კრიტერიუმების შემოღება, რომლებიც არ ეყრდნობა აბსოლუტურ ჭეშმარიტებას (ე.ი. ცნობილ ჭდეებს, მონიშვნებს).

ეს უფრო რთული ამოცანაა, ვიდრე კონტროლირებად სწავლებაში (მასწავლებლის დახმარებით), სადაც მიზანი ბევრად უფრო ნათელია. მაგრამ იგი არ არის შეუძლებელი და არსებობს მონაცემთა დაყოფის ალგორითმის ეფექტურობის მაჩვენებლების ფართო სპექტრი.

ამის შემდეგ მიიჩნევა, რომ მონაცემთა მოუნიშნავი $\mathcal{D} = \left\{ \vec{x}^1, \vec{x}^2, \ldots, \vec{x}^n \right\}$ კრებული (მასივი), რომელიც შეიცავს n წერტილს $\mathcal{X}$ სივრცეში, დავყავით K რაოდენობის $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_K$ კლასტერად. ამასთან ერთად შემოგვაქვს d მანძილი $\mathcal{X}$–ზე და იმ კლასტერის $k(\vec{x})$ ინდექსი, რომელსაც მივაკუთვნეთ $\vec{x}$ წერტილი.

2.1 კლასტერების სტრუქტურა

დაყოფის ალგორითმით მიღებული კლასტერების ხარისხის შესაფასებლად შეიძლება დავეყრდნოთ მსგავსი დაკვირვებების ერთად დაჯგუფების სურვილს. ამრიგად, დაკვირვებები, რომლებიც ერთსა და იმავე კლასტერს მიეკუთვნება, ერთმანეთთან ახლოს იმყოფება, ხოლო არამსგავსი დაკვირვებები უნდა ეკუთვნოდეს სხვადასხვა კლასტერს.

ამ ცნებათა რაოდენობრივად შეფასებისათვის დაგვჭირდება *ცენტროიდი*, (ინგლ. centroid), რომელიც კლასტერის *ბარიცენტრია* (ინგლ. barycenter or barycentre), ე.ი. კლასტერის გეომეტრიული ცენტრი მათემატიკაში და მასათა ცენტრის სინონიმი ფიზიკაში

განსაზღვრება 12.1 (ცენტროიდი და მედოიდი) $\mathcal{C}$ კლასტერის ცენტროიდი ეწოდება წერტილს, რომელიც შემდეგი სახით განისაზღვრება :

$$\vec{\mu}_{\mathcal{C}} = \frac{1}{|\mathcal{C}|} \sum_{\vec{x} \in \mathcal{C}} \vec{x} \, .$$

მედოიდი — ეს *ცენტროიდისადმი* უახლოესი კლასტერის წერტილია (იგი შეიძლება ერთადერთი არც იყოს და ამ შემთხვევაში მისი არჩევა ნებისმიერად ხდება). იგი წარმოადგენს კლასტერს :

$$\vec{m}_{\mathcal{C}} = \arg\min_{\vec{x} \in \mathcal{C}} d\left(\vec{x}, \vec{\mu}_{\mathcal{C}}\right).$$

ამრიგად, მედოიდი (კლასტერულ ანალიზში) არის ობიექტი, რომელიც ეკუთვნის მონაცემთა ნაკრებს ან კლასტერს და განსხვავება (მაგალითად, კოორდინატების მიხედვით) სხვა ობიექტებისგან მონაცემთა ნაკრებში ან კლასტერში მინიმალურია. მედოიდი თავისი არსით ჰგავს ცენტროიდს, მაგრამ მისგან განსხვავებით არის ობიექტი, რომელიც კლასტერს ეკუთვნის და, როგორც წესი, გამოიყენება იმ შემთხვევებში, როცა შეუძლებელია კლასტერის საშუალო კოორდინატების ან მასათა ცენტრის გამოთვლა.

შენიშვნა

მედოიდების ტიპური გამოყენებაა k – მედოიდთა კლასტერიზაციის ალგორითმი, რომელიც ჰგავს k – საშუალოთა ალგორითმს, მაგრამ მისგან განსხვავებით ყოველ იტერაციაზე იძებნება კლასტერების ცენტრები არა როგორც წერტილთა საშუალო, არამედ როგორც წერტილთა მედოიდები. სხვანაირად რომ ვთქვათ, კლასტერის ცენტრი აუცილებლად უნდა იყოს მისი ერთ-ერთი წერტილი.

ის ფაქტი, რომ მსგავსი დაკვირვებები მიეკუთვნება ერთსა და იმავე კლასტერს, შეიძლება იყოს გადმოცემული *ჰომოგენობის* (*ერთგვაროვნობის*, ერთგვარობის, ერთნაირობის, თანაგვარობის) ახალი ცნებით, რაც ნაჩვენებია ნახატზე 12.1 :

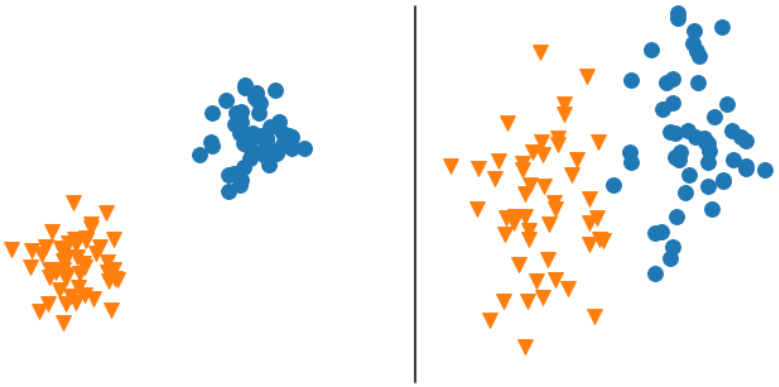
განსაზღვრება 12.2 (ჰომოგენობა, ერთგვაროვნობა) $\mathcal{C}_k$ კლასტერის *ჰომოგენობა* (*ერთგვაროვნობა*), ანუ *tightness* ინგლისურად, ეწოდება ამ კლასტერის დაკვირვებების მისივე ცენტროიდიდან დაშორებათა მანძილების საშუალო მნიშვნელობას :

$$T_k = \frac{1}{|\mathcal{C}_k|} \sum_{\vec{x} \in \mathcal{C}_k} d\left(\vec{x}, \vec{\mu}_k\right).$$

აქ $\vec{\mu}_k$ არის $\mathcal{C}_k$ კლასტერის ცენტროიდი.

$\mathcal{D}$ სიმრავლის (ნაკრების) კლასტერიზაციის *გლობალური ჰომოგენობა* (*საერთო ერთგვაროვნობა*) გამოითვლება როგორც ცალკეულ კლასტერთა ერთგვაროვნობის საშუალო მნიშვნელობა :

$$T = \frac{1}{K} \sum_{k=1}^{K} T_k \,.$$



ნახატი 12.1 – მარცხენა ზოლში ნაჩვენები ორი კლასტერი ერთგვაროვანია და ახლოს მდებარეობს ერთმანეთთან : ისინი შედგენილია კიდევ უფრო მცირე მანძილით დაშორებული წერტილებით. პირიქით, მარჯვენა ზოლში ნაჩვენები ორი კლასტერი ნაკლებ ერთგვაროვანია.

კლასტერების ერთმანეთისგან დაშორების რაოდენობრივი შეფასება შეგვიძლია ნახატზე 12.2 ნაჩვენები *სეპარაბელურობის* (*გამოყოფილობის, განცალკევებულობის, განკერძოებულობის*) კრიტერიუმის გამოყენებით.

განსაზღვრება 12.3 (სეპარაბელურობა, გამოყოფილობა, განცალკევებულობა, განკერძოებულობა) $\mathcal{C}_k$ და $\mathcal{C}_l$ კლასტერების *სეპარაბელურობა* (გამოყოფილობა, განცალკევებულობა, განკერძოებულობა), რასაც ინგლისურად separability ცნება შეესაბამება, — ეს არის მანძილი ამ კლასტერების ცენტროიდებს შორის :

$$S_{kl} = d\left(\vec{\mu}_k, \vec{\mu}_l\right).$$

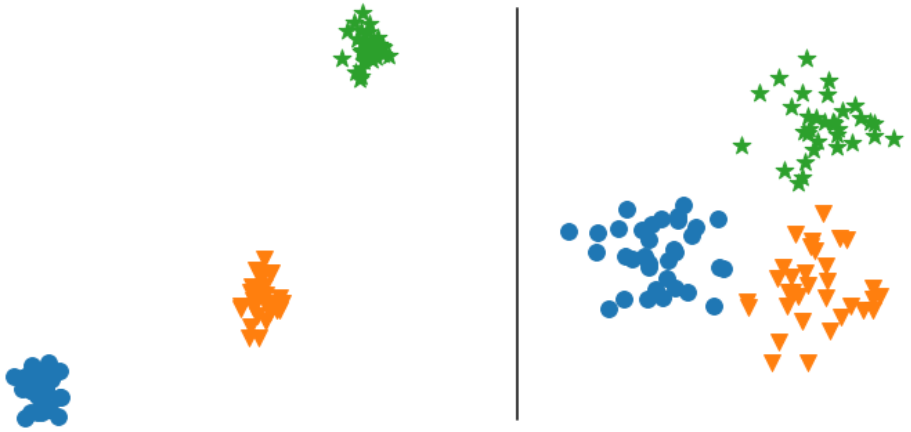
მონაცემთა $\mathcal{D}$ ნაკრების (სიმრავლის) კლასტერიზაციის *გლობალური სეპარაბელურობა (გლობალური გამოყოფილობა)* გამოითვლება როგორც წყვილებში კლასტერთა გამოყოფილობის საშუალო მნიშვნელობა :

$$S = \frac{2}{K\left(K-1\right)} \sum_{k=1}^{K} \sum_{k=k+1}^{K} S_{kl} \,.$$

ნაცვლად იმისა, რომ ცალ-ცალკე განვიხილოთ ორი კრიტერიუმი : სეპარაბელურობის, რომელიც სასურველია იყოს მაღალი და ერთგვაროვნობის, რომელიც სასურველია იყოს დაბალი, შეიძლება მათი ერთმანეთთან შედარება *დევის-ბოლდინის ინდექსის* (*Davies-Bouldin index*) საშუალებით.

განსაზღვრება 12.4 (დევის-ბოლდინის ინდექსი) $\mathcal{C}_k$ კლასტერის *დევის-ბოლდინის ინდექსი* ეწოდება შემდეგ სიდიდეს :

$$\mathcal{D}_k = \max_{l \neq k} \frac{T_k + T_l}{S_{kl}}.$$



ნახატი 12.2 – მარცხენა ზოლში ნაჩვენები სამი კლასტერი კარგად არის განცალკევებული, მაშინ როცა მარჯვენა ზოლში ნაჩვენები კლასტერები ერთმანეთთან ახლოს არის განლაგებული.

$\mathcal{D}$ სიმკვრივის (ნაკრების) კლასტერიზაციის *დევის-ბოლდინის გლობალური* (*საერთო*) *ინდექსი* გამოითვლება როგორც ცალკეულ კლასტერთა დევის-ბოლდინის ინდექსების საშუალო მნიშვნელობა :

$$\mathcal{D} = \frac{1}{\mathcal{C}} \sum_{k=1}^{K} \mathcal{D}_k .$$

განცალკევებულობისა და ერთგვაროვნობის გათვალისწინების კიდევ ერთი ხერხია *სილუეტის კოეფიციენტის* გამოთვლა, რომელიც რაოდენობრივად განსაზღვრავს თითოეული დაკვირვებისათვის მიეკუთვნება იგი თუ არა საჭირო კლასტერს.

განსაზღვრება 12.5 (სილუეტის კოეფიციენტი) $\vec{x} \in \mathcal{D}$ დაკვირვების *სილუეტის კოეფიციენტი* ეწოდება სიდიდეს, რომელიც შემდეგი თანაფარდობით განისაზღვრება :

$$s(\vec{x}) = \frac{b(\vec{x}) - a(\vec{x})}{\max\left(a(\vec{x}), b(\vec{x})\right)},$$

სადაც $a(\vec{x})$ არის $\vec{x}$ დაკვირვების საშუალო მანძილი იმ კლასტერის ყველა დანარჩენი ელემენტიდან, რომელსაც იგი მიეკუთვნება, ხოლო $b(\vec{x})$ არის უმცირესი მნიშვნელობა, რომლის მიღება $a(\vec{x})$ მანძილისთვის გახდებოდა შესაძლებელი, თუ a ერთ-ერთ სხვა კლასტერში იქნებოდა წარმოდგენილი :

$$a(\vec{x}) = \frac{1}{\left|\mathcal{C}_{k(\vec{x})}\right| - 1} \sum_{\vec{u} \in \mathcal{C}_{k(\vec{x})}, \vec{u} \neq \vec{x}} d(\vec{u}, \vec{x}); \quad b(\vec{x}) = \min_{l \neq k(\vec{x})} \frac{1}{\left|\mathcal{C}_l\right|} \sum_{\vec{u} \in \mathcal{C}_l} d(\vec{u}, \vec{x}).$$

კლასტერიზაციის *სილუეტის გლობალური* (*საერთო*) კოეფიციენტი — ეს არის მისი სილუეტის საშუალო კოეფიციენტი :

$$s = \frac{1}{n} \sum_{i=1}^{n} s(\vec{x}^i).$$

$\vec{x}$ დაკვირვების სილუეტის კოეფიციენტი მით უფრო ახლოსაა $1-$თან, რაც უფრო მისაღებია (გამართლებულია) მისი (ე.ი. $\vec{x}-$ის) მიკუთვნება $\mathcal{C}_{k(\vec{x})}$ კლასტერისადმი.

2.2 კლასტერების სტაბილურობა

კიდევ ერთი მნიშვნელოვანი კრიტერიუმია კლასტერების *სტაბილურობა*. არსებითად, ელოდებიან იმავე კლასტერების მიღებას, თუ ხდება რამდენიმე დაკვირვების ამოღება, შეშფოთება ან დაყოფის ალგორითმის განსხვავებული ინიციალიზაცია.

ეს კრიტერიუმი შეიძლება იყოს გამოყენებული ალგორითმის ჰიპერპარამეტრის ასარჩევად : თუ დაყოფის სხვადასხვანაირი ინიციალიზაციის გამოყენებისას ძალიან განსხვავებული კლასტერები მიიღება, ეს შეიძლება მოწმობდეს იმის თაობაზე, რომ ჰიპერპარამეტრები არჩეულია უხეიროდ.

2.3 საექსპერტო ცოდნა

დაბოლოს, ზოგჯერ გვაქვს ნაწილობრივ იმ კლასებით მონიშნული მონაცემების ნაკრები, რომლებიც სასურველი იქნებოდა კლასტერიზაციის საშუალებით დაგვედგინა. ეს შეიძლება იყოს დოკუმენტების ერთობლიობა, რომელთა ქვესიმრავლე მონიშნულია თემის მიხედვით, ან იმ გამოსახულებათა მონაცემების ბაზა, რომელთა ქვესიმრავლე მონიშნულია წარმოდგენილ ობიექტთა შესაბამისად.

ასე რომ შეგვიძლია დაყოფის ალგორითმის მუშაობის შედეგის შეფასება ზუსტად ისევე, როგორც მრავალკლასიანი კლასიფიკაციის ალგორითმის შედეგის. მაგრამ არის ერთი განსხვავება : კლასტერიზაციის დროს მნიშვნელობა არ აქვს k კლასის ობიექტების ყოფნას პირველ, მეორე ან $k-$ურ კლასტერში. ამიტომ ჩვენ აუცილებლად მოგვიწევს შეფასება იმისა, თუ შეესაბამება მონაცემების დაყოფა კლასტერიზაციის ალგორითმით მონაცემების დაყოფას ჭდეთა საშუალებით. სწორედ ამის საშუალებას იძლევა *რენდის ინდექსი* (ინგლ. *Rand Index* ან უბრალოდ *Rand*), რომელიც ამ დასახელებას მასაჩუსეტსის ტექნოლოგიური ინსტიტუტის

ბიოსტატისტიკის ასისტენტ-პროფესორის უილიამ რენდის (William M. Rand, assistant professor of biostatistics, Massachusetts Institute of Technology) პატივსაცემად ატარებს.

ახლა დავუშვათ, რომ $\vec{x}^i$ დაკვირვებები მონიშნულია $y^i \in \{1,2,\ldots,\mathcal{C}\}$ ჭდეებით.

განსაზღვრება 12.6 (რენდის ინდექსი) *რენდის ინდექსი* ეწოდება იმ დაკვირვებათა წყვილების ნაწილს, რომლებიც მიეკუთვნება ან ერთსა და იმავე კლასს და იმყოფება ერთ კლასტერში, ან სხვადასხვა კლასს და იმყოფება ორ სხვადასხვა კლასტერში :

$$\mathrm{RI} = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{l=i+1}^{n} \delta\left(k\left(\vec{x}^i\right) = k\left(\vec{x}^l\right)\right)\ \delta\left(y^i = y^l\right) + \delta\left(k\left(\vec{x}^i\right) \neq k\left(\vec{x}^l\right)\right) \delta\left(y^i \neq y^l\right).$$

ბიოინფორმატიკაში, სადაც ხშირად ძნელია ობიექტების გამოკვლევა ცოდნის უკმარისობის გამო, მიღებულია კლასტერიზაციის შეფასება მისი შედარებით *ონტოლოგიასთან*, ესე იგი ობიექტების (მაგალითად გენების) კლასიფიკაციასთან კატეგორიებით, რომლებიც აღიწერება საერთო ლექსიკონით და ორგანიზებულია იერარქიულად.

კლასტერიზაციის შესაბამისობის შეფასება ონტოლოგიასთან შეიძლება (*გენების ნაკრების*) *გამდიდრების ანალიზით* (ინგლ. gene set enrichment analysis), რომელიც მდგომარეობს იმის შეფასებაში, რამდენად მეტია კლასტერში ობიექტები ონტოლოგიის კატეგორიიდან, ვიდრე ეს მოსალოდნელია შემთხვევითობის პირობებში.

დავუშვათ, რომ მონაცემები შემოდის ჰიპერგეომეტრიული განაწილებიდან და მიზანი მდგომარეობს იმაში, რომ $\mathcal{C}_k$ კლასტერისათვის დადგინდეს $\mathcal{G}$ კატეგორია და $t \in \mathbb{N}$ ზღურბლი :

$$\mathbb{P}\left[\left|\mathcal{G} \bigcap \mathcal{C}_k\right| \geq t\right] = 1 - \sum_{s=1}^{t} \frac{\binom{|\mathcal{G}|}{s}\binom{n-|\mathcal{G}|}{|\mathcal{C}_k|-s}}{\binom{n}{|\mathcal{C}_k|}}. \tag{12.1}$$

მართლაც, წევრი ჯამის ნიშნის ქვეშ არის იმის ალბათობა, რომ, როცა $|\mathcal{C}_k|$ ელემენტის ამოღება ხდება n ელემენტთა შორის, მაშინ s რაოდენობა მიეკუთვნება $\mathcal{G}$ კატეგორიას.

რაკი ჩვენ გავეცანით მონაცემთა დაყოფის ალგორითმების შეფასების სხვადასხვა ხერხს, ახლა შეგვიძლია თავად ალგორითმების გაცნობაც.

ეს ალგორითმები ცდილობს ჰომოგენობის და სეპარაბელურობის იმ კრიტერიუმების ოპტიმიზაციას, რომლებსაც ჩვენ ახლახან შევეხეთ. ვინაიდან ამის ზუსტად განხორციელება შეუძლებელია, ჩვენ უნდა გავაკეთოთ ეს მიახლოებით.

ამ თავში ჩვენ ფოკუსირებას მოვახდენთ კლასტერიზაციის ალგორითმების სამ ძირითად ოჯახზე. ესენია : იერარქიული კლასტერიზაცია, ცენტროიდული კლასტერიზაცია და კლასტერიზაცია სიმკვრივით.

3 იერარქიული კლასტერიზაცია

იერარქიული კლასტერიზაცია აყალიბებს ცალკეულ კლასტერებს რეკურსიის გზით : ეს გულისხმობს მონაცემთა დაყოფას ამ დაყოფის ზომის ყველა შესაძლო მასშტაბისთვის მრავალდონიან იერარქიაში.

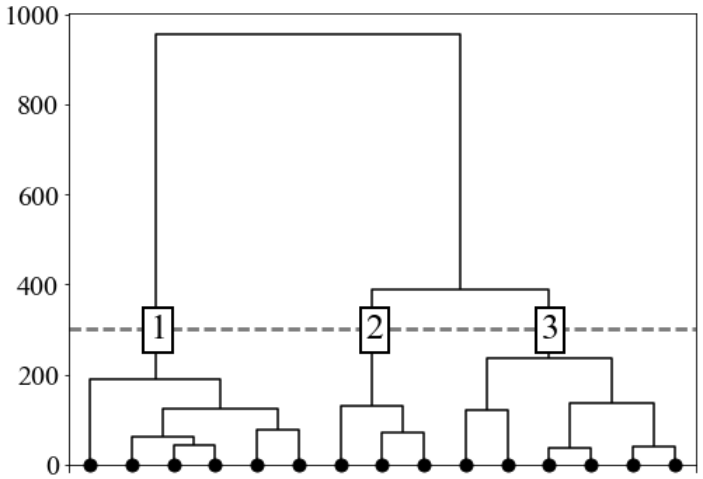
3.1 დენდროგრამა

იერარქიული კლასტერიზაციის შედეგი შეიძლება იყოს წარმოდგენილი *დენდროგრამის* სახით. იგი წარმოადგენს ხეს n ფოთლით და მათ შორის თითოეული შეესაბამება დაკვირვებას. ხის ყოველი კვანძი კი ასახავს ერთ კლასტერს :

- ფესვი — ეს ყველა დაკვირვების შემცველი კლასტერია ;
- თითოეული ფოთოლი — ეს ერთი დაკვირვების შემცველი კლასტერია ;
- კლასტერები ერთითა და იმავე მშობლით ჯგუფდება ერთ კლასტერში ზედა დონეზე ;
- კლასტერი იყოფა თავის შვილებად ქვედა დონეზე. ამიტომ ჩვენ ყველაზე უფრო გვაინტერესებს შუალედური კვანძები.

დაბოლოს, ხის შტოს სიგრძე მანძილის პროპორციულია იმ ორ კლასტერს შორის, რომლებსაც ეს შტო აკავშირებს.

ნახატზე 12.3 ნაჩვენებია დენდროგრამის მაგალითი. თუ n რიცხვი ძალიან დიდია ხის წარმოსადგენად მთლიანობაში, ამ ხეს, ჩვეულებრივ, ჭრიან და წარმოადგენენ მხოლოდ ფესვურ ნაწილს არჩეულ დონეზე.



ნახატი 12.3 - დენდროგრამის მაგალითი. პუნქტირული ხაზის დონეზე გაჭრის გზით სამი კლასტერი მიიღება. აბსცისათა ღერძზე ხის ყოველი ფოთოლო შეესაბამება ერთ დაკვირვებას.

იერარქიული კლასტერიზაციის ალგორითმის შედეგების ვიზუალურად ასახვის სიმარტივე ნიშნავს, რომ ეს ალგორითმი ფართოდ გამოიყენება მრავალ სფეროში, მაგალითად, ბიოინფორმატიკაში.

3.2 გამაერთიანებელი ან გამყოფი აგება

იერარქიული კლასტერიზაცია შეიძლება იყოს *შემკრები* ან *გამყოფი*.

გამაერთიანებელი (აგლომერაციული) კლასტერიზაცია (*ინგლ. bottom-up clustering*), ანუ *კლასტერიზაცია ქვემოდან ზევით*, იწყება დენდროგრამის ფოთლების განხილვით : თავდაპირველად თითოეული დაკვირვება ქმნის კლასტერს $1-$ის ტოლი ზომით. ალგორითმის ყოველ იტერაციაზე ხდება ორი უახლოესი კლასტერის დადგენა და მათი გაერთიანება ერთ კლასტერად ვიდრე ნაპოვნი არ იქნება ყველა n დაკვირვების შემცველი ერთადერთი კლასტერი.

გამყოფი კლასტერიზაცია (*ინგლ. top-down clustering*), ანუ *კლასტერიზაცია ზემოდან ქვევით*, — ეს საწინააღმდეგო მიდგომაა. თავდაპირველად განიხილება ერთი კლასტერი, ყველა დაკვირვების შემცველი დენდროგრამის ფესვი. ყოველ იტერაციაზე კლასტერი ორზე იყოფა, ვიდრე თითოეული კლასტერი მხოლოდ ერთადერთი დაკვირვების შემცველი გახდება.

შემდგომ ჩვენ ყურადღებას გამაერთიანებელ (აგლომერაციულ) მიდგომაზე გავამახვილებთ.

3.3 ბმის ფუნქციები

ორი უახლოესი კლასტერის განსაზღვრა მათი გაერთიანების მიზნით მოითხოვს კლასტერებს შორის მანძილის დადგენას; კლასტერიზაციაში მიღებული ტერმინებით იგი ცნობილია როგორც *ბმის ფუნქცია* (ინგლ. *linkage function*). არსებობს რამდენიმე შესაძლო მიდგომა, თითოეული მათგანი ეფუძნება d მანძილს $\mathcal{X}$ სივრცეში.

უპირველეს ყოვლისა შესაძლებელია ორი კლასტერის გაერთიანების (აგლომერაციის) არჩევა, თუ მათი ორი ელემენტი ახლოსაა ერთმანეთთან. ასეთ შემთხვევაში გამოიყენება *განმარტოებული (განმხოლოებული) ბმა* (ინგლ. *single linkage*).

განსაზღვრება 12.7 (განმარტოებული ბმა) *განმარტოებული ბმა* (ინგლ. *single linkage*) ეწოდება ორ კლასტერს შორის მანძილს, რომელიც განისაზღვრება შემდეგი თანაფარდობით :

$$d_{\text{single}}\left(\mathcal{C}_k, \mathcal{C}_l\right) = \min_{(\vec{u},\vec{v})\in\mathcal{C}_k\times\mathcal{C}_l} d\left(\vec{u},\vec{v}\right).$$

✿

ჩვენ ასევე შეგვიძლია ორი კლასტერის გაერთიანების (აგლომერაციის) არჩევა, თუ მათი *ყველა* ელემენტი ახლოსაა ერთმანეთთან. ასეთ შემთხვევაში გამოიყენება *სრული* (ზოგჯერ იხმარება სიტყვა *აბსოლუტური*) *ბმა*, რომელიც წარმოადგენს მაქსიმალურ მანძილს პირველი კლასტერის ელემენტსა და მეორე კლასტერის ელემენტს შორის.

განსაზღვრება 12.8 (სრული ბმა) *სრული ბმა* (ინგლ. *complete linkage* ან *absolute linkage*), ანუ *ყველაზე უფრო დაშორებულ მეზობელთა მეთოდი* ეწოდება შემდეგი თანაფარდობით განსაზღვრულ მანძილს ორ კლასტერს შორის :

$$d_{\text{complete}}\left(\mathcal{C}_k, \mathcal{C}_l\right) = \max_{(\vec{u},\vec{v})\in\mathcal{C}_k\times\mathcal{C}_l} d\left(\vec{u},\vec{v}\right).$$

✿

შუალედური მიდგომა მდგომარეობს *საშუალო მანძილის* (*the average distance*) განხილვაში პირველ კლასტერის ელემენტსა და მეორე კლასტერის ელემენტს შორის. სწორედ ეს არის *საშუალო ბმა* (ინგლ. *average linkage* ან *group average*).

განსაზღვრება 12.9 (საშუალო ბმა) *საშუალო ბმა* (ინგლ. *average linkage*) ეწოდება ორ კლასტერს შორის მანძილს, რომელიც განისაზღვრება თანაფარდობით :

$$d_{\text{average}}\left(\mathcal{C}_k, \mathcal{C}_l\right) = \frac{1}{|\mathcal{C}_k|}\frac{1}{|\mathcal{C}_l|}\sum_{\vec{u}\in\mathcal{C}_k}\sum_{\vec{v}\in\mathcal{C}_l} d\left(\vec{u},\vec{v}\right) \ .$$

ამ მანძილისთვის ზოგჯერ ასევე ხმარობენ *UPGMA* დასახელებას, რაც წარმოადგენს აკრონიმს (აბრევიატურის ნაირსახეობას, რომელიც შექმნილია *Unweighted Pair-Group Method using arithmetic Averages* ინგლისური ფრაზის სიტყვათა საწყისი ბგერებით) და ნიშნავს შეუწონელ წყვილთა ჯგუფების მეთოდს საშუალო არითმეტიკულის გამოყენებით. ხსენებული აბრევიატურა შემოტანილია სნიტისა და სოკელის მიერ 1973 წელს მათ წიგნში : Sneath, P.H.A. and Sokal, R.R., Numerical Taxonomy: The Principles and Practice of Numerical Classification, San Francisco, W.H. Freeman & Co. Ltd (division of Macmillan Publishers), 573 pages.

საშუალო ბმის ალტერნატივა არის *ცენტროიდული ბმა*, რომელიც ითვალისწინებს მანძილს კლასტერთა ცენტროიდებს შორის.

განსაზღვრება 12.10 (ცენტროიდული კავშირი) *ცენტროიდული კავშირი* (ინგლ. *centroid linkage*) ეწოდება ორ კლასტერს შორის მანძილს, რომელიც განისაზღვრება შემდეგი თანაფარდობით :

$$d_{\text{centroid}}\left(\mathcal{C}_k, \mathcal{C}_l\right) = d\left(\frac{1}{|\mathcal{C}_k|}\sum_{\vec{u}\in\mathcal{C}_k}\vec{u}, \frac{1}{|\mathcal{C}_l|}\sum_{\vec{v}\in\mathcal{C}_l}\vec{v}\right) .$$

ზოგჯერ ამ მანძილს ასევე ეწოდება *შეუწონელ წყვილთა ჯგუფების მეთოდიც ცენტროიდის გამოყენებით* და აღინიშნება *UPGMC* აკრონიმით, გამომდინარე ინგლისური ფრაზიდან — *Unweighted Paired Group Method with Centroid*.

ზემოთ მოცემული ბმის ფუნქციები გამოიყენება კლასტერების *სეპარაბელურობის* უზრუნველსაყოფად. და პირიქით, შეიძლება მიზანი იყოს კლასტერთა *ჰომოგენურობის* (*ერთგვაროვნობის*) მაქსიმიზაციის მიღწევა : ეს გულისხმობს $T_k = \frac{1}{|\mathcal{C}_k|}\sum_{\vec{x}\in\mathcal{C}_k} d\left(\vec{x},\vec{\mu}_k\right)$ გამოსახულების მინიმიზაციას. თუ d — ევკლიდეს მანძილია, მაშინ $T_k = \frac{1}{|\mathcal{C}_k|}\sum_{\vec{x}\in\mathcal{C}_k}\left\|\vec{x}-\vec{\mu}\right\|_2$.

უორდის (Joe H. Ward, Jr.) კლასტერიზაცია, რომელიც პირველად იყო აღწერილი ხსენებული მეცნიერის მიერ 1963 წლის მარტში მის ნაშრომში «Hierarchical Grouping to Optimize an Objective Function» (Journal of the American Statistical Association, Vol. 58, No. 301, pp. 236–244), იყენებს ანალოგიურ ფორმულირებას : მიზანია ორი კლასტერის ისეთნაირად გაერთიანება, რომ მოხდეს შედეგის შიდაკლასტერული დისპერსიის მინიმიზაცია.

განსაზღვრება 12.11 (ინერცია) $\mathcal{C}$ კლასტერის *შიდაკლასტერული დისპერსია*, ანუ უბრალოდ *ინერცია* ეწოდება შემდეგ სიდიდეს :

$$\mathrm{Var}_{\mathrm{in}}\left(\mathcal{C}\right)=\frac{1}{\left|\mathcal{C}\right|}\sum_{\vec{x}\in\mathcal{C}}\left\|\vec{x}-\vec{\mu}\right\|_2^2.$$

წერტილთა $\mathcal{D}$ ღრუბლის (მასივის) კლასტერული სისტემის *გლობალური* (*საერთო*) *ინერცია*, ამრიგად, მოიცემა კლასტერების ინერციათა ჯამით :

$$\mathrm{V}=\sum_{k=1}^{K}\frac{1}{\left|\mathcal{C}_k\right|}\sum_{\vec{x}\in\mathcal{C}_k}\left\|\vec{x}-\vec{\mu}_k\right\|_2^2.$$

3.4 კლასტერების რაოდენობის არჩევა

იერარქიული კლასტერიზაციის უპირატესობა მდგომარეობს იმაში, რომ საჭირო არ არის წინასწარ კლასტერების რაოდენობის დადგენა, ამიტომ შესაძლებელია ყველა შესაძლებლობათა გამოკვლევა დენდროგრამის გასწვრივ. მაგრამ ეს გადაწყვეტილება, როგორც წესი, უნდა იყოს მიღებული გარკვეულ ეტაპზე.

დენდროგრამა შეიძლება იყოს გამოყენებული იმ «დონის» გამოსარკვევად, რომელზეც კლასტერები მკვეთრად არის დისტანცირებული (დაშორებული) ერთმანეთისგან. ყოველთვის უნდა გვახსოვდეს, რომ შტოს სიგრძე ორი კლასტერის ერთმანეთისგან გამომყოფი მანძილის პროპორციულია.

ალტერნატიული გადაწყვეტილებაა (ეფექტურობის ზომის საშუალებით, ისეთის, როგორიცაა სილუეტის კოეფიციენტი) სხვადასხვა ნაპოვნ დანაყოფთა — ე.ი. დენდროგრამის სხვადასხვა კვანძების — შეფასება.

3.5 ალგორითმული სირთულე

იერარქიული კლასტერიზაციის ალგორითმული სირთულე მაღალია : ყოველ იტერაციაზე იმის გადასაწყვეტად, თუ რომელი კლასტერები გადავაჯგუფოთ, ჩვენ მოგვიწევს ორ-ორი მანძილის გამოთვლა მონაცემთა ნაკრების დაკვირვებათა ყველა წყვილს შორის, $\mathcal{O}(pn^2)$ სირთულის პირობებში. ალტერნატივას წარმოადგენს ამ მანძილების შენახვა მეხსიერებაში, რათა შესაძლებელი იყოს მათი ხელმეორედ გამოყენება, რასაც კვადრატული ხასიათის სირთულე აქვს დაკვირვებათა რაოდენობის მიხედვით.

ამიტომ იერარქიული კლასტერიზაცია უფრო ადაპტირებულია მონაცემთა ისეთ ნაკრებთან, რომელიც ნიმუშების ძალიან დიდ რაოდენობას არ შეიცავს.

4 k-საშუალოთა მეთოდი

სხვადასხვა მასშტაბში ყველა შესაძლო დაყოფის გამოკვლევის ნაცვლად შეიძლება უფრო მიზანშეწონილი აღმოჩნდეს კლასტერების K რიცხვის მოცემა, რაც ამცირებს გამოთვლათა დროს.

ზემოთ ჩვენ ვნახეთ, რომ, უორდის (Joe H. Ward, Jr.) კლასტერიზაცია მიზნად ისახავს შიდაკლასტერული დისპერსიის მინიმიზაციას ერთგვაროვანი კლასტერების შესაქმნელად. ეგრეთ წოდებული $k-$საშუალოთა მეთოდი, რომელიც в 1957 წელს იყო შემოთავაზებული

ჰუგო შტაინჰაუსის (Hugo Steinhaus) მიერ, ცდილობს ამ პრობლემის გადაჭრას კლასტერების ფიქსირებული რიცხვისათვის. მიზანი იმაში მდგომარეობს, რომ ნაპოვნი იქნეს დაკვირვებათა ისეთი განაწილება K კლასტერთა მიხედვით, რომელიც მინიმუმს ანიჭებს შიდაკლასტერულ დისპერსიას :

$$\arg \min_{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_K} \sum_{k=1}^{K} \sum_{\vec{x} \in \mathcal{C}_k} \left\| \vec{x} - \vec{\mu}_k \right\|_2^2 . \tag{12.2}$$

მაგრამ ამ ამოცანის ზუსტი ამონახსნის მიღება შეუძლებელია. ამიტომ ჩვენ ვიყენებთ ევრისტიკას — *ლოიდის ალგორითმს*, რომელიც 1982 წელს აღმოაჩინა სტიუარტ ლოიდმა (Stuart Lloyd).

4.1 ლოიდის ალგორითმი

განსაზღვრება 12.12 (ლოიდის ალგორითმი) $\mathbb{R}^p$-ში n დაკვირვების მოცემით და კლასტერთა K რაოდენობის გათვალისწინებისას, *ლოიდის ალგორითმი* შემდეგნაირად მუშაობს :

1. n დაკვირვებიდან K რაოდენობის $\vec{\mu}_1, \vec{\mu}_2, \ldots, \vec{\mu}_K$ დაკვირვებათა არჩევა, რომლებიც საწყის ცენტროიდებად იქნება გამოყენებული ;

2. თითოეული $\vec{x}^i \in \mathcal{D}$ დაკვირვების მიწერა ცენტროიდისადმი, რომელთანაც იგი ყველაზე ახლოს არის :

$$k\left(\vec{x}^i\right) = \arg \min_{k=1,\ldots,K} \left\| \vec{x}^i - \vec{\mu}_k \right\|_2 ;$$

3. თითოეულ კლასტერთა ცენტროიდების გადაანგარიშება :

$$\vec{\mu}_k = \frac{1}{\left|\mathcal{C}_k\right|} \sum_{\vec{x}^i \in \mathcal{C}_k} \vec{x}^i ;$$

4. მე-2 — მე-3 ოპერაციათა გამეორება კრებადობის მიღწევამდე, ესე იგი მიზნობრივ დანიშნულებათა ცვლილების შეწყვეტამდე.

ლოიდის ალგორითმი ახორციელებს *ხარბ სტრატეგიას* და, მიუხედავად იმისა, რომ იგი საკმაოდ სწრაფად იკრიბება, გამორიცხული არ არის მისი მოხვედრა ლოკალური მინიმუმის წერტილზე. ამიტომ, შესაძლოა, ღირდეს რამდენჯერმე მისი გაშვება და მინიმალური შიდაკლასტერული დისპერსიის მქონე ამონახსნის შენახვა, შენარჩუნება.

თუ ჩვენ უნდა განვახორციელოთ ლოიდის ალგორითმის იტერირება t-ჯერ და ვიცით, რომ Kn მანძილის გამოთვლის ღირებულება p განზომილების ვითარებაში $\mathcal{O}(npK)$ რიგის სიდიდეს შეადგენს, მაშინ, ცხადია, ლოიდის ალგორითმის ალგორითმული სირთულეც $\mathcal{O}(npK)$ მნიშვნელობის იქნება. K და t სიდიდეები, ჩვეულებრივ, ძალიან უმნიშვნელოა n პარამეტრთან შედარებით და, ამრიგად, მოცემული ალგორითმი დაკვირვებათა რიცხვის მიხედვით არის *წრფივი*, რაც განასხვავებს მას იერარქიული კლასტერიზაციისგან, რომლის ღი-

რებულება *კვადრატულია* n პარამეტრის მიხედვით : ჩვენ ჩავანაცვლეთ მანძილების გამოანგარიშება $\vec{x}^i$ დაკვირვებიდან დანარჩენ $n-1$ წერტილამდე მონაცემთა ნაკრებში (სიმრავლეში) ამ დაკვირვების მანძილის გამოთვლით K ცენტროიდამდე.

4.2 კლასტერების ფორმა

$k-$საშუალოთა (ინგლ. $k-$means) ალგორითმის ფორმულირების (იხ. განტოლება 12.2) შესაბამისად, ყოველი C_k კლასტერი შედგენილია დაკვირვებებით $\mathcal{D}-$დან, რომლებიც ყველაზე უფრო ახლოს იმყოფება $\vec{\mu}_k$ ცენტროიდთან. ამიტომ ისინი ქმნის *ვორონოის დიაგრამას* (Георгий Феодосьевич Вороной, 1868-1908, იხ. 8.1.2 ქვედანაყოფი). სახელდობრ, ეს ნიშნავს, რომ $k-$საშუალოთა (ინგლ. $k-$means) ალგორითმით ნაპოვნი კლასტერები აუცილებლად ამოზნექილია. როგორც ცნობილია, სიბრტყეზე წერტილთა რომელიღაც სასრული S სიმრავლის ვორონოის დიაგრამა წარმოადგენს სიბრტყის განსაკუთრებულ დაყოფას, რომლის დროსაც ამ დაყოფის არე ქმნის ისეთი წერტილების სიმრავლეს, რომლებიც უფრო ახლოსაა S სიმრავლის ერთ ელემენტთან, ვიდრე სიმრავლის ნებისმიერ სხვა წერტილთან. ცნობილია აგრეთვე ვორონოის მოზაიკისა და *დირიხლეს დაყოფის* (გერმ. Johann Peter Gustav Lejeune Dirichlet, 1805-1859) სახელწოდებითაც

აბერანტული მონაცემები

სტატისტიკაში აბერაცია, ნორმიდან გადახრა, ანომალია, ამოვარდნა (ინგლ. outlier) — ეს არის მნიშვნელობა ან დაკვირვება, რომელიც «დაშორებულია» იმავე მოვლენაზე განხორციელებული სხვა დაკვირვებისგან, უფრო ზუსტად, იგი ძლიერ განსხვავდება გაზომვებისას «ჩვეულებრივ» დანაკვირვები და გაზომილი მნიშვნელობებისგან. ასეთი ამოვარდნა შეიძლება იყოს განპირობებული შესასწავლი მოვლენისათვის დამახასიათებელი ცვალებადობით, არასტაბილურობით ან სულაც წარმოადგენდეს ექსპერიმენტულ შეცდომას. უკანასკნელ შემთხვევაში ზოგჯერ მისი გადაგდება ხდება — ამოღება დაკვირვებათა სტატისტიკიდან.

$k-$საშუალოთა (ინგლ. $k-$means) ალგორითმი მგრძნობიარეა ამოვარდნების მიმართ : მათ უკან მიჰყვება საკუთარი კლასტერის გაჩენა. თუ $\vec{x}^i$ დაკვირვება ძალიან დაშორებულია სხვა დაკვირვებისგან, ასეთ შემთხვევაში იგი აღმოჩნდება თავის საკუთარ კლასტერში, მაშინ როცა დანარჩენ მონაცემთა სიმრავლე დაყოფილი იქნება $K-1$ კლასტერად.

მიუხედავად ამისა, ეს თვისება ძალზე საინტერესოა, ვინაიდან $k-$საშუალოთა ალგორითმი გამოყენების ქმედით ინსტრუმენტს იძლევა სწორედ ამოვარდნების — თავიანთ კლასტერებში განმხოლოებულ, განმარტოებულ დაკვირვებათა — გამოსავლენად.

4.3 ვარიანტები : კლასტერიზაციის k - means ალგორითმის ნაირსახეობები

$k-$საშუალოთა++ მეთოდი

$k-$საშუალოთა (ინგლ. $k-$means) ალგორითმი სტოქასტიკურია : გამოყენებული ინიციალიზაციის ტიპის შესაბამისად შეიძლება სხვადასხვა შედეგი იქნეს მიღებული და ზოგიერთს ამ შედეგიდან შეიძლება გაცილებით უფრო დიდი ინერცია გააჩნდეს, ვიდრე ოპტიმალურ გადაწყვეტილებას.

ამ პრობლემის თავიდან ასაცილებლად $k-$საშუალოთა++ (ინგლ. $k-$means++) ალგორითმი ცენტროიდების ინიციალიზაციას ისეთნაირად იწყებს, რომ ისინი იყოს განაწილებული მონაცემებზე რაც შეიძლება უფრო ფართოდ. უფრო ზუსტად და დაწვრილებით, პროცედურა შემდეგ ორ ბიჯს გულისხმობს :

1. პირველი $\vec{u}^1$ ცენტროიდის შემთხვევითი არჩევა დაკვირვებებს შორის $\mathcal{D}$-დან ;

2. k ინდექსის ყველა $k=2,\ldots,K$ მნიშვნელობისათვის :

- $k-$ური $\vec{u}^k$ ცენტროიდის არჩევა $\mathcal{D}\setminus\vec{u}^k$-დან კანონის შესაბამისად, რომელიც $\vec{u}^{k-1}$-მდე მანძილის კვადრატის პროპორციულია, ე.ი. $\vec{u}^k$ ცენტროიდს ექნება უფრო მეტი შანსი იყოს დაშორებული $\vec{u}^{k-1}$ ცენტროიდისგან.

ეს მიდგომა არ აძლევს $k-$საშუალოთა (ინგლ. $k-$means) მეთოდს დეტერმინირებულ ხასიათს, მაგრამ «უარესი» გადაწყვეტილებების თავიდან აცილებას უწყობს ხელს.

k - საშუალოთა მეთოდი ბირთვით

$k-$საშუალოთა (ინგლ. $k-$means) მეთოდი მოითხოვს, რომ მონაცემები იყოს აღწერილი ევკლიდეს სივრცეში და ამასთან ერთად მას შეუძლია მხოლოდ ამოზნექილი კლასტერების ფორმირება, რაც არასასურველი შეზღუდვა შეიძლება იყოს.

მაგრამ ლოიდის ალგორითმის მიმართ, საბედნიეროდ, შეიძლება *ბირთვული ხრიკის* (ინგლ. *core trick* ან *kernel trick*) გამოყენება (იხ. ქვედანაყოფი 10.3.3).

მტკიცებულება. დავუშვათ, რომ $k:\mathcal{X}\times\mathcal{X}\to\mathbb{R}$ არის ბირთვი, ასევე არსებობს ჰილბერტის (Hilbert) $\mathcal{H}$ სივრცე და ისეთი $\phi:\mathcal{X}\to\mathcal{H}$ აპლიკაცია (არგუმენტისადმი ფუნქციის გამოყენების ოპერაცია), რომ დაკვირვებათა ყველა $\vec{x},\vec{x}'\in\mathcal{X}\times\mathcal{X}$ წყვილისათვის $k\left(\vec{x},\vec{x}'\right)=\left\langle\phi\left(\vec{x}\right),\phi\left(\vec{x}'\right)\right\rangle_{\mathcal{H}}$ თანაფარდობას აქვს ადგილი.

ლოიდის ალგორითმის გამოსაყენებლად $\left\{\phi\left(\vec{x}^1\right),\phi\left(\vec{x}^2\right),\ldots,\phi\left(\vec{x}^n\right)\right\}$ სახეთა მიმართ $\mathcal{D}$ სიმრავლიდან $\mathcal{H}$ სივრცეში, ყოველ იტერაციაზე უნდა გამოითვალოს მანძილი $\phi\left(\vec{x}^i\right)$-დან თითოეულ ცენტროიდამდე $\vec{h}_1,\vec{h}_2\ldots,\vec{h}_K$ სიმრავლიდან K კარდინალური რიცხვით (სიმძლავრით).

$\vec{h}_k$ ცენტროიდის ადგილმდებარეობა (პოზიცია) გამოითვლება როგორც იმ დაკვირვებათა სახეების საშუალო, რომლებიც მიეკუთვნება $\mathcal{C}_k$ კლასტერს :

$$\vec{h}_k\frac{1}{\left|\mathcal{C}_k\right|}\sum_{\vec{x}^i\in\mathcal{C}_k}\phi\left(\vec{x}^i\right).$$

მანძილი $\vec{x}^i$ დაკვირვების სახიდან ცენტროიდამდე გამოითვლება შემდეგი თანაფარდობით :

$$\left\|\phi\left(\vec{x}^i-\vec{h}_k\right)\right\|_2^2=\left\|\phi\left(\vec{x}^i-\frac{1}{|\mathcal{C}_k|}\sum_{\vec{u}\in\mathcal{C}_k}\phi(\vec{u})\right)\right\|_2^2$$

$$=\left\langle\phi\left(\vec{x}^i\right)-\frac{1}{|\mathcal{C}_k|}\sum_{\vec{u}\in\mathcal{C}_k}\phi(\vec{u}),\phi\left(\vec{x}^i\right)-\frac{1}{|\mathcal{C}_k|}\sum_{\vec{u}\in\mathcal{C}_k}\phi(\vec{u})\right\rangle_{\mathcal{H}}$$

$$=k\left(\vec{x}^i,\vec{x}^i\right)-\frac{2}{|\mathcal{C}_k|}\sum_{\vec{u}\in\mathcal{C}_k}k\left(\vec{u},\vec{x}^i\right)+\frac{1}{|\mathcal{C}_k|^2}\sum_{\vec{u}\in\mathcal{C}_k}\sum_{\vec{v}\in\mathcal{C}_k}k(\vec{u},\vec{v}).$$

ასე რომ, ლოიდის ალგორითმის მე-(2) ბიჯი შეიძლება იყოს გადაწერილი ϕ აპლიკაციის გამოუყენებლად და ცენტროიდების ცხადად გადაანგარიშების აუცილებლობის გარეშე მე-(3) ბიჯზე.

□

შენიშვნა

$k-$საშუალოთა ბირთვული ვერსია, ჩვეულებრივ, არ გვაძლევს კლასტერთა ცენტროიდების გაგების საშუალებას, რადგან ისინი ცხოვრობს განმეორებითი აღწერის $\mathcal{H}$ სივრცეში, რომელიც მიუწვდომელია ϕ აპლიკაციის უცოდნელად.

5 კლასტერიზაცია გაუსის ნარევის მოდელით

კლასტერიზაცია გაუსის ნარევის მოდელით მდგომარეობს K გაუსის განაწილებიდან ამოღებული მონაცემების მოდელირებაში : ამრიგად, ეს არის *გენერაციული* მოდელი — იგულისხმება, რომ ჩვენი დაკვირვებები *გაჩენილია, წარმოშობილია* ალბათური კანონით, რომლის ფორმა მოცემულია ჩვენ მიერ და ყველა მომდევნო მოქმედება მიზნად ისახავს მისი პარამეტრების შეფასებას.

განსაზღვრება 12.13 (გაუსის ნარევის მოდელი) დავუშვათ, რომ $\mathbb{R}^p$ სივრციდან მოცემულია $\mathcal{D}=\left\{\vec{x}^1,\vec{x}^2,\ldots,\vec{x}^n\right\}$ წერტილების ღრუბელი, სიმრავლე. მაშინ ჩვენ ვლაპარაკობთ *გაუსის ნარევის მოდელზე*, ანუ GMM-ზე (ინგლისური ფრაზიდან *Gaussian Mixture Model*), როცა $\mathcal{D}$ სიმრავლის მოდელირება ხდება როგორც $p-$განზომილების X შემთხვევითი ვექტორის, რომლის f სიმკვრივე არის *გაუსიანების ნარევი*, სხვა სიტყვებით, გაუსიანების წრფივი კომბინაცია :

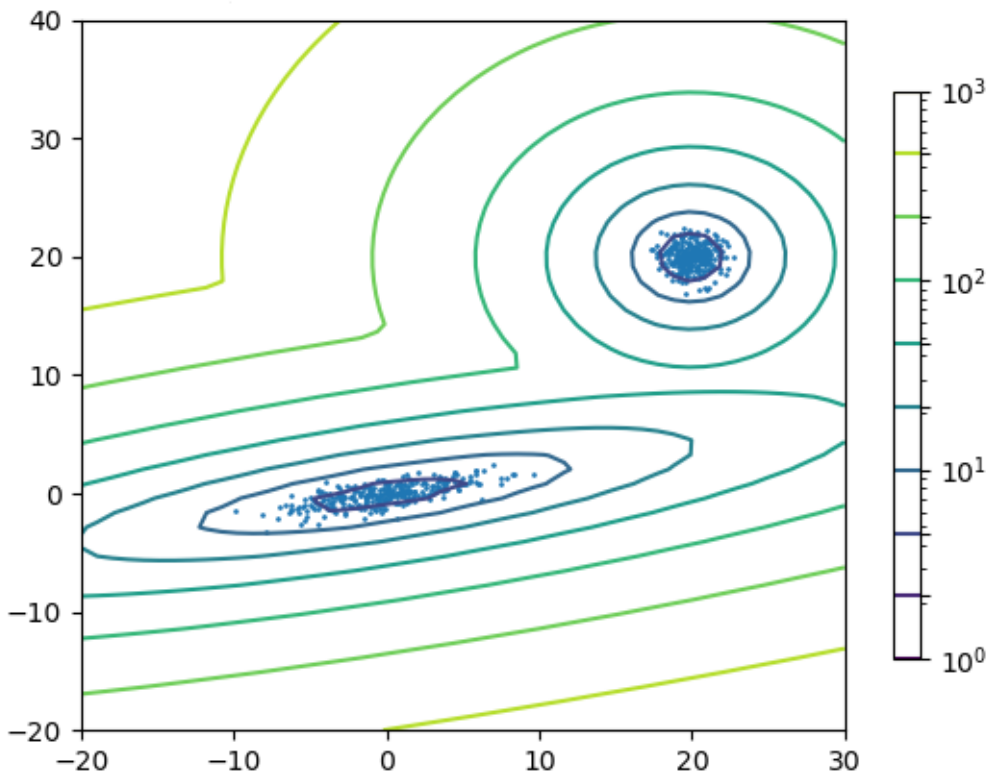
$$f\left(\vec{x}\right)=\sum_{k=1}^{K}\pi_k f_k\left(\vec{x}\right), \tag{12.3}$$

სადაც K არის ნარევის კომპონენტების რაოდენობა, π_k — თითოეული კომპონენტის წილი, $\sum_{k=1}^{K}\pi_k=1$ და f_k — მრავალგანზომილებიანი გაუსიანის სიმკვრივეები.

ამიტომ ვთვლით, რომ არსებობს დისკრეტული ლატენტური (ფარული) Z შემთხვევითი ცვლადი მნიშვნელობებით $\{1,2,\ldots,K\}$ სიმრავლეში და იგი განსაზღვრავს, რომელ კლასტერს მიეკუთვნება დაკვირვება. ასე რომ, $\vec{x}^i$ დაკვირვება არის გაუსიანის რეალიზაცია $\vec{\mu}_k$ მათემატიკური ლოდინით და Σ_k დისპერსიით, სადაც k ცვლადი თავად წარმოადგენს Z შემთხვევითი სიდიდის რეალიზაციას.

განსაზღვრება 12.14 (კლასტერიზაცია გაუსის ნარევის მოდელით) *კლასტერიზაცია გაუსის ნარევის მოდელით* ეწოდება დაყოფის ალგორითმს, რომელიც მდგომარეობს მათემატიკური ლოდინების მაქსიმიზაციის მეთოდით K გაუსიანის ნარევის $\pi_k, \vec{\mu}_k, \Sigma_k$ პარამეტრთა განსაზღვრაში მათი დანაკვირვები მნიშვნელობების განხილვისას სტატისტიკურ ანარჩევად.

ნახატი 12.4 ამ კონცეფციის ილუსტრაციას იძლევა ორგანზომილებიან მონაცემებზე. გაუსის ნარევის $k-$ური კომპონენტის $\vec{\mu}_k$ მათემატიკური ლოდინი C_k კლასტერის ცენტროიდია.



ნახატი 12.4 - კლასტერიზაცია გაუსის ნარევის მოდელით (Gaussian Mixture model).

ნავარაუდევია, რომ მონაცემები (წერტილები) აღებულია ორი გაუსიანის ნარევიდან. დონეთა წირები (მოდელის ერთნაირ ალბათობათა ზედაპირები) წარმოადგენს ამ მოდელის ლოგარითმულ დამაჯერებლობას ; რაც უფრო მუქია ფერი, მით მეტია დამაჯერებლობა.

5.1 მათემატიკური ლოდინის ოპტიმიზაცია

თუ მოცემულია $\{\pi_k, \vec{\mu}_k, \Sigma_k\}_{k=1,\ldots,K}$ პარამეტრები, მაშინ რაღაც $\vec{x}$ დაკვირვების ლოგარითმული

დამაჯერებლობა შემდეგი გამოსახულებით განისაზღვრება :

$$\ln L\left(\vec{x};\left\{\pi_k,\vec{\mu}_k,\Sigma_k\right\}_{k=1,\ldots,K}\right)=\ln\sum_{k=1}^{K}\pi_k\,f_{\mathcal{N}}\left(\vec{x};\vec{\mu}_k,\Sigma_k\right),\tag{12.4}$$

სადაც $f_{\mathcal{N}}\left(\vec{x};\vec{\mu}_k,\Sigma_k\right)$ არის $\vec{\mu}_k\in\mathbb{R}^p$ მათემატიკური ლოდინისა და $\Sigma_k\in\mathbb{R}^{p\times p}$ კოვარიაციის მქონე მრავალგანზომილებიანი გაუსიანის სიმკვრივე $\vec{x}$-ით.

პრაქტიკულად 12.4 დამაჯერებლობის ლოგარითმის მინიმიზაცია ძნელია, რადგან იგი ამოზნექილი არ არის. ამიტომ, ჩვეულებრივ, გამოიყენება ეგრეთ წოდებული *მათემატიკური ლოდინების მაქსიმიზაციის* ალგორითმი (ინგლ. EM algorithm, ანუ an expectation–maximization algorithm), რომელიც მდგომარეობს (E) ბიჯსა და (M) ბიჯს შორის მონაცვლეობაში ამ პროცედურის კრებადობის მიღწევამდე, სადაც (E) არის $z^1,z^2,\ldots,z^n$ ლატენტური (ფარული) ცვლადების აპოსტერიორული განაწილების მიმართ ლოგარითმული დამაჯერებლობის ლოდინის შეფასების ეტაპი, ხოლო (M) — ამ ლოდინის მაქსიმიზაციის ეტაპი.

უფრო ზუსტად რომ ვთქვათ, $\left(\vec{x},z\right)$ წყვილის ლოგარითმული დამაჯერებლობა შემდეგნაირად მოიცემა :

$$\ln L\left(\vec{x},z;\left\{\pi_k,\vec{\mu}_k,\Sigma_k\right\}_{k=1,\ldots,K}\right)=\ln\mathbb{P}\left(Z=z\right)f_{X|Z=z}\left(\vec{x}\right)=\ln\pi_z+\ln f_{\mathcal{N}}\left(\vec{x};\vec{\mu}_z,\Sigma_z\right).\tag{12.5}$$

ჩავთვალოთ, რომ $\mathcal{Z}=\left(z^1,z^2,\ldots,z^n\right)$ ლატენტური (ფარული) ცვლადებია, რომლებიც მონაცემთა $\mathcal{D}$ ნაკრებს (სიმრავლეს) შეესაბამება. მაშინ $\left(\mathcal{D},\mathcal{Z}\right)$ წყვილის 12.5 ლოგარითმული დამაჯერებლობის მათემატიკური ლოდინი ლატენტური ცვლადების $\mathbb{P}\left(Z=z\mid X=\vec{x}\right)$ აპოსტერიორული განაწილების მიმართ შემდეგი სახით ჩაიწერება :

$$\begin{aligned}
&\mathbb{E}\left[\sum_{i=1}^{n}\ln L\left(\vec{x}^i,z^i;\left\{\pi_k,\vec{\mu}_k,\Sigma_k\right\}_{k=1,\ldots,K}\right)\right]=\sum_{i=1}^{n}\mathbb{E}\left[L\left(\vec{x}^i,z^i;\left\{\pi_k,\vec{\mu}_k,\Sigma_k\right\}_{k=1,\ldots,K}\right)\right]=\\
&=\sum_{i=1}^{n}\sum_{k=1}^{K}\mathbb{P}\left(Z=k\mid X=\vec{x}^i\right)\ln L\left(\vec{x}^i,k;\left\{\pi_k,\vec{\mu}_k,\Sigma_k\right\}_{k=1,\ldots,K}\right)=\\
&=\sum_{i=1}^{n}\sum_{k=1}^{K}\mathbb{P}\left(Z=k\mid X=\vec{x}^i\right)\left(\ln\pi_k+\ln f_{\mathcal{N}}\left(\vec{x}^i,\vec{\mu}_k,\Sigma_k\right)\right).
\end{aligned}$$

როგორც ცნობილია, სტატისტიკაში *ლატენტური*, ანუ *ფარული* ცვლადები წარმოადგენს ისეთ ცვლადებს, რომლებიც არ შეიძლება იყოს გაზომილი ცხადი სახით და განისაზღვრება მხოლოდ მათემატიკური მოდელებიდან დაკვირვებადი ცვლადების გამოყენებით.

აპოსტერიორული $\mathbb{P}\left(Z=k\mid X=\vec{x}^i\right)$ განაწილება მოიცემა ბაიესის კანონით :

$$\mathbb{P}\left(Z=k\mid X=\vec{x}^i\right)=\frac{f_{\mathcal{N}}\left(\vec{x}^i;\vec{\mu}_k,\Sigma_k\right)\mathbb{P}\left(Z=k\right)}{f\left(\vec{x}\right)}=\frac{f_{\mathcal{N}}\left(\vec{x}^i;\vec{\mu}_k,\Sigma_k\right)\pi_k}{\sum_{l=1}^{K}\pi_l\,f_{\mathcal{N}}\left(\vec{x}^i;\vec{\mu}_l,\Sigma_l\right)}.$$

ამრიგად, მათემატიკური ლოდინების მაქსიმიზაციის პროცედურა მდგომარეობს :

1. $\pi_k, \vec{\mu}_k, \Sigma_k$ მნიშვნელობათა შემთხვევით ინიციალიზაციაში $\pi_k^{(0)}, \vec{\mu}_k^{(0)}, \Sigma_k^{(0)}$ მნიშვნელობებით ყველა $k = 1, \ldots, K$ სიდიდისათვის.

2. $(t+1)$ იტერაციაში ალგორითმის კრებადობის დასრულებამდე, რაც გულისხმობს :

- (E) : აპოსტერიორული ალბათობების გამოთვლას $k = 1, \ldots, K$ და $i = 1, \ldots, n$ სიდიდეებისათვის ფორმულით

$$\tau_{ki}^{(t)} = \frac{f_{\mathcal{N}}\left(\vec{x}^i; \vec{\mu}_k^{(t)}, \Sigma_k^{(t)}\right)\pi_k^{(t)}}{\sum_{l=1}^{K} \pi_l^{(t)} f_{\mathcal{N}}\left(\vec{x}^i; \vec{\mu}_l^{(t)}, \Sigma_l^{(t)}\right)};$$

- (M) : ყველა $k = 1, \ldots, K$ ინდექსისათვის $\pi_k^{(t+1)}, \vec{\mu}_k^{(t+1)}, \Sigma_k^{(t+1)}$ სიდიდეთა იმ მნიშვნელობების განსაზღვრას, რომლებიც ლოგარითმული დამაჯერებლობის მათემატიკურ ლოდინს მაქსიმუმს ანიჭებს

$$\pi_k^{(t+1)}, \vec{\mu}_k^{(t+1)}, \Sigma_k^{(t+1)} = \arg\max_{\pi_k, \vec{\mu}_k, \Sigma_k} \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ki}^{(t)} \left(\ln \pi_k + \ln\, f_{\mathcal{N}}\left(\vec{x}^i; \mu_k, \Sigma_k\right)\right).$$

(M) ბიჯზე π_k, $\vec{\mu}_k$ და Σ_k სიდიდეები ჩნდება ცალკეული წრფივი წევრების სახით და შეიძლება იყოს განსაზღვრული ერთმანეთისგან დამოუკიდებლად (ცალ-ცალკე). π_k სიდიდისათვის განისაზღვრება ექსპერიმენტთა K რიცხვის შესაბამისი ბინომიალური განაწილების წარმატებათა ალბათობის მაქსიმალური დამაჯერებლობის შეფასება, ხოლო $\vec{\mu}_k$ და Σ_k სიდიდეებისათვის — გაუსიანის პარამეტრთა მაქსიმალური დამაჯერებლობის შეფასება :

$$\left.\begin{aligned} \pi_k^{(t+1)} &= \frac{n_k^{(t)}}{n} \\ \vec{\mu}_k^{(t+1)} &= \frac{1}{n_k^{(t)}} \sum_{i=1}^{n} \tau_{ki}^{(t)} \vec{x}^i \\ \Sigma_k^{(t+1)} &= \frac{1}{n_k^{(t)}} \sum_{i=1}^{n} \tau_{ki}^{(t)} \left(\vec{x}^i - \vec{\mu}_k^{(t+1)}\right)\left(\vec{x}^i - \vec{\mu}_k^{(t+1)}\right)^{\mathrm{T}} \end{aligned}\right\},$$

სადაც $n_k^{(t)} = \sum_{i=1}^{n} \tau_{ki}^{(t)}$ არის $\mathcal{C}_k$ კლასტერის სავარაუდო ზომაა (t) იტერაციაზე.

5.2 კავშირი k-საშუალოთა მეთოდთან

კლასტერიზაცია გაუსის ნარევის მოდელის გამოყენებით შეიძლება იყოს განხილული როგორც $k-$საშუალოთა (ინგლ. $k-$means) მეთოდის განზოგადება.

თეორემა 12.1 (გაუსიანთა ნარევისა და k-საშუალოთა ეკვივალენტობა) თუ განიხილება იგივურად ტოლი ერთნაირი კოვარიაციული მატრიცების მქონე გაუსიანები და მათი ნარევი

მდგენელთა ტოლი წილებით, ანუ პროპორციებით (როცა $\pi_k = \frac{1}{K}$ ყველა $k = 1, \ldots, K$ მნიშვნელობისთვის), მაშინ კლასტერიზაცია ასეთი გაუსიანების ნარევის მოდელის გამოყენებით k–საშუალოთა მეთოდის ეკვივალენტურია.

მტკიცებულება $(\mathcal{D}, 2)$-ის ლოგარითმული დამაჯერებლობა მოიცემა თანაფარდობით

$$\ln L\left(\mathcal{D}, \mathcal{Z}; \{\pi_k, \vec{\mu}_k, \Sigma_k\}_{k=1,\ldots,K}\right) = \sum_{i=1}^{n} \ln \frac{1}{K} f_N\left(\vec{x}^i; \vec{\mu}_{z^i}, I_p\right) =$$
$$= \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{I}_{z^i=k} \ln \frac{1}{K} f_N\left(\vec{x}^i; \vec{\mu}_{z^i}, I_p\right).$$

$f_N\left(\vec{x}^i; \vec{\mu}_{z^i}, I_p\right)$ გამოსახულების ჩანაცვლებისას მისი შესაბამისი მნიშვნელობით, ლოგარითმული დამაჯერებლობის მაქსიმიზაცია, ამრიგად, შემდეგი თანაფარდობის მინიმიზაციის ეკვივალენტური ხდება :

$$\sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{I}_{z^i=k} \left(\vec{x}^i - \vec{\mu}_{z^i}\right)^{\mathrm{T}} \left(\vec{x}^i - \vec{\mu}_{z^i}\right),$$

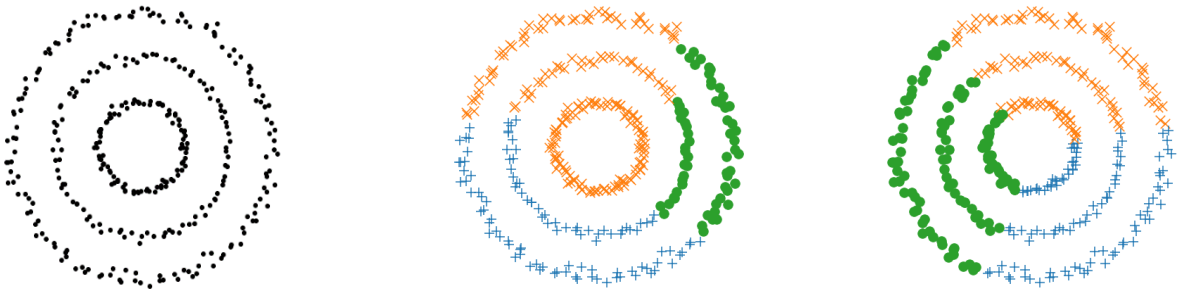
საიდანაც k–საშუალოთა მეთოდის კრიტერიუმი მოიპოვება (განტოლება 12.2).

□

გამოთვლათა ხანგრძლივობის შესამცირებლად შესაფასებელ პარამეტრთა რაოდენობის შემცირების გზით ისეთი ძლიერი და ხისტი დაშვებების გამოუყენებლად, როგორც ეს k–საშუალოთა მეთოდისთვის არის დამახასიათებელი, ზოგჯერ ვარაუდობენ მხოლოდ იმას, რომ სხვადასხვა კოვარიაციული მატრიცა დიაგონალურია, მაგრამ იდენტურობის პირობა არ მოითხოვება.

