Primjena linearne regresije i strojnog učenja u izradi kreditnog skoringa u bankarstvu

Gavrić, Marko

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Marko Gavrić

PRIMJENA LINEARNE REGRESIJE I STROJNOG UČENJA U IZRADI KREDITNOG SKORINGA U BANKARSTVU

Diplomski rad

Voditelj rada:
prof.dr.sc.Siniša Slijepčević

Zagreb, srpanj 2023.

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Zahvaljujem mentoru doc.dr.sc. Siniši Slijepčeviću na velikoj pomoći pri pisanju ovog rada, na uloženom vremenu, strpljenju i trudu.
Iznimno sam zahvalan svojim roditeljima što su mi pružali podršku i omogućili mi ovaj uspjeh.
Veliko hvala sestri Branki na bezuvjetnoj podršci i razumijevanju tokom svih ovih godina.
Također se zahvaljujem prijateljima i kolegama na nezaboravnim trenucima tokom studija.
Za kraj, posebno hvala mojoj najvećoj motivaciji pri pisanju ovog rada Aleksandri i B.

Sadržaj

Uvod

U današnje digitalno doba, podaci igraju ključnu ulogu u svim aspektima poslovanja. Od malih start-up poduzeća do velikih korporacija, sve više organizacija prepoznaje važnost prikupljanja, analize i korištenja podataka u svrhu donošenja efikasnijih poslovnih odluka. Stoga, strojno učenje i umjetna inteligencija postaju ključne tehnologije pri radu s ogromnim količinama podataka. One mogu znatno poboljšati učinkovitost, donošenje odluka, predviđanje trendova i optimizaciju procesa.

Cilj ovog diplomskog rada je istražiti primjenu strojnog učenja u bankarstvu, fokusirajući se na kreditni skoring kojim se procjenjuje kreditna sposobnost potencijalnih klijenata, odnosno zajmotražitelja. Rad se sastoji od 4 poglavlja. U prvom poglavlju navedeni su pojmovi iz vjerojatnosti i statistike koji su nužni za daljnje razumijevanje rada. Drugo poglavlje bavi se linearnom regresijom, jednom od najpopularnijih statističkih metoda za proučavanje odnosa dviju varijabli. Nadalje, treće poglavlje predstavlja uvod u strojno učenje i njegove modele s naglaskom na boosting algoritam. Posljednje poglavlje prikazuje primjenu prethodno navedenih metoda u izračunu kreditnog skoringa.

Poglavlje 1

Osnovni koncepti statistike

Pojam statistika izveden je iz latinskog izraza *statisticum collegium* (vijeće država) te talijanske riječi *statista* (državnik ili političar). Od početka 19. stoljeća smatra se da je statistika grana primijenjene matematike koja se bavi prikupljanjem, uređivanjem, analizom, sažimanjem, interpretiranjem i prezentiranjem velikog broja podataka te izradom predviđanja na temelju tih podataka [1]. Podaci se mogu dobiti promatranjem ili iz statističkoga pokusa te se smatraju statističkima samo ako su prikupljeni prema nacrtu statističkoga pokusa. Potrebno je raspolagati dovoljnim brojem podataka kako bi do izražaja došle osobitosti istraživanih pojava, odnosno statističke zakonitosti.

Populacija i uzorak su temeljni pojmovi statistike. Sljedeće definicije su preuzete iz [1], poglavlje 1.2.

Definicija 1.0.1. ***Populacija*** *ili osnovni skup je skup podataka svih jedinki ili objekata od interesa.*

Populacija može imati konačno ili beskonačno mnogo objekata. Primjerice ako želimo analizirati visinu studenata studija matematike onda je populacija, i to konačna, skup mjerenih visina svih studenata matematike. Primjer populacije sa beskonačno mnogo objekata jesu svi rezultati eksperimentalnih mjerenja koji bi mogli biti opaženi ako se mjerenja provedu beskonačno mnogo puta pod istim uvjetima. U praksi obično nije moguće pridobiti informacije o sveukupnoj populaciji. Stoga je primarni cilj statistike sakupljanje i proučavanje podskupa populacije kako bi se dobile informacije o nekim specifičnim karakteristikama populacije koje su od interesa. Takav podskup naziva se uzorkom. Neovisno o tome je li populacija konačna ili beskonačna, promatrani uzorak je uvijek konačan.

Definicija 1.0.2. ***Uzorak*** *je podskup populacije za koji se mjere ili skupljaju podaci koji nas zanimaju.*

Dva osnovna tipa statistike jesu deskriptivna i inferencijalna statistika. Deskriptivna statistika opisuje dane podatke. Tu je uzorak jednak populaciji te se na njemu primjenjuju

metode koje se sastoje od uređivanja, sažimanja i prikazivanja podataka u obliku tablica, grafikona i dijagrama. Inferencijalna statistika izvodi zaključke i donosi odluke o populaciji pomoću promatranog uzorka. Također, ona koristi teoriju vjerojatnosti.

Definicija 1.0.3. ***Statistički zaključak*** *je procjena, predviđanje, odluka ili generalizacija o populaciji na temelju informacija sadržanih u uzorku.*

1.1 Deskriptivna statistika

Vrste podataka

Podaci se mogu klasificirati na nekoliko načina. Promotrit ćemo dvije vrste klasifikacija, jedna se temelji na tome jesu li podaci mjereni na numeričkoj skali ili ne, a druga na tome jesu li podaci prikupljeni u istom ili različitom vremenskom razdoblju. Obzirom na skalu po kojoj su podaci mjereni, razlikujemo kvantitativne i kvalitativne podatke. Kvantitativni ili numerički podaci su opažanja mjerena na numeričkoj skali. Za ne numeričke podatke koji se mogu svrstati samo u jednu od skupina kategorija kaže se da su kvalitativni ili kategorički podaci. Mjerenja kvantitativnih podataka mogu biti diskretna ili neprekidna, dok kvalitativne podatke dijelimo na nominalne i ordinalne. S druge strane, podatke dijelimo na podatke presjeka i podatke vremenskih serija. Podaci presjeka su podaci u kojima se prikuplja više varijabli za pojedini element u istom vremenskom trenutku ili u istom vremenskom periodu, a podaci prikupljeni o istom elementu ili istoj varijabli u različitim vremenskim trenucima ili za različita vremenska razdoblja nazivaju se podaci vremenske serije.

Uzorkovanje

U ovom poglavlju slijedimo [1], poglavlje 1.3. U svakoj statističkoj analizi važno je da se populacija jasno definira u skladu sa ciljevima studije. Kada je u studiju uključena cijela populacija, to se naziva popisna studija (eng. *census study*) jer se podaci prikupljaju o svakom članu populacije. Općenito nije moguće dobiti informacije o cijeloj populaciji radi njene veličine ili isplativosti (financijske ili vremenske). Mali, ali pažljivo odabran, uzorak može se koristiti za predstavljanje populacije. Dobar uzorak dobivamo prikupljanjem podataka samo od pojedinih članova populacije uz očuvanje svih njenih značajnih karakteristika. Uzorak nazivamo reprezentativnim ako po svojim osnovnim karakteristikama nalikuje na populaciju, suprotno uzorak nazivamo pristranim (eng. *biased*). Pouzdanost ili točnost zaključaka koji se odnose na populaciju ovisi kvaliteti izabranoga uzorka, odnosno reprezentira li populaciju dovoljno dobro. Dostupne su mnoge metode uzorkovanja, u nastavku navodimo nekoliko jednostavnih, ali često korištenih metoda.

Definicija 1.1.1. *Uzorak odabran na način da svaki element populacije ima jednake šanse biti izabran naziva se* ***jednostavnim slučajnim uzorkom****. Ekvivalentno, svaki mogući uzorak veličine n ima jednaku šansu da bude odabran.*

Neke prednosti jednostavnog slučajnog uzorkovanja jesu mogućnost procjene veličine uzorka za propisanu razinu pogreške, nemogućnost pristranosti istraživača i nepotreba za poznavanjem populacije.

Definicija 1.1.2. ***Slučajni sustavni uzorak*** *je uzorak u kojem se odabire svaki k-ti element nakon slučajno odabranog prvog elementa, gdje je k broj dobiven dijeljenjem veličine populacije s veličinom uzorka.*

Slučajno sustavno uzorkovanje ima široku primjenu jer ga je lako provesti, no bitno je napomenuti da ukoliko postoji korelacija između uzastopnih elemenata ili neka periodična struktura, tada ova metoda uzorkovanja može dovesti do pristranosti.

Definicija 1.1.3. *Stratificirano uzorkovanje je modifikacija jednostavnog slučajnog uzorkovanja i slučajnog sustavnog uzorkovanja s ciljem dobivanja reprezentativnijeg uzorka, ali uz cijenu kompliciranijeg postupka. Elementi populacije grupiraju se u homogene skupine ili stratume na temelju jednog ili više čimbenika zatim uzorkovanjem iz svakog od stratuma dobivamo uzorak kojeg nazivamo* ***slučajni stratificirani uzorak****.*

Za razliku od ostalih metoda, stratificiranim se uzorkovanjem osigurava reprezentativnost obzirom na nama relevantne faktore populacije čime se smanjuje pogreška uzorkovanja. Također, ovo se uzorkovanje često koristi kada jedan ili više stratuma u populaciji imaju nisku učestalost u odnosu na ostale stratume.

Definicija 1.1.4. *U klasterskom uzorkovanju jedinice uzorkovanja su prirodno pojavljujuće grupe elemenata koje se nazivaju klasteri.* ***Slučajni uzorak klastera*** *se dobiva jednostavnim slučajnim uzorkom klastera (grupa) te zatim uzorkovanjem svih elemenata unutar odabaranih klastera.*

Za razliku od stratificiranog uzorkovanja, gdje istraživač raspolaže relevantnim podacima o populaciji i stvara stratume, u klasterskom uzorkovanju grupe (klasteri) nisu formirane od strane istraživača već prirodno postoje.

Odabir metode uzorkovanja ovisi o prirodi problema ili istraživanja, dostupnosti dovoljno dobrih izvora podataka iz kojih uzorkujemo, proračunu ili raspoloživim financijskim sredstvima, željenoj razini točnosti te metodi prikupljanja podataka (ankete, intervjui, ...). Uz odabir metode kojom se uzorkuje, jedno od glavnih pitanja je odabir veličine uzorka. Odabir prave veličine uzorka je važan jer može utjecati na valjanost i točnost rezultata. Idealna veličina uzorka trebala bi biti dovoljno velika da točno predstavlja populaciju, ali dovoljno mala da se može prikupiti i analizirati. Veličina također ovisi o čimbenicima kao

što su varijabilnost populacije te željena statistička pouzdanost rezultata, odnosno tolerirana količina pogreške.

Grafičko prikazivanje podataka

U podacima leži izvor našeg statističkog znanja. Jedan od načina upoznavanja podataka jesu tabelarni i grafički prikazi. Takvi prikazi su vrlo važan dio deskriptivne statistike iz razloga što vizualno prenose informacije. U poslovnom svijetu su grafički prikazi svakodnevni statistički alat za donošenje odluka te razumijevanje procesa, problema i rješenja.

Vrijednosti kvalitativne varijable jesu kategorije, a mjere kojima opisujemo zastupljenost pojedine kategorije u uzorku jesu frekvencija kategorije te relativna frekvencija kategorije. Podatke možemo prikazivati tabelarno i grafički, posebno korisni grafički prikazi kvalitativnih varijabli su stupičasti dijagram, *Pareto* dijagram i kružni dijagram.

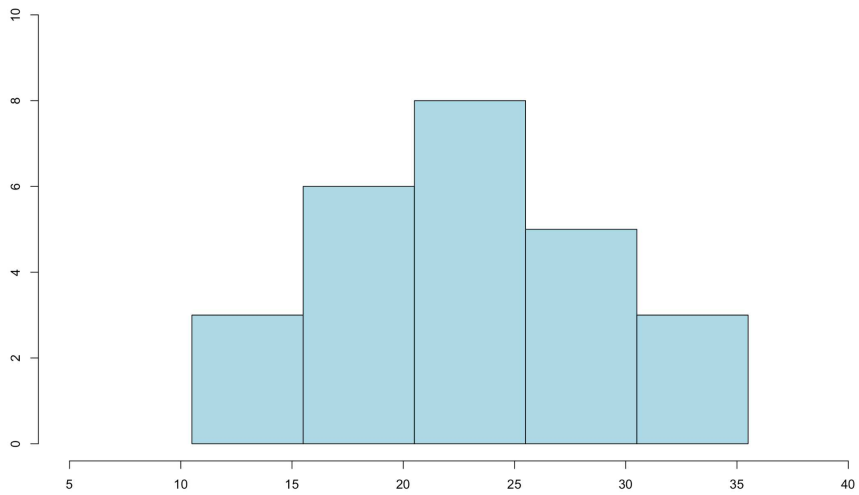
Kvantitativne varijable mogu biti diskretne i neprekidne. Ukoliko su kvantitativne varijable diskretne s malo mogućih vrijednosti, tada podatke možemo opisati gore navedenim metodama opisivanja kvalitativnih podataka. U suprotnom, ako numerička varijabla prima mnogo međusobno različitih vrijednosti, za grafički prikaz podataka koristimo posebnu vrstu stupičastog dijagrama zvanog histogram.