6 კლასტერიზაცია სიმკვრივის მიხედვით

არაამოზნექილი კლასტერების მიღების კიდევ ერთი მეთოდია კლასტერიზაცია სიმკვრივით. ავიღოთ ნახატზე 12.5 ნაჩვენები მაგალითი : თითქოს ბუნებრივია მონაცემების გაყოფა სამ კონცენტრულ წრედ, რისი გაკეთება შეუძლებელია როგორც აგლომერაციული, ასევე k–საშუალოთა მეთოდისთვის, ვინაიდან ეს კლასტერები არ არის ამოზნექილი.



(*a*) ბუნებრივი ჩანს ამ მონაცემების დაყოფა სამ კონცენტრულ წრედ.

(*b*) დაყოფა სამ კლასტერად აგლომერაციული კლასტერიზაციით (საშუალო ბმა).

(c) დაყოფა სამ კლასტერად k – საშუალოთა მეშვეობით.

ნახატი 12.5 - მოტივაცია კლასტერიზაციისათვის სიმკვრივის მიხედვით.

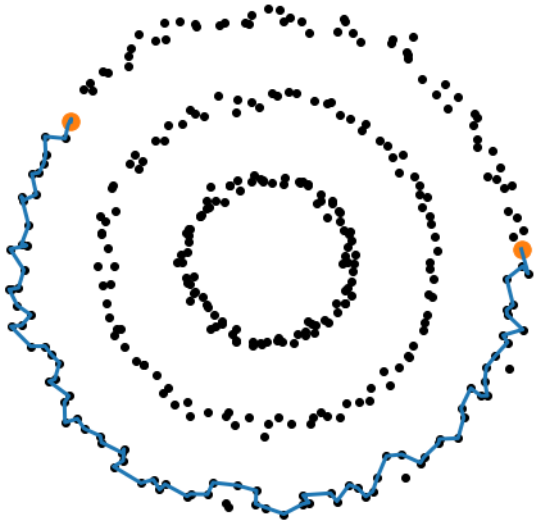
სწორედ ამ პრობლემის გადაჭრას ცდილობს კლასტერიზაცია სიმკვრივით, როცა შენიშნავს, რომ მოცემულ წერტილთან უახლოესი წერტილები იმყოფება ერთსა და იმავე წრეწირზე და არა სხვაზე. იდეა იმაში მდგომარეობს, რომ მოვახდინოთ ისეთ დაკვირვებათა კლასტერების ფორმირება, რომლებიც ახლოსაა ერთმანეთთან იმ თვალსაზრისით, რომ, თუ ერთისა და იმავე $\mathcal{C}_k$ კლასტერის ორი, $\vec{x}^i$ და $\vec{x}^l$, ელემენტი შეიძლება იყოს დაშორებული ერთმანეთისგან, მაშინ არსებობს $\vec{u}^1, \vec{u}^2, \ldots, \vec{u}^m \in \mathcal{C}_k$ ელემენტების ისეთი თანამიმდევრობა, რომ $\vec{u}^1$ არის $\vec{x}^i$-ის ახლობლად, $\vec{u}^m$ არის $\vec{x}^l$-ის ახლობლად და ნებისმიერი $t \in \{1, 2, \ldots, m\}$-თვის $\vec{u}^t$ არის $\vec{u}^{t-1}$-ის ახლობლად. ეს იდეა ასახულია ნახატზე 12.6.

ამ იდეის ფორმალიზებისათვის ჩვენ რამდენიმე განსაზღვრება დაგვჭირდება.

განსაზღვრება 12.15 (ϵ – მიდამო) დავუშვათ, რომ $\mathcal{D} = \{\vec{x}^1, \vec{x}^2, \ldots, \vec{x}^n\}$ არის $\mathcal{X}$ – ის ჯგუფებად დასაყოფ ელემენტთა ნაკრები (სიმრავლე), d — მანძილი $\mathcal{X}$ – ზე, ხოლო $\epsilon > 0$. დაკვირვებათა სიმრავლეს $\mathcal{D}$ – დან, რომელთა მანძილი $\vec{x}$ – მდე ნაკლებია ϵ – ზე, ეწოდება $\vec{x} \in \mathcal{X}$ ელემენტის ϵ – *მიდამო* (სამეზობლო) :

$$\mathcal{N}_\epsilon(\vec{x}) = \{\vec{u} \in \mathcal{D} : d(\vec{x}, \vec{u}) < \epsilon\}.$$

განსაზღვრება 12.16 (შიდა წერტილი) დავუშვათ, რომ მოცემულია $n_{\min} \in \mathbb{N}$ რიცხვი. მაშინ ამბობენ, რომ $\vec{x} \in \mathcal{X}$ ელემენტი არის *შიდა წერტილი*, ანუ, ინგლისურად, *core point* (*ბირთვის წერტილი*), თუ ამ წერტილის ϵ – მიდამო შეიცავს $n_{\min}$ – ზე არანაკლებ $\mathcal{N}_\epsilon(\vec{x}) \geq n_{\min}$ ელემენტს.



ნახატი 12.6 - უახლოეს მეზობლებს შორის არსებობს გზა, რომელიც ერთი წერტილიდან მეორე წერტილისკენ გადაადგილების საშუალებას იძლევა იმავე კლასტერში.

განსაზღვრება 12.17 (ბმა სიმკვრივით) ამბობენ, რომ $\vec{x}$ და $\vec{v} \in \mathcal{X}$ *ბმულია სიმკვრივით*, თუ ისინი შეიძლება იყოს «შეერთებული» («გადაბმული», «შეჭიდული») $\epsilon-$ მიდამოთა თანამიმდევრობით, რომელთა შორის ნებისმიერი შეიცავს სულ ცოტა $n_{\min}$ ელემენტს. ფორმალურად ეს ნიშნავს, რომ არსებობს $\mathcal{D}$ სიმრავლის ელემენტთა $m \in \mathbb{N}$ რიცხვი და $\vec{u}^1, \vec{u}^2, \ldots, \vec{u}^m$ ისეთი თანამიმდევრობა, რომ $\vec{u}^1$ არის შიდა წერტილი $\mathcal{N}_\epsilon\left(\vec{x}\right)-$ დან, $\vec{u}^2$ არის შიდა წერტილი $\mathcal{N}_\epsilon\left(\vec{u}^1\right)-$ დან, $\vec{u}^m$ არის შიდა წერტილი $\mathcal{N}_\epsilon\left(\vec{u}^{m-1}\right)-$ დან, $\vec{v}$ არის შიდა წერტილი $\mathcal{N}_\epsilon\left(\vec{u}^m\right)-$ დან.

DBSCAN ალგორითმი (ინგლ. Density-Based Spatial Clustering of Applications with Noise — სიმკვრივეზე დაფუძნებული სივრცითი კლასტერიზაცია ხმაურების შემცველი გამოყენებებისათვის), რომელიც შემოთავაზებული იყო 1966 წელს მარტინ ესტერის (Martin Ester), ჰანს-პეტერ კრიგელის (Hans-Peter Kriegel), იორგ სანდერის (Jörg Sander) და სიაოვეი სიუის (Xiaowei Xu) მიერ (Ester et al., 1996), ახორციელებს მონაცემთა დაყოფას იმ წერტილთა კლასტერების შექმნით, რომლებიც შეიძლება იყოს დაშორებული ერთმანეთისგან სიმკვრივის მიხედვით. ეს დაყოფის სწორედ ის პოპულარული ალგორითმია, რომელსაც 2014 წელს მიღებული აქვს KDD (Knowledge Discovery and Data Mining — ცოდნის მოპოვება და მონაცემების ამოღება) პრესტიჟული საერთაშორისო კონფერენციის «*Test of time award*» ჯილდო მეცნიერული წვლილისთვის, რომელმაც გაუძლო დროის გამოცდას.

განსაზღვრება 12.18 (DBSCAN) *DBSCAN* (ინგლ. *Density-Based Spatial Clustering of Applications with Noise*) სახელწოდებას მონაცემთა დაყოფის შემდეგი პროცედურა ატარებს :

- ინიციალიზაცია : მონახულებული ელემენტების კრებული (სიმრავლე) $\mathcal{V}=0$, კლასტერების სია $\mathcal{K}=0$, აბერანტულ (ამოვარდნილ, მნიშვნელოვნად გადახრილ, ანომალიურ) დაკვირვებათა სია $\mathcal{A}=0$.
- ყოველი $\vec{x} \in \mathcal{D} \setminus \mathcal{V}$ ელემენტისთვის :
  1. $\mathcal{N}_\epsilon\left(\vec{x}\right)$-ის აგება.
  2. თუ $\mathcal{N}_\epsilon\left(\vec{x}\right) < n_{\min}$, მაშინ $\vec{x}$ ელემენტი ჩაითვალოს (დროებით) აბერანტულ (ამოვარდნილ, მნიშვნელოვნად გადახრილ, ანომალიურ) დაკვირვებად :

$$\mathcal{A} \leftarrow \mathcal{A} \bigcup \{\vec{x}\}.$$

  წინააღმდეგ შემთხვევაში :

  - $\mathcal{C} \leftarrow \{\vec{x}\}$ კლასტერის შექმნა ;
  - ამ კლასტერის გადიდება $\text{grow\_cluster}\left(\mathcal{C}, \mathcal{N}_\epsilon\left(\vec{x}\right), \epsilon, n_{\min}\right)$ პროცედურით.

3. კლასტერების სიაში $\mathcal{C}$ კლასტერის დამატება : $\mathcal{K} \leftarrow \mathcal{K} \cup \mathcal{C}$.

4. $\mathcal{C}$ კლასტერის ყველა ელემენტის მონიშვნა მონახულებულ ელემენტად : $\mathcal{V} \leftarrow \mathcal{V} \cup \mathcal{C}$.

grow_cluster$\left(\mathcal{C}, \mathcal{N}_{\epsilon}\left(\vec{x}\right), \epsilon, n_{\min}\right)$ პროცედურა განსაზღვრულია შემდეგი სახით :

ყველა $\vec{u} \in \mathcal{N} \setminus \mathcal{V}$ ობიექტისათვის :

– $\mathcal{N}' \leftarrow \mathcal{N}_{\epsilon}\left(\vec{u}\right)$ სიდიდის შექმნა ;

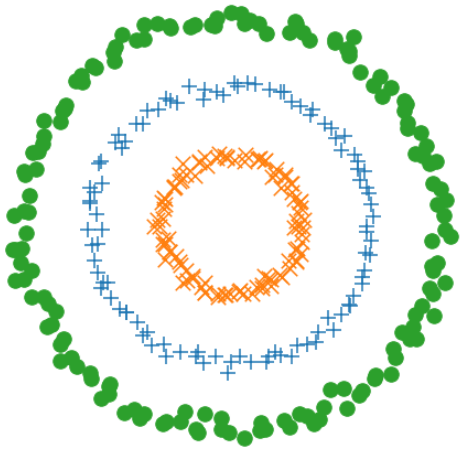
– თუ $\left|\mathcal{N}'\right| \geq n_{\min}$: $\mathcal{N}$ –-ის ისეთნაირად განახლება, რომ განხილვისას მხედველობაში იყოს მიღებული $\mathcal{N}'$ – ის ელემენტები : $\mathcal{N} \leftarrow \mathcal{N} \cup \mathcal{N}'$;

– თუ $\vec{u}$ არ მიეკუთვნება არცერთ სხვა კლასტერს, მისი დამატება $\mathcal{C}$ –-ში : $\mathcal{C} \leftarrow \mathcal{C} \cup \{\vec{u}\}$;

– თუ ადრე $\vec{u}$ კლასიფიცირებული იყო როგორც აბერანტული დაკვირვება, მისი ამოღება ანომალურ დაკვირვებათა სიიდან : $\mathcal{A} \leftarrow \mathcal{A} \setminus \{\vec{u}\}$.

DBSCAN-ის ერთ-ერთ უპირატესობას წარმოადგენს მისი მედეგობა, (მდგრადობა, სიმყარე, სიმტკიცე, გამძლეობა, ამტანობა) კლასტერის ფორმირებისას იდენტიფიცირებულ (გამოვლენილ) აბერანტულ (ამოვარდნილ, მნიშვნელოვნად გადახრილ, ანომალიურ) დაკვირვებათა მიმართ.

განზომილების ჭირი ან, თუ გენებოთ, უბედურება, შეჩვენება (ინგლ. curse of dimensionality), რაზეც საუბარი ადრე გვქონდა (იხ. პუნქტი 11.1.3), ართულებს DBSCAN ალგორითმის გამოყენებას ძალიან დიდი განზომილებების შემთხვევებში : მეზობელ კლასტერებში წარმოდგენილი იქნება მხოლოდ მათი ცენტრები. უფრო მეტიც, ვინაიდან სიმკვრივე განისაზღვრება ϵ და $n_{\min}$ პარამეტრებით, DBSCAN ალგორითმი ვერ შეძლებს განსხვავებულ სიმკვრივეთა მქონე კლასტერების პოვნას. მაგრამ DBSCAN ალგორითმი არ ითხოვს კლასტერების რაოდენობის წინასწარ დადგენას და გამოთვლით ასპექტში ეფექტურია. მისი გამოყენება $12.5\left(a\right)$ ნახატის მონაცემებისათვის ნაჩვენებია ნახატზე 1.27.



ნახატი 12.7 –12.5(a) ნახატის მონაცემების დაყოფა DBSCAN-ის საშუალებით.

7 სპექტრული კლასტერიზაცია

სპექტრული კლასტერიზაცია — ეს კლასტერიზაციის კიდევ ერთი მეთოდია, რომელიც ეყრდნობა დაკვირვებათა სიმკვრივეს. იგი იყენებს *მსგავსების გრაფის* აგებას, რომელშიც n წვერო (კვანძი) დაკვირვებებს შეესაბამება, ხოლო წიბოები (რკალები) წარმოადგენს მსგავსებას დაკვირვებათა შორის. $S \in \mathbb{R}^{n \times n}$ მსგავსების მატრიცის საფუძველზე ასეთი გრაფი შეიძლება იყოს აგებული რამდენიმე ხერხით :

- *სრული გრაფი*, რომლის (i,l) წიბო შეწონილია მსგავსების მატრიცის S_{il} ელემენტით ;
- *სრული შეკვეცილი (წაკვეთილი) გრაფი* (ინგლ. complete truncated graph), საიდანაც ამოღებულია ყველა წიბო, რომლისთვისაც მსგავსების მატრიცის S_{il} ელემენტი ფიქსირებულ ზღურბლზე ნაკლებია ;
- *k-უახლოეს მეზობელთა გრაფი*, რომელშიც თითოეული წვერო შეერთებულია თავის k უახლოეს მეზობელთან ;
- *ε-მიდამოს (ε-სამეზობლოს) გრაფი*, რომელშიც თითოეული წვერო შეერთებულია ყველა წვეროსთან თავის ε-მიდამოში (*ε-სამეზობლოში*).

უკანასკნელ სამ შემთხვევაში წიბოები შეიძლება იყოს შეწონილი მსგავსების მატრიცის S_{ij} ელემენტებით. ისიც დასაშვებია, რომ წიბო უბრალოდ არსებობდეს ან არა.

განსაზღვრება 12.9 (მოსაზღვრეობის, მომიჯნავეობის მატრიცა) დავუშვათ, რომ $\mathcal{G}$ არის გრაფი n წვეროთი (კვანძით). *მოსაზღვრეობის, მომიჯნავეობის მატრიცა* (ინგლ. adjacency matrix) ეწოდება ისეთ კვადრატულ სიმეტრიულ $A \in \mathbb{R}^{n \times n}$ მატრიცას, რომელშიც ყველა $i,l = 1,\ldots,n$ ინდექსებისათვის, A_{il} ელემენტი აკმაყოფილებს შემდეგ პირობას :

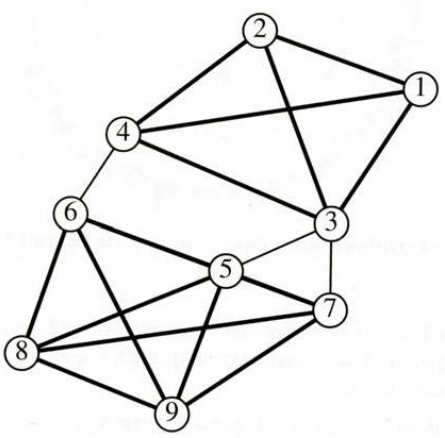
$$A_{il} = \begin{cases} 0 \text{ - თუ არ არის } i \text{ და } l \text{ წვეროების შემაერთებელი წიბო} \\ i \text{ და } l \text{ წვეროების შემაერთებელი წიბოს წონა - სხვა შემთხვევაში} \end{cases}. \tag{12.6}$$

კერძოდ, A მატრიცის დიაგონალი ნულია.

სრული მსგავსების გრაფის შემთხვევაში, $A = S$.

ჩვენი მონაცემებისთვის ავაგოთ მსგავსების გრაფი და შევეცადოთ კომპონენტთა K რიცხვის განსაზღვრა ამ გრაფში, ე.ი. დავადგინოთ იმ ქვეგრაფების რაოდენობა, რომელთა წვეროები ძლიერ არის დაკავშირებული ერთმანეთთან დიდი წონის წიბოებით.

ეს იდეა ილუსტრირებულია ნახატზე 12.8.



ნახატი 12.8 - მსგავსები გრაფი (სრული წაკვეთილი/შეკვეცილი გრაფი) 9 დაკვირვების შემცველი მონაცემთა ნაკრებისათვის. წიბოების სისქე მათი წონის პროპორციულია. აქ ჩვენ ორ გაერთიანებას ვხედავთ : ერთი შედგენილია დაკვირვებებით 1-4, მეორე — დაკვირვებებით 5-9.

7.1 გრაფის ლაპლასიანი

მსგავსების გრაფის კომპონენტების განსაზღვრისათვის ჩვენ გამოყენებული გვექნება ამ გრაფის *ლაპლასიანად* წოდებული მატრიცის თვისებები.

განსაზღვრება 12.20 (გრაფის ლაპლასიანი) დავუშვათ, რომ $\mathcal{G}$ არის გრაფი n წვეროთი, ხოლო A — მისი მოსაზღვრეობის/მომიჯნავეობის გრაფი. ამ გრაფის *ლაპლასიანი* ეწოდება ისეთ $L \in \mathbb{R}^{n\times n}$ მატრიცას, რომელშიც მისი L_{il} ელემენტი ყველა $i,l = 1,\ldots,n$ მნიშვნელობისათვის განისაზღვრება შემდეგი ფორმულით :

$$L_{il} = \begin{cases} \sum_{k=1}^{n} A_{ik} & \text{თუ } i = l \\ -A_{il} & \text{წინააღმდეგ შემთხვევაში} \end{cases}. \tag{12.7}$$

ყურადღება უნდა მივაქციოთ იმ გარემოებას, რომ $\sum_{k=1}^{n} A_{ik}$ არის i წვეროს *ხარისხი*, ე.ი. მისი მეზობლების რაოდენობა. ამრიგად, ელემენტები L მატრიცის დიაგონალზე არის $\mathcal{G}$ გრაფის წვეროთა ხარისხები. ელემენტების ჯამი L მატრიცის სტრიქონში უდრის 0 — ს.

თეორემა12.2 გრაფის ლაპლასიანი — ეს *განსაზღვრული* (*ნახევრადგანსაზღვრული*) *დადებითი* მატრიცაა. (გაიხსენეთ მათემატიკის კურსიდან, რომ *ნახევრადგანსაზღვრული დადებითი მატრიცა* არის *განსაზღვრული დადებითი მატრიცა*).

მტკიცებულება მართლაც, ყველა $\vec{v} \in \mathbb{R}^n$-თვის, გვაქვს:

$$\left.\begin{aligned}\vec{v}^{\mathrm{T}}L\vec{v} &= \sum_{i=1}^{n}\sum_{l=1}^{n} v_i v_l L_{il} = \sum_{i=1}^{n} v_i^2 A_{ii} - \sum_{i=1}^{n}\sum_{l\neq i} v_i v_l A_{il} \\ &= \sum_{i=1}^{n}\sum_{l=1}^{n} v_i^2 A_{il} - \sum_{i=1}^{n}\sum_{l=1}^{n} v_i v_l A_{il} \qquad (\text{რადგან } A_{ll}=0) \\ &= \frac{1}{2}\sum_{i=1}^{n}\sum_{l=1}^{n}\left(v_i^2 - v_l^2\right)A_{il} - \sum_{i=1}^{n}\sum_{l=1}^{n} v_i v_l A_{il} \quad (\text{რადგან } A_{il}=A_{li}) \\ &= \frac{1}{2}\sum_{i=1}^{n}\sum_{l=1}^{n} A_{il}\left(v_i - v_l\right)^2 \geq 0 \qquad (\text{რადგან } A_{il}\geq 0)\end{aligned}\right\}.$$

□

ინტუიციიდან გამომდინარე, მსგავსად იმისა, როგორც ფუნქციის ლაპლასიანი განსაზღვრავს ამ ფუნქციის ცვალებადობას, გრაფის ლაპლასიანი უნდა განსაზღვრავდეს იმ ფუნქციის ცვალებადობას, რომელიც აკავშირებს რიცხვს გრაფის თითოეულ წვეროსთან. უფრო ზუსტად, მივიღოთ ასეთი ფუნქციისათვის $f:\{1,\ldots,n\}\mapsto\mathbb{R}$ აღნიშვნა.

მაშინ $f^{\mathrm{T}}Lf=\frac{1}{2}\sum_{i=1}^{n}\sum_{l=1}^{n}A_{il}\left(f(i)-f(l)\right)^2=\frac{1}{2}\sum_{i\sim l}A_{il}\left(f(i)-f(l)\right)^2$, სადაც $i\sim l$ ნოტაცია იმაზე მიუთითებს, რომ გრაფში i და l დაკავშირებულია. ამ რიცხვის შესანარჩუნებლად დაბალ დონეზე, $f(i)$ და $f(l)$ მნიშვნელობები უნდა იყოს ახლოს ერთმანეთთან, როცა (i,l) წიბოს A_{il} წონა დიდია.

თეორემა 12.3 გრაფის ლაპლასიანის საკუთარი 0 მნიშვნელობის ჯერადობა არის ამ გრაფის ბმულ კომპონენტთა რიცხვი.

მტკიცებულება დავიწყოთ იმით, რომ $\left(0,\vec{1}_n\right)$ გამოსახულება წარმოადგენს L მატრიცის განსაკუთრებულ წყვილს :

$$\left(0,\vec{1}_n\right)=\left(\text{საკუთარი მნიშვნელობა, საკუთარი ვექტორი}\right),$$

სადაც $\vec{1}_n$ არის n ზომის ვექტორი, რომელშიც შემავალი თითოეული ელემენტი უდრის 1-ს. ასე რომ, ნებისმიერი $i=1,\ldots,n$ მნიშვნელობისთვის, $\left(L\vec{1}_n\right)_i=\sum_{l=1}^{n}L_{il}=0$ და, მაშასადამე, $L\vec{1}_n=0\vec{1}_n$.

დავუშვათ, რომ L არის K ბმული კომპონენტით შედგენილი გრაფის ლაპლასიანი. ასევე დავუშვათ, რომ წვეროები მოწესრიგებულია (დალაგებულია) იმ კომპონენტის შესაბამისად, რომელსაც ისინი მიეკუთვნება, მაშინ L მატრიცა არის ბლოკურად დიაგონალური ; იგი შედგენილია K ბლოკით და ამ ბლოკების რიცხვში თითოეული არის ერთ-ერთი ბმული კომპონენტის ლაპლასიანი. ნებისმიერ ბლოკს აქვს ნულოვანი საკუთარი მნიშვნელობა, რომელიც საკუთარ ვექტორს შეესაბამება და ამ საკუთარი ვექტორის ყველა მნიშვნელობა ტოლია, ე.ი. ერთმანეთს უდრის. ამრიგად, L მატრიცას აქვს სულ ცოტა იმდენივე ნულოვანი საკუთარი მნიშვნელობა, რამდენიც ბმული კომპონენტი გააჩნია მას, ხოლო თითოეული

ბმული კომპონენტის შესაბამისი საკუთარი ვექტორი ამ კომპონენტის ინდიკატორული ვექტორია.

დავუშვათ, რომ $\vec{u}$ ახლა არის საკუთარი ვექტორი ნულის ტოლი საკუთარი მნიშვნელობით. მაშინ $\vec{u}^{\mathrm{T}} L\vec{u} = 0$ და, ამრიგად :

$$\sum_{i \neq l} A_{il} \left(u_i - u_l\right)^2 = 0 .$$

ვინაიდან მოსაზღვრეობის (მომიჯნავეობის) მატრიცის A_{il} ელემენტი მკაცრად დადებითია, თუ i და l ბმულია, ვასკვნით, რომ $\vec{u}$ ვექტორის მდგენელები (ელემენტები), რომლებიც ერთსა და იმავე ბმულ კომპონენტს მიეკუთვნება, ტოლია. ამიტომ არ არსებობს იმაზე მეტი ნულოვანი საკუთარი მნიშვნელობა, რაც უკვე იყო დადგენილი ჩვენ მიერ.

□

სპექტრული კლასტერიზაცია

ამრიგად, თუ $\left\{\vec{x}^1, \vec{x}^2, \ldots, \vec{x}^n\right\}$ მონაცემთა ნაკრების (სიმრავლის) მსგავსების გრაფი შედგება K ბმული კომპონენტისგან, ამ მსგავსების გრაფის ლაპლასიანის ნულოვანი საკუთარი მნიშვნელობების შესაბამისი K საკუთარი ვექტორი არის ბმულ კომპონენტთა ინდიკატორული ვექტორები. ამიტომ მონაცემების პროეცირება ამ K საკუთარ ვექტორზე საშუალებას იძლევა განისაზღვროს, დადგინდეს, რომელ კომპონენტს მიეკუთვნება ისინი.

პრაქტიკულად იშვიათად ხდება, რომ მსგავსების გრაფი შედგება K ცალკეული ბმული კომპონენტისგან (ეს შეუძლებელია, მაგალითად, თუ ჩვენ ვიყენებთ სრულ გრაფს მსგავსებით, რომელიც ყოველთვის მკაცრად დადებითია ; იხ. ასევე ნახატი 12.8). ამიტომ, ნაცვლად ამისა, ჩვენ ვიყენებთ K უმცირეს საკუთარ მნიშვნელობას, რაც არ მოითხოვს კლასტერიზაციის k – საშუალოთა მეთოდით განხორციელებას კლასტერების იდენტიფიკაციისათვის (განსაზღვრისათვის). ამის თეორიული დასაბუთება შეიძლება ვიპოვოთ ჯერ კიდევ 2007 წელს გამოცემულ სახელმძღვანელოში, რომელიც ქალბატონ ულრიკე ფონ ლუქსბურგის (Ulrike von Luxburg, Eberhard Karls University of Tübingen, Faculty of Science, Department of Computer Science, Research Group : Theory of Machine Learning) კალამს ეკუთვნის.

განსაზღვრება 12.21 (სპექტრული კლასტერიზაცია) *სპექტრული კლასტერიზაცია* ანუ *სპექტრული დაყოფა* (ინგლ. *spectral clustering*) ეწოდება მასივის ელემენტთა (დაკვირვებათა) დანაწილების შემდეგ პროცედურას :

1. მსგავსების გრაფის აგება $\left\{\vec{x}^1, \vec{x}^2, \ldots, \vec{x}^n\right\}$ დაკვირვებათა შორის.

2. ამ გრაფის L ლაპლასიანის გამოთვლა.

3. L ლაპლასიანის K უმცირეს საკუთარ მნიშვნელობათა შესაბამისი K საკუთარი ვექტორის იდენტიფიცირება.

4. მონაცემთა პროეცირება ნაპოვნ K განზომილებაზე და ამ მონაცემების $\left\{\vec{z}^1, \vec{z}^2, \ldots, \vec{z}^n\right\}$ სახეთა (ანასახების) მიღება.

5. $\left\{\vec{z}^1, \vec{z}^2, \ldots, \vec{z}^n\right\}$-ის დაყოფა $k-$საშუალოთა (ინგლ. $k-$means) ალგორითმის გამოყენებით.

ამიტომ სპექტრული კლასტერიზაცია არ აკეთებს რაიმე დაშვებებს (კერძოდ, დაშვებებს ამოზნექილობის შესახებ) კლასტერთა ფორმის შესახებ. ეს მოდგომა, როგორც წესი, საკმაოდ ძვირადღირებულია გამოთვლათა დროის თვალსაზრისით, თუ, რა თქმა უნდა, გრაფის *მოსაზღვრეობის* (მომიჯნავეობის) A *მატრიცა* არ არის პარსიმონიული, ე.ი. დამზოგველი, ეკონომიური (ინგლ. parsimonious). ამიტომ პრაქტიკაში უფრო მიზანშეწონილია მსგავსების სრული გრაფის გამოყენება, რომელიც მიღებულია წაკვეთით, ასევე $k-$უახლოესი მეზობლის ან $\varepsilon-$მიდამოს ალგორითმებით.

8 საკვანძო მომენტები

- კლასტერიზაცია, ანუ მონაცემთა დაყოფა ჯგუფებად, მიმართულია კლასების გამოვლენაზე ჭდეთა გამოუყენებლად.
- ჭდეთა არარსებობის პირობებში დაყოფის ხარისხი შეიძლება იყოს შეფასებული სეპარაბელურობისა და ერთგვაროვნობის კრიტერიუმების საფუძველზე.
- იერარქიული კლასტერიზაცია მონაცემების დაყოფას იტერაციულად ახორციელებს. შედეგები შეიძლება იყოს ასახული დენდროგრამაზე.
- კლასტერიზაცია $k-$საშუალოთა მეთოდით ხორციელდება ლოიდის ალგორითმით ან რომელიმე ერთე-ერთით მისი ვარიენტებიდან. იგი ამოზნექილი K კლასტერის ეფექტურად პოვნის საშუალებას იძლევა.
- $k-$საშუალოთა მეთოდი კლასტერიზაციის კერძო შემთხვევაა გაუსის ნარევის მოდელით (ერთნაირი პროპორციები, ერთნაირი კოვარიაციული მატრიცები და იდენტურად ტოლი).
- $k-$საშუალოთა ბირთვული ვერსია შეიძლება იყოს გამოყენებული არაამოზნექილი კლასტერების გამოსავლენად.
- კლასტერიზაცია სიმკვრივით და, კერძოდ, DBSCAN (Density-based spatial clustering of applications with nois) ალგორითმი, შეიძლება იყოს გამოყენებული მონაცემთა ნაკრებების (მონაცემთა სიმრავლეების) მკვრივი არეების გამოსავლენად, ე.ი. დაკვირვებების, რომლებსაც შეუძლია არაამოზნექილი სიმრავლის შექმნა, მაგრამ ერთმანეთთან ახლოს არის.
- სპექტრული კლასტერიზაცია ასევე შეიძლება იყოს გამოყენებული მკვრივი არეების გამოსავლენად მონაცემთა ნაკრებში. DBSCAN (Density-based spatial clustering of applications with nois) მეთოდისგან განსხვავებით კლასტერების რაოდენობა წინასწარ არის განსაზღვრული.

დამატებითი ინფორმაცია

- ჩვენ მიერ განხილული დაყოფის ალგორითმები თითოეულ დაკვირვებას უკავშირებს მხოლოდ ერთ კლასტერს. არსებობს ალგორითმები, რომლებიც ცნობილია როგორც *არამკაფიო კლასტერიზაცია* (ინგლ. *fuzzy clustering*). ეს ალგორითმები ყოველ დაკვირვებასთან და ყოველ კლასტერთან აკავშირებს ალბათობას იმისა, რომ დაკვირვება სწორედ ამ კლასტერს მიეკუთვნება.
- ჯეინისა და დაბესის წიგნი (Jain და Dubes, 1988) კლასტერიზაციის ძალიან კარგ შესავალს წარმოადგენს. ურიგო არ იქნება ასევე იხილოთ სიუისა და ვუნშ II-ის სტატია (Xu და Wunsch II 2005).

9 ბიბლიოგრაფია

1. Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pages 226–231, Portland (OR). AAAI Press.

2. Jain, A. K. and Dubes, R. C. (1988). *Algorithms for Clustering Data*. Prentice Hall, New York.

3. Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2) :129–137.

4. Steinhaus, H. (1957). Sur la division des corps matériels en parties. *Bulletin de l'Académie polonaise des Sciences*,4(12) :801–804.

5. Xu, R. and Wunsch II, D. (2005). Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16 :645– 678.

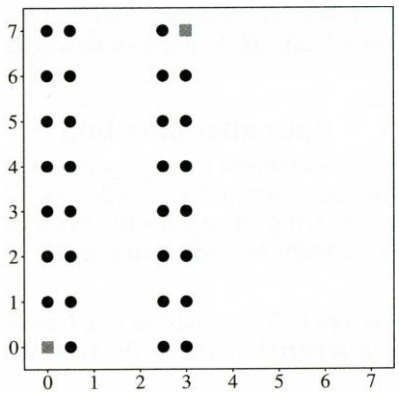
10 სავარჯიშოები

12.1 ელენეს აქვს n დაკვირვება, რომლებიც წარმოდგენილია p ბინარული ცვლადით.

1. იგი იყენებს იერარქიული კლასტერიზაციის ალგორითმს ერთი რგოლით. როგორია მის მიერ მიღებული დენდროგრამის მაქსიმალური სიღრმე ?

2. ახლა, ვთქვათ, ელენე სრული ბმულობის გამოყენებაზე გადადის. ამასთან ერთად დავუშვათ, რომ $p \square n$. შეცვლის ეს პასუხს ?

12.2 ბექას სურს მონაცემების გაყოფა ორ განზომილებად, რომლებიც ნაჩვენებია ნახატზე ქვემოთ. საწყის ცენტროიდებად იგი ორ კვადრატს იყენებს.



1. როგორი კლასტერები მიიღება ალგორითმის ერთი იტერაციის შემდეგ ?

2. იცვლება კლასტერები მეორე იტერაციის შემდეგ ?

12.3 რა მსგავსება და რა განსხვავება არის k-უახლოესი მეზობლის და $\mathrm{K}-$საშუალოთა მეთოდებს შორის ?

12.4 კატო იყენებს კლასტერიზაციის DBSCAN (ინგლ. **D**ensity-**B**ased **S**patial **C**lustering of **A**pplications with **N**oise - ხმაურის შემცველი გამოყენებების სივრცული კლასტერიზაცია სიმკვრივის საფუძველზე) მეთოდს თავის მონაცემებზე და რწმუნდება, რომ მიღებული აქვს ძალიან ფრაგმენტირებული კლასტერიზაცია, რომელიც შედგება ერთმანეთთან საკმაოდ ახლოს განლაგებული მცირე კლასტერების სიმრავლეებისგან. რომელი პარამეტრ(ებ)ის შეცვლა შეუძლია კატოს და როგორ კლასტერების ნაკლები რაოდენობის მისაღებად ?

12.5 დურუმ განახორციელა ლოიდის (Stuart P. Lloyd) ალგორითმის საკუთარი ვერსია და შენიშნა, რომ დაკვირვებათა ხელახლა განაწილებისას კლასტერების მიხედვით ალგორითმი პოულობს იმ გამყოფს, რომელსაც დურუ უკვე იყენებდა წინა ეტაპზე. ნორმალურია ეს ?

12.6 ელისოს სურს იერარქიული კლასტერიზაციის ალგორითმის გამოყენება ბინარული ცვლადებით წარმოდგენილ მონაცემებზე : რომელი მანძილი უნდა გამოიყენოს მან ?

სავარჯიშოთა ამონახსნები

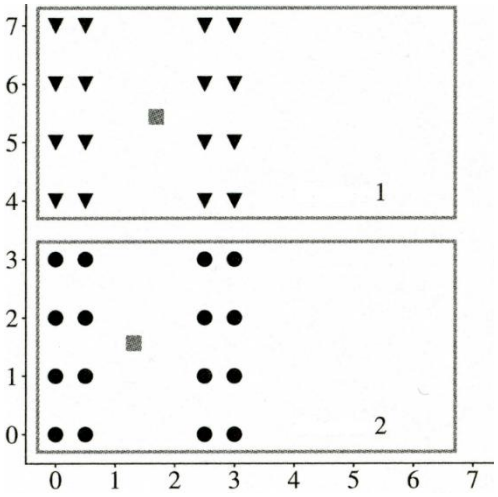
12.1

1. მაქსიმუმ n : არსებობს ერთი კლასტერი, რომლის ზომა იზრდება 1-ით აგლომერაციული ალგორითმის ყოველ იტერაციაზე.

2. ბინარული ცვლადების გამოყენებისას შეუძლებელია, რომ დაკვირვებები ერთმანეთისგან ძალიან დაშორებული იყოს (რაც აუცილებელია იმისათვის, რომ არ მოხდეს ერთზე მეტი დაკვირვების შემცველი ორი კლასტერის აგლომერაცია - შეგროვება, თავმოყრა, შეჯგუფება). ამიტომ სიღრმე n-ზე ნაკლები იქნება.

12.2.

1.



2. არა.

12.3 ორივე მეთოდი აყალიბებს დაკვირვებათა სივრცის ამოზნექილ დაყოფას და ორივე მეთოდს აქვს ერთი k ჰიპერპარამეტრი. დანარჩენ ასპექტში კი ეს ძალიან განსხვავებული მეთოდებია (ერთი არის კონტროლირებადი, ხოლო მეორე — არაკონტროლირებადი ; k პარამეტრს არ აქვს მათთვის ერთნაირი შინაარსი ; და ასე შემდეგ).

12.4 ε-ის გაზრდა გამოიწვევს ε–მეზობლობის ზომების გადიდებას და, მაშასადამე, კლასტერების ზომებისაც. $n_{\min}$-ის შემცირება გაზრდის შიდა წერტილების რაოდენობას და, ამრიგად, კლასტერების ზომებსაც.

12.5 არა (იმ შემთხვევის გამოკლებით, როცა ამონახსნის კრებადობა განხორციელდა და როცა დაყოფის ცვლილება უკვე შეწყდა). მართლაც, გლობალური შიდაკლასტერული დისპერსია მცირდება ყოველ იტერაციასთან ერთად ; ამიტომ შეუძლებელია დავუბრუნდეთ წინა კონფიგურაციას, თუ, რა თქმა უნდა, იგი მუდმივი არ არის.

12.6 ბინარულ მონაცემებს, ჩვეულებრივ, შეესაბამება ჰემინგის მანძილი.

ლექცია 13 ამოზნექილი ოპტიმიზაციის პრინციპები

შინაარსი

ლექციათა ამ კურსში მანქანური სწავლების მრავალ მოდელს ჩვენ ვაგებთ როგორც ოპტიმიზაციის ამოცანას.

ჩვეულებრივ, ეს არის ემპირიული რისკის მინიმიზაცია ან დისპერსიის მაქსიმიზაცია, ხშირად გარკვეული შეზღუდვების პირობებში. ამ ამოცანათა ეფექტურად გადაწყვეტა ზოგად შემთხვევაში ძნელია. მაგრამ მანქანური სწავლების კონტექსტში განსახილველი ფუნქციები ხშირად *ამოზნექილია* (ინგლისურად : *convex*), რაც მნიშვნელოვნად ამარტივებს ამოცანას. მოცემული კურსის მიზანია ოპტიმიზაციის ეგრეთ წოდებული ამოზნექილი ამოცანებისა და მათი ამოხსნის მეთოდების აღწერა.

მიზნები

— ოპტიმიზაციის ამოზნექილი ამოცანის ამოცნობა ;

— ოპტიმიზაციის ამოზნექილი ამოცანის ზუსტი ამ მიმართული დაშვების ალგორითმის საშუალებით ამოხსნა;

— კვადრატული ოპტიმიზაციის დუალური ამოცანის ფორმულირება ;

— კარუშის-კუნის-ტაკერის (Karush-Kuhn-Tucker, *KKT*) პირობათა ჩაწერა კვადრატული

ოპტიმიზაციის ამოცანისთვის.

1 ამოზნექილობა

დავიწყოთ ამოზნექილობის ცნების განსაზღვრებით ჯერ სიმრავლისათვის, ხოლო შემდეგ გადავიდეთ ამავე ცნების განმარტებაზე ფუნქციისთვის.

1.1 ამოზნექილი სიმრავლე

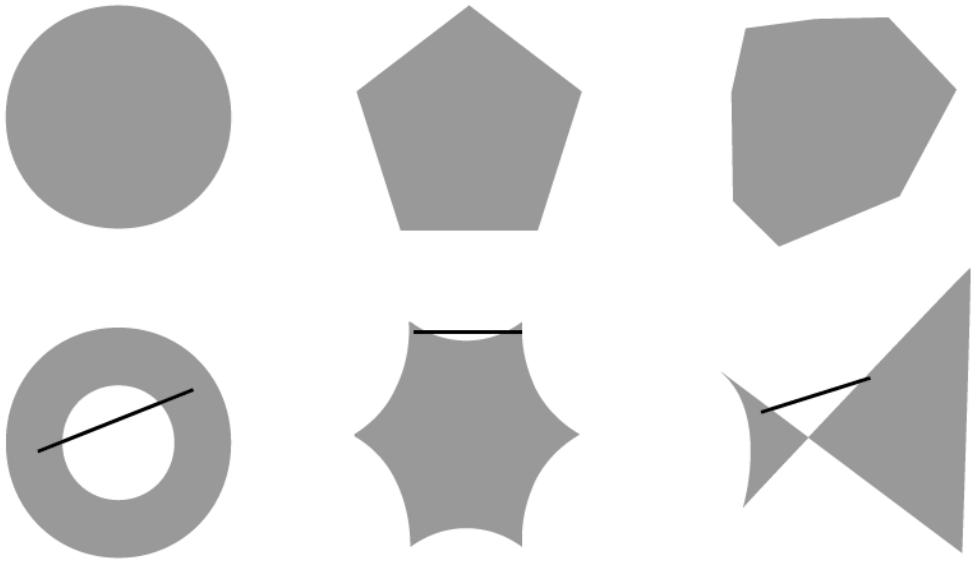
განსაზღვრება 1.1 (ამოზნექილი სიმრავლე) $S \subseteq \mathbb{R}^n$ სიმრავლის შესახებ ამბობენ, რომ იგი *ამოზნექილია*, მაშინ და მხოლოდ მაშინ, როცა ნებისმიერ $\vec{u}, \vec{v} \in S$ და $t \in [0,1]$ პირობებში ადგილი აქვს შემდეგ თანაფარდობას :

$$t\vec{u} + (1-t)\vec{v} \in S.$$

სხვა სიტყვებით რომ ვთქვათ, წრფის $[\vec{u}, \vec{v}]$ მონაკვეთი მთლიანად მიეკუთვნება S სიმრავლეს.

⚙

ნახატზე 13.1 ნაჩვენებია ამოზნექილ და არაამოზნექილ სიმრავლეთა რამდენიმე მაგალითი $\mathbb{R}^2$ სივრცეში.



ნახატი 13.1 – სამი სიმრავლე $\mathbb{R}^2$ სივრციდან, რომლებიც ნაჩვენებია ზედა რიგში, არის ამოზნექილი ; სამი სიმრავლე ქვედა რიგში არ არის ამოზნექილი და თითოეული მათ შორის წარმოდგენილია წრფის მონაკვეთით, რომელიც სიმრავლის ორ წერტილს აკავშირებს, მაგრამ მთლიანად ამ სიმრავლეში ჩართული არ არის.

1.2 ამოზნექილი ფუნქცია

განსაზღვრება 1.2 (ამოზნექილი ფუნქცია) დავუშვათ, რომ $\mathcal{U} \subseteq \mathbb{R}^n$. მაშინ $f : \mathcal{U} \to \mathbb{R}$ ფუნქციას *ამოზნექილი* ეწოდება, როცა :

- f ფუნქციის განსაზღვრების $\mathcal{U}$ არე ამოზნექილი სიმრავლეა ;
- ნებისმიერ $\vec{u}, \vec{v} \in S$ და $t \in [0,1]$ პირობებში სრულდება

$$f\left(t\vec{u}+(1-t)\vec{v}\right)\le t\,f\left(\vec{u}\right)+(1-t)\,f\left(\vec{v}\right)$$

უტოლობა, ე.ი. $[\vec{u},\vec{v}]$ მონაკვეთზე f ფუნქცია განთავსებულია $\left[\left(\vec{u},f\left(\vec{u}\right)\right),\left(\vec{v},f\left(\vec{v}\right)\right)\right]$ მონაკვეთის ქვემოთ.

თუ უტოლობა მკაცრია ყველა $\vec{u}\neq\vec{v}$ მნიშვნელობისთვის $\mathcal{U}$ არეში და ამასთან ერთად $t\in\left]0,1\right[$, მაშინ ლაპარაკობენ *მკაცრად ამოზნექილი* ფუნქციის შესახებ. ამოზნექილი ფუნქციის სიმრუდე აღემატება აფინური ფუნქციის ასეთივე მახასიათებელს.

დაბოლოს, იმ შემთხვევაში, როცა არსებობს დადებითი $k>0$ სიდიდე, ისეთი რომ $f-\frac{k}{2}\|\vec{u}\|_2^2$ ფუნქცია მკაცრად ამოზნექილია, მაშინ ამბობენ, რომ f ფუნქცია *ძლიერ ამოზნექილია.*

მაგალითი

შემდეგი ფუნქციები ამოზნექილია :

- $f:\mathbb{R}\to\mathbb{R},u\mapsto u^{2a}\quad a\in\mathbb{N};$
- $f:\mathbb{R}_+^*\to\mathbb{R},u\mapsto u^a\quad a\notin\left]0,1\right[;$
- $f:\mathbb{R}\to\mathbb{R},u\mapsto u^{au}\quad a\in\mathbb{R};$
- $f:\mathbb{R}_+^*\to\mathbb{R},u\mapsto-\log(au)\quad a\in\mathbb{R};$
- $f:\mathbb{R}^n\to\mathbb{R},\vec{u}\mapsto\vec{a}^T\vec{u}+b\quad\vec{a}\in\mathbb{R}^n,b\in\mathbb{R};$
- $f:\mathbb{R}^n\to\mathbb{R},\vec{u}\mapsto\frac{1}{2}\vec{u}^TQ\vec{u}+\vec{a}^T\vec{u}+b\quad\vec{a}\in\mathbb{R}^n,b\in\mathbb{R},Q\ge0;$
- $f:\mathbb{R}^n\to\mathbb{R},\vec{u}\mapsto\|\vec{u}\|_p=\left(\sum_{i=1}^n|u_i|^p\right)^{\frac{1}{p}}\quad p\ge1.$

ამოზნექილი ფუნქციის საწინააღმდეგო ფუნქციას *ჩაზნექილი* (ინგლისურად : *concave*) ეწოდება.

განსაზღვრება 1.3 (ჩაზნექილი ფუნქცია) დავუშვათ, რომ $\mathcal{U}\subseteq\mathbb{R}^n$. მაშინ $f:\mathcal{U}\to\mathbb{R}$ ფუნქციას ეწოდება *ჩაზნექილი* (ინგლისურად : *concave*) მაშინ და მხოლოდ მაშინ, როცა $-f$ არის ამოზნექილი :

- $\mathcal{U}$ ამოზნექილი სიმრავლეა ;
- როგორიც არ უნდა იყოს $\vec{u},\vec{v}\in\mathcal{U}$ და $t\in[0,1]$, ადგილი აქვს უტოლობას

$$f\left(t\vec{u}+(1-t)\vec{v}\right)\ge t\,f\left(\vec{u}\right)+(1-t)\,f\left(\vec{v}\right).$$

თუ უტოლობა მკაცრია ყველა $\vec{u} \neq \vec{v}$ მნიშვნელობისთვის $\mathcal{U}$ არეში და თანაც $t \in]0,1[$, მაშინ ლაპარაკობენ *მკაცრად ჩაზნექილი* ფუნქციის შესახებ.

შემდეგი ოპერაციები არ არღვევს ამოზნექილობას :

- წრფივი დადებითი კომბინაცია : თუ $f_1, f_2, \ldots, f_k : \mathbb{R}^n \to \mathbb{R}$ ფუნქციები ამოზნექილია, და ამასთან ერთად $a_1, a_2, \ldots, a_k > 0$, მაშინ

  $$a_1 f_1 + a_1 f_2 + \cdots + a_k f_k$$

  გამოსახულება ამოზნექილია.

- მაქსიმიზაცია : თუ $f_1, f_2, \ldots, f_k : \mathbb{R}^n \to \mathbb{R}$ ფუნქციები ამოზნექილია, მაშინ

  $$\vec{u} \mapsto \max_{1,\ldots,k} f_k(\vec{u})$$

  გამოსახულება ამოზნექილია.

- ნაწილობრივი მინიმიზაცია : თუ $f : \mathbb{R}^n \to \mathbb{R}$ არის ამოზნექილი, და ამასთან ერთად $\mathcal{C} \subseteq \mathbb{R}^n$ ამოზნექილი სიმრავლეა, მაშინ

  $$u_1, u_2, \ldots, u_{k-1} \mapsto \min_{v \in \mathcal{C}} f_k(u_1, u_2, \ldots, u_{k-1}, v)$$

  არის ამოზნექილი.

2 ამოზნექილი ოპტიმიზაციის ამოცანები

2.1 ფორმულირება და ლექსიკა

განსაზღვრება 1.4 (ამოზნექილი ოპტიმიზაცია) თუ მხედველობაში მივიღებთ $\mathcal{U} \subseteq \mathbb{R}^n$ პირობას და ამოზნექილ $f : \mathcal{U} \to \mathbb{R}$ ფუნქციას, მაშინ *ამოზნექილი ოპტიმიზაციის ამოცანა* შეიძლება ვუწოდოთ შემდეგ პრობლემას :

$$\min_{\vec{u} \in \mathcal{U}} f(\vec{u}).$$

ამ დროს თავად f -ს ეწოდება *მიზნობრივი ფუნქცია*, ზოგჯერ *ღირებულების (დანახარჯის) ფუნქცია* ან კიდევ *კრიტერიული ფუნქცია.*

$\vec{u}^* \subseteq \mathcal{U}$ წერტილს, რომელიც ამოწმებს $f(\vec{u}^*) \leq f(\vec{u}) \; \forall \vec{u} \in \mathcal{U}$ პირობას, ეწოდება f -ის *მინიმუმის წერტილი* $\mathcal{U}$ -ზე, მაშინ როცა $f(\vec{u}^*)$ სიდიდეს ეწოდება f -ის *მინიმუმი* $\mathcal{U}$ -ზე.

განსაზღვრება 1.5 (ამოზნექილი ოპტიმიზაცია შეზღუდვათა პირობებში) დავუშვათ, რომ მოცემულია ორი, m და r, დადებითი მთელი რიცხვი, $\mathcal{U}, \mathcal{U}_1, \ldots, \mathcal{U}_m, \mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_r \subseteq \mathbb{R}^n$ სიმრავლეები, ერთი ამოზნექილი $f : \mathcal{U} \to \mathbb{R}$ ფუნქცია, m ასევე ამოზნექილი $g_i : \mathcal{U}_i \to \mathbb{R}$

ფუნქცია და r აფინური $h_j : \mathcal{V}_i \to \mathbb{R}$ ფუნქცია, მაშინ *ამოზნექილი ოპტიმიზაციის ამოცანა შეზღუდვათა პირობებში* ეწოდება შემდეგ ამოცანას :

$$\min_{\vec{u}\in\mathcal{D}} f(\vec{u}) \text{ შეზღუდვათა პირობებში}$$

$$g_i(\vec{u}) \le 0 \quad \forall i = 1,\ldots,m$$

$$h_j(\vec{u}) = 0 \quad \forall j = 1,\ldots,r,$$

სადაც $\mathcal{D} = \mathcal{U} \bigcap_{i=1}^{m} \mathcal{U}_i \bigcap_{j=1}^{r} \mathcal{V}_j$ არის განსაზღვრების საერთო არე ყველა განსახილველი ფუნქციისთვის.

f არის *მიზნობრივი ფუნქცია*. $g_i \le 0$ სახის m შეზღუდვა *უტოლობაა*. $h_j = 0$ სახის r შეზღუდვა კი *ტოლობას* წარმოადგენს.

$\vec{v} \in \mathcal{D}$ წერტილს, რომელიც ამოწმებს ყველა შეზღუდვას, ეწოდება *დასაშვები წერტილი* ან *მიღწევადი* (*რეალიზებადი*, *შესრულებადი*) *წერტილი* (ინგლისურად : *feasible point*), ხოლო დასაშვებ წერტილთა სიმრავლეს — *დასაშვები რეგიონი*, *შეზღუდვათა დაკმაყოფილების არე* ან კიდევ, სხვანაირად, *შეზღუდვების ზონა* (ინგლისურად : *feasible domain*). ეს სიმრავლე არის ამოზნექილი.

2.2 ლოკალური და გლობალური ექსტრემუმები

ტერმინი «ექსტრემუმი» გამოიყენება განურჩევლად როგორც მაქსიმუმის, ასევე მინიმუმის აღსანიშნავად. ექსტრემუმები შეიძლება იყოს გლობალური, ესე იგი ანიჭებდეს მაქსიმუმს (შესაბამისად მინიმუმს) ფუნქციას მისი განსაზღვრების არეში, ან ლოკალური, ესე იგი ოპტიმალური თავის ახლობლობაში. ფორმალურად :

განსაზღვრება 1.6 (გლობალური ექსტრემუმი) დავუშვათ, რომ $\mathcal{U} \subseteq \mathbb{R}^n$. $f : \mathcal{U} \to \mathbb{R}$ და $\vec{u}^* \in \mathcal{U}$. ამბობენ, რომ $\vec{u}^*$ არის f ფუნქციის *გლობალური მინიმუმის წერტილი* ან *აბსოლუტური მინიმუმის წერტილი* $\mathcal{U}$ -ზე, თუ

$$f(\vec{u}^*) \le f(\vec{u}) \quad \forall \vec{u} \in \mathcal{U}.$$

$\vec{u}^*$ არის f ფუნქციის *გლობალური მაქსიმუმის წერტილი* $\mathcal{U}$ -ზე, თუ იგი არის $-f$ ფუნქციის *გლობალური მინიმუმის წერტილი* $\mathcal{U}$ -ზე.

განსაზღვრება 1.7 (ლოკალური ექსტრემუმი) დავუშვათ, რომ $\mathcal{U} \subseteq \mathbb{R}^n$, $f : \mathcal{U} \to \mathbb{R}$ და $\vec{u}^* \in \mathcal{U}$. მაშინ ამბობენ, რომ $\vec{u}^*$ არის f ფუნქციის *ლოკალური მინიმუმის წერტილი*, ანუ *ფარდობითი მინიმუმის წერტილი* $\mathcal{U}$ -ზე, თუ არსებობს $\vec{u}^*$ წერტილის $\mathcal{V}$ მახლობლობა $\mathbb{R}^n$ სივრცეში, ისეთი, რომ

$$f\left(\vec{u}^*\right) \le f\left(\vec{u}\right) \quad \forall \vec{u} \in \mathcal{U} \bigcap \mathcal{V}.$$

$\vec{u}^*$ არის f ფუნქციის *ლოკალური მაქსიმუმის წერტილი* $\mathcal{U}-$ზე, თუ იგი არის $-f$ ფუნქციის ლოკალური მინიმუმის წერტილი $\mathcal{U}-$ზე.

ამოზნექილობის ცნება ძალიან მნიშვნელოვანია ოპტიმიზაციისას, რადგან ამოზნექილობის შემთხვევაში ლოკალური მინიმუმი გარანტირებულად გლობალური მინიმუმიც არის.

თეორემა 1.1 *დავუშვათ, რომ* $\mathcal{U} \subseteq \mathbb{R}^n$, $\mathcal{U} \to \mathbb{R}$, *ხოლო* $\vec{u}^* \in \mathcal{U}$ *არის* f *ფუნქციის ლოკალური მინიმუმის წერტილი* $\mathcal{U}-$ *ზე. მაშინ*

- *თუ* f *ამოზნექილი ფუნქციაა, მაშინ* $\vec{u}^*$ *გლობალური მინიმუმის წერტილია*;
- *თუ* f *მკაცრად ამოზნექილი ფუნქციაა, მაშინ* $\vec{u}^*$ *გლობალური მინიმუმის ერთადერთი წერტილია.*

მტკიცებულება. თუ $\vec{u}^*$ ლოკალური მინიმუმის წერტილია, მაშინ არსებობს $\epsilon > 0$, ისეთი რომ ყველა $\vec{u}-$თვის $\left\|\vec{u}-\vec{u}^*\right\|_2 \le \epsilon$ პირობის შესრულებისას :

$$f\left(\vec{u}^*\right) \le f\left(\vec{u}\right). \tag{13.1}$$

დავუშვათ, რომ $\vec{v} \in \mathcal{U}$ და იგი განსხვავებულია $\vec{u}^*-$გან. დავუშვათ ასევე, რომ f ამოზნექილი ფუნქციაა. ვუჩვენოთ, რომ $f\left(\vec{v}\right) \ge f\left(\vec{u}^*\right)$. შესაძლებელია ორი შემთხვევა : ან $\vec{v}$ მიეკუთვნება $\vec{u}^*$ წერტილის სამეზობლოს, სხვანაირად, $\left\|\vec{v}-\vec{u}^*\right\|_2 \le \epsilon$, ასეთ შემთხვევაში $f\left(\vec{v}\right) \ge f\left(\vec{u}^*\right)$ (13.1) ფორმულის თანახმად, ან $\vec{v}$ ამ სამეზობლოს მიღმაა და ასეთ შემთხვევაში

$$\left\|\vec{v}-\vec{u}^*\right\|_2 > \epsilon. \tag{13.2}$$

ახლა განვსაზღვროთ $\vec{u} = \left(1-\lambda\right)\vec{u}^*$, სადაც $\lambda = \frac{1}{2}\frac{\epsilon}{\left\|\vec{v}-\vec{u}^*\right\|_2}$. ამიტომ, (13.2) ფორმულის თანახმად, $0 < \lambda < 1$.

ცხადია, რომ $\mathcal{U}$ სიმრავლის ამოზნექილობის გამო, $\vec{u} \in \mathcal{U}$.

ვინაიდან

$$\left\|\vec{u}-\vec{u}^*\right\|_2 = \left\|\left(1-\lambda\right)\vec{u}^* + \lambda\vec{v} - \vec{u}^*\right\|_2 = \lambda\left\|\vec{v}-\vec{u}^*\right\|_2 = \frac{\epsilon}{2} < \epsilon,$$

მაშასადამე $\vec{u}$ მიეკუთვნება $\vec{u}^*$ წერტილის სამეზობლოს.

(13.1) განტოლების თამამად კი $f\left(\vec{u}\right) \ge f\left(\vec{u}^*\right)$.

f -ის ამოზნექილობის გამო,

$$f(\vec{u}) \le (1-\lambda) f(\vec{u}^*) + \lambda f(\vec{v}) = f(\vec{u}^*) + \lambda \left(f(\vec{v}) - f(\vec{u}^*) \right).$$

რადგან $f(\vec{u}^*) \le f(\vec{u})$, მივიღებთ :

$$f(\vec{u}^*) \le f(\vec{u}^*) + \lambda \left(f(\vec{v}) - f(\vec{u}^*) \right),$$

საიდანაც ვასკვნით (რაკი $\lambda > 0$), რომ

$$f(\vec{v}) \ge f(\vec{u}^*).$$

ამრიგად, ყველა $\vec{v} \in \mathcal{U}$ – თვის, რომელიც განსხვავებულია $\vec{u}^*$ – გან, $f(\vec{v}) \ge f(\vec{u}^*)$ და ამიტომ $\vec{u}^*$ არის f – ის გლობალური მინიმუმი.

თუ f მკაცრად ამოზნექილია, ასეთივე მსჯელობა ჩატარდება მკაცრი უტოლობებისთვის და მიიღება რომ $\vec{u}^*$ არის f -ის ერთადერთი გლობალური მინიმუმი.

□

გადავიდეთ ამოზნექილი ოპტიმიზაციის ამოცანის ამონახსნის მიღებაზე.

3 ამოზნექილი ოპტიმიზაცია შეზღუდვის გარეშე

დავიწყოთ ამოზნექილი ოპტიმიზაციის ამოცანებით, რომლებიც შეზღუდვებს არ შეიცავს. მიუხედავად იმისა, რომ ამოზნექილი ოპტიმიზაციის ამოცანა ზოგად შემთხვევაში რთულია, არსებობს მრავალი მეთოდი, რომლის გამოყენება შესაძლებელია ძალიან კარგი ხარისხის მიახლოებითი რიცხვითი ამონახსნის ეფექტურად მიღებისთვის, განსაკუთრებით უწყვეტობისა და დიფერენცირებადობის შესახებ გარკვეულ დაშვებათა შემოღებისას, რომლებიც სტატისტიკური სწავლების კონტექსტში დასმული ამოცანებისთვის ხშირად ადვილი შესამოწმებელია.

3.1 ამოზნექილობის სხვადასხვა დამახასიათებელი თვისება

ამოზნექილი და დიფერენცირებადი ფუნქცია ნებისმიერი თავისი მხების ზემოთ ძევს, როგორც ეს ნაჩვენებია ნახატზე 13.2. ფორმალურად :

თეორემა 1.2 პირველი გვარის დამახასიათებელი თვისება *დავუშვათ, რომ* $\mathcal{U} \subseteq \mathbb{R}^n$ *და* $f : \mathcal{U} \to \mathbb{R}$ არის C^1 კლასის ფუნქცია *(სხვანაირად რომ ვთქვათ, უწყვეტი, რომლის ყველა კერძო წარმოებული არსებობს და ასევე უწყვეტია). ასეთ პირობებში* f *ფუნქცია ამოზნექილია მაშინ და მხოლოდ მაშინ, თუ*

- $\mathcal{U}$ *ამოზნექილია ;*
- $\vec{u}, \vec{v} \in \mathcal{U}$ *წერტილებისგან დამოუკიდებლად,*

$$f(\vec{v}) \ge f(\vec{u}) + \nabla f(\vec{u})^{\mathrm{T}} (\vec{v} - \vec{u}). \qquad (13.3)$$

მტკიცებულება. ვთქვათ, $t \in [0,1]$ და $\vec{u}, \vec{v} \in \mathcal{U}$. ასევე დავუშვათ, რომ f ამოზნექილია. მაშინ

$$\lim_{t \to 0^+} \frac{1}{t}\Big(f\left(\vec{u} + t\left(\vec{v} - \vec{u}\right)\right) - f\left(\vec{u}\right)\Big) = \nabla f\left(u\right)^{\mathrm{T}}\left(\vec{v} - \vec{u}\right). \tag{13.4}$$

f ფუნქციის ამოზნექილობის ძალით, შესაძლებელია ვწეროთ, რომ

$$f\left(\vec{u} + t\left(\vec{v} - \vec{u}\right)\right) = f\left(\left(1-t\right)\vec{u} + t\vec{v}\right) \le \left(1-t\right) f\left(\vec{u}\right) + t\, f\left(\vec{v}\right), \tag{13.5}$$

და, ამრიგად,

$$\frac{1}{t}\Big(f\left(\vec{u} + t\left(\vec{v} - \vec{u}\right)\right) - f\left(\vec{u}\right)\Big) \le f\left(\vec{v}\right) - f\left(\vec{u}\right). \tag{13.6}$$

თუ განტოლებაში 13.4 ზღვარზე გადავალთ, მაშინ უტოლობა 13.3 მიიღება.

და პირიქით, თუ მოცემულია, რომ $\vec{u}, \vec{v} \in \mathcal{U}$ და $t \in]0,1[$, შემოვიღოთ $\vec{w} = t\vec{u} + \left(1-t\right)\vec{v}$. ვინაიდან $\mathcal{U}$ სიმრავლე ამოზნექილია, ამიტომ

$$\vec{w} \in \mathcal{U}.$$

თუ უტოლობა 13.3 სრულდება, მაშინ :

$$\left.\begin{aligned} f\left(\vec{u}\right) &\ge f\left(\vec{w}\right) + \nabla f\left(\vec{w}\right)^{\mathrm{T}}\left(\vec{u} - \vec{w}\right) \\ f\left(\vec{v}\right) &\ge f\left(\vec{w}\right) + \nabla f\left(\vec{w}\right)^{\mathrm{T}}\left(\vec{v} - \vec{w}\right) \end{aligned}\right\}.$$

თუ t-ზე გამრავლებულ პირველ უტოლობას მივუმატებთ $\left(1-t\right)$-ზე გამრავლებულ მეორე უტოლობას, მივიღებთ

$$t\, f\left(\vec{u}\right) + \left(1-t\right) f\left(\vec{v}\right) \ge f\left(\vec{w}\right) + \nabla f\left(\vec{w}\right)^{\mathrm{T}}\left(\vec{v} - \vec{w}\right) \tag{13.7}$$

უტოლობას და, ამრიგად, f ფუნქციის ამოზნექილობის მტკიცებულებასაც.

□

თეორემიდან 1.2 გამომდინარეობს, რომ *ამოზნექილი და დიფერენცირებადი f ფუნქცია მინიმალურია იქ, სადაც მისი გრადიენტი ნულის ტოლი ხდება.*

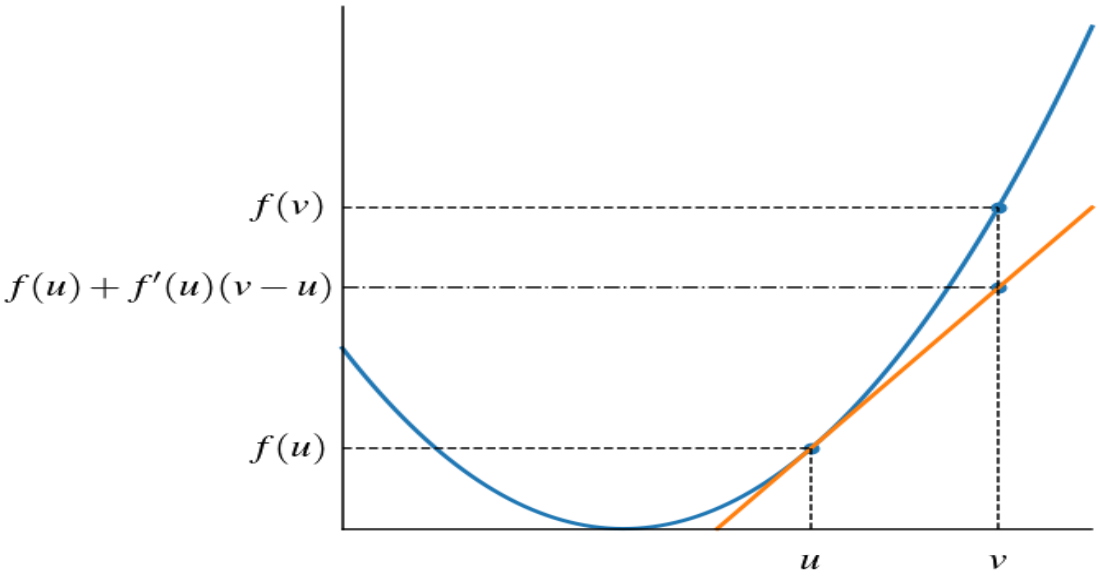
ფორმალურად :

თეორემა 1.3 *დავუშვათ, რომ $\mathcal{U} \subseteq \mathbb{R}^n$ და $f : \mathcal{U} \to \mathbb{R}$ არის $\mathcal{C}^1$ კლასის ამოზნექილი ფუნქცია. დავუშვათ ასევე, რომ $\vec{u}^* \in \mathcal{U}$.*

მაშინ შემდეგი წინადადებები ეკვივალენტურია :

1. $\vec{u}^*$ არის f ფუნქციის მინიმუმის წერტილი $\mathcal{U}$ სიმრავლეზე ;

2. $\nabla f\left(\vec{u}^*\right)=0.$



ნახატი 13.2 – ამოზნექილი f ფუნქცია ნებისმიერი თავისი მხების ზემოთ ძევს.

3.2 მეორე რიგის ამოზნექილობის დახასიათება

$\mathcal{C}^2$ კლასის ამოზნექილი ფუნქციები ხასიათდება იმით, რომ მათი ჰესიანი დადებითად ნახევრადგანსაზღვრულია.

სხვა სიტყვებით რომ ვთქვათ, ამ ფუნქციებს ყველგან დადებითი სიმრუდე აქვს.

ეს თვისება ქვემოთ წარმოდგენილ ზოგიერთ ალგორითმში გამოიყენება.

თეორემა 1.4 მეორე რიგის დამახასიათებელი თვისება *დავუშვათ, რომ* $\mathcal{U}\subseteq\mathbb{R}^n$ *და* $f:\mathcal{U}\to\mathbb{R}$ *არის* $\mathcal{C}^2$ *კლასის ფუნქცია (სხვანაირად რომ ვთქვათ, უწყვეტი, ორჯერ დიფერენცირებადი, პირველი და მეორე რიგის ასევე უწყვეტი წარმოებულებით). ასეთ პირობებში* f *ფუნქცია ამოზნექილია მაშინ და მხოლოდ მაშინ, თუ*

- $\mathcal{U}$ *ამოზნექილია ;*
- $\vec{u}\in\mathcal{U}$ *წერტილისგან დამოუკიდებლად* $\nabla^2 f\left(\vec{u}\right)\geq 0.$

მტკიცებულება. დავუშვათ, რომ $\vec{u}\in\mathcal{U}$ და $\vec{v}\subseteq\mathbb{R}^n$. ჩავთვალოთ, რომ $I=\left\{t\in\mathbb{R}\,|\,\vec{u}+t\vec{v}\in\mathcal{U}\right\}$ და განვსაზღვროთ $\phi:I\to\mathbb{R},\ t\mapsto f\left(\vec{u}+t\vec{v}\right)$. მაშინ I არის ინტერვალი და ϕ ფუნქცია ამოზნექილია.

მართლაც, დავუშვათ, რომ I შეიცავს თუნდაც ორ ცალკეულ t_1 და t_2 წერტილს. (წინააღმდეგ

შემთხვევაში, I — ეს ან *სინგლეტონია*, ე.ი. ერთადერთი ელემენტის შემცველი სიმრავლე, ან ცარიელი ინტერვალი.

ასე რომ, ეს ინტერვალია და, უფრო მეტიც, ϕ ფუნქციის ამოზნექილობა ტრივიალურია.)

დავუშვათ, რომ $s \in]0,1[$ და $\vec{w} = \vec{u} + \left(st_1 + (1-s)t_2\right)v$. მაშინ, $\vec{w} = s\vec{u}_1 + (1-s)\vec{u}_2$, სადაც $\vec{u}_1 = \vec{u} + t_1\vec{v}$ და $\vec{u}_2 = \vec{u} + t_2\vec{v}$.

ვინაიდან $t_1 \neq t_2, v_1 \neq v_2$ და $\mathcal{U}$ ამოზნექილია, ამიტომ $w \in \mathcal{U}$. მაშასადამე, $\left(st_1 + (1-s)t_2\right) \in I$. ეს მსჯელობა სამართლიანია ნებისმიერი $s \in]0,1[$ სიდიდისთვის, ამიტომ I არის ინტერვალი. გარდა ამისა, f ფუნქციის ამოზნექილობის გამო,

$$\begin{aligned}\phi\left(st_1 + (1-s)t_2\right) = f\left(s\vec{u}_1 + (1-s)\vec{u}_2\right) &\leq sf(\vec{u}_1) + (1-s)\,f(\vec{u}_2) \\ &= s\phi(t_1) + (1-s)\phi(t_2)\end{aligned}$$

და, ამრიგად, ϕ ფუნქცია ამოზნექილია.

ვინაიდან f არის $\mathcal{C}^2$ კლასის ფუნქცია $\mathcal{U}$ სიმრავლეზე, ხოლო ϕ არის $\mathcal{C}^2$ კლასის ფუნქცია I ინტერვალზე, გვაქვს :

$$\phi''(t) = \vec{v}^{\mathrm{T}}\nabla^2 f\left(\vec{u} + t\vec{v}\right)\vec{v}, \quad \forall t \in I. \tag{13.8}$$

ϕ ფუნქციის ამოზნექილობის გამო და 1.2 თეორემის ძალით, ϕ' წარმოებული ზრდადი ფუნქციაა და ამიტომ $\phi'' \geq 0$. კერძოდ, $\phi'' \geq 0$ უტოლობიდან ცალსახად გამომდინარეობს, რომ $\vec{v}^{\mathrm{T}}\nabla^2 f\left(\vec{u}\right)\vec{v} \geq 0$. ვინაიდან მსჯელობები სამართლიანია ყველა $\vec{u} \in \mathcal{U}$ და ყველა $\vec{v} \in \mathbb{R}^n$ ვითარებაში, ამიტომ $\nabla^2 f\left(\vec{u}\right) \geq 0$.

და პირიქით, დავუშვათ, რომ $\nabla^2 f\left(\vec{u}\right)$ დადებითად ნახევრადგანსაზღვრულია ნებისმიერ $\vec{u} \in \mathcal{U}$ წერტილზე. ავიღოთ ორი $\vec{u}_1, \vec{u}_2 \in \mathcal{U}$ წერტილი, დავუშვათ, რომ $\vec{u} = \vec{u}_2$ და $\vec{v} = \vec{u}_1 - \vec{u}_2$, ასევე განვსაზღვროთ $\phi : t \mapsto f\left(\vec{u} + t\vec{v}\right)$ ფუნქცია ინტერვალზე $I = \{t \in \mathbb{R} \mid \vec{u} + t\vec{v} \in \mathcal{U}\}$. მაშინ, (13.8) განტოლების თანახმად, ϕ'' დადებითია I ინტერვალზე და, ამრიგად, ϕ ფუნქცია ამოზნექილია I ინტერვალზე. დაბოლოს, I შეიცავს $t_1 = 1$ და $t_2 = 0$ წერტილებს. მაშინ ნებისმიერი $s \in]0,1[$ სიდიდისათვის, სრულდება $s = st_1 + (1-s)t_2$ თანაფარდობა და

$$f\left(s\vec{u}_1 + (1-s)\vec{u}_2\right) = \phi(s) \leq s\phi(t_1) + (1-s)\phi(t_2) = s\,f\left(\vec{u}_1\right) + (1-s)f\left(\vec{u}_2\right). \tag{13.9}$$

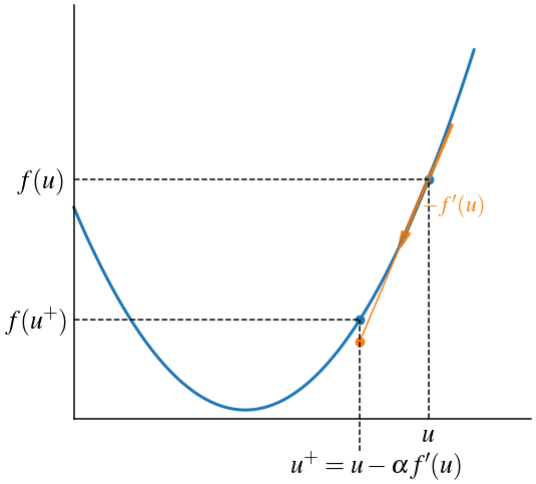
ასე რომ, f ფუნქცია ამოზნექილია, რისი დამტკიცებაც გვინდოდა.

□

3.3 გრადიენტული ალგორითმი

იმ შემთხვევაში, როცა ამოზნექილი ფუნქცია, რომლიც მინიმიზაცია გვაქვს დასახული, დიფერენცირებადია, ყოველთვის ადვილი არ არის მისი გრადიენტის განულების წერტილის

პოვნა. მაშინ შეიძლება მიახლოებითი ამონახსნის პოვნა, თუ გავიხსენებთ, რომ ∇f გრადიენტი f ფუნქციის მაქსიმალური ზრდის მიმართულებას იძლევა. ამრიგად, თუ შემთხვევითად შეირჩევა $\vec{u}$ წერტილი f –ის არეში და $\nabla f \neq 0$, ხოლო t საკმარისად მცირე დადებითი ნამდვილი რიცხვია, მაშინ $\left(\vec{u} - t\nabla f\left(\vec{u}\right)\right)$ უფრო ახლოს იქნება მინიმუმის u^* წერტილთან, ვიდრე u . ეს კონცეფცია ასახულია ნახატზე 13.3.



ნახატი 13.3 – გრადიენტული ალგორითმის იტერაცია გადააადგილებს u -ს $-\alpha f'(u)$ მანძილით.

ამ დაკვირვებიდან გამომდინარეობს *მიმართული დაშვების ალგორითმებად* წოდებული დანაწესების მთელი ოჯახი, რომლებიდანაც ამ განყოფილების დარჩენილ ნაწილში წარმოდგენილია მხოლოდ ზოგიერთი მაგალითი, ვინაიდან ისინი ყველაზე უფრო ხშირად გამოიყენება. ამ ალგორითმებს შორის უმარტივესია *გრადიენტული ალგორითმი*, რომელსაც ზოგჯერ *გრადიენტულ დაშვებასაც* უწოდებენ პირდაპირი ანალოგიით თავისი დასახელების ინგლისურ — *gradient descent* — ვარიანტთან.

განსაზღვრება 1.8 (გრადიენტული ალგორითმი) დავუშვათ, რომ $\mathcal{U} \subseteq \mathbb{R}^n$ და $f : \mathcal{U} \to \mathbb{R}$ არის $\mathcal{C}^1$ კლასის ფუნქცია. როცა მოცემულია $\alpha > 0$ *ბიჯი* და $\epsilon > 0$ *ტოლერანტობა (დასაშვები გადახრა, უზუსტობა)*, შემდეგ ალგორითმს ეწოდება *გრადიენტული ალგორითმი* :

1. შემთხვევითად $\vec{u}$ წერტილის არჩევა $\mathcal{U}$ სიმრავლიდან.
2. ვიდრე სრულდება $\left\|\nabla f\left(\vec{u}\right)\right\|_2^2 > \epsilon$ პირობა, გაგრძელდეს $\vec{u}$ წერტილის განახლება :

$$\vec{u} \leftarrow \vec{u} - \alpha\nabla f\left(\vec{u}\right). \tag{13.10}$$

$\vec{u}$ არის მაშინ f ფუნქციის $\mathcal{U}$ – ზე მყოფი გლობალური მინიმუმის წერტილის აპროქსიმაცია.

რაც უფრო ნაკლებია ტოლერანტობა (დასაშვები გადახრა, უზუსტობა), მით უფრო ახლოსაა რიცხობრივად მინიმუმის წერტილი გლობალური მინიმუმის წერტილთან.

მკითხველთა საყურადღებოდ

α ბიჯის ზომა გრადიენტული ალგორითმის უმნიშვნელოვანესი პარამეტრია. თუ α პარამეტრი ძალიან მცირეა, ალგორითმის კრებადობის მიღწევა ხანგრძლივი პროცესი იქნება. და, პირიქით, თუ α პარამეტრი ძალიან დიდია, მაშინ $\vec{u}$ -ს მნიშვნელობა დაიწყებს რხევას, ოსცილირებას გლობალური მინიმუმის მიდამოებში და ალგორითმი შეიძლება განშლადიც კი აღმოჩნდეს.

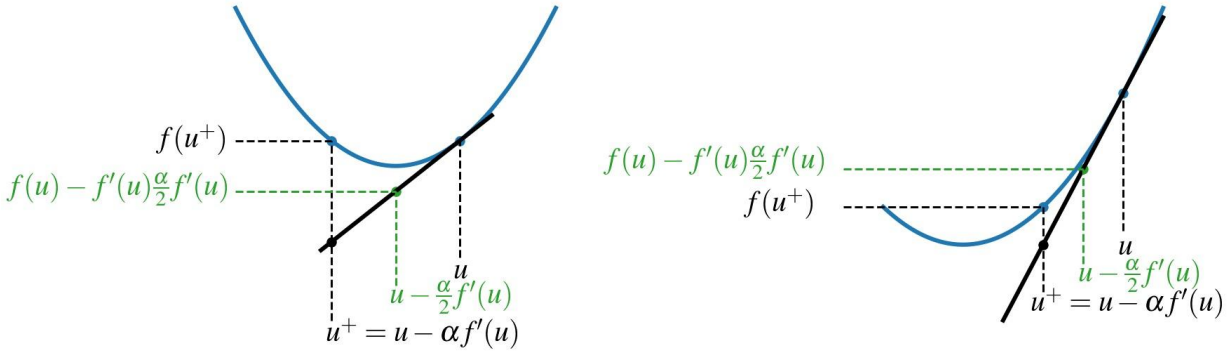
ამიტომ ხშირად გამოიყენება *ადაპტირებადი ზომის ბიჯი*, რომელიც იცვლება ყოველ იტერაციაზე. იგი იწყება შედარებით დიდი მნიშვნელობებით და თანდათანობით მცირდება ამონახსნისადმი მიახლოებისას.

3.4 წრფივი ძებნა განმეორებითი მარშრუტიზაციით

გრადიენტული ალგორითმის ბიჯის ზომის ადაპტაციის ხშირად გამოყენებული მეთოდია *წრფივი ძებნა უკან დაბრუნებით,* ანუ *ბეკტრეკინგი* (ინგლ. backtracking). ეს მეთოდი ეფუძნება 13.4 ნახატზე ნაჩვენებ იმ დაკვირვებას, რომ

$$f\left(\vec{u}-\alpha\nabla f\left(\vec{u}\right)\right)\le f\left(\vec{u}\right)-\frac{\alpha}{2}\nabla f\left(\vec{u}\right)^{\mathrm{T}}\nabla f\left(\vec{u}\right)$$

უტოლობის სამართლიანობისას, სავარაუდოდ, α საკმარისად მცირეა და $\vec{u}$ რჩება ერთსა და იმავე მხარეზე გლობალური მინიმუმის მიმართ ყოველი იტერაციის შემდეგ.



(a) როცა $f\left(u-\alpha f'(u)\right)>f(u)-f'(u)\frac{\alpha}{2}f'(u)$, მაშინ α ბიჯი ძალზე დიდია და $u-\alpha f'(u)$ აღმოჩნდება მინიმუმის წერტილის მიმართ მეორე მხრიდან. ასე რომ, საჭირო გახდება α პარამეტრის შემცირება.

(b) როცა $f\left(u-\alpha f'(u)\right)\le f(u)-f'(u)\frac{\alpha}{2}f'(u)$, მაშინ α ბიჯი საკმარისად მცირეა იმისთვის, რომ $u-\alpha f'(u)$ იყოს მინიმუმის წერტილსა და u – ს შორის.

ნახატი 13.4 – $f\left(u-\alpha f'(u)\right)$ გამოსახულების შედარება $f(u)-f'(u)\frac{\alpha}{2}f'(u)$ გამოსახულებასთან საშუალებას იძლევა განისაზღვროს, თუ არის α პარამეტრის მნიშვნელობა ძალიან დიდი.

განსაზღვრება 1.9 (წრფივი ძებნა უკან დაბრუნებით) დავუშვათ, რომ $\mathcal{U}\subseteq\mathbb{R}^n$ და $f:\mathcal{U}\to\mathbb{R}$ არის $\mathcal{C}^1$ კლასის ფუნქცია. მივიჩნიოთ, რომ *საწყისი ბიჯი* $\alpha>0$, *რედუქციის (დაყვანის, შემცირების) კოეფიციენტი* $\beta\in]0,1[$ და *ტოლერანტობა (მისაღები გადახრა, უზუსტობა)* $\in>0$, მაშინ *წრფივი ძებნა უკან დაბრუნებით* ანუ *BLS ძებნა* (ინგლისური ფრაზიდან *Backtracking Line Search*) ეწოდება შემდეგ ალგორითმს :

1. შემთხვევითად $\vec{u}$ წერტილის არჩევა $\mathcal{U}$ სიმრავლიდან.

2. ვიდრე სრულდება $\left\|\nabla f\left(\vec{u}\right)\right\|_2^2 > \in$ პირობა :

— თუ $f\left(\vec{u}-\alpha\nabla f\left(\vec{u}\right)\right) > f\left(\vec{u}\right)-\frac{\alpha}{2}\nabla f\left(\vec{u}\right)^{\mathrm{T}}\nabla f\left(\vec{u}\right)$, მაშინ შემცირდეს ბიჯი : $\alpha \leftarrow \beta\alpha$;

— განახლდეს $\vec{u} : \vec{u} \leftarrow \vec{u} - \alpha\nabla f\left(\vec{u}\right)$.

მაშინ და მხოლოდ მაშინ $\vec{u}$ არის $\mathcal{U}$-ზე f ფუნქციის გლობალური მინიმუმის წერტილის აპროქსიმაცია (მიახლოება).

3.5 ნიუტონის მეთოდი

როცა სამინიმიზაციო ამოზნექილი f ფუნქცია მიეკუთვნება $\mathcal{C}^2$ კლასს, არსებობს ბიჯის ადაპტაციის გაცილებით უფრო ეფექტური საშუალებები. ნიუტონის მეთოდი ერთ-ერთია მათ შორის და ეფუძნება f ფუნქციის ტეილორის მეორე რიგის დაშლას $\vec{u}$ არგუმენტით. თუ გავითვალისწინებთ $\vec{u} \in \mathcal{U}$ თანაფარდობას, შეგვიძლია $g : \mathcal{U} \to \mathbb{R}$ ფუნქცია განვსაზღვროთ შემდეგი სახით :

$$g\left(\vec{v}\right) = f\left(\vec{u}\right) + \nabla f\left(\vec{u}\right)^{\mathrm{T}}\left(\vec{v}-\vec{u}\right) + \frac{1}{2}\left(\vec{v}-\vec{u}\right)^{\mathrm{T}}\nabla^2 f\left(\vec{u}\right)\left(\vec{v}-\vec{u}\right). \tag{13.11}$$

g ფუნქციის მინიმიზაციისთვის საკმარისია უბრალოდ მისი გრადიენტის განულება

$$\nabla g\left(\vec{v}\right) = \nabla f\left(\vec{u}\right) + \nabla^2 f\left(\vec{u}\right)\left(\vec{v}-\vec{u}\right). \tag{13.12}$$

და, ამრიგად, დადგინდება, რომ g ფუნქცია მინიმალურია წერტილზე

$$\vec{v}^* = \vec{u} - \left(\nabla^2 f\left(\vec{u}\right)\right)^{-1}\nabla f\left(\vec{u}\right). \tag{13.13}$$

ამრიგად, ნიუტონის მეთოდი გულისხმობს $\alpha = \left(\nabla^2 f\left(\vec{u}\right)\right)^{-1}$ სიდიდის გამოყენებას ბიჯის სახით. ეს მეთოდი გულისხმობს, რომ ჰესიანი ინვერტირებადია, რასაც, კერძოდ, ადგილი აქვს ძლიერ ამოზნექილი ფუნქციებისთვის. თუ ეს ასე არ არის, მაშინ ჩვენ შეგვიძლია ცოტა ხმაურის დამატება ჰესიანში ($\in \cdot I_n$ შესაკრების შეტანით მცირე დადებითი $\in > 0$ სიდიდით), იმისათვის, რომ ვაქციოთ იგი ინვერტირებად ფორმად.

განსაზღვრება 1.10 (ნიუტონის მეთოდი) დავუშვათ, რომ $\mathcal{U} \subseteq \mathbb{R}^n$ და $f : \mathcal{U} \to \mathbb{R}$ არის $\mathcal{C}^1$ კლასის ფუნქცია. ასევე მივიჩნიოთ, რომ მოცემულია $\in > 0$ *ტოლერანტობა* (ანუ დასაშვები გადახრა, უზუსტობა). მაშინ *ნიუტონის მეთოდი* შემდეგ ალგორითმს ეწოდება :

1. შემთხვევითად ხდება $\vec{u}$ წერტილის არჩევა $\mathcal{U}$ სიმრავლეზე.

2. ვიდრე სრულდება $\left\|\nabla f\left(\vec{u}\right)\right\|_2^2 > \in$ პირობა :

— გამოითვლება ბიჯი : $\alpha = \left(\nabla^2 f\left(\vec{u}\right)\right)^{-1}$.

— ხდება $\vec{u}$ წერტილის განახლება : $\vec{u} \leftarrow \vec{u} - \alpha \nabla f\left(\vec{u}\right)$.

მაშინ $\vec{u}$ არის $\mathcal{U}$-ზე f ფუნქციის გლობალური მინიმუმის წერტილის აპროქსიმაცია (მიახლოება).

3.6 ნიუტონის მეთოდი შეუღლებული გრადიენტით

ნიუტონის მეთოდი განსახორციელებლად შეიძლება იყოს ზორზოხი, უშველებელი გრადიენტის შექცეული სიდიდის გამოთვლის აუცილებლობის გამო.

ამის გამოსასწორებლად, *ნიუტონის მეთოდი შეუღლებული გრადიენტით* მდგომარეობს δ ამონახსნის სიდიდის გამოთვლაში ყოველ ეტაპზე შემდეგი ტოლობიდან :

$$\nabla^2 f\left(\vec{u}\right)\delta = \nabla f\left(\vec{u}\right). \tag{13.14}$$

მაშინ $\vec{u}$ წერტილის განახლების წესი გრადიენტულ ალგორითმში იძენს $\vec{u} \leftarrow \vec{u} - \delta$ სახეს.

განტოლება 13.14 წარმოადგენს $A\vec{x} - \vec{b} = 0$ სახის ამოცანას, სადაც $A \geq 0$ (1.4 თეორემის თანახმად). სწორედ ამ განტოლების ამოხსნა ხდება ეგრეთ წოდებული შეუღლებული გრადიენტის მეთოდით, რომელიც შემოთავაზებული იყო 1950-ან წლებში კორნელიუს ლანცოშის (Cornelius Lanczos), ედუარდ შტიფელის (Eduard Stiefel) და მაგნუს ჰესტენესის (Magnus Hestenes) მიერ ნაშრომში (Hestenes and Stiefel, 1952).

ამ მეთოდის ცენტრალური იდეა $\mathbb{R}^n$ ბაზისის აგება არის, რომელიც შედგენილია A მატრიცის მიმართ შეუღლებული ვექტორებისგან :

$$\left\{\vec{v}_1, \vec{v}_2, \ldots, \vec{v}_n\right\}, \text{ სადაც } \vec{v}_i^{\mathrm{T}} A \vec{v}_j = 0 \ \ \forall i \neq j\,. \tag{13.15}$$

განსაზღვრება 1.11 (შეუღლებული გრადიენტის მეთოდი) თუ მხედველობაში მივიღებთ დადებით ნახევრად უსასრულო $A \in \mathbb{R}^{n \times n}$ მატრიცასა და $b \in \mathbb{R}^n$ ვექტორს, *შეუღლებული გრადიენტის მეთოდი* (ინგლ. *conjugate gradient*) შემდეგი ალგორითმის ფორმით ჩაიწერება :

1. ინიციალიზაცია :
   — $\vec{x}^{(0)} \in \mathbb{R}^n$ საწყისი წერტილის შემთხვევითად არჩევა.
   — $\vec{r}_0 = \vec{v}_0 = \vec{b} - A\vec{x}^{(0)}$ საწყისი მნიშვნელობის განსაზღვრა.
2. $t = 1, \ldots, n$ მნიშვნელობებისთვის :
   (a) $\vec{x}$-ის განახლება :

$$\vec{x}^{(t)} = \vec{x}^{(t-1)} + \frac{\vec{r}_{t-1}^{\mathrm{T}} \vec{r}_{t-1}}{\vec{v}_{t-1}^{\mathrm{T}} A \vec{v}_{t-1}} \vec{v}_{t-1}\,. \tag{13.16}$$

   (b) ნაშთის განახლება :

$$\vec{r}_t = \vec{b} - A\vec{x}^{(t)}\,. \tag{13.17}$$

(c) $\vec{v}$ -ის განახლება :

$$\vec{v}_t = \vec{r}_t + \frac{\vec{r}_t^{\mathrm{T}}\vec{r}_t}{\vec{r}_{t-1}^{\mathrm{T}}\vec{r}_{t-1}}\vec{v}_{t-1}\,. \tag{13.18}$$

$\vec{x}^{(n)}$ ამონახსნია, რომელიც იძებნებოდა.

❁

თეორემა 1.5 *შეუღლებული გრადიენტის მეთოდი ამტკიცებს, რომ* $\vec{v}_i^{\mathrm{T}} A\vec{v}_j = 0 \;\; \forall i \neq j$.

❁

მტკიცებულება. დავუშვათ, რომ $\alpha_t = \frac{\vec{r}_{t-1}^{\mathrm{T}}\vec{r}_{t-1}}{\vec{v}_{t-1}^{\mathrm{T}}A\vec{v}_{t-1}}$ და $\beta_t = \frac{\vec{r}_t^{\mathrm{T}}\vec{r}_t}{\vec{r}_{t-1}^{\mathrm{T}}\vec{r}_{t-1}}$.

დავიწყოთ იმით, რომ ჩავწეროთ ზოგიერთი ზემოთ მოყვანილი განმარტება გარკვეულად შეცვლილი სახით გამოსაყენებლად მოგვიანებით. ჩავანაცვლოთ $\vec{x}^{(t)}$ განტოლებაში (13.17) მისი განსაზღვრებით (13.16). მაშინ მივიღებთ, რომ $\vec{r}_t = \vec{b} - A\vec{x}^{(t-1)} - \alpha_t A\vec{v}_{t-1}$ და, ამრიგად, გვექნება :

$$\vec{r}_t = \vec{r}_{t-1} - \alpha_t A\vec{v}_{t-1}. \tag{13.19}$$

ან, ამის ტოლფასი ფორმით,

$$A\vec{v}_{t-1} = \frac{1}{\alpha_t}\left(\vec{r}_{t-1} - \vec{r}_t\right). \tag{13.20}$$

გარდა ამისა, (13.18) განტოლების თანახმად

$$\vec{r}_t = \vec{v}_t - \beta\vec{v}_{t-1}. \tag{13.21}$$

დაბოლოს, α_t და β_t განმარტებათა შესაბამისად,

$$\beta_t\vec{v}_{t-1}^{\mathrm{T}}A\vec{v}_{t-1} = \frac{\vec{r}_t^{\mathrm{T}}\vec{r}_t}{\alpha_t}. \tag{13.22}$$

ახლა რეკურსიის გამოყენებით ჩვენ ვუჩვენებთ, რომ ნებისმიერი $t = 1,\ldots,n$ და ნებისმიერი $i = 1,\ldots,t$ მნიშვნელობებისთვის ადგილი აქვს $\vec{r}_{t-i} = 0$ და $\vec{v}_t^{\mathrm{T}}A\vec{v}_{t-i} = 0$ თანაფარდობებს.

დავიწყოთ $t = 1$ (და, ამრიგად, $i = 1$) შემთხვევით. საჭიროა ვაჩვენოთ, რომ

$$\vec{r}_1^{\mathrm{T}}\vec{r}_0 = 0$$

და

$$\vec{v}_1^{\mathrm{T}}A\vec{v}_0 = 0\,.$$

(13.19) განტოლების თანახმად,

$$\vec{r}_1^{\,\mathrm{T}}\vec{r}_0 = \vec{r}_0^{\,\mathrm{T}}\vec{r}_0 - \alpha_1 A\vec{v}_0^{\,\mathrm{T}}\vec{r}_0 = \vec{r}_0^{\,\mathrm{T}}\vec{r}_0\left(1 - \frac{\vec{v}_0^{\,\mathrm{T}}\vec{r}_0}{\vec{v}_0^{\,\mathrm{T}}\vec{v}_0}A\right).$$

ვინაიდან $\vec{r}_0 = \vec{v}_0$, ზემოთ მოყვანილი წილადი არის 1 და ჩვენ ვიღებთ $\vec{r}_1^{\,\mathrm{T}}\vec{r}_0 = 0$ ტოლობას.

$\vec{v}_1$ – ის განმარტების თანახმად განტოლებიდან (13.18), გვაქვს, რომ $\vec{v}_1^{\,\mathrm{T}}A\vec{v}_0 = \vec{r}_1^{\,\mathrm{T}}A\vec{v}_0 + \beta_1\vec{v}_0^{\,\mathrm{T}}A\vec{v}_0$. (13.20) განტოლებიდან კი $\vec{r}_1^{\,\mathrm{T}}A\vec{v}_0 = \frac{1}{\alpha_1}\vec{r}_1^{\,\mathrm{T}}\left(\vec{r}_0 - \vec{r}_1\right)$. ხოლო (13.22) განტოლებიდან გამომდინარეობს, რომ $\beta_1\vec{v}_0^{\,\mathrm{T}}A\vec{v}_0 = \frac{\vec{r}_1^{\,\mathrm{T}}\vec{r}_1}{\alpha_1}$. ამრიგად, $\vec{v}_1^{\,\mathrm{T}}A\vec{v}_0 = \frac{\vec{r}_1^{\,\mathrm{T}}\vec{r}_0}{\alpha_1} = 0$, ვინაიდან ჩვენ მიერ ახლახან იყო ნაჩვენები, რომ $\vec{r}_1^{\,\mathrm{T}}\vec{r}_0 = 0$. ამით $t = 1$ შემთხვევის განხილვა დასრულებულია.

ახლა კი დავუშვათ, რომ $t > 1$ და რომ ყველა $u < t$ და ყველა $i = 1,\ldots,u$ მნიშვნელობისთვის სრულდება $\vec{r}_u^{\,\mathrm{T}}\vec{r}_{u-i} = 0$ და $\vec{v}_u^{\,\mathrm{T}}A\vec{v}_{u-i} = 0$ თანაფარდობები.

უპირველეს ყოვლისა ვაჩვენოთ, რომ $\vec{r}_t^{\,\mathrm{T}}\vec{r}_{t-i} = 0$ ყველა $i = 1,\ldots,t$ სიდიდისთვის. თუ გამოვიყენებთ (13.19) განტოლებას და იმ ფაქტსაც, რომ $A = A^{\mathrm{T}}$ (რადგან $A \geq 0$), მაშინ მივიღებთ :

$$\vec{r}_t^{\,\mathrm{T}}\vec{r}_{t-i} = \vec{r}_{t-1}^{\,\mathrm{T}}\vec{r}_{t-i} - \alpha_t\vec{v}_{t-1}^{\,\mathrm{T}}A\vec{r}_{t-i}. \tag{13.23}$$

შესაძლებელია სამი შემთხვევა :

– თუ t ცვლადს შევცვლით 1-ით, ხოლო α_t ცვლადს — მისი განსაზღვრებით (13.23) განტოლებაში, მაშინ გვექნება :

$$\vec{r}_t^{\,\mathrm{T}}\vec{r}_{t-1} = \vec{r}_{t-1}^{\,\mathrm{T}}\vec{r}_{t-1}\left(1 - \frac{\vec{v}_{t-1}^{\,\mathrm{T}}A\vec{r}_{t-1}}{\vec{v}_{t-1}^{\,\mathrm{T}}A\vec{v}_{t-1}}\right).$$

(13.21) განტოლების თანახმად,

$$\vec{v}_{t-1}^{\,\mathrm{T}}A\vec{r}_{t-1} = \vec{v}_{t-1}^{\,\mathrm{T}}A\vec{v}_{t-1} - \beta_{t-1}\vec{v}_{t-1}^{\,\mathrm{T}}A\vec{v}_{t-2}.$$

$\vec{v}_{t-1}^{\,\mathrm{T}}A\vec{v}_{t-2} = 0$ და, მაშასადამე, $\vec{v}_{t-1}^{\,\mathrm{T}}A\vec{r}_{t-1} = \vec{v}_{t-1}^{\,\mathrm{T}}A\vec{v}_{t-1}$ რეკურენტულობის ჰიპოთეზებიდან გამომდინარე, ზემოთ მოყვანილი $\frac{\vec{v}_{t-1}^{\,\mathrm{T}}A\vec{r}_{t-1}}{\vec{v}_{t-1}^{\,\mathrm{T}}A\vec{v}_{t-1}}$ წილადი არის 1, რაც გვაძლევს საშუალებას დავასკვნათ, რომ $\vec{r}_t^{\,\mathrm{T}}\vec{r}_{t-1} = 0$.

– თუ $1 < i < t$, ჩვენ შეგვიძლია შევცვალოთ $\vec{r}_{t-i}$ (13.23) განტოლების მეორე პოზიციაში თავისი მნიშვნელობით (13.21) თანაფარდობიდან და, ამრიგად, მივიღოთ :

$$\vec{r}_t^{\,\mathrm{T}}\vec{r}_{t-i} = \vec{r}_{t-1}^{\,\mathrm{T}}\vec{r}_{t-i} - \alpha_t\vec{v}_{t-1}^{\,\mathrm{T}}A\vec{v}_{t-i} - \alpha_t\beta_{t-i}\vec{v}_{t-1}^{\,\mathrm{T}}A\vec{v}_{t-i-1}.$$

თითოეული ამ ჯამის წევრებიდან უდრის ნულს ჩვენი რეკურენტული დაშვებების (ჰიპოთეზების) პირობებში, ამიტომ შეგვიძლია დავასკვნათ, რომ $\vec{r}_{t-1}^{\,\mathrm{T}}\vec{r}_{t-i} = 0$.

– დაბოლოს, თუ $i=t$, მაშინ საქმე ეხება $\vec{r}_{t-1}^{\mathrm{T}}\vec{r}_0-\alpha_t\vec{v}_{t-1}^{\mathrm{T}}A\vec{r}_0$ თანაფარდობის შესწავლას. ვინაიდან $\vec{r}_0=\vec{v}_0$, ამ ჯამის ორივე წევრი ნულის ტოლია რეკურენტულობის ჩვენი დაშვებების თანახმად და $\vec{r}_t^{\mathrm{T}}\vec{r}_0=0$.

ახლა ჩვენ შეგვიძლია ყურადღება გადავიტანოთ $\vec{v}_t^{\mathrm{T}}A\vec{v}_{t-i}$ გამოსახულებაზე. (13.18) განტოლების თანახმად, გვექნება :

$$\vec{v}_t^{\mathrm{T}}A\vec{v}_{t-i}=\vec{r}_t^{\mathrm{T}}A\vec{v}_{t-i}+\beta_t\vec{v}_{t-1}^{\mathrm{T}}A\vec{v}_{t-i}. \tag{13.24}$$

(13.20) განტოლების შესაბამისად, ამ ჯამის პირველი წევრი შეიძლება შემდეგი სახით ჩაიწეროს :

$$\vec{r}_t^{\mathrm{T}}A\vec{v}_{t-i}=\frac{1}{\alpha_{t-i+1}}\vec{r}_t^{\mathrm{T}}\left(\vec{r}_{t-i}-\vec{r}_{t-i+1}\right).$$

აქ ორი შემთხვევაა შესაძლებელი :

– თუ $i>1$, მაშინ $\vec{r}_t^{\mathrm{T}}\vec{r}_{t-i}$ და $\vec{r}_t^{\mathrm{T}}\vec{r}_{t-i+1}$ ორივე ნულის ტოლია, როგორც ეს ახლახან ვაჩვენეთ ამ მტკიცებულების პირველ ნაწილში. $\vec{v}_{t-1}^{\mathrm{T}}A\vec{v}_{t-i}$ გამოსახულება ასევე ნულს უდრის რეკურენტულობის ჩვენი ჰიპოთეზების თანახმად.

– თუ $i=1$, მაშინ ჯამის პირველი წევრი (13.24) განტოლებაში იქნება $\vec{r}_t^{\mathrm{T}}A\vec{v}_{t-1}=-\frac{\vec{r}_t^{\mathrm{T}}\vec{r}_t}{\alpha_t}$. მეორე წევრი მოცემულია უშუალოდ (13.22) განტოლებით და პირველის საწინააღმდეგო იქნება.

ამრიგად, დამოუკიდებლად იმისა, თუ როგორია i, სრულდება $\vec{v}_t^{\mathrm{T}}A\vec{v}_{t-i}=0$ ტოლობა.

□

შენიშვნა

$\mathbb{R}^n$ ბაზისის მაფორმირებელი $\vec{v}_t$ ვექტორები შეუღლებული გრადიენტის მეთოდის $n-$ ური იტერაციის შემდეგ, აკმაყოფილებს $\vec{v}_t=0$ თანაფარდობებს.

3.7 კვაზინიუტონური მეთოდები

შესაძლოა, რომ ჰესიანის გამოთვლამ დიდი რესურსები მოითხოვოს. ასეთ შემთხვევაში გამოიყენება ეგრეთ წოდებული *კვაზი ნიუტონის მეთოდები*, რომლებიც გერმანელი მათემატიკოსის ოტო ჰესეს (Ludwig Otto Hesse 1811–1874) მატრიცის შექცევის ჩანაცვლების საშუალებას იძლევა გარკვეული მიახლოებით ნიუტონის მეთოდში. ეს აპროქსიმაცია იტერაციის გზით გამოითვლება.

განსაზღვრება 1.12 (კვაზი-ნიუტონის მეთოდი) დავუშვათ, რომ $\mathcal{U}\subseteq\mathbb{R}^n$ და $f:\mathcal{U}\to\mathbb{R}$ არის $\mathcal{C}^1$ კლასის ფუნქციები. მაშინ $\epsilon>0$ ტოლერანტობის (დასაშვები გადახრის) გათვალისწინებით *კვაზი ნიუტონის მეთოდს* უწოდებენ შემდეგი სახის ალგორითმს :

1. $\vec{u}$ ვექტორის შემთხვევითი არჩევა $\mathcal{U}$ სივრცეში.
2. ჰესიანის განზომილების მქონე იგივეობის $W : W^{(0)} = I$ მატრიცის ინიციალიზაცია.
3. იმ დრომდე, სანამ $\left\|\nabla f\left(\vec{u}\right)\right\|_2^2 > \in$:
   - $t : t \leftarrow \left(t+1\right)$ ნაზრდის მიცემა,
   - $\vec{u} : \vec{u}^{(t)} = \vec{u}^{(t)} - W^{(t-1)}$ განახლება,
   - $\left(\nabla^2 f\left(\vec{u}\right)\right)^{-1}$ გამოსახულების W აპროქსიმაციის განახლება.

მაშინ $\vec{u}^{(t)}$ არის f ფუნქციის გლობალური მინიმუმის წერტილის აპროქსიმაცია $\mathcal{U}$ სივრცეში.

შექცეული $\nabla^2 f\left(\vec{u}\right)^{-1}$ ჰესიანის აპროქსიმაციისათვის იძებნება $W^{(t)}$ მატრიცა, რომელიც არის :

- სიმეტრიული და დადებითად ნახევრადგანსაზღვრული;
- წინა აპროქსიმაციასთან მიახლოებული, ე.ი. ისეთი, რომ $\left\|W^{(t)} - W^{(t-1)}\right\|_F$ მინიმალური იყოს (აქ $\left\|...\right\|_F$ აღნიშნავს მატრიცის ფრობენიუსის, ანუ, რაც იგივეა, ევკლიდეს, ნორმას, სხვანაირად რომ ვთქვათ, კვადრატულ ფესვს ამ მატრიცის ელემენტთა კვადრატების ჯამიდან);
- ისეთი, რომ $W^{(t)}\left(\nabla f\left(\vec{u}^{(t)}\right) - \nabla f\left(\vec{u}^{(t-1)}\right)\right) = \vec{u}^{(t)} - \vec{u}^{(t-1)}$, ეს უკანასკნელი პირობა, რომელსაც *მკვეთის განტოლება* ეწოდება, უშუალოდ გამომდინარეობს ∇f ფუნქციის ტეილორის მწკრივად პირველი რიგის დაშლიდან $\vec{u}^{(t-1)}$ წერტილზე.

ერთ-ერთი ყველაზე ფართოდ გამოყენებული დღეს კვაზი ნიუტონის მეთოდებს შორის და იმავდროულად განსაკუთრებულად წარმატებული არის *BFGS* მეთოდი, რომელიც ასეთ სახელწოდებას ატარებს ჩარლზ ჯორჯ ბროიდენის (Charles George **B**royden), როჯერ ფლეტჩერის (Roger **F**letcher), დონალდ გოლდფარბის (Donald **G**oldfarb) და დევიდ შანოს (David **S**hanno) საპატივცემულოდ, ვინაიდან სწორედ ამ პირებმა წარმოადგინეს ერთმანეთისგან დამოუკიდებლად იგი 1970 წელს.

ამ მეთოდში, შექცეული ჰესიანის (ჰესეს შექცეული მატრიცის) იტერაციული აპროქსიმაცია მოიცემა ფორმულით

$$W^{(t)} = W^{(t-1)} - \frac{d_t \delta_t^{\mathrm{T}} W^{(t-1)} + W^{(t-1)} \delta_t d_t^{\mathrm{T}}}{\langle \delta_t, d_t \rangle} + \left(1 + \frac{\langle \delta_t, W^{(t-1)} \delta_t \rangle}{\langle \delta_t, d_t \rangle}\right) \frac{d_t d_t^{\mathrm{T}}}{\langle \delta_t, d_t \rangle}, \qquad (13.25)$$

სადაც $d_t = \vec{u}^{(t)} - \vec{u}^{(t-1)}$ და $\delta_t = \nabla f\left(\vec{u}^{(t)}\right) - \nabla f\left(\vec{u}^{(t-1)}\right)$.

არსებობს მეხსიერების ტევადობისადმი ნაკლებად მომთხოვნი ვერსია, რომელსაც L-BFGS ეწოდება (Limited-memory BFGS ფრაზიდან გამომდინარე, ე.ი. *BFGS მეთოდი შეზღუდული მეხსიერებით*).

BFGS მეთოდის ხსენებული L-BFGS ვერსია საშუალებას იძლევა თავიდან ავიცილოთ მეხსიერებაში მთლიანი $W^{(t)}$ მატრიცის შენახვის აუცილებლობა.

3.8 სტოქასტიკური გრადიენტის ალგორითმი

იმ შემთხვევაში, როცა n რიცხვი ძალიან დიდია, თავად გრადიენტი შეიძლება გახდეს გამოსათვლელად ძვირად ღირებული. მაგრამ სტატისტიკურ სწავლებაში ხშირად გვხვდება მინიმიზაციისთვის განკუთვნილი ისეთი ფუნქციები, რომლებიც შეიძლება დაიშალოს და ჩაიწეროს n უფრო მარტივი ფუნქციის ჯამის სახით :

$$f\left(\vec{u}\right)=\sum_{i=1}^{n} f_i\left(\vec{u}\right). \tag{13.26}$$

მაშინ მისი $\nabla f\left(\vec{u}\right)$ გრადიენტი ასევე განიცდის დეკომპოზიციას შესაკრებ ფუნქციათა $\nabla f_i\left(\vec{u}\right)$ გრადიენტების ჯამად :

$$\nabla f\left(\vec{u}\right)=\sum_{i=1}^{n} \nabla f_i\left(\vec{u}\right). \tag{13.27}$$

მაგალითი

ეს ხდება მაშინ, მაგალითად, როცა გვჭირდება უმცირეს კვადრატთა ჯამის მინიმიზაცია (იხ. მეხუთე ლექციის პუნქტი 1.2). ხსენებული ჯამი ასეთი სახით ჩაიწერება :

$$f\left(\vec{\beta}\right)=\sum_{i=1}^{n}\left(y^i-\phi\left(\vec{x}^i \mid \vec{\beta}\right)\right)^2,$$

სადაც ϕ არის მაპროგნოზირებელი (პროგნოზის, წინასწარმეტყველების) ფუნქცია.

სტოქასტიკური გრადიენტული ალგორითმი გამოთვლათა დაჩქარების საშუალებას იძლევა ყოველ იტერაციაზე მხოლოდ ერთ-ერთი, შემთხვევითად არჩეული, f_i ფუნქციის გამოყენებით : $\sum_{i=1}^{n}\nabla f_i\left(\vec{u}\right)$ ჩანაცვლდება $\nabla f_k\left(\vec{u}\right)$ გრადიენტით.

განსაზღვრება 1.13 (სტოქასტიკური გრადიენტული ალგორითმი) დავუშვათ, რომ $\mathcal{U}\subseteq\mathbb{R}^n$ და $f:\mathcal{U}\to\mathbb{R}$ არის $\mathcal{C}^1$ კლასის ფუნქცია, რომლის დაშლა შემდეგი სახით ხდება :

$$f\left(\vec{u}\right)=\sum_{i=1}^{n} f_i\left(\vec{u}\right). \tag{13.28}$$

მოცემული $\alpha>0$ *ბიჯისა* და $\epsilon>0$ *ტოლერანტობის (დასაშვები გადახრის)* გათვალისწინებით, *სტოქასტიკური გრადიენტული ალგორითმი* ეწოდება შემდეგ ალგორითმს :

1. $\vec{u}$ ვექტორის შემთხვევითი არჩევა $\mathcal{U}$ სივრცეში.
2. იმ დრომდე, სანამ $\left\|\nabla f\left(\vec{u}\right)\right\|_2^2 > \in$:
   — $\{1,2,\ldots,n\}$ რიცხვთა შორის რომელიღაც k რიცხვის შემთხვევითად არჩევა.
   — $\vec{u}:\vec{u}\leftarrow\vec{u}-\alpha\nabla f_k\left(\vec{u}\right)$ განახლება.

მაშინ $\vec{u}$ არის f ფუნქციის გლობალური მინიმუმის წერტილის აპროქსიმაცია $\mathcal{U}$ სივრცეში.

ცხადია, რომ ამავე კონტექსტში შეიძლება იყოს გამოყენებული ადაპტირებადი ბიჯიც.

3.9 კოორდინატული დაშვება

ზემოთ განხილული მეთოდები გულისხმობს, რომ ფუნქცია, რომლის მინიმიზაცია უნდა მოხდეს, დიფერენცირებადია. ანალოგიური მეთოდი შეიძლება იყოს გამოყენებული არადიფერენცირებადი ფუნქციის შემთხვევაშიც, თუ შესაძლებელია მისი დაშლა და წარმოდგენა ჯამის სახით, სადაც მოცემული იქნება $\mathcal{C}^1$ კლასის ერთი ამოზნექილი ფუნქცია და სხვა n ფუნქცია — თითოეული მხოლოდ ერთი კოორდინატით ამოზნექილი. შემდეგ ფუნქციის მინიმიზაცია ყოველი ცალკეული კოორდინატით ხდება.

განსაზღვრება 1.14 (კოორდინატული დაშვება) დავუშვათ, რომ $\mathcal{U}\subseteq\mathbb{R}^n$ და $f:\mathcal{U}\to\mathbb{R}$ არის შემდეგი სახის ფუნქცია :

$$f:\vec{u}\mapsto g\left(\vec{u}\right)+\sum_{i=1}^{n}h_i\left(u_i\right),$$

სადაც g არის $\mathcal{C}^1$ კლასის რაღაც ამოზნექილი ფუნქცია და n რაოდენობის ყველა h_i $(i=1,2,\ldots,n)$ ფუნქციაც ასევე ამოზნექილია.

კოორდინატული დაშვება (ინგლ. *coordinate descent*) ეწოდება შემდეგ ალგორითმს :

1. $\vec{u}$ ვექტორის შემთხვევითი არჩევა $\mathcal{U}$ სივრცეში.
2. იმ დრომდე, სანამ $\left\|\nabla f\left(\vec{u}\right)\right\|_2^2 > \in$, $\vec{u}$ ვექტორის განახლება :
   — $u_1^{(t)}$ არის $u\mapsto f\left(u,u_2^{(t-1)},\ldots,u_n^{(t-1)}\right)$-ის მინიმალური წერტილი ;
   — $u_2^{(t)}$ არის $u\mapsto f\left(u_1^{(t-1)},u,\ldots,u_n^{(t-1)}\right)$-ის მინიმალური წერტილი ;
   — …
   — $u_n^{(t)}$ არის $u\mapsto f\left(u_1^{(t-1)},u_2^{(t-1)},\ldots,u\right)$-ის მინიმალური წერტილი.

მაშინ $\vec{u}$ არის f ფუნქციის გლობალური მინიმუმის წერტილის აპროქსიმაცია $\mathcal{U}$ სივრცეში.

4 ამოზნექილი ოპტიმიზაცია შეზღუდვათა არსებობისას

ამ ნაწილში ჩვენ შევჩერდებით შემდეგი სახის ოპტიმიზაციის ამოცანებზე :

$$(\mathcal{P}):\quad \left.\begin{array}{ll} \min_{\vec{u}\in\mathcal{U}} & f(\vec{u}) \text{ შეზღუდვებისას} \\ & g_i(\vec{u}) \le 0 \quad \forall i = 1,\dots,m \\ & h_j(\vec{u}) = 0 \quad \forall j = 1,\dots,r \end{array}\right\}, \tag{13.29}$$

სადაც f, g_i, h_j ფუნქციები მიჩნეულია $\mathcal{C}^1$ კლასის ნამდვილ სიდიდეებად.

ასეთი ტიპის ამოცანათა ამოსახსნელად მრავალი მეთოდი არსებობს. ქვემოთ ჩვენ დაწვრილებით განვიხილავთ ლაგრანჟის მამრავლების მეთოდს.

4.1 ლაგრანჟიანი

$(\mathcal{P})$ პრობლემის გადასაჭრელად, ჩვენ ახლა შემოვიტანთ მის *ლაგრანჟიანს.*

განსაზღვრება 1.15 (ლაგრანჟიანი) დავუშვათ, რომ $(\mathcal{P})$ არის მინიმიზაციის პრობლემა (13.29) ფორმით მოცემული შეზღუდვებით. $(\mathcal{P})$ ამოცანის *ლაგრანჟიანი* ეწოდება ფუნქციას

$$L : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^r \to \mathbb{R}$$
$$\vec{u}, \vec{\alpha}, \vec{\beta} \mapsto f(\vec{u}) + \sum_{i=1}^{m} \alpha_i g_i(\vec{u}) + \sum_{j=1}^{r} \beta_j h_j(\vec{u}).$$

$\alpha_1, \alpha_2, \dots \alpha_m, \beta_1, \beta_2, \dots, \beta_r$ ცვლადებს უწოდებენ *ლაგრანჟის მამრავლებს* ან, სხვანაირად, *დუალურ ცვლადებსაც.*

ლაგრანჟიანი საშუალებას გვაძლევს განვსაზღვროთ *ლაგრანჟის დუალურ ფუნქციად* წოდებული ფუნქცია, რომლის მაქსიმუმი $(\mathcal{P})$ პრობლემის ამონახსნისთვის ქვედა საზღვარს იძლევა.

განსაზღვრება 1.16 (ლაგრანჟის დუალური ფუნქცია) დავუშვათ, რომ $(\mathcal{P})$ არის მინიმიზაციის პრობლემა (13.29) ფორმით მოცემული შეზღუდვების პირობებში, ხოლო L — მისი ლაგრან-ჟიანი. მაშინ *ლაგრანჟის დუალური ფუნქცია* ეწოდება შემდეგ ფუნქციას :

$$Q : \mathbb{R}^m \times \mathbb{R}^r \to \mathbb{R}$$
$$\vec{\alpha}, \vec{\beta} \mapsto \inf_{\vec{u}\in\mathcal{U}} L\left(\vec{u}, \vec{\alpha}, \vec{\beta}\right).$$

L ლაგრანჟიანის $\inf_{\vec{u}\in\mathcal{U}} L\left(\vec{u}, \vec{\alpha}, \vec{\beta}\right)$ ინფიმუმი (*ზუსტი ქვედა საზღვარი*, ანუ *უდიდესი ქვედა საზღვარი*) $\vec{u}$ ვექტორით არის ისეთი $q^* \in \mathbb{R}$ უდიდესი მნიშვნელობა, რომელიც $q^* \le L\left(\vec{u}, \vec{\alpha}, \vec{\beta}\right)$ პირობას აკმაყოფილებს ნებისმიერი $u \in \mathcal{U}$-თვის და პოტენციურად იგი $(-\infty)$-ის ტოლიც კი შეიძლება იყოს.

✿

ლაგრანჟის დუალურ ფუნქციას ის უპირატესობა აქვს, რომ იგი ჩაზნექილ ფუნქციად რჩება $(\mathcal{P})$ პრობლემის ამოზნექილობისგან დამოუკიდებლად.

თეორემა 1.6 *(13.29) სახის შეზღუდვებიანი მინიმიზაციის ამოცანის ლაგრანჟის დუალური ფუნქცია ჩაზნექილია.*

✿

მტკიცებულება. დავუშვათ, რომ მოცემულია $\left(\vec{\alpha}_1, \vec{\beta}_1\right)$, $\left(\vec{\alpha}_2, \vec{\beta}_2\right)$ და $0 \le \lambda \le 1$. ასევე ავიღოთ $\vec{\alpha} = \lambda\vec{\alpha}_1 + (1-\lambda)\vec{\alpha}_2$ და $\vec{\beta} = \lambda\vec{\beta}_1 + (1-\lambda)\vec{\beta}_2$. ასე რომ $Q\left(\vec{\alpha}, \vec{\beta}\right) = \inf\limits_{\vec{u}\in\mathcal{U}} f(\vec{u}) + \alpha^{\mathrm{T}}\vec{g} + \vec{\beta}^{\mathrm{T}}\vec{h}$, სადაც $\vec{g} = \left(g_1(\vec{u}), g_2(\vec{u}), \ldots, g_m(\vec{u})\right)$ და $\vec{h} = \left(h_1(\vec{u}), h_2(\vec{u}), \ldots, h_r(\vec{u})\right)$. ამრიგად,

$$\begin{aligned} Q\left(\vec{\alpha}, \vec{\beta}\right) &= \inf_{\vec{u}\in\mathcal{U}} f(\vec{u}) + \lambda\vec{\alpha}_1^{\mathrm{T}}\vec{g} + (1-\lambda)\vec{\alpha}_2^{\mathrm{T}}\vec{g} + \lambda\vec{\beta}_1^{\mathrm{T}}\vec{h} + (1-\lambda)\vec{\beta}_2^{\mathrm{T}}\vec{h} \\ &= \inf_{\vec{u}\in\mathcal{U}} \lambda\left(f(\vec{u}) + \vec{\alpha}_1^{\mathrm{T}}\vec{g} + \vec{\beta}_1^{\mathrm{T}}\vec{h}\right) + (1-\lambda)\left(f(\vec{u}) + \vec{\alpha}_2^{\mathrm{T}}\vec{g} + \vec{\beta}_2^{\mathrm{T}}\vec{h}\right). \end{aligned}$$

უკანასკნელი ტოლობა მიღებულია იმის გათვალისწინებით, რომ $f(\vec{u}) = \lambda f(\vec{u}) + (1-\lambda) f(\vec{u})$.

დაბოლოს,

$$\left.\begin{aligned} Q\left(\vec{\alpha}, \vec{\beta}\right) &\ge \lambda \inf_{\vec{u}\in\mathcal{U}}\left(f(\vec{u}) + \vec{\alpha}_1^{\mathrm{T}}\vec{g} + \vec{\beta}_1^{\mathrm{T}}\vec{h}\right) + (1-\lambda)\inf_{\vec{u}\in\mathcal{U}}\left(f(\vec{u}) + \vec{\alpha}_2^{\mathrm{T}}\vec{g} + \vec{\beta}_2^{\mathrm{T}}\vec{h}\right) \\ &\ge \lambda Q\left(\vec{\alpha}_1, \vec{\beta}_1\right) + (1-\lambda) Q\left(\vec{\alpha}_2, \vec{\beta}_2\right) \end{aligned}\right\}.$$

ამრიგად, Q ჩაზნექილია.

□

4.2 სუსტი დუალობა

ლაგრანჟის $(\mathcal{P})$ შეზღუდვებიანი მინიმიზაციის ამოცანის დუალური ფუნქცია $(\mathcal{P})$ პრობლემის ამონახსნისთვის ქვედა საზღვარს იძლევა.

თეორემა 1.7 *დავუშვათ, რომ p^* არის $(\mathcal{P})$ პრობლემის ამონახსნი. მაშინ, როგორიც არ უნდა იყოს $\alpha_1, \alpha_2, \ldots, \alpha_m \ge 0$ და $\beta_1, \beta_2, \ldots, \beta_r \in \mathbb{R}$ სიდიდეთა მნიშვნელობები, $Q(\alpha, \beta) \le p^*$.*

✿

მტკიცებულება. დავუშვათ, რომ $\vec{u}$ — დასაშვები წერტილია. $g_i(\vec{u}) \le 0$ ყველა $i = 1, \ldots, m$ სიდიდისათვის და $h_j(\vec{u}) = 0$ ყველა $j = 1, \ldots, r$ სიდიდისათვის.

ამრიგად, $L\left(\vec{u}, \vec{\alpha}, \vec{\beta}\right) = f(\vec{u}) + \sum\limits_{i=1}^{m} \alpha_i g_i(\vec{u}) + \sum\limits_{j=1}^{r} \beta_j h_j(\vec{u})$ და, მაშასადამე, ნებისმიერი დასაშვები წერტილისთვის — $L\left(\vec{u}, \vec{\alpha}, \vec{\beta}\right) \le f(\vec{u})$. აქედან კი ჩვენ შეგვიძლია დავასკვნათ, რომ

$$\inf_{\vec{u}\in\mathbb{R}^n} L\left(\vec{u},\vec{\alpha},\vec{\beta}\right) \le \inf_{\vec{u}\in\mathbb{R}^n} f\left(\vec{u}\right).$$

□

ვინაიდან $Q\left(\alpha,\beta\right) \le p^*$ ყველა $\vec{\alpha}\in\mathbb{R}_+^m, \vec{\beta}\in\mathbb{R}^r$ სიდიდისათვის, ამიტომ Q-ის მაქსიმუმი $\mathbb{R}_+^m \times \mathbb{R}^r$ სივრცეში საუკეთესოა p^*-ის ქვედა საზღვრებს შორის.

განსაზღვრება 1.17 (ლაგრანჟის დუალური ამოცანა) დავუშვათ, რომ $\left(\mathcal{P}\right)$ არის (13.29) სახის მინიმიზაციის ამოცანა შეზღუდვებით, L — მისი ლაგრანჟიანი, ხოლო Q — მისი ლაგრანჟის დუალური ფუნქცია. მაშინ ოპტიმიზაციის შემდეგ ამოცანას *ლაგრანჟის დუალური ამოცანა* ეწოდება :

$$\left(\mathcal{Q}\right): \left.\begin{array}{rl} \max_{\vec{\alpha}\in\mathbb{R}^m,\vec{\beta}\in\mathbb{R}^r} & Q\left(\vec{\alpha},\vec{\beta}\right) \\ \text{შეზღუდვებით } \alpha_i \ge 0 & \forall i=1,\ldots,m \end{array}\right\}.$$

ამის საპირისპიროდ $\left(\mathcal{P}\right)$ პრობლემას *პირდაპირი* (*პრიმალური*) *ამოცანა* ეწოდება.

დაბოლოს, მაქსიმიზაციის $\vec{\alpha}^*, \vec{\beta}^*$ წერტილს *ლაგრანჟის ოპტიმალური მამრავლები* ან *ოპტიმალური დუალური ცვლადები* ეწოდება.

✿

ვინაიდან Q არის *ჩაზნექილი*, შეზღუდვებიანი ოპტიმიზაციის ნებისმიერი ამოცანის დუალური ამოცანა არის *ამოზნექილი* ოპტიმიზაციის შეზღუდვებიანი ამოცანა.

თეორემა 1.8 (სუსტი დუალობა) *დავუშვათ, რომ $\left(\mathcal{P}\right)$ არის (13.29) სახით მოცემული მინიმიზაციის ამოცანა შეზღუდვებით, ხოლო $\left(\mathcal{Q}\right)$ — მისი დუალური ლაგრანჟიანი. ვუწოდოთ d^* სიდიდეს $\left(\mathcal{Q}\right)$ ლაგრანჟის დუალური ამოცანის ამონახსნი, ხოლო p^* სიდიდეს — $\left(\mathcal{P}\right)$ პირდაპირი ამოცანის ამონახსნი. მაშინ $d^* \le p^*$.*

ეს სწორედ ისაა, რასაც სუსტი დუალობა ეწოდება.

✿

4.3 ძლიერი დუალობა

თუ დუალური ამოცანის ამონახსნი პირდაპირი ამოცანის მინიმიზაციას ახდენს, ზოგად შემთხვევაში ეს ამონახსნები ერთმანეთს არ უდრის. თუმცა, ფუნქციათა ზოგიერთი კლასისთვის, შესაძლებელია $d^* = p^*$ ტოლობა. ასეთ შემთხვევაში ლაპარაკობენ *ძლიერი დუალობის* შესახებ.

შენიშვნა

დუალობა არის პრინციპი, რომლის თანახმად ოპტიმიზაციის ამოცანა შეიძლება განიხილებოდეს ორი თვალსაზრისით : ან როგორც პირდაპირი ამოცანა (primal), ან როგორც მისი დუალური ამოცანა (dual). დუალური ამოცანის ამონახსნი იძლევა (მინიმიზაციის დროს) პირდაპირი ამოცანის ქვედა საზღვარს. მაგრამ, ზოგად შემთხვევაში, პირდაპირი და დუალური ამოცანების ოპტიმალური ამონახსნების მიზნობრივ ფუნქციათა მნიშვნელობები ერთმანეთისგან განსხვავებულია. ამ მნიშვნელობათა განსხვავებას, თუ იგი შეინიშნება, *დუალობის წყვეტა* ეწოდება. ამოზნექილი ოპტიმიზაციის ამოცანებისთვის დუალობის წყვეტა ნულს უდრის შეზღუდვათა რეგულარობის გარკვეული პირობების შესრულებისას.

განსაზღვრება 1.18 (ძლიერი დუალობა) დავუშვათ, რომ $(\mathcal{P})$ არის (13.29) სახით მოცემული მინიმიზაციის ამოცანა შეზღუდვებით, ხოლო $(\mathcal{Q})$ — მისი დუალური ლაგრანჟიანი. ვუწოდოთ d^* სიდიდეს $(\mathcal{Q})$ ლაგრანჟის დუალური ამოცანის ამონახსნი , ხოლო p^* სიდიდეს — $(\mathcal{P})$ პირდაპირი ამოცანის ამონახსნი. ლაპარაკობენ *ძლიერი დუალობის* შესახებ, თუ $d^* = p^*$.

სლეიტერის პირობა — ეს საკმარისი (თუმცა არა აუცილებელი) პირობაა ძლიერი დუალობის გარანტიისთვის და იგი ამერიკელმა მათემატიკოსმა მორტონ ლინკოლნ სლეიტერმა (Morton Lincoln Slater, 1921-2002) დაამტკიცა («*Lagrange Multipliers Revisited*», Cowles Commission[4] Discussion Paper : Mathematics No. 403, November 7, 1950).

თეორემა 1.9 (სლეიტერის პირობა) *დავუშვათ, რომ $(\mathcal{P})$ არის (13.29) სახის მინიმიზაციის ამოცანა შეზღუდვებით. თუ $(\mathcal{P})$ ამოზნექილია, სხვანაირად, თუ $f, g_1, g_2, \ldots, g_m$ და $h_1, h_2, \ldots, h_r$ ფუნქციები ამოზნექილია და არსებობს თუნდაც ერთი დასაშვები წერტილი, რომლისთვისაც უტოლობის ტიპის g_i არააფინური შეზღუდვები მკაცრად მოწმდება, მაშინ ძლიერი დუალობა გარანტირებულია.*

ამრიგად, შეზღუდვებიანი ამოზნექილი ოპტიმიზაციის ამოცანათა დიდი რაოდენობა შეიძლება იყოს ამოხსნილი დუალური ამოცანის საშუალებით, რომლის $(\alpha_i \geq 0 \quad \forall i = 1, \ldots, m)$ შეზღუდვები მარტივად გასათვალისწინებელია.

კერძოდ, შეიძლება იყოს გამოყენებული *გრადიენტის პროექციის მეთოდები* (ინგლ. the Gradient Projection Methods), რომლებიც ხასიათდება დაშვების მიმართულების ალგორითმის

[4] იელის უნივერსიტეტის ალფრედ კოულზის (Alfred Cowles III, 1891 – 1984) მიერ 1932 წელს დაარსებული ფონდი ეკონომიკური დარგის კვლევებში (Cowles Commission for Research in Economics at Yale University) მიზნად ისახავს ეკონომიკის სფეროში კვლევათა ჩატარებას და წახალისებას. ქოულზის ფონდი ცდილობს ხელი შეუწყოს ანალიზის მკაცრ ლოგიკურ, მათემატიკურ და სტატისტიკურ განვითარებას და გამოყენებას. ქოულზის ფონდი ფინანსურ მხარდაჭერას უწევს გამოკვლევებს, მოწვეულ პედაგოგებს, პოსტსადოქტორო სტიპენდიებს, სემინარებს და დოქტორანტებს. ქოულზის ფონდი რეგულარულად აფინანსებს კონფერენციებს და აქვეყნებს სამუშაო დოკუმენტების კრებულებს, ასევე რეპრინტებისა და მონოგრაფიების სერიებს.

თანამიმდევრული იტერაციების «დაბრუნებით» შეზღუდვათა არეში მასზე პროეცირების გზით.

4.4 კარუშ-კუნ-ტაკერის პირობები

ძლიერი დუალობის პირობებში, თუ f, g_i და h_i ფუნქციები დიფერენცირებადია, უილიამ კარუშმა (William Karush) 1939 წლის თავის გამოუქვეყნებელ სამაგისტრო დისერტაციაში, ხოლო მოგვიანებით ჰაროლდ კუნმა (Harold W. Kuhn) და ალბერტ ტაკერმა (Kuhn and Tucker, 1951) წამოაყენეს პირობითი (შეზღუდვებიანი) ოპტიმიზაციის ამოცანის ამონახსნის ოპტიმალურობის საკმარის პირობათა მთელი სიმრავლე.

თეორემა 1.10 (კარუშ-კუნ-ტაკერის პირობები) *დავუშვათ, რომ* $(\mathcal{P})$ *არის (13.29) სახის შეზღუდვებიანი მინიმიზაციის ამოცანა, ხოლო* $(\mathcal{Q})$ *— მისი დუალური ამოცანა.*

დავუშვათ, რომ გვაქვს $\mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^r$ *სივრცის* $\left(\vec{u}^*, \vec{\alpha}^*, \vec{\beta}^*\right)$ *ტრიპლეტი (სამეული), რომელიც ამოწმებს კარუშ-კუნ-ტაკერის (Karush-Kuhn-Tucker) პირობებად წოდებულ შემდეგ მოთხოვნებს :*

- *პირდაპირი ამოცანის დასაშვებობა:* $g_i\left(\vec{u}^*\right) \le 0$ *ყველა* $i = 1, \ldots, m$ *მნიშვნელობისთვის და* $h_i\left(\vec{u}^*\right) = 0$ *ყველა* $j = 1, \ldots, r$ *მნიშვნელობისთვის;*
- *დუალური ამოცანის დასაშვებობა:* $\alpha_i^* \ge 0$ *ყველა* $i = 1, \ldots, m$ *მნიშვნელობისთვის;*
- *შეზღუდვათა კომპლემენტარობა (ურთიერთშევსებადობა, ურთიერთდამატებითობა) :* $\alpha_i^* \cdot g_i\left(\vec{u}^*\right) = 0$ *ყველა* $i = 1, \ldots, m$ *მნიშვნელობისთვის;*
- *სტაციონარობა:* $\nabla f\left(\vec{u}^*\right) + \sum_{i=1}^{m} \alpha_i^* \nabla g_i\left(\vec{u}^*\right) + \sum_{j=1}^{r} \beta_j^* \nabla h_j\left(\vec{u}^*\right)$.

მაშინ $\vec{u}^*$ *არის* $(\mathcal{P})$ *პირდაპირი ამოცანის მინიმიზაციის წერტილი, ხოლო* $\left(\vec{\alpha}^*, \vec{\beta}^*\right)$ *წარმოადგენს* $(\mathcal{Q})$ *დუალური ამოცანის მაქსიმიზაციის წერტილს.*

✿

მტკიცებულება. *დავუშვათ, რომ გვაქვს* $\left(\vec{u}^*, \vec{\alpha}^*, \vec{\beta}^*\right)$ ტრიპლეტი, რომელიც ამოწმებს კარუშ-კუნ-ტაკერის პირობებს. პირდაპირი დასაშვებობის პირობა გულისხმობს, რომ $\vec{u}^*$ არის დასაშვები. $\vec{u} \mapsto L\left(\vec{u}, \vec{\alpha}^*, \vec{\beta}^*\right)$ ფუნქცია ამოზნექილია $\vec{u}$ -ზე, მართლაც :

$$L\left(\vec{u}^*, \vec{\alpha}^*, \vec{\beta}^*\right) = f\left(\vec{u}\right) + \sum_{i=1}^{n} \alpha_i^* g_i\left(\vec{u}\right) + \sum_{j=1}^{r} \beta_j^* h_j\left(\vec{u}\right),$$

სადაც f და g_i ფუნქციები ამოზნექილია, α_i^* სიდიდეები დადებითია, ხოლო h_j ფუნქციები აფინურია *(როგორც ცნობილია, ერთი ან მრავალი ცვლადის ფუნქციას აფინური ეწოდება, თუ*

შესაძლებელია მისი წარმოდგენა ამ ცვლადების წრფივ კომბინაციად, რომელსაც ემატება რაღაც კონსტანტა). ამიტომ სტაციონარობის პირობა გულისხმობს, რომ $\vec{u}^*$ ახდენს $L\left(\vec{u},\vec{\alpha}^*,\vec{\beta}^*\right)$ ფუნქციის მინიმიზაციას.