Definicija 1.1.5. ***Histogram*** *je grafički prikaz nekog skupa podataka koji se sastoji od međusobno susjednih pravokutnika s po jednom stranicom na osi apscisa. Pritom se površine dijelova odnose kao (relativne) frekvencije podataka, a visine iznad pojedine apscise su te (relativne) frekvencije po jednoj stranici apscise, tj. površine podijeljene sa širinom pojedine skupine podataka.*

Stoga, u histogramu odnose među frekvencijama ne pokazuju ordinate već površine, a ukupna površina histograma jednaka je ukupnom broju izmjerenih vrijednosti, odnosno 1 ukoliko se radi o relativnim frekvencijama. Ukoliko su širine skupina podataka jednake, onda su visine pravokutnika histograma proporcionalne frekvencijama. Primjerice, sljedeće podatke želimo prikazati pomoću histograma. Primijetimo kako se iz histograma lako uočavaju središte podataka, raspršenost podataka, asimetrija u podacima, *outlier*-i, zvonoliki oblik i slično. Slijede grafički prikazi podataka konstruirani u *R*-u pomoću *echarts4r* paketa.

11	19	24	30	12	20	25	29	15	21
24	31	16	23	25	26	32	17	22	26
35	18	24	18	27					

Tablica 1.1: Rezultati mjerenja nečistoće vode, udio po milijunu



Slika 1.1: Histogram frekvencija

Numerički opis podataka

S ciljem prikazivanja skupa podataka, u ovom potpoglavlju ćemo razmotriti neke od najčešće korištenih numeričkih karakteristika skupa podataka. Pretpostavimo da imamo uzorak s numeričkim vrijednostima $x_1, x_2, \ldots, x_n$. Numeričke karakteristike povezane s ovim skupom podataka jesu mjere lokacije i mjere raspršenja podataka. Tri najvažnije mjere lokacije, odnosno centralne tendencije, jesu aritmetička sredina, medijan i mod.

Definicija 1.1.6. ***Aritmetička sredina*** *uzorka $x_1, x_2, \ldots, x_n$ definirana je kao*

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i. \tag{1.1}$$

Definicija 1.1.7. ***Medijan*** *je broj koji se nalazi u sredini sortirane liste podataka (ili je aritmetička sredina srednjih dvaju podataka). Malo preciznije, to je broj sa svojstvom da 50% svih podataka ima vrijednost bar koliko on iznosi. Vertikala povučena u medijanu dijeli histogram na dva dijela jednake površine.*

Definicija 1.1.8. ***Mod*** *je iznos koji se najčešće pojavljuje, tj. to je vrijednost s najvećom frekvencijom. On ne mora postojati, ni biti jedinstveno određen. Može se opisati i kao najtipičnija vrijednost uzorka.*

Uz mjere lokacije, odnosno srednje vrijednosti skupa podataka, važno svojstvo distribucije podataka je i kako su podaci raspršeni, često u odnosu na neku srednju vrijednost.

Definicija 1.1.9. ***Raspon*** *je razlika maksimalne i minimalne vrijednosti u uzorku.*

Definicija 1.1.10. ***Prvi (donji) kvartil*** *je broj od kojega je 25% podataka manje ili je njemu jednako.* ***Treći (gornji) kvartil*** *je broj od kojega je 75% podataka manje ili je njemu jednako. Konačno,* ***interkvartil*** *skupa podataka je razlika gornjeg i donjeg kvartila, $IQR = q_U - q_L$.*

Definicija 1.1.11. ***Varijanca*** *uzorka je mjera raspršenja podataka koja predstavlja prosječno kvadratno odstupanje podataka od njihove aritmetičke sredine i dana je formulom,*

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2. \tag{1.2}$$

Definicija 1.1.12. ***Standardna devijacija*** *uzorka jest drugi korijen varijance, tj.:*

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2}. \tag{1.3}$$

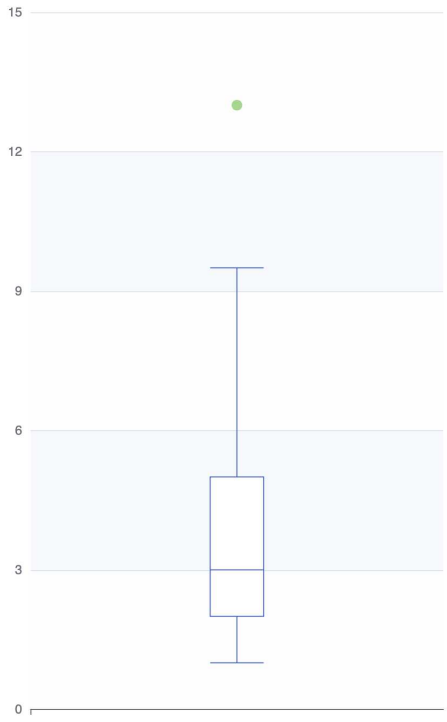
Za grafički prikaz distribucije skupa numeričkih podataka koristi se dijagram pravokutnika (engl. *box and whisker*). Iz njega se direktno može očitati medijan, donji i gornji kvartil, interkvartil, raspon, ekstremne vrijednosti i simetrija. Na slici 1.2 nalazi se primjer dijagrama pravokutnika. Iz njega vidimo da je medijan 3, donji kvartil 2, gornji kvartil 5.5, interkvartil 3.5, donji brk se proteže do 1, a gornji do 10.75 ($1.5(IQR)$), imamo i jednu ekstremnu vrijednost koja iznosi 13.

1.2 Teorija vjerojatnosti

Teorija vjerojatnosti je matematička disciplina čiji je zadatak formirati i proučavati matematički model nekog danog slučajnog pokusa (eksperimenta). Koncept vjerojatnosti zauzima važnu ulogu u procesu donošenja odluka, bilo da se radi o problemu u poslovnom svijetu, inženjerstvu, politici, znanosti ili svakodnevnom životu [2]. Matematički modeli teorije vjerojatnosti omogućuju nam da predvidimo određene pojave iz nužno nepotpunih informacija dobivenih tehnikama uzorkovanja. Teorija vjerojatnosti je ta koja omogućuje prijelaz iz deskriptivne statistike u inferencijalnu. Zapravo, teorija vjerojatnosti je najvažniji alat u statističkom zaključivanju. [1]

Prostor elementarnih događaja

Svaki pokus definiran je odnosom uzroka i posljedica, a pretpostavke za realizaciju pokusa su ponavljanje pokusa proizvoljno konačno mnogo puta te poznavanje svih mogućih



Slika 1.2: Dijagram pravokutnika

ishoda. Ishodi pokusa jedini su objekti koji nam služe za izgradnju matematičkog modela. Obzirom na ishod, postoje dvije vrste pokusa deterministički i slučajni pokus. U determinističkom pokusu je ishod jednoznačno određen uvjetima pokusa, dok u slučajnom pokusu nije. Osnovna pretpostavka slučajnog pokusa je da svako izvođenje pokusa mora dati ishod, tj. događaj koji odgovara jednom i samo jednom elementarnom događaju. Slučajni pokus je definiran svojim osnovnim ishodima koji se međusobno isključuju i zovu se elementarni događaji, a označavaju se malim grčkim slovima $\omega_1, \omega_2, \omega_3, \ldots$. Skup $\Omega = \{\omega_i : \omega_i = \text{elementarni događaj}, i = 1, 2, \ldots, n, \ldots\}$ je ne prazan skup i zove se prostor elementarnih događaja. Cijeli prostor elementarnih događaja Ω je siguran događaj koji se mora dogoditi u svakom izvođenju pokusa, dok je prazan skup $\emptyset$ nemoguć događaj.

Definicija 1.2.1. ***Slučajni događaj*** *je podskup prostora elementarnih događaja. Slučajni događaji označavaju se velikim tiskanim slovima latinice* $A, B, \ldots \subseteq \Omega$.

Definicija 1.2.2. *Elementarni događaj koji pripada događaju A zove se* ***povoljan događaj*** *za A. Pojavljivanje tog elementarnog događaja u pokusu povlači da se dogodio događaj A.*

Definicije vjerojatnosti

Klasična definicija vjerojatnosti

Neka je $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ konačan skup gdje su svi elementarni događaji jednako mogući te neka događaj A ima m povoljnih elementarnih događaja, $A \subseteq \Omega$, $|A| = m$. Vjerojatnost svakog elementarnog događaja je $P(\omega_i) = \frac{1}{|\Omega|}$, a vjerojatnost događaja A definira se kao $P(A) = \frac{|A|}{|\Omega|}$.

Aksiomatska definicija vjerojatnosti

Neka je Ω prostor elementarnih događaja. Partitivni skup ili skup svih podskupova od Ω zovemo skup svih mogućih događaja slučajnog pokusa. Podskup $\mathcal{A} \subseteq P(\Omega)$ zovemo familija događaja iz Ω.

Definicija 1.2.3. *Neka familija događaja $\mathcal{F} \subseteq P(\Omega)$ ima svojstva:*

i) $\emptyset \in \mathcal{F}$,

ii) $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$,

iii) Ako je $A_i \in \mathcal{F}, i \in \mathbb{N} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

*Takvu familiju skupova $\mathcal{F}$ zovemo **sigma algebra** događaja (σ-algebra). Ako je Ω konačan skup, onda je i svaka σ-algebra $A \subseteq P(\Omega)$ konačna i naziva se algebra događaja.*

Definicija 1.2.4. *Neka je Ω prostor elementarnih događaja slučajnog pokusa i neka je $\mathcal{F}$ σ-algebra skupova na Ω. Funkcija $P : \mathcal{F} \to \mathbb{R}$ zove se **vjerojatnost** na $\mathcal{F}$ ako vrijedi:*

(P1) $P(A) \geq 0, A \subseteq \mathcal{F}$ *(svojstvo nenegativnosti),*

(P2) $P(\Omega) = 1$ *(svojstvo normiranosti),*

(P3) $A_i \in \mathcal{F}, i \in \mathbb{N}, A_i \cap A_j = \emptyset, i \neq j \Rightarrow P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$ *(svojstvo prebrojive aditivnosti).*

Vjerojatnosnim prostorom zovemo uređenu trojku $(\Omega, \mathcal{F}, P)$, gdje je $\mathcal{F}$ σ-algebra na Ω, a P vjerojatnost na Ω. Ako vrijedi da je Ω prebrojiv ili konačan skup elementarnih događaja, onda je $(\Omega, \mathcal{F}, P)$ diskretni vjerojatnosni prostor. Slijede svojstva funkcije vjerojatnosti, dokaz teorema preskačemo.

Teorem 1.2.5. *Neka je $(\Omega, \mathcal{F}, P)$ vjerojatnosni prostor. Tada za funkciju vjerojatnosti P vrijedi:*

i) $P(\emptyset) = 0$

ii) $A_i \in \mathcal{F}, i \in \{1, \ldots, n\}, A_i \cap A_j = \emptyset, i \neq j \Rightarrow P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P\left(A_j\right)$ *(svojstvo konačne aditivnosti)*

iii) $A, B \in \mathcal{F}, A \subseteq B \Rightarrow P(A) \leq P(B)$ *(svojstvo monotonosti)*

iv) $A \in \mathcal{F} \Rightarrow P\left(A^c\right) = 1 - P(A)$

v) $A, B \in \mathcal{F} \Rightarrow P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Slučajne varijable

Elementarni događaji su događaji koji mogu biti rezultat nekog razmatranog pokusa. U većini slučajeva, pokus može sadržavati brojne karakteristike koje se mogu mjeriti te osoba koja provodi pokus bira specifične karakteristike pokusa na koje će se usredotočiti. Matematički se to može predočiti funkcijom koja slučajno, ovisno o ishodu pokusa, postiže određene vrijednosti te se iz toga razloga naziva slučajnom varijablom. Preciznije, neka je Ω vjerojatnosni prostor tada je slučajna varijabla funkcija $X : \Omega \to \mathbb{R}$. Važno je napomenuti da u definiciji slučajne varijable, vjerojatnost ne igra nikakvu ulogu. Međutim, za svaku vrijednost ili skup vrijednosti slučajne varijable postoje određeni događaji, a kroz te događaje povezujemo vrijednosti slučajne varijable s mjerama vjerojatnosti.

Definicija 1.2.6. ***Funkcija distribucije*** *slučajne varijable X je funkcija*

$$F_X(x) = P(X \leq x). \tag{1.4}$$

Vjerojatnost $P(X \leq x)$ je vjerojatnost da slučajna varijabla poprimi vrijednost manju ili jednaku vrijednosti x:

$$P(X \leq x) = P\left(\bigcup_{y \leq x} X^{-1}(y)\right).$$

Slučajna varijabla X je diskretna slučajna varijabla ako može poprimiti samo konačno ili prebrojivo mnogo vrijednosti. S druge strane, pretpostavimo da postoji nenegativna realna funkcija $f : \mathbb{R} \to [0, \infty\rangle$ tako da za svaki interval $[a, b]$,

$$P(X \in [a, b]) = \int_a^b f(t)dt.$$

Tada X nazivamo neprekidnom slučajnom varijablom, a funkciju f funkcijom gustoće varijable X. Funkcija f je funkcija gustoće ako zadovoljava sljedeće uvijete:

$$f(x) \geq 0, \forall x$$
$$\int_{-\infty}^{\infty} f(x)dx = 1.$$

Neka je $x \in \langle -\infty, +\infty \rangle$, funkcija distribucije diskretne slučajne varijable X dana je s

$$F_X(x) = P(X \leq x) = \sum_{\forall y \leq x} P(X = y),$$

dok je za neprekidnu slučajnu varijablu Y dana s

$$F_Y(x) = P(Y \leq x) = \int_{-\infty}^{x} f(t)dt.$$

Jedan od najkorisnijih koncepata u teoriji vjerojatnosti je koncept očekivanja slučajne varijable. Očekivana vrijednost može se promatrati kao točka ravnoteže distribucije vjerojatnosti na realnoj skali, ili uobičajeno rečeno, presjek. Očekivanu vrijednost nazivamo još i očekivanje ili matematičko očekivanje te ju označavamo s μ. Očekivanu vrijednost diskretne slučajne varijable označavamo s $\mathbb{E}[X]$, a definiramo je kao

$$\mu = \mathbb{E}[X] = \sum_{\forall x} xP(x), \tag{1.5}$$

pod uvjetom da je $\sum_{\forall x} |x|P(x) < \infty$. Dakle, očekivanje diskretne slučajne varijable je prosjek svih vrijednosti varijable X. Očekivana vrijednost neprekidne slučajne varijable X s funkcijom gustoće f definirana je s

$$\mu = \mathbb{E}[X] = \int_{-\infty}^{\infty} xf(x)dx, \tag{1.6}$$

pod uvjetom da je $\int_{-\infty}^{\infty} |x|f(x)dx < \infty$.