ლაგრანჟის დუალური ფუნქციის განმარტებით, $Q\left(\vec{\alpha}^*,\vec{\beta}^*\right)=L\left(\vec{u}^*,\vec{\alpha}^*,\vec{\beta}^*\right)=f\left(\vec{u}^*\right)$. მართლაც, შეზღუდვათა კომპლემენტარობის (ურთიერთშევსებადობის, ურთიერთდამატებითობის) პირობა გულისხმობს $\sum_{i=1}^{n}\alpha_i^*\,g_i\left(\vec{u}^*\right)=0$ ტოლობას, ხოლო პირდაპირი ამოცანის დასაშვებობის პირობა — $\sum_{j=1}^{r}\beta_j^*h_j\left(\vec{u}^*\right)=0$ ტოლობას.

დავუშვათ, რომ p^* არის $\left(\mathcal{P}\right)$ ამოცანის ამონახსნი, ხოლო d^* — დუალური $\left(\mathcal{Q}\right)$ ამოცანის ამონახსნი. d^* სიდიდის განმარტების თანახმად, $f\left(\vec{u}^*\right)=Q\left(\vec{\alpha}^*,\vec{\beta}^*\right)\le d^*$ და, სუსტი დუალობიდან გამომდინარე, $d^*\le p^*$. ამრიგად, $f\left(\vec{u}^*\right)\le p^*$ და, მაშასადამე, $f\left(\vec{u}^*\right)=p^*$: $\vec{u}^*$ არის $\left(\mathcal{P}\right)$ ამოცანის მინიმიზაციის წერტილი, ხოლო ყველა წინა უტოლობები არის ტოლობები და, კერძოდ, $Q\left(\vec{\alpha}^*,\vec{\beta}^*\right)=d^*$, ხოლო $\left(\vec{\alpha}^*,\vec{\beta}^*\right)$ წარმოადგენს $\left(\mathcal{Q}\right)$ ამოცანის მაქსიმიზაციის წერტილს.

□

გეომეტრიულად შეზღუდვათა სტაციონარობის და კომპლემენტარობის (ურთიერთშევსებადობის, ურთიერთდამატებითობის) პირობები შეიძლება შემდეგნაირად იყოს გაგებული. განვიხილოთ ამოზნექილი ოპტიმიზაციის ამოცანა უტოლობის ტიპის ერთი შეზღუდვით :

$$\left.\begin{array}{ll}\min\limits_{\vec{u}\in\mathbb{R}^n} & f\left(\vec{u}\right)\\ & g\left(\vec{u}\right)\le 0 \quad \text{შეზღუდვისას}\end{array}\right\}. \tag{13.30}$$

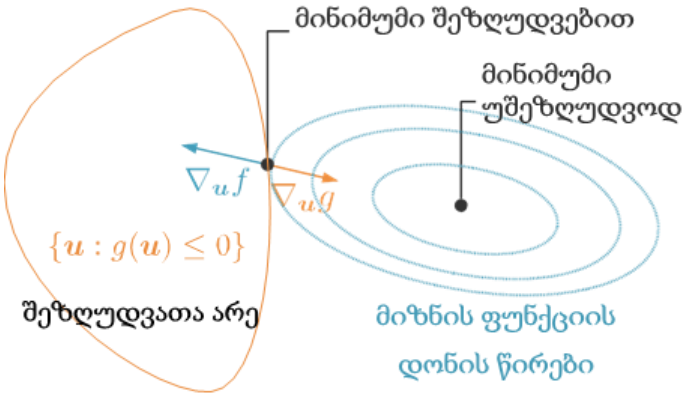
დავუშვათ, რომ $\vec{u}_0$ უპირობო მინიმიზაციის წერტილია : $\vec{u}_0=\min\limits_{\vec{u}\in\mathbb{R}^n} f\left(\vec{u}\right)$.

შესაძლებელია ორი შემთხვევა :

- ან $\vec{u}_0$ მიეკუთვნება შეზღუდვათა არეს და $\vec{u}^*=\vec{u}_0$ წარმოადგენს $\left(\mathcal{P}\right)$ ამოცანის მინიმიზაციის წერტილს (შემთხვევა 1);
- ან $\vec{u}_0$ არ მიეკუთვნება შეზღუდვათა არეს (შემთხვევა 2).

პირველ შემთხვევაში — $\nabla f\left(\vec{u}^*\right)=0$ (ვინაიდან $\vec{u}^*$ უპირობო ოპტიმიზაციის ამოცანის ამონახსნია) და $g\left(\vec{u}^*\right)\le 0$ (ვინაიდან $\vec{u}^*$ მიეკუთვნება შეზღუდვათა არეს).

მეორე შემთხვევაში, რომელიც ნაჩვენებია ნახატზე 13.5, მინიმიზაციის $\vec{u}^*$ წერტილი იმყოფება შეზღუდვათა უბნის $\left(g\left(\vec{u}^*\right)=0\right)$ საზღვარზე (კიდეზე).



ნახატი 13.5 - მინიმიზაციის წერტილი იმყოფება შეზღუდვათა უბნის საზღვარზე f ფუნქციის ერთ-ერთი დონის წირისადმი მხებ წერტილზე, სადაც ∇g და ∇f გრადიენტები პარალელურია და მიმართულია ურთიერთსაწინააღმდეგოდ.

ეს იმიტომ ხდება, რომ f ფუნქცია, რომელიც ამოზნექილია და მინიმუმს ამ უბნის გარეთ იძენს, იზრდება ხსენებულ წერტილთან მიახლოებისას. უფრო ზუსტად რომ ვთქვათ, $\vec{u}^*$ იმყოფება შეზღუდვათა უბნის კიდისა და f ფუნქციის ერთ-ერთი დონის წირის გადაკვეთაზე. ამ წერტილზე, ვინაიდან f-ის გრადიენტი მიმართულია დონის წირისადმი გავლებული მხების გასწვრივ, ხოლო g ფუნქციის გრადიენტი მიმართულია უბნის კიდისადმი გავლებული მხების გასწვრივ, ამიტომ ∇f და ∇g პარალელურია. უფრო მეტიც, f იზრდება, როცა შორდება $\vec{u}_0$-ს, და g უარყოფითია შეზღუდვათა უბნის შიგნით. ამრიგად, ∇f და ∇g ურთიერთსაწინააღმდეგოდ არის მიმართულია. მაშასადამე, არსებობს $\alpha \geq 0$ ისეთი, რომ $\nabla f\left(\vec{u}^*\right) = -\alpha \nabla g\left(\vec{u}^*\right)$.

ამრიგად, ეს ორი შემთხვევა საბოლოო ფორმით შემდეგნაირად შეიძლება იყოს წარმოდგენილი :

$$\nabla f\left(\vec{u}^*\right) + \alpha \nabla g\left(\vec{u}^*\right) = 0 \text{ და } \alpha g\left(\vec{u}^*\right) = 0,$$

სადაც ან $\alpha = 0$ და $g\left(\vec{u}^*\right) \leq 0$ (შემთხვევა 1), ან $g\left(\vec{u}^*\right) = 0$ და $\alpha \geq 0$ (შემთხვევა 2).

ამრიგად, ნაპოვნია $\nabla f\left(\vec{u}^*\right) + \alpha \cdot \nabla g\left(\vec{u}^*\right) = 0$ სტაციონარობის პირობა და შეზღუდვათა კომპლემენტარობის (ურთიერთშევსებადობის, ურთიერთდამატებითობის) $\alpha \cdot g\left(\vec{u}^*\right) = 0$ პირობა.

4.5 კვადრატული დაპროგრამება

მანქანურ სწავლებაში ხშირად წამოიჭრება შეზღუდვებიანი ამოზნექილი ოპტიმიზაციის პრობლემა, რომლის კერძო შემთხვევას ამოზნექილი კვადრატული დაპროგრამება წარმოადგენს — *convex quadratic programming* (*convex QP*) ინგლისურად. ამ შემთხვევაში სამინიმიზაციო f ფუნქცია *კვადრატულია*, ხოლო შეზღუდვები — *აფინური*.

განსაზღვრება 1.19 როცა მოცემულია n, m, r სამი დადებითი მთელი რიცხვი, დადებითი ნახევრადგანსაზღვრული $Q \in \mathbb{R}^{n\times n}$ მატრიცა, $\vec{a}, \vec{b}_1, \ldots, \vec{b}_m, \vec{c}_1, \ldots, \vec{c}_r$ ვექტორები $\mathbb{R}^n$ სივრციდან და $d, k_1, \ldots, k_m, l_1, \ldots, l_r \in \mathbb{R}$ კონსტანტები, *კვადრატული ამოზნექილი ოპტიმიზაციის ამოცანა* შემდეგი სახის ამოცანას ეწოდება :

$$\left.\begin{aligned} &\min_{\vec{u}\in\mathbb{R}^n} \frac{1}{2}\vec{u}^{\mathrm{T}} Q\vec{u} + \vec{a}^{\mathrm{T}}\vec{u} + d \quad \text{შემდეგი შეზღუდვების პირობებში} \\ &\vec{b}_i^{\mathrm{T}}\vec{u} + k_i \quad \forall i = 1,\ldots,m \\ &\vec{c}_j^{\mathrm{T}}\vec{u} + l_j \quad \forall j = 1,\ldots,r \end{aligned}\right\}.$$

ასე რომ, საქმე ეხება ამოზნექილი ოპტიმიზაციის ამოცანას, რომელიც სლეიტრის (Slater) პირობებს აკმაყოფილებს და ამ ამოცანის ამონახსნის არე მრავალწახნაგას (პოლიედრს) წარმოადგენს.

არსებობს ასეთი სახის ამოცანათა ამოხსნის მრავალი მეთოდი, მაგალითად შიდა წერტილების მეთოდები ან აქტივაციის მეთოდები, რომლებსაც ზოგჯერ აქტიურ სიმრავლეთა (ინგლ. *active set*) მეთოდებსაც უწოდებენ. არაერთი პროგრამული უზრუნველყოფა და ბიბლიოთეკა ახორციელებს ამ მიდგომებს. მაგალითის სახით შეიძლება დასახელდეს CPLEX, CVXOPT, CGAL და ძალიან ბევრი სხვა.

დამატებითი ინფორმაცია

- ნაშრომი Boyd and Vandenberghe (2004) წარმოადგენს პრაქტიკულად ამომწურავ ცნობარს ამოზნექილი ოპტიმიზაციის შესახებ. რაც შეეხება ამ პრობლემის რიცხვით ასპექტებს, ძალიან საინტერესოა Bonnans et al (1997) ან Nocedal and Wright (2006).
- უფრო დაწვრილებითი ინფორმაციის მიღება ალგორითმებზე, რომლებიც იყენებს და-შვებებს სტოქასტიკური მიმართულებებით, შეიძლება Léon Bottou (2012) სტატიიდან.
- უფრო დაწვრილებითი ინფორმაციის მიღება ლაგრანჟის დუალობის შესახებ შეიძლება Luenberger (1969) წიგნის მერვე თავიდან და ასევე Bertsekas et al. (2003) ნაშრომიდან.

ბიბლიოგრაფია

1. Dimitri P. Bertsekas with Angelia Nedic and Asuman E. Ozdaglar (2003). *Convex Analysis and Optimization*, Athena Scientific, Belmont, Massachusetts. ISBN: 1-886529-45-0, 560 pages

2. Dimitri P. Bertsekas (2009), Massachusetts Institute of Technology, *Convex Optimization Theory*, Athena Scientific, Belmont, Massachusetts, ISBN-10: 1-886529-31-0, ISBN-13: 978-1-886529-31-1, 444 pages, http://web.mit.edu/dimitrib/www/Convex_Theory_Entire_Book.pdf

3. Bonnans, J., Gilbert, J., Lemaréchal, C., et Sagastizábal, C. (1997). *Optimisation numérique : aspects théoriques et pratiques*. Springer-Verlag, Berlin Heidelberg.

4. Bottou, L. (2012). Stochastic gradient descent tricks. http://leon.bottou.org/publications/pdf/tricks-2012.pdf .

5. Boyd, S. et Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press. https://web.stanford.edu/~boyd/cvxbook/ .

6. Hestenes, M. et Stiefel, E. (1952). Methods of conjugate gradients for solving linear systems. *Journal of Research of the National Bureau of Standards*, 19(6).

7. Kuhn, H. W. et Tucker, A. W. (1951). Nonlinear programming. In Neyman, J., editor, *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pages 481–492. University of California Press.

8. Luenberger, D. G. (1969). *Optimization by Vector Space Methods*. John Wiley & Sons.

9. Nocedal, J. et Wright, S. J. (2006). *Numerical Optimization*. Springer, Berlin, New York, 2nd edition.

10. Slater, M. (1950). Lagrange multipliers revisited. *Cowles Commission Discussion Paper*, number 403.

ლექცია 14 წერტილოვანი შეფასების კონცეფციები

შინაარსი

ლექციათა ამ კურსში ჩვენ რეგულარულად ვიყენებთ ინფენციულ სტატისტიკას. ინფერენციული სტატისტიკა მდგომარეობს იმაში, რომ აკეთოს დასკვნები გარკვეულ პოპულაციაზე, ერთობლიობაზე მისი ანარჩევის შესწავლის საფუძველზე. ამ ლექციაში მოცემულია ზოგიერთი წარმოდგენა ინფერენციული სტატისტიკისა და, კერძოდ, შეფასების შესახებ.

მიზნები

— წერტილოვანი შეფასების პრობლემის გაცნობა ;

— შემფასებლის არჩევა, სახელდობრ, მის ისეთ თვისებათა განსაზღვრის გზით, როგორიცაა წანაცვლება ან სიზუსტე ;

— შემფასებლის შემოთავაზება, ვთქვათ, დამაჯერებლობის მაქსიმიზაციით.

1. ინფერენციული სტატისტიკა

სტატისტიკა არის გამოყენებითი მათემატიკური მეთოდების ნაკრები იმ მოვლენათა აღსაწერად და გასაანალიზებლად, რომელთა ბუნება შეუძლებელს ხდის მათ ამომწურავ შესწავლას. ეს მეთოდები საშუალებას იძლევა შევისწავლოთ მონაცემები ან დაკვირვებები, რომლებიც შედგენილია პირთა ან ეკვივალენტურ ობიექტთა ერთობლიობის ერთი ან რამდენიმე მახასიათებლით.

ეკვივალენტური პირების ან ობიექტების ამ ერთობლიობას *პოპულაცია* ეწოდება. ეს შეიძლება იყოს პოპულაცია ამ სიტყვის «ჩვეულებრივი» გაგებით (მაგალითად, საფრანგეთის მთელი მოსახლეობა ან გარკვეულ ტერიტორიაზე ერთი სახის ცხოველთა ყველა ინდივიდი), ასევე, გაცილებით ფართო გაგებით, იმ ობიექტების უფრო ზოგადი ნაკრები, რომელთა შესწავლას ვცდილობთ (მაგალითად, საამწყობო კონვეიერზე წარმოებული ყველა დეტალი ან ნაწილაკთა

ნაკრები ფიზიკაში).

პოპულაციის თითოეულ ელემენტს *ინდივიდუუმი* ეწოდება.

მახასიათებლებს, რომლებიც იზომება ყოველი ამ ინდივიდუუმისთვის, *ცვლადები* ეწოდება ; ინდივიდუუმებს, რომლებისთვისაც იყო გაზომილი ეს მახასიათებლები, *დაკვირვებები* ეწოდება. გარკვეული ცვლადის n დაკვირვებათა $(x_1, x_2, ..., x_n)$ სიმრავლეს *სტატისტიკური მწკრივი* ეწოდება.

მაგალითი

დავუშვათ, რომ ვსწავლობთ ერთი (მცხრალი, მარტივი, არანაკიანი — 365 დღის შემცველი) წელიწადის კლიმატურ მონაცემებს მეტეოსადგურიდან. ამ მონაცემებში ყოველდღიური ჩანაწერით ასახულია 8 გაზომვა : მაქსიმალური, მინიმალური და საშუალო ტემპერატურის; ქარის მაქსიმალური სიჩქარის; მზის აქტივობის; ნალექის საერთო რაოდენობის; მინიმალური და მაქსიმალური ატმოსფერული წნევის. პოპულაცია შეიცავს 365 *ინდივიდს*, რომელთა შორის თითოეული აღიწერება 8 *ცვლადით*. ახლა დავუშვათ, რომ მონაცემთა მხოლოდ ნახევარი გვაქვს (ვთქვათ, ყოველი მეორე დღისთვის). მაშინ ჩვენ 182 *დაკვირვება* გვექნება. მზის აქტივობის მნიშვნელობები ამ 182 დღისთვის ქმნის *სტატისტიკურ მწკრივს*.

როცა შესასწავლი პოპულაცია ძალიან დიდია დაკვირვების განსახორციელებლად თითოეულ ინდივიდუუმზე, მაშინ პოპულაციის მხოლოდ და მხოლოდ ნაწილი შეისწავლება. ამ ნაწილს *ანარჩევი* ანუ *ნიმუში* ეწოდება (იხ. მე-14 ლექციის პუნქტი 1.1). ამ შემთხვევაში ასევე ლაპარაკობენ *გამოკითხვაზე* განსხვავებით *აღწერისგან*, რომელიც წარმოადგენს ყველა ინდივიდის კვლევას პოპულაციაში.

ინფერენციალური სტატისტიკა, რომელიც ასევე ცნობილია *გადაწყვეტილებათა სტატისტიკის* ან *სტატისტიკური დანასკვის* სახელწოდებებითაც, გულისხმობს დასკვნების გაკეთებას ერთობლიობაზე (პოპულაციაზე) ამ ერთობლიობიდან (პოპულაციიდან) მიღებული ანარჩევის შესწავლის საფუძველზე. დანაკვირვები მონაცემები განიხილება როგორც ანარჩევი ერთობლიობიდან. ერთი სიტყვით, ლაპარაკია ანარჩევზე დანაკვირვებ თვისებათა გავრცობის შესახებ პოპულაციაზე. სტატისტიკური დანასკვი მნიშვნელოვნად ეყრდნობა ალბათობებს : დაკვირვებები განიხილება შემთხვევით სიდიდეთა რეალიზაციებად, რაც ამ შემთხვევით სიდიდეთა ალბათურ მახასიათებლებთან მიახლოების საშუალებას იძლევა ანარჩევზე დადგენილი მაჩვენებლების გამოყენებით.

1.1 ანარჩევი

ანარჩევის უპირატესობის გამოსაყენებლად ჩვენ მოგვიწევს შემდეგი დაშვებების გაკეთება :

- პოპულაციის რაოდენობა (რიცხვი, რიცხობრიობა) უსასრულოა.

- პოპულაციაზე გაზომილი ცვლადები შეიძლება იყოს განხილული შემთხვევით სიდიდეებად და მათი ზომები წარმოადგენს რეალიზაციებს. ალბათობის კანონები, რომლებსაც ემორჩილება ეს ცვლადები, შეიძლება მიეკუთვნებოდეს ცნობილ ოჯახს (მაგალითად, გაუსის კანონს, პუასონის კანონს) ან იყოს სრულიად უცნობი. პირველ შემთხვევაში, რომელიც ჩვენ აქ

გვაინტერესებს, ლაპარაკობენ *პარამეტრულ ინფერენციულ სტატისტიკაზე* ; მეორეში კი — *არაპარამეტრულ ინფერენციულ სტატისტიკაზე.*

ამრიგად, ინფერენციული სტატისტიკის მიზანს წარმოადგენს ხსენებული *შემთხვევითი სიდიდეების ალბათური კანონების დადგენა.* კერძოდ, *შეფასება* მდგომარეობს ანარჩევის გამოყენებაში შესასწავლ შემთხვევით სიდიდეთა კანონების პარამეტრთა განსასაზღვრავად (მაგალითად, p პარამეტრის ბერნულის განაწილებაში, ინდექსისა და მასშტაბის პარამეტრის გამა განაწილებაში) ან ამ სიდიდეთა მახასიათებლების გამოსაანგარიშებლად (ვთქვათ, მათემატიკური ლოდინის, დისპერსიის, უმაღლესი რიგის მომენტის, კვანტილის — მნიშვნელობის, რომელსაც მოცემული შემთხვევითი სიდიდე არ აღემატება ფიქსირებული ალბათობით).

ამ ლექციის დარჩენილ ნაწილში ნაგულისხმევი იქნება, რომ გამოკითხვის გზით შედგენილი ანარჩევი მიღებულია *მარტივი შემთხვევითი შერჩევით* :

ინდივიდები აიღება პოპულაციიდან ალალბედზე, შემთხვევითად, დაბრუნების გარეშე საწყის მდგომარეობაში. თითოეულ ინდივიდს პოპულაციიდან აქვს ანარჩევში მოხვედრის ერთნაირი $1/N$ ალბათობა, სადაც N — პოპულაციის ზომაა (არ უნდა დაგვავიწყდეს, რომ $N \to \infty$) და ინდივიდები ხვდება ანარჩევში ერთმანეთისგან დამოუკიდებლად.

შენიშვნა

შემთხვევითი ანარჩევის საფუძველზე დასკვნების გამოტანამდე მნიშვნელოვანია დავრწმუნდეთ, რომ იგი რეპრეზენტატულია შესასწავლი პოპულაციისათვის.

მაგალითად, პირველი კლინიკური გამოკვლევები, რომლებმაც გვიჩვენა ასპირინის ეფექტურობა მიოკარდიუმის ინფარქტის რისკის შემცირებაში პაციენტებთან რისკის ჯგუფიდან, ჩატარებული იყო, ძირითადად, მამაკაცებით შედგენილ ანარჩევებზე; მხოლოდ მოგვიანებით სამედიცინო საზოგადოებამ გააცნობიერა, რომ ასპირინის ეფექტურობა ქალებში გაცილებით უფრო დაბალია.

ერთნაირი n ზომის ორი $(x_1, x_2, \dots, x_n)$ და $(x'_1, x'_2, \dots, x'_n)$ ანარჩევი ერთისა და იმავე პოპულაციიდან სხვადასხვაგვარი იქნება.

ამ განსხვავების მოდელირება ხდება დაშვებით, რომ x_i და x'_i არის ერთისა და იმავე X_i შემთხვევითი ცვლადის რეალიზაცია.

$(X_1, X_2, \dots, X_n)$ — შემთხვევითი ვექტორია. მისი კომპონენტები დამოუკიდებელი და იდენტურად განაწილებული სიდიდეებია (ინგლისურად : independent and identically distributed – iid).

განსაზღვრება 14.1 (შემთხვევითი ანარჩევი და ანარჩევი) დავუშვათ, რომ X არის შემთხვევითი ცვლადი. მაშინ X ცვლადის *შემთხვევითი ანარჩევი* ეწოდება X ცვლადის n დამოუკიდებელი და ერთნაირად განაწილებული ასლების $(X_1, X_2, \dots, X_n)$ მწკრივს, ხოლო $(x_1, x_2, \dots, x_n)$ *ანარჩევი* წარმოადგენს $(X_1, X_2, \dots, X_n)$ მწკრივის რეალიზაციას.

ამრიგად, *შემთხვევითი ანარჩევი* არის *შემთხვევითი ვექტორი*, მაშინ როცა *ანარჩევი* წარმოადგენს *სტატისტიკურ მწკრივს*.

მაგალითი

ანარჩევის საშუალო მნიშვნელობა, $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, არის M_n შემთხვევითი ცვლადის რეალიზაცია. ეს შემთხვევითი M_n ცვლადი კი განისაზღვრება თანაფარდობით

$$M_n = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

რომელიც წარმოადგენს $\left(X_1,\, X_2,\, \ldots,\, X_n\right)$ შემთხვევითი ანარჩევის ფუნქციას.

2 წერტილოვანი შეფასება

დავუშვათ, რომ $\left(\Omega, \mathcal{A}, \mathbb{P}\right)$ არის ალბათური სივრცე, E — განზომადი სივრცე, ხოლო X — შემთხვევითი სიდიდე მნიშვნელობებით E სივრცეში. პრაქტიკულად, ამ ლექციის დარჩენილ ნაწილში, ჩვენ მიერ განხილული იქნება ნამდვილი შემთხვევითი სიდიდეები (როცა E უდრის $\mathbb{R}$-ს ან $\mathbb{R}$-ის ნაწილს, ისეთს, როგორიც არის $\mathbb{R}_+$ ან $\mathbb{N}$), მაგრამ აქ წარმოდგენილი იდეები შეიძლება იყოს გავრცელებული $\mathbb{R}^d$-ზე ან უფრო რთულ სივრცეებზე.

დავუშვათ, რომ $\left(X_1,\, X_2,\, \ldots,\, X_n\right)$ არის X ცვლადის შემთხვევითი ანარჩევი. X_i სიდიდეები დამოუკიდებელია და ერთნაირად განაწილებული, იმავე $\mathbb{P}_X$ განაწილებით, რომელიც X ცვლადს ახასიათებს. დავუშვათ ასევე, რომ $\left(x_1, x_2, \ldots, x_n\right)$ ანარჩევია, სხვა სიტყვებით რომ ვთქვათ, ამ შემთხვევითი ანარჩევის რეალიზაცია. დაბოლოს, დავუშვათ, რომ $\theta \in \mathbb{R}$ წარმოადგენს დეტერმინებულ (სხვა სიტყვებით რომ ვთქვათ, *არაშემთხვევით*) სიდიდეს, რომელიც დამოკიდებულია მხოლოდ $\mathbb{P}_X$-ზე. წერტილოვანი შეფასების მიზანი მდგომარეობს იმაში, რომ — რაც შეიძლება უფრო ზუსტად — მოხდეს θ მნიშვნელობის აპროქსიმაცია.

მაგალითად, თუ ვარაუდობენ, რომ X ემორჩილება ალბათობის განაწილების ექსპონენციალურ კანონს (გავიხსენოთ, რომ ვიმყოფებით პარამეტრული ინფერენციული სტატისტიკის კონტექსტში), მაშინ θ შეიძლება იყოს ამ კანონის პარამეტრი და ასევე მისი ერთ-ერთი მომენტი, კვანტილი და ა.შ.

განსაზღვრება 14.2 (შემფასებელი და შეფასება) θ სიდიდის *შემფასებელი* ეწოდება $\left(X_1,\, X_2,\, \ldots,\, X_n\right)$ შემთხვევითი ანარჩევის *სტატისტიკას*, ესე იგი შემთხვევით ცვლადს — $\left(X_1,\, X_2,\, \ldots,\, X_n\right)$-ის ფუნქციას: θ სიდიდის Θ_n შემფასებელი შეიძლება იყოს განსაზღვრული ფორმულით

$$\Theta_n \;=\; g\left(X_1,\, X_2,\, \ldots,\, X_n\right), \quad g : E^n \to \mathbb{R}.$$

დავუშვათ, რომ მოცემულია X შემთხვევითი ცვლადის $(x_1, x_2, \dots, x_n)$ ანარჩევი, მაშინ θ სიდიდის *შეფასება* ეწოდება სიდიდეს

$$\hat{\Theta}_n = g(x_1, x_2, \dots, x_n) \in \mathbb{R},$$

რომელიც, ამრიგად, წარმოადგენს Θ_n *შემფასებლის* რეალიზაციას.

მაგალითი

განვიხილოთ X როგორც შემთხვევითი ცვლადი ინტეგრირებადი კვადრატით (სხვა სიტყვებით, კეთდება დაშვება, რომ X ცვლადს აქვს მათემატიკური ლოდინი და დისპერსია) m მათემატიკური ლოდინით და σ^2 დისპერსიით.

X *ცვლადის ემპირიული საშუალო მნიშვნელობა* არის შემდეგი სახით განსაზღვრული M_n შემთხვევითი ცვლადი :

$$M_n = \frac{1}{n}\sum_{i=1}^{n} X_i \,. \tag{14.1}$$

M_n არის m-ის *შემფასებელი*: თუ მოცემულია $(x_1, x_2, \dots, x_n)$ ანარჩევი, მაშინ

$$\hat{m}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$$

მნიშვნელობა იძლევა *შეფასებას* m სიდიდისთვის.

ზემოთ მოცემულ მაგალითში არ მოიპოვება იმის რაიმე მტკიცებულება, რომ M_n არის *კარგი შემფასებელი* m სიდიდისთვის.

მართალია, შეგვეძლო $(2/n)\sum_{i=1}^{n} X_i$ ან $(1/n)\sum_{i=1}^{n} X_i^2$ თანაფარდობათა არჩევა $X-$ის მათემატიკური ლოდინის შემფასებლებად, თუმცა ეს ნაკლებად ბუნებრივად გამოიყურება. მომდევნო პუნქტში შემფასებლის ზოგიერთი თვისება განისაზღვრება, რომლებიც გამოყენებული იქნება მისი სარგებლიანობის (ვარგისობის, გამოსადეგობის) რაოდენობრივი შეფასების მისაღებად. ეს თვისებები საშუალებას გვაძლევს მათემატიკურად დავასაბუთოთ ჩვენი ინტუიციური მიგნება, რომ M_n არის უკეთესი შემფასებელი m-ისთვის, ვიდრე, ვთქვათ, $2M_n$.

3 შემფასებლის (შემფასებელი მაჩვენებლის) თვისებები

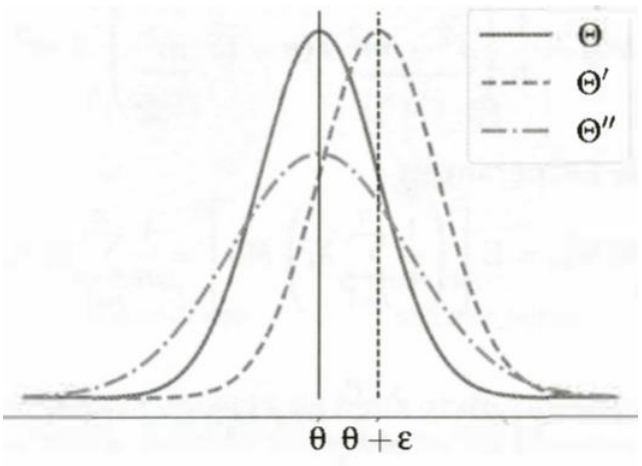
ამ პუნქტში ჩვენ ჯერ კიდევ ვიხილავთ ნამდვილი X შემთხვევითი ცვლადის $n \in \mathbb{N}^*$ ზომის $(X_1, X_2, \dots, X_n)$ შემთხვევით ანარჩევს. ხსენებული ცვლადი ხასიათდება $\mathbb{P}_X$ განაწილების კანონით და Θ_n შემფასებლით θ-თვის. მაგრამ ახლა ჩვენი მიზანია Θ_n შემფასებელზე (შემფასებელ მაჩვენებელზე) შეჩერება უფრო დაწვრილებით.

3.1 შემფასებლის წანაცვლება

განსაზღვრება 13.3 (შემფასებლის წანაცვლება) θ სიდიდის Θ_n შემფასებლის *წანაცვლება* (ინგლისურად : *the bias*) განსაზღვრულია თანაფარდობით

$$\mathrm{B}(\Theta_n) = \mathbb{E}\left[\Theta_n\right] - \theta\,. \tag{14.2}$$

ითვლება, რომ Θ_n *წაუნაცვლებელია*, თუ $\mathrm{B}(\Theta_n) = 0$, სხვა სიტყვებით, თუ $\mathbb{E}\left[\Theta_n\right] = \theta$.



ნახატი 14.1 - θ სიდიდის სამი შემფასებლის განაწილება.

ნახატზე 14.1 ნაჩვენებია ერთისა და იმავე θ სიდიდის სამი შემფასებლის განაწილება. აქ იგულისხმება, რომ განაწილება არის გაუსის. Θ და Θ'' შემფასებლები წაუნაცვლებელია. Θ' შემფასებელი კი წანაცვლებულია : მისი მათემატიკური ლოდინი უდრის $\theta + \varepsilon \equiv \theta + \in$.

მაგალითი

განვიხილოთ X ცვლადი $\left(X \in \mathcal{L}^2\right)$ ინტეგრირებადი კვადრატით, m მათემატიკური ლოდინით და σ^2 დისპერსიით. X ცვლადის ემპირიული დისპერსია — ეს S_n შემთხვევითი სიდიდეა, რომელიც განისაზღვრება ფორმულით

$$S_n = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - M_n\right)^2, \tag{14.3}$$

სადაც M_n — ემპირიული საშუალოა, როგორც ეს განსაზღვრულია ზემოთ. S_n — ეს σ^2 დისპერსიის შემფასებელია.

იმავდროულად, მისი წანაცვლება $\left(-\frac{1}{n}\right)\sigma^2$ სიდიდეს შეადგენს. მართლაც,

$$\mathbb{E}\left[S_n\right]=\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left(X_i-M_n\right)^2\right]=\frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{E}\left[X_i^2\right]+\mathbb{E}\left[M_n^2\right]-2\mathbb{E}\left[X_iM_n\right]\right)$$

$$=\mathbb{E}\left[X_i^2\right]+\mathbb{E}\left[M_n^2\right]-\frac{2}{n}\sum_{i=1}^{n}\mathbb{E}\left[X_iM_n\right].$$

დისპერსიის განსაზღვრების თანახმად, $\sigma^2=\mathbb{E}\left[X^2\right]-\mathbb{E}\left[X\right]^2$, ასე რომ

$$\mathbb{E}\left[X^2\right]=\sigma^2+m^2.$$

გარდა ამისა, $\mathbb{E}\left[M_n^2\right]=m^2+\frac{\sigma^2}{n}$, ვინაიდან M_n სიდიდის დისპერსია არის

$$\mathbb{V}\left[M_n\right]=\mathbb{E}\left[M_n^2\right]-\mathbb{E}\left[M_n\right]^2=\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)^2\right]-m^2=\frac{1}{n^2}\mathbb{E}\left[\sum_{i=1}^{n}X_i\sum_{i=j}^{n}X_j\right]-m^2$$

$$=\frac{1}{n^2}\mathbb{E}\left[\sum_{i=1}^{n}\left(X_i^2++\sum_{j\neq i}^{n}X_iX_j\right)\right]-m^2=\frac{1}{n}\left(\mathbb{E}\left[X^2\right]+\sum_{j\neq i}^{n}\mathbb{E}\left[X\right]^2\right)-m^2$$

მათემატიკური ლოდინის წრფივობის წყალობით და იმიტომაც, რომ $i\neq j$ პირობებში X_i და X_j სიდიდეები დამოუკიდებელია, რის გამოც $\mathbb{E}\left[X_iX_j\right]=\mathbb{E}\left[X_i\right]\mathbb{E}\left[X_j\right]=\mathbb{E}\left[X\right]^2$. მაშასადამე,

$$\mathbb{V}\left[M_n\right]=\frac{1}{n}\left(\underbrace{\sigma^2+m^2}_{\mathbb{E}[X^2]}+\underbrace{\left(n-1\right)m^2}_{\mathbb{E}[X]^2}\right)-m^2=\frac{\sigma^2}{n}.$$

დაბოლოს, მათემატიკური ლოდინის წრფივობის ძალით

$$\mathbb{E}\left[M_n\right]^2=\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)M_n\right]=\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[X_iM_i\right],$$

ასე რომ

$$\mathbb{E}\left[M_n^2\right]-\frac{2}{n}\sum_{i=1}^{n}\mathbb{E}\left[X_iM_n\right]=-\mathbb{E}\left[M_n^2\right].$$

ამრიგად, მიიღწევა :

$$\mathbb{E}\left[S_n\right]=\left(\sigma^2+m^2\right)-\left(m^2+\frac{\sigma^2}{n}\right)=\frac{n-1}{n}\sigma^2.$$

აქედან გამომდინარე, შემოდის *კორექტირებული ემპირიული დისპერსია*, რომელიც განისაზღვრება

$$S_n^*=\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-M_n\right)^2 \tag{14.4}$$

თანაფარდობით და წაუნაცვლებელია.

მიუხედავად ამისა, ემპირიული დისპერსიის წანაცვლება ისწრაფვის ნულისკენ, როცა $n \to \infty$. ამ შემთხვევაში ლაპარაკობენ *ასიმპტოტურად წაუნაცვლებელ* შემფასებელზე, ე.ი. შემფასებელ მაჩვენებელზე.

3.2 შემფასებლის (შემფასებელი მაჩვენებლის) სიზუსტე

კიდევ ერთხელ გადავავლოთ თვალი ნახატს 14.1. ორივე Θ და Θ'' შეფასება წაუნაცვლებელია. მაგრამ Θ'' შეფასების დისპერსია მეტია; ერთ-ერთს მისი რეალიზაციებიდან აქვს Θ-თან შედარებით მეტი ალბათობა აღმოჩნდეს θ-დან დაშორებული. ამრიგად, Θ'' არის *ნაკლებად ზუსტი*, ვიდრე Θ.

ჩაითვლება, რომ წაუნაცვლებელი შემფასებელი (შემფასებელი პარამეტრი) მით უფრო ზუსტია, რაც უფრო მცირეა მისი დისპერსია. წანაცვლებული შემფასებლის (შემფასებელი მაჩვენებლის) ზოგად შემთხვევაში წანაცვლება ასევე უნდა იყოს გათვალისწინებული სიზუსტის განსაზღვრისას. წანაცვლებულ შემფასებელს (შემფასებელ მაჩვენებელს) მცირე დისპერსიით შეუძლია უკეთესი (ე.ი. ჭეშმარიტ მნიშვნელობასთან უფრო ახლოს მყოფი) შეფასებების მოცემა, ვიდრე ნაკლებად წანაცვლებულ შემფასებელს (შემფასებელ მაჩვენებელს), მაგრამ უფრო დიდი დისპერსიით.

ამიტომ ზოგადი წერტილოვანი შემფასებლის სიზუსტის რაოდენობრივი შეფასებისთვის ჩვენ მის საშუალო კვადრატულ შეცდომას (ინგლ. Mean Square Error, MSE) ვიყენებთ, რომელიც შემდეგი სახით განისაზღვრება :

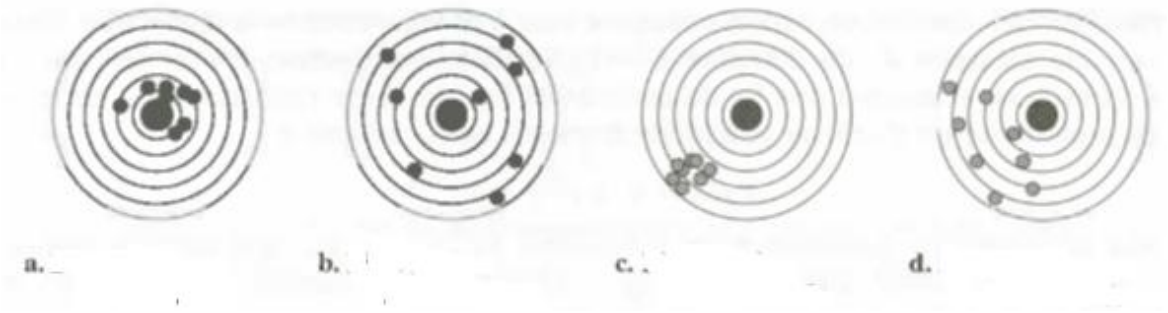
$$\mathrm{MSE}\left(\Theta_n\right) = \mathbb{E}\left[\left(\Theta_n - \theta\right)^2\right] = \mathbb{V}\left[\Theta_n - \theta\right] + \mathbb{E}\left[\Theta_n - \theta\right]^2 = \mathbb{V}\left[\Theta_n\right] + \mathrm{B}\left(\Theta_n\right)^2. \qquad (14.5)$$

შეიძლება გავიხსენოთ, რომ აქ $\mathrm{B}\left(\Theta_n\right)$ არის θ სიდიდის Θ_n შემფასებლის *წანაცვლება* (ინგლისურად : *the bias*) და იგი მოცემულია თანაფარდობით (14.2) : $\mathrm{B}(\Theta_n) = \mathbb{E}\left[\Theta_n\right] - \theta$.

რაც უფრო მცირეა შემფასებლის (შემფასებელი მაჩვენებლის) საშუალო კვადრატული შეცდომა, მით უფრო ზუსტი იქნება იგი.

სავსებით შესაძლებელია, რომ წანაცვლებულ შეფასებას წაუნაცვლებელზე უკეთესი სიზუსტე გააჩნდეს, თუ უკანასკნელს უფრო დიდი დისპერსია აქვს ! ამას ეწოდება *კომპრომისი წანაცვლებასა და დისპერსიას შორის.*

ნახატი 14.2 შემფასებლის წანაცვლებისა და დისპერსიის ცნებათა ილუსტრირებას იძლევა სატყორცნი შუბების (ისრების) ნაკრებთან ანალოგიის გამოყენებით.



a. მცირე წანაცვლება, მცირე დისპერსია ; b. მცირე წანაცვლება, დიდი დისპერსია ;

c. დიდი წანაცვლება, მცირე დისპერსია ; d. დიდი წანაცვლება, დიდი დისპერსია.

ნახატი 14.2 - წანაცვლებისა და დისპერსიის კონცეფციების ილუსტრაცია ანალოგიით სატყორცნ შუბებთან. შესაფასებელი სიდიდე არის მიზნის ცენტრი; შუბები არის შეფასებები. წრეწირებიდან თითოეული წარმოაჩენს განსხვავებულ შემფასებელს.

მაგალითი

დავუშვათ, რომ X არის შემთხვევითი ცვლადი $\left(X \in \mathcal{L}^2\right)$ ინტეგრირებადი კვადრატით, m მათემატიკური ლოდინით და σ^2 დისპერსიით. დავუშვათ ასვე, რომ $X_1, X_2, \ldots, X_n$ არის დამოუკიდებელი და ერთნაირად განაწილებული სიდიდეები, იმავე განაწილებით, რომელიც X სიდიდეს აქვს.

ემპირიული დისპერსიის წანაცვლების გამოანგარიშებისას ჩვენ დავინახეთ, რომ M_n სიდიდისათვის დისპერსია $\mathbb{V}\left[M_n\right] = \dfrac{\sigma^2}{n}$ თანაფარდობით აისახება.

ამრიგად, M_n ემპირიული საშუალოს საშუალო კვადრატულ შეცდომას (ინგლ. Mean Square Error, MSE) შემდეგი სახე აქვს :

$$\mathrm{MSE}\left(M_n\right) = \mathbb{V}\left[M_n\right] + \mathrm{B}\left(M_n\right)^2 = \frac{\sigma^2}{n}.$$

ეს შეცდომა მცირდება X სიდიდის დისპერსიასთან ერთად და მით ნაკლებ მნიშვნელობას იძენს, რაც უფრო დიდი იქნება დაკვირვებათა რიცხვი.

4 შეფასება დამაჯერებლობის მაქსიმუმით

ამ პუნქტში ჩვენ ყოველთვის ვიხილავთ X ნამდვილი შემთხვევითი ცვლადის $n \in \mathbb{N}^*$ ზომის $\left(X_1, X_2, \ldots, X_n\right)$ შემთხვევით ანარჩევს და $\theta \in \mathcal{S} \subseteq \mathbb{R}$ სიდიდეს, რომელიც უნდა შეფასდეს. X ცვლადის ალბათური განაწილების კანონი $\mathbb{P}_X$ სიმბოლოთი აღვნიშნოთ.

ჩვენ ეს-ეს არის ვნახეთ, როგორ უნდა დახასიათდეს Θ_n შემფასებელი (შემფასებელი მაჩვენებელი), რათა მოხდეს საუკეთესო შემფასებლის, საუკეთესო შემფასებელი მაჩვენებლის არჩევა რამდენიმედან. მაგრამ როგორ უნდა განხორციელდეს შემფასებლის, ე.ი. შემფასებელი მაჩვენებლის, *შეთავაზება* θ სიდიდისთვის ?

დავუშვათ, რომ $(x_1, x_2, \ldots, x_n)$ წარმოადგენს $(X_1, X_2, \ldots, X_n)$ შემთხვევითი ანარჩევის რეალიზაციას. მეთოდი, რომელსაც ჩვენ განვიხილავთ, მდგომარეობს ანარჩევის დამაჯერებლობის მაქსიმიზაციაში, სხვა სიტყვებით, θ სიდიდის მნიშვნელობის შეფასებისას სწორედ ამ ანარჩევის დაკვირვების ალბათობის მაქსიმუმის მიღწევაში.

მაგალითი

დავუშვათ, რომ გვაინტერესებს დედაქალაქში უფროსკლასელთა წარმატებები და ჩვენს ხელთ არის დაკვირვებანი ამ ქალაქის რამდენიმე საშუალო სკოლიდან.

ორი — «წარმატება» ან «წარუმატებლობა» — დაკვირვების მოდელირება ხდება როგორც ალბათობის $\mathbb{P}_X$ კანონით განაწილებული X შემთხვევითი ცვლადის რეალიზაცია $E = [0,1]$ სიმრავლეზე. ამ დროს 0 შეესაბამება «წარუმატებლობას», ხოლო 1 — «წარმატებას». კლასიკური არჩევანია ალბათობის ამ კანონისთვის p პარამეტრიანი ბერნულის განაწილების გამოყენება:

$$\mathbb{P}_X(X = x) = p^x (1-p)^{1-x}.$$

არსებული დაკვირვებები შეადგენს $(x_1, x_2, \ldots, x_n)$ ანარჩევს, რომელიც დამოუკიდებელი და ერთნაირად (*X შემთხვევითი ცვლადის ალბათობათა იმავე $\mathbb{P}_X$ კანონით*) განაწილებული კომპონენტების შემთხვევითი $(X_1, X_2, \ldots, X_n)$ ანარჩევის რეალიზაცია არის.

ვეცადოთ ბერნულის კანონის p პარამეტრის შეფასება ამ ანარჩევით.

დავუშვათ, რომ საწყისი ანარჩევი შეიცავს $n = 500$ მოწაფეს და მათ შორის $b = 450$ პირს გამოცდა ჩაბარებული აქვს.

$p = 50\%$ მნიშვნელობა ნაკლებად დამაჯერებელი არის; $p = 90\%$ მნიშვნელობა კი — გაცილებით უფრო დამაჯერებელი. სწორედ ეს არის ის ცნება, რომლის ფორმალიზებას ვაპირებთ მომდევნო მასალაში.

4.1 დამაჯერებლობა

$(x_i, x_2, \ldots, x_n)$ ანარჩევის დამაჯერებლობა განსაზღვრავს, შესაძლებლობის შესაბამისად, რამდენად სარწმუნოა ამ ანარჩევის დაკვირვება შესაფასებელი სიდიდის მნიშვნელობიდან გამომდინარე.

განსაზღვრება 8.4 (ანარჩევის ალბათობა) დავუშვათ, რომ $(x_i, x_2, \ldots, x_n)$ არის X შემთხვევითი სიდიდის ანარჩევი. ნებისმიერი $t \in S$ პარამეტრისთვის აღვნიშნოთ $\mathbb{P}_{X;t}$ სიმბოლოთი X ცვლადის ალბათური კანონი, რომელიც პარამეტრიზებულია t სიდიდით. დავუშვათ ასევე, რომ $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$-ზე არსებობს μ ზომა, ისეთი, რომ $\mathbb{P}_{X;t} = f_{t\mu}$, ან $f_t : \mathbb{R} \mapsto \mathbb{R}_+$ ფორმით აღწერილი $\mathbb{P}_{X;t}$ კანონი μ – განზომადია.

იმ შემთხვევაში, როცა X ცვლადი დისკრეტულია, μ წარმოადგენს თვლად ზომას, ხოლო f_t — შემთხვევითი დისკრეტული X ცვლადის *ალბათობის მასის ფუნქციას* (probability mass function, pmf)[5].

თუ X შემთხვევითი ცვლადი მოცემულია ალბათობათა სიმკვრივით (ე.ი. უწყვეტია), მაშინ μ არის ე.წ. ლებეგის (Henri Léon Lebesgue, 1875-1941) ზომა, ხოლო f_t — შემთხვევითი X ცვლადის სიმკვრივე. ასე რომ, $(x_i, x_2\ ,\ \ldots\ , x_n)$ ანარჩევის დამაჯერებლობა არის t არგუმენტის ფუნქცია, რომელიც განისაზღვრება შემდეგი თანაფარდობით :

$$L\left(x_1, x_2, \ldots, x_n; t\right) = \prod_{i=1}^{n} f_t\left(x_i\right). \tag{14.6}$$

ხშირად გვხვდება ასეთი აღნიშვნაც :

$$L\left(x_1, x_2, \ldots, x_n; t\right) = \prod_{i=1}^{n} \mathbb{P}_t\left(X = x^i\right), \tag{14.7}$$

მაგრამ იგი მცდარია, როცა X შემთხვევით სიდიდეს აქვს ალბათობათა განაწილების სიმკვრივე. ამ შემთხვევაში $\mathbb{P}_t\left(X = x^i\right) = 0$.

უნდა აღინიშნოს, რომ $\left(X_1, X_2, \ldots, X_n\right)$ შემთხვევითი ანარჩევის ალბათური კანონი ჩაიწერება $\mathbb{P}_{X_1, X_2, \ldots} = \prod_{i=1}^{n} \mathbb{P}_{X;t}$ ფორმით, რადგან X_i — დამოუკიდებელი და ერთნაირად განაწილებული შემთხვევითი სიდიდეებია.

4.2 შეფასება მაქსიმალური დამაჯერებლობის მეთოდით

განსაზღვრება 14.5

დავუშვათ, რომ $\left(x_1\ ,\ x_2\ ,\ \ldots\ ,\ x_n\right)$ არის X შემთხვევითი სიდიდის ანარჩევი. ნებისმიერი $t \in \mathcal{S}$ პარამეტრისთვის აღვნიშნოთ $\mathbb{P}_{X;t}$ სიმბოლოთი t პარამეტრის მნიშვნელობით პარა-მეტრიზებული X შემთხვევითი სიდიდის განაწილება.

მაშინ შეფასებად θ სიდიდისთვის *დამაჯერებლობის მაქსიმუმით* (ინგლისურად : *maximum likelihood estimate,* MLE) უწოდებენ $\hat{\theta}_{\text{MLE}}$ სიდიდეს, რომელიც $\left(x_1\ ,\ x_2\ ,\ \ldots\ ,\ x_n\right)$ ანარჩევის დამაჯერებლობას მაქსიმუმს ანიჭებს :

---

[5] *ალბათობის მასის ფუნქცია* (ინგლ. PMF, probability mass function) იძლევა იმის ალბათობას, რომ დისკრეტული შემთხვევითი ცვლადი ზუსტად უდრის რაღაც მნიშვნელობას. ზოგჯერ მას ასევე უწოდებენ სიმკვრივის დისკრეტულ ფუნქციასაც. ალბათობათა მასის ფუნქცია ხშირად არის ალბათობათა დისკრეტული განაწილების განსაზღვრის ძირითადი საშუალება და ასეთი ფუნქციები არსებობს როგორც სკალარული, ასევე მრავალგანზომილებიანი შემთხვევითი სიდიდეებისთვის, როცა მათი მნიშვნელობების არე დისკრეტულია. ალბათობის მასის ფუნქცია განსხვავდება ალბათობის სიმკვრივის ფუნქციისგან (ინგლ. PDF, probability density function) იმით, რომ უკანასკნელი დაკავშირებულია უწყვეტ და არა დისკრეტულ შემთხვევით სიდიდეებთან. ალბათობის დასადგენად ალბათობის სიმკვრივის ფუნქცია უნდა იყოს ინტეგრირებული ინტერვალზე.

$$\hat{\theta}_{\text{MLE}} \in \arg\max_{t \in \mathcal{S}} \prod_{i=1}^{n} f_t\left(x_i\right), \tag{14.8}$$

სადაც *მაქსიმიზაციის არგუმენტი* (argmax ან arg max) არის არგუმენტის ის მნიშვნელობა, რომელზეც მოცემული გამოსახულება მაქსიმუმს აღწევს.

მაქსიმალური დამაჯერებლობის შეფასება θ სიდიდისათვის არის ნამდვილი შემთხვევითი $\hat{\theta}_{\text{MLE}}$ ცვლადი, რომლის მნიშვნელობა, როცა $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$, მოცემულია $\hat{\theta}_{\text{MLE}}$ სიდიდით.

გამოთვლათა გასამარტივებლად მაქსიმიზაციის არჩევა ხშირად ხდება არა უშუალოდ დამაჯერებლობიდან, არამედ დამაჯერებლობის ლოგარითმიდან, რომელსაც ზოგჯერ *ლოგარითმულ დამაჯერებლობას* უწოდებენ :

$$\hat{\theta}_{\text{MLE}} \in \arg\max_{t \in \mathcal{S}} \prod_{i=1}^{n} \ln f_t\left(x_i\right). \tag{14.9}$$

4.3 ბერნულის განაწილების პარამეტრის მაქსიმალური დამაჯერებლობის შეფასება

ეს უმარტივესი დისკრეტული განაწილება შვეიცარიელი მათემატიკოსის იაკობ ბერნულის (Jakob/Jacques Bernoulli, 1654/1655-1705) სახელს ატარებს.

ალბათობათა თეორიასა და სტატისტიკაში ბერნულის განაწილება ალბათობათა დისკრეტული განაწილებაა, რომელიც ნებისმიერი ბუნების შემთხვევითი ექსპერიმენტის მოდელირებას ახორციელებს, თუ წინასწარ არის ცნობილი წარმატების ან წარუმატებლობის ალბათობა.

თეორემა 2.1

დავუშვათ, რომ X შემთხვევით ცვლადს აქვს ბერნულის განაწილება p პარამეტრით. დავუშვათ ასევე, რომ $\left\{x^1, x^2, \ldots, x^n\right\}$ არის X ცვლადის ანარჩევი, ხოლო $b = \sum_{i=1}^{n} x_i$ — ერთიანების (1) რაოდენობა ამ ანარჩევში. მაქსიმალური დამაჯერებლობის $\hat{p}_{\text{MLE}}$ შეფასება p პარამეტრისათვის ანარჩევის საშუალო მნიშვნელობაა :

$$\hat{p}_{\text{MLE}} = \frac{b}{n}. \tag{14.10}$$

მტკიცებულება

p პარამეტრისათვის მაქსიმალური დამაჯერებლობის შეფასება არის :

$$\begin{aligned}\hat{p}_{\text{MLE}} &= \arg\max_{t \in [0,1]} \sum_{i=1}^{n} \ln \mathbb{P}_{x;t}\left(X = x_i\right) = \arg\max_{t \in [0,1]} \sum_{i=1}^{n} \ln\left(t^{x_i}\left(1-t\right)^{1-x_i}\right) \\ &= \arg\max_{t \in [0,1]} \sum_{i=1}^{n} x_i \ln t + \left(n - \sum_{i=1}^{n} x_i\right) \ln\left(1-t\right).\end{aligned}$$

ფუნქცია, რომლის მინიმიზაციას ვცდილობთ, არის ℓ :

$$\ell : t \mapsto \sum_{i=1}^{n} x_i \ln t + \left(n - \sum_{i=1}^{n} x_i \right) \ln(1-t).$$

თუ $b = n$, მაშინ $\ell(t) = n \ln t$ და ℓ ფუნქცია მაქსიმალურია, როცა $t = 1$.

თუ $b = 0$, მაშინ $\ell(t) = n \ln(1-t)$ და ℓ ფუნქცია მაქსიმალურია, როცა $t = 0$.

დაბოლოს, თუ $0 < b < n$, მაშინ ფუნქცია, რომლის მინიმიზაცია გვსურს, ჩაზნექილია, ამიტომ შეგვიძლია ხსენებული ფუნქციის მაქსიმიზაცია მისი წარმოებულის განულების გზით :

$$\frac{d\ell}{dt} = \sum_{i=1}^{n} x_i \frac{1}{t} - \left(n - \sum_{i=1}^{n} x_i \right) \frac{1}{1-t},$$

რაც გვაძლევს

$$(1 - \hat{p}_{\text{MLE}}) \left(\sum_{i=1}^{n} x_i \right) - \hat{p}_{\text{MLE}} \left(n - \sum_{i=1}^{n} x_i \right) = 0$$

თანაფარდობას და, ამრიგად, b რიცხვის ყველა შესაძლო მნიშვნელობისათვის

$$\hat{p}_{\text{MLE}} = \frac{b}{n}.$$

მაგალითი

დავუბრუნდეთ სასკოლო გამოცდაში წარმატების მაგალითს. მაქსიმალური დამაჯერებლობის შეფასება p პარამეტრისათვის ეს უბრალოდ ანარჩევის საშუალო მნიშვნელობაა. სახელდობრ, ხსენებულ მაგალითში $p = 450 / 500 = 90\%$.

4.4 ნორმალური განაწილების პარამეტრთა მაქსიმალური დამაჯერებლობის შეფასება

თეორემა 2.2

დავუშვათ, რომ X შემთხვევით სიდიდეს აქვს ნორმალური განაწილება μ და σ პარამეტრებით, ხოლო ალბათობათა სიმკვრივე განისაზღვრება შემდეგი ფორმულით :

$$p_X(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right).$$

დავუშვათ ასევე, რომ $\{x^1, x^2, \ldots, x^n\}$ არის X შემთხვევითი სიდიდის ანარჩევი. მაქსიმალური დამაჯერებლობის შეფასება μ პარამეტრისათვის ამ ანარჩევის საშუალო მნიშვნელობაა :

$$\hat{\mu}_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^{n} x^i,$$

ხოლო მაქსიმალური დამაჯერებლობის შეფასება σ პარამეტრისათვის — ანარჩევის სტანდარტული (საშუალო კვადრატული) გადახრა :

$$\hat{\sigma}^2_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n}\left(x^i - \hat{\mu}_{\text{MLE}}\right)^2 .$$

მტკიცებულება

ანარჩევის დამაჯერებლობის ლოგარითმს შემდეგი სახე აქვს :

$$\sum_{i=1}^{n}\ln p_X\left(x^i\right) = \sum_{i=1}^{n} -\log\left(\sigma\sqrt{2\pi}\right) - \frac{\left(x^i-\mu\right)^2}{2\sigma^2} .$$

ეს ჩაზნექილი ფუნქციაა, რომლის გრადიენტი μ პარამეტრით

$$\frac{1}{\sigma^2}\left(n\mu - \sum_{i=1}^{n} x^i\right)$$

გამოსახულებას გვაძლევს.

თუ ამ გრადიენტს მივანიჭებთ ნულოვან მნიშვნელობას, მაშინ მივიღებთ :

$$\hat{\mu}_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n} x^i .$$

დამაჯერებლობის ლოგარითმის გრადიენტი σ პარამეტრით კი შემდეგი სახისაა :

$$\frac{\partial L}{\partial \sigma} = -n\sigma^2 + \sum_{i=1}^{n}\left(x^i - \hat{\mu}_{\text{MLE}}\right)^2 .$$

ახლა თუ ამ გრადიენტს მივანიჭებთ ნულოვან მნიშვნელობას, მაშინ გვექნება :

$$\hat{\sigma}^2_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n}\left(x^i - \hat{\mu}_{\text{MLE}}\right)^2 .$$

4.5 სტოქასტიკური ვექტორის შემთხვევა

ალბათობის განსაზღვრება შეიძლება იყოს გავრცელებული სტოქასტიკური ვექტორის შემთხვევაზე. დავუშვათ, რომ (U,V) არის ისეთი შემთხვევითი ვექტორების წყვილი, სადაც U სიმკვრივიანი შემთხვევითი ვექტორია, ხოლო V — დისკრეტული შემთხვევითი ვექტორი. $p_{U;\vec{t}}$ ნოტაციით მოცემულ სიდიდეს ეწოდება U ვექტორის $\vec{t}$ პარამეტრთა ვექტორით პარამეტრიზებული სიმკვრივე, ხოლო $\mathbb{P}_{V;\vec{\tau}}$ სიმბოლოთი აღინიშნება V ვექტორის ალბათობათა განაწილების კანონი, რომელიც პარამეტრიზებულია $\vec{\tau}$ პარამეტრთა ვექტორით. დავუშვათ, რომ $\left(\left(\vec{u}^1,\vec{v}^1\right),\left(\vec{u}^2,\vec{v}^2\right),\ldots,\left(\vec{u}^n,\vec{v}^n\right)\right)$ წარმოადგენს (U,V) წყვილის ანარჩევს.

ამ ანარჩევის დამაჯერებლობა მოიცემა თანაფარდობით :

$$L\left(\left(\vec{u}^1,\vec{v}^1\right),\left(\vec{u}^2,\vec{v}^2\right),\ldots,\left(\vec{u}^n,\vec{v}^n\right);\vec{t},\vec{\tau}\right) = \prod_{i=1}^{n} p_{U;\vec{t}}\left(\vec{u}^i\right)\mathbb{P}_{V;\vec{\tau}}\left(\vec{v}^i\right). \qquad (14.11)$$

ასეთი ზოგადი ფორმულირება დამაჯერებლობის ჩაწერის საშუალებას გვაძლევს შემდეგ შემთხვევებში :

a) $\left(\vec{x}^1,\vec{x}^2,\ldots,\vec{x}^n\right)$ არის p განზომილების დისკრეტული შემთხვევითი ვექტორის ანარჩევი და ამ ვექტორის ალბათობათა განაწილების $\mathbb{P}_{X;\vec{\tau}}$ კანონი პარამეტრიზებულია $\vec{\tau}$ ვექტორით. მაშინ :

$$L\left(\vec{x}^1,\vec{x}^2,\ldots,\vec{x}^n;\vec{t}\right)=\prod_{i=1}^{n}\mathbb{P}_{X;\vec{\tau}}\left(\vec{x}^i\right)=\prod_{i=1}^{n}\mathbb{P}_{X;\vec{\tau}}\left(x_1^i,x_2^i,\ldots,x_p^i\right). \tag{14.12}$$

b) $\left(\vec{x}^1,\vec{x}^2,\ldots,\vec{x}^n\right)$ არის p განზომილების სიმკვრივიანი შემთხვევითი ვექტორის ანარჩევი და ამ ვექტორის ალბათობათა განაწილების სიმკვრივე მოცემულია $p_{X;\vec{t}}$ ნოტაციით. ასე რომ :

$$L\left(\vec{x}^1,\vec{x}^2,\ldots,\vec{x}^n;\vec{t}\right)=\prod_{i=1}^{n}p_{X;\vec{t}}\left(\vec{x}^i\right)=\prod_{i=1}^{n}p_{X;\vec{t}}\left(x_1^i,x_2^i,\ldots,x_p^i\right). \tag{14.13}$$

c) $\left(\left(\vec{x}^1,\vec{x}^2,\ldots,\vec{x}^n\right),y^i\right)$ ვექტორთა (X,Y) წყვილის ანარჩევია და ამ წყვილში X არის p განზომილების დისკრეტული შემთხვევითი ვექტორი, რომლის ალბათობათა განაწილების $\mathbb{P}_{X;\vec{t}}$ კანონი პარამეტრიზებულია $\vec{t}$ ვექტორით, ხოლო Y არის დისკრეტული შემთხვევითი ცვლადი, რომლის $\mathbb{P}_{Y;\vec{\tau}}$ კანონი პარამეტრიზებულია $\vec{\tau}$ ვექტორით. ამ შემთხვევაში :

$$\left.\begin{aligned}&L\left(\left(\vec{x}^1,y^1\right),\left(\vec{x}^2,y^2\right),\ldots,\left(\vec{x}^n,y^n\right);\vec{t},\vec{\tau}\right)=\prod_{i=1}^{n}\mathbb{P}(X=\vec{x}^i,Y=y^i)=\\&=\prod_{i=1}^{n}\mathbb{P}\left(Y=y^i\mid X=\left(x_1^i,x_2^i,\ldots,x_p^i\right)\right)\mathbb{P}\left(X=\left(x_1^i,x_2^i,\ldots,x_p^i\right)\right)\end{aligned}\right\}. \tag{14.14}$$

d) $\left(\left(\vec{x}^1,\vec{x}^2,\ldots,\vec{x}^n\right),y^i\right)$ ვექტორთა (X,Y) წყვილის ანარჩევია და ამ წყვილში X არის სიმკვრივიანი შემთხვევითი ვექტორი p განზომილებით და $p_{\vec{t}}$ ალბათობათა განაწილების სიმკვრივით, ხოლო Y დისკრეტული შემთხვევითი ცვლადია.

ამრიგად :

$$\left.\begin{aligned}&L\left(\left(\vec{x}^1,y^1\right),\left(\vec{x}^2,y^2\right),\ldots,\left(\vec{x}^n,y^n\right);\vec{t}\right)=\prod_{i=1}^{n}p_{\vec{t}}\left(\vec{x}^i\right)\mathbb{P}\left(Y=y^i\right)=\\&=\prod_{i=1}^{n}p_{\vec{t}}\left(\vec{x}_1^i,\vec{x}_2^i,\ldots,\vec{x}_p^i\right)\mathbb{P}\left(Y=y^i\right)=\\&=\prod_{i=1}^{n}\mathbb{P}\left(Y=y^i\mid X=\left(\vec{x}_1^i,\vec{x}_2^i,\ldots,\vec{x}_p^i\right)\right)p_{\vec{t}}\left(\vec{x}_1^i,\vec{x}_2^i,\ldots,\vec{x}_p^i\right)\end{aligned}\right\}. \tag{14.15}$$

e) $\left(\left(\vec{x}^1,\vec{x}^2,\ldots,\vec{x}^n\right),y^i\right)$ ვექტორთა (X,Y) წყვილის ანარჩევია და ამ წყვილში X არის სიმკვრივიანი შემთხვევითი ვექტორი p განზომილებით და $p_{X;\vec{t}}$ სიმკვრივით, ხოლო Y სიმკვრივიანი შემთხვევითი ცვლადია $p_{Y;\vec{\tau}}$ სიმკვრივით. ასე რომ :

$$\left.\begin{aligned} L\left(\left(\vec{x}^1,y^1\right),\left(\vec{x}^2,y^2\right),\ldots,\left(\vec{x}^n,y^n\right);\vec{t}\right)&=\prod_{i=1}^{n}p_{X;\vec{t}}\left(\vec{x}^i\right)p_{Y;\vec{\tau}}\left(y^i\right)=\\ &=\prod_{i=1}^{n}p_{X;\vec{t}}\left(\vec{x}_1^i,\vec{x}_2^i,\ldots,\vec{x}_p^i\right)p_{Y;\vec{\tau}}\left(y^i\right)=\\ &=\prod_{i=1}^{n}p_{Y|X=\left(\vec{x}_1^i,\vec{x}_2^i,\ldots,\vec{x}_p^i\right);\vec{t},\vec{\tau}}\left(y^i\right)p_{X;\vec{t}}\left(\vec{x}_1^i,\vec{x}_2^i,\ldots,\vec{x}_p^i\right)\end{aligned}\right\}. \tag{14.16}$$

5 ბაიესის შეფასება

დავუშვათ, რომ θ პარამეტრის იმ მნიშვნელობათა შესახებ, რომლებიც ამ პარამეტრმა შეიძლება მიიღოს, გვაქვს გარკვეული საფუძვლიანი ვარაუდები. ეს ინფორმაცია შეიძლება ძალიან სასარგებლო იყოს, განსაკუთრებით მაშინ, როცა დაკვირვებათა რაოდენობა მცირეა.

რომ ვისარგებლოთ ამით, ჩვენ მივმართავთ ნადვილ შემთხვევით Θ ცვლადს (მნიშვნელობებით $\mathcal{S}-$ზე), რომლისთვისაც ალბათობათა განაწილების $\mathbb{P}_\Theta$ კანონი არის *აპრიორული*, ე.ი. ანარჩევის დაკვირვებამდე განსაზღვრული ან დადგენილი. ახლა კი ვისარგებლოთ ბაიესის ფორმულით, რათა გამოვსახოთ *აპოსტერიორული*, ე.ი. Θ ცვლადის შემთხვევითი ანარჩევით განპირობებული, *კანონი*.

თუ Θ დისკრეტულია, მაშინ ბაიესის ფორმულა კანონის ჩაწერის საშუალებას გვაძლევს $\Theta\,|\,X_1,X_2,\ldots,X_n-$თვის :

$$\mathbb{P}_{\Theta|X_1,X_2,\ldots,X_n}\left(\Theta=t\right)=\frac{\mathbb{P}_\Theta\left(t\right)\prod_{i=1}^{n}f_t\left(x_i\right)}{\sum_{u\in\mathcal{S}}\mathbb{P}_\Theta\left(u\right)\prod_{i=1}^{n}f_u\left(x_i\right)}. \tag{14.17}$$

აქ, როგორც უკვე ვიცით, თუ X ცვლადი დისკრეტულია, მაშინ f_t არის შემთხვევითი დისკრეტული X ცვლადის *ალბათობის მასის ფუნქცია,* ხოლო თუ X შემთხვევითი ცვლადი მოცემულია ალბათობათა სიმკვრივით (ე.ი. უწყვეტია), მაშინ f_t არის შემთხვევითი X ცვლადის სიმკვრივე.

თუ Θ ცვლადს აქვს g_Θ სიმკვრივე, მაშინ $\Theta\,|\,X_1,X_2,\ldots,X_n$ ცვლადს ასევე აქვს სიმკვრივე და ბაიესის ფორმულა მისი სიმკვრივის ჩაწერის საშუალებას გვაძლევს შემდეგი ფორმით :

$$p_{\Theta|X_1,X_2,\ldots,X_n}\left(t\right)=\frac{g_\Theta(t)\prod_{i=1}^{n}f_t\left(x_i\right)}{\int_{\mathcal{S}}p_\Theta\left(u\right)\prod_{i=1}^{n}f_u\left(x_i\right)du}. \tag{14.18}$$

სხვა სიტყვებით რომ ვთქვათ, ანარჩევზე დაკვირვება Θ ცვლადის აპრიორული კანონის კორექტირების საშუალებას იძლევა ამ ცვლადის აპოსტერიორულ კანონში. სწორედ ეს იდეა უდევს საფუძვლად *ბაიესურ დანასკვს.*

5.1 შეფასება აპოსტერიორული მაქსიმუმით

განსაზღვრება 8.6 (შეფასება აპოსტერიორული მაქსიმუმით) Θ ცვლადის *შეფასება აპოსტერიორული მაქსიმუმით* განისაზღვრება როგორც ისეთი მნიშვნელობა $\mathcal{S}-$დან,

რომელიც Θ ცვლადის აპოსტერიორული კანონის მაქსიმიზებას (მაქსიმიზაციას) ახდენს. მაშასადამე, ამ ცვლადის დისკრეტულ შემთხვევაში :

$$\hat{\theta}_{\text{MAP}} \in \arg\max_{t \in \mathcal{S}} \mathbb{P}_{\Theta|X_1, X_2, \ldots, X_n}\left(\Theta = t\right), \tag{14.19}$$

ხოლო ის შემთხვევა, როცა Θ ცვლადს სიმკვრივე აქვს, შემდეგი თანაფარდობით აისახება :

$$\hat{\theta}_{\text{MAP}} \in \arg\max_{t \in \mathcal{S}} \; p_{\Theta|X_1, X_2, \ldots, X_n}\left(t\right). \tag{14.20}$$

შენიშვნა

მაქსიმალური აპოსტერიორული შეფასება ემთხვევა მაქსიმალური დამაჯერებლობის შეფასებას, თუ აპრიორულ განაწილებად *თანაბარი განაწილება* გამოიყენება.

5.2 ბაიესის შეფასება

აპოსტერიორული მაქსიმუით შეფასება შეზღუდულია იმ შემთხვევაში, როცა აპოსტერიორული განაწილება არის მულტიმოდალური (მაქსიმუმის რამდენიმე პიკის, ანუ მოდის შემცველი); მართლაც, ყველაზე დიდი მოდის გამოვლენა, იდენტიფიცირება შეიძლება ძნელი აღმოჩნდეს გრადიენტული ალგორითმებით. გარდა ამისა, აპოსტერიორული მაქსიმუით შეფასება ითვალისწინებს აპოსტერიორული განაწილების მხოლოდ ერთ წერტილს და არ განიხილავს მას მთლიანობაში.

განსაზღვრება 8.7 (ბაიესის შეფასება)

ბაიესის შეფასება θ ცვლადისათვის განისაზღვრება როგორც ისეთი მნიშვნელობა $\mathcal{S}-$დან, რომელიც დანახარჯების გარკვეული ფუნქციის მინიმიზებას (გამინიმუმებას) ახდენს მათემატიკური ლოდინის მიხედვით Θ ცვლადის აპოსტერიორულ კანონზე. ეს განსაზღვრება ზოგადია და დამოკიდებულია დანახარჯთა გამოყენებულ ფუნქციაზე. ლოგიკურია ყველაზე გავრცელებულ განსაზღვრებაზე შეჩერება და, აქედან გამომდინარე, (მე-14 ლექციის 3.2 პუნქტში მოცემული) საშუალო კვადრატული შეცდომის ამოქმედება პრაქტიკულ ინსტრუმენტად.

ბაიესის შეფასება θ *ცვლადის საშუალო კვადრატული შეცდომისთვის,* ამრიგად, განისაზღვრება შემდეგი ფორმულით :

$$\hat{\theta}_{\text{Bayes}} \in \arg\max_{t \in \mathcal{S}} \mathbb{E}\left[\left(\left(\Theta \mid X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\right) - t\right)^2\right]. \tag{14.21}$$

თეორემა 2.3

ბაიესის შეფასება საშუალო კვადრატული შეცდომისთვის არის Θ სიდიდის აპოსტერიორული განაწილების მათემატიკური ლოდინი :

$$\hat{\theta}_{\text{Bayes}} = \mathbb{E}\left[\Theta \mid X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\right]. \tag{14.22}$$

მტკიცებულება

დავუშვათ, რომ $t \in \mathcal{S}$. შემოვიღოთ $W = \left[\Theta \mid X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\right]$ აღნიშვნა ამ გრძელი ჩანაწერის გასამარტივებლად. მაშინ :

$$\mathbb{E}\left[\left(W-t\right)^2\right] = \mathbb{E}\left[W^2\right] + t^2 - 2t\,\mathbb{E}\left[W\right] = \mathbb{E}\left[W^2\right] + \left(\mathbb{E}\left[W\right] - t\right)^2 - \mathbb{E}\left[W\right]^2 .$$

ვინაიდან არც $\mathbb{E}\left[W^2\right]$ და არც $\mathbb{E}\left[W\right]^2$ დამოკიდებული არ არის t-ზე, ამიტომ $\hat{\theta}_{\text{Bayes}}$ მიიღება $\mathbb{E}\left[\left(W-t\right)^2\right]$ მათემატიკური ლოდინის მინიმიზაციის გზით და, მაშასადამე, $\hat{\theta}_{\text{Bayes}} = \mathbb{E}\left[W\right]$.

მაგალითი

კვლავ დავუბრუნდეთ სასკოლო გამოცდის წარმატებულობის კოეფიციენტის მაგალითს. ახლა დავუშვათ, რომ p არის ისეთი Θ შემთხვევითი ცვლადის რეალიზაცია, რომელიც ემორჩილება ბეტა-კანონს (α, β) პარამეტრებით და რომლის სიმკვრივე, წინა შემთხვევის მსგავსად აღვნიშნოთ g_Θ სიმბოლოთი.

$\alpha, \beta > 0$ პარამეტრების მქონე ბეტა-კანონის $0 \le u \le 1$ ჩაკეტილ შუალედზე (მონაკვეთზე, სეგმენტზე) განსაზღვრული სიმკვრივე მოიცემა ფორმულით :

$$f_{\alpha,\beta}\left(u\right) = \frac{u^{\alpha-1}\left(1-u\right)^{\beta-1}}{B\left(\alpha,\beta\right)}, \tag{14.23}$$

სადაც $B\left(\alpha,\beta\right) = \dfrac{\Gamma\left(\alpha\right)\Gamma\left(\beta\right)}{\Gamma\left(\alpha+\beta\right)}$, ხოლო Γ — გამა-ფუნქციაა. ამ განაწილების მათემატიკური ლოდინია $\dfrac{\alpha}{\alpha+\beta}$. ბაიესის შეფასების გამოსათვლელად p რეალიზაციისათვის აუცილებელია $\left(\Theta \mid X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\right)$ ცვლადის კანონის ცოდნა. ვინაიდან Θ ცვლადს აქვს სიმკვრივე, ისევე როგორც $W = \left[\Theta \mid X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\right]$ ცვლადსაც, ამიტომ ბაიესის კანონი, იმ დაშვებით, რომ $X_i\,(i = 1, 2, \ldots, n)$ სიდიდეები დამოუკიდებელი და ერთნაირად განაწილებულია, W ცვლადის სიმკვრივის ჩაწერის საშუალებას შემდეგი სახით იძლევა :

$$\left.\begin{aligned} g_{\Theta \mid X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n}(t) &= \frac{\mathbb{P}\left(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n \mid \Theta = t\right) g_\Theta\left(t\right)}{\mathbb{P}\left(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\right)} \\ &= \frac{1}{\mathbb{P}\left(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\right) B\left(\alpha,\beta\right)} \prod_{i=1}^{n} t^{x_i}\left(1-t\right)^{1-x_i}\, t^{\alpha-1}\left(1-t\right)^{\beta-1} \\ &= \frac{1}{\mathbb{P}\left(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\right) B\left(\alpha,\beta\right)} t^{b+\alpha-1}(1-t)^{n-b+\beta-1} \end{aligned}\right\}.$$

ამ გამოსახულებაში ადვილი ამოსაცნობია ახალი ბეტა-კანონის სიმკვრივე. ასე რომ, $\left(\Theta \mid X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\right)$ ემორჩილება ბეტა-კანონს $\left(b+\alpha\right)$ და $\left(n-b+\beta\right)$ პარამეტრებით.

მაშასადამე, ბაიესის შეფასება p რეალიზაციისათვის ასეთი იქნება :

$$\hat{p}_{\text{Bayes}} = \mathbb{E}\left[\Theta \mid X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\right] = \frac{\left(b+\alpha\right)}{\left(b+\alpha\right)+\left(n-b+\beta\right)} = \frac{b+\alpha}{n+\alpha+\beta}.$$

ეს პირველი ტოლობა მიიღება ფორმულიდან, რომელიც ბეტა-კანონის მათემატიკურ ლოდინს იძლევა.

შენიშვნა

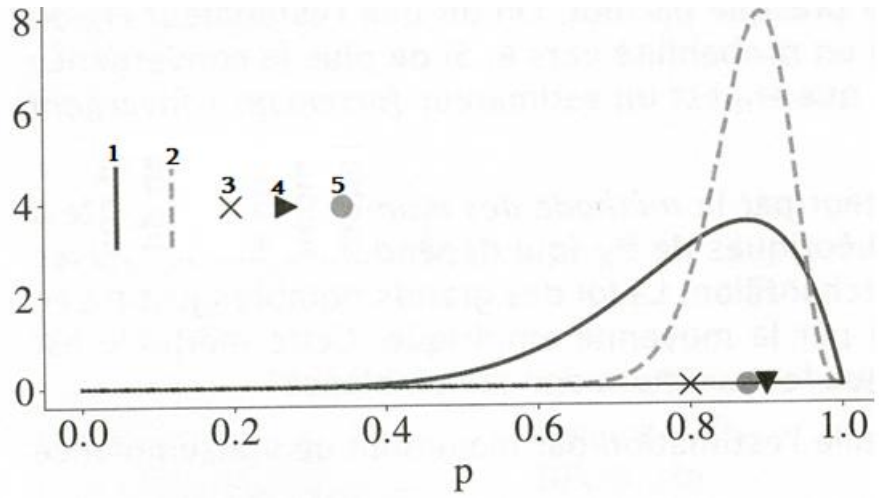
ზემოთ მოყვანილ მაგალითში მიღებული ბაიესური შეფასება შეიძლება შემდეგი სახით იყოს გადაწერილი :

$$\hat{p}_{\text{Bayes}} = \frac{\alpha+\beta}{n+\alpha+\beta}\,\mathbb{E}\left[\Theta_n\right] + \frac{n}{n+\alpha+\beta}\,\hat{p}_{\text{MLE}}\,.$$

ამრიგად, p პარამეტრის ბაიესური შეფასება არის ამ პარამეტრის აპრიორული განაწილების $\mathbb{E}\left[\Theta_n\right]$ მათემატიკური ლოდინის და $\hat{p}_{\text{MLE}}$ მაქსიმალური დამაჯერებლობის შეფასების (ინგლ. maximum likelihood estimate, MLE) წრფივი კომბინაცია.

გარდა ამისა, აპრიორული მათემატიკური ლოდინის $\frac{\alpha+\beta}{n+\alpha+\beta}$ მულტიპლიკაციური კოეფიციენტი ანარჩევის n ზომის კლებადი ფუნქციაა, მაშინ როცა დამაჯერებლობის მაქსიმუმით დადგენილი შეფასების $\frac{n}{n+\alpha+\beta}$ მულტიპლიკაციური კოეფიციენტი ანარჩევის n ზომის ზრდადი ფუნქციაა. ასე რომ, რაც უფრო დიდია ანარჩევის ზომა, მით უფრო მეტ ნდობას იჩენს ბაიესური შემფასებელი მონაცემებისადმი და შორდება პარამეტრის აპრიორულ მათემატიკურ ლოდინს, რომელთანაც მეტი სიახლოვე აქვს მცირე ანარჩევის დროს.

ნახატი 14.3 ამ მაგალითის ილუსტრირებას იძლევა.



ნახატი 14.3 — აპრიორული და აპოსტერიორული კანონი p პარამეტრისთვის მაგალითში წარმატების შესახებ სასკოლო გამოცდაზე. თუ არ გვაქვს მონაცემები, მაშინ $p = 0.80$, ე.ი. ასეთია წარმატების აპრიორული მათემატიკური ლოდინი (ჯვარი). მხოლოდ ანარჩევის გამოყენებისას კი $p = 0.90$, ე.ი. ასეთია შეფასება დამაჯერებლობის მაქსიმუმით (სამკუთხედი). ბაიესური შეფასება (წრე) არის შუალედური ამ ორ შეფასებას შორის.

აქ მიღებულია შემდეგი პირობითი აღნიშვნები : 1 - აპრიორული განაწილება ; 2 - აპოსტერიორული განაწილება ; 3 - აპრიორული მათემატიკური ლოდინი ; 4 – მაქსიმალური დამაჯერებლობის შეფასება (MLE, Maximum Likelihood Estimate) ; 5 - ბაიესის შეფასება.

შენიშვნა

განსახილველად ბეტა განაწილების არჩევა წინა მაგალითში შემთხვევითი არ ყოფილა. ბაიესურ დანასკვში აპრიორული კანონი და აპოსტერიორული კანონი, როგორც ამბობენ, *შეუღლებულია*, თუ ისინი ალბათობათა განაწილების ერთსა და იმავე ოჯახს მიეკუთვნება. კერძოდ, ბერნულის დამაჯერებლობისთვის ბეტა-კანონი შეუღლებულია თავის თავთან.

დამატებითი ინფორმაცია

- *ჰიპოთეზის ტესტები*, რომლებიც საშუალებას იძლევა ვუარყოთ ან დავადასტუროთ ალბათური კანონების, მათი პარამეტრების ან მახასიათებლების შესახებ გამოთქმული სავარაუდო წინადადებები, ასევე არის ინფერენციული სტატისტიკის ნაწილი. მაგალითად, აუცილებელია გადავწყვიტოთ, თუ დამაჯერებელია, რომ ცვლადის მათემატიკური ლოდინი გარკვეულ მნიშვნელობას აღემატება ; ან რომ ცვლადი ემორჩილება განაწილების ნორმალურ კანონს.

- შესაძლებელია ანარჩევის შედგენის სხვა მეთოდიც მარტივი შემთხვევითი ანარჩევის მეთოდის გარდა, ვთქვათ, *სტრატიფიცირებული* შემთხვევითი ანარჩევის მეთოდი, რომლის გამოყენებისას პოპულაცია იყოფა სტრატებად გარკვეული მახასიათებლის (მაგალითად, ასაკობრივი ჯგუფის) შესაბამისად, ხოლო ანარჩევის ფორმირება ხდება მარტივი შემთხვევითი არჩევით თითოეულ სტრატაში. ამრიგად, ყოველი სტრატისთვის მიიღება ანარჩევის ისეთი ზომა, რომელიც პროპორციულია პოპულაციაში სათანადო სტრატის ზომის. სხვა სიტყვებით რომ ვთქვათ, ყველა ინდივიდუუმს არ აქვს ერთნაირი ალბათობა იყოს არჩეული და ამოღებული პოპულაციიდან : ეს დამოკიდებულია იმ სტრატის ზომაზე, რომელსაც ისინი მიეკუთვნება.

- ბუნებრივია ასევე სურვილი, რომ შემფასებელი უახლოვდებოდეს ჭეშმარიტ მნიშვნელობას მით უკეთ, რაც მეტია შესაფასებელი სტრატის ზომა და, ამრიგად, ანარჩევის მოცულობაც დიდია. აქ ჩვენ ვსაუბრობთ $\left(\Theta_n\right)_{n\in\mathbb{N}^*}$ ნამდვილ შემთხვევით ცვლადთა მწკრივის კონვერგენციაზე, კრებადობაზე, სწრაფვაზე θ ნამდვილი მნიშვნელობისკენ ; ფაქტობრივად ლაპარაკია კრებადობაზე Θ შემთხვევით ცვლადთან, რომელიც θ სიდიდის ტოლია თითქმის ყველგან. ამბობენ, რომ θ სიდიდის Θ_n შემფასებელი *კრებადია*, თუ იგი *ალბათობით იკრიბება* $\theta-$თან. თუ ამასთან ერთად კრებადობა თითქმის უსაფრთხოა, მაშინ ლაპარაკობენ, რომ Θ_n არის θ სიდიდის *ძლიერ კრებადი* შემფასებელი.

- რა თქმა უნდა, შეიძლება აიგოს შემფასებელი *მომენტების მეთოდით*, რომელიც იმაში მდგომარეობს, რომ $\mathbb{P}_X$ თეორიული მომენტები (რომლებიც, ამრიგად, დამოკიდებულია $\theta-$ზე) ემთხვეოდეს ანარჩევის ემპირიულ მომენტებს. დიდ რიცხვთა კანონი მართლაც იძლევა შესაძლებლობას მიახლოებული გახდეს საშუალო მნიშვნელობა ემპირიული მნიშვნელობით. ეს მეთოდი, როგორც წესი, ნაკლებად ზუსტია, ვიდრე მიდგომა დამაჯერებლობის მაქსიმუმით.

- შელდონ როსის ნაშრომში (Sheldon M. Ross, 1987) დაწვრილებით აღწერილია შეფასება მაქსიმალური დამაჯერებლობის პრინციპით და ბაიესური შეფასება.

- იმისათვის, რომ გავერკვიოთ ალბათობათა ცნებებისა და წარმოდგენების საფუძვლებში და გავაღრმავოთ აღწერითი (დესკრიფციული) სტატისტიკის ჩვენი ცოდნა როგორც ინფერენციული სტატისტიკის, შეიძლება მივმართოთ ჟილბერ საპორტას (Gilbert Saporta, 2011) ან ჟან-პიერ ლეკუტრის (Jean-Pierre Lecoutre, 2019) ნაშრომებს და ასევე ომარ ფურთუხიას ამოცანათა კრებულს (ომარ ფურთუხია, 2020).

ბიბლიოგრაფია

1. Ross, S. M. (1987). *Introduction to Probability and Statistics for Engineers and Scientists.* Wiley, New York, 492 pages. https://minerva.it.manchester.ac.uk/~saralees/statbook3.pdf

2. Saporta, G. (2011). *Probabilités, analyse des données et statistique* – 3^{e} edition. Editions Technip , 622 pages

3. Lecoutre, J.-P. (2019). *Statistique et probabilités* – 7^{e} édition. Dunod Eco Sup, 314 pages.

4. ომარ ფურთუხია (2020). *ამოცანათა კრებული სტატისტიკური დასკვნების თეორიაში* (შემოკლებული ვარიანტი), თბილისი : ივანე ჯავახიშვილის სახელობის თბილისის სახელმწიფო უნივერსიტეტის გამომცემლობა, 214 გვერდი.

ლექცია 15 სწავლება განმტკიცებით

შინაარსი

1 განმტკიცებით სწავლების ცნება

1.1 ძირითადი განსაზღვრებები და ამოცანის დასმა

ამ ლექციაში ჩვენ განვიხილავთ მანქანური სწავლების კიდევ ერთ შტოს, კიდევ ერთ ტიპს — სწავლებას განმტკიცებით. მაგრამ დავიწყებთ, ბუნებრივია, მოტივირებით და განსხვავებებით წინა ლექციებში განხილული ყველა სხვა ტიპისგან.

ამრიგად, გავიხსენოთ, რასთან გვაქვს საქმე, როცა, ვთქვათ, გადასაწყვეტია სწავლების ამოცანა მასწავლებლით. ჩვენ გვაქვს პრედიქტორების (ანუ შემავალი ცვლადების) $X_1, X_2, \ldots, X_p$ კრებული და ჭდეთა (ანუ სწორი პასუხების) Y კრებული. ჩვენი ამოცანაა «შეძენილი გამოცდილების» (საწვრთნელ მონაცემებზე დამყარებული გამოცდილების) განზოგადება «მთელ სივრცეზე» (ნებისმიერ ტესტურ მონაცემებზე). მაგალითად, რეგრესიის ამოცანის გადაწყვეტისას $p=1$ შემთხვევაში, ჩვენ ხელთ გვაქვს მხოლოდ (უხეშად რომ ვთქვათ) n წყვილთა $(x_1, y_1), (x_1, y_2), \ldots, (x_n, y_n)$ მონაცემების კრებული. ჩვენი მიზანი კი არის $Y = f(X)$ სახის დამოკიდებულების აგება, რომელსაც შეუძლია იწინასწარმეტყველოს «სწორი პასუხი» უკვე ნებისმიერი X – თვის X_1 პრედიქტორის მნიშვნელობათა შესაძლო სიმრავლიდან.

მსგავსი რამ ხდება კლასიფიკაციის ამოცანაშიც: უბრალოდ გაიხსენეთ საწვრთნელი ობიექტების მხოლოდ სასრული რიცხვის კლასების ცოდნისას კლასიფიკაციის შედეგად ნიშნების მთელი სივრცე (ობიექტების უსასრულო რიცხვით) «ფერადდება» ამა თუ იმ კლასის

შესატყვისი სხვადასხვა ფერით. ეს სწორედ ისაა, რასაც ჩვენ ვუწოდეთ განზოგადება მთელ სივრცეზე. გასაგებია, რომ ანალოგიური მსჯელობები სამართლიანია უმასწავლებლოდ სწავლების განხილვისას.

მაგრამ რეალურ ცხოვრებაში სწავლება არც თუ ყოველთვის ხდება სქემით, რომელშიც ფაქტობრივად ცნობილია «სწორი პასუხები». ვთქვათ, მეცნიერთა მიერ ახალი სამკურნალო პრეპარატის შექმნა, ან მათემატიკოსების მიერ მორიგი ძალიან ეშმაკური ფორმულის მიღება, — ეს პროცესებია, რომლებსაც პასუხები არ აქვს, გარდა ამ პროცესების დასკვნით ეტაპებს: ეფექტურად მუშაობს სამკურნალო პრეპარატი, თუ არა; მიღებულ ფორმულას აქვს პრაქტიკული მნიშვნელობა, თუ არა. თავად შექმნის პროცესი კი (იქნება ეს ფორმულა თუ წამალი), ალბათ, აღიწერება დაახლოებით ასე: რაღაცას ვაკეთებთ, მოქმედებათა შედეგად რაღაც ხდება, ამ შედეგს როგორღაც ვაფასებთ, ვუფიქრდებით, ხოლო შემდეგ ვაგრძელებთ რაღაცის კეთებას და ვცდილობთ მივუახლოვდეთ სასურველ შედეგს. ცხადია, რომ თავისუფლად, შეუბოჭავად აღწერილ ასეთ ლოგიკას ყოველდღიურ ცხოვრებაში და ყოველ ნაბიჯზე ვაწყდებით უთვალავი პროცესის სახით:

1. მაგალითად, როგორ სწავლობს ბავშვი ფეხზე დგომას, როცა უკვე ნასწავლი აქვს ხოხვა (პროცესის აღწერა შეიძლება არც იყოს ძალიან ზუსტი)? ჯერ იგი მრავალჯერ ცდილობს მუხლებიდან ადგომას, თუმცა, გასაგები მიზეზების გამო მუდამ ეცემა. მაგრამ გარკვეული დროის შემდეგ იგი ხვდება (ცხადია, გამომდინარე გამოცდილებიდან), რომ, როცა იგი დედის ხელს ჩაეჭიდება, ფეხზე დადგომა გაცილებით უფრო ადვილი ხდება. ესე იგი, ადგომისას რაღაცას უნდა ჩაეჭიდო.
მაგიდას, სკამს და თუნდაც კედელს — ყველაფერი გამოდგება. მოგიწევს მხოლოდ ოდნავ აცოცება და, როგორც ფრანგები იტყვიან ხოლმე, voilà tout – ადამიანმა ფეხზე დადგომა ისწავლა!

2. ან როგორ მიმდინარეობს საფეხბურთო მატჩი? ბუნებრივია, თითოეულ გუნდს აქვს წინასწარ მოფიქრებული და საკუთარი გამოცდილებიდანაც გამომდინარე ტაქტიკა, რომელიც ითვალისწინებს საკუთარი თავისა და მოწინააღმდეგეთა როგორც ძლიერ, ასევე სუსტ მხარეებს. მაგრამ თამაშის მსვლელობა სავსეა მოულოდნელობებით და მისი წინასწარმეტყველება ძალიან ძნელია: რატომღაც სწორედ დღეს თავდამსხმელს შეუძლია საკმაოდ ადვილად მიუახლოვდეს მეტოქის კარს სწორედ მარჯვენა ფლანგიდან. ეტყობა მცველები სხვა რამეზე ფიქრობენ. ასე რომ, ეს ხელსაყრელია და აზრი აქვს შეტევების, ძირითადად, სწორედ ამ ფლანგიდან განხორციელებას.
რა თქმა უნდა, შესაძლებელია კიდევ მრავალი სხვა მაგალითის მოყვანა, მაგრამ იდეა, ალბათ, ცხადია. ამიტომ ახლა გადავიდეთ ამოცანის გარკვეულ ფორმალურ დასმაზე და შემოვიტანოთ განსაზღვრებები ყველგან მომდევნო მასალაში გამოსაყენებლად.

ამრიგად, უხეშად რომ ვთქვათ, სწავლება განმტკიცებით — ეს შემდეგ პარადიგმაზე დაფუძნებული მანქანური სწავლების მიმართულებაა: გასაწვრთნელ მოდელს არ აქვს დეტალური ცნობები მის გარშემო არსებულ სისტემაზე, მაგრამ შეუძლია აწარმოოს მასში (ე.ი. ამ სისტემაში) გარკვეული მოქმედებები და გააანალიზოს სისტემის რეაქცია ამ მოქმედებებზე. სხვა სიტყვებით რომ ვთქვათ, გააანალიზოს განმტკიცებები. ცხადია, როგორც მოდელი ახდენს ზემოქმედებას სისტემაზე, ასევე სისტემაც ახდენს ზემოქმედებას მოდელზე, ამიტომ წყვილი «სისტემა-მოდელი» ბადებს შემდეგ განსაზღვრებას, რომელიც საკმაოდ პირობით ხასიათს ატარებს.

განსაზღვრება 1.1.1 *მოდელს, რომლის სწავლება განმტკიცებით სწავლების ამოცანის ფარგლებში ხდება, ხშირად აგენტს უწოდებენ. სისტემას აგენტის გარშემო კი ხშირად გარემო ეწოდება.*

გადავიდეთ აგენტის გარემოსთან თანამოქმედების პროცესის ფორმალურ აღწერაზე.

1. დავუშვათ, რომ აგენტი და გარემო თანამოქმედებს $t \in \{0,1,2,\ ...,n\}, n \in \mathbb{N} \cup \{0,\infty\}$ დროის დისკრეტულ მომენტებში : თანამოქმედებათა რაოდენობა შეიძლება იყოს როგორც სასრული (თუ n სასრულია), ასევე უსასრულოც.
2. ყოველ ბიჯზე აგენტი იმყოფება გარკვეულ $s_t \in S$ მდგომარეობაში, სადაც S — ყველა შესაძლო მდგომარეობათა სიმრავლეა.
3. s_t მდგომარეობის საფუძველზე აგენტი ირჩევს გარკვეულ $a_t \in A(s_t)$ მდგომარეობას, სადაც $A(s_t)$ — აგენტისათვის s_t მდგომარეობაში მისაწვდომი მოქმედებების სიმრავლეა.
4. a_t მოქმედების განხორციელების შემდეგ გარემო გენერირებს r_{t+1} ჯილდოს და გადაჰყავს აგენტი s_{t+1} მდგომარეობაში.
5. ბიჯები 3-4 მეორდება, თუ არ არის მიღწეული გაჩერების რომელიმე კრიტერიუმი.

შემდგომ ჩვენ დაწვრილებით ავხსნით, რას ნიშნავს სიტყვები «გენერირებს», «გადაჰყავს» და აღვწერთ მთელ სქემას მათემატიკური კორექტულობით. ახლა კი, ვითარების ზოგადი გაგების განსამტკიცებლად, მოვიყვანთ გარემოსა და აგენტს შორის დიალოგის ძალიან პირობით მაგალითს:

1. **გარემო:** აგენტო, შენ იმყოფები s_0 მდგომარეობაში. შენთვის მისაწვდომია მოქმედებები $3,4,5\left(A(s_0)=\{3,4,5\}\right)$ ნომრებით.
აგენტი: ვახორციელებ მოქმედებას ნომრით $4\ (a_0=4)$.

2. **გარემო:** შენ გეძლევა 3 ერთეულის $(r_1=3)$ ტოლი ჯილდო და გადადიხარ s_1 მდგომარეობაში. ახლა შენთვის მისაწვდომია მოქმედებები $1,4,5\ (A(s_1)=\{1,4,5\})$ ნომრებით.
აგენტი: ვასრულებ მოქმედებას $5\ (a_1=5)$ ნომრით.

3. **გარემო:** შენ გეძლევა -1 ერთეულის ტოლი ჯილდო $(r_2=-1)$ და გადადიხარ s_2 в მდგომარეობაში. შენთვის მისაწვდომია … და ასე შემდეგ.

როგორც ვხედავთ, ჯილდოები (ანუ განმტკიცებები) შეიძლება იყოს როგორც დადებითი, ასევე უარყოფითიც. აგენტის ამოცანა — დაგროვებულ ჯილდოთა მაქსიმიზაცია, ეყრდნობა შემდეგ ვარაუდს (ეგრეთ წოდებულ ჰიპოთეზას ჯილდოს შესახებ — Reward Hypothesis): «ყველა მიზანი შეიძლება იყოს აღწერილი მოსალოდნელი დაგროვებული ანაზღაურების მაქსიმიზაციის გზით» (ინგლ. : «*All goals can be described by the maximisation of expected cumulative reward*»). ამ სიტყვებით გადმოცემული ფრაზის არსს ჩვენ ოდნავ მოგვიანებით დავუბრუნდებით.

ამრიგად, იმედი გვაქვს, რომ ურთიერთქმედების სქემა გამჭვირვალეა და სრულიად ნათელიც. მაგრამ რჩება კიდევ ერთი პრობლემა, რომლის გადაჭრას, საერთოდ და მთლიანობაში, მიეძღვნება ამ ლექციის მთელი მომდევნო მასალა : როგორ უნდა მოხდეს a_t მოქმედების არჩევა t ბიჯზე შესაძლო მოქმედებათა $A(s_t)$ სიმრავლიდან ? ასე ჩვენ ბუნებრივად

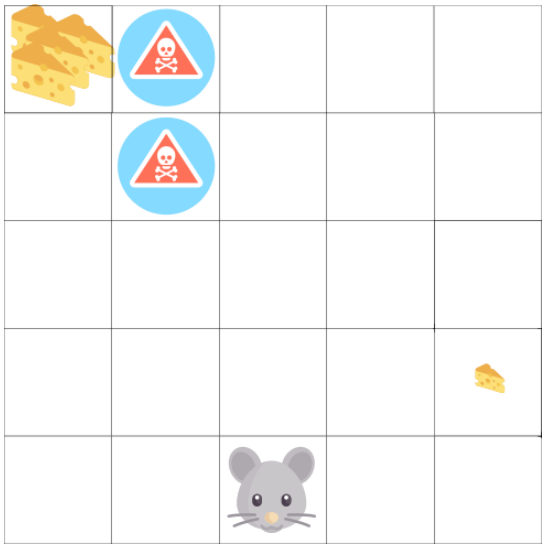
მივდივართ სტრატეგიის ცნებამდე.

განსაზღვრება 1.1.2 *ხელმისაწვდომი a მოქმედების დროის t მომენტში (იმ შემთხვევაში, როცა აგენტი $s_t \in S$ მდგომაროებაში იმყოფება) არჩევის $\pi_t(a \mid s_t)$ ალბათობათა სიმრავლეს აგენტის სტრატეგია ეწოდება.*

სხვა სიტყვებით რომ ვთქვათ, მაშინ, როცა s_t მდგომარეობაში დროის t მომენტისათვის მისაწვდომია ერთზე მეტი მოქმედება $A(s_t)$ მასივიდან, განსახორციელებელი მოქმედების არჩევა ხდება შემთხვევითად, ალბათურად : $\pi_t(a \mid s_t)$ სტრატეგიის შესაბამისად. როგორ უნდა იყოს მოცემული ისეთი სტრატეგია, რომელიც დაგროვებულ ანაზღაურებათა მაქსიმუმს უზრუნველყოფს ? სწორედ ეს არის ის «*მილიონად შეფასებული შეკითხვა*»[6], რომელზეც პასუხის გაცემას ჩვენ შევეცდებით.

1.2 კომპრომისი «შესწავლა-გამოყენება»

ვიდრე კონკრეტულ მაგალითებზე და ამოცანებზე გადავიდოდეთ, ყურადღება მივაქციოთ ეგრეთ წოდებულ პრობლემას (ანუ კონფლიქტს, კომპრომისს, დილემას) შესწავლასა (კვლევასა, დაზვერვასა) და გამოყენებას (ექსპლუატაციას) შორის, რომელიც ინგლისურად «exploration-exploitation» სახელწოდებით არის ცნობილი. გამოყენების პრინციპი დამყარებულია შემდეგ იდეაზე : ყოველ ბიჯზე სასურველია საუკეთესო (ან ყველაზე მომგებიანი, ხელსაყრელი) გადაწყვეტილების მიღება არსებული ინფორმაციის საფუძველზე. შესწავლის პრინციპი კი ცვლის კონცეფციას : შეგროვდეს რაც შეიძლება მეტი ინფორმაცია (შესაძლოა) უფრო წარმატებული გამოყენებისათვის შემდგომში. ხსენებული კონფლიქტის არსი საკმაოდ ადვილია გავიგოთ ნახატიდან 15.1.



[6] ეს არის თამაშის ტიპის შოუ სახელგანთქმულ პირთათვის. მოწვეულ ვარსკვლავებს შეუძლიათ ფულის მოგება პირად საიდუმლოებათა გამჟღავნებისათვის. სტუმარმა უნდა უპასუხოს წამყვანის ათ ძალიან არასასიამოვნო შეკითხვას. უკანასკნელი შეკითხვა — ყველაზე ძნელია და მართალი პასუხისათვის მოწვეულს 1 მილიონი ამერიკული დოლარი ეძლევა. მხოლოდ ყველაზე პატიოსანი და უშიშარი სტუმრებს შორის ბედავს პასუხის გაცემას ამ შეკითხვაზე და უმძიმესი სიმართლის მოყოლას, რომელსაც წლების მანძილზე იგი უმალავდა თავის ფანებს, თაყვანისმცემლებს.

დავუშვათ, რომ ერთი სვლის შედეგად თაგვი (ჩვენი აგენტი) ახორციელებს ერთ გადაადგილებას (მოქმედებას). თამაში ჩერდება, ან როცა თაგვი შეჭამს ყველს, ან როცა ხვდება ხაფანგში, ან როცა მას ამოეწურება სვლების მარაგი (ვთქვათ, მათი რაოდენობა სულ 10-ს შეადგენს). ტარდება ექსპერიმენტის (თამაშის) გარკვეული სასრული რაოდენობა, რის შედეგადაც მოპოვებული ყველის საერთო რაოდენობა გაანგარიშდება. ცხადია, რაც უფრო მეტი ყველია მოპოვებული ყველა თამაშის შედეგად — მით უკეთესია. გარდა ამისა, არ მიაღწიო ყველს (ყველა სვლის ამოწურვისას) და არ მოხვდე ხაფანგში (არ მოკლა აგენტი) — შედეგია, რომელიც ბევრად უკეთესია, ვიდრე ხაფანგში მოხვედრა (ესე იგი აგენტის მოკვლა).

ნახატიდან ცხადია, რომ თაგვისათვის ყველაფერზე უფრო ადვილი (და, ალბათ, სწრაფად განსახორციელებელი და ასევე უსაფრთხო) არის ყოველთვის მიაღწიოს ყველის მცირე ნაჭერს, რომელიც ახლოსაა განთავსებული, — რადგან იგი გარანტირებულად მიიღებს ყველს — მართალია, მცირე ნაჭერს, მაგრამ მაინც სასიამოვნო ჯილდოს — ყოველი თამაშის შედეგად. ასეთი მიდგომის პირობებში (რომელიც ხშირად სახელდება მიდგომად «დღევანდელი კვერცხი ხვალინდელ ქათამს უნდა გერჩიოს» ან «უკეთესია ტიტი ხელში, ვიდრე წერო ცაში») ყველის მნიშვნელოვანი მარაგები, რომლებიც ხაფანგების მახლობლობაშია, მთლიანად უფასურდება. აღწერილი ისტორია სხვა არაფერია, თუ არა «გამოყენების» პრინციპი.

მეორე მხრივ, თაგვს შეუძლია მსხვერპლიც გაიღოს, შეწიროს წამიერი შედეგი (და, შესაძლოა, თავისი სიცოცხლეც) იმისათვის, რომ დაზვეროს, შეისწავლოს გარემო და შეეცადოს იპოვოს გზა, თუნდაც უფრო რთული და სახიფათო, მაგრამ მიმავალი უფრო მნიშვნელოვანი ჯილდოსკენ, და, შედეგად გაზარდოს ყველის ნადავლი (ანუ ჯილდო) «საშუალოდ».

ამრიგად, თეზისების სახით კომპრომისი შესწავლასა და გამოყენებას შორის შეიძლება იყოს დახასიათებული შემდეგი პუნქტებით :

1. აუცილებელია საკმარისი ინფორმაციის მიღება გარემოზე, იმისათვის რომ გადაწყვეტილებები (და, მაშასადამე, შედეგები) «საშუალოდ» საუკეთესო იყოს.
2. უკეთესი გრძელვადიანი სტრატეგიისათვის დასაშვები (და გარდაუვალიც კი) არის მოკლევადიანი დანაკარგებიც.

აღწერილი პრობლემა შესწავლასა და გამოყენებას შორის — ცარიელ ადგილზე არ აღმოცენებულა უსაფუძვლო გამონაგონად. ეს პრობლემა რეგულარულად ჩნდება რეალურ ცხოვრებაშიც. მაგალითად, დავუშვათ, რომ თქვენ გადაწყვიტეთ ვახშმობა რესტორანში. *გამოყენება* — ეს შემოწმებული რესტორნის არჩევაა, რომელიც სტაბილურად იმსახურებს თქვენს ნდობასა და მოწონებას როგორც კერძების ხარისხით, ასევე მათი ფასებითაც. *შესწავლა* (*და ზვერვა*) — ეს რომელიღაც ახალი დაწესებულების არჩევა და მოსინჯვაა : იქნებ იქ უკეთესი აღმოჩნდეს ?

ამრიგად, ვიმედოვნებთ, რომ როგორც პრობლემა, ასევე განმტკიცებით (წახალისებით) სწავლების ამოცანის დასმა ცხადი და გამჭვირვალეა. ასე რომ ახლა ერთ-ერთი ყველაზე გასაგები და პოპულარული ამოცანის განხილვაზე გადავიდეთ — ამოცანაზე k – ხელიანი ბანდიტის შესახებ.

1.3 ამოცანის დასმა k – ხელიან ბანდიტზე

როგორც უკვე ვთქვით, ერთ-ერთ ყველაზე მარტივ მაგალითს, რომელიც განმტკიცებით

(წახალისებით) სწავლების იდეის ილუსტრირებას იძლევა, არის ე.წ. ამოცანა $k-$ხელიანი ბანდიტის შესახებ. ამოცანის სახელწოდება სათავეს იღებს ანალოგიიდან სათამაშო ავტომატთან (რომელიც კარგად არის ცნობილი როგორც «ცალხელა ბანდიტი»). ჩვენს შემთხვევაში ჩავთვალოთ, რომ საქმე გვაქვს ავტომატთან, რომელსაც k ბერკეტი (ხელი) აქვს.

ბუნებრივი იქნება, რომ დავიწყოთ გარკვევით : რა არის მოცემულ ამოცანაში აგენტი, რა არის სტრატეგია და ასე შემდეგ.

აგენტი — ეს ადამიანია, რომლისთვისაც დროის ნებისმიერ t მომენტში და ნებისმიერ $s \in S$ მდგომარეობაში მისაწვდომია k ნებისმიერი მოქმედება ერთსა და იმავე მოქმედებათა რიცხვიდან — k ბერკეტიდან ერთ-ერთის არჩევა.

რა თქმა უნდა, ყველა შემთხვევაში ეს არჩევანი ხორციელდება რომელიღაც თავისი π_t სტრატეგიის (პოლიტიკის) შესაბამისად. დროის t მომენტში a_t მოქმედების არჩევის შემდეგ აგენტი იღებს წინასწარ უცნობ r_{t+1} ჯილდოს, რომელიც ნაკარნახევია გარემოს მიერ — $k-$ხელიანი ავტომატით, და გადადის შემდეგ მდგომარეობაში.

შენიშვნა 1.3.1 *ცალკე აღსანიშნავია, რომ ვინაიდან აგენტის შესაძლებლობები (ანუ $A(s)$ შესაძლო მოქმედებათა სიმრავლე) ნებისმიერ $s \in S$ მდგომარეობაში ერთნაირია — k ბერკეტიდან ერთ-ერთის არჩევა, ამიტომ ჩნდება ეჭვი, რომ აგენტის s მდგომარეობებიც ყოველ ბიჯზე ერთნაირია (ანუ საერთოდ, მხოლოდ ერთადერთი მდგომარეობა არსებობს). ასეთი ვარაუდი, ზოგად შემთხვევაში, არასწორია, რადგან ყოველ $s_t \in S$ მდგომარეობაში დროის t მომენტში, აგენტი მოქმედებს გარკვეული, საზოგადოდ, სხვადასხვა π_t სტრატეგიის შესაბამისად.*

ახლა კი დავსვათ ბუნებრივი შეკითხვა : რა მოსაზრებებიდან გამომდინარე უნდა ხდებოდეს ავტომატის სახელურის ჩამოწევა? ჩვენ რომ წინასწარ ვიცოდეთ ყოველი მოქმედების ფასეულობა, მაშინ ლოგიკური ხარბი (მაგრამ არაშორსმჭვრეტელური, წინდაუხედავი) სტრატეგია იქნებოდა ასეთი :

დროის t მომენტში ავირჩიოთ ის მოქმედება (ის ბერკეტი), რომლის არჩევისათვის r_{t+1} ჯილდო მაქსიმალური იქნება. მაგრამ მომავალი ჩვენ არ ვიცით, ამიტომ ჩვენი მოქმედებებით მხოლოდ წარსულში მიღწეული შედეგების შეფასება შეგვიძლია : «ჩამოწიე» ეს სახელური — მიიღე ამდენი, მეორე სახელური — ამდენი, და ასე შემდეგ. ამიტომ გონივრულ ნაბიჯად გამოიყურება კიდევ ერთი სპეციალური განსაზღვრების შემოტანა.

განსაზღვრება 1.3.1 *დავუშვათ, რომ $q_0(a) \in \mathbb{R}$ არის a მოქმედების ღირებულების საწყისი შეფასება (რომელსაც მკვლევარი ადგენს). მაშინ a მოქმედების ღირებულების შეფასება დროის $t \geq 1$ მომენტში ეწოდება შემდეგ სიდიდეს :*

$$q_t(a) = \begin{cases} q_0(a), & a \text{ მოქმედება არც ერთხელ არჩეულა} \\ \dfrac{\sum\limits_{i=0}^{t-1} r_{t+1} \mathrm{I}(a_i = a)}{\sum\limits_{i=0}^{t-1} \mathrm{I}(a_i = a)}, & \text{წინააღმდეგ შემთხვევაში} \end{cases},$$

სადაც $\mathrm{I}(a_i = a)$ არის $a_i = a$ ხდომილობის (ხდომილების) ინდიკატორი (მაჩვენებელი) : იგი

უდრის ერთს, თუ i – ურ ბიჯზე არჩეულია a მოქმედება, და ნულს წინააღმდეგ შემთხვევაში.

შენიშვნა 1.3.2 *ჩაწერილი გამოსახულება (იმ შემთხვევაში, როცა a მოქმედება არჩეულია თუნდაც ერთხელ) — იმ ჯილდოთა საშუალო არითმეტიკულია, რომლებიც მიღებულია a მოქმედების არჩევისას დროის t მომენტამდე.*

განვიხილოთ მაგალითი.

მაგალითი 1.3.1 *დავუშვათ, რომ გარკვეული თამაშის პროცესში აგენტი ყოველ ნაბიჯზე ახორციელებს ერთ-ერთ მოქმედებას ორი შესაძლო — a და b — მოქმედებიდან გარკვეული სტრატეგიის შესაბამისად და ყოველ t ბიჯზე იღებს გარკვეულ r_{t+1} ჯილდოს. მონაცემები (ზემოთ შემოღებული აღნიშვნების გამოყენებით) წარმოდგენილია ცხრილში.*

ბიჯი	*მოქმედება*	*ჯილდო*
$t = 0$	$a_0 = b$	$r_1 = 3$
$t = 1$	$a_1 = a$	$r_2 = 0$
$t = 2$	$a_2 = a$	$r_3 = 8$
$t = 3$	$a_3 = b$	$r_4 = 1$
$t = 4$	$a_4 = b$	$r_5 = 5$

ვიპოვოთ a და b მოქმედებათა ღირებულებების შეფასებები მეხუთე ბიჯზე. როგორც ვხედავთ, a მოქმედება $2-$ჯერ იყო არჩეული, ამ დროს მიღებული იყო 0 და 8 ერთეულის ტოლი ჯილდოები. ასეთ პირობებში a მოქმედების ღირებულება მეხუთე ბიჯზე შეიძლება შემდეგი გზით იყოს მოპოვებული (როგორც ზემოთ ვთქვით — როგორც ადრე მიღებული ჯილდოების საშუალო არითმეტიკული მნიშვნელობა) :

$$q_5(a) = \frac{0+8}{2} = 4.$$

ცხადია, იგივე მიიღება «ორიგინალური» თანაფარდობის გამოყენებით საკუთრივ განმარტებიდან :

$$q_5(a) = \frac{\sum_{i=0}^{5-1} r_{i+1} \mathrm{I}(a_i = a)}{\sum_{i=0}^{5-1} \mathrm{I}(a_i = a)} = \frac{3\cdot 0 + 0\cdot 1 + 8\cdot 1 + 1\cdot 0 + 5\cdot 0}{0+1+1+0+0} = 4.$$

ანალოგიურად მოიპოვება b მოქმედების ღირებულების შეფასება მეხუთე ბიჯზე :

$$q_5(b) = \frac{3+1+5}{3} = 3.$$

შეიძლება შევნიშნოთ, რომ a მოქმედების ღირებულების შეფასება მეხუთე ბიჯზე, უფრო მაღალია, ვიდრე b მოქმედების ღირებულების შეფასება იმავე პირობისათვის, მაგრამ ყველაფერი შეიძლება შეიცვალოს, თუ მიღებულ ჯილდოებზე მეტი რაოდენობის ინფორმაციას დავაგროვებთ.

აღწერილი მიდგომის გამოყენება მოქმედებათა ღირებულების შესაფასებლად იწვევს გარკვეულ ზედმეტ დანახარჯებს გამოთვლებზე და ინფორმაციის შენახვაზე. მნელი არ არის, მაგალითად, იმის გაგება, რომ ღირებულების შესაფასებლად გვიხდება ინფორმაციის შენახვა წინათ მიღებული ყველა ანაზღაურების შესახებ. გონივრულია დავფიქრდეთ, ხომ არ არსებობს ნაკლებ დანახარჯებთან დაკავშირებული ხერხი ? თურმე, არსებობს. იგი შემდეგი ლემით განისაზღვრება.

ლემა 1.3.1 (ღირებულების შეფასების რეკურენტული ფორმულა) *დავუშვათ, რომ a მოქმედება $(t-1)$ რაოდენობის ბიჯზე არჩეულია ზუსტად $n-ჯერ$, სადაც $n \le (t-1)$ და $q_t(a)$ — ამ მოქმედების ღირებულების შეფასებაა t ბიჯზე. მაშინ, თუ იგი არჩეულია t ბიჯზე და მისთვის მიღებულია r_{t+1} ჯილდო, ადგილი აქვს შემდეგ თანაფარდობას :*

$$q_{t+1}(a) = q_t(a) + \frac{1}{n+1}(r_{t+1} - q_t(a)),$$

წინააღმდეგ შემთხვევაში $q_{t+1}(a) = q_t(a)$, ამასთან ერთად $q_0(a) \in \mathbb{R}$ — ღირებულების საწყისი შეფასებაა.

შენიშვნა 1.3.3 *ზემოთ შემოტანილი ფორმულის გამოყენება მხოლოდ ორი პარამეტრის შენახვის საშუალებას იძლევა : მოქმედების ღირებულების მიმდინარე მნიშვნელობის და ამ მოქმედების აქტივაციის რიცხვის.*

მტკიცებულება. შემთხვევა, როცა t ბიჯზე a მოქმედება არჩეული არ არის, აშკარაა — მოქმედების ღირებულების შეფასება არ იცვლება. განსახილველად მნიშვნელოვანია, პირველ რიგში, ისეთი შემთხვევა, როცა t ბიჯზე a მოქმედება მაინც არჩეულია. განსაზღვრების თანახმად,

$$q_t(a) = \frac{\sum_{i=0}^{t-1} r_{i+1} \mathrm{I}(a_i = a)}{\sum_{i=0}^{t-1} \mathrm{I}(a_i = a)} = \frac{\sum_{i=0}^{t-1} r_{i+1} \mathrm{I}(a_i = a)}{n},$$

სადაც უკანასკნელი ტოლობა სამართლიანია შემდეგი პირობის ძალით : a მოქმედება t ბიჯამდე არჩეული იყო n ჯერ. მაგრამ მაშინ $nq_t(a)$ — ეს იმ ანაზღაურებათა ჯამია, რომლებიც მიღებულია a მოქმედების არჩევისათვის $(t-1)$ ბიჯამდე ჩათვლით. რადგან მოქმედება არჩეულია t ბიჯზეც, ამიტომ, ვინაიდან ამისათვის მიღებულია r_{t+1} ჯილდო (და ახლა იგი არჩეულია $(n+1)$-ჯერ), მისი ღირებულების შეფასებისათვის გვაქვს :

$$\left.\begin{aligned} q_{t+1}(a) &= \frac{1}{n+1}\left(r_{t+1} + nq_t(a)\right) = \frac{1}{n+1}\left(r_{t+1} + (n+1)q_t(a) - q_t(a)\right) \\ &= q_t(a) + \frac{1}{n+1}\left(r_{t+1} - q_t(a)\right) \end{aligned}\right\}.$$

□

შენიშვნა 1.3.4 *მიღებულ ჯილდოთა საშუალო არითმეტიკულის გარდა (კერძოდ იმისათვის, რომ დინამიკაში ხდებოდეს სტრატეგიის ცვლილება), ხშირად იყენებენ $q_{t+1}(a)$ სიდიდის*

მიღების შემდეგ წესსაც :

$$q_{t+1}(a) = q_t(a) \ + \alpha_t(r_{t+1} - q_t(a)) \, .$$

კერძოდ, თუ $\alpha_t = \alpha$ — გარკვეული კონსტანტაა, მაშინ ჩვენ ვიღებთ ექსპონენციალურ მოსრიალე საშუალოს :

$$q_{t+1}(a) \ = \alpha r_{t+1} + \ (1-\alpha) q_t(a) \, .$$

ეს თანაფარდობა გვიჩვენებს, რომ, თუ $\alpha \in (0,1)$, მაშინ განსაკუთრებით ფასეული ხდება წამიერი ჯილდოები (ძალაში შედის ერთხელ უკვე ნახსენები პრინციპი «დღევანდელი კვერცხი ხვალინდელ ქათამს უნდა გერჩიოს» ან «უკეთესია ტიტი ხელში, ვიდრე წერო ცაში», რომლებიც იმ აზრის მატარებელია, რომ პატარა, მაგრამ საიმედო საქმე უკეთესია, ვიდრე მომხიბვლელი, თუმცა შორეული პერსპექტივა), ხოლო თუ $\alpha > 1$ — პირიქით. ახლა ამაზე ჩვენ დაწვრილებით არ შევჩერდებით.

1.4 ხარბი სტრატეგია

ამრიგად, გადავდივართ t ბიჯზე მოქმედებათა არჩევის სხვადასხვა სტრატეგიის განხილვაზე. ალბათ, ჯილდოს მაქსიმიზაციის პირველი და ყველაზე გულუბრყვილო ხერხი — ეს t ბიჯზე ისეთი მოქმედების (ან ისეთ მოქმედებათა) არჩევაა, რომლის (ან რომელთა) ღირებულების შეფასება მაქსიმალურია. დავუშვათ, რომ A_t — იმ მოქმედებათა სიმრავლეა s_t მდგომარეობაში, რომლებსაც აქვს ღირებულების მაქსიმალური (და, აქედან გამომდინარე, ერთნაირი) შეფასება. მათემატიკის ენაზე რომ ვილაპარაკოთ, A_t სიმრავლე — ისეთი სიმრავლეა, რომელიც შემდეგ პირობას აკმაყოფილებს :

$$A_t = \underset{a \in A(s_t)}{\text{Argmax}} \, q_t(a) \, .$$

ვინაიდან მომდევნო არჩევანს ვახორციელებთ მხოლოდ მოქმედების ღირებულების შეფასების მნიშვნელობიდან გამომდინარე, ხოლო ყველა მოქმედება A_t სიმრავლიდან t ბიჯზე ერთნაირად ღირებულია (ხოლო სხვები საერთოდ არ განიხილება), ამიტომ ლოგიკურია, რომ ნებისმიერი $a \in A_t$ მოქმედების არჩევა ხდებოდეს თანაბარი ალბათობით, რის გამოც :

$$\pi_t(a \mid s_t) = \begin{cases} \dfrac{1}{|A_t|}, & a \in A_t \\ 0, & a \notin A_t \end{cases} ,$$

სადაც $|A_t|$ — ელემენტების რაოდენობაა A_t სიმრავლეში. მაშასადამე, ღირებულების მაქსიმალური შეფასების მქონე მოქმედებათა არჩევა $\dfrac{1}{|A_t|}$ ალბათობით ხდება, დანარჩენი მოქმედება კი იგნორირებული რჩება.

განსაზღვრება 1.4.1 *სტრატეგიას, რომელიც ზემოთ იყო შემოტანილი, ხარბი სტრატეგია ეწოდება.*

შეიძლება ხარბი სტრატეგიის სრულიად საწინააღმდეგოც განვიხილოთ — ეგრეთ წოდებული აბსოლუტურად კვლევითი სტრატეგია.

განსაზღვრება 1.4.2 *თუ სწავლების შედეგად ყოველ t ბიჯზე ნებისმიერი მოქმედება $A(s_t)$ სიმრავლიდან — რომელიც s_t მდგომარეობაში მისაწვდომ ყველა მოქმედებას შეიცავს — აირჩევა ერთნაირი ალბათობით, მაშინ ასეთ სტრატეგიას აბსოლუტურად კვლევითი ეწოდება.*

«შესწავლა-გამოყენება» კომპრომისის გადმოსახედიდან ხარბი სტრატეგია — ეს წმინდა წყლის გამოყენებაა, შესწავლის ნებისმიერი შესაძლებლობის გარეშე. აბსოლუტურად კვლევითი სტრატეგია კი პასუხისმგებლობას იღებს უფრო მუდმივ შესწავლაზე. განვიხილოთ ხარბი სტრატეგიის გამოყენება შემდეგ მაგალითზე.

მაგალითი 1.4.1 *პატარა მაღაზიის მფლობელი იხილავს შემოსავლის გაზრდის შესაძლებლობებს. სადღაც მან ამოიკითხა ინფორმაცია იმის შესახებ, რომ საქონლის განლაგებამ მაღაზიაში შეიძლება გამოიწვიოს მყიდველის იმპულსური სტიმულირება გააკეთოს საყიდლები. მაგალითად, ფუნთუშა რომ ჩაის გვერდით განათავსო, მაშინ ფუნთუშას ყიდვის ალბათობა იზრდება, რაც ზრდის შემოსავალს. მფლობელმა გადაწყვიტა სამი სხვადასხვა კონფიგურაციის შემოწმება და მაღაზიის შემოსავლის გაკონტროლება სამივე შემთხვევაში.*

ამრიგად, განვსაზღვროთ ამოცანის ყველა ელემენტი. დავუშვათ, რომ «ბერკეტი» — ეს ერთ-ერთია სამი არჩეული (a,b,c) კონფიგურაციიდან და t — იმ დღის ნომერი, როცა ტარდება ექსპერიმენტი, ჯილდო — მყიდველის მიერ დღის განმავლობაში დახარჯულ თანხათა ჯამი (მაღაზიის შემოსავალი). დავუშვათ, რომ ყოველი თამაში შეიცავს 1000 იტერაციას (ბიჯს) (ფაქტობრივად — ეს იმ დღეთა რაოდენობით გამოსახული ექსპერიმენტის ხანგრძლივობაა). ჯილდოები a, b ან c კონფიგურაციათა არჩევისას — ეს ანარჩევებია ξ_1, ξ_2, ξ_3 გენერალური ერთობლიობებიდან, რომლებსაც ნორმალური $N_{2,1}$, $N_{2.5,1}$ და $N_{3,1}$ განაწილებები აქვს, შესაბამისად. ფორმალურად ეს იმას ნიშნავს, რომ მოცემულ მაგალითში ჯილდოები შეიძლება უარყოფითიც იყოს და ისინი შეიძლება იყოს ინტერპრეტირებული ან როგორც შეცდომები მონაცემებში, ან როგორც მაღაზიის დანაკარგები : დაკარგული, ვადაგადაცილებული, გატეხილი საქონელი და ასე შემდეგ.

ახლა კი, იმის შესახებ, როგორ მოქმედებს აგენტი ხარბი სტრატეგიის გამოყენებისას ? დავუშვათ, რომ $q_0(a) = q_0(b) = q_0(c) = 0$. რადგან დროის $t = 0$ მომენტში ყველა მოქმედების (კონფიგურაციის) ღირებულების შეფასებები ერთნაირია, ამიტომ არჩევანი მათ შორის კეთდება შემთხვევითად და, მაშასადამე, ნებისმიერო სამი კონფიგურაციიდან არჩეული იქნება ერთნაირი ალბათობით. შედეგად, სტრატეგია ნულოვან ბიჯზე ასეთია :

$$\pi_0(a \mid s_0) = \pi_0(b \mid s_0) = \pi_0(c \mid s_0) = \frac{1}{3}.$$

ვთქვათ, რომ არჩეულია ღირებულების c კონფიგურაცია c (ესე იგი $a_0 = c$). მაშინ ამ კონფიგურაციის და მხოლოდ ამ კონფიგურაციის ღირებულების შეფასება დადებითი გახდება (ნავარაუდევია, რომ მაღაზიის შემოსავალი ექსპერიმენტის პირველ დღეს მაინც დადებითია) და, მაშასადამე, ალგორითმის შესაბამისად, პირველ იტერაციაზე კვლავ იქნება არჩეული c

კონფიგურაცია (ესე იგი $a_1 = c$) და ასე შემდეგ, ასე რომ სტრატეგია ასეთია :

$$\pi_t(c \mid st) = 1, \quad t \geq 1 .$$

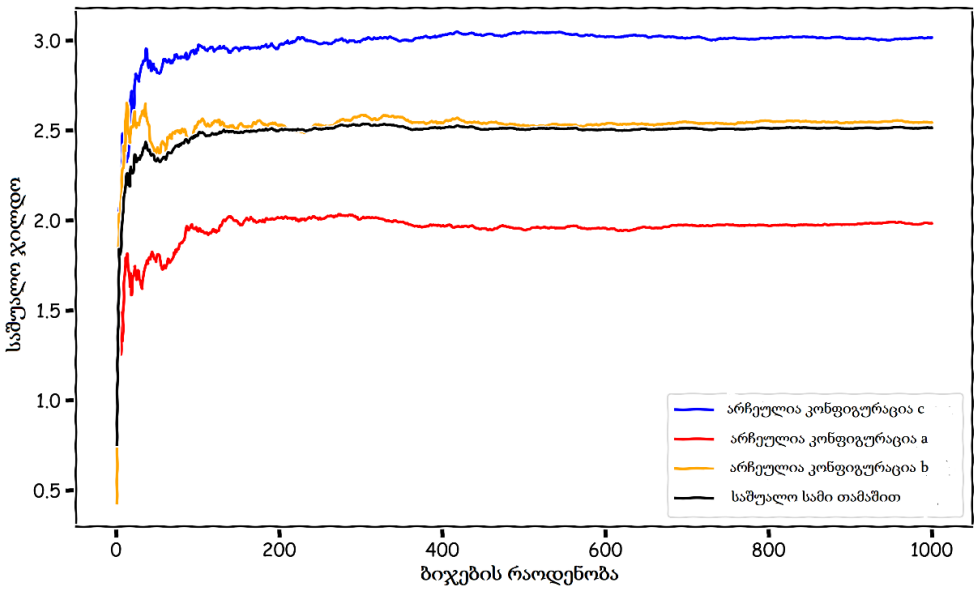
შედეგად, ჩვენს მაგალითში (ხარბი სტრატეგიის შესაბამისად) ყოველთვის მოხდება იმ კონფიგურაციის არჩევა, რომელიც იყო არჩეული (შემთხვევითად) ნულოვან ბიჯზე.

ნახატზე 15.2 წარმოდგენილია სამი თამაშის შედეგი. ლურჯი მრუდი შეესაბამება სიტუაციას, როცა პირველ თამაშში თავდაპირველად არჩეული იყო $a_0 = c$ მოქმედება და, როგორც ახლახანს აღვნიშნეთ, მისივე არჩევა ხდებოდა ყველა მომდევნო ბიჯზე ($a_t = c, t \in \{0,1,...,999\}$) საკუთრივ გრაფიკი გვიჩვენებს ჯილდოს საშუალო მნიშვნელობას ანუ, რაც ჩვენს შემთხვევაში ერთი და იგივეა, c მოქმედების ღირებულების შეფასებას ყოველ ბიჯზე.

მეორე თამაშში პირველ ბიჯზე a კონფიგურაციის არჩევა მოხდა, ესე იგი $a_0 = a$ (წითელი მრუდი). გრაფიკი კვლავ უჩვენებს საშუალო ჯილდოს სიდიდეს ასეთი საწყისი კონფიგურაციის არჩევისას (და კვლავ, $a_t = a, t \in \{0,1,...,999\}$).

მესამე თამაშში თავდაპირველად (ისევე, როგორც მომდევნო შემთხვევებში) არჩეული იყო b მოქმედება. საშუალო ჯილდოს მნიშვნელობები ამ შემთხვევაში გადმოცემულია ნარინჯის ფერით. როგორც ვხედავთ, ჯილდოს მაქსიმალურად შესაძლო საშუალო მნიშვნელობა მიღწეულია მხოლოდ ერთ შემთხვევაში სამიდან. გარდა ამისა, ერთ შემთხვევაში სამიდან მიღწეულია ჯილდოს მინიმალური საშუალო მნიშვნელობა.

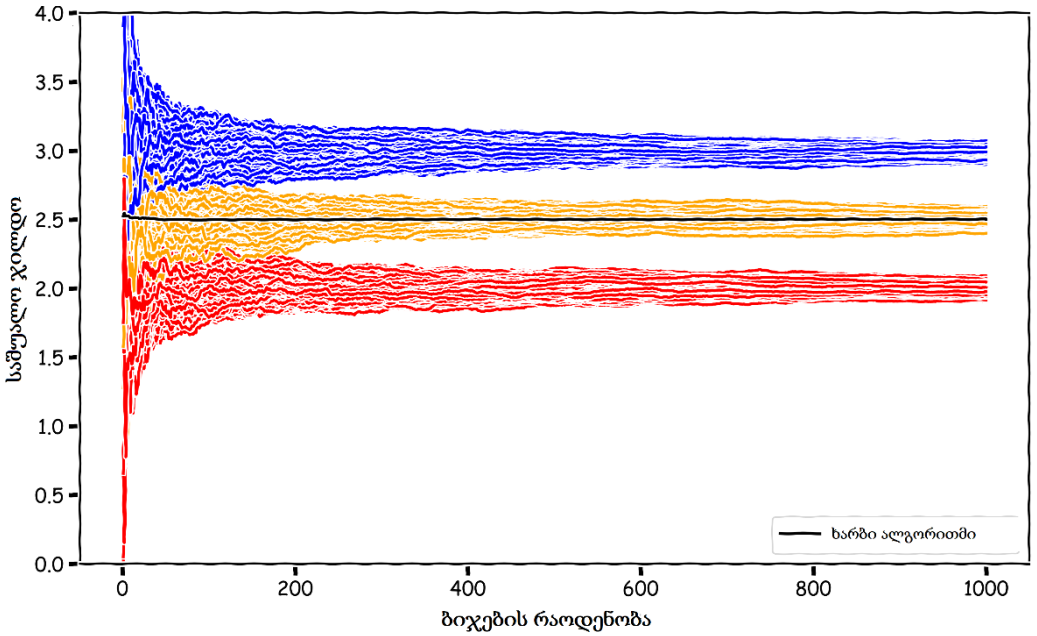
ამ მაგალითში აღწერილი «თამაშები», პრაქტიკაში შეიძლება იყოს რეალიზებული შემდეგნაირად : არის გარკვეული ქსელის სამი მაღაზია, რომლებიც ერთნაირ პირობებში იმყოფება. თითოეული მათგანი დამოუკიდებლად 1000 დღის განმავლობაში იყენებს ერთადერთ კონფიგურაციას და აგასებს საშუალო შემოსავალს. შემდეგ ხდება შეფასებათა შედარება და გამოიტანება დასკვნა იმის თაობაზე, რომელი კონფიგურაცია მუშაობს უკეთესად (თუ ასეთი კონფიგურაცია არსებობს).



ნახატი 15.2 - საშუალო შემოსავალი როგორც არჩეული კონფიგურაციის ფუნქცია.

გასაგებია, რომ საუკეთესო საწყისი კონფიგურაციის არჩევა ჩვენს მაგალითში — შემთხვევითობის საკითხია. სწორედ ამიტომ, თავად სტრატეგიის შეფასების თვალსაზრისით, ჩვენთვის საინტერესოა, როგორია ალგორითმის ქცევა «საშუალოდ». ამისათვის ატარებენ თამაშების სერიას. ნახატზე 15.2 შავი ფერით ილუსტრირებულია საშუალო ჯილდოს საშუალო მნიშვნელობა სამი თამაშის მიხედვით თითოეულ ბიჯზე. ნახატზე 15.3 წარმოდგენილია ხარბი სტრატეგიის გამოყენების შედეგები 1000 თამაშის ჩატარებისას 1000 იტერაციით თითოეულში. შავი წირით აღნიშნულია საშუალო ჯილდოს საშუალო მნიშვნელობა ყველა თამაშის მიხედვით. რაც შეეხება მაგალითს მაღაზიის კონფიგურაციების შესახებ, აღწერილი მიდგომა შეიძლება იყოს რეალიზებული შემდეგი ხერხით : არის გარკვეული ქსელის 1000 მაღაზია, რომლებიც ერთნაირ პირობებში იმყოფება. თითოეულ მაღაზიაში ხდება სამი კონფიგურაციიდან ერთ-ერთის შეფასება 1000 დღის განმავლობაში. მაშინ ხარბი სტრატეგიის გამოყენებისას, «საშუალოდ» თითოეული მაღაზია მოახდენს ყოველდღიური შემოსავლის გენერირებას 2.5 ერთეულის დონეზე.