Raspršenje vrijednosti slučajne varijable X oko srednje vrijednosti μ mjerimo varijancom koja se definira kao

$$\sigma^2 = Var(X) = \mathbb{E}\left(X - \mu\right)^2, \tag{1.7}$$

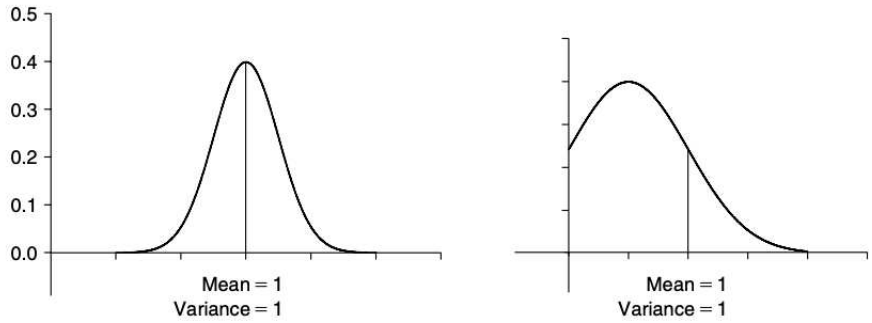
pri čemu s desne strane imamo prosječnu vrijednost kvadrata otklona. Nadalje, drugi korijen varijance nazivamo standardna devijacija i označavamo s

$$\sigma_X = \sqrt{Var(X)}. \tag{1.8}$$

Iako su očekivanje i standardna devijacija značajne deskriptivne mjere distribucije, one ne daju jedinstvenu karakterizaciju distribucije. Dvije distribucije mogu imati isto očekivanje i varijancu, ali mogu biti vrlo različite (vidi sliku 1.3). Za bolju aproksimaciju distribucije slučajne varijable trebaju nam viši momenti.

Neka je X slučajna varijabla, k prirodan broj, a c neki realan broj. K-ti moment od X oko c je broj

$$\mathbb{E}[(X - c)^k].$$



Slika 1.3: Funkcije distribucije dviju slučajnih varijabli

Momenti oko ishodišta ($c = 0$) jednostavno se nazivaju momentima, dok su momenti oko matematičkog očekivanja centralni momenti. Na primjer, matematičko očekivanje je prvi moment, a varijanca drugi centralni moment. Standardizirani treći centralni moment, tzv. koeficijent asimetrije (eng. *skewness*), slučajne varijable X definira se kao

$$\alpha_3(X) = \mathbb{E}[\left(\frac{X-\mu}{\sigma}\right)^3] = \frac{\mathbb{E}(X-\mu)^3}{\sigma^3} = \frac{\mu_3}{\mu_2^{\frac{3}{2}}}.$$

Distribucija od X je simetrična ako je $\alpha_3(X) = 0$, negativno je asimetrična ako je $\alpha_3(X) < 0$, a pozitivno asimetrična ako je $\alpha_3(X) > 0$. Standardizirani četvrti centralni moment, tzv. koeficijent spljoštenosti (eng. *kurtosis*), definira se kao

$$\alpha_4 = \frac{\mathbb{E}(X-\mu)^4}{\sigma^4}.$$

Koeficijent spljoštenosti temelji se na veličini repova distribucije, pozitivna vrijednost ukazuje na premalo opažanja u repovima, a negativna ukazuje na previše opažanja u repu distribucije.

Primjeri važnih distribucija

Slučajne varijable često se klasificiraju funkcijama distribucije, stoga su one jedno od najvažnijih svojstava slučajnih varijabli. U ovom potpoglavlju, prateći poglavlje 3 iz [1], navodimo distribucije neprekidnih slučajni varijabli koje se prirodno pojavljuju u mnogim područjima primjene.

Normalna distribucija

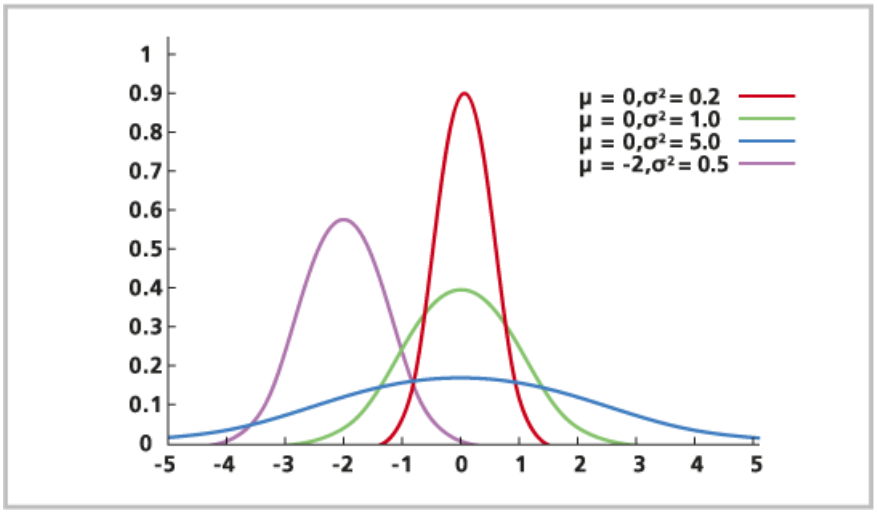
Za kontinuiranu slučajnu varijablu X kažemo da ima normalnu ili Gaussovu razdiobu s parametrima μ i σ^2, pišemo $X \sim N\left(\mu, \sigma^2\right)$, ako je njezina funkcija gustoće zadana formulom:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \tag{1.9}$$

gdje je σ standardna devijacija, σ^2 varijanca, a μ očekivanje. Takvu slučajnu varijablu X zovemo normalna slučajna varijabla. Specijalno, ako je $\mu = 0$ i $\sigma^2 = 1$, normalnu slučajnu varijablu zovemo standardna normalna slučajna varijabla. Želimo li izračunati vjerojatnost da je normalna slučajna varijabla $N(\mu, \sigma^2)$ manja ili jednaka nekom broju a, onda ćemo morati poznavati funkciju distribucije slučajne varijable N, tj. morat ćemo izračunati integral

$$\mathrm{P}(X \leq a) = \int_{-\infty}^{a} f(x)dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{a} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx,$$

što je ne trivijalni integral koji ovisi o tri parametra (a, μ, σ). Umjesto toga, tabelirana je samo funkcija distribucije standardne normalne slučajne varijable $N(0, 1)$, a sve ostale normalne slučajne varijable se jednostavnom transformacijom svode na tu normalnu slučajnu varijablu. Spomenutu transformaciju zovemo standardizacija ($Z = \frac{X-\mu}{\sigma}$).



Slika 1.4: Krivulje normalne razdiobe

Promatrajući sliku 1.4 vidimo da vrh krivulje leži na samoj očekivanoj vrijednosti μ. Krivulja je simetrična s obje strane, a njeni krajevi padaju u zvonoliki oblik te se asimptotski približavaju apscisi, a to znači da se dodiruju u beskonačnosti. Normalna krivulja

ima dvije točke infleksije, koje su od očekivane vrijednosti udaljene za iznos standardne devijacije σ. Stoga je međusobna udaljenost ovih točaka 2σ. Za manje iznose standardne devijacije krivulja je strmija od, primjerice, krivulja s većim iznosom devijacije. Moguće je donijeti zaključak da se širina razdiobe povećava kako se povećava i vrijednost σ.

Gama distribucija

Gama distribucija našla je primjene u različitim područjima. Primjerice, u inženjerstvu se gama distribucija koristi u istraživanju pouzdanosti sustava. Kažemo da slučajna varijabla X ima gama distribuciju s parametrima $\alpha > 0$ i $\beta > 0$, i pišemo $X \sim \Gamma(\alpha, \beta)$, ako je strogo pozitivna ($\text{Im}\, X = \langle 0, +\infty \rangle$) i gustoća razdiobe je

$$f_X(x) = \begin{cases} \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} & \text{za } x > 0 \\ 0 & \text{inače}, \end{cases} \tag{1.10}$$

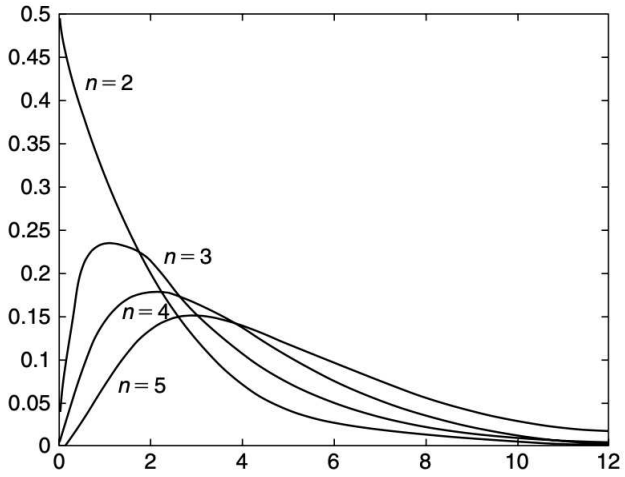
gdje je $\Gamma(\alpha) = \int_0^{+\infty} t^{\alpha-1} e^{-t} dt$ (Γ-funkcija). Za Gama distribuciju vrijedi:

$$\mathbb{E}[X] = \alpha\beta, \quad Var(X) = \alpha\beta^2.$$

Ako je $X \sim \Gamma(1, \beta)$, tada kažemo da X ima eksponencijalnu distribuciju s parametrom β i pišemo $X \sim Exp(1/\beta)$. Može se pokazati da se $\Gamma(k, \beta)$-razdioba, gdje je k prirodan broj, može interpretirati kao zbroj od k nezavisnih $Exp(1/\beta)$-distribuiranih slučajnih varijabli. Drugim riječima, slučajna varijabla s tom gama razdiobom se interpretira kao vrijeme čekanja da se dogodi točno k događaja u Poissonovom procesu s intenzitetom $1/\beta$. Eksponencijalne slučajne varijable često se koriste za modeliranje vijeka trajanja elektroničkih komponenti poput osigurača, za analizu preživljavanja, analizu pouzdanosti te razvoj modela osiguravajućih rizika. Još jedan poseban slučaj gama distribucije koji je posebno koristan u problemima statističke inferencije je distribucija hi-kvadrata. Neka je n prirodan broj. Slučajna varijabla X ima hi-kvadrat distribuciju s n stupnjeva slobode ako i samo ako je X gama distribuirana slučajna varijabla s parametrima $\alpha = n/2$ i $\beta = 2$. Pišemo, $X \sim \chi^2(n)$. Slika 1.5 prikazuje ovisnost hi-kvadrat distribucije o broju stupnjeva slobode n.

Granični teoremi

Granični teoremi igraju vrlo važnu ulogu u proučavanju teorije vjerojatnosti i njenim primjenama. Mnoge slučajne varijable koje susrećemo u prirodi imaju distribucije bliske normalnoj distribuciji te su upravo ta pojednostavljenja modeliranja moguća zbog različitih graničnih teorema. Prvo navodimo Čebiševljev teorem, koji je koristan rezultat za dokazivanje ostalih graničnih teorema. On daje donju granicu za površinu ispod krivulje između dvije točke koje se nalaze na suprotnim stranama srednje vrijednosti i jednako su udaljene



Slika 1.5: Hi-kvadrat funkcije gustoće s različitim stupnjevima slobode

od nje. Snaga ovog rezultata leži u tome što nam nije potrebno poznavati distribuciju osnovne populacije, već samo zahtijeva postojanje srednje vrijednosti i varijance. Dokaz je dan u [1], poglavlje 3.5.

Teorem 1.2.7. ***(Čebiševljev teorem)*** *Pretpostavimo da slučajna varijabla X ima konačnu srednju vrijednost μ i konačnu varijancu σ^2. Tada za svaki $K > 0$ vrijedi*

$$\mathbb{P}\left(|X - \mu| \geq K\right) \leq \frac{\sigma^2}{K^2}.$$

Čebiševljevu nejednakost koristimo u dokazu sljedećeg rezultata, koji se naziva slabim zakonom velikih brojeva. U njemu se tvrdi da ako je veličina uzorka (n) velika, sredina uzroka rijetko odstupa od sredine distribucije X, što se u statistici naziva populacijskom sredinom.

Teorem 1.2.8. ***(Slabi zakon velikih brojeva)*** *Neka su $X_1, \ldots, X_n$ skup međusobno nezavisnih slučajnih varijabli s $E(X_i) = \mu$ i $Var(X_i) = \sigma^2 < \infty$. Tada za svaki $\varepsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| \geq \varepsilon\right) = 0, \tag{1.11}$$

gdje je $S_n = \sum_{i=1}^{n} X_i$.

Dokaz. Neka je $\varepsilon > 0$ proizvoljan. Uočimo da je

$$E\left[\frac{S_n}{n}\right] = \frac{1}{n}\sum_{i=1}^{n} E\left[X_i\right] = \mu,$$

gdje jednakosti slijede iz pretpostavki nezavisnosti i $E(X_i) = \mu, \forall i$. Slično se pokazuje da je $\mathrm{Var}(S_n/n) < \infty$,

$$\mathrm{Var}\left(\frac{S_n}{n}\right) = \frac{1}{n^2}\,\mathrm{Var}(\sum_{i=1}^{n} X_i) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}.$$

Stoga, možemo primijeniti teorem 1.2.7

$$\mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| \geq \varepsilon\right) \leq \frac{\mathrm{Var}\left(\frac{S_n}{n}\right)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2}.$$

Kada n teži u beskonačnost, slijedi 1.11. □

Teorem 1.2.9. ***(Jaki zakon velikih brojeva)*** *Neka su $X_1, \ldots, X_n$ skup nezavisnih i jednako distribuiranih varijabli za koje je $\mathrm{E}[|X_i|] < \infty$ te $\mathrm{E}[X_i] = \mu$, $\forall i \in \{1, 2, \ldots, n\}$. Tada vrijedi*

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{S_n}{n} = \mu\right) = 1, \tag{1.12}$$

odnosno kažemo da S_n/n konvergira gotovo sigurno prema μ.

Navedeni rezultati u osnovi kažu da možemo početi s pokusom čiji se ishod ne može predvidjeti s potpunom sigurnošću i uzimajući prosjek, možemo dobiti pokus čiji se ishod može predvidjeti s visokom točnošću. Zakoni velikih brojeva se široko koriste u primjenama u osiguranju, statistici i proučavanju nasljeđivanja. Jedan od najvažnijih rezultata teorije vjerojatnosti koji je našao primjene u statistici je centralni granični teorem. Na njemu se zasniva statističko zaključivanje, tj.inferencijalna statistika, o populacijskim srednjim vrijednostima i proporcijama na osnovi velikih uzoraka neovisno o populacijskoj distribuciji. Upravo je centralni granični teorem jedan od uzroka važnosti normalne razdiobe u statistici. Detaljan dokaz teorema se nalazi u [8], na stranicama 152 i 153.

Teorem 1.2.10. ***(Centralni granični teorem)*** *Neka je $X_1, X_2, \ldots$ niz nezavisnih jednako distribuiranih slučajnih varijabli s konačnim matematičkim očekivanjem μ i konačnom varijancom $\sigma^2 > 0$. Nadalje, neka je $\overline{X}_n := (X_1 + X_2 + \cdots + X_n)/n$ za sve prirodne brojeve n. Tada za sve $a < b$ vrijedi*

$$\lim_{n\to+\infty} \mathbb{P}\left(a \leq \frac{\overline{X}_n - \mu}{\sigma}\sqrt{n} \leq b\right) = \Phi(b) - \Phi(a), \tag{1.13}$$

gdje je $\Phi(x)$ funkcija distribucije jedinične normalne razdiobe.

Drugim riječima, kažemo da niz slučajnih varijabli $\left(\overline{X}_n - \mu\right)\sqrt{n}/\sigma$ konvergira po distribuciji jediničnoj normalnoj razdiobi kada n teži u beskonačnost, i pišemo

$$\frac{\overline{X}_n - \mu}{\sigma}\sqrt{n} \xrightarrow{D} N(0,1), \quad n \to \infty$$

Slučajna varijabla

$$Z = \frac{\overline{X}_n - \mu}{\sigma}\sqrt{n}$$

je standardizirana verzija od aritmetičke sredine $\overline{X}_n = \overline{X}$. Nadalje, Z je i standardizirana verzija od $\sum_{i=1}^{n} X_i$ jer je

$$\frac{\overline{X} - \mu}{\sigma}\sqrt{n} = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}}.$$

Dakle, centralni granični teorem kaže da standardizirana verzija aritmetičke sredine, odnosno zbroja, od n nezavisnih jednako distribuiranih slučajnih varijabli s konačnom nenul varijancom ima aproksimativno jediničnu normalnu razdiobu za velike n. Prirodno se postavlja pitanje koliko n mora biti velik da bi aproksimacija normalnom razdiobom bila zadovoljavajuća. Prema [2], poglavlje 5 slijedi: "Obično se uzima $n \geq 30$, ali potpuni odgovor bi glasio da veličina od n ovisi o obliku razdiobe slučajnih varijabli X_i, preciznije, je li simetrična, a ako nije, koliko je asimetrična. Ako je razdioba od X_i približno simetrična, tada i $n = 10$ može biti dovoljno velik, a ako je distribucija znatno asimetrična, tada se mora uzeti barem $n = 50$".

1.3 Uzoračke razdiobe

Informacije o populacijskoj razdiobi varijable koju izučavamo dobivamo iz uzorka uzetog iz te populacije. Na primjer, procjenu populacijske srednje vrijednosti, odnosno utvrđivanje istinitosti hipoteza o populacijskoj razdiobi donosimo na osnovi vrijednosti iz uzorka. Stoga, od interesa nam je razdioba statistike izračunate iz slučajnog uzorka. Budući da je uzorak skup slučajnih varijabli $X_1, \ldots, X_n$, slijedi da je statistika kao funkcija uzorka također slučajna. Vjerojatnosnu razdiobu takve statistike nazivamo uzoračka razdioba (distribucija). Uzoračke razdiobe pružaju vezu između teorije vjerojatnosti i statističkog zaključivanja. Mogućnost određivanja razdiobe statistike kritični je dio u konstrukciji i evaluaciji statističkih postupaka [1]. Važno je primijetiti da postoji razlika između distribucije populacije iz koje je uzorak uzet i uzoračke distribucije. Općenito, populacija ima distribuciju koja se naziva populacijskom distribucijom i obično je nepoznata, dok statistika ima uzoračku distribuciju koja se obično razlikuje od populacijske distribucije. Uzoračka razdioba statistike pruža teorijski model relativnog frekvencijskog histograma za moguće vrijednosti statistike koje bismo promatrali opetovanim uzorkovanjem. Sada predstavljamo, već poznate, definicije u terminima slučajnih varijabli slijedeći poglavlje 4 u [1].

Definicija 1.3.1. ***Uzorak*** *je skup slučajnih varijabli* $X_1, \ldots, X_n$.

Definicija 1.3.2. ***Slučajni uzorak*** *je skup nezavisnih jednako distribuiranih slučajnih varijabli $X_1, \ldots, X_n$.*

Definicija 1.3.3. ***Statistika*** *je funkcija slučajnog uzorka koja ne sadrži nepoznate parametre.*

Na primjer, dvije statistike su uzoračka sredina i uzoračka varijanca

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X} \right)^2 .$$

Primijetimo da je zbog centralnog graničnog teorema uzoračka razdioba uzoračke sredine $\overline{X}$ za velike uzorke aproksimativno normalna, bez obzira na populacijsku razdiobu izučavane varijable X, uz jedini uvjet da je populacijska varijanca konačna i nije jednaka nuli.

Definicija 1.3.4. ***Uzoračka razdioba*** *je vjerojatnosna razdioba statistike.*

Uzoračke distribucije statistika normalnog uzorka

Uzoračka razdioba neke statistike ovisi o distribuciji populacije iz koje su uzorci uzeti. U ovom potpoglavlju pratimo [1], poglavlje 4.2, i navodimo uzoračke razdiobe nekih statistika koje se temelje na slučajnom uzorku iz normalne distribucije. Te statistike se koriste u mnogim statističkim postupcima koji su vrlo važni u rješavanju problema u praksi. Sljedeći rezultat uspostavlja distribuciju linearnih kombinacija nezavisnih normalnih slučajnih varijabli.

Teorem 1.3.5. *Neka su $X_1, \ldots, X_n$ nezavisne slučajne varijable gdje je svaka slučajna varijabla X_i normalno distribuirana sa očekivanjem μ_i i varijancom σ_i^2 te neka su $a_1, \ldots, a_n$ realni brojevi. Tada je uzoračka distribucija od $Y = \sum_{i=1}^{n} a_i X_i$ normalna sa parametrima očekivanja $\mu_Y = \sum_{i=1}^{n} a_i \mu_i$ i varijance $\sigma_Y^2 = \sum_{i=1}^{n} a_i^2 \sigma_i^2$.*

Dokaz je dan u [1] na stranici 191. Ako su u teoremu 1.3.5 $a_i = 1/n$, $\mu_i = \mu$ i $\sigma_i^2 = \sigma^2$, dobivamo sljedeći rezultat, koji daje distribuciju uzoračke sredine.

Korolar 1.3.6. *Neka je $X_1, \ldots, X_n$ slučajni uzorak duljine n iz populacije s parametrima očekivanja μ i varijance σ^2. Tada je*

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

normalno distribuirana sa očekivanjem $\mu_{\overline{X}} = \mu$ i varijancom $\sigma^2_{\overline{X}} = \sigma^2/n$.

Iz korolara 1.3.6 vidimo da $\overline{X}$ ima normalnu uzoračku razdiobu, $\overline{X} \sim N(\mu, \sigma^2/n)$, prema tome standardizirana verzija Z od $\overline{X}$ ima jediničnu normalnu razdiobu

$$Z = \frac{\overline{X} - \mu}{\sigma} \sqrt{n} \sim N(0, 1). \tag{1.14}$$

Usporedimo li ovaj zaključak sa teoremom 1.2.10 vidimo da, u slučaju normalno distribuiranog uzorka, konvergencija standardiziranih verzija aritmetičkih sredina prelazi u identitet.

Sada ćemo navesti neke distribucije koje se mogu izvesti iz normalne distribucije. One igraju vrlo važnu ulogu u inferencijalnim problemima.

Hi-kvadrat distribucija

Hi-kvadrat razdioba koristi se u inferencijalnim problemima koji se bave varijancom. Slijedeći teoremi su dokazani u [1], poglavlje 4.2.1 *Chi-Square Distribution*.

Teorem 1.3.7. *Neka su $X_1, \ldots, X_k$ nezavisne χ^2 slučajne varijable s $n_1, \ldots, n_k$ stupnjeva slobode, respektivno. Tada je suma $V = \sum_{i=1}^{k} X_i$ hi-kvadrat distribuirana s $n_1 + n_2 + \cdots + n_k$ stupnjeva slobode.*

Teorem 1.3.8. *Ako je X standardizirana normalna slučajna varijabla, tada je X^2 hi-kvadrat slučajna varijabla s jednim stupnjem slobode.*

Teorem 1.3.9. *Neka je slučajni uzorak $X_1, \ldots, X_n$ iz normalne populacije $N(\mu, \sigma^2)$. Tada su $Z_i = (X_i - \mu)/\sigma$, $i = 1, \ldots, n$ nezavisne jedinične normalne slučajne varijable i*

$$\sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 \tag{1.15}$$

ima χ^2-distribuciju s n stupnjeva slobode. Specijalno, ako su $X_1, \ldots, X_n$ nezavisne jedinične normalne slučajne varijable, tada $Y^2 = \sum_{i=1}^{n} X_i^2$ ima hi-kvadrat distribuciju s n stupnjeva slobode.

Studentova *t*-distribucija

Ako su slučajne varijable $X_1, \ldots, X_n$ normalno distribuirane s očekivanjem μ i varijancom σ^2. Ako je σ poznata, onda znamo da je $\sqrt{n}((\overline{X} - \mu)/\sigma)$ ima $N(0, 1)$ distribuciju. Međutim, ako σ nije poznata, kao što je obično slučaj, tada se rutinski zamjenjuje uzoračkom standardnom devijacijom S. Ako je veličina uzorka velika, mogli bismo

pretpostaviti da je $S \approx \sigma$ pa primjenom centralnog graničnog teorema dobivamo da je $\sqrt{n}((\overline{X} - \mu)/S)$ približno $N(0,1)$ distribuirana. No, ako je slučajni uzorak mali, tada je distribucija $\sqrt{n}((\overline{X} - \mu)/S)$ dana s tzv. Studentovom t-distribucijom, ili jednostavno t-distribucijom. Preciznije, ako su Y i Z dvije nezavisne slučajne varijable, $Y \sim \chi^2(n)$ i $Z \sim N(0,1)$, onda za

$$T = \frac{Z}{\sqrt{Y/n}} \tag{1.16}$$

kažemo da ima (Studentovu) t-distribuciju s n stupnjeva slobode.

Teorem 1.3.10. *Ako su $\overline{X}$ i S^2 srednja vrijednost i varijanca slučajnog uzorka veličine n iz normalne populacije s očekivanjem μ i varijancom σ^2, onda*

$$T = \frac{\overline{X} - \mu}{s/\sqrt{n}}$$

ima t-distribuciju s $(n-1)$ stupnjeva slobode.

Dokaz se nalazi u [1] pod dokazom teorema 4.2.9.

Fisherova F-distribucija

F-distribuciju je razvio R. Fisher za proučavanje ponašanja varijanci slučajnih uzoraka uzetih iz dviju nezavisnih normalnih populacija. U praktičnim problemima može nas zanimati jesu li populacijske varijance jednake ili ne, na temelju slučajnih uzoraka. Poznavanje odgovora na takvo pitanje također je važno pri odabiru odgovarajućih statističkih metoda za proučavanje populacijskih stvarnih srednjih vrijednosti. Ako su U i V dvije nezavisne slučajne varijable, $U \sim \chi^2(n_1)$ i $V \sim \chi^2(n_2)$, tada varijabla

$$F = \frac{U/n_1}{V/n_2}$$

ima Fisherovu ili F-razdiobu s (n_1, n_2) stupnjeva slobode. Pišemo, $F \sim F(n_1, n_2)$.

Teorem 1.3.11. *Neka su S_1^2 i S_2^2 uzoračke varijance dvaju nezavisnih uzoraka duljine n_1, odnosno n_2, iz normalno distribuiranih populacija s varijancama σ_1^2, odnosno σ_2^2. Tada vrijedi:*

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1 - 1, n_2 - 1). \tag{1.17}$$

Teorem se dokazuje u [1] na stranicama 203 i 204.