შენიშვნა 1.4.1 *შემდგომ, სტრატეგიათა შესადარებლად ჩვენ ასევე ჩავატარებთ* 1000 *თამაშს* 1000 *იტერაციით თითოეულში, ხოლო შედეგებს გავასაშუალოებთ. ვინაიდან მიღებული შედეგები არ იქნება სტრატეგიის უშუალო გამოყენების შედეგები, არამედ იქნება სტრატეგიის ათასჯერადი გამოყენების გასაშუალოება, ამიტომ ვიხმართ ტერმინს ალგორითმი (მაგალითად, ხარბი ალგორითმი).*



ნახატი 15.3 - ხარბი ალგორითმის ილუსტრაცია

რა თქმა უნდა, ხარბი ალგორითმის არსებით ნაკლოვანებას განხილულ მაგალითში წარმოადგენს ის, რომ გარემოს კვლევა წყდება პირველივე ბიჯის შემდეგ, რაც, ცხადია, სერიოზულ გავლენას ახდენს შედეგებზე. როგორც უკვე აღვნიშნავდით, ნახატზე 15.2 ჩანს, რომ «ხარბი» პრინციპი განხილულ მაგალითში მაღაზიებზე მხოლოდ ერთ შემთხვევაში სამიდან გვაძლევს მაქსიმალურად შესაძლო საშუალო შემოსავალს. უფრო მეტიც, ერთ

შემთხვევაში სამი შემთხვევიდან იგი გვაძლევს მინიმალურ შესაძლო საშუალო შემოსავალს. ამის მიზეზია — ყოველნაირი შესწავლის, კვლევის, დაზვერვის აკრძალვა. ამ მდგომარეობის გამოსწორება ხელს უწყობს $\varepsilon-$ხარბი სტრატეგია.

1.5 $\varepsilon-$ხარბი სტრატეგია

გასაგები ხდება, რომ მეტ-ნაკლებად დამაკმაყოფილებელი შედეგის მისაღებად ურიგო არ იქნება აბსოლუტურად კვლევითი და ხარბი სტრატეგიების რაღაც ნარევის გამოყენება. ხომ ნათელია, რაც უფრო უკეთ ესმის აგენტს, როგორ არის მოწყობილი გარემო, მით უფრო მომგებიანი მოქმედებების განხორციელება შეუძლია მას. იმავდროულად, კვლევა კვლევისათვის — მოკლებულია ინტერესს, ასე რომ «სიხარბის» მთლიანად გამორიცხვა, ცხადია, შეუძლებელია, რადგან პრინციპი «როცა გაძლევენ, აიღე», ჯერ არავის გაუუქმებია.

განსაზღვრება 1.5.1 *დავუშვათ, რომ* $0<\varepsilon<1$. *მაშინ* $\pi_t(a\mid s_t)$ *სტრატეგიას, რომელიც დროის* t *მომენტში* $(1-\varepsilon)$ *ალბათობით მოქმედებს როგორც ხარბი, ხოლო* ε *ალბათობით — რო-გორც აბსოლუტურად კვლევითი,* $\varepsilon-$ *ხარბი ეწოდება.*

ამრიგად, შემოტანილი სტრატეგია — ეს მცდელობაა უზრუნველვყოთ კომპრომისი გამოყენებასა და შესწავლას შორის. სასარგებლოა იმის აღნიშვნაც, რომ $\varepsilon=0$ ან $\varepsilon=1$ პირობათა შესრულებისას (თუმცა ეს აკრძალულია განსაზღვრებაში), $\varepsilon-$ხარბი სტრატეგია ხდება ხარბი და აბსოლუტურად კვლევითი, შესაბამისად.

პრაქტიკისათვის სასარგებლოა $\varepsilon-$ხარბი სტრატეგიის მოცემა შემდეგი სახით :

$$\pi_t\left(a\mid s_t\right)=\begin{cases}\dfrac{1-\varepsilon}{\left|A_t\right|}+\dfrac{\varepsilon}{\left|A_t\left(s_t\right)\right|}, & a\in A_t\\ \dfrac{\varepsilon}{\left|A_t\left(s_t\right)\right|}, & a\notin A_t\end{cases},$$

სადაც

$$A_t=\operatorname*{Arg\,max}_{a\in A(S_t)} q_t\left(a\right).$$

შენიშვნა 1.5.1 *გავარკვიოთ, რატომ არის მოცემული ფორმულები შეთანხმებული ზემოთ ჩამოყალიბებულ განსაზღვრებასთან. ამისათვის ჯერ პასუხი გავცეთ შეკითხვას : როგორია* $a\in A_t$ *მოქმედების არჩევის ალბათობა ? ამ მოქმედების არჩევა* $\dfrac{1}{\left|A_t\right|}$ *ალბათობით ხდება, თუ გამოიყენება ხარბი სტრატეგია, ხოლო* $\dfrac{1}{\left|A(s_t)\right|}$ *ალბათობით, თუ გამოიყენება აბსოლუტურად კვლევითი სტრატეგია. იმის გამო, რომ ხარბი სტრატეგიის არჩევა* $(1-\varepsilon)$ *ალბათობით ხდება, ხოლო აბსოლუტურად კვლევითის —* ε *ალბათობით, ვიღებთ* $a\in A_t$ *მოქმედების არჩევის შემდეგ ალბათობას :*

$$\frac{1-\varepsilon}{|A_t|}+\frac{\varepsilon}{|A(s_t)|}.$$

მსგავსად ამისა, $a \notin A_t$ მოქმედება შეიძლება იყოს არჩეული მხოლოდ იმ შემთხვევაში, როცა არჩეულია აბსოლუტურად კვლევითი სტრატეგია, თანაც ამ შემთხვევაში მისი არჩევა $\frac{1}{|A(s_t)|}$ ალბათობით ხდება. ვინაიდან აბსოლუტურად კვლევითი სტრატეგიის არჩევა ε ალბათობით ხდება, ამიტომ $a \notin A_t$ მოქმედების არჩევის ალბათობა უდრის შემდეგ სიდიდეს :

$$\frac{\varepsilon}{|A(s_t)|}.$$

ვუჩვენოთ, რომ შემოტანილი განსაზღვრება კორექტულია, ესე იგი ვუჩვენოთ, რომ შემოტანილი «სტრატეგია» არის სტრატეგია ადრე შემოღებული განსაზღვრების აზრით.

ლემა 1.5.1 *$\varepsilon-$ ხარბი სტრატეგია წარმოადგენს სტრატეგიას.*

მტკიცებულება. საკმარისია დავამტკიცოთ, რომ

$$\sum_{a\in A(s_t)} \pi_t(a \mid s_t)=1.$$

ეს კი მართლაც ასეა, რადგან :

$$\sum_{a\in A(s_t)} \pi_t(a \mid s_t)=|A_t|\left(\frac{1-\varepsilon}{|A_t|}+\frac{\varepsilon}{|A(s_t)|}\right)+\left(|A(s_t)|-|A_t|\right)\left(\frac{\varepsilon}{|A(s_t)|}\right)=1.$$

□

ხარბი და $\varepsilon-$ხარბი ალგორითმების შედარება მაგალითისათვის მაღაზიათა სხვადასხვა კონფიგურაციით (1000 თამაში 1000 იტერაციით თითოეულში) წარმოდგენილია ნახატზე 15.4.

უნდა შევნიშნოთ, რომ მეტი ინფორმაციულობისათვის ეს და მომდევნო გრაფიკები აგებული იქნება სათავით $t=1$ ბიჯიდან.

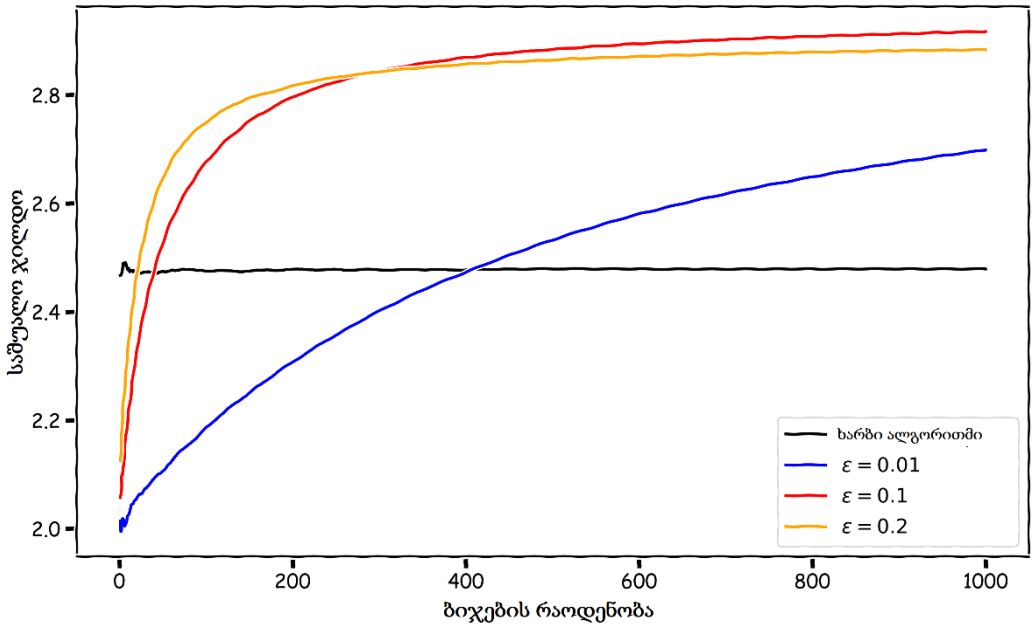
ისიც უნდა აღინიშნოს, რომ $\varepsilon-$ის სხვადასხვა მნიშვნელობის გამოყენებისას, შედეგები რამდენადმე განსხვავებულია.

ჩანს, მაგალითად, რომ $\varepsilon=0.2$ მნიშვნელობის პირობებში (ესე იგი იმ შემთხვევაში, როცა მეტი «თავისუფლება» ეძლევა აბსოლუტურად კვლევით სტრატეგიას), საშუალო ჯილდოს მაღალი მნიშვნელობები გაცილებით უფრო სწრაფად მიიღწევა, ვიდრე $\varepsilon=0.1$ და $\varepsilon=0.01$ შემთხვევაში.

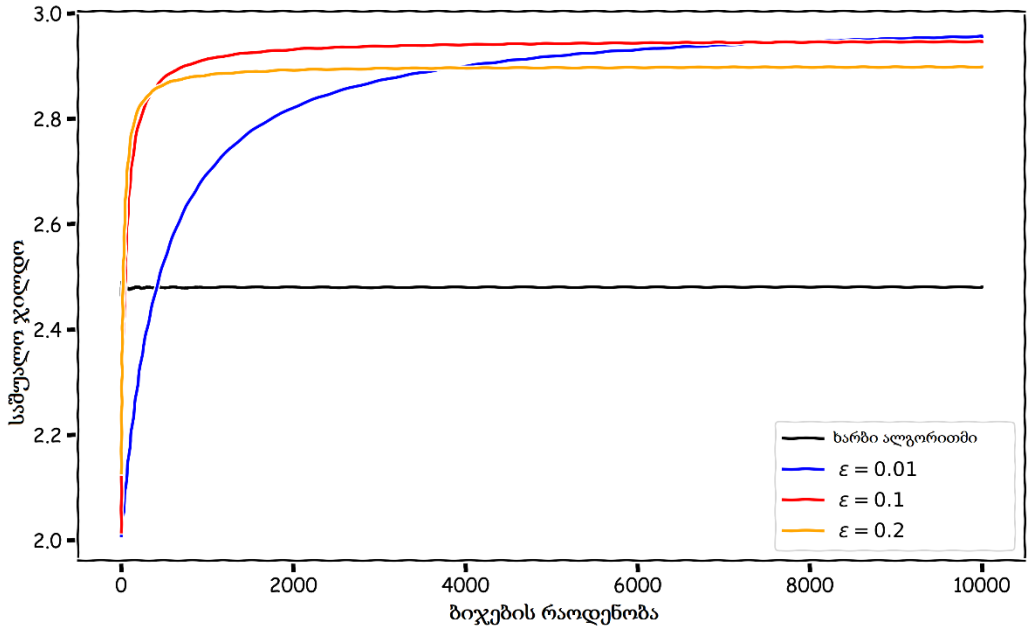
მაგრამ, როცა მოქმედებათა ღირებულებები საკმარისად შესწავლილია (დაწყებული მესამასე ბიჯიდან, დაახლოებით), ალგორითმი $\varepsilon=0.2$ მნიშვნელობაზე იწყებს წაგებას $\varepsilon=0.1$ შემთხვევასთან იმის გამო, რომ მისი კვლევითი მოქმედებების წილი უფრო მაღალია და უკვე საკმარისი სიზუსტით შეფასებული ღირებულების ჭეშმარიტად ოპტიმალური მოქმედების

არჩევა უფრო იშვიათად ხდება.

$\varepsilon = 0.01$ შემთხვევას გაცილებით მეტი დრო ესაჭიროება ყველა მოქმედების ღირებულებათა შესაფასებლად. ამ ალგორითმის პოტენციალი მჟღავნდება არა 1000, არამედ 10000 ბიჯის გამოყენებისას (ნახატი 15.5).



ნახატი 15.4 - ε — ხარბი ალგორითმების შედარება.



ნახატი 15.5 - ε — ხარბი ალგორითმების შედარება.

თუ დავაკვირდებით, შეიძლება შევნიშნოთ, რომ საბოლოო ჯამში $\varepsilon = 0.01$ შემთხვევისათვის საშუალო ჯილდოს მნიშვნელობა კონკურენტთა მაჩვენებლებზე მაღალი ხდება, დაწყებული, დაახლოებით, $t = 8000$ ბიჯიდან.

შენიშვნა 1.5.2 *მნიშნელოვანია აღინიშნოს, რომ თითოეული განხილული* ε *– ხარბი*

ალგორითმებიდან მუშაობს გაცილებით უკეთ, ვიდრე ხარბი ალგორითმი. უფრო მეტიც, საშუალო ჯილდოს მნიშვნელობები ყოველ განხილულ შემთხვევაში საკმაოდ სწრაფად იკრიბება საშუალო მაქსიმუმის ახლოს მდებარე მნიშვნელობასთან — სამთან. ზოგიერთი გრაფიკის ნახტომისებრი ქცევა ბიჯების მცირე რაოდენობისას განპირობებულია უზუსტო ცნობებით მოქმედებათა ღირებულების შეფასებაზე (ცნობათა დაგროვება ჯერ კიდევ ვერ მოასწრეს). მათი დაზუსტებისათვის მოქმედებათა არჩევა დიდწილად ქაოსურად ხდება და სწორედ ეს იწვევს საშუალო ჯილდოს ნახტომებს.

შენიშვნა 1.5.3 *დაბოლოს, გასაკეთებელია შემდეგი ევრისტიკული შენიშვნა : დროთა განმავლობაში (ესე იგი ბიჯთა t რაოდენობის ზრდისას) აზრი აქვს ε პარამეტრის შემცირებას, რადგან მოქმედებათა ღირებულებების შეფასებები უფრო ზუსტი ხდება და აზრი აქვს მეტი ყურადღება დაეთმოს გამოყენებას და არა შესწავლას. მაგალითად, ε სიდიდის t დროზე დამოკიდებულება შეიძლება იყოს მოცემული შემდეგი ფუნქციის სახით :*

$$\varepsilon(t) = \frac{1}{1 + t \cdot \beta},$$

სადაც t – დროის მიმდინარე მომენტია, ხოლო $\beta \in (0,1)$ — გარკვეული კოეფიციენტია, რომელსაც ევალება ε სიდიდის შემცირების სისწრაფის რეგულირება. β კოეფიციენტის როლი შეიძლება დავაკისროთ, მაგალითად, $\frac{1}{k}$ წილადს, სადაც k — განდიტის ხელების რაოდენობაა. ფუნქციის არჩევის ექსკლუზიური უფლება, რა თქმა უნდა, მკვლევარს ენიჭება.

1.6 Softmax-სტრატეგია

$\varepsilon-$ხარბი სტრატეგიების გამოყენებას განსაზღვრული ნაკლოვანება აქვს. იმ შემთხვევაში, როცა «შესწავლის» პროცესი მიმდინარეობს, მოქმედებათა არჩევა ერთნაირი ალბათობით ხდება, რაც ნამდვილად არ ახდენს დადებით გავლენას საბოლოო მიზანზე — წახალისებათა უდიდეს ჯამზე (რადგან მოქმედებებს ღირებულებათა მაღალი და დაბალი შეფასებებით ერთნაირი შანსი აქვს იყოს არჩეული). პრინციპი, რომლითაც ხელმძღვანელობს softmax-სტრატეგია, შემდეგნაირად შეიძლება იყოს ჩამოყალიბებული : დანაკარგების შემცირება შესწავლისას t იტერაციაზე უფრო იშვიათი არჩევით იმ a მოქმედებების, რომლებსაც აქვს ღირებულების დაბალი $q_t(a)$ შეფასება. ამის მისაღწევად, ყოველი მოქმედებისათვის გამოითვლება საკუთარი «წონითი კოეფიციენტი» და სწორედ მის საფუძველზე ხდება მოქმედების არჩევა. უფრო ზუსტად რომ ვთქვათ, სტრატეგია აღიწერება შემდეგი განსაზღვრებით.

განსაზღვრება 1.6.1 *სტრატეგიას, რომლისთვისაც $a_t \in A(s_t)$ მოქმედების არჩევის ალბათობა დროის t მომენტში უდრის*

$$\left.\pi_t(a \mid s_t) = \frac{e^{q_t(a)/\tau}}{\sum\limits_{a \in A(s_t)} e^{q_t(a)/\tau}}, \quad \tau > 0\right\}$$

მნიშვნელობას, softmax-სტრატეგია. ეწოდება.

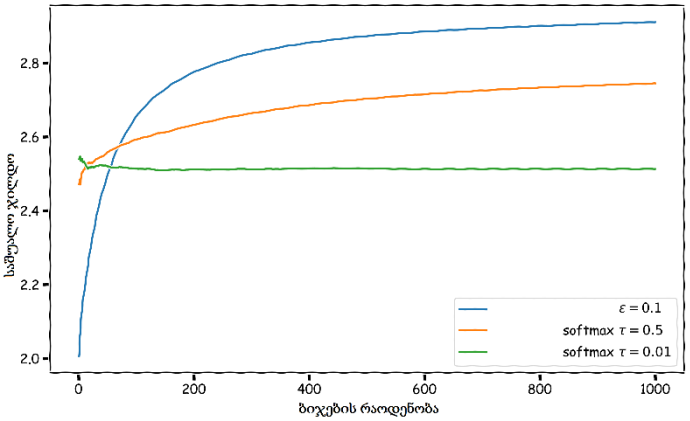
$\tau > 0$ პარამეტრი, რომელსაც კიდევ ტემპერატურასაც უწოდებენ, მოდელის პარამეტრია. τ ტემპერატურის დიდ მნიშვნელობებზე სტრატეგია უფრო კვლევით ხასიათს იძენს, მცირე

მნიშვნელობებზე კი — უახლოვდება ხარბს. ნახატზე 15.6 (რომელიც კვლავ აგებულია $t \geq 1$ პირობის გათვალისწინებით) წარმოდგენილია $\varepsilon = 0.1$ პარამეტრის მქონე $\varepsilon-$ხარბი ალგორითმისა და softmax ალგორითმის შედარება τ ტემპერატურის სხვადასხვა მნიშვნელობაზე.

ადვილად გასაგებია, რომ კონკრეტულად ჩვენს შემთხვევაში $\varepsilon-$ხარბი ალგორითმი უკეთეს შედეგს იძლევა. ამასთან ერთად, ამა თუ იმ მეთოდის არჩევის უფლებას, ჩვეულებრივ, მკვლევარს უტოვებენ. შევნიშნავთ ასევე, რომ $\tau = 0.01$ შემთხვევაში softmax-ალგორითმი ძალიან უახლოვდება ხარბს.

შენიშვნა 1.6.1 ზემოთ მოცემულ ალბათობათა განაწილებას ფიზიკაში ხშირად უწოდებენ გიბსისა და ბოლცმანის განაწილებას, რადგან იგი დაკავშირებულია ამერიკელი მეცნიერის ჯოსაია გიბსისა *(Josiah Willard Gibbs,1839–1903) და ავსტრიელი ფიზიკოსის და ფილოსოფოსის ლუდვიგ ბოლცმანის (Ludwig Eduard Boltzmann, 1844-1906) სახელებთან. კიდევ ერთხელ ხაზს ვუსვამთ იმ გარემოებას, რომ* $\tau \to +\infty$ *შემთხვევაში ჩვენ ვიღებთ* $a \in A(s_t)$ *მოქმედებათა არჩევის თითქმის ერთნაირ ალბათობებს და, მაშასადამე, სტრატეგია მიისწრაფვის აბსოლუტურად კვლევითისკენ. იმ შემთხვევაში კი, როცა* $\tau \to 0+$*, სტრატეგია მიისწრაფვის ხარბისკენ, ვინაიდან ყველა ექსპონენტთა შორის მეტი წონა აქვს იმას, რომლისთვისაც* $q_t(a)$ *მეტია.*

შენიშვნა 1.6.2 ისევე, როგორც $\varepsilon-$ *ხარბი სტრატეგიის შემთხვევაში, აზრი აქვს* τ *პარამეტრის შემცირებას დროის განმავლობაში (ბიჯების ზრდისას). როგორ ვაკეთოთ ეს და ვაკეთოთ კი საერთოდ — საკითხებია, რომლებიც გადაეცემა მკვლევარს ექსკლუზიური გადაწყვეტილების მისაღებად.*



ნახატი 15.6 - ხარბი და softmax ალგორითმების შედარება

1.7 ნდობის ინტერვალის ზედა საზღვრის მეთოდი

განვიხილოთ კომპრომისის პოვნის კიდევ ერთი ხერხი «შესწავლა-გამოყენება» პრობლემის გადაწყვეტისას. ამ მეთოდმა მიიღო მაქსიმალური ზედა შეფასების მეთოდის, ანუ ნდობის ინტერვალის ზედა საზღვრის მეთოდის (Upper Confidence Bound, UCB) სახელწოდება. მეთოდის იდეა შეიძლება ასეთნაირად აღვწეროთ :

რაც უფრო იშვიათად იყო არჩეული რომელიღაცა a მოქმედება, მით ნაკლებ ზუსტია მისი ღირებულების შეფასება (მოქმედება ხომ ნაკლებად შესწავლილია). აღწერილი პრობლემის გადასაწყვეტად სავსებით გონივრული ჩანს a მოქმედების ღირებულების ახალ შეფასებად განხილვა ღირებულების მიმდინარე შეფასებისა და ჭეშმარიტ მნიშვნელობასთან ამ შეფასების

«სიახლოვის» ჯამის — ჭეშმარიტი ღირებულებისათვის გარკვეული ნდობის ინტერვალის მარჯვენა საზღვრის.

აღვწეროთ ეს მეთოდი ფორმალურად.

დავუშვათ, რომ აგენტი იმყოფება s_t მდგომარეობაში, სადაც $t \geq 0$. ასევე დავუშვათ, რომ $N_0(a) = 0$ და $N_t(a)$ — იმ შემთხვევათა რაოდენობაა, როცა მოხდა a მოქმედების არჩევა $(t-1)$ ბიჯამდე ჩათვლით და ამასთან ერთად $(t \geq 1)$.

1. თუ არსებობს $a \in A(s_t)$ მოქმედება, ისეთი რომ $N_t(a) = 0$, მაშინ

$$A_t = \{a \in A(s_t) : \ N_t(a) = 0\}$$

და $a_t \in A_t$ მოქმედების არჩევის სტრატეგია ასეთია :

$$\pi_t(a \mid s_t) = \begin{cases} \dfrac{1}{|A_t|}, a \in A_t \\ 0, \quad a \notin A_t \end{cases}.$$

2. წინააღმდეგ შემთხვევაში

$$A_t = \underset{a \in A(s_t)}{\operatorname{Arg\,max}} \left\{ q_t(a) + \delta \sqrt{\frac{\ln t}{N_t(a)}} \right\}$$
$$\delta > 0$$

და $a_t \in A_t$ მოქმედების არჩევის სტრატეგია ასეთია :

$$\pi_t(a \mid s_t) = \begin{cases} \dfrac{1}{|A_t|}, \ a \in A_t \\ 0, \quad a \notin A_t \end{cases}.$$

სხვა სიტყვებით რომ ვთქვათ, თუ მოქმედება არასოდეს იყო არჩეული, მაშინ სწორედ მისი არჩევა ხდება შემდეგ მოქმედებად (თუ ასეთი მოქმედება რამდენიმეა, არჩევა მათ შორის შემთხვევით ხორციელდება). თუ ასეთი მოქმედებები არ არსებობს, მაშინ არჩევაზე გავლენას ახდენს, ერთი მხრივ, მოქმედების $q_t(a)$ ღირებულება, ხოლო მეორე მხრივ — მასზე ჩატარებული კვლევის ხარისხი.

გარკვეული თვალსაზრისით სწორედ $\delta > 0$ პარამეტრი ასახავს ხსენებულ ხარისხს და მის წვლილს. უფრო დაწვრილებითი ახსნა-განმარტებები შეიძლება მოიპოვოთ დამატებით მასალაში ამ საკითხთან დაკავშირებით.

ლემა 1.7.1 *დავუშვათ, რომ $X_1, X_2, \ldots, X_n$ — ანარჩევია ξ განერალური ერთობლიობიდან, ამასთან ერთად $X_i \in [0,1]$ და, გარდა ამისა,*

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i .$$

მაშინ სამართლიანია ფინელი სტატისტიკოსის ვასილი ჰაფდინგის (Wassily Hoeffding 1914–

1991) მიერ დამტკიცებული უტოლობა, რომელიც იძლევა იმ ალბათობის ზედა საზღვარს, რომ შემთხვევით სიდიდეთა ჯამი გადაიხრება თავისი მათემატიკური ლოდინიდან :

$$\mathrm{P}(\overline{X} - \mathrm{E}\xi \geq \varepsilon) \leq e^{-2n\varepsilon^2} \ .$$

დავუშვათ, რომ $U \geq 0$. თუ გადავწერთ უტოლობას ჩვენს აღნიშვნებში და ასევე ვისარგებლებთ ამ უკანასკნელის სიმეტრიულობით, მაშინ მივიღებთ :

$$\mathrm{P}\left(\mathrm{E}\xi - q_t\left(a\right) \geq U\right) \leq e^{-2N_t(a)U^2} .$$

სხვანაირად,

$$\mathrm{P}\left(\mathrm{E}\xi \geq \underbrace{\left(q_t\left(a\right)+U\right)}_{\text{ჯამი}}\right) \leq e^{-2N_t(a)U^2} .$$

ამრიგად, ჩვენ შევაფასეთ ზემოდან ალბათობა იმისა, რომ ჭეშმარიტი «ღირებულება» მეტია ორი შესაკრების ჯამზე, სადაც პირველი შესაკრები არის მოქმედების ღირებულების შეფასება, ხოლო მეორე — გარკვეული U რიცხვი, რომელსაც ეწოდება ნდობის ინტერვალის ზედა საზღვარი (ინგლ. Upper Confidence Bound, UCB). შემოვიღოთ აღნიშვნა :

$$\mathrm{p} = e^{-2N_t(a)U} ,$$

საიდანაც

$$\ln \mathrm{p} = -2N_t\left(a\right)U^2 \Rightarrow U = \sqrt{\frac{\ln \mathrm{p}}{-2N_t\left(a\right)}} .$$

p შეიძლება ავირჩიოთ, მაგალითად, როგორც t არგუმენტის ფუნქცია და ამასთან გონივრულია შემდეგი დაშვება :

$$\mathrm{p}\left(t\right) \underset{t\to+\infty}{\to} 0 .$$

მაშინ, თუ დავუშვებთ, რომ

$$\mathrm{p} = \frac{1}{t^{2\delta^2}} ,$$

მივიღებთ გამოსახულებას :

$$U = \delta\sqrt{\frac{\ln t}{N_t\left(a\right)}} ,$$

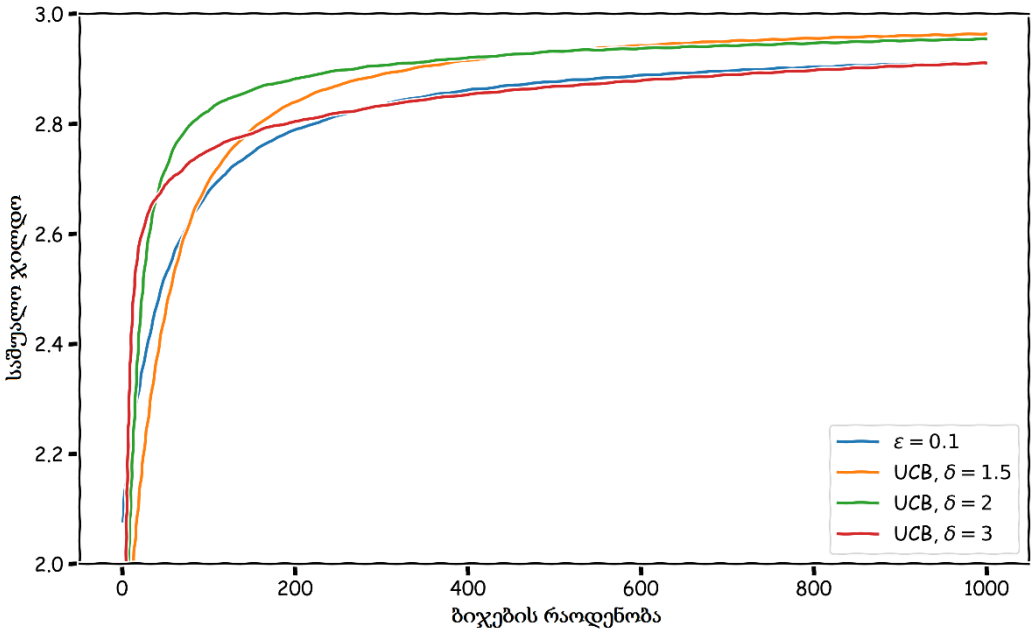
რომელიც გამოიყენება ზემოთ შემოთავაზებულ ალგორითმში.

შენიშვნა 1.7.1 *ჰაფდინგის უტოლობის პირობებში იგულისხმება, რომ ანარჩევის ნებისმიერი X_i ელემენტი აკმაყოფილებს $X_i \in [0,1]$ მოთხოვნას, ამიტომ მეთოდის გამოყენების კორექტულობისათვის საჭიროა წინასწარი ნორმირების ჩატარება.*

ნახატზე 15.7 წარმოდგენილია სხვადასხვა δ პარამეტრის მქონე UCB მეთოდისა და $\varepsilon-$ხარბი ალგორითმის შედარების შედეგები.

ნათლად შესამჩნევია, რომ დაკვირვების პირობებში რაც უფრო დიდია δ, მით უფრო ადრე გადის ალგორითმი პირობით «პროდუქტიულობის პლატოზე».

იმ შემთხვევაში, როცა $\delta=1.5$, კვლევითი მოქმედებების წილი ნაკლებია, ვიდრე მაშინ, როცა $\delta=2$ და $\delta=3$, ამიტომ საჭიროა მეტი დრო კვლევაზე, შესწავლაზე, რაც, მიუხედავად ამისა, დადებითი გავლენის მატარებელია გრძელვადიან პერსპექტივაში.



ნახატი 15.7 - $\varepsilon-$ხარბი და UCB ალგორითმების შედარება.

შენიშვნა 1.7.2 *აღსანიშნავია ასევე, რომ აზრი აქვს δ პარამეტრის შემცირებას დროის განმავლობაში, იმისათვის რომ მოხდეს მახვილის წანაცვლება კვლევიდან (შესწავლიდან) გამოყენებაზე (ექსპლუატაციაზე) და, ამრიგად, გაიზარდოს საშუალო ჯილდო (ანაზღაურება).*

1.8 ოპტიმისტური საწყისი შეფასებები

აქამდე განხილულ ყველა შემთხვევაში ამა თუ იმ მოქმედების არჩევა ემყარებოდა მისი ღირებულების შეფასებას. ამასთან ერთად ყველა მოქმედებისათვის საწყისი შეფასებები მიჩნეული იყო ტოლ სიდიდეებად იმის გამო, რომ აგენტს საწყის ეტაპზე არ გააჩნია რაიმე ინფორმაცია ამა თუ იმ მოქმედებათა ღირებულებების შესახებ. მაგალითად, $\varepsilon-$ხარბი ალგორითმის გამოყენებისას მოქმედებათა ღირებულებების შეფასებები ზუსტდება თავად ამ მოქმედებათა არჩევის კვალდაკვალ.

მსგავსი იდეა გამოიყენება მეთოდშიც, რომელმაც «ოპტიმისტურ საწყის შეფასებათა მეთოდის » დასახელება მიიღო. ოპტიმიზმი, რომელზეც ეს დასახელება ამახვილებს ყურადღებას, მდგომარეობს იმაში, რომ ინიციალიზაციის დროს თითოეულ მოქმედებას ენიჭება ღირებულების აშკარად გადაჭარბებული მნიშვნელობა. შევეცადოთ ავხსნათ ნათქვამი იმ მაგალითზე, რომელსაც განვიხილავდით.

გავიხსენოთ, რომ ჯილდოების მნიშვნელობათა გენერირება ხდებოდა $\mathcal{N}_{2,1}, \mathcal{N}_{2.5,1}, \mathcal{N}_{3,1}$ ნორმალურ განაწილებათა მქონე გენერალური ერთობლიობებიდან. ამრიგად, თუ საწყის მნიშვნელობებად მივიჩნევთ $q_0(a) = q_0(b) = q_0(c) = 10$ სიდიდეებს (რაც აშკარად მეტია როგორც 2-ზე, ასევე 2.5-ზე და 3-ზეც), მაშინ ასეთი შეფასებები შეიძლება ჩაითვალოს ძალიან ოპტიმისტურად — ისინი დაიწყებს შემცირებას მოდელის სწავლებასთან ერთად. მოქმედებებს ავირჩევთ ხარბი სტრატეგიის საშუალებით. განვიხილოთ რამდენიმე ბიჯი.

ღირებულებათა შეფასებებს შევიტანთ შესაბამის ცხრილში.

ბიჯი	$q_t(a)$	$q_t(b)$	$q_t(c)$
$t = 0$	10	10	10

ვინაიდან ყველა მოქმედების ღირებულებათა შეფასებები ერთნაირია, ხდება ნებისმიერი მოქმედების არჩევა, მაგალითად, c მოქმედების. დავუშვათ, რომ ჯილდო $r_1 = 2.8$ მნიშვნელობისაა, მაშინ

$$q_1(c) = q_0(c) + \frac{1}{1+1}(r_1 - q_0(c)) = 10 + \frac{(2.8-10)}{2} = 6.4.$$

დავამატოთ მნიშვნელობა ცხრილში :

ბიჯი	$q_t(a)$	$q_t(b)$	$q_t(c)$
$t = 0$	10	10	10
$t = 1$	10	10	6.4

ახლა, ხარბი ალგორითმის შესაბამისად, აგენტი აირჩევს ერთ-ერთ მოქმედებას a ან b მოქმედებათა წყვილიდან, რადგან ღირებულების მათი მიმდინარე შეფასება უფრო მაღალია. დავუშვათ, რომ არჩეულია a მოქმედება და მიღებულია $r_2 = 2.2$ ჯილდო. მაშინ

$$q_2(a) = q_1(a) + \frac{1}{1+1}(r_2 - q_1(a)) = 10 + \frac{(2.2-10)}{2} = 6.1.$$

შევიტანოთ მიღებული შედეგი ცხრილში.

ბიჯი	$q_t(a)$	$q_t(b)$	$q_t(c)$
$t = 0$	10	10	10
$t = 1$	10	10	6.4
$t = 2$	6.1	10	6.4

შემდეგ, ცხადია, არჩეული იქნება b მოქმედება.

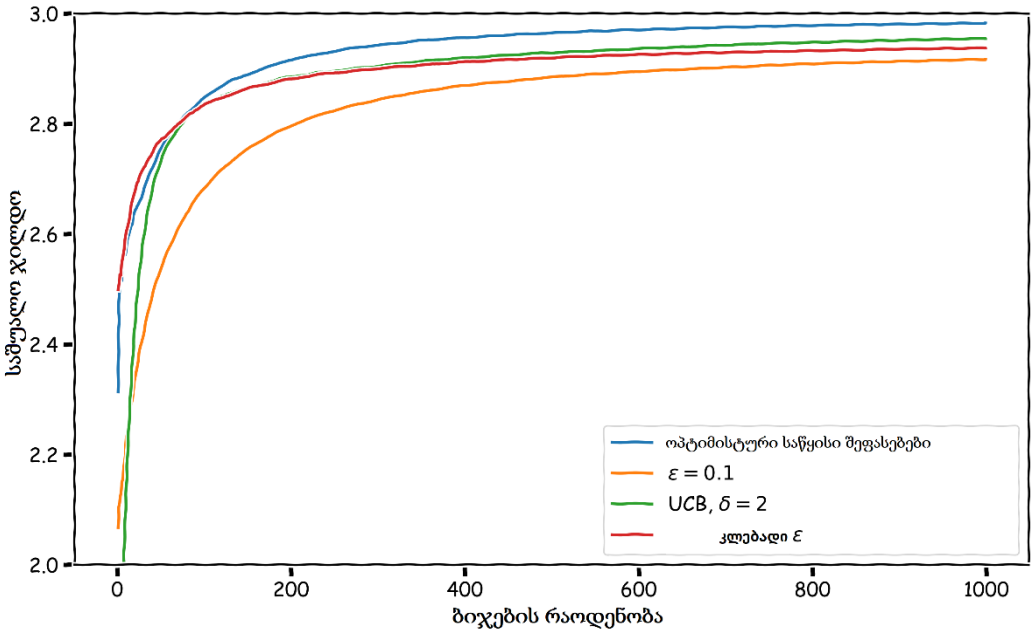
ამ პროცესში, აგენტს მოუწევს ერთი მოქმედებიდან მეორე მოქმედებაზე გადართვა, თუ იმედგაცრუებული აღმოჩნდება მიღებული ანაზღაურებით.

ყველა მოქმედება მოსინჯული იქნება რამდენჯერმე, რაც ჭეშმარიტ მნიშვნელობებთან

საკმარისად მიახლოებულ შეფასებათა აგების საშუალებას უზრუნველყოფს.

ძნელი არ არის იმის შემჩნევა, რომ «ოპტიმიზმის» გამო პირველ ბიჯებზე ხდება შესწავლა, ხოლო როცა ღირებულებათა შეფასებები დაზუსტებულია, ყურადღება გადაიტანება გამოყენებაზე.

ოპტიმისტურ საწყის შეფასებათა მქონე ხარბი ალგორითმის და ადრე განხილული ალგორითმების (1000 თამაში, თითოეული 1000 ბიჯით, აგება $t \geq 1$ პირობით) შედარება წარმოდგენილია ნახატზე 15.8.



ნახატი 15.8 - სხვადასხვა ალგორითმის შედარება.

ადვილი შესამჩნევია, რომ სწავლების დასაწყისში ოპტიმისტური ალგორითმი აგებს ε – ხარბ ალგორითმთან კლებადი ε – ით, რადგან «იკვლევს», «სწავლობს» გარემოს, მაგრამ გარკვეული დროის შემდეგ მასზე გაცილებით უფრო მაღალ შედეგებს გვიჩვენებს.

2 ამოცანის ზოგადი დასმა

2.1 აგენტისა და გარემოს თანამოქმედება

ადრე, k – ხელიან ბანდიტებზე საუბრისას, ჩვენ უკვე აღვწერეთ «თითებზე» აგენტისა და გარემოს თანამოქმედების სქემა. ამ ადრინდელ აღწერაში ლამის ყველაზე მთავარი, თუ საკითხს მკაცრად მივუდგებით, — ეს აგენტის მიერ სტრატეგიის არჩევაა დროის t მომენტში. მოდით კიდევ ერთხელ დაწვრილებით განვიხილოთ მთელი პროცესი.

დროის ყოველ დისკრეტულ $t \in \{0,1,2,\ ...,n\}, n \in N \cup \{0,\infty\}$ მომენტში, აგენტი იმყოფება გარკვეულ $s_t \in S$ მდგომარეობაში, სადაც S — ყველა შესაძლო მდგომარეობის სიმრავლეა. $\pi_t(a \mid s_t)$ სტრატეგიის შესაბამისად, აგენტი ირჩევს ერთ-ერთ შესაძლო $a_t \in A(s_t)$ მოქმედებას, სადაც $A(s_t)$ — აგენტისათვის s_t მდგომარეობაში მისაწვდომ მოქმედებათა სიმრავლეა. იმ შემთხვევაში, თუ ნებისმიერ ბიჯზე აგენტისათვის ყველა მოქმედებაა მისაწვდომი, დავწერთ $a_t \in A$, სადაც $A = \bigcup_{s \in S} A(s)$ — ყველა შესაძლო მოქმედების სიმრავლეა. მოქმედების

განხორციელების შემდეგ აგენტი იღებს $r_{t+1} \in \mathbb{R}$ ჯილდოს და გადადის s_{t+1} მდგომარეობაში. შედეგად, სწავლების მთელი ისტორია შეიძლება განიხილებოდეს შემდეგი სამეულების (s_i მდგომარეობა, a_i მოქმედება, r_i ჯილდო) თანამიმდევრობად :

$$\left.\begin{array}{l} s_0, a_0, r_1; \\ s_1, a_1, r_2; \\ \ldots,\ldots,\ldots; \end{array}\right\}.$$

შენიშვნა 2.1.1 *ცალკე აღსანიშნავია, რომ, ვინაიდან მოქმედებათა არჩევა ხდება სტრატეგიის შესაბამისად (ალბათურად და არა დეტერმინირებულად), ამიტომ ჯილდო, ისევე, როგორც აგენტის შემდეგი მდგომარეობა, გარკვეული აზრით შემთხვევითია. ამიტომ შეიძლება ვთქვათ, რომ გარემო წარმოქმნის როგორც r_{t+1} ჯილდოს a_t მოქმედებისათვის, ასევე აგენტის მომდევნო s_{t+1} მდგომარეობასაც ამას წინათ არჩეულ მოქმედებათა და მდგომარეობათა შესაბამისად. ეს კი იმას ნიშნავს, რომ s_t და r_t შემთხვევითი სიდიდეებია, თანაც*

$$\left.\begin{array}{l} s_{t+1} \sim p(s \mid s_t, a_t, r_t, s_{t-1}, a_{t-1}, r_{t-1}, \ldots, s_0, a_0) \\ r_{t+1} \sim p(s \mid s_t,\ a_t,\ r_t,\ s_{t-1},\ a_{t-1},\ r_{t-1}, \ldots,\ s_0,\ a_0) \end{array}\right\},$$

სადაც p — რომელიღაც ალბათური განაწილებაა.

შედეგად, იმ დასკვნამდე მივდივართ, რომ განსახილველი სამეული (აგენტის მდგომარეობა, მოქმედების არჩევის სტრატეგია, ჯილდო არჩეული მოქმედებისათვის) ყოველ მომდევნო ბიჯზე დამოკიდებულია ყოველ წინა ბიჯზე. პრობლემის ასეთი დასმა უკიდურესად რთულია როგორც თეორიის თვალსაზრისით (საქმე გვაქვს მრავალგანზომილებიან ერთობლივ განაწილებებთან), ასევე პრაქტიკის თვალსაზრისითაც : წინა ინფორმაციის უზარმაზარი რაოდენობის შენახვის აუცილებლობა. ამიტომ ამ ლექციის ფარგლებში ჩვენ გავიადვილებთ ამოცანას და ჩავთვლით, რომ მომდევნო ბიჯი დამოკიდებულია ვითარებათა მხოლოდ მიმდინარე მდგომარეობაზე და წარსული გამოცდილება შესაძლებელია საერთოდ არ გავითვალისწინოთ. შემოვიტანოთ შემდეგი განსაზღვრება.

განსაზღვრება 2.1.1 *გადაწყვეტილებათა მიღების მარკოვის პროცესი (გმმპ) ეწოდება (S, A, R, P) ოთხეულს, სადაც შემდეგი აღნიშვნებია მიღებული :*

1. S *— აგენტის შესაძლოა მდგომარეობათა სიმრავლე,*
2. A *— აგენტისათვის მისაწვდომ მოქმედებათა სიმრავლე,*
3. $R : S \times A \to \mathbb{R}$ *— წახალისების ფუნქცია (ანუ მოსალოდნელი ჯილდო, ანაზღაურება) s მდგომარეობიდან s' მდგომარეობაში გადასვლისას a მოქმედების არჩევის შედეგად,*
4. $P : S \times A \to \Pi(S)$ *— მდგომარეობებს შორის გადასვლის ფუნქცია, სადაც $\Pi(S)$ — ალბათობათა განაწილებების სიმრავლე აგენტის შესაძლო მდგომარეობათა S სიმრავლეზე.*

ამასთან ერთად მომდევნო გადასვლების ალბათობები დამოკიდებული არ არის წინა გადასვლათა ისტორიაზე, ესე იგი

$$\left.\begin{array}{l} s_{t+1} \sim \kappa(s_t, a_t) \in \Pi(S) \\ P_\kappa(s_{t+1} = s' \mid s_t, a_t, r_t, s_{t-1}, a_{t-1}, r_{t-1}, \ldots, s_0, a_0) = P_\kappa(s_{t+1} = s' \mid s_t, a_t) \end{array}\right\}.$$

ამრიგად, $\kappa(s_t, a_t)$ — ეს s_t მდგომარეობიდან გარკვეულ მდგომარეობაში გადასვლის ალბათობათა განაწილებაა a_t მოქმედების განხორციელების შედეგად, გადაწყვეტილებათა მიღების მარკოვის პროცესი — ეს მოდელია, რომლის გამოყენებისას აგენტს გარკვეული თვალსაზრისით «მეხსიერება არ აქვს» : t ბიჯზე s' მდგომარეობაში გადასვლის ალბათობა დამოკიდებულია მხოლოდ მიმდინარე s_t მდგომარეობასა და არჩეულ a_t მოქმედებაზე.

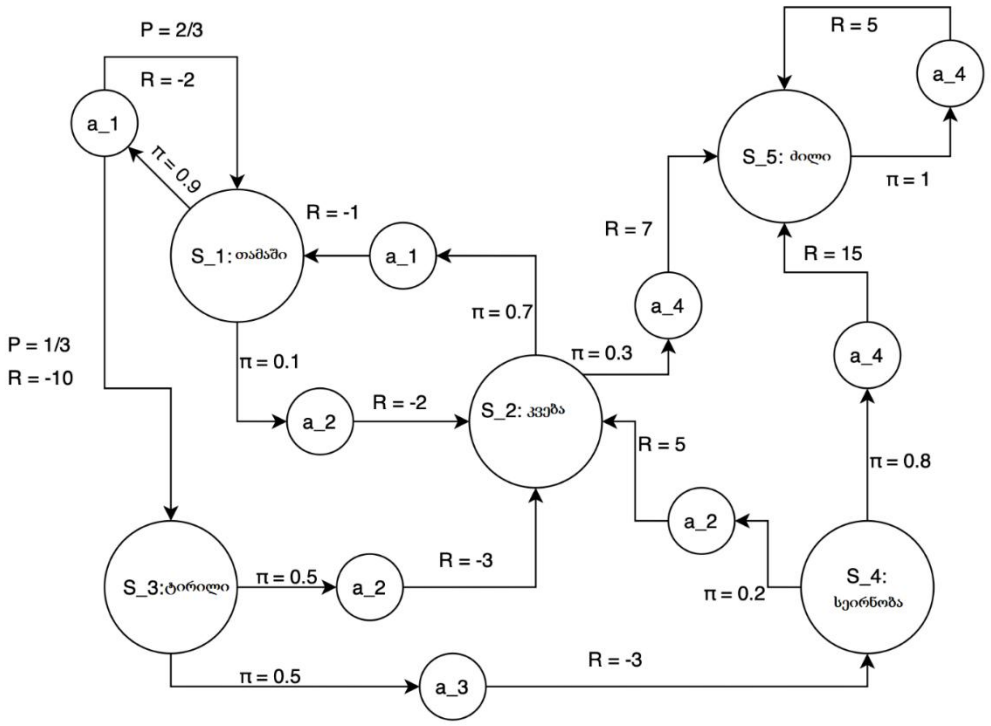
შენიშვნა 2.1.2 (!) *განსაზღვრების თითოეული პუნქტის უფრო დაწვრილებით ახსნამდე, შემდეგ მნიშვნელოვან მომენტს შევეხოთ. სწავლება განმტკიცებით იყენებს თავის საფუძველში გადაწყვეტილებათა მიღების მარკოვის პროცესს, მაგრამ არ აღიწერება «პირდაპირი სახით» მის მიერ. განსხვავება ისაა, რომ აგენტი ყოველ* s *მდგომარეობაში ირჩევს გარკვეულ* $a \in A(s)$ *მოქმედებას* $\pi(a \mid s)$ *ალბათობით (ესე იგი მოქმედებს რაღაც სტრატეგიის შესაბამისად). ამიერიდან აგენტის სტრატეგია ჩვენ მიერ მიჩნეული იქნება გადაწყვეტილებათა მიღების მარკოვის პროცესის განუყოფელ, შემადგენელ ნაწილად ამ გარემოების ცალკე დამატებითი ხაზგასმის გარეშე.*

ახლა კი დავუბრუნდეთ განსაზღვრების ყოველ პუნქტს დაწვრილებით. აგენტის მდგომარეობათა სიმრავლის და მისაწვდომ მოქმედებათა სიმრავლის ცნებები ჩვენთვის უკვე ცნობილია. R — ეს ფუნქციაა, რომელიც მოსალოდნელ ჯილდოს, ანაზღაურებას იძლევა $a \in A(s_t)$ მოქმედების არჩევისათვის s_t მდგომარეობაში ყოფნისას. უხეშად რომ ვთქვათ, როცა შესასვლელს მიეწოდება მიმდინარე მდგომარეობა და არჩეული მოქმედება, მაშინ R ფუნქციას გამოაქვს რიცხვი — მოსალოდნელი ჯილდო ასეთი «სვლის» დროს. P ფუნქცია კი იძლევა მდგომარეობებს შორის გადასვლათა ალბათობების განაწილებას იმ შემთხვევაში, როცა აგენტი იმყოფება s_t მდგომარეობაში და არჩევს a_t მოქმედებას.

მაგალითი 2.1.1 *ნახატზე 15.9 მოცემულია გადაწყვეტილებათა მიღების მარკოვის პროცესის მაგალითი.*

მოცემულ შემთხვევაში აგენტია ბავშვი, რომელსაც შეუძლია იყოს ერთ-ერთ მდგომარეობაში ხუთი მდგომარეობიდან (დიდი წრეები) :

s_1 : *თამაში.* s_2 : *კვება.* s_3 : *ტირილი.* s_4 : *სეირნობა.* s_5 : *ძილი.*



ნახატი 15.9 - გადაწყვეტილებათა მიღების მარკოვის პროცესი, რომელშიც აგენტი არის ბავშვი.

დავუშვათ, რომ აგენტისათვის მისაწვდომია მხოლოდ ოთხი მოქმედება (პატარა წრეები):

a_1 : *თამაშში ჩართვა.*

a_2 : *საჭმლის მოთხოვნა.*

a_3 : *სასეირნოდ წასვლა.*

a_4 : *დასაძინებლად დაწოლა.*

ახლა კი იმის შესახებ, როგორ წავიკითხოთ და გავიგოთ ინფორმაცია, რომელსაც ნახატი 15.9 იძლევა ? მდგომარეობათა შესაბამისი წრეებიდან გამომავალ ისრებს მკითხველი მიჰყავს ამ მდგომარეობაში მისაწვდომ მდგომარეობებისკენ (ამ ისრებთან განთავსებული π რიცხვები ნაკარნახევია აგენტის სტრატეგიით : ისინი უჩვენებს ისეთი მოქმედების არჩევის ალბათობას, რომლისკენაც მიდის ისარი იმ მდგომარეობაში ყოფნისას, საიდანაც ეს ისარია გამოსული). შემდეგ, მოქმედების არჩევის დასრულებისას, ჩნდება ალბათობათა განაწილება S სიმრავლეზე : P რიცხვები ისრებთან, რომლებიც მოქმედებათა შესაბამის წრეებიდან გამოდის, უჩვენებს როგორი ალბათობით და რომელ მდგომარეობაში შეიძლება გადავიდეს აგენტი და ასევე როგორი R ჯილდოს მიღება შეუძლია მას. ასე, მაგალითად, s_1 მდგომარეობიდან a_1 მოქმედების არჩევისას (0.9 ალბათობით), აგენტი გადადის s_3 მდგომარეობაში $(1/3)$ – ის ტოლი ალბათობით $r = -10$ ჯილდოს შეძენით და უბრუნდება $(2/3)$ – ის ტოლი ალბათობით s_1 მდგომარეობას, რისთვისაც (-2) ანაზღაურებას იღებს.

s_1 *მდგომარეობიდან გადასვლის ალბათობათა განაწილება* a_1 *მოქმედების არჩევისას — ეს*

არის სწორედ $\kappa(s_1, a_1)$ განაწილება გადაწყვეტილებათა მიღების მარკოვის პროცესის განსაზღვრებიდან. ჩვენს მაგალითში იგი მოიცემა შემდეგი ცხრილით

	s_1	s_3
P	$2/3$	$1/3$

იმავდროულად, იმავე მდგომარეობაში a_2 მოქმედების არჩევისას (ალბათობით 0.1), ავტომატურად ხდება გადასვლა ($P=1$ ალბათობით) s_2 მდგომარეობაში $r=-2$ ჯილდოს მიღებით, სხვანაირად რომ ვთქვათ, $\kappa(s_1, a_2)$ განაწილება მოიცემა შემდეგი ცხრილით :

	s_2
P	1

და ასე შემდეგ.

ზოგად შემთხვევაში მდგომარეობათა S და მოქმედებათა A სიმრავლეები შეიძლება იყოს როგორც სასრული, ასევე უსასრულოც. ამ ლექციაში ჩვენ შემოვიფარგლებით შემთხვევით, როცა ორივე ეს სიმრავლე სასრულია. ასეთ პროცესებს სპეციალური სახელწოდებაც კი აქვს.

განსაზღვრება 2.1.2 *გადაწყვეტილებათა მიღების მარკოვის პროცესს, რომელშიც A და S სიმრავლეები სასრული არის, ფინიტური ეწოდება.*

აღსანიშნავია კიდევ ერთი მნიშვნელოვანი განსაზღვრება.

განსაზღვრება 2.1.3 *თუ T ბიჯზე გარემო გადადის მდგომარეობაში, სადაც აგენტისა და გარემოს ურთიერთქმედება (თამაში) წყდება, მაშინ ამ მდგომარეობას ტერმინალური ეწოდება.*

მაგალითად, თამაშში «ჯვრები-ნულები» (სასრული ზომის მინდვრით) ტერმინალური მდგომარეობა მიიღწევა ან მაშინ, როცა ერთ-ერთი მოთამაშე იმარჯვებს, ან როცა ცხრილის ყველა უჯრედი შევსებულია.

ეს მაგალითი რომ გასაგები გახდეს, საკმარისია ვიცოდეთ, რომ «ჯვრები-ნულები» — ლოგიკური თამაშია ორ მოწინააღმდეგეს შორის კვადრატულ მინდორზე.

მოთამაშეები რიგ-რიგობით ათავსებენ 3×3 მინდვრის თავისუფალ უჯრედებში ნიშნებს : ერთ-ერთი — ყოველთვის ჯვრებს, ხოლო მეორე — ყოველთვის ნულებს.

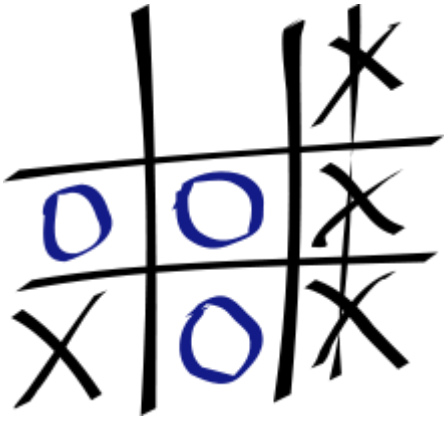
ის მოთამაშე, რომელიც მიყოლებით პირველი ჩაამწკრივებს თავის სამ ფიგურას ვერტიკალის, ჰორიზონტალის ან დიდი დიაგონალის გასწვრივ, იგებს თამაშს.

თუ მოთამაშეებმა შეავსეს ყველა (ცხრა) უჯრედი და აღმოჩნდა, რომ არც ერთ ვერტიკალზე, ჰორიზონტალზე ან დიდ დიაგონალზე არ არის სამი ერთნაირი ნიშანი, მაშინ პარტია ყაიმით მთავრდება.

პირველ სვლას აკეთებს ის მოთამაშე, რომელსაც ჯვრების დასმა ერგო წილად. ჩვეულებრივ, პარტიის დასრულების შემდეგ გამარჯვებული მხარე ახორციელებს (ნულებით ან ჯვრებით

შედგენილი) თავისი უწყვეტი სამეულის გადახაზვას.

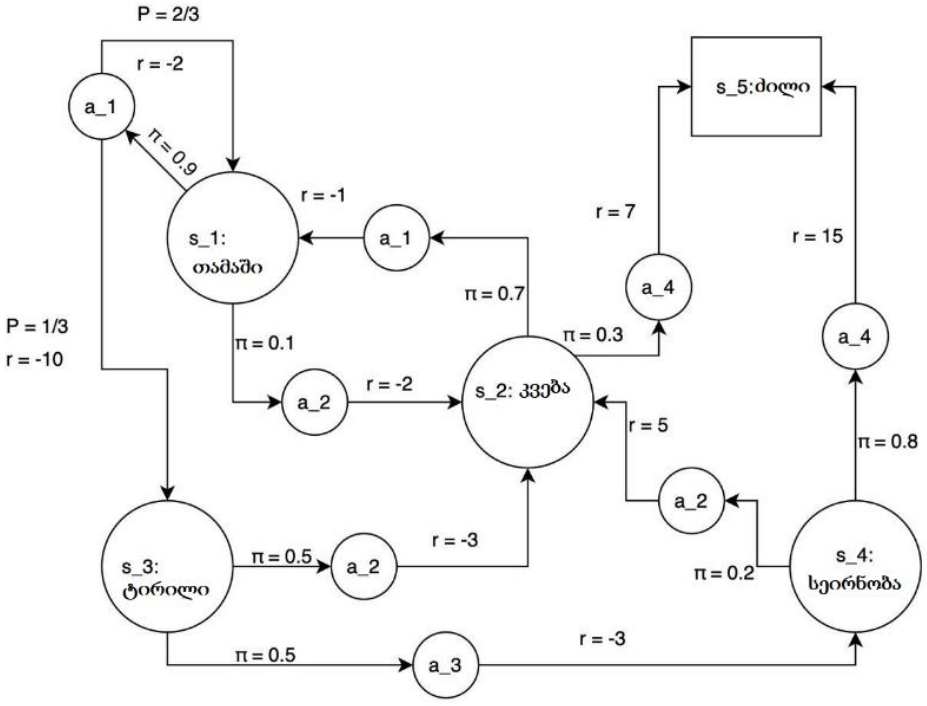
ასე რომ, მოგებული პარტია შემდეგნაირად შეიძლება გამოიყურებოდეს :



მაგალითი 2.1.2 *ოდნავ შევცვალოთ ამას წინათ განხილული მაგალითი — ცვლილება ეხება* s_5 *მდგომარეობას. ახლა ეს მდგომარეობა ტერმინალური გახდა — ამ მდგომარეობიდან არც ერთი მოქმედებაა მისაწვდომი. მშობლებს შეიძლიათ დასვენება.*

ამრიგად, ძირითადი განსაზღვრებები შემოტანილია.

თუმცა გასარკვევი დარჩა მნიშვნელოვანი საკითხი : რა უნდა შეფასდეს «თამაშის» შედეგად ?



ნახატი 15.10 - გადაწყვეტილებათა მიღების მარკოვის პროცესი ტერმინალური მდგომარეობით.

განსაზღვრება 2.1.4 *დავუშვათ, რომ ჩვენ ვიმყოფებით* t *ბიჯზე, ხოლო აგენტისა და გარემოს თანამოქმედება ბიჯების უსასრულო რაოდენობას გულისხმობს. მაშინ მოსალოდნელი სარგებელი* t *ბიჯის შემდეგ ეწოდება შემდეგ სიდიდეს :*

$$G_t = r_{t+1} + r_{t+2} + \ldots = \sum_{k=0}^{\infty} r_{t+k+1} \ .$$

თუ განიხილება ურთიერთქმედება, რომელიც გულისხმობს ტერმინალური მდგომარეობის მიღწევას ბიჯების $t \in \{0,1,2,...,T\}$ სასრული რაოდენობის პირობებში, მაშინ მოსალოდნელი სარგებელი ეწოდება შემდეგ სიდიდეს :

$$G_t = r_{t+1} + r_{t+2} + \ldots + r_T \ .$$

ალბათ, შემოღებული განსაზღვრება — ერთ-ერთი ყველაზე ბუნებრივია : არაფერია იმაზე ადვილი, რომ უბრალოდ შეკრიბო მომავალში მისაღები ყველა (თუმცა, ჯერ უცნობია, როგორი) ჯილდო.

შენიშვნა 2.1.3 *ცალკე აღსანიშნავია, რომ აგენტისა და გარემოს უსასრულო თანამოქმედება (ესე იგი $T = +\infty$) — დაკვირვებათა, მონაცემთა ძირითადი სიმრავლისაგან მკვეთრად განსხვავებული, გადახრილი, ამოვარდნილი მოვლენა კი არა, არამედ ძალიან ხშირი სიტუაციაა.*

ამასთან, ხშირად აგენტის ეგრეთ წოდებული «სიცოცხლის ხანგრძლივობა» განსაზღვრული არ არის. მაგალითად, ლაბირინთიდან გამოსვლა შეიძლება სვლათა სასრული (წინასწარ უცნობი!) რაოდენობის მიღწევისას დასრულდეს და შეიძლება საერთოდ არ დამთავრდეს (თუ აგენტს გზა აებნა და იგი დაიკარგა).

იმავდროულად, ჯამური ჯილდოს მაქსიმიზაცია — საკმაოდ გულუბრყვილო საქმედ გამოიყურება. მაგალითად, შეიძლება ვაგროვოთ «შესაკრები ჯილდოები» აუჩქარებლად გადაადგილებისას გამოსასვლელისკენ ლაბირინთიდან, თუმცა, ალბათ, უფრო ლოგიკური იქნებოდა იქედან რაც შეიძლება სწრაფად გამოსვლა და ქცევის, მოქმედების სახელმძღვანელო აზრად, იდეად შემდეგი ფრაზის გამოყენება : «*რაც უფრო მალე გექნება სარგებელი, მით უკეთესია*». ასეთ ვითარებაში გაცილებით უკეთესი იქნება შემდეგი ცნების გამოყენება :

განსაზღვრება 2.1.5 *დავუშვათ, რომ ვართ t ბიჯზე, ხოლო აგენტისა და გარემოს ურთიერთქმედება გულისხმობს ბიჯების უსასრულო რიცხვს. მაშინ დაყვანილი მოსალოდნელი სარგებელი ეწოდება შემდეგ სიდიდეს :*

$$G_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \ldots = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \ ,$$

სადაც $\gamma \in [0,1]$ სიდიდეს დისკონტირების კოეფიციენტი ეწოდება.

თუ განიხილება თანამოქმედება, რომელიც გულისხმობს ტერმინალური მდგომარეობის მიღწევას ბიჯების $t \in \{0,1,2,...,T\}$ სასრული რიცხვის პირობებში, მაშინ დაყვანილი მოსალოდნელი სარგებელი ეწოდება შემდეგ სიდიდეს :

$$G_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \ldots + \gamma^{T-t-1} r_T \ .$$

დაყვანილ მოსალოდნელ სარგებელს ასევე შეიძლება «*ჯილდოს გაუფასურების*» პრინციპიც ვუწოდოთ. აღწერილი მიდგომის შესაბამისად, აგენტი ცდილობს მოქმედებათა ისეთნაირად არჩევას, რომ გაუფასურებული მიღებული ჯილდოების ჯამი უდიდესი იყოს. ესე იგი ყოველ

ბიჯზე აგენტი ცდილობს ისეთი a_t მოქმედების არჩევას, რომ უდიდესი შესაძლო ჯილდო ძალიან არ გაუფასურდეს (რაც, გასაგები უნდა იყოს, დაყოვნების გამო ხდება).

ამის აღწერა სხვანაირადაც შეიძლება. დისკონტირების γ კოეფიციენტი პასუხისმგებელია იმაზე, რამდენად მნიშვნელოვანია ყოველი მომავალი სარგებელი, ესე იგი არეგულირებს აგენტის «შორსმჭვრეტელობას» «წინდახედულობას», «გამჭრიახობას», მოვლენათა გათვალისწინების უნარს. თუ $\gamma = 0$, მაშინ აგენტი ყურადღებას არ აქცევს ყველა მომდევნო სარგებელს და ცდილობს მიმდინარე სარგებლის მაქსიმიზაციას. პერსპექტივაში ამან შეიძლება გამოიწვიოს მოსალოდნელი სარგებლის დონის დაწევა. იმავდროულად, დისკონტირების γ კოეფიციენტი, რომელიც ახლოსაა ერთთან, ზრდის მომავალ ჯილდოთა მნიშვნელოვანებას.

შენიშვნა 2.1.4 *უნდა აღინიშნოს, რომ მოსალოდნელი სარგებელი ბიჯების როგორც სასრული, ასევე უსასრულო რიცხვის შემთხვევაში, მიიღება ზემოთ მოცემულიდან, როცა* $\gamma = 1$.

სამართლიანია შემდეგი ლემა.

ლემა 2.1.1 *თუ* $\gamma \in (0,1)$ *და* $r_{t+1}, r_{t+2}, r_{t+3}, \ldots$ *მიმდევრობა შეზღუდულია, მაშინ დაყვანილ*

$$G_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}$$

სარგებელს სასრული მნიშვნელობა აქვს.

მტკიცებულება. ვინაიდან $\{r_{t+k+1}\}\big|_{k=0}^{\infty}$ მიმდევრობა შეზღუდულია, ამიტომ $|r_{t+k+1}| \le C$.

მაშინ

$$\left|\gamma^k r_{t+k+1}\right| \le C\gamma^k .$$

რადგან მწკრივი γ^k ზოგადი წევრით $\gamma \in (0,1)$ შემთხვევაში კრებადია (როგორც გეომეტრიული პროგრესია), ამიტომ თავდაპირველი მწკრივი იკრიბება აბსოლუტურად და, მაშასადამე, იკრიბება.

□

2.2 მოქმედებათა და მდგომარეობათა ღირებულების ფუნქციები

იმისათვის, რომ აგენტ ესმოდეს, როგორ უნდა იქცეოდეს გარემოსთან თანამოქმედებისას, ანუ როგორ სტრატეგიას უნდა იყენებდეს იგი, გარდაუვალი ხდება აგენტის მდგომარეობის აღმწერი გარკვეული აპარატის არსებობა. სწავლებაში განმტკიცებით გამოიყენება მიდგომა, რომელშიც ყოვე; მდგომარეობას ენიჭება რაღაც რიცხვითი სიდიდე — ამ მდგომარეობის ღირებულება. მდგომარეობის ღირებულება განკუთვნილია ამ მდგომარეობის დასახასიათებლად დაყვანილი სარგებლის თვალსაზრისით.

განსაზღვრება 2.2.1 *დავუშვათ, რომ* π *— აგენტის სტრატეგიაა.* s *მდგომარეობის ღირებულების ფუნქცია* π *სტრატეგიის გამოყენებისას ეწოდება შემდეგი სახის ფუნქციას :*

$$v_\pi(s) = \mathrm{E}_\pi\left(G_t \mid s_t = s\right) = \mathrm{E}_\pi\left(\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid s_t = s\right), \quad s \in S.$$

მდგომარეობათა ღირებულების ფუნქცია განსაზღვრავს დაყვანილი სარგებლის საშუალო მნიშვნელობას, თუ ჩვენ ვიწყებთ s მდგომარეობაში და ვხელმძღვანელობთ π სტრატეგიით. ყოველი მდგომარეობის ღირებულების ცოდნისას, ლოგიკურია ისეთი სტრატეგიის გამოყენება, რომელიც მიგვიყვანს მდგომარეობამდე მაქსიმალური ღირებულებით.