Poglavlje 2

Linearna regresija

U ovom poglavlju, proučavamo odnos između varijabli pomoću analize regresije. Cilj nam je stvoriti model koji se može koristiti u svrhe predviđanja i proučavati inferencijalne postupke kada je prisutna jedna ovisna i nekoliko nezavisnih varijabli. Preciznije, zanima nas distribucija neke istaknute varijable Y, koje se naziva ovisna ili odzivna varijabla, u ovisnosti o vrijednostima koje poprimaju druge, nezavisne ili prediktorske varijable $X_1, \ldots, X_k$. Glavna značajka regresijske analize je sposobnost davanja izjava o varijablama nakon provedenih kontroliranih promjena poznatih prediktorskih varijabli. Na primjer, neka (x, y) označava visinu i težinu odrasle osobe. Naše bi zanimanje moglo biti pronaći vezu između visine i težine iz uzoraka mjerenja n pojedinaca. Postupak pronalaženja matematičke jednadžbe koja najbolje odgovara šumovitim podacima poznat je kao analiza regresije [1].

Ako pretpostavimo da $Var(Y \mid X_1 = x_1, \ldots, X_k = x_k)$ ne ovisi o vrijednostima prediktora $x_1, \ldots, x_k$, onda možemo zapisati

$$Y = f(X_1, \ldots, X_k) + \varepsilon,$$

pri čemu je izraz $f(X_1, \ldots, X_k)$ očekivanje odaziva Y u ovisnosti o vrijednostima prediktora $X_1, \ldots, X_k$, a ε je slučajna varijabla s očekivanjem 0 koja opisuje grešku, tj. odstupanje od očekivanja. Normalna distribucija greške je imala ključnu ulogu u razvoju regresijske analize i najčešća je pretpostavka, ali od interesa je proširiti ideje regresije i na druge modele podataka. Stoga se Y može smatrati da ima određenu determinističku komponentu, $\mathbb{E}(Y)$, i slučajnu komponentu, ε. Prilikom modeliranja odziva Y, nužno je uzeti u obzir oblik funkcije f, tj. pretpostavljamo da f pripada nekom skupu funkcija $\mathcal{F}$. Ako bismo dopustili da f bude proizvoljan, došlo bi do prevelike prilagodbe modela. Točnije, regresijska funkcija bi savršeno prolazila kroz sve trening podatke, no to bi rezultiralo velikom pogreškom na testnim podacima. Postoje različite forme regresije: jednostavna linearna, nelinearna, višestruka i druge. Jedan od najjednostavnijih modela $\mathcal{F}$ je linearni model gdje

je očekivanje odaziva neka linearna kombinacija prediktorskih varijabli:

$$\mathbb{E}[Y] = f(X_1, \ldots, X_k) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k.$$

Takav model nazivamo višestrukim linearnim regresijskim modelom, precizna definicija dana je u [1] kao Definicija 8.2.1.

Razlozi za primjenu linearne regresije su:

- jednostavna prilagodba modela metodom najmanjih kvadrata,
- jednostavna interpretacija veze odzivne i prediktorskih varijabli,
- jednostavno računanje vrijednosti raznih procjenitelja i njihovih distribucija,
- mogućnost konstrukcije intervala pouzdanosti za procijenjene parametre modela,
- mogućnost predviđanja budućih vrijednosti odaziva te konstrukciju predikcijskih intervala za njih.

2.1 Jednostavna linearna regresija

Kako bismo razumjeli osnovne koncepte regresijske analize, razmotrimo model koji se sastoji od odzivne varijable Y opisane samo jednom prediktorskom varijablom X. Takav model, oblika

$$Y = \beta_0 + \beta_1 X + \varepsilon, \tag{2.1}$$

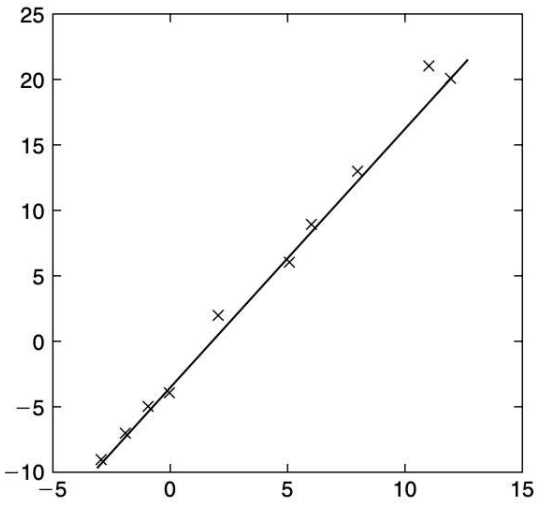
nazivamo jednostavnim linearnim regresijskim modelom, gdje je β_0 presjek pravca sa y-osi, β_1 nagib pravca te ε greška. Ovaj osnovni linearni model pretpostavlja postojanje linearnog odnosa između varijabli X i Y uz prisutnost raspršenosti.

Jednostavna linearna regresija ima prednost što se može lako prikazati na dvodimenzionalnom grafu. Na ilustraciji 2.1, preuzetoj iz [1] sa stranice 414, vidimo da je problem jednostavne linearne regresije zapravo optimalna prilagodba pravca danom skupu podataka, odnosno pronalaženje "najboljih" procjenitelja za β_0 i β_1.

Metoda najmanjih kvadrata

U ovom potpoglavlju opisujemo najčešće korištenu tehniku za određivanje regresijske krivulje zvanu metoda najmanjih kvadrata. Neka su $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ n opažanja s odgovarajućim pogreškama ε_i, $i = 1, 2, \ldots, n$. Drugim riječima,

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \ldots, n.$$



Slika 2.1: Regresijski pravac prilagođen metodom najmanjih kvadrata

Pretpostavimo da su greške ε_i, $i = 1, 2, \ldots, n$ nezavisne i jednako distribuirane s očekivanjem 0 i varijancom σ^2. Jedan od načina za ispitivanje koliko dobro pravac opisuje skup podataka jest određivanje u kojoj mjeri podaci odstupaju od pravca. Uz navedene pretpostavke, za dani x vrijedi da je očekivana vrijednost od Y jednaka $\mathbb{E}(Y) = \beta_0 + \beta_1 x$. Stoga, traženi regresijski pravac je procijenjena od $\mathbb{E}(Y)$ i dan je s

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x,$$

gdje su $\hat{\beta}_0$ i $\hat{\beta}_1$ procjenitelji parametara β_0 i β_1, respektivno. Slijedi da se, za neko opažanje (x_i, y_i), procijenjena vrijednost od y_i dobiva kao

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i. \tag{2.2}$$

Odstupanje između opažene vrijednosti y_i i njenog predviđanja $\hat{y}_i$ zovemo i-ti rezidual i definiramo kao

$$e_i = (y_i - \hat{y}_i) = \left[y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i \right) \right].$$

Primijetimo da su reziduali zapravo udaljenosti između opaženih i predviđenih vrijednosti na osi y. U metodi najmanjih kvadrata cilj nam je prilagoditi regresijski pravac, tj. pronaći $\hat{\beta}_0$ i $\hat{\beta}_1$, tako da je suma kvadrata reziduala minimalna za dani skup podataka. U tu svrhu definiramo funkciju koja određuje kvalitetu prilagodbe,

$$\text{SSE} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left[y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i \right) \right]^2. \tag{2.3}$$

Dakle, u metodi najmanjih kvadrata, trebamo odrediti parametre $\hat{\beta}_0$ i $\hat{\beta}_1$ tako da je SSE minimum.

Propozicija 2.1.1. *Metodom najmanjih kvadrata slijedi da su procjenitelji za koeficijente β_0 i β_1 dani s*

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2} \\ \hat{\beta}_0 &= \overline{y}-\hat{\beta}_1\overline{x}, \end{aligned} \tag{2.4}$$

gdje je $\overline{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$ i $\overline{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$.

Dokaz. Ako SSE doseže minimum, tada su parcijalne derivacije SSE u odnosu na β_0 i β_1 jednake nuli:

$$\begin{cases} \frac{\partial \mathrm{SSE}}{\partial \beta_0}=0 \\ \frac{\partial \mathrm{SSE}}{\partial \beta_1}=0 \end{cases} \Leftrightarrow \begin{cases} \sum_{i=1}^{n}-2\left[y_i-(\beta_0+\beta_1 x_i)\right]=0 \\ \sum_{i=1}^{n}-2x_i\left[y_i-(\beta_0+\beta_1 x_i)\right]=0 \end{cases}$$

$$\Leftrightarrow \begin{cases} \sum_{i=1}^{n}\left[y_i-(\beta_0+\beta_1 x_i)\right]=0 \\ \sum_{i=1}^{n}x_i\left[y_i-(\beta_0+\beta_1 x_i)\right]=0 \end{cases}$$

rješavanjem dobivenih jednadžbi dobivamo:

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2} \\ \hat{\beta}_0 &= \overline{y}-\hat{\beta}_1\overline{x}. \end{aligned}$$

Nismo sigurni postiže li funkcija $\mathrm{SSE}(\beta_0,\beta_1)$ u dobivenom ekstremu maksimalnu ili minimalnu vrijednost, stoga provjeravamo dovoljan uvjet ekstrema koji kaže da diferencijabilna funkcija $\mathrm{SSE}\left(\hat{\beta}_0,\hat{\beta}_1\right)$ postiže minimum u točki ekstrema T ako vrijedi da su determinante minora Hessove matrice pozitivne, odnosno

$$\frac{\partial^2 \mathrm{SSE}}{\partial \beta_0^2}\cdot\frac{\partial^2 \mathrm{SSE}}{\partial \beta_1^2}-\left(\frac{\partial^2 \mathrm{SSE}}{\partial \beta_0 \partial \beta_1}\right)^2>0 \quad \text{i} \quad \frac{\partial^2 \mathrm{SSE}}{\partial \beta_0^2}>0.$$

Izračunom druge parcijalne derivacije dobivamo

$$\frac{\partial\,\mathrm{SSE}^2}{\partial^2\beta_0}(\beta_0,\beta_1)=2n, \quad \frac{\partial\,\mathrm{SSE}^2}{\partial^2\beta_1}(\beta_0,\beta_1)=2\sum_{i=1}^{n}x_i^2, \quad \frac{\partial\,\mathrm{SSE}^2}{\partial\beta_0\partial\beta_1}(\beta_0,\beta_1)=2\sum_{i=1}^{n}x_i.$$

Uvrštavanjem dobivamo da su obje determinante pozitivne te zaključujemo da funkcija $\mathrm{SSE}(\beta_0,\beta_1)$ postiže minimum. □

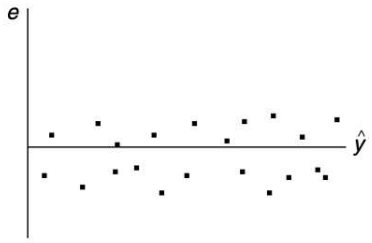
Možemo se zapitati zašto smo odabrali metodu najmanjih kvadrata za minimalizaciju udaljenosti podataka od regresijskog pravca. Razlog tome je što je funkcija SSE diferencijabilna, što znači da rješenje postoji, jedinstveno je i može se izraziti u zatvorenoj formi. S druge strane, postoji i metoda L_1-regresije koja promatra sumu apsolutnih vrijednosti reziduala, ali apsolutna vrijednost nije neprekidno diferencijabilna funkcija. Također, metoda najmanjih kvadrata je popularna jer postoje dokazi o njezinoj optimalnosti te, ako pretpostavimo normalnu distribuciju greške ε_i, možemo odrediti točne distribucije procijenjenih koeficijenata i raznih testnih statistika.

Pouzdanost regresije

Nakon što odredimo linearni model, prirodno se nameće pitanje: koliko dobro pravac odgovara podacima? Jedan od načina procijene pouzdanosti modela je pomoću reziduala

$$\hat{e}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i.$$

Što je model bolji, to bi rezidual $\hat{e}_i$ trebao biti "bliži" slučajnoj grešci ε s očekivanjem 0. Nadalje, ako prikažemo reziduale u odnosu na nezavisnu varijablu na x-osi, oni ne bi trebali pokazivati prepoznatljivi obrazac. Idealno bi to trebalo izgledati kao horizontalna mutna slika, kao što je prikazano na slici 2.2 preuzetoj iz [1] sa stranice 421. Bilo kakav simetrični trend na grafu ukazuje na nepouzdani regresijski model.



Slika 2.2: Vidimo reziduali ne pokazuju nikakav odnos prema vrijednostima x

Dok nam grafovi reziduala daju vizualni prikaz kvalitete prilagodbe, numerička mjera pouzdanosti regresije dobiva se računanjem koeficijenta determinacije. Koeficijent determinacije, oznake R^2, mjeri za koliki dio sveukupne varijacije je odgovorna regresijska funkcija, tj. proporcija varijance odzivne varijable koja je objašnjena prediktorskom varijablom. Kada bismo željeli predvidjeti vrijednost ovisne varijable bez ikakvog znanja o nezavisnoj varijabli, najbolje bi bilo koristiti prosjek mjerenja ovisne varijable. Međutim, s obzirom na to da imamo poznate vrijednosti nezavisne varijable, naša predikcija može

biti značajno preciznija. U slučaju kada nemamo prediktorsku varijablu promatramo kvadrat udaljenosti svakog mjerenja ovisne varijable od srednje vrijednosti svih mjerenja, tj. $(y_i - \overline{y})^2$, dok u suprotnom promatramo kvadratno odstupanje od regresijskog pravca $(y_i - \hat{y}_i)^2$. Koeficijent determinacije je dan s

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y}_i)^2} \in [0, 1]\,.$$

Maksimalna vrijednost koju koeficijent determinacije može postići je 1 i ona predstavlja slučaj kada svi podaci leže na regresijskom pravcu. Drugi ekstrem je 0, a to se događa kada je regresijski pravac dan s $y = \overline{y}$, tj. kada nema prediktorskih varijabli koje opisuju ovisnu varijablu. Što je R^2 bliže 1, to se veći postotak raspršenosti može objasniti pomoću prediktorske varijable. Ne postoje općenite smjernice o tome koliko minimalno koeficijent determinacije treba biti kako bi regresija bila pouzdana, niti postoje testovi za R^2. Stoga, moramo imati na umu da veličina R^2 ne znači nužno da je prilagođeni model dobar ili loš. Primjerice, ako je varijanca grešaka σ^2 velika, veći je i udio varijance odaziva koji ne možemo objasniti uz pomoću regresije te je stoga i R^2 manji.

Svojstva regresijskih procjenitelja

Definicija 2.1.2. *Procjenitelj $\hat{\theta}$ je nepristrani procjenitelj za parametar θ ako vrijedi*

$$E(\hat{\theta}) = \theta.$$

Definicija 2.1.3. *Ako je procjenitelj $\hat{\theta}$ linearna kombinacija uzoraka i ima varijancu koja je manja ili jednaka varijanci od bilo kojeg drugog procjenitelja koji je također linearna kombinacija uzoraka, tada se $\hat{\theta}$ naziva najboljim linearnim nepristranim procjeniteljem (eng. BLUE - best linear unbiased estimate) za θ.*

U ovom potpoglavlju govorimo o svojevrsnoj optimalnosti metode najmanjih kvadrata. Utvrditi ćemo pretpostavke pod kojima metoda najmanjih kvadrata daje najtočnije rezultate. Postoje četiri pretpostavke poznate kao Gauss-Markovljevi uvjeti. Važno je napomenuti da ove pretpostavke neće uvijek biti u potpunosti ispunjene u stvarnom svijetu zbog nesavršene distribucije podataka. Unatoč tome, model linearne regresije s idealnim uvjetima može se koristiti kao referenca za usporedbu s drugim stvarnijim modelima.

Definicija 2.1.4. ***Gauss-Markovljevi uvjeti*** *za slučajne greške $\varepsilon_i, i = 1, 2, \ldots, n$ su:*

1. $E(\varepsilon_i) = 0, i = 1, 2, \ldots, n$
2. $\varepsilon_i \sim N(0, \sigma^2)$
3. $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$ *za* $i \neq j, i, j = 1, 2, \ldots, n$

4. $\text{Var}(\varepsilon_i) = \sigma^2, i = 1, 2, \ldots, n$

Teorem 2.1.5. ***(Gauss - Markovljev teorem)*** *Neka je $Y = \beta_0 + \beta_1 x + \varepsilon$ jednostavni linearni regresijski model. Ako vrijede Gauss-Markovljevi uvjeti 2.1.4, tada su procjenitelji metode najmanjih kvadrata, $\hat{\beta}_0$ i $\hat{\beta}_1$, najbolji linearni nepristrani procjenitelji.*

Dokaz teorema je u [9]. Nadalje, poznavanje distribucija procjenitelja najmanjih kvadrata $\hat{\beta}_0$ i $\hat{\beta}_1$ nužno je za donošenje statističkih zaključaka o njima. Teorem u nastavku daje uzoračku razdiobu procjenitelja najmanjih kvadrata, dokaz tog rezultata je dan u [1], Teorem 8.2.1.

Teorem 2.1.6. *Neka je $Y = \beta_0 + \beta_1 x + \varepsilon$ jednostavni linearni regresijski model s $\varepsilon \sim N(0, \sigma^2)$ i neka su greške ε_i međusobno nezavisne. Tada*

(a) $\hat{\beta}_0$ i $\hat{\beta}_1$ imaju normalne distribucije.

(b) Očekivanje i varijanca su jednake

$$\mathbb{E}\left(\hat{\beta}_0\right) = \beta_0, \quad Var\left(\hat{\beta}_0\right) = \left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)\sigma^2$$

i

$$\mathbb{E}\left(\hat{\beta}_1\right) = \beta_1, \quad Var\left(\hat{\beta}_1\right) = \frac{\sigma^2}{S_{xx}},$$

gdje je $S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2$. Posebno, procjenitelji najmanjih kvadrata $\hat{\beta}_0$ i $\hat{\beta}_1$ su nepristrani procjenitelji za β_0 i β_1, respektivno.

Procjena varijance grešaka

Osim regresijskih koeficijenata, potrebno je procijeniti i varijancu greške kako bi se model u potpunosti procijenio. Varijanca greške potrebna je za sve testove i intervale pouzdanosti. Iz teorema 2.1.6 vidimo da što je varijanca slučajne greške ε veća, to će greške u procjeni parametara modela β_0 i β_1 biti veće.

Teorem 2.1.7. *Ukoliko vrijede Gauss-Markovljevi uvjeti 2.1.4, statistika*

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2.5}$$

je nepristrani procjenitelj parametra zajedničke varijance σ^2.

U dokazu se koriste tvrdnje iz 2.2, 2.4 i 2.1.6, a cijeli dokaz je dan u [11]. Ideja je da ako su koeficijenti β_0 i β_1 dobro procijenjeni, tada će reziduali otprilike odgovarati slučajnim greškama ε_i.

2.2 Intervali pouzdanosti regresijskih parametara

Procjenitelji $\hat{\beta}_0$ i $\hat{\beta}_1$ su izračunati iz uzorka, što povlači da se u slučaju drugačijeg uzorka mijenja procjena. Stoga, su nam od interesa intervali pouzdanosti parametara nagiba β_1 i presjeka β_0 linearnog regresijskog modela.

Iz teorema 2.1.6 slijedi,

$$Z_1 = \frac{\hat{\beta}_1 - \beta_1}{\frac{\sigma}{\sqrt{S_{xx}}}} \sim N(0,1).$$

Također, može se pokazati da je SSE/σ^2 nezavisna od $\hat{\beta}_1$ i ima hi-kvadrat distribuciju s $n-2$ stupnja slobode. Tada iz 1.16 imamo

$$t_{\beta_1} = \frac{Z_1}{\sqrt{\frac{\left(\frac{SSE}{\sigma^2}\right)}{n-2}}} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{MSE}{S_{xx}}}},$$

koja ima t-distribuciju s $n-2$ stupnja slobode i gdje je $\mathrm{MSE} = \mathrm{SSE}/(n-2)$. Slijedi da je $(1-\alpha)100\%$ pouzdani interval za β_1 dan s

$$\left(\hat{\beta}_1 - t_{\alpha/2,n-2}\sqrt{\frac{MSE}{S_{xx}}}, \hat{\beta}_1 + t_{\alpha/2,n-2}\sqrt{\frac{MSE}{S_{xx}}}\right), \tag{2.6}$$

pri čemu $t_{\alpha/2,n-2}$ predstavlja $(1-\alpha)\%$ kvantil Studentove t-distribucije s $n-2$ stupnja slobode. Slično,

$$Z_0 = \frac{\hat{\beta}_0 - \beta_0}{\sigma\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{yy}}\right)} \sim N(0,1)$$

i

$$t_{\beta_0} = \frac{Z_0}{\sqrt{\frac{\frac{SSE}{\sigma^2}}{n-2}}}$$

pa slijedi da je $(1-\alpha)100\%$ pouzdani interval za β_0 dan s

$$\left(\hat{\beta}_0 - t_{\alpha/2,n-2}\left[MSE\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)\right]^2, \hat{\beta}_0 + t_{\alpha/2,n-2}\left[MSE\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)\right]^2\right). \tag{2.7}$$

Sada uvodimo statističko testiranje hipoteza, fokusirati ćemo se samo na obostrani test. Kako bismo saznali da li je razumno da vrijednost $b \in \mathbb{R}$ bude nagib pravca, provodimo

testiranje:

$$H_0 : \beta_1 = b$$
$$H_1 : \beta_1 \neq b$$

Testna statistika i njena nul-distribucija dane su s

$$T_{H_0} = \frac{\hat{\beta}_1 - b}{\hat{\sigma}_{\hat{\beta}_1}} \sim t_{n-2}.$$

Kako bismo odredili područje prihvaćanja i odbijanja hipoteze, koristimo nul-distribuciju testne statistike, što je Studentova t-distribucija s $n - 2$ stupnja slobode, te promotrimo p-vrijednost kako bi odlučili prihvaćamo li ili odbijamo nul-hipotezu. Lako se pokaže da je regija prihvaćanja jednaka gore navedenom intervalu pouzdanosti, stoga testiranje provodimo tako da provjerimo nalazi li se ta vrijednost unutar intervala pouzdanosti. Hipotezom $H_0 : \beta_1 = 0$ testiramo značajnost regresije. Ako je $\beta_1 = 0$, zaključujemo da nema značajne linearnu povezanost između X i Y, dakle nezavisna varijabla X nije važna za predviđanje vrijednosti Y. Drugim riječima, pitanje važnosti nezavisna varijable u regresijskom modelu prevedeno je u uže pitanje testa hipoteze $H_0 : \beta_1 = 0$.

Testovi značajnosti odsječka β_0 nisu od velikog praktičnog značaja. Naime, u praksi je pravilo da se regresija ne provodi bez slobodnog člana. Čak, i u slučaju kada nul-hipoteza nije odbijena te postoji mogućnost da je slobodan član jednak nuli, ostavljamo ga u modelu. Uklanjanje odsječka iz modela predstavlja restrikciju koja zahtijeva da regresijski pravac prolazi kroz ishodište, što može dovesti do loše prilagodbe modela. Ovo je posebno često u slučajevima kada su vrijednosti prediktora X u trening podacima daleko od nule te se javlja nelinearna ovisnost između odziva Y i prediktora X, iako linearni model možda i dobro lokalno opisuje trening podatke. Primjeri testiranja hipoteza za oba parametra nalaze se u [1] od stranice 430 do 433.

2.3 Intervali pouzdanosti predviđanja

Cilj linearne regresije je predvidjeti vrijednost odaziva Y na temelju dane vrijednosti prediktorske varijable X. Budući da je $\mathbb{E}[Y|X = x] = \beta_0 + \beta_1 x$, možemo procijeniti vrijednost odaziva u točki x kao

$$\hat{y}(x) = \hat{\beta}_0 + \hat{\beta}_1 x.$$

Jedino je smisleno predvidjeti vrijednosti odaziva unutar domene prediktorskih varijabli trening skupa podataka, to jest, onih koje smo koristili za prilagođavanje modela, što se naziva interpolacija. S druge strane, ekstrapolacija podrazumijeva predviđanje izvan raspona vrijednosti trening skupa, što je rizično jer nemamo sigurnost da će linearan model biti primjenjiv tamo.

Interval pouzdanosti odaziva

Regresijski koeficijenti su slučajne varijable pa je samim time i regresijski pravac $\beta_0 + \beta_1 x$ slučajna varijabla. Pomoću rezultata dobivenih u 2.2 definiramo pouzdani interval za odaziv.

Definicija 2.3.1. *Za fiksni $x = x_0$ vrijedi da je $(1-\alpha)100\%$ pouzdani interval za $\beta_0 + \beta_1 x$ dan s*

$$\left(\hat{\beta}_0 + \hat{\beta}_1 x_0\right) \pm t_{\alpha/2} s_e \sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}}, \tag{2.8}$$

gdje je

$$s_e = \sqrt{\frac{S_{yy} - (S_{xy})^2}{(n-2)S_{xx}}}.$$

Interval pouzdanosti predikcije

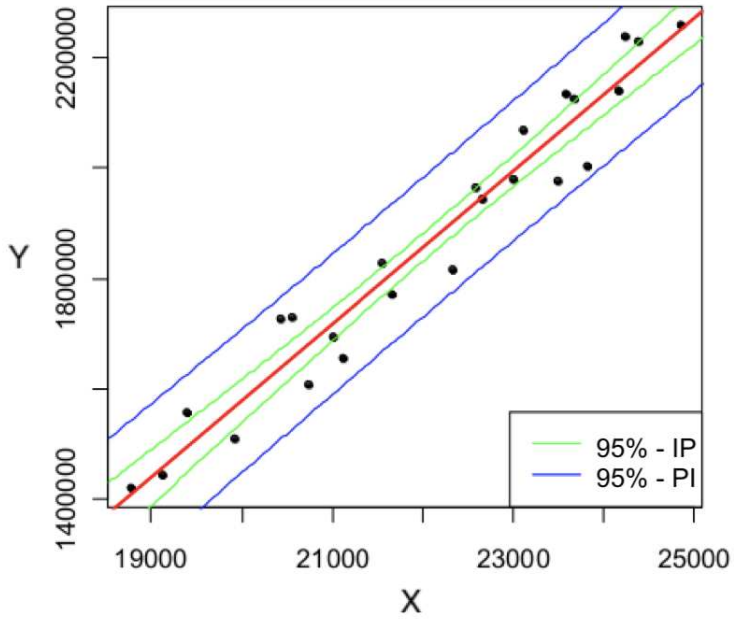
Interval 2.8 nam govori gdje se nalazi $\mathbb{E}[Y|X = x]$, no to nije valjani interval pouzdanosti za sam odaziv $Y|X = x$ iz razloga što je promatrana vrijednost Y dodatno raspršena oko regresijskog pravca radi greške ε. U [1], na stranicama 437 i 438, je pokazano da greška predviđanja određene vrijednosti od Y, uz dani x, ima normalnu distribuciju sa očekivanjem nula i varijancom $\left[1 + \frac{1}{n} + \frac{(x-\overline{x})^2}{S_{xx}}\right]\sigma^2$.