შენიშვნა 2.2.1 *ფორმალურად (ღირებულების ფუნქციის განსაზღვრებაში) ჩვენ წინაშეა სხვა არაფერია თუ არა პირობითი მათემატიკური ლოდინი — $s \in S$ მდგომარეობის ფუნქცია. s მდგომარეობის ცვლილებისას იცვლება $\pi(a \mid s)$ სტრატეგია და ამით იცვლება $v_\pi(s)$ მახასიათებელიც.*

ღირებულების შეფასების საკითხისადმი შეიძლება უფრო პედანტურადაც მიდგომა : დაყვანილი სარგებლის შეფასების განხორციელება სტარტის აღებისას არა იმდენად s მდგომარეობიდან, არამედ a მოქმედების არჩევისას ასევე. ამრიგად, მივადექით მოქმედების ღირებულების ცნებას.

განსაზღვრება 2.2.2 *დავუშვათ, რომ π — აგენტის სტრატეგიაა. a მოქმედების ღირებულების ფუნქცია s მდგომარეობაში π სტრატეგიის გამოყენებისას შემდეგი სახის ფუნქციას ეწოდება :*

$$q_\pi(s,a) = \mathrm{E}_\pi\left(G_t \mid s_t = s, a_t = a\right) = \mathrm{E}_\pi\left(\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid s_t = s, a_t = a\right).$$

კიდევ ერთხელ გავიმეოროთ, რომ თავისი არსით მოქმედების ღირებულების ფუნქცია ძალიან ჰგავს მდგომარეობის ღირებულების ფუნქციას მხოლოდ იმ განსხვავებით, რომ ჩვენ ვსაზღვრავთ დაყვანილ სარგებელს, თუ ვიწყებთ s მდგომარეობაში, ვირჩევთ a მოქმედებას და მივყვებით π სტრატეგიას.

თითქოს, ყველაფერი ლოგიკურია და, რომ ვიცოდეთ ყველაფერი გარემოსა და აგენტის შესახებ, მაშინ საკმარისი იქნებოდა ისეთი a მოქმედების არჩევა, რომელიც $q(s,a)$ მახასიათებლის მაქსიმიზაციას ახდენს. რა თქმა უნდა, ეს შეუძლებელია, მაგრამ, თურმე, ღირებულების ფუნქციებისათვის რეკურსიული თანაფარდობები სრულდება.

ლემა 2.2.1 *დავუშვათ, რომ $v_\pi(s)$ და $q_\pi(s,a)$ — მდგომარეობათა და მოქმედებათა ფუნქციებია, შესაბამისად. ასევე დავუშვათ, რომ ცნობილია მთელი ინფორმაცია გარემოზე : ცნობილია s მდგომარეობიდან s' მდგომარეობაში a მოქმედების შესრულებისას ყველა გადასვლის ალბათობა*

$$\mathcal{P}_{ss'}^a = \mathrm{P}\left(s_{t+1} = s' \mid s_t = s, a_t = a\right)$$

და ასევე ცნობილია ყველა მოსალოდნელი ჯილდო (პრემია, ანაზღაურება)

$$\mathcal{R}_{ss'}^a = \mathrm{E}_\pi\left(r_{t+1} \mid s_t = s, a_t = a, s_{t+1} = s'\right).$$

მაშინ სამართლიანია შემდეგი თანაფარდობები :

$$v_\pi(s) = \sum_{a\in A(s)} \pi(a\,|\,s) \sum_{s'\in S} \mathcal{P}^a_{ss'} \left(\mathcal{R}^a_{ss'} + \gamma v_\pi(s') \right)$$

და

$$q_\pi(s,a) = \sum_{s'\in S} \mathcal{P}^a_{ss'} \left(\mathcal{R}^a_{ss'} + \gamma \sum_{a'\in A(s')} \pi(a'\,|\,s') q_\pi(s',a') \right) = \sum_{s'\in S} \mathcal{P}^a_{ss'} \left(\mathcal{R}^a_{ss'} + \gamma v_\pi(s') \right).$$

მტკიცებულება. 1. დავამტკიცოთ პირველი ტოლობა. განსაზღვრების თანახმად, მათემატიკური ლოდინის წრფივობის თვისების გამოყენებით, მივიღებთ :

$$v_\pi(s) = \mathrm{E}_\pi(G_t\,|\,s_t = s) = \mathrm{E}_\pi\left(\sum_{k=0}^{\infty} \gamma^k r_{t+k+1}\,|\,s_t = s \right) = \mathrm{E}_\pi(r_{t+1}\,|\,s_t = s) + \gamma \mathrm{E}_\pi\left(\sum_{k=0}^{\infty} \gamma^k r_{t+k+2}\,|\,s_t = s \right).$$

განვიხილოთ მიღებულ შესაკრებთა წყვილი ცალ-ცალკე. დავიწყოთ პირველით. გარკვეულ s მდგომარეობაში ყოფნისას, ყოველი მოქმედების არჩევა ხდება $\pi(a\,|\,s)$ ალბათობით, გადასვლას s' მდგომარეობაში ადგილი აქვს $\mathcal{P}^a_{ss'}$ ალბათობით და ამასთან «გაიცემა» მოსალოდნელი $\mathcal{R}^a_{ss'}$ პრემია (ჯილდო). მაშინ

$$E_\pi(r_{t+1}\,|\,s_t = s) = \sum_{a\in A(s)} \sum_{s'\in S} \pi(a\,|\,s) \mathcal{R}^a_{ss'} \mathcal{P}^a_{ss'}.$$

მსგავსად ამისა, თუ ჩავანაცვლებთ $s_t = s$ ტოლობას $s_{t+1} = s'$ ტოლობით, მივიღებთ :

$$E_\pi\left(\sum_{k=0}^{\infty} \gamma^k r_{t+k+2}\,|\,s_t = s \right) = \sum_{a\in A(s)} \sum_{s'\in S} \pi(a\,|\,s) \mathrm{E}_\pi\left(\sum_{k=0}^{\infty} \gamma^k r_{t+k+2}\,|\,s_{t+1} = s' \right) \mathcal{P}^a_{ss'}.$$

შეკრებით და გადაჯგუფებით გვექნება:

$$v_\pi(s) = \sum_{a\in A(s)} \pi(a\,|\,s) \sum_{s'\in S} \mathcal{P}^a_{ss'} \left(\mathcal{R}^a_{ss'} + \gamma \mathrm{E}_\pi\left(\sum_{k=0}^{\infty} \gamma^k r_{t+k+2}\,|\,s_{t+1} = s' \right) \right).$$

რადგან $v_\pi(s') = \mathrm{E}_\pi\left(\sum_{k=0}^{\infty} \gamma^k r_{t+k+2}\,|\,s_{t+1} = s' \right)$, ამიტომ ვიღებთ :

$$v_\pi(s) = \sum_{a\in A(s)} \pi(a\,|\,s) \sum_{s'\in S} \mathcal{P}^a_{ss'} \left(\mathcal{R}^a_{ss'} + \gamma v_\pi(s') \right).$$

2. ანალოგიურად მტკიცდება თანაფარდობა $q_\pi(s,a)$ მახასიათებლისათვის. განსაზღვრების თანახმად, თუ გამოვიყენებთ მათემატიკური ლოდინის წრფივობის თვისებას, შემდეგ ტოლობას მივიღებთ :

$$\left.\begin{aligned} q_\pi(s,a) = \mathrm{E}_\pi(G_t\,|\,s_t = s, a_t = a) = \mathrm{E}_\pi\left(\sum_{k=0}^{\infty} \gamma^k r_{t+k+1}\,|\,s_t = s, a_t = a \right) \\ = \mathrm{E}_\pi(r_{t+1}\,|\,s_t = s, a_t = a) + \gamma \mathrm{E}_\pi\left(\sum_{k=0}^{\infty} \gamma^k r_{t+k+2}\,|\,s_t = s, a_t = a \right) \end{aligned}\right\}.$$

ვინაიდან a მოქმედება არჩეულია, ამიტომ, ჯერ ერთი,

$$\mathrm{E}_\pi\left(r_{t+1} \mid s_t = s, a_t = a\right) = \sum_{s' \in S} \mathcal{P}_{ss'}^a \mathcal{R}_{ss'}^a ,$$

მეორეც და,

$$\left.\begin{aligned} \mathrm{E}_\pi\left(\sum_{k=0}^{\infty} \gamma^k r_{t+k+2} \mid s_t = s, a_t = a\right) &= \sum_{s' \in S} \mathcal{P}_{ss'}^a \mathrm{E}_\pi\left(\sum_{k=0}^{\infty} \gamma^k r_{t+k+2} \mid s_{t+1} = s', a_t = a\right) \\ &= \sum_{s' \in S} \mathcal{P}_{ss'}^a q_\pi\left(s', a\right) = \sum_{s' \in S} \mathcal{P}_{ss'}^a \sum_{a' \in A(s')} \pi\left(a' \mid s'\right) q_\pi\left(s', a'\right) \end{aligned}\right\}.$$

შედეგად,

$$q_\pi\left(s, a\right) = \sum_{s' \in S} \mathcal{P}_{ss'}^a \left(\mathcal{R}_{ss'}^a + \gamma \sum_{a' \in A(s')} \pi\left(a' \mid s'\right) q_\pi\left(s', a'\right)\right) = \sum_{s' \in S} \mathcal{P}_{ss'}^a \left(\mathcal{R}_{ss'}^a + \gamma v\left(s'\right)\right).$$

□

ცხადია, რომ, თუ გვინდა მდგომარეობათა და მოქმედებათა ღირებულებების ფუნქციათა მიღება, პრობლემა დადის $|S|$ განტოლებათა სისტემამდე $|S|$ უცნობით, როგორც ერთ შემთხვევაში, ასევე მეორეშიც.

შენიშვნა 2.2.2 *მიღებულ განტოლებებს უწოდებენ ბელმანის განტოლებებს ღირებულების ფუნქციებისათვის π სტრატეგიის გამოყენებისას. მდგომარეობათა ღირებულების ფუნქციის შემთხვევაში ბელმანის განტოლება ასახავს გარკვეული მდგომარეობის ღირებულების დამოკიდებულებას მომდევნო მდგომარეობათა ღირებულებებისგან. მოქმედებათა ღირებულების ფუნქციათა შემთხვევაში — წყვილის (მდგომარეობა, მოქმედება) ღირებულების დამოკიდებულებას მომდევნო ასეთი წყვილების ღირებულებებისგან.*

განტოლებათა იდეის (ისევე, როგორც გამოყენების ტექნიკის) გაგება შეიძლება ნახატიდან 15.11. დავუშვათ, რომ ჩვენთვის (სადღაციდან) ცნობილია ყველა მდგომარეობათა ღირებულებები, ხოლო მიმდინარე მდგომარეობად მოცემულ შემთხვევაში s_1 მდგომარეობა განიხილება, მაშინ

$$v_\pi\left(s\right) = \sum_{a \in A(s)} \pi\left(a \mid s\right) \sum_{s' \in S} \mathcal{P}_{ss'}^a \left(\mathcal{R}_{ss'}^a + \gamma v_\pi\left(s'\right)\right)$$

თანაფარდობის გამოყენებით $\gamma = 0.8$ მნიშვნელობისათვის, მივიღებთ :

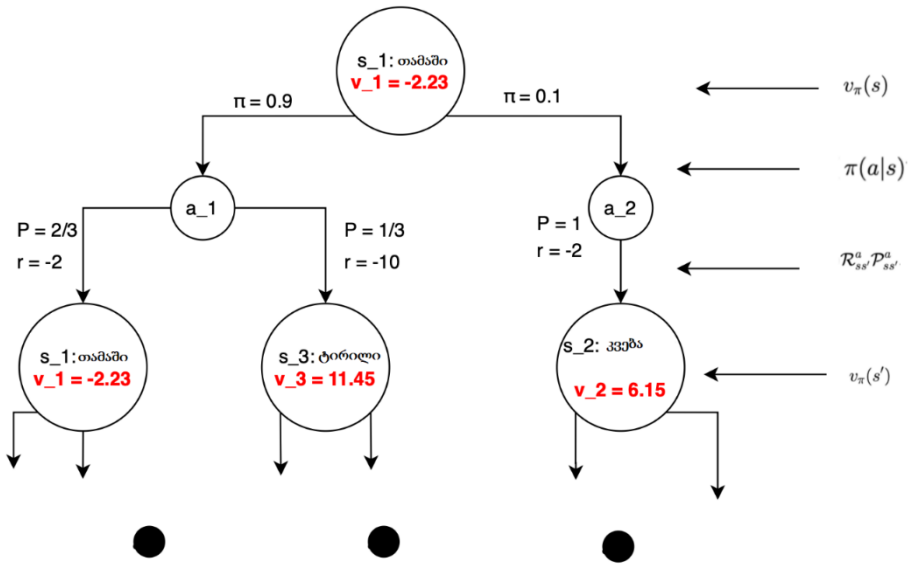
$$-2.23 = 0.9\left(\frac{1}{3}\left(-10 + 0.8 \cdot 11.45\right) + \frac{2}{3}\left(-2 + 0.8 \cdot \left(-2.23\right)\right)\right) + 0.1 \cdot 1 \cdot \left(-2 + 0.8 \cdot 6.15\right),$$

რაც სწორი ტოლობაა.

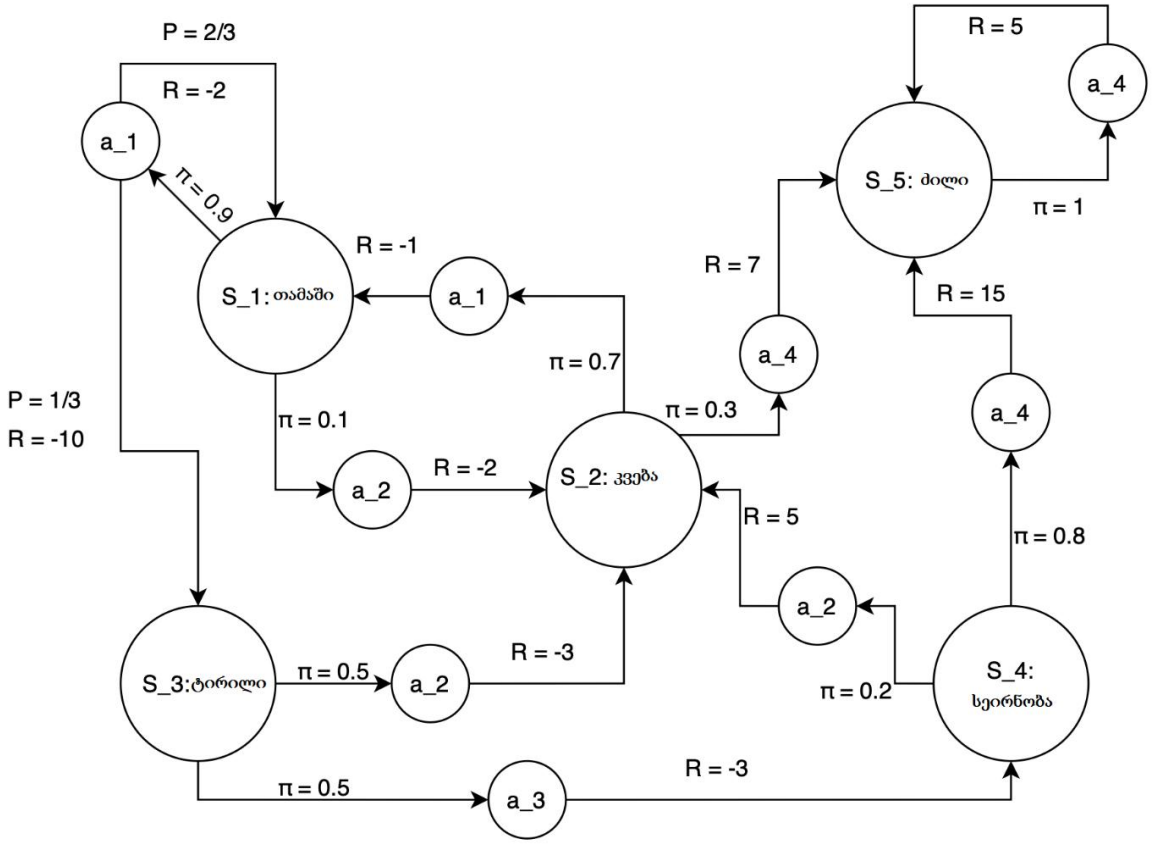
ცხადია, ძირითადი საკითხი, რომელიც ახლა წამოიჭრება, ასეთია : როგორ მივიღოთ მდგომარეობათა ღირებულებები ? განვიხილოთ ეს შემდეგ მაგალითზე.

2.3 მაგალითი 1

ადრე ჩვენ ვნახეთ, როგორ არის დაკავშირებული ერთმანეთთან ღირებულებები — მიმდინარე s მდგომარეობის და ნებისმიერი მომდევნო s' მდგომარეობის, რომელშიც შეიძლება გადავიდეს აგენტი. ახლა კი განვიხილოთ მაგალითზე, როგორ მივიღოთ ყველა მდგომარეობათა ღირებულებები ბელმანის განტოლების საშუალებით კონკრეტული π სტრატეგიის გამოყენებისას. მაგალითის საწყისი მონაცემები წარმოდგენილია ნახატზე 15.12 (სადაც ისინი ნასესხებია უკვე ნაცნობი ნახატიდან 15.9).



ნახატი 15.11 - გადაწყვეტილებათა მიღების მარკოვის პროცესი.



ნახატი 15.12 - საწყისი მონაცემები.

დისკონტირების კოეფიციენტის მნიშვნელობად ავიღოთ $\gamma = 0.8$ სიდიდე და მხედველობაში მივიღოთ, რომ

$$v_\pi(s) = \sum_{a \in A(s)} \pi(a \mid s) \sum_{s' \in S} \mathcal{P}_{ss'}^a \left(\mathcal{R}_{ss'}^a + \gamma v_\pi(s') \right),$$

რის შემდეგაც ჩავწეროთ ბელმანის განტოლება მდგომარეობისათვის «თამაში» :

$$v_\pi(s_1) = 0.9\left(\frac{1}{3}\left(-10 + 0.8 v_\pi(s_3)\right) + \frac{2}{3}\left(-2 + 0.8 v_\pi(s_1)\right)\right) + 0.1 \cdot 1 \cdot \left(-2 + 0.8 v_\pi(s_2)\right).$$

ვინაიდან s_1 მდგომარეობისგან განსხვავებულ ყველა სხვა მდგომარეობაში კონკრეტული მოქმედების განხორციელებისას გადასვლა მომდევნო მდგომარეობაში დეტერმინირებულია (ხორციელდება ალბათობით 1), ამიტომ შიდა ჯამი $s' -$ით შეიცავს $1-$ზე გამრავლებულ მხოლოდ ერთ შესაკრებს. ამ დაკვირვების გათვალისწინებით, ანალოგიურად ჩავწეროთ ყველა დანარჩენი განტოლება :

$$\left.\begin{aligned} v_\pi(s_2) &= 0.7(-1 + 0.8 v_\pi(s_1)) + 0.3(7 + 0.8 v_\pi(s_5)) \\ v_\pi(s_3) &= 0.5(-3 + 0.8 v_\pi(s_2)) + 0.5(-3 + 0.8 v_\pi(s_4)) \\ v_\pi(s_4) &= 0.2\left(5 + 0.8 v_\pi(s_2)\right) + 0.8\left(15 + 0.8 v_\pi(s_5)\right) \\ v_\pi(s_5) &= 5 + 0.8 v_\pi(s_5) \end{aligned}\right\}.$$

თუ ამ სისტემას ამოვხსნით, მივიღებთ მდგომარეობის ღირებულების (უახლოეს ასეულამდე დამრგვალებულ) შემდეგ მნიშვნელობებს (შედეგები ასევე წარმოდგენილია ნახატზე 15.13) :

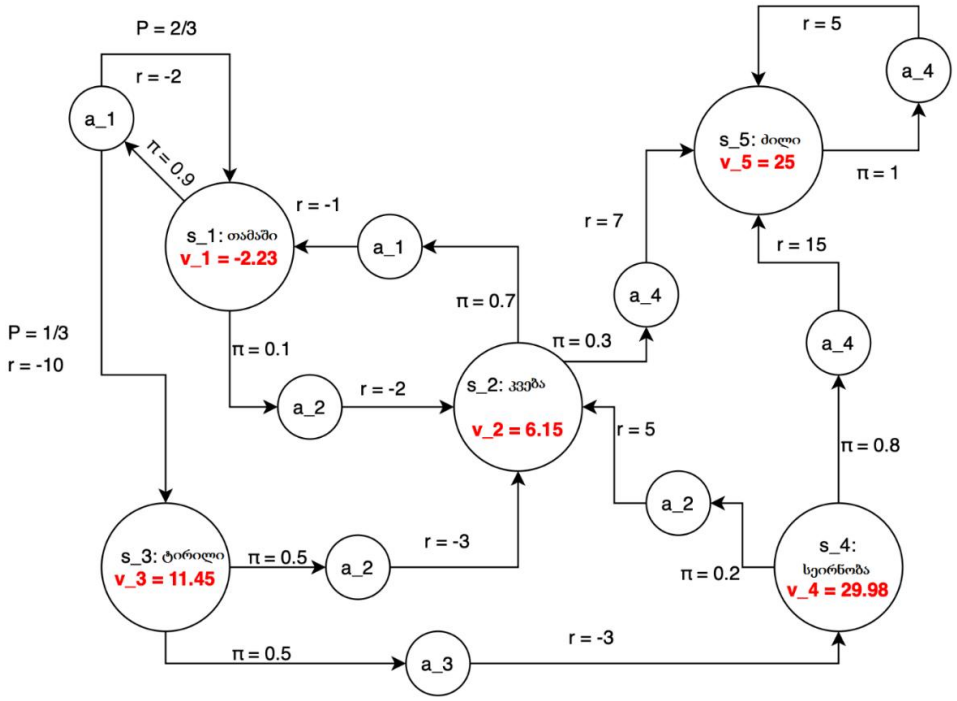
$$\left.\begin{aligned} v_\pi(s_1) &= -2.23 \\ v_\pi(s_2) &= 6.15 \\ v_\pi(s_3) &= 11.45 \\ v_\pi(s_4) &= 29.98 \\ v_\pi(s_5) &= 25. \end{aligned}\right\}.$$

შენიშვნა 2.3.1 *ცხადია, რომ მდგომარეობათა ღირებულებების მნიშვნელობები პირდაპირ არის დამოკიდებული აგენტის მიერ არჩეულ სტრატეგიაზე. სტრატეგიის შეცვლისას, საზოგადოდ, შეიცვლება მდგომარეობათა ღირებულებებიც. მაგალითად, თუ დავუშვებთ, რომ* $\pi(a_1 \mid s_1) = 0.5,\ \pi(a_2 \mid s_1) = 0.5$, *ხოლო დანარჩენი პარამეტრი უწინდელი დავტოვოთ, მაშინ მდგომარეობათა შემდეგ (უახლოეს ასეულამდე დამრგვალებულ) ღიღებულებებს მივიღებთ :*

$$\left.\begin{aligned} v_\pi(s_1) &= 2.59 \\ v_\pi(s_2) &= 8.85 \\ v_\pi(s_3) &= 12.72 \\ v_\pi(s_4) &= 30.42 \\ v_\pi(s_5) &= 25. \end{aligned}\right\}.$$

შეიძლება ის ფაქტიც აღვნიშნოთ, რომ ამ შემთხვევაში არც ერთი მდგომარეობის ღირებულება იმაზე ნაკლები არ აღმოჩენილა, ვიდრე შესაბამისი ღირებულებებია წინა სტრატეგიის გამოყენებისას. ოდნავ მოგვიანებით ჩვენ სპეციალურად დავუბრუნდებით მსგავსი

სიტუაციის უფრო დაწვრილებით განხილვას.



ნახატი 15.13 - მდგომარეობათა ღირებულებები არჩეული სტრატეგიის თანახმად.

2.4 სტრატეგიათა შეფასება

განხილულ მაგალითში ჩვენ გამოვიყენეთ ბელმანის განტოლება მდგომარეობათა ღირებულების დასადგენად. ხუთი მდგომარეობის შემთხვევისათვის წრფივ განტოლებათა სისტემის ფორმირება და ამოხსნა განსაკუთრებულ სიძნელეს არ წარმოადგენს. მაგრამ პრაქტიკაში აგენტის შესაძლო მდგომარეობათა რაოდენობა შეიძლება საკმაოდ დიდი იყოს. როგორც უკვე აღვნიშნავდით, თუ აგენტისათვის მისაწვდომ მდგომარეობათა რაოდენობა უდრის $n-$ს, მაშინ ჩნდება n უცნობიან n განტოლებათა სისტემის ამოხსნის ამოცანა.

ამიტომ, შესაძლებელია, დასაფიქრებელიც კი იყოს გამოთვლათა გამარტივების რაიმე ხერხის არსებობის საკითხი. თურმე, არსებობს.

მაგრამ ჯერ ...

ცოტა რამ მატრიცულ განტოლებათა და ნორმირებულ სივრცეთა შესახებ

აღმოჩნდა, რომ ბელმანის განტოლებათა სისტემის ამონახსნი შეიძლება ჩაიწეროს ცხადი სახით. გადავწეროთ

$$v_\pi(s) = \sum_{a \in A(s)} \pi(a \mid s) \sum_{s' \in S} \mathcal{P}^a_{ss'} \left(\mathcal{R}^a_{ss'} + \gamma v_\pi(s') \right), s \in S$$

განტოლებათა სისტემა შემდეგი მატრიცული აღნიშვნების გამოყენებით :

1. დავუშვათ, რომ v_π — $|S|$ სიმაღლის ვექტორ-სვეტია, რომლის $j-$ური ელემენტი $v_\pi(s_j),\ j \in \{1, 2, \ldots, |S|\}$ სიდიდეს უდრის.
2. დავუშვათ, რომ r_π — $|S|$ სიმაღლის ვექტორ-სვეტია, რომლის $j-$ური ელემენტი

$\mathrm{E}_\pi(r_{t+1} \mid s_t = s_j),\ j \in \{1, 2,\ ..., |S|\}$ სიდიდეს უდრის.

3. ასევე ჩავთვალოთ, რომ T_π — $|S| \times |S|$ ზომის მატრიცაა, ელემენტებით

$$\left(T_\pi\right)_{ij} = \mathrm{P}_\pi\left(s_{t+1} = s_j \mid s_t = s_i\right) = \sum_{a \in A(s_i)} \pi\left(a \mid s_i\right) \mathcal{P}^a_{s_i s_j}.$$

ასეთ აღნიშვნებში ბელმანის განტოლებათა სისტემა გადაიწერება შემდეგი სახით :

$$v_\pi = r_\pi + \gamma T_\pi v_\pi.$$

ცხადია, რომ მაშინ

$$\left(I - \gamma T_\pi\right) v_\pi = r_\pi \Rightarrow v_\pi = \left(I - \gamma T_\pi\right)^{-1} r_\pi\ .$$

მიღებული გამოსახულების გამოსაკვლევად, უპირველეს ყოვლისა, გავიხსენოთ ნორმირებული სივრცის ცნება.

განსაზღვრება 2.4.1 (ნორმირებული სივრცე) *დავუშვათ, რომ X — წრფივი სივრცეა. ნორმა ეწოდება $\|\cdot\| : X \to \mathbb{R}$ ფუნქციას, რომელიც შემდეგ აქსიომებს აკმაყოფილებს :*

1. $\forall x \in X \Rightarrow \|x\| \geq 0$, *თანაც* $\|x\| = 0 \Leftrightarrow x = 0$.
2. $\|\lambda x\| = |\lambda| \|x\|,\ \lambda \in \mathbb{R},\ x \in X$.
3. $\|x + y\| \leq \|x\| + \|y\|,\ \ x, y \in X$.

შენიშვნა 2.4.1 *გასაგებია, რომ ნორმა ზოგად შემთხვევაში განაზოგადებს სიგრძის ცნებას. თუ $X = \mathbb{R}$, მაშინ შეიძლება იმის დაშვება, რომ $\|x\| = |x|$.*

თუ, მაგალითად, $X = \mathbb{R}^n$, მაშინ $x = (x_1, x_2, ..., x_n) \in \mathbb{R}^n$ ვექტორის ნორმად შეიძლება განვიხილოთ ამ ვექტორის ჩვენთვის კარგად ნაცნობი სიგრძე :

$$\|x\| = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}\ .$$

შენიშვნა 2.4.2 *ნამდვილ რიცხვთა n განზომილების $\mathbb{R}^n$ სივრცეში შეიძლება სხვა ნორმის განხილვაც, მაგალითად, ასეთის :*

$$\|x\|_p = \sqrt[p]{|x_1|^p + |x_2|^p + \cdots + |x_n|^p},\ p \in \mathbb{N}.$$

ვექტორის ჩვენთვის კარგად ნაცნობი სიგრძე — ჩაწერილი ნორმის კერძო შემთხვევაა $p = 2$ ტოლობის პირობებში.

შემდგომ ჩვენ გამოვიყენებთ ნორმას, როცა $p = +\infty$:

$$\|x\|_\infty = \max_{i \in \{1, \ldots, n\}} |x_i|.$$

ახლა კი გადავიდეთ ოპერატორის ნორმის ცნებაზე.

განსაზღვრება 2.4.2 *დავუშვათ, რომ X — წრფივი ნორმირებული სივრცეა და $A : X \to X$ — წრფივი ოპერატორია. მაშინ A ოპერატორის ნორმა ეწოდება შემდეგ სიდიდეს :*

$$\|A\| = \sup_{x:\,\|x\|=1} \|Ax\|.$$

ამრიგად, ოპერატორის ნორმა გარკვეული თვალსაზრისით — ეს რიცხვია, რომელიც ასახავს ერთეულოვანი სიგრძის ვექტორის A ოპერატორის მოქმედებით გამოწვეული «გაჭიმვის» მაქსიმალურ მნიშვნელობას.

უნდა აღნიშნოს, რომ სასრული განზომილების სივრცეში ოპერატორის ნორმაც ყოველთვის სასრულია.

შენიშვნა 2.4.3 *ასევე უნდა აღინიშნოს, რომ ნორმის განსაზღვრებიდან უმალვე გამომდინარეობს შემდეგი უტოლობა :*

$$\|Ax\| \le \|A\|\|x\|,\ \forall x \in X.$$

შენიშვნა 2.4.4 *კარგად არის ცნობილი, რომ ნებისმიერი $A : X \to X$ წრფივი ოპერატორი შეიძლება იყოს გაიგივებული თავის მატრიცასთან (იმ შემთხვევაში, როცა X – ში ფიქსირებულია რომელიმე ბაზისი).*

ამ დროს ოპერატორთა ჯამს და ნამრავლს შეესაბამება სათანადო მატრიცათა ჯამი და ნამრავლი, ხოლო შებრუნებულ ოპერატორს — შებრუნებული მატრიცა. შედეგად, ოპერატორის შებრუნებადობა მისი მატრიცის შებრუნებადობის ტოლფასია გარკვეულ ბაზისში.

მაგალითი 2.4.1 *ვიპოვოთ $\|x\|_\infty$ ნორმის გამოყენებისას $A : \mathbb{R}^n \to \mathbb{R}^n$ ოპერატორის ნორმა. დავუშვათ, რომ ფიქსირებულ ბაზისში ოპერატორთან შესაბამისობაშია A მატრიცა a_{ij} ელემენტებით, სადაც $i, j \in \{1, 2, ..., n\}$. მაშინ,*

$$\|Ax\|_\infty = \max_{i\in\{1,\ldots,n\}} \left|\sum_{j=1}^{n} a_{ij}x_j\right| \le \max_{i\in\{1,\ldots,n\}} \sum_{j=1}^{n} |a_{ij}|\,\|x\|_\infty .$$

ასე რომ, $x \neq 0$ შემთხვევაში, $y = \dfrac{x}{\|x\|_\infty}$ ნოტაციის შემოღებით გვექნება :

$$\|Ay\| \le \max_{i\in\{1,\ldots,n\}} \sum_{j=1}^{n} |a_{ij}|,\ \|y\|_\infty = 1.$$

თუ ჩვენ დავამტკიცებთ, რომ ჩაწერილი შეფასება მიიღწევა, მაშინ დამტკიცებული აღმოჩნდება შემდეგი თანაფარდობაც :

$$\|A\| = \max_{i\in\{1,\ldots,n\}} \sum_{j=1}^{n} |a_{ij}|.$$

დავუშვათ, რომ მატრიცის თუნდაც ერთი ელემენტი არ უდრის ნულს (წინააღმდეგ შემთხვევაში მისი ნორმა არის 0 და ტოლობა, ცხადია, მიიღწევა). ამასთან ერთად ასევე დავუშვათ, რომ

$$i_0 = \arg\max_{i\in\{1,\dots,n\}} \sum_{j=1}^{n} \left|a_{ij}\right|$$

— რომელიღაც ისეთი i ცვლადია, რომელზეც მაქსიმუმი მიიღწევა. მაშინ x – ის როლში ავიღოთ $x = \left(\operatorname{sign} a_{i_0 1}, \operatorname{sign} a_{i_0 2}, \dots, \operatorname{sign} a_{i_0 n}\right)$ ვექტორი. ცხადია, რომ $\|x\|_\infty = 1$ და

$$Ax = \sum_{j=1}^{n} a_{i_0 j} \operatorname{sign} a_{i_0 j} = \sum_{j=1}^{n} \left|a_{i_0 j}\right| = \max_{i\in\{1,\dots,n\}} \sum_{j=1}^{n} \left|a_{i_0 j}\right|.$$

ლემა 2.4.1 (ოპერატორის შებრუნებადობის შესახებ) *იმისათვის რომ წრფივი $A: X \to X$ ოპერატორი, სადაც X — ნორმირებული სივრცეა $\|\cdot\|$ ნორმით, შებრუნებადი იყოს, აუცილებელი და საკმარისია მოიძებნოს ისეთი $m > 0$, რომ*

$$\forall x \in X \Rightarrow \|Ax\| \geq m\|x\|.$$

მტკიცებულება. ჯერ დავამტკიცოთ **საკმარისობა**. ჩაწერილი უტოლობის თანახმად, $x_1 \neq 0$ პირობიდან გამომდინარეობს $Ax_1 \neq 0$ შედეგი :

$$x_1 \neq 0 \Rightarrow Ax_1 \neq 0.$$

ოპერატორის წრფივობის შემთხვევაში სწორედ ეს ნიშნავს შებრუნებადობას.

ახლა კი დავამტკიცოთ **აუცილებლობა**. დავუშვათ, რომ ოპერატორი შებრუნებადია, მაშინ :

$$y = Ax,\ x = A^{-1}y \text{ და } \left\|A^{-1}y\right\| \leq \left\|A^{-1}\right\|\|y\| \Rightarrow \|x\| \leq \left\|A^{-1}\right\|\|Ax\| \Rightarrow \|Ax\| \geq \left\|A^{-1}\right\|^{-1}\|x\|,$$

სადაც $m = \left\|A^{-1}\right\|^{-1}$.

□

ლემა 2.4.2 (სისტემის ამონახსნის არსებობის შესახებ) *დავუშვათ. რომ $\gamma \in (0,1)$. $(I - \gamma T_\pi)$ მატრიცა შექცევადია, მაშასადამე, ბელმანის განტოლებათა სისტემის ამონახსნი არსებობს და იგი ერთადერთია, ამასთან*

$$v_\pi = (I - \gamma T_\pi)^{-1} r_\pi.$$

მტკიცებულება. მატრიცის შექცევადობა (შებრუნებადობა) იმ ოპერატორის შექცევადობის (შებრუნებადობის) ტოლფასია, რომელიც მოცემულია ფიქსირებულ ბაზისში $(I - \gamma T_\pi)$ მატრიცით. ვინაიდან $\|I\| = 1$ და $\|T_\pi\| = 1$ (რაკი იხილება $R^{|S|}$ სივრცე $\|x\|_\infty$ ნორმით), ამიტომ

$$\|(I - \gamma T_\pi)x\| \geq \left|\|Ix\| - \gamma\|T_\pi x\|\right| \geq (1-\gamma)\|x\|.$$

ამრიგად, ოპერატორი შექცევადია და ლემა დამტკიცებულია.

□

შემოვიტანოთ შემდეგი მნიშვნელოვანი განსაზღვრება.

განსაზღვრება 2.4.3 *დავუშვათ, რომ* $A: X \to X$ *არის წრფივი ოპერატორი, ხოლო* X — *ნორმირებული სივრცეა.* A *ოპერატორს შეკუმშვა ეწოდება, თუ* $\exists \alpha \in (0,1)$, *რომ*

$$\|Ax - Ay\| \le \alpha \|x - y\| \ .$$

გავიხსენოთ ამ განსაზღვრებაში გამოყენებული $\exists$ კვანტორის არსი : არსებობს $(0,1)$ ინტერვალზე, ანუ ღია $(0,1)$ შუალედში მოცემული α პარამეტრის სულ ცოტა ერთი მნიშვნელობა მაინც, ისეთი რომ $\|Ax - Ay\| \le \alpha \|x - y\|$ გამოსახულება სწორია.

ფაქტობრივად შეკუმშვა ამცირებს მანძილს ასახვებს შორის.

განსაზღვრება 2.4.4 *დავუშვათ, რომ* $A: X \to X$ — *წრფივი ოპერატორია. მაშინ* $x^* \in X$ *წერტილს, რომლისთვისაც*

$$Ax^* = x^*,$$

A *ოპერატორის უძრავი წერტილი ეწოდება.*

ოპერატორის უძრავი წერტილები — ეს ისეთი წერტილებია, რომლებიც საკუთარ თავში გადადის ოპერატორის მოქმედების შედეგად.

თეორემა 2.4.1 (შემკუმშავ ასახვათა უმარტივესი მეთოდი) *დავუშვათ, რომ* $A: \mathbb{R}^n \to \mathbb{R}^n$ — *წრფივი ოპერატორია და წარმოადგენს შეკუმშვას* α *პარამეტრით, ასევე დავუშვათ, რომ* $\|\cdot\|$ — *ნებისმიერო ნორმაა* $\mathbb{R}^n$ *სივრცეში. მაშინ* A *ოპერატორს აქვს ერთადერთი უძრავი* x^* *წერტილი, თანაც*

$$x_{k+1} = Ax_k, \quad x_0 \in \mathbb{R}^n$$

მიმდევრობა, როცა $k \to +\infty$, *იკრიბება* x^* *წერტილზე. გარდა ამისა,*

$$\|x^* - x_k\| \le \frac{\alpha^k}{1-\alpha} \|x_1 - x_0\|.$$

მტკიცებულება. ჯერ დავრწმუნდეთ, რომ x_k მიმდევრობა *ფუნდამენტურია.* აქედან კი გაკეთდება დასკვნა მისი კრებადობის შესახებ (შემოღებული ნორმისაგან დამოუკიდებლად).

გავიხსენოთ, რომ ფუნდამენტური მიმდევრობა, ანუ საკუთარ თავში კრებადი მიმდევრობა (ფრანგულ მათემატიკურ წყაროებში გვხვდება კოშის მიმდევრობის სახელწოდებაც) — მეტრიკული სივრცის წერტილების მიმდევრობაა ისეთი, რომ ნებისმიერი არანულოვანი მოცემული მანძილისთვის არსებობს მიმდევრობის ელემენტი, რომლიდანაც დაწყებული მიმდევრობის ყველა ელემენტი დაშორებულია ერთმანეთისგან მოცემულზე ნაკლები მანძილით.

სამკუთხედის უტოლობის გამოყენების შედეგად, გვექნება :

$$\|x_{k+p} - x_k\| \le \|x_{k+p} - x_{k+p-1}\| + \|x_{k+p-1} - x_{k+p-2}\| + \cdots + \|x_{k+1} - x_k\|.$$

ვისარგებლოთ x_k მიმდევრობის განსაზღვრებით, მაშინ მივიღებთ, რომ

$$\|x_{k+p} - x_k\| \le \|Ax_{k+p-1} - Ax_{k+p-2}\| + \cdots + \|Ax_k - Ax_{k-1}\| \le$$
$$\left(\alpha^{k+p-1} + \cdots + \alpha^k\right)\|x_1 - x_0\| = \frac{\alpha^k\left(1-\alpha^p\right)}{1-\alpha}\|x_1 - x_0\| \le \frac{\alpha^k}{1-\alpha}\|x_1 - x_0\|,$$

რაც ამტკიცებს კიდეც ფუნდამენტურობას. საიდანაც ვასკვნით, რომ მიმდევრობა კრებადია. დავუშვათ, რომ $p \to +\infty$, მაშინ

$$\|x^* - x_k\| \le \frac{\alpha^k}{1-\alpha}\|x_1 - x_0\|.$$

დავამტკიცოთ ერთადერთობა. დავუშვათ, რომ x^* და y^* — ორი უძრავი წერტილია. მაშინ

$$\|x^* - y^*\| = \|Ax^* - Ay^*\| \le \alpha\|x^* - y^*\|.$$

აქედან კი, $\|x^* - y^*\| = 0$ და $x^* = y^*$.

სტრატეგიათა იტერაციული შეფასება

ამრიგად, ჩვენ უკვე მზად ვართ ჩამოვაყალიბოთ თეორემა სტრატეგიის იტერაციული შეფასების შესახებ.

თეორემა 2.4.2 (სტრატეგიათა იტერაციული შეფასების შესახებ) *დავუშვათ, რომ $v_\pi(s)$ არის $s \in S$ მდგომარეობის ღირებულების ფუნქცია. ამასთან ერთად ასევე მივიჩნიოთ, რომ $v_0(s) \in \mathbb{R}$ — ნებისმიერო რიცხვია, თუ s მდგომარეობა არ არის ტერმინალური და $v_0(s) = 0$ — წინააღმდეგ შემთხვევაში.*

მაშინ

$$v_{k+1}(s) = \sum_{a \in A(s)} \pi(a \mid s) \sum_{s' \in S} \mathcal{P}^a_{ss'}\left(\mathcal{R}^a_{ss'} + \gamma v_k(s')\right)$$

მიმდევრობა იკრიბება $v_\pi(s)$ ღირებულების ფუნქციისკენ, როცა $k \to +\infty$.

მტკიცებულება. *განვიხილოთ ეგრეთ წოდებული ბელმანის ოპერატორი*

$$Bv = r_\pi + \gamma T_\pi v.$$

დავამტკიცოთ, რომ ეს ოპერატორი არის შეკუმშვა $\|\cdot\|_\infty$ ნორმის განხილვისას. ვინაიდან ამ შემთხვევაში $\|T_\pi\| = 1$, ამიტომ

$$\|Bv - Bv'\|_\infty = \|\gamma T_\pi(v - v')\| \le \gamma\|v - v'\|.$$

□

დაგვრჩა შემკუმშავ ასახვათა ზემოთ დამტკიცებული მეთოდის გამოყენება.

შენიშვნა 2.4.5 *შეიძლება მივიღოთ ცდომილების შეფასება ჭეშმარიტი ღირებულების ჩანაცვლებისას მისი მიახლოებით :*

$$\left|v_\pi(s)-v_k(s)\right| \le \frac{\gamma^k}{1-\lambda}\max_{s\in S}\left|v_1(s)-v_0(s)\right|.$$

აქედან კი გამომდინარეობს, რომ მდგომარეობათა ღირებულების შესაფასებლად გარკვეული სტრატეგიის გამოყენებისას შეიძლება შემდეგი ალგორითმით მოქმედება :

ალგორითმი 1. სტრატეგიის იტერაციული შეფასება

1: შესაფასებელი სტრატეგიის არჩევა

2: ღირებულებათა $v_0(s)=0$ ინიციალიზაცია ყველა $s\in S$ მდგომარეობისათვის

3: $k=0$ ინიციალიზაცია

4: **შეასრულეთ**

5: $\Delta \leftarrow 0$

6: **ამისთვის** $s\in S$ **შეასრულეთ**

7: $v^* \leftarrow v_k(s)$

8: $v_{k+1}(s) \leftarrow \sum_{a\in A(s)} \pi(a\mid s)\sum_{s'\in S}\mathcal{P}^a_{ss'}\left(\mathcal{R}^a_{ss'}+\gamma v_k(s')\right)$

9: $\Delta \leftarrow \max(\Delta, \left|v^* - v_{k+1}(s)\right|)$

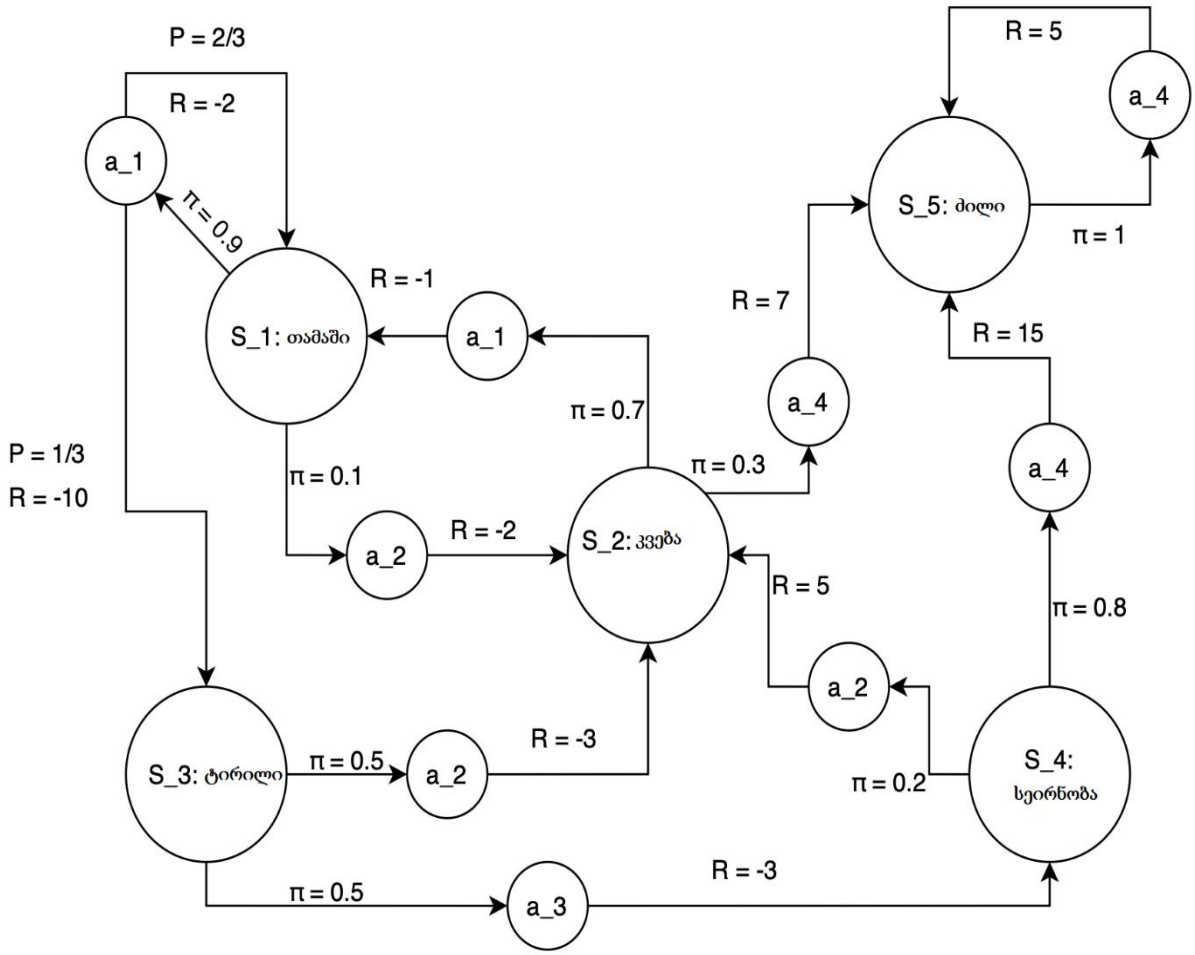
10: **ციკლის დასასრული**

11: $k \leftarrow k+1$

12: **იმეორეთ ვიდრე** $\Delta<\theta$ (მცირე დადებითი რიცხვი)

13: **დაიბრუნეთ** v_k

გამოვიყენოთ აღწერილი ალგორითმი უკვე ნაცნობი მაგალითისათვის. გადაწყვეტილებათა მიღების მარკოვის პროცესის (გმმპ) სქემა მოცემულია ნახატზე 15.14.



ნახატი 15.14 - საწყისი მონაცემები.

პირველ ეტაპზე მდგომარეობათა ღირებულებების ყველა მნიშვნელობა მიჩნეული იყო ნულის ტოლად.

შემდეგ, ყოველ ბიჯზე ღირებულებათა მნიშვნელობების განახლება ხდებოდა წინად მიღებული ბელმანის განტოლებების შესაბამისად.

სახელდობრ :

$$\left.\begin{aligned}
v_{k+1}(s_1) &= 0.9\left(\frac{1}{3}\left(-10+0,8v_k(s_3)\right)+\frac{2}{3}\left(-2+0.8v_k(s_1)\right)\right)+0.1\cdot 1\cdot\left(-2+0.8v_k(s_2)\right)\\
v_{k+1}(s_2) &= 0.7(-1+0.8v_k(s_1))+0.3(7+0.8v_k(s_5))\\
v_{k+1}(s_3) &= 0.5(-3+0.8v_k(s_2))+0.5(-3+0.8v_k(s_4))\\
v_{k+1}(s_4) &= 0.2\left(5+0.8v_k(s_2)\right)+0.8\left(15+0.8v_k(s_5)\right)\\
v_{k+1}(s_5) &= 5+0.8v_k(s_5)
\end{aligned}\right\}.$$

უახლოეს ასეულამდე დამრგვალებულ გამოთვლათა შედეგები წარმოდგენილია ცხრილში ქვემოთ :

ბიჯი	$v_k(s_1)$	$v_k(s_2)$	$v_k(s_3)$	$v_k(s_4)$	$v_k(s_5)$

0	0	0	0	0	0
1	–4.4	1.4	–3	13	5
2	–7.12	0.14	2.76	16.42	9
3	–7.14	–0.43	3.62	18.78	12.2
. . .	. . .	. . .	. . .	. . .	. . .
36	–2.24	6.14	11.45	29.98	24.99
37	–2.23	6.15	11.45	29.98	24.99
38	–2.23	6.15	11.45	29.98	24.99
39	–2.23	6.15	11.45	29.98	25

შევნიშნავთ, რომ მოთხოვნილი საჭირო სიზუსტის (ასეულამდე თანხვედრით) მისაღებად აუცილებელი გახდა იტერაციათა საკმაოდ დიდი რაოდენობა.

ცდომილება არ აღემატება შემდეგ მნიშვნელობას :

$$13 \cdot \frac{0.8^{39}}{1-0.8} \approx 0.011.$$

2.5 ოპტიმალური სტრატეგიები

განმტკიცებით სწავლების გადმოსახედიდან, ჩვენთვის არა იმდენად მდგომარეობებია საინტერესო. რამდენადაც სტრატეგიები, რომლებიც უზრუნველყოფს მაქსიმალურ დაყვანილ სარგებელს. თუ ბელმანის განტოლების ამონახსნი ნაპოვნი იქნება თითოეული სტრატეგიისათვის, შესაძლებელია საუკეთესოს არჩევა, ანუ ეგრეთ წოდებული ოპტიმალური სტრატეგიის. რას ნიშნავს ეს სიტყვები ? ჯერ შემოვიღოთ განსაზღვრებები :

$$v^*(s) = \max_{\pi} v_\pi(s), \quad \forall s \in S,$$

თანაფარდობას მდგომარეობათა ღირებულების ოპტიმალური ფუნქცია ეწოდება.

განსაზღვრება 2.5.2 *ფუნქციას*

$$q^*(s,a) = \max_{\pi} q_\pi(s,a), \quad \forall s \in S, \forall a \in A(s),$$

მოქმედებათა ღირებულების ოპტიმალური ფუნქცია ეწოდება.

ამრიგად, შემოღებული ფუნქციები ასახავს, რამდენად მაქსიმალური შეიძლება იყოს მდგომარეობის და (ან) მოქმედების ღირებულება.

შენიშვნა 2.5.1 *ვაჩვენოთ, რომ ზემოთ მოცემული მაქსიმუმები გადაწყვეტილებათა მიღების ფინიტური მარკოვის პროცესის შემთხვევაში არსებობს. მაგალითად, რატომ არის $v^*(s)$ კორექტულად განსაზღვრული ფუნქცია ? ბელმანის განტოლების თანახმად,*

$$\left.\begin{aligned} v_\pi(s) &= \sum_{a\in A(s)} \pi(a\mid s)\sum_{s'\in S} \mathcal{P}^a_{ss'}\left(\mathcal{R}^a_{ss'} + \gamma v_\pi(s')\right) \\ &\le \sum_{a\in A(s)} \pi(a\mid s)\max_{a\in A(s)}\left(\sum_{s'\in S} \mathcal{P}^a_{ss'}\left(\mathcal{R}^a_{ss'} + \gamma v_\pi(s')\right)\right) \\ &= \max_{a\in A(s)}\left(\sum_{s'\in S} \mathcal{P}^a_{ss'}\left(\mathcal{R}^a_{ss'} + \gamma v_\pi(s')\right)\right) = \max_{a\in A(s)} q_\pi(s,a) \end{aligned}\right\}.$$

ვინაიდან $A(s)$ სიმრავლე სასრულია, ამიტომ ცხადია, რომ სტრატეგია, რომელზეც მიიღწევა მარჯვნივ მიღებული გამოსახულება შემდეგი სახისაა (იმ შემთხვევაში, როცა მაქსიმუმი მიიღწევა ერთ მოქმედებაზე) :

$$\pi(a\mid s) = \begin{cases} 1, & a = \arg\max\limits_{a\in A(s)} q_\pi(s,a) \\ 0, & \text{წინააღმდეგ შემთხვევაში} \end{cases}.$$

თუ მაქსიმუმი მიიღწევა რამდენიმე მოქმედებაზე, მაშინ შესაძლებელია ნებისმიერის არჩევა მათ შორის. ამრიგად, ჩვენი ფუნქცია შეფასებულია ზემოდან და ნაპოვნია სტრატეგია, რომელზეც ეს შეფასება მიიღწევა.

მაშასადამე, სწორედ ეს არის საძიებელი მაქსიმუმი.

ასეთივე მსჯელობათა ჩატარება სასარგებლოა სტრატეგიათა ტერმინებში.

განსაზღვრება 2.5.3 *ამბობენ, რომ π სტრატეგია π' სტრატეგიაზე უარესი არ არის და წერენ*

$$\pi \ge \pi'$$

უტოლობას, თუ ყველა $s \in S$ მდგომარეობისათვის

$$v_\pi(s) \ge v_{\pi'}(s).$$

შენიშვნა 2.5.2 *ხშირად ამბობენ, რომ შემოტანილი $\ge$ თანაფარდობა ინდუცირებს (ან ამყარებს) ნაწილობრივ წესრიგს ყველა π სტრატეგიათა სიმრავლეზე. ყურადღება მიაქციეთ, შეიძლება ისე მოხდეს, რომ რომელიღაც ორი — π_1 და π_2 — სტრატეგია არაშესადარისია.*

ეს უბრალოდ იმას ნიშნავს, რომ მოიძებნება ისეთი $s_1, s_2 \in S$, რომ

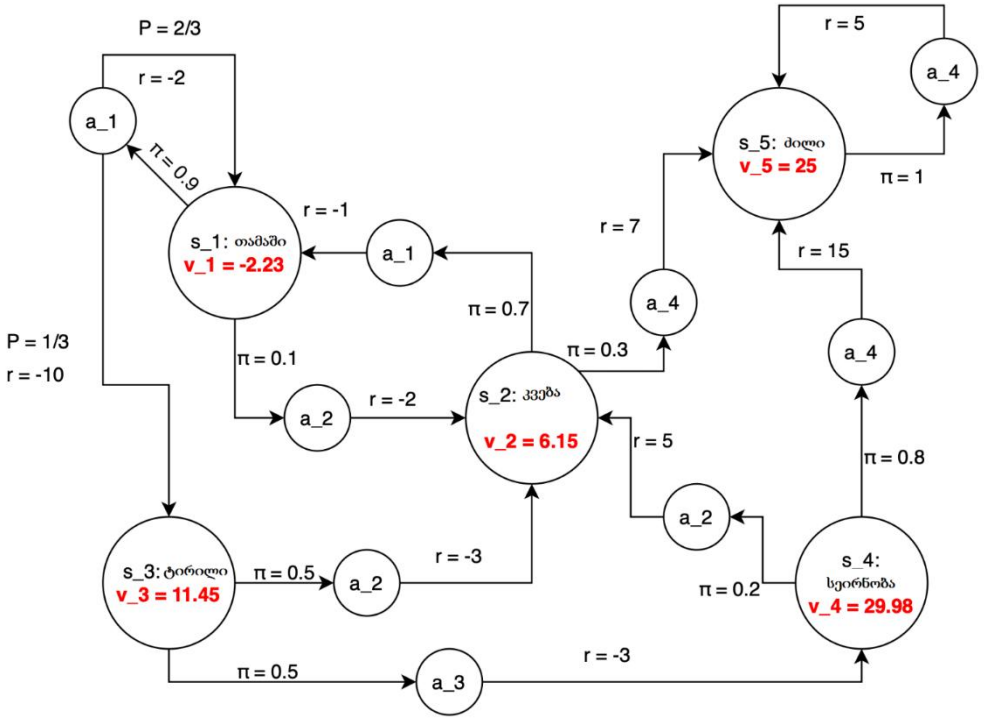
$$v_{\pi_1}(s_1) \le v_{\pi_2}(s_1),$$

მაგრამ

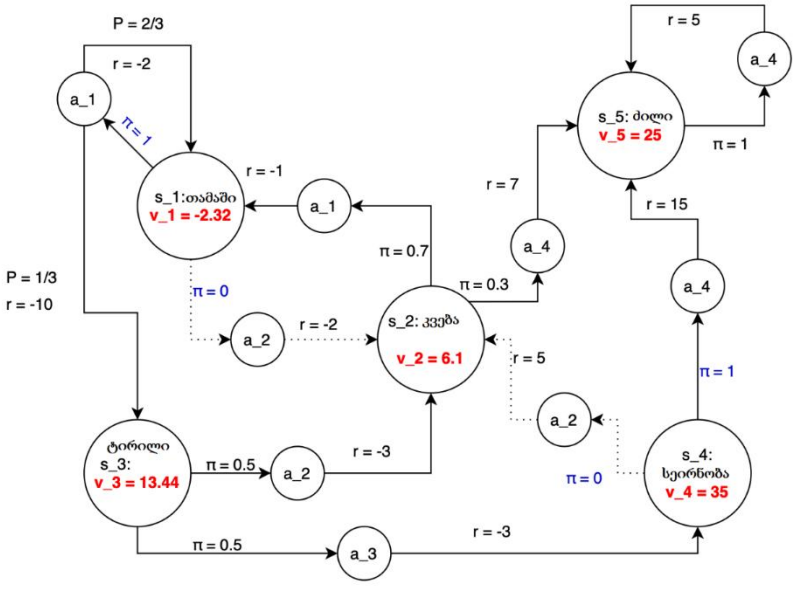
$$v_{\pi_1}(s_2) \ge v_{\pi_2}(s_2).$$

მაგალითი 2.5.1 *თუ უკვე განხილულ მაგალითში ორიგინალური სტრატეგიის (ნახატი 15.15ა) ნაცვლად π' სტრატეგიას გამოვიყენებთ, სადაც $\pi'(a_1 \mid s_1) = 1$, $\pi'(a_2 \mid s_1) = 0$, $\pi'(a_4 \mid s_4) = 1$, $\pi'(a_2 \mid s_4) = 0$ (ნახატი 15.15ბ), მაშინ მივიღებთ მდგომარეობათა შემდეგ ღირებულებებს :*

სტრატეგია	$v(s_1)$	$v(s_2)$	$v(s_3)$	$v(s_4)$	$v(s_5)$
π	–2.23	6.15	11.45	29.98	25
π'	–2.32	6.1	13.44	35	25



ა)



ბ)

ნახატი 15.15 - არაშესადარის სტრატეგიათა მაგალითი.

ცხრილიდან ადვილი შესამჩნევია, რომ s_1 და s_2 მდგომარეობებისათვის ღირებულების უფრო

მაღალი მნიშვნელობები მიიღება π სტრატეგიის გამოყენებისას, ხოლო s_3 და s_4 მდგომარეობებისათვის სიტუაცია საპირისპიროა.

ცხადია, სასურველია ისეთი სტრატეგიის პოვნა, რომელიც ყველა დანარჩენზე უარესი არ იქნება.

განსაზღვრება 2.5.4 *π^* სტრატეგიას ოპტიმალური ეწოდება, თუ ნებისმიერი π სტრატეგიისათვის შესრულებულია შემდეგი უტოლობა :*

$$\pi^* \geq \pi.$$

შენიშვნა 2.5.3 *რა თქმა უნდა, შემოღებული განსაზღვრება ერთბაშად მრავალ შეკითხვას იწვევს : ყოველთვის თუ არსებობს ოპტიმალური სტრატეგია და თუ იქნება ოპტიმალური სტრატეგია ერთადერთი ? თურმე, ზოგად შემთხვევაში პასუხი ორივე შეკითხვაზე არის ცალსახა : «არა». ჩვენ დავუბრუნდებით ამას ოდნავ მოგვიანებით.*

იმავდროულად, როცა გვაქვს ოპტიმალური π^* სტრატეგია, ლოგიკურია განხილვაში $v_{\pi^*}(s)$ და $q_{\pi^*}(s,a)$ ფუნქციების შემოტანაც, რომლებიც არის კანდიდატები მდგომარეობათა და მოქმედებათა ღირებულების ოპტიმალური ფუნქციის როლზე, შესაბამისად. მოდით გამოვარკვიოთ მათი კავშირი. ჯერ ერთი, დავადგინოთ, რომ $v_{\pi^*}(s)$ და $q_{\pi^*}(s,a)$ ფუნქციები ოპტიმალური სტრატეგიის არსებობისას განსაზღვრულია კორექტულად (ინგლ. *well-defined function*), ესე იგი, სხვადასხვა ოპტიმალურ სტრატეგიაზე ისინი ერთსა და იმავე მნიშვნელობას იძლევა.

ლემა 2.5.1 *ნებისმიერო ორი ოპტიმალური π_1^* და π_2^* სტრატეგიისათვის*

$$v_{\pi_1^*} \equiv v_{\pi_2^*} \text{ და } q_{\pi_1^*} \equiv q_{\pi_2^*}.$$

მტკიცებულება. დავამტკიცოთ, მაგალითად, პირველი იგივეობა. რადგან π_1^* — ოპტიმალური სტრატეგიაა, ამიტომ, განსაზღვრების თანახმად,

$$v_{\pi_1^*}(s) \geq v_\pi(s), \quad \forall s \in S,$$

კერძოდ, როცა $\pi = \pi_2^*$, აქედან გამომდინარეობს, რომ $v_{\pi_1^*}(s) \geq v_{\pi_2^*}(s)$. თუ π_1^* და π_2^* სიდიდეებს შევუცვლით ადგილებს, მივიღებთ $v_{\pi_2^*}(s) \geq v_{\pi_1^*}(s)$, უტოლობას, რითაც მტკიცებულებაც სრულდება.

□

ახლა კი ვუპასუხოთ შეკითხვაზე : ყოველთვის არსებობს ოპტიმალური სტრატეგია ? თურმე, გადაწყვეტილებათა მიღების მარკოვის პროცესის *ფინიტურობის* შემთხვევაში, პასუხი დადებითია.

გარდა ამისა სამართლიანია შემდეგი (და შეიძლება ითქვას, სავსებით მოსალოდნელი) თეორემა.

თეორემა 2.5.1 *დავუშვათ, რომ გადაწყვეტილებათა მიღების მარკოვის პროცესი ფინიტურია.*

მაშინ :

1. *არსებობს ოპტიმალური π^* სტრატეგია ;*
2. *ნებისმიერი ოპტიმალური სტრატეგიისათვის $v_{\pi^*} \equiv v^*$;*
3. *ნებისმიერი ოპტიმალური სტრატეგიისათვის $q_{\pi^*} \equiv q^*$.*

ამრიგად, თურმე, მდგომარეობათა და მოქმედებათა ღირებულების ოპტიმალური ფუნქციები — ეს მდგომარეობათა და მოქმედებათა ღირებულების შესაბამისი ფუნქციებია იმ შემთხვევისათვის, როცა გამოიყენება ოპტიმალური სტრატეგია.

მტკიცებულება. მტკიცებულება გამომდინარეობს იქედან, რაც იყო ნაჩვენები ადრე. ოპტიმალური π^* სტრატეგიის როლზე გამოდგება, მაგალითად, შემდეგი სტრატეგია (თუ მაქსიმუმი მიიღწევა მხოლოდ ერთ მოქმედებაზე) :

$$\pi(a \mid s)\begin{cases}1, & a = \arg\max\limits_{a \in A(s)} q_\pi(a) \\ 0, & \text{წინააღმდეგ შემთხვევაში}\end{cases}.$$

თუ მაქსიმუმი რამდენიმე მოქმედებაზე მიიღწევა, მაშინ მათ შორის ნებისმიერის არჩევა შეიძლება. შემდგომი მსჯელობები ცხადია.

□

ანალოგიურად იმისა, როგორც იყო ნაპოვნი ბელმანის განტოლებები $v_\pi(s)$ და $q_\pi(s,a)$ მახასიათებლებისათვის, ახლა დავადგენთ ასეთ განტოლებებს მდგომარეობათა და მოქმედებათა ღირებულების $v^*(s)$ და $q^*(s,a)$ ოპტიმალური ფუნქციებისათვისაც.

ლემა 2.5.2 *დავუშვათ, რომ $v^*(s)$ და $q^*(s,a)$ — მდგომარეობათა და მოქმედებათა ღირებულების ოპტიმალური ფუნქციებია შესაბამისად, მაშინ ადგილი აქვს შემდეგ თანაფარდობებს :*

$$v^*(s) = \max_{a \in A(s)} \sum_{s' \in S} \mathcal{P}^a_{ss'} \left(\mathcal{R}^a_{ss'} + \gamma v^*(s') \right),$$

$$q^*(s,a) = \sum_{s' \in S} \mathcal{P}^a_{ss'} \left(\mathcal{R}^a_{ss'} + \gamma \max_{a' \in A(s')} q^*(s',a') \right).$$

მტკიცებულება. როგორც გავარკვიეთ, ღირებულების ოპტიმალური ფუნქციები — ეს ღირებულების ფუნქციებია ოპტიმალურ π^* სტრატეგიაზე, ესე იგი, თუ მაქსიმუმი მიიღწევა ერთ მოქმედებაზე, სტრატეგიაზე

$$\pi(a \mid s) = \begin{cases}1, & a = \arg\max\limits_{a \in A(s)} q_\pi(a) \\ 0, & \text{წინააღმდეგ შემთხვევაში}\end{cases}.$$

ვინაიდან ზოგად შემთხვევაში

$$v_\pi(s) = \sum_{a \in A(s)} \pi(a \mid s) \sum_{s' \in S} \mathcal{P}^a_{ss'} \left(\mathcal{R}^a_{ss'} + \gamma v_\pi(s') \right),$$

ამიტომ ოპტიმალური π^* სტრატეგიისათვის

$$v^*(s) = \max_{a \in A(s)} \sum_{s' \in S} \mathcal{P}_{ss'}^a \left(\mathcal{R}_{ss'}^a + \gamma v^*(s') \right).$$

მსგავსი ხერხით მტკიცდება მეორე თანაფარდობაც.

□

შენიშვნა 2.5.4 *მნიშნელოვანია გვესმოდეს, რომ მდგომარეობათა ღირებულების ოპტიმალური ფუნქციის ცოდნა უმალვე გვაძლევს ოპტიმალურ სტრატეგიასაც : საჭიროა იმ მდგომარეობაში გადასვლა, რომლის ღირებულებაც მეტია.*

შენიშვნა 2.5.5 *შენიშვნის სახით ხაზს გავუსვამთ იმ გარემოებას, რომ მიღებულ განტოლებებს ბელმანის ოპტიმალურობის განტოლებებს უწოდებენ მდგომარეობათა და მოქმედებათა ღირებულების მიმართ, შესაბამისად. ბელმანის განტოლებებისგან განსხვავებით, მიღებული სისტემა უკვე არ არის წრფივ ალგებრულ განტოლებათა სისტემა, ამიტომ მისი ამოხსნა — ცალკე პრობლემაა.*

საბედნიეროდ, თურმე, ადრე განხილული იტერაციული მიდგომა ამ შემთხვევაშიც გვეხმარება.