Definicija 2.3.2. *Interval u kojem bi se trebala nalaziti vrijednost odaziva Y je*

$$\left(\hat{\beta}_0 + \hat{\beta}_1 x\right) \pm t_{\alpha/2} S \sqrt{1 + \frac{1}{n} + \frac{(x-\overline{x})^2}{S_{xx}}}, \tag{2.9}$$

gdje je $t_{\alpha/2}$ $(1-\alpha)\%$ kvantil Studentove t-distribucije s $(n-2)$ stupnja slobode i $S^2 = \dfrac{\text{SSE}}{n-2}$. Taj interval se naziva $(1-\alpha)100\%$ predikcijski interval.

Graf 2.3 prikazuje odnos intervala pouzdanosti prilagođene vrijednosti i predikcijskog intervala. Područje unutar zelenih linija označava interval pouzdanosti regresijskog pravca, odnosno područje unutar kojeg se otprilike nalazi točan pravac regresije. Plave linije označavaju predikcijski interval budućih opažanja, tj. područje u kojem se otprilike mogu očekivati budući podaci. Važno je primijetiti kako je predikcijski interval širi zbog dodatne nesigurnosti uzrokovane greškom ε.



Slika 2.3: Ilustracija preuzeta iz izvora [10].

2.4 Korelacija

Regresijskim modelom možemo procijeniti veličinu promjene ovisne varijable zbog određenih promjena nezavisnih varijabli. Nakon što utvrdimo postojanje povezanosti između varijabli, zanima nas koliko je ta povezanost jaka. Linearna povezanost između dviju slučajnih varijabli naziva se korelacijom, snagu te veze opisujemo koeficijentom korelacije ρ. Neka su X i Y slučajne varijable, koeficijent korelacije je dan s $\rho = \sigma_{XY}/(\sigma_X\sigma_Y)$. Primijetimo, ρ je pozitivan ako i samo ako X i Y rastu zajedno (povećava se nagib), dok je ρ negativan ako i samo ako Y opada s rastom X-a (opadajući nagib). U slučaju kada je $\rho = 0$, nema veze između X-a i Y-a. Zaključujemo, koeficijent korelacije može se koristiti za utvrđivanje koliko dobro linearni regresijski model odgovara podacima.

Neka je $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ slučajni uzorak iz bivarijantne normalne distribucije. Uzorački korelacijski koeficijent r je procjenitelj maksimalne vjerodostojnosti od ρ te je dan s

$$\begin{aligned} R &= \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}} \\ &= \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}}. \end{aligned} \tag{2.10}$$

Pomoću Cauchy-Schwartzove nejednakost može se pokazati da je $-1 \leq R \leq 1$. Primijetimo da su brojnici od R i $\hat{\beta}_1$ jednaki. Dodatno nazivnici su im nenegativni pa slijedi da su istog predznaka. Ako je vrijednost od R blizu ili jednaka nuli, to implicira gotovo ne postojeći linearni odnos između x i y. S druge strane, što je R bliži 1 ili -1, jači je linearni odnos između x i y. Ako je $R > 0$, vrijednosti y rastu kako vrijednosti x rastu, a skup podataka naziva se pozitivno koreliranim. Slično, ako je $R < 0$, vrijednosti y padaju kako vrijednosti x rastu, a skup podataka naziva se negativno koreliranim. U praktičnim primjenama, R se koristi kao pokazatelj smislenosti razvijanja linearnih regresijskih modela. Nepisano pravilo je da ukoliko je $R > 0.30$ ili $R < -0.30$, nastavljamo s razvojem linearnog regresijskog modela. Međutim, poželjna je mnogo veća apsolutna vrijednost. Izračun vjerojatnosne distribucije od R je veoma kompleksan, iz [1] slijedi da za veliki uzorak, varijabla

$$z = \frac{1}{2}\ln\left(\frac{1+R}{1-R}\right)$$

ima aproksimativno normalnu distribuciju s očekivanjem $\mu_z = (1/2)\ln[(1+\rho)(1-\rho)]$ i varijancom $\sigma_z = 1/(n-3)$. Dakle, na velikim slučajnim uzorcima, možemo provoditi testove hipoteza, primjerice da je $\rho = 0$, uz pomoć aproksimativne testne statistike

$$\begin{aligned} Z &= \frac{z - \mu_z}{\sigma_z} \\ &= \frac{(1/2)\ln\left(\frac{1+R}{1-R}\right) - (1/2)\left(\frac{1+\rho}{1-\rho}\right)}{\frac{1}{\sqrt{n-3}}}. \end{aligned}$$

2.5 Opći linearni regresijski model

Većina primjena regresijske analize u stvarnom životu koristi modele koji su složeniji od jednostavnog linearnog regresijskog modela iz razloga što je odzivna varijabla rijetko posljedica samo jedne prediktorske varijable. Stoga želimo uključiti dodatne potencijalne nezavisne varijable u modeliranje. Linearni model u kojem je zavisna varijabla opisana pomoću $k(>1)$ nezavisnih dan je s

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \varepsilon, \tag{2.11}$$

gdje je $\varepsilon \sim N(0, \sigma^2)$. U općem linearnom regresijskom modelu imamo složeniji utjecaj više prediktorskih varijabli na odzivnu varijablu, što može uključivati i međusobni utjecaj više prediktora. Osim toga, postaje izazovno odrediti koje su prediktorske varijable značajne i koje zaista utječu na odzivnu varijablu. Nadalje, želimo prilagoditi taj model

podacima. Kao i dosad, iz podataka metodom najmanjih kvadrata procjenjujemo regresijske parametre tako da rješenje bude na neki način optimalno.

Neka su $y_1, y_2, \ldots, y_n$ n nezavisnih opservacija od Y. Tada je svaki y_i u višedimenzionalnom linearnom modelu dan s

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i, \quad i = 1, 2, \ldots, n.$$

Elegantno možemo preoblikovati gornje jednadžbe u matrični oblik

$$Y = X\beta + \varepsilon,$$

gdje su

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_k \end{bmatrix}$$

vektor odgovora, matrica dizajna, vektor koeficijenata i vektor greške, respektivno.

Procjenitelji najmanjih kvadrata $\hat{\beta}_i$ od β_i za $i = 1, 2, \ldots, k$ su oni koji minimiziraju sumu kvadrata greške,

$$\begin{aligned} \text{SSE} &= \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left[y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k \right) \right]^2 \\ &= (Y - X\hat{\beta})^T (Y - X\hat{\beta}) \\ &= Y^T Y - Y^T X\hat{\beta} - (X\hat{\beta})^T Y + (\hat{\beta}X)^T X\hat{\beta}. \end{aligned}$$

Definicija 2.5.1. *Kaže se da je matrica $A \in M_n(F)$ regularna ako postoji matrica $B \in M_n(F)$ takva da vrijedi $AB = BA = I$. U tom slučaju se matrica B zove multiplikativni inverz ili inverzna matrica od A i označava s A^{-1}. Matrica $A \in M_n(F)$ se naziva singularnom matricom ako nema multiplikativni inverz.*

Propozicija 2.5.2. *Uz pretpostavku da je $X^T X$ regularna matrica, rješenje regresijskog problema pomoću metode najmanjih kvadrata je jedinstveno i dano s*

$$\hat{\beta} = \left(X^T X \right)^{-1} X^T Y. \tag{2.12}$$

Dokaz. Za minimum $\hat{\beta}$ funkcije SSE vrijedi da je $\nabla \text{SSE}(\hat{\beta}) = 0$, tj.

$$\begin{aligned} &\frac{\partial}{\partial \beta} \left(Y^T Y - Y^T X\beta - \beta^T X^T y + X^T \beta^T X\beta \right) = 0 \Leftrightarrow \\ &-2X^T (Y - X\hat{\beta}) = 0 \Leftrightarrow \\ &\left(X^T X \right) \hat{\beta} = X^T Y \end{aligned}$$

Ukoliko je $X^T X$ regularna, posljednja jednadžba ima jedinstveno rješenje

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T Y.$$

□

Naravno, jedinstvenost rješenja imamo samo u slučaju kad je matrica X punog ranga, tj. kada su prediktorske varijable linearno nezavisne. Problem nastaje kada je matrica dizajna singularna, tj. nije punog ranga, a najčešći uzroci su duplicirane i cirkularne varijable te slučaj u kojem postoji više prediktora nego podataka.

Poglavlje 3

Strojno učenje

3.1 Definicija i podjela

Definiciju strojnog učenja preuzimamo iz [5], "Kažemo da računalni program uči iz iskustva E u odnosu na neki skup zadataka T i s obzirom na mjeru uspješnosti izvođenja P, ako se povećava uspješnost obavljanja zadataka T, kroz iskustvo E, mjerena mjerom uspješnosti P". Disciplina strojnog učenja je bazirana na pitanju kako konstruirati kompjuterski program koji automatski poboljšava iskustvo. Na primjer, pretpostavimo da želimo izgraditi model koji predviđa cijenu automobila na temelju određenih atributa. Za ispravnu identifikaciju problema učenja, treba utvrditi sljedeće značajke: zadatak, mjeru uspješnosti koja se treba unaprijediti i izvor iskustva. U nastavku opisujemo svaku od ovih značajki:

- Iskustvo (E): Skup podataka s informacijama o automobilima, uključujući atribute poput godine proizvodnje, snage motora, broja kilometara i cijene.
- Zadatak (T): Predviđanje cijene automobila na temelju atributa godine proizvodnje, snage motora i broja kilometara.
- Mjera uspješnosti (P): To može biti srednja kvadratna pogreška (MSE) koja mjeri koliko dobro model predviđa stvarne cijene automobila na temelju atributa.

Kroz strojno učenje, možemo trenirati model koristeći iskustvo (E) kako bismo postigli što nižu pogrešku predviđanja cijene automobila (P). Nakon što je model istreniran, može se koristiti za predviđanje cijene novih automobila na temelju njihovih atributa.

Algoritmi strojnog učenja grade matematički model temeljen na podacima iz uzorka, koji se nazivaju trening podaci. Taj model omogućava algoritmu da donese predviđanja, odnosno odluke bez da je eksplicitno programiran za to. Podatke obično dijelimo na skup za treniranje (učenje) i skup za testiranje u omjeru 70/30 ili 80/20, kako bi mogli provjeriti učinkovitost istreniranog modela. Slijedi jednostavni način izgradnje modela strojnog

učenja. Prvo uzimamo neki skup podataka za koji znamo odgovor te podijelimo na skupove za trening i testiranje. Zatim treniramo algoritam na trening skupu te provučemo test podatke kroz algoritam i očitujemo rezultate.

Tri su glavne podjele strojnog učenja:

(i) **Nadzirano učenje (eng. *supervised learning*)**
Podaci su dani u parovima $(ulaz, izlaz) = (x, y)$ te treba pronaći procjenitelj $\hat{y} = f(x)$. Cilj je napraviti model koji će raditi predikcije na još neviđenim primjerima (modeli klasifikacije i regresije).

(ii) **Nenadzirano učenje (eng. *unsupervised learning*)**
Dani su podaci bez ciljne vrijednosti, a treba naći pravilnost u podacima (grupiranje, otkrivanje outlier-a, smanjenje dimenzionalnosti)

(iii) **Učenje s podrškom (eng. *reinforcement learning*)**
Učenje optimalne strategije na temelju pokušaja s ciljem maksimizacije kumulativne nagrade.

3.2 Teorija u pozadini

U ovom potpoglavlju upoznajemo teoriju potrebnu za razvoj prediktivnih modela. Promotrimo prvo modele koji predviđaju kvantitativne izlaze. Neka je $X \in \mathbb{R}^p$ slučajni ulazni vektor i $Y \in \mathbb{R}$ slučajna varijabla izlaza s zajedničkom distribucijom $\mathrm{P}(X, Y)$. Cilj je pronaći funkciju $f(X)$ koja može predvidjeti Y na temelju ulaza X. Ovakav pristup zahtijeva funkciju gubitka $L(Y, f(X))$ koja kažnjava pogreške u predviđanju, a najčešći izbor je kvadratna greška $L(Y, f(X)) = (Y - f(X))^2$. To nas dovodi do kriterija za odabir f, poznatog kao očekivana kvadratna greška predikcije

$$\begin{aligned} \mathrm{EPE}(f) &= \mathrm{E}(Y - f(X))^2 \\ &= \int [y - f(x)]^2 P(dx, dy). \end{aligned} \tag{3.1}$$

Koristeći svojstvo uvjetne vjerojatnosti na X, EPE se može izraziti kao

$$\mathrm{EPE}(f) = \mathrm{E}_X \mathrm{E}_{Y|X} \left([Y - f(X)]^2 \mid X \right).$$

Minimizacija EPE po točkama rezultira rješenjem

$$f(x) = \mathrm{E}(Y \mid X = x), \tag{3.2}$$

što je uvjetno očekivanje, odnosno tzv. funkcija regresije. Dakle, najbolje predviđanje, mjereno prosječnom kvadratnom greškom, vrijednosti Y u bilo kojoj točki $X = x$ je uvjetno očekivanje.

Primjerice, metode najbližih susjeda (eng. *nearest-neighbors methods*), ili jednostavno *kNN*, nastoje izravno primijeniti ovaj postupak koristeći trening podatke. Za svaku točku x računamo prosjek svih vrijednosti y_i gdje je odgovarajuća ulazna vrijednost $x_i = x$. U [4], poglavlje 2.4, to se definira kao

$$\hat{f}(x) = \text{Ave}\left(y_i \mid x_i \in N_k(x)\right),$$

gdje Ave predstavlja prosjek, a $N_k(x)$ označava susjedstvo koje sadrži k točaka u T najbližih x. Vidimo da se u ovoj metodi vrše se dvije aproksimacije:

- očekivanje se aproksimira prosjekom nad podacima iz uzorka,
- uvjetna vjerojatnost u određenoj točki proširuje se na okolinu ciljne točke.