თეორემა 2.5.2 *დავუშვათ, რომ $v^*(s)$ არის $s \in S$ მდგომარეობის ღირებულების ოპტიმალური ფუნქცია. ასევე დავუშვათ, რომ $v_0(s) \in \mathbb{R}$ — ნებისმიერო რიცხვია, თუ s მდგომარეობა არ არის ტერმინალური და $v_0(s) = 0$ — წინააღმდეგ შემთხვევაში. მაშინ მიმდევრობა*

$$v_{k+1}(s) = \max_{a \in A(s)} \sum_{s' \in S} \mathcal{P}_{ss'}^a \left(\mathcal{R}_{ss'}^a + \gamma v_k(s') \right)$$

იკრიბება $v^*(s)$ – ისაკენ, როცა $k \to +\infty$.

მტკიცებულება. განვიხილოთ ბელმანის ოპტიმალური B^* ოპერატორი, რომელიც შემდეგი წესით მოქმედებს :

$$B^* v = \max_{a \in A} \sum_{s' \in S} \mathcal{P}_{ss'}^a \left(\mathcal{R}_{ss'}^a + \gamma v(s') \right),$$

სადაც

$$v = \left(v(s_1), v(s_2), \ldots v\left(s_{|S|}\right), \right)^T.$$

განვიხილოთ $B^* v$ და $B^* v'$ მახასიათებლები. გადაწყვეტილებათა მიღების მარკოვის მეთოდის ფინიტურობის გამო, მოიპოვება a და a' მოქმედებები (აუცილებელი არ არის სხვადასხვა), ისეთი რომ

$$B^* v = \sum_{s' \in S} \mathcal{P}_{ss'}^a \left(\mathcal{R}_{ss'}^a + \gamma v(s) \right),$$

$$B^* v' = \sum_{s' \in S} \mathcal{P}_{ss'}^{a'} \left(\mathcal{R}_{ss'}^{a'} + \gamma v'(s') \right).$$

გარდა ამისა, $B^*v' \geq \sum_{s' \in S} \mathcal{P}^a_{ss'} \left(\mathcal{R}^a_{ss'} + \gamma v'(s') \right)$, საიდანაც

$$B^*v - B^*v' \leq \sum_{s' \in S} \mathcal{P}^a_{ss'} \left(\mathcal{R}^a_{ss'} + \gamma v(s') \right) - \sum_{s' \in S} \mathcal{P}^a_{ss'} \left(\mathcal{R}^a_{ss'} + \gamma v'(s') \right)$$
$$= \gamma \sum_{s' \in S} P^a_{ss'} \left(v(s') - v'(s') \right) \leq \gamma \left\| v - v \right\|_\infty .$$

თუ v და v' სიდიდეებს ადგილებს შევუცვლით, მივიღებთ რომ

$$\left| B^*v - B^*v' \right| \leq \gamma \left\| v - v' \right\|_\infty \Rightarrow \left\| B^*v - B^*v' \right\|_\infty \leq \gamma \left\| v - v' \right\|_\infty .$$

ამის შემდეგ ის დაგვრჩენია, რომ გამოვიყენო შემკუმშავ ასახვათა უმარტივესი მეთოდი.

□

ადრე გაკეთებულის ანალოგიურად, შეიძლება ჩამოვაყალიბოთ მდგომარეობათა ოპტიმალური ფუნქციების იტერაციული ძებნის ალგორითმი.

ალგორითმი 2. ოპტიმალური სტრატეგიის იტერაციული შეფასება

1: შესაფასებელი სტრატეგიის არჩევა

2: ღირებულებათა $v_0(s) = 0$ ინიციალიზაცია ყველა $s \in S$ მდგომარეობისათვის

3: $k = 0$ ინიციალიზაცია

4: **შეასრულეთ**

5: $\Delta \leftarrow 0$

6: $s \in S$ მდგომარეობი-**სათვის შეასრულეთ**

7: $v^* \leftarrow v_k(s)$

8: $v_{k+1}(s) \leftarrow \max_{a \in A(s)} \sum_{s' \in S} \mathcal{P}^a_{ss'} \left(\mathcal{R}^a_{ss} + \gamma v_k(s') \right)$

9: $\Delta \leftarrow max\ (\Delta, \left| v^* - v_{k+1}(s) \right|)$

10: **ციკლის დასასრული**

11: $k \leftarrow k + 1$

12: **იმეორეთ ვიდრე** $\Delta < \theta$ (მცირე დადებითი რიცხვი)

13: **დაიბრუნეთ** v_k

შენიშვნა 2.5.6 *მსგავსად იმისა, როგორც ეს წინათ იყო გაკეთებული, ალგორითმის ყოველ ბიჯზე შეიძლება შეფასდეს ცდომილება ოპტიმალური სტრატეგიის ჭეშმარიტ და მიახლოებით მნიშვნელობას შორის.*

დავუბრუნდეთ ჩვენთვის უკვე ნაცნობ ერთი დღის მაგალითს ბავშვის ცხოვრებიდან (ნახატი 15.16) და გამოვარკვიოთ, რა მოქმედებათა განხორციელება ღირს აგენტისათვის დაყვანილი სარგებლის მაქსიმიზაციისათვის.

მდგომარეობათა ღირებულებების მნიშვნელობები ოპტიმალური სტრატეგიის გამოყენების შემთხვევისათვის წარმოდგენილია ნახატზე 15.17.

დავრწმუნდეთ იმაში, რომ მიღებული მნიშვნელობები ნამდვილად შეესაბამება ოპტიმალურ სტრატეგიას, სახელდობრ, შევამოწმოთ ყოველი მდგომარეობისათვის შემდეგი ტოლობის შესრულება :

$$v^*(s) = \max_{a \in A(s)} \sum_{s' \in S} \mathcal{P}_{ss'}^a \left(\mathcal{R}_{ss'}^a + \gamma v^*(s') \right).$$

$v^*(s_1)$ მახასიათებლისათვის გვაქვს :

$$19.6 = \max \left\{ \left(\frac{1}{3}(-10 + 0.8 \cdot 25) + \frac{2}{3}(-2 + 0.8 \cdot 19.6) \right), (-2 + 0.8 \cdot 27) \right\},$$

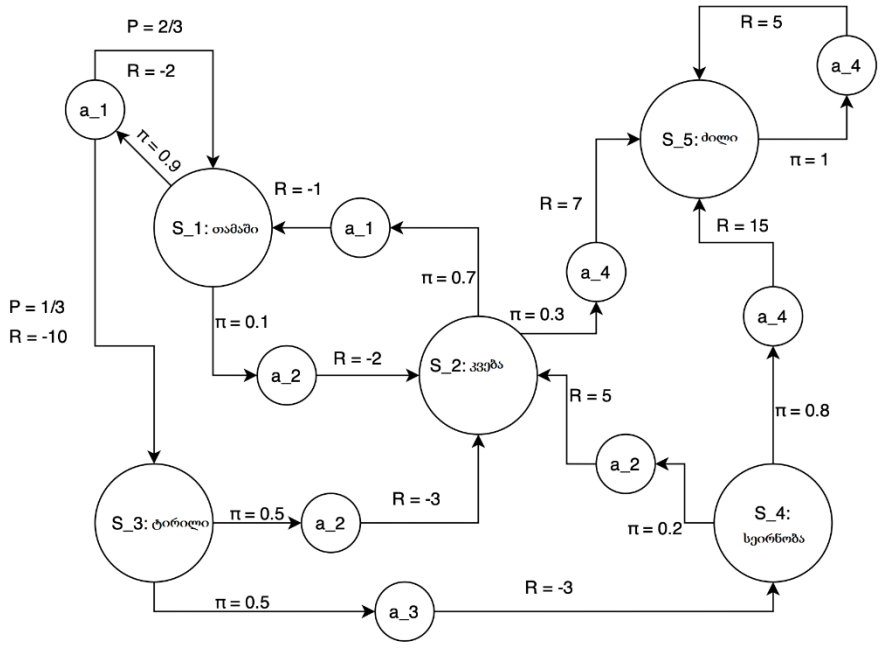
ესე იგი

$$19.6 = \max \{12.45, 19.6\},$$

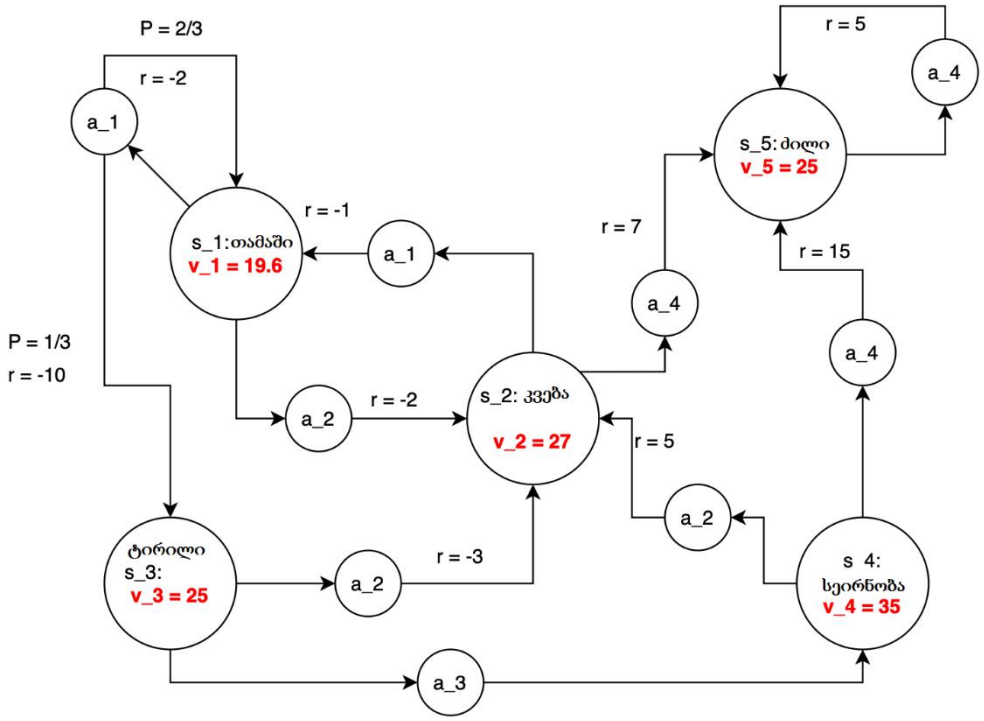
რაც სწორია. ამის მსგავსად, $v^*(s_2)$ მახასიათებლისათვის :

$$27 = \max \{(-1 + 0.8 \cdot 19.6),\ (7 + 0.8 \cdot 25)\} = \max \{14.68,\ 27\},$$

რაც ასევე სწორია და ასე შემდეგ.

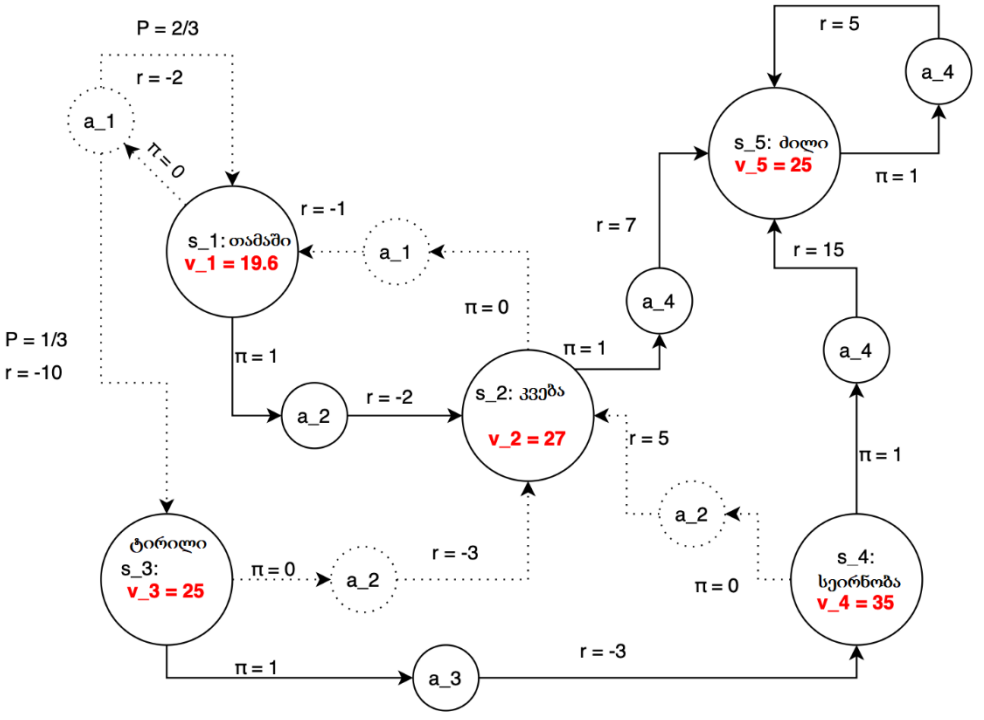


ნახატი 15.16 - საწყისი მონაცემები.



ნახატი 15.17 - მდგომარეობათა ღირებულებები ოპტიმალური სტრატეგიის გამოყენებისას.

უნდა შევნიშნოთ, რომ მდგომარეობათა ღირებულების ოპტიმალური ფუნქციის მნიშვნელობათა ცოდნისას, სტრატეგიის ფორმირება საკმაოდ ადვილად ხდება : უბრალოდ საჭიროა გადასვლა მდგომარეობაში ღირებულების მაქსიმალური მნიშვნელობით (ნახატი 15.18).



ნახატი 15.18 - ოპტიმალური სტრატეგია.

ნახეთ, ახლა ჩვენი სტრატეგია გახდა მთლიანად დეტერმინირებული : დროის ნებისმიერ მომენტში აგენტმა ზუსტად იცის, რა უნდა მოიმოქმედოს.

მაგალითად, «თამაშის» მდგომარეობაში ყოფნისას, საჭიროა საკვების მიღება და წასვლა დასაძინებლად, ხოლო ატირებისას — ჯერ სასეირნოდ წასვლა და შემდეგ დაძინება.

მნიშვნელოვანი რჩება საკითხი იმის შესახებ, როგორ მივიღოთ მდგომარეობათა ღირებულების საჭირო მნიშვნელობები. ზუსტი მნიშვნელობების მისაღებად შეიძლება ბელმანის ოპტიმალურობის განტოლებათა სისტემის ამოხსნა. ჩვენი მაგალითისათვის ამ სისტემას შემდეგი სახე ექნება :

$$\left.\begin{aligned}
v^*(s_1) &= \max\left\{\left(\frac{1}{3}\left(-10+0.8v^*(s_3)\right)+\frac{2}{3}\left(-2+0.8v^*(s_1)\right)\right),\left(-2+0.8v^*(s_2)\right)\right\}\\
v^*(s_2) &= \max\left\{\left(0.7\left(-1+0.8v^*(s_1)\right)\right),\left(0.3\left(7+0.8v^*(s_5)\right)\right)\right\}\\
v^*(s_3) &= \max\left\{\left(0.5\left(-3+0.8v^*(s_2)\right)\right),\left(0.5\left(-3+0.8v^*(s_4)\right)\right)\right\}\\
v^*(s_4) &= \max\left\{\left(0.2\left(5+0.8v^*(s_2)\right)\right),\left(0.8\left(15+0.8v^*(s_5)\right)\right)\right\}\\
v^*(s_5) &= 5+0.8v^*(s_5)
\end{aligned}\right|.$$

ცხადია, რომ ეს სისტემა წრფივი უკვე არ არის და, მაშასადამე, მისი ამონახსნის პოვნა ასე ადვილი არ იქნება. სწორედ აქ გამოგვადგება ადრე განხილული იტერაციული მეთოდი. დავუშვათ, რომ პირველ ბიჯზე ღირებულებათა ყველა მნიშვნელობა უდრის ნულს. გამოთვლათა შედეგები (ასეულამდე დამრგვალებით) წარმოდგენილია ქვემოთ ცხრილში.

ბიჯი	$v_k(s_1)$	$v_k(s_2)$	$v_k(s_3)$	$v_k(s_4)$	$v_k(s_5)$
0	0	0	0	0	0
1	–2	7	–3	15	5
2	3.6	11	9	19	9
3	6.8	14.2	12.2	22.2	12.2
. . .	. . .	. . .	. . .	. . .	. . .
38	19.59	26.99	24.99	34.99	24.99
39	19.6	27	25	35	25

უნდა აღინიშნოს, რომ სიზუსტის მისაღები დონის — ასეულებში თანხვედრის — მისაღწევად ჩვენს შემთხვევაში საჭირო გახდა ორმოცამდე იტერაცია, რაც სავსებით გამართლებულია.

2.6 SARSA

სწავლების ზემოთ განხილულ მეთოდებს საკმაოდ სერიოზული ნაკლოვანება აქვს. ხსენებულ მეთოდებში იგულისხმება, რომ აგენტს აქვს წარმოდგენა გარემოს კონფიგურაციის შესახებ. სხვა სიტყვებით რომ ვთქვათ, ბელმანის განტოლებაში

$$v_\pi(s)=\sum_{a\in A(s)}\pi(a\mid s)\sum_{s'\in S}\mathcal{P}_{ss}^{a}\left(\mathcal{R}_{ss'}^{a}+\gamma v_\pi(s')\right),$$

იგულისხმება, რომ ცნობილია $\mathcal{P}_{ss}^{a}$ განაწილებაც, ესე იგი s მდგომარეობიდან s' მდგომარეობაში გადასვლის ალბათობები a მოქმედების განხორციელებისას, ასევე ყველა

მოსალოდნელი პრემიაც (ჯილდო. ანაზღაურება), ესე იგი $\mathcal{R}_{ss'}^{a}$ მახასიათებელიც. რეალურად ეს თითქმის ყოველთვის შეუძლებელია. გამოსავალი ერთია — ზოგიერთი შეფასებები უნდა ხდებოდეს განხორციელებული მოქმედებებისა და მიღებული შედეგების საფუძველზე. იდეა ძალიან ჰგავს იმ იდეას, რომელიც ადრე იყო განხილული მრავალხელიან ბანდიტთან დაკავშირებით. გავიხსენოთ რეკურენტული ფორმულა გარკვეული მოქმედების ღირებულების დასადგენად ექსპონენციალური მოსრიალე საშუალოს საფუძველზე :

$$q_{t+1}(a) = q_t(a) + \alpha\left(r_{t+1} - q_t(a)\right),$$

სადაც $\alpha \in (0,1]$. ასეთი მიდგომა შეიძლება იყოს გამოყენებული მდგომარეობათა ღირებულების შესაფასებლადაც. მას *დროითი სხვაობა* (ინგლ. TD, Temporal Difference) ეწოდება. თუ გავითვალისწინებთ, რომ

$$v_\pi(s) = \mathrm{E}_\pi\left(G_t \mid s_t = s\right) = \mathrm{E}_\pi\left(\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid s_t = s\right) = \mathrm{E}_\pi\left(r_{t+1} + \gamma v_\pi(s') \mid s_t = s\right),$$

შეიძლება s_t მდგომარეობის ღირებულების შეფასების ხერხის მიღება :

$$v(s_t) \leftarrow v(s_t) + \alpha\left(r_{t+1} + \gamma v(s_{t+1}) - v(s_t)\right).$$

მოცემულ შემთხვევაში აგენტი იმყოფება s_t მდგომარეობაში და გარკვეული სტრატეგიით ხელმძღვანელობისას ასრულებს a_t მოქმედებას, იღებს r_{t+1} ჯილდოს და გადადის s_{t+1} მდგომარეობაში. ახალი s_{t+1} მდგომარეობის $v(s_{t+1})$ ღირებულება ასევე ცნობილია და, ამრიგად, შეიძლება s_t მდგომარეობის ღირებულების განახლება. სხვა სიტყვებით რომ ვთქვათ, მომდევნო ბიჯხე ხდება ღირებულების შეფასების მნიშვნელობის განახლება წინა ბიჯისათვის.

მდგომარეობათა ღირებულების შეფასება, მართალია, მნიშვნელოვანია, მაგრამ თვითმიზანს არ წარმოადგენს, რადგან ჯილდო მოაქვს მოქმედებებს, მათი არჩევა კი დამოკიდებულია სტრატეგიაზე. ამიტომ სწავლებისას აუცილებელია სტრატეგიის ოპტიმიზირებაც. ამ მიზნებისათვის მოქმედებათა ფუნქციის $q(s,a)$ შეფასების მისაღებად შეიძლება დროითი სხვაობის იდეის გამოყენება, სახელდობრ :

$$q(s_t, a_t) \leftarrow q(s_t, a_t) + \alpha\left(r_{t+1} + \gamma q(s_{t+1},\ a_{t+1}) - q(s_t,\ a_t)\right).$$

ასევე ადვილი შესამჩნევია, რომ სწავლების პროცესში მეორდება ერთი და იგივე თანამიმდევრობა : მდგომარეობა (s), მოქმედება (a), ჯილდო (r), მდგომარეობა (s), ... ესე იგი

$$s, a, r, s, a, r, s, a, r, \ \ldots,$$

ამიტომ ამ მეთოდმა *SARSA* დასახელება მიიღო. თუ ცნობილია თითოეული მოქმედების ღირებულების შეფასება, შესაძლებელია გამოვიყენოთ, მაგალითად, ხარბი სტრატეგია მორიგი მოქმედების ასარჩევად, სხვანაირად, სტრატეგია, რომელიც მოცემულია შემდეგი სახით :

$$A_t = \arg\max_{a \in A(s_t)} q_t(a), \quad \pi(a \mid s_t) = \begin{cases} \frac{1}{|A_t|}, & a \in A_t \\ 0, & \text{წინააღმდეგ შემთხვევაში} \end{cases},$$

თუმცა, მიუხედავად ამისა, ლოგიკურია, რომ გარკვეული ყურადღება ეთმობოდეს შესწავლასაც, კვლევას, დაზვერვას. აი ასეთი მიზნებისათვის უფრო მომგებიანია უწინ განხილული ε – ხარბი სტრატეგია :

$$\pi(a \mid s_t) = \begin{cases} \frac{1-\varepsilon}{|A_t|} + \frac{\varepsilon}{|A(s_t)|}, & a \in A_t \\ \frac{\varepsilon}{|A(s_t)|}, & a \notin A_t \end{cases}.$$

ქვემოთ მოცემულია *SARSA* ალგორითმის ფორმალური აღწერა.

ალგორითმი 3. ალგორითმი SARSA

1: დავუშვათ, რომ S არის ყველა მდგომარეობათა სიმრავლე, ხოლო $A(s), s \in S$ — მხოლოდ s მდგომარეობაში მისაწვდომ მოქმედებათა სიმრავლე

2: მოხდეს $q(s,a), s \in S$ მახასიათებლის ინიციალიზაცია, სადაც s არის არატერმინალური მდგომარეობა, ხოლო $a \in A(s)$ — ნებისმიერი მოქმედება

3: მოხდეს α და γ პარამეტრების ინიციალიზაცია

4: ყოველი თამაში-**სათვის** მეორდებოდეს **შეასრულეთ**

5: არატერმინალური s_0 მდგომარეობის ინიციალიზაცია შემთხვევითად

6: აირჩიეთ a_0 მოქმედება $\pi_0(a \mid s_0)$ სტრატეგიის თანახმად

7: $t \leftarrow 0$

8: თამაშის ყოველი t ბიჯი-**სათვის**, ვიდრე არ შესრულებულა გაჩერების კრიტერიუმი ან ვიდრე s_t — არატერმინალური მდგომარეობაა, **შეასრულეთ**

9: შეასრულეთ a_t, იპოვეთ r_{t+1}, გადადით s_{t+1}-ში

10: **თუ** s_{t+1} — ტერმინალური მდგომარეობაა, **მაშინ**

11: $q(s_t, a_t) \leftarrow q(s_t, a_t) + \alpha\left(r_{t+1} - q(s_t, a_t)\right)$

12: **წინააღმდეგ შემთხვევაში**

13: მოხდეს a_{t+1} მოქმედების არჩევა $\pi_{t+1}(a \mid s_{t+1})$ სტრატეგიის საფუძველზე

14: $q(s_t, a_t) \leftarrow q(s_t, a_t) + \alpha\left(r_{t+1} + \gamma q(s_{t+1}, a_{t+1}) - q(s_t, a_t)\right)$

15: **პირობის დასასრული**

16: $t \leftarrow (t+1)$

17: **ციკლის დასასრული**

18: **ციკლის დასასრული**

ამრიგად, ალგორითმის მიხედვით, პირველ ეტაპზე ნებისმიერად ხდება ყველა მისაწვდომი $a \in A(s)$ მოქმედების ღირებულების ინიციალიზაცია თითოეული არატერმინალური $s \in S$ მდგომარეობისათვის.

შემდეგ ტარდება თამაშების სერია. თითოეულ თამაშში აგენტის საწყისი არატერმინალური მდგომარეობის ინიციალიზაცია ხდება შემთხვევითად.

ამის მერე, ყოველი თამაშის ფარგლებში თითოეულ t ბიჯზე აგენტი s_t მდგომარეობაში ყოფნისას ირჩევს გარკვეულ a_t მოქმედებას არჩეული სტრატეგიის შესაბამისად, გადადის მომდევნო s_{t+1} მდგომარეობაში და იღებს r_{t+1} ჯილდოს.

თუ s_{t+1} მდგომარეობა — არატერმინალურია, მაშინ, სტრატეგიის თანახმად, ხდება a_{t+1} მოქმედების არჩევა, ხოლო $q(s_t, a_t)$ მახასიათებლის მნიშვნელობის განახლება ხორციელდება შემდეგი ფორმულით :

$$q(s_t, a_t) \leftarrow q(s_t, a_t) + \alpha\left(r_{t+1} + \gamma q(s_{t+1}, a_{t+1}) - q(s_t, a_t)\right).$$

ამის შემდეგ ხდება გადასვლა თამაშის ახალ იტერაციაზე.

თუ s_{t+1} მდგომარეობა ტერმინალურია, მაშინ $A(s_{t+1}) = \varnothing$ და

$$q(s_t, a_t) \leftarrow q(s_t, a_t) + \alpha\left(r_{t+1} - q(s_t, a_t)\right).$$

თამაში მთავრდება იმ შემთხვევაში, თუ გადასვლა მოხდა ტერმინალურ მდგომარეობაში ან შესრულდა გაჩერების კრიტერიუმი.

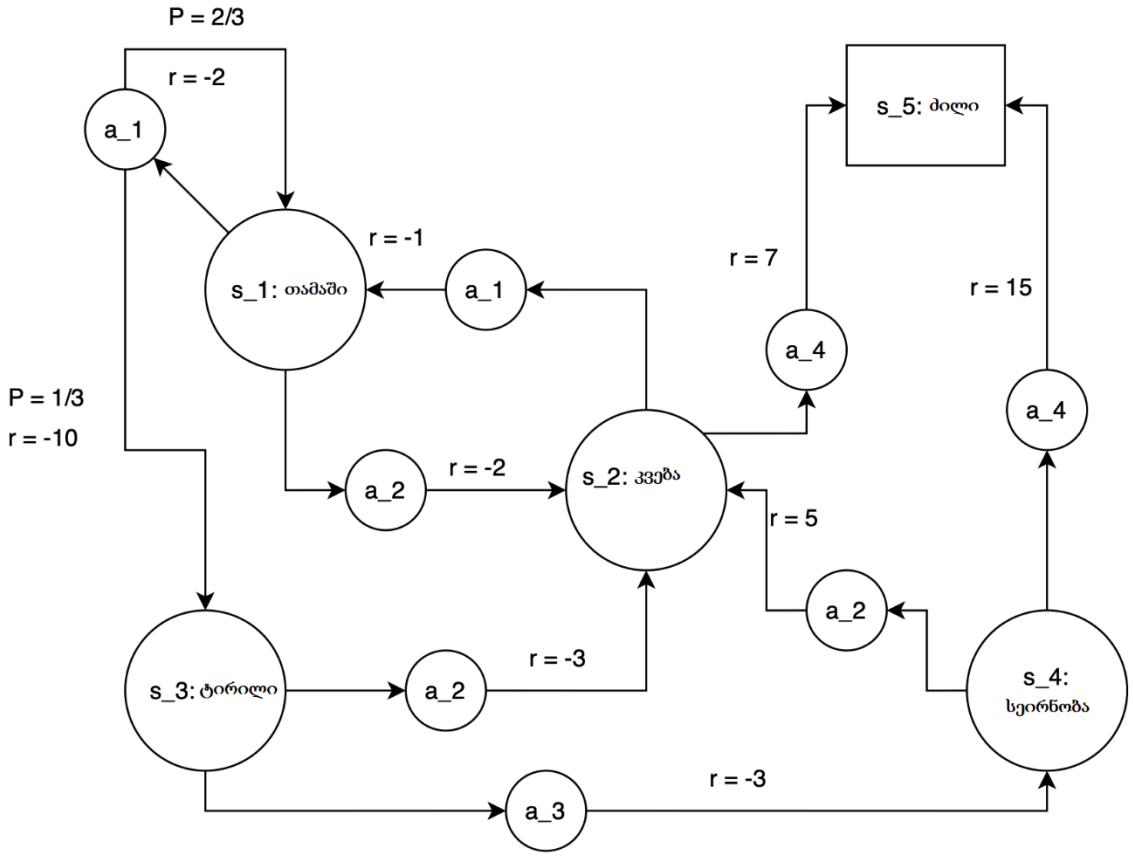
გაჩერების კრიტერიუმის შესრულების მაგალითებად შეიძლება გვესახებოდეს შემდეგი სიტუაციები :

- მიღწეულია მიღებულ ჯილდოთა ჯამის მოცემული დონე ;
- მიღწეულია მიღებულ ჯილდოთა საშუალო მნიშვნელობის მოცემული დონე ;
- მიღწეულია დახარჯულ ბიჯთა რიცხვის მოცემული დონე ;
- და მისთანანი.

განვიხილოთ *SARSA* ალგორითმის გამოყენება უკვე ნაცნობი, თუმცა ოდნავ მოდიფიცირებული მაგალითისათვის.

სახელდობრ, ჩავთვალოთ, რომ მდგომარეობა «ძილი» ტერმინალურია (ესე იგი, ამ მდგომარეობაში გადასვლისას თამაში დასრულებულია).

სქემა წარმოდგენილია ნახატზე 15.19.



ნახატი 15.19 - გარემოს აღწერა.

დავუშვათ, რომ $\gamma = 0.8, \alpha = 0.1$. სტრატეგიად გამოყენებული გვექნება ε – ხარბი სტრატეგია $\varepsilon = 0.1$ მნიშვნელობით. აღსანიშნავია ასევე, რომ ჩვენთვის ახლა უცნობია როგორც s მდგომარეობიდან s' მდგომარეობაში გადასვლის ალბათობები a მოქმედების შესრულების შედეგად, ასევე ამ შემთხვევაში მისაღები ჯილდოებიც. ყველა მნიშვნელობის შეფასება მოგვიხდება თამაშის მსვლელობის პროცესში. მოხერხებულობის მიზნით შევადგინოთ მოქმედებათა ღირებულებების მდგომარეობებზე დამოკიდებულების ცხრილი და განვახორციელოთ საწყისი ინიციალიზაცია (სახელდობრ, მივანიჭოთ ყველა შესაძლო ღირებულებას ნულის ტოლი მნიშვნელობა). ყურადღება მივაქციოთ იმ გარემოებასაც, რომ ჩვენ მიერ გამოყენებული × ნოტაცია მოქმედების არჩევის შეუძლებლობას უკავშირდება მიმდინარე მდგომარეობაში.

	a_1	a_2	a_3	a_4
s_1	0	0	×	×
s_2	0	×	×	0
s_3	×	0	0	×
s_4	×	0	×	0
s_5	×	×	×	×

დავუშვათ, რომ ჩვენ ვიწყებთ s_1 მდგომარეობაში. აქ ჩვენთვის მისაწვდომია a_1 და a_2 მოქმედებები და თანაც მოცემული მომენტისათვის $q(s_1, a_1) = q(s_1, a_2) = 0$. ჩვენი სტრატეგიის თანახმად, მოცემულ შემთხვევაში მისაწვდომ მოქმედებათა შორის ნებისმიერი მოქმედების

არჩევა მხოლოდ თავისუფლად ხდება.

დავუშვათ, რომ არჩეულია a_2 მოქმედება. მაშინ აგენტი გადადის s_2 მდგომარეობაში და იღებს $r=-2$ ჯილდოს. შემდეგ კი, s_2 მდგომარეობაში ყოფნისას, საჭიროა მომდევნო მოქმედების არჩევა. ჩვენთვის მისაწვდომია a_1 და a_4 მოქმედებები. ვინაიდან ღირებულებები ერთნაირია (უდრის ნულს), ვირჩევთ ნებისმიერს, მაგალითად, a_4 მოქმედებას.

ახლა შესაძლებელია $q(s_1,a_2)$ მახასიათებლის განახლება :

$$\left.\begin{array}{c} q(s_1,a_2)\leftarrow q(s_1,a_2)+\alpha\left(r+\gamma q(s_2,a_4)-q(s_1,a_2)\right) \\ =0+0.1\cdot(-2+0.8\cdot 0-0)=-0.2 \end{array}\right\}.$$

შევიტანოთ მიღებული მნიშვნელობა ცხრილში.

	a_1	a_2	a_3	a_4
s_1	0	–0.2	×	×
s_2	0	×	×	0
s_3	×	0	0	×
s_4	×	0	×	0
s_5	×	×	×	×

ვაგრძელებთ ალგორითმის შესრულებას :

მიმდინარე მდგომარეობა : s_2 ;

არჩეული მოქმედება : a_4 ;

შემდეგი მდგომარეობა : s_5 ;

ჯილდო : 7 .

უნდა აღინიშნოს, რომ, ვინაიდან s_5 — ტერმინალური მდგომარეობაა, ამიტომ თამაში დამთავრებულია.

მაშინ

$$q(s_2,a_4)\leftarrow 0+0.1\cdot(7-0)=0.7.$$

	a_1	a_2	a_3	a_4
s_1	0	–0.2	×	×
s_2	0	×	×	0.7
s_3	×	0	0	×
s_4	×	0	×	0
s_5	×	×	×	×

კიდევ რამდენიმე თამაშის ჩატარებისას, შემდეგ ცხრილს მივიღებთ :

	a_1	a_2	a_3	a_4
s_1	–0.32	0.88	×	×
s_2	–0.02	×	×	6.5
s_3	×	–0.09	–0.19	×
s_4	×	0.99	×	6.14
s_5	×	×	×	×

ამრიგად, შესრულებული მოქმედების შეფასებისას ჩვენ რეგულარულად გვიხდება მომავალში, თითქოს, ჩახედვა და შემდეგი მოქმედების წინასწარი არჩევა, რაც, საბოლოო ჯამში, გავლენას ახდენს მიმდინარე მოქმედების შეფასებაზე.

შენიშვნა 2.6.1 *უნდა აღინიშნოს, რომ უკანასკნელი ცხრილის საფუძველზე შეიძლება ოპტიმალური სტრატეგიის ჩამოყალიბება (მოქმედებათა ხარბად არჩევისას), რომელიც ემთხვევა ადრე ნაპოვნ თეორიულს (ნახატი 15.20).*

s_1 *მდგომარეობაში ყოფნისას, საჭიროა* a_2 *მოქმედების არჩევა, რადგან*

$$q(s_1, a_2) = 0.88 > q(s_1, a_1) = -0.32.$$

s_2 *მდგომარეობაში ყოფნისას კი საჭიროა* a_4 *მოქმედების არჩევა, ვინაიდან*

$$q(s_2, a_4) = 6.5 > q(s_2, a_1) = -0.02$$

და მისთანანი.

შენიშვნა 2.6.2 *ჩვენს მაგალითში თამაშის მსვლელობისას მიღებული ჯილდოები ერთისა და იმავე მნიშვნელობებისაა* s , s' *და* a *მახასიათებლებისგან დამოუკიდებლად.*

უფრო ზოგად შემთხვევაში კი ეს შეიძლება ასე არც იყოს.

2.7 Q – სწავლება

განვიხილოთ სწავლების კიდევ ერთი ალგორითმი *დროითი სხვაობის* (ინგლ. TD, Temporal Difference) საფუძველზე.

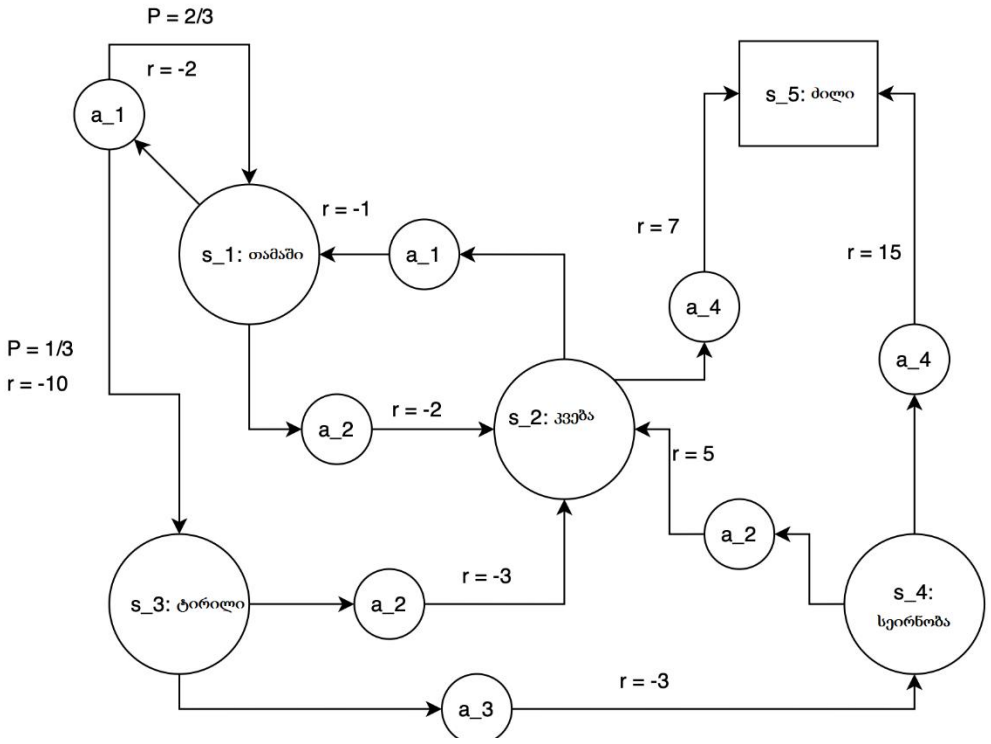
ადრე, ჩვენ მიერ, მიღებული იყო ოპტიმალურობის ბელმანის განტოლებები, კერძოდ, განტოლება მოქმედებათა ღირებულების ოპტიმალური ფუნქციისათვის.

მას შემდეგ სახე ჰქონდა :

$$q^*(s,a) = \sum_{s' \in S} \mathcal{P}^a_{ss'} \left(\mathcal{R}^a_{ss'} + \gamma \max_{a' \in A(s')} q^*(s', a') \right).$$

Q – სწავლების მეთოდის იდეა მდგომარეობს ღირებულების ოპტიმალური ფუნქციის შემდეგი სახით აპროქსიმაციაში :

$$q(s_t, a_t) \leftarrow q(s_t, a_t) + \left(r_{t+1} + \gamma \max_{a \in A(s_{t+1})} q(s_{t+1}, a) - q(s_t, a_t) \right).$$



ნახატი 15.20 - ოპტიმალური სტრატეგია.

ღირს, ალბათ, ერთი წამით მაინც დავფიქრდეთ და შევეკითხოთ საკუთარ თავს : რაში მდგომარეობს განსხვავება *SARSA* ალგორითმსა და $Q-$სწავლების მეთოდს შორის ? საქმე ისაა, რომ $q(s_t, a_t)$ მახასიათებლის განახლებისათვის *SARSA* ალგორითმში გამოიყენება $q(s_{t+1}, a_{t+1})$ მნიშვნელობა, ესე იგი s_{t+1} მდგომარეობაში — რომელიღაც დადგენილი $\pi_{t+1}(a \mid s_{t+1})$ სტრატეგიის შესაბამისად — არჩეული a_{t+1} მოქმედების ღირებულება. $Q-$სწავლებაში კი ოპტიმალური სტრატეგია გამოიყენება, ვინაიდან ის მოქმედება აირჩევა, რომლისთვისაც მიიღწევა $\max\limits_{a \in A(s_{t+1})} q(s_{t+1}, a)$. შედეგად, მოქმედების ღირებულების ფუნქციის მნიშვნელობის განახლებისას ჩვენ ვეყრდნობით არა არჩეულ სტრატეგიას (შესაძლოა, არაოპტიმალურსაც კი), არამედ ოპტიმალურს, მაგრამ მომდევნო ბიჯზე (sic!). თუ მიმდინარე სტრატეგია ოპტიმალურია, მაშინ ალგორითმები ერთნაირად მუშაობს. ჩავწეროთ $Q-$სწავლების ალგორითმი ფორმალურად.

ალგორითმი 4. $Q-$სწავლების ალგორითმი

1: დავუშვათ, რომ S არის ყველა მდგომარეობათა სიმრავლე, ხოლო $A(s), s \in S$ — მხოლოდ s მდგომარეობაში მისაწვდომ მოქმედებათა სიმრავლე

2: მოხდეს $q(s, a), s \in S$ მახასიათებლის ინიციალიზაცია, სადაც s არის არატერმინალური მდგომარეობა, ხოლო $a \in A(s)$ — ნებისმიერი მოქმედება

3: მოხდეს α და γ პარამეტრების ინიციალიზაცია

4: ყოველი თამაში-**სათვის** მეორდებოდეს **შეასრულეთ**

5: არატერმინალური s_0 მდგომარეობის ინიციალიზაცია შემთხვევითად

6: $t \leftarrow 0$

7: თამაშის ყოველი t ბიჯი-**სათვის**, ვიდრე არ შესრულებულა გაჩერების კრიტერიუმი ან ვიდრე s_t — არატერმინალური მდგომარეობაა, **შეასრულეთ**

8: აირჩიეთ a_t მოქმედება $\pi_t(a \mid s_t)$ სტრატეგიის თანახმად

9: შეასრულეთ a_t, იპოვეთ r_{t+1}, გადადით s_{t+1}-ში

10: **თუ** s_{t+1} — ტერმინალური მდგომარეობაა, **მაშინ**

11: $$q(s_t, a_t) \leftarrow q(s_t, a_t) + \alpha\left(r_{t+1} - q(s_t, a_t)\right)$$

12: **წინააღმდეგ შემთხვევაში**

13: $$q(s_t, a_t) \leftarrow q(s_t, a_t) + \alpha\left(r_{t+1} + \gamma \max_{a \in A(s_{t+1})} q(s_{t+1}, a) - q(s_t, a_t)\right)$$

14: **პირობის დასასრული**

15: $t \leftarrow (t+1)$

16: **ციკლის დასასრული**

17: **ციკლის დასასრული**

ჯერ ერთი, როგორც ეს იყო უკვე აღნიშნული, განსხვავება *SARSA* ალგორითმისგან მდგომარეობს მოქმედების ღირებულების შეფასების ხერხში. მეორეც და, განსხვავებულია მოქმედების არჩევა : *SARSA* ალგორითმში მიმდინარე იტერაციაზე იმყოფებოდა როგორც მიმდინარე, ასევე მომდევნო მოქმედება, მაშინ როცა $Q-$სწავლებაში მხოლოდ მიმდინარე მოქმედების არჩევა ხდება. თუ s_{t+1} მდგომარეობა — არატერმინალურია, მაშინ $q(s_t, a_t)$ მახასიათებლის განახლება შემდეგი ფორმულის თანახმად ხორციელდება :

$$q(s_t, a_t) \leftarrow q(s_t, a_t) \ + \alpha\left(r_{t+1} + \gamma \max_{a \in A(s_{t+1})} q(s_{t+1}, a) - q(s_t, a_t)\right).$$

შემდეგ თამაშის ახალ იტერაციაზე გადასვლა ხდება. თუ s_{t+1} მდგომარეობა ტერმინალურია, მაშინ $A(s_{t+1}) = \varnothing$ და

$$q(s_t, a_t) \leftarrow q(s_t, a_t) + \alpha\left(r_{t+1} - q(s_t, a_t)\right).$$

თამაში იმ შემთხვევაშ მთავრდება, თუ დაკმაყოფილებულია გაჩერების კრიტერიუმი ან განხორციელდა ტერმინალურ მდგომარეობაში გადასვლა.

გავარკვიოთ $Q-$სწავლების ალგორითმისა და *SARSA* ალგორითმის განსხვავებები კვლავ იმავე მაგალითზე. დავუშვათ, რომ სწავლების რომელიღაც ეტაპზე აგენტი მოხვდება s_1 მდგომარეობაში, აირჩევს a_2 მოქმედებას, ხოლო შეფასებათა ცხრილს ექნება შემდეგი სახე :

	a_1	a_2	a_3	a_4
s_1	–0.32	0.88	×	×
s_2	–0.02	×	×	6.5
s_3	×	–0.09	–0.19	×
s_4	×	0.99	×	6.14
s_5	×	×	×	×

მაშინ $q(s_1, a_2)$ შეფასების დაზუსტება გამოყენებული სტრატეგიისგან დამოუკიდებლად შემდეგნაირად იქნება ნაწარმოები :

$$q(s_1, a_2) \leftarrow 0.88 + 0.1 \cdot (-2 + 0.8 \cdot 6.5 - 0.88) \approx 1.12,$$

ვინაიდან

$$q(s_2, a_4) = max\big(q(s_2, a_1), q(s_2, a_4)\big).$$

SARSA ალგორითმში ეს ყოველთვის ასე არ არის. მაგალითად, ε – ხარბი სტრატეგიის გამოყენების შედეგად s_2 მგომარეობაში არჩეული რომ ყოფილიყო არა a_4 მოქმედება, არამედ a_1 მოქმედება, მაშინ $q(s_1, a_2)$ ღირებულების განახლებისათვის გამოყენებული იქნებოდა $q(s_2, a_1)$ მნიშვნელობა. Q – სწავლების შემთხვევაში სტრატეგია (მაგალითად, ε – ხარბი) გამოიყენება მხოლოდ მიმდინარე მოქმედების არჩევისას და არ მონაწილეობს ღირებულების შეფასების განახლებაში.

2.8 მაგალითი 2

შევჩერდეთ განხილული ალგორითმების გამოყენების მაგალითებზე. ჩავატაროთ ისეთი თამაშის სიმულირება, რომელშიც აგენტისათვის საჭიროა ლაბირინთიდან გამოსასვლელის პოვნა. შესაძლო ამოხსნის მაგალითი წარმოდგენილია ნახატზე 15.21.

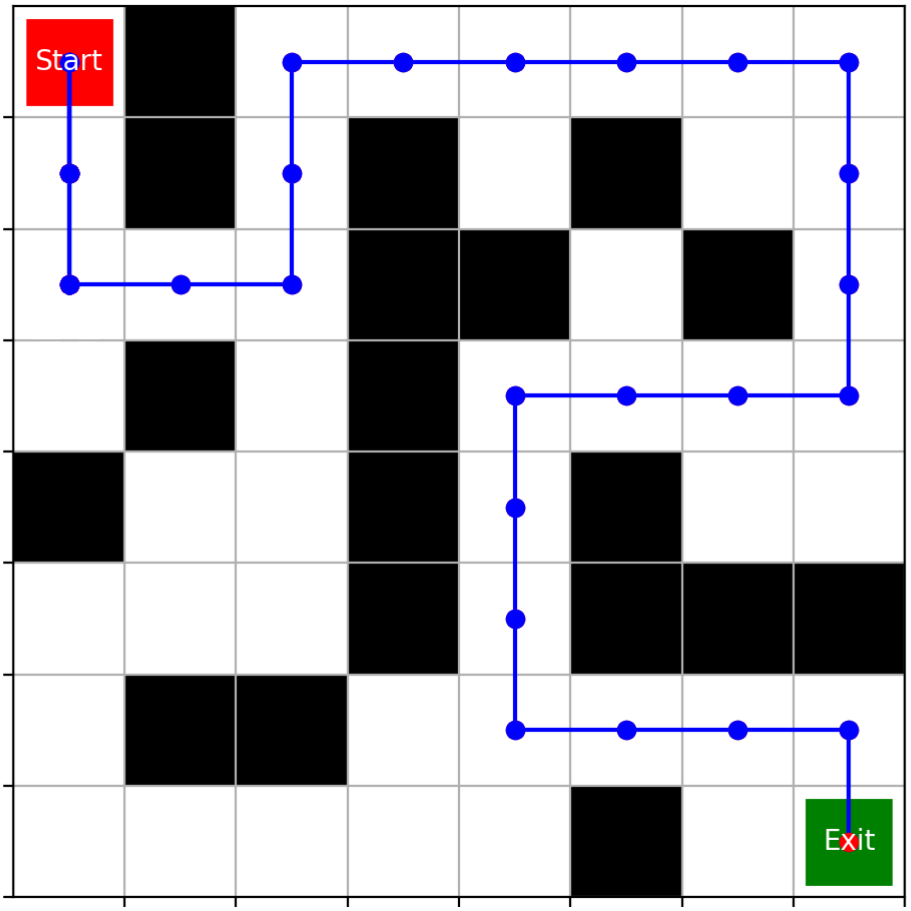
თამაშის პროცესში აგენტი თავსდება სასტარტო წერტილზე და მისთვის მისაწვდომია გადაადგილებები ოთხი მიმართულებით : ზევით, მარჯვნივ, ქვევით და მარცხნივ. აგენტის მიზანია — ლაბირინთიდან გამოსვლის წერტილამდე მიღწევა. ყოველი მოქმედების შედეგად აგენტი იღებს ჯილდოს. თამაშის პროცესში ჯილდოთა შეკრება ხდება. ყოველ ბიჯს თან ახლავს ჯილდო (-0.05). თუ აგენტი გადაადგილდა უკვე მონახულებულ უჯრედში, მაშინ ჯილდო შეადგენს (-0.25) სიდიდეს. თუ მოქმედების შედეგად აგენტს წინ ეღობება კედელი, მაშინ ჯილდო შეადგენს (-0.75). ლაბირინთიდან გამოსვლას თან ახლავს ჯილდო $(+10)$. ერთი თამაშის ფარგლებში ტერმინალური მდგომარეობა ყოველთვის მიიღწევა :

ან მოთამაშე პოულობს გამოსასვლელს და მაშინ თამაში გამარჯვებით მთავრდება, ან მიღებულ ჯილდოთა ჯამი აღმოჩნდება მოცემულ ზღურბლურ მნიშვნელობაზე ნაკლები.

მოცემულ შემთხვევაში ზღურბლური მნიშვნელობა განისაზღვრება შემდეგი თანაფარდობით :

$$-\frac{1}{2} \cdot (\text{ლაბირინთის ზომა}) = -\frac{1}{2} \cdot (8 \times 8) = -\frac{1}{2} \cdot 64 = -32.$$

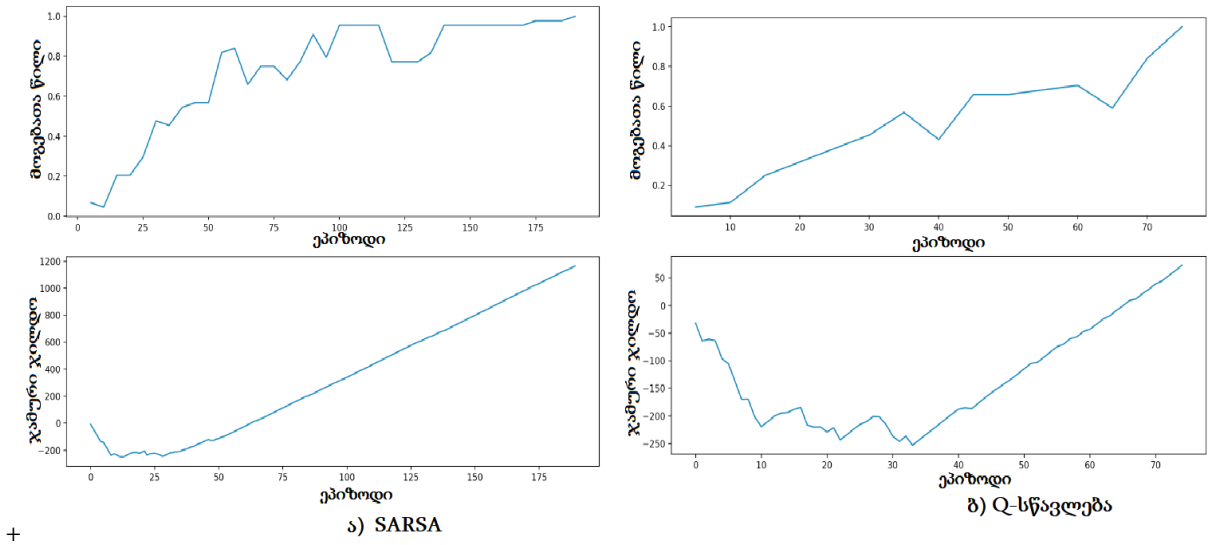
სწავლებისათვის გამოიყენებოდა $\varepsilon-$ხარბი ალგორითმი $\varepsilon=0.1$ მნიშვნელობით. დისკონტირების γ კოეფიციენტის მნიშვნელობაა 0.9, ხოლო α პარამეტრი $0.1-$ის ტოლია.



ნახატი 15.21 – ლაბირინთიდან გამოსვლის მაგალითი.

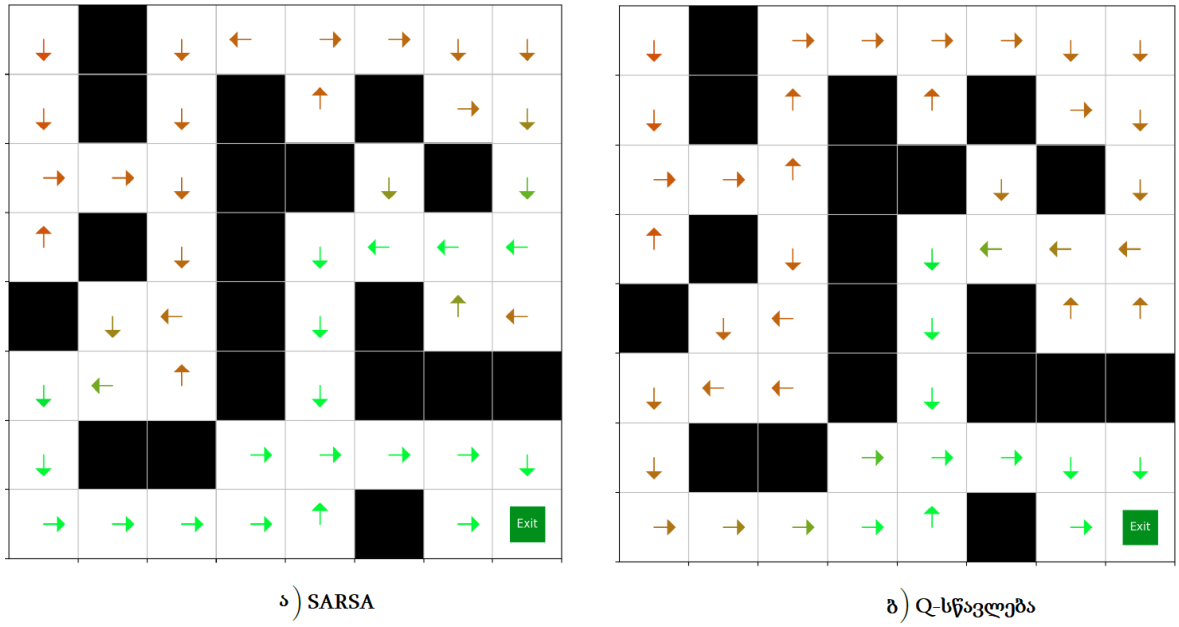
აგენტისთვის მისაწვდომია 44 საწყისი მდგომარეობა (თეთრი უჯრედები, ლაბირინთიდან გამოსვლის უჯრედის გამორიცხვით). ეპიზოდად მიიჩნევა 44 თამაშის სერია. თითოეულ თამაშში ყოველი მდგომარეობა (უჯრედი) ზუსტად ერთხელ გამოიყენება საწყისი (სასტარტო) მდგომარეობის (უჯრედის) როლში. სწავლება მთავრდება, თუ აგენტს აქვს გამოსასვლელისკენ (თუნდაც არაოპტიმალური) გზის პოვნის ცოდნა ნებისმიერი მდგომარეობიდან, რომელიც კი ხსენებულ 44 მისაწვდომ მდგომარეობათა რიცხვს მიეკუთვნება. სხვა სიტყვებით რომ ვთქვათ, ნებისმიერ მდგომარეობაში მოხვედრისას, აგენტმა ზუსტად იცის, რა მოქმედებები უნდა განახორციელოს, რომ გარანტირებულად გამოვიდეს ლაბირინთიდან. $Q-$სწავლების ალგორითმს დასჭირდა სულ 76 ეპიზოდი 44 თამაშით თითოეულში, მაშინ როცა *SARSA* ალგორითმს — 188 ეპიზოდი. მიღებულ ჯილდოთა (ანაზღაურებათა) ჯამს, ჩვეულებრივ, *ჯამურ ჯილდოს* უწოდებენ. შესაბამისი გრაფიკები წარმოდგენილია ნახატებზე 15.22.ა და 15.22.ბ *SARSA* და $Q-$სწავლების ალგორითმებისათვის, შესაბამისად. დადებითი მნიშვნელობების არეში გასასვლელად ჯამური ჯილდოს თვალსაზრისით თითოეულ ალგორითმს დაახლოებით 60 ეპიზოდი დასჭირდა. გარდა ამისა, ხსენებულ ნახატებზე წარმოდგენილია მოგებულ თამაშთა წილის გრაფიკები თითოეული ეპიზოდისათვის. მოგებულ თამაშთა წილი წარმოადგენს ეპიზოდის ფარგლებში მოპოვებულ გამარჯვებათა რაოდენობის შეფარდებას ეპიზოდების რიცხვთან, ჩვენს შემთხვევაში, ორმოცდაოთხთან.

აღსანიშნავია, რომ Q – სწავლების ალგორითმი ამ შემთხვევაში უზრუნველყოფს მოგებული თამაშების წილის უფრო მდოვრე, ნელა ზრდას.



ნახატი 15.22 - სწავლების ალგორითმების შედარება.

ყველა მდგომარეობისათვის მოქმედებათა ღირებულებების მიღებულ მნიშვნელობათა საფუძველზე შეიძლება ოპტიმალური სტრატეგიის ჩამოყალიბება. მიღებული სტრატეგიები *SARSA* და Q – სწავლების ალგორითმებისათვის წარმოდგენილია ნახატებზე 15.23.ა და 15.23.ბ, შესაბამისად.



ნახატი 15.23 - *SARSA* და Q – სწავლების ალგორითმთა სტრატეგიები.

შეიძლება ε – ხარბი სტრატეგიების გამოყენებასთან დაკავშირებული ზოგიერთი საინტერესო განსხვავების დანახვა. დახედეთ მდგომარეობას $(3,3)$ (უჯრედი, რომელიც შეესაბამება მესამე სტრიქონისა და მესამე სვეტის გადაკვეთას). თუ ვიწყებთ ამ მდგომარეობაში, მაშინ

სტრატეგია, რომელიც მიღებულია *SARSA* მეთოდის თანახმად, გვთავაზობს წავიდეთ ქვევით და მაშინ გამოსასვლელი ნაპოვნი აღმოჩნდება 17 ბიჯის შემდეგ. $Q-$სწავლების საფუძველზე მიღებული სტრატეგია, პირიქით, გვირჩევს ზევით წასვლას, მაგრამ მაშინ ლაბირინთიდან გამოსასვლელი ნაპოვნი იქნება 21 ბიჯის შემდეგ. ამ მდგომარეობისათვის *SARSA* უფრო ეფექტური აღმოჩნდა. მეორე მხრივ, *SARSA* ალგორითმისათვის შეიძლება აღმოვაჩინოთ გარკვეული უცნაურობა უჯრედში $(6,3)$ (მე-6 სტრიქონისა და მე-3 სვეტის გადაკვეთა). დაბოლოს, განსხვავებები, რომლებიც ორ ალგორითმს შორის შეიმჩნევა, როგორც, მაგალითად, მოძრაობის დაწყებისას $(5,8)$ უჯრედიდან, ზოგჯერ, შეიძლება, როლს არც თამაშობდეს.

3 საკვანძო მომენტები

1. კლასიკური მიდგომები სწავლებისადმი განმტკიცებით მოცემულია რიჩარდ სატონისა და ენდრიუ ბარტოს ცნობილ წიგნში Richard S. Sutton, Andrew G. Barto (2018). წიგნი განკუთვნილია ტექნიკური უნივერსიტეტების სტუდენტებისათვის, დამპროექტებლებისათვის, რომლებიც სპეციალიზებულია მანქანურ სწავლებაზე და ხელოვნურ ინტელექტზე, ასევე არატექნიკური პროფესიების წარმომადგენლებისათვისაც აღწერილი მეთოდების გამოსაყენებლად თავიანთ საქმიანობაში. არსებობს ამ წიგნის რუსული თარგმანიც: Саттон Ричард С., Барто Эндрю Г. Обучение с подкреплением, второе издание, издательство: ДМК-Пресс, 2020, 552 с.

2. მნიშვნელოვანი პრობლემებია უნივერსალური პრიორიტეტები, კოლმოგოროვის სირთულე ინდუქციასა და წინასწარმეტყველებისათვის. ამ პრობლემებთან გასაცნობად შეიძლება დავასახელოთ Marcus Hutter (2007), Ray J. Solomonoff (1964) და Marcus Hutter (2005).

3. ძალიან პერსპექტიულია ხისებრი ძებნა მონტე-კარლოს მეთოდით და აქ პირველ რიგში შეიძლება დავასახელოთ ნაშრომი Sylvain Gelly, David Silver (2011) და ასევე David Silver et al. (2017).

4. ცხადია. რომ მკითხველისათვის ძალზე საინტერესო იქნება დროითი სხვაობის ალგორითმის კრებადობასთან დაკავშირებული გამოკვლევები. ამ პრობლემას ეძღვნება კლასიკური ნაშრომი John N. Tsitsiklis, Benjamin Van Roy (1997). ხოლო პრაქტიკულად ახალ შედეგებს შეიძლება გავეცნოთ ნაშრომიდან Yann Ollivier (2018).

დასასრულ, აუცილებლად უნდა აღინიშნოს ის გარემოება, რომ სწავლება განმტკიცებით (წახალისებით) — ეს, შესაძლოა, მანქანური სწავლების ერთ-ერთი ყველაზე საინტერესო და დიდი იმედების მომცემი მიმართულებაა. სწავლების ასეთ ტიპში კარგად ეხამება ერთმანეთს როგორც სწავლება მასწავლებლით, ასევე სწავლება მასწავლებლის გარეშე. ერთი მხრივ, ადამიანის ჩარევა საჭირო არ არის, აგენტი დამოუკიდებლად სწავლობს. მეორე მხრივ, მასწავლებლის როლში, გარკვეული თვალსაზრისით, გარემო გამოდის.

კიდევ ერთი დადებითი მხარე ისაა, რომ სწავლება შეიძლება ხდებოდეს დინამიურად ცვალებად გარემოში.

ცხადია, რომ ამ ლექციაში, მისი შეზღუდული ფორმატის ფარგლებში, ჩვენ განვიხილეთ განმტკიცებით (წახალისებით) სწავლების მეთოდთა მცირე ნაწილი, რომელთა რაოდენობა სინამდვილეში საკმაოდ დიდია. ამასთან ერთად სრულიად ახალი გადაწყვეტილებები და

იდეები ამ სფეროში შესაშური რეგულარობით ქვეყნდება.

ძვირფასო მკითხველო, ამით მანქანური სწავლების ჩვენი კურსი დასრულდა. იმედი გვაქვს, რომ თქვენ მიერ მიღებული ცოდნა დაგეხმარებათ პროფესიული და ყოფითი ამოცანების გადაწყვეტაში. გარდა ამისა ჩვენ უსაზღვროდ გაგვიხარდება, თუ ეს კურსი აღძრავს თქვენში სურვილს დაუთმოთ დრო ხელოვნური ინტელექტის სხვადასხვა მიმართულების შემდგომ შესწავლას. წარმატებებს გისურვებთ!

4 ბიბლიოგრაფია

1. Richard S. Sutton, Andrew G. Barto, Reinforcement Learning : An Introduction, Second Edition, MIT Press, Cambridge, MA, 2018, 552 pages.
2. Marcus Hutter, On universal prediction and Bayesian confirmation, Theoretical Computer Science, 384, 2007, pages 33-48.
3. Ray J. Solomonoff, A Formal Theory of Inductive Inference. Part I and Part II, Information and Control, 7, 1964, 1-22; 224-254.
4. Marcus Hutter, Universal Artificial Intelligence: Sequential Decisions Based On Algorithmic Probability, Springer, 2005, 298 pages.
5. Sylvain Gelly, David Silver, Monte-Carlo tree search and rapid action value estimation in computer Go, Artificial Intelligence, Volume 175, Issue 11, 2011, pages 1856-1875.
6. David Silver et al., Mastering the game of Go without human knowledge, Nature, Volume 550, 2017, pages 354–359.
7. John N. Tsitsiklis, Benjamin Van Roy, An Analysis of Temporal-Difference Learning with Function Approximation, IEEE Transactions on Automatic Control, vol. 42, no. 5, may 1997, pages 674-690
8. Yann Ollivier, Approximate Temporal Difference Learning is a Gradient Descent for Reversible Policies, 2018 : https://arxiv.org/pdf/1805.00869.pdf

რედაქცია - პროფ. ო. ნამიჩეიშვილი

კომპიუტერული დაკაბადონება - ავტორების მიერ.

სტამბური გამოცემა - გოჩა დალაქიშვილი

გადაეცა წარმოებას 15.02.2024. ხელმოწერილია ოფსეტური ქაღალდის
ზომა 70 x 100 1/16. პირობითი ნაბეჭდი თაბახი 23,4. ტირაჟი 8 ეგზ.

Verba volant,
scripta manent

სტუ-ს „IT კონსალტინგის სამეცნიერო ცენტრი“,
თბილისი, მ.კოსტავას 77

არჩილ ფრანგიშვილი — ტექნიკის მეცნიერებათა დოქტორი, პროფესორი, საქართველოს მეცნიერებათა ეროვნული აკადემიის ნამდვილი წევრი, საქართველოს საინჟინრო აკადემიის პრეზიდენტი, რიგი საერთაშორისო და საზღვარგარეთის აკადემიის აკადემიკოსი, ინფორმატიკისა და გამოთვლითი ტექნიკის, კონფლიქტოლოგიის, მართვის პროცესებისა და სისტემების თეორიის საყოველთაოდ აღიარებული სპეციალისტი, გამოჩენილი მეცნიერი, პედაგოგი და საზოგადო მოღვაწე.

ოლეგ ნამიჩეიშვილი — ტექნიკის მეცნიერებათა დოქტორი, პროფესორი, საქართველოს და საერთაშორისო საინჟინრო აკადემიათა ნამდვილი წევრი, საქართველოს საბუნებისმეტყველო მეცნიერებათა აკადემიის ნამდვილი წევრი, ინფორმაციული სისტემების, გამოთვლითი და ელექტრონული ტექნიკის, ხელოვნური ნეირონული ქსელების, ფიზიკური მოვლენების კომპიუტერული მოდელირებისა და საიმედოობის მათემატიკური თეორიის ცნობილი მკვლევარი.

ჟუჟუნა გოგიაშვილი — ფიზიკა-მათემატიკის მეცნიერებათა კანდიდატი, საქართველოს ტექნიკური უნივერსიტეტის ასოცირებული პროფესორი, ფართო მეცნიერული ინტერესებით საიმედოობის მათემატიკური თეორიის, ხელოვნური ნეირონული ქსელების, ფიზიკურ მოვლენათა კომპიუტერული მოდელირების, ინფორმაციული ტექნოლოგიების, საინჟინრო ეთიკისა და ხელოვნური ინტელექტის სფეროებში.