Kada je veličina trening uzorka N velika, očekujemo da su točke iz susjedstva bliske točki x, te kako k raste, prosjek postaje stabilniji. Uz blage pretpostavke o distribuciji $\mathrm{P}(X, Y)$, može se pokazati da se $\hat{f}(x)$ približava $\mathrm{E}(Y \mid X = x)$ kako N i k teže u beskonačnost uz $k/N \to 0$.

Obzirom na pokazano, možemo se pitati ima li potrebe za daljnjim istraživanjem. Ipak, često ne raspolažemo s vrlo velikim uzorcima te postoji mogućnost dobivanja stabilnije procjene nekim drugim prikladnijim modelom u usporedbi s najbližim susjedima. Također, porastom dimenzije p povećava se i metrička veličina algoritma, prethodno spomenuta konvergencija će i dalje vrijediti, ali stopa konvergencije opada (više o ovome u slijedećem potpoglavlju). Stoga, oslanjanje samo na metodu najbližih susjeda može dovesti do značajnih neuspjeha.

Pretpostavimo sada da je regresijska funkcija $f(x)$ približno linearna u svojim argumentima, $f(x) \approx x^T\beta$. Teorijski možemo dobiti koeficijente β kao

$$\beta = \left[\mathrm{E}\left(XX^T\right)\right]^{-1} \mathrm{E}(XY).$$

Ubacivanjem ovog linearnog modela u očekivanu grešku predikcije (EPE) uočavamo da rješenje metode najmanjih kvadrata zamjenjuje očekivanja s prosjecima na trening podacima. Zaključujemo da obje metode aproksimiraju uvjetna očekivanja koristeći prosjeke, no znatno se razlikuju u pretpostavkama modela, odnosno pretpostavkama aproksimacije funkcije $f(x)$. Mnoge fleksibilnije i modernije metode rješavaju problem procjene uvjetnog očekivanja nametanjem nekih, često nerealističnih, modelnih pretpostavki. Na primjer, aditivni modeli pretpostavljaju da $f(X) = \sum_{j=1}^{p} f_j\left(X_j\right)$, upravo prethodna aditivnost omogućuje da problemi procjene uvjetnih očekivanja u visokim dimenzijama nestanu.

Pristup ostaje isti u slučaju kada je varijabla izlaza kategorička varijabla G, jedino što se mijenja je funkcija gubitka. Neka procjena $\hat{G}$ poprima vrijednosti iz skupa $\mathcal{G}$. Tada se funkcija gubitka može prikazati matricom $\mathbf{L}$ veličine $K \times K$, gdje je $K = \mathrm{card}(\mathcal{G})$. Matrica

L ima nule na dijagonali i nenegativna je svugdje drugdje, pri čemu je $L(k, l)$ kazna za klasifikaciju opservacije iz klase $\mathcal{G}_k$ u klasu $\mathcal{G}_l$. Najčešće se koristi funkcija gubitka "nula-jedan", gdje kazna za svaku krivu klasifikaciju iznosi 1. Očekivana greška predikcije je

$$\text{EPE} = \text{E}[L(G, \hat{G}(X))],$$

gdje je očekivanje u odnosu na zajedničku distribuciju $\text{P}(G, X)$. Ponovno, koristeći svojstvo uvjetne vjerojatnosti i minimizacijom po točkama slijedi

$$\hat{G}(x) = \text{argmin}_{g \in \mathcal{G}} \sum_{k=1}^{K} L(\mathcal{G}_k, g)\, \text{P}\left(\mathcal{G}_k \mid X = x\right),$$

a ukoliko je funkcija gubitka "nula-jedan" vrijedi

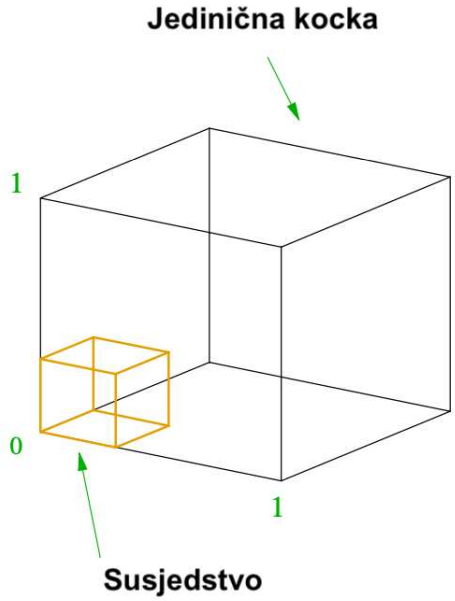
$$\hat{G}(X) = \mathcal{G}_k, \quad \text{ako je } \text{P}\left(\mathcal{G}_k \mid X = x\right) = \max_{g \in \mathcal{G}} \text{P}\left(g \mid X = x\right).$$

Ovo rješenje, poznato kao Bayesov klasifikator, kaže da klasificiramo u najvjerojatniju klasu koristeći uvjetnu (diskretnu) distribuciju $\text{P}(G \mid X)$.

3.3 Lokalne metode u višim dimenzijama

U prethodnom poglavlju smo iskazali dva prediktivna modela, stabilan, ali pristran linearni model te manje stabilan i manje pristran *kNN* model. Čini se da bismo, s dovoljno velikim skupom podataka za treniranje, uvijek mogli aproksimirati teoretski optimalno uvjetno očekivanje koristeći metodu najbližih susjeda, budući da bi trebali moći pronaći relativno veliko susjedstvo opažanja blizu bilo koje točke x [4]. Međutim, ovaj pristup ne prolazi u višim dimenzijama, a pojam je poznat kao **prokletstvo dimenzionalnosti** (eng. *curse of dimensionality*). Prateći [4], poglavlje 2.5, u nastavku razmatramo nekoliko primjera.

Pretpostavimo da su ulazni podaci *kNN* modela uniformno distribuirani u p-dimenzionalnoj jediničnoj hiperkocki i da želimo opisati susjedstvo oko neke točke x koje obuhvaća r udjela opažanja, kao što je prikazano na slici 3.1. Budući da se radi o jediničnoj hiperkocki, slijedi da je očekivana duljina ruba $e_p(r) = r^{1/p}$. Primjerice, u deset dimenzionalnom prostoru vrijedi $e_{10}(0.01) = 0.63$ i $e_{10}(0.1) = 0.80$, tj. za izgradnju susjedstva koje obuhvaća od 1% do 10% podataka, moramo uzeti od 63% do 80% raspona svake ulazne varijable. Zaključujemo da se veličina susjedstva značajno povećava s brojem dimenzija te ono nije više toliko lokalno. Također, smanjenje udjela r ne ide u korist jer dovodi do veće varijance u procjenama. Još jedna posljedica rijetkog uzorkovanja u visokodimenzionalnim prostorima je da se većina podataka nalazi bliže rubu prostora uzorka nego bilo kojem drugom podatku. Ta činjenica predstavlja problem jer je predviđanje mnogo izazovnije na rubovima trening uzorka.



Slika 3.1: Susjedstvo unutar jedinične kocke

Još jedna manifestacija prokletstva dimenzionalnosti je proporcionalnost gustoće uzorka s $N^{1/p}$, gdje je p dimenzija prostora ulaznih varijabli, a N veličina uzorka. Dakle, ako $N_1 = 100$ predstavlja gust uzorak za jednodimenzionalne ulazne varijable, tada je $N_{10} = 100^{10}$ veličina uzorka potrebna za istu gustoću uzorka u 10-dimenzionalnom prostoru ulaza. Dakle, u višim dimenzijama vrijedi da svi mogući trening uzorci rijetko nastanjuju ulazni prostor.

Pretpostavimo da pomoću metode 1-*NN* želimo predvidjeti y_0 u točki x_0 za neku funkciju f te neka je $\mathcal{T}$ skup za treniranje. Dodatno, ako pretpostavimo da imamo na raspolaganju n skupova za trening, tada možemo izračunati očekivanu grešku predviđanja u x_0 kao prosjek greški po svima njima. Budući da je problem deterministički, to je zapravo srednja kvadratna greška (MSE) procjene $f(x_0)$:

$$\begin{aligned}\mathrm{MSE}(x_0) &= \mathrm{E}_{\mathcal{T}}\left[f(x_0) - \hat{y}_0\right]^2 \\ &= \mathrm{E}_{\mathcal{T}}\left[\hat{y}_0 - \mathrm{E}_{\mathcal{T}}(\hat{y}_0)\right]^2 + \left[\mathrm{E}_{\mathcal{T}}(\hat{y}_0) - f(x_0)\right]^2 \\ &= \mathrm{Var}_{\mathcal{T}}(\hat{y}_0) + \mathrm{Bias}^2(\hat{y}_0).\end{aligned}$$

Pristranost i varijanca ovise jedna o drugoj te obje pridonose pogrešci modela. Modeli sa malom pristranošću najčešće imaju visoku varijancu te obratno. To jedan od glavnih problema prilikom izgradnje modela, izgraditi što bolji model na temelju odnosa varijance i pristranosti. Gornja dekompozicija na varijancu i kvadratnu pristranost je uvijek moguća

i često korisna, a naziva se dekompozicija varijance i pristranosti (eng. *bias-variance decomposition*) kojoj je cilj pronaći optimalnu hipotezu. U manjim dimenzijama s dovoljno velikim $\mathcal{T}$, najbliži susjed je vrlo blizu x_0 pa su i pristranost i varijanca male. Kako dimenzija raste, najbliži susjed odstupa od ciljne točke te pristranost i varijanca nastaju. Složenost funkcije f može eksponencijalno rasti s dimenzijom, a za zadržavanje točnosti procjene kao u nižim dimenzijama potreban je eksponencijalni rast veličine trening skupa.

S druge strane, pretpostavimo da je veza između Y i X linearna,

$$Y = X^T\beta + \varepsilon,$$

gdje $\varepsilon \sim N\left(0, \sigma^2\right)$ i prilagođavamo model metodom najmanjih kvadrata trening podacima. Za proizvoljnu točku x_0, imamo $\hat{y}_0 = x_0^T\hat{\beta}$, gdje je $\hat{\beta} = (X^TX)^{-1}X^Ty$. Očekivana kvadratna greška predikcije je

$$\begin{aligned}
\mathrm{EPE}(x_0) &= E(y_0 - \hat{y}_0)^2 \\
&= E\left[(y_0 - E[y_0 \mid x_0] + E[y_0 \mid x_0] - E[\hat{y}_0 \mid x_0] + E[\hat{y}_0 \mid x_0] - \hat{y}_0)^2\right] \\
&= E(y_0 - E[y_0 \mid x_0])^2 + (E[y_0 \mid x_0] - E[\hat{y}_0 \mid x_0])^2 + E(\hat{y}_0 - E[\hat{y}_0 \mid x_0])^2 \\
&= \mathrm{Var}(y_0 \mid x_0) + \mathrm{Bias}^2(\hat{y}_0) + \mathrm{Var}(\hat{y}_0).
\end{aligned}$$

Iz nepristranosti procjena metodom najmanjih kvadrata i činjenice da je $\hat{y}_0 = x_0^T\beta + \sum_{i=1}^N l_i(x_0)\varepsilon_i$, gdje je $l_i(x_0)$ i-ti element $X(X^TX)^{-1}x_0$, slijedi:

$$\mathrm{EPE}(x_0) = \sigma^2 + x_0^T E(X^TX)^{-1}x_0\sigma^2.$$

Ako je X slučajno odabran iz distribucije s $\mathrm{E}(X) = 0$ i N je velik, tada $X^TX \to N\,\mathrm{Cov}(X)$ i

$$\begin{aligned}
\mathrm{E}_{x_0}\mathrm{EPE}(x_0) &\sim \mathrm{E}_{x_0}x_0^T\,\mathrm{Cov}(X)^{-1}x_0\sigma^2/N + \sigma^2 \\
&= \mathrm{trag}\left[\mathrm{Cov}(X)^{-1}\,\mathrm{Cov}(x_0)\right]\sigma^2/N + \sigma^2 \\
&= \sigma^2(p/N) + \sigma^2.
\end{aligned}$$

Primijetimo da očekivani EPE raste linearno s povećanjem dimenzionalnosti p, uz nagib σ^2/N. Ako je N dovoljno velik i/ili varijanca σ^2 mala, porast varijance postaje zanemariv. Nametanjem nekih snažnih restrikcija na klasu modela izbjegli smo problem dimenzionalnosti.

Hastie et al. u [4] zaključuju slijedeće, "Oslanjajući se na stroge pretpostavke, linearni model nema pristranost i ima zanemarivu varijancu, dok je pogreška kod 1-najbližeg susjeda značajno veća. Međutim, kada pretpostavke ne vrijede, 1-najbliži susjed može dominirati.". Izuzev strogih linearnih modela i izuzetno fleksibilnih *kNN* modela, postoje različiti modeli razvijeni s ciljem izbjegavanja eksponencijalnog rasta složenosti funkcija u visokim dimenzijama pomoću raznih pretpostavki.

3.4 Statistički modeli i aproksimacija funkcija

Cilj je pronaći pouzdanu aproksimaciju $\hat{f}(x)$ funkcije $f(x)$ koja opisuje prediktivni odnos između ulaznih i izlaznih varijabli. Ranije smo pokazali da minimizacija očekivane kvadratne greške kvantitativnih (regresijskih) predikcija rezultira regresijskom funkcijom $f(x) = \mathrm{E}(Y \mid X = x)$ te da metode najbližih susjeda mogu procijeniti tu vezu, ali imaju ograničenja u višim dimenzijama. Kako bi se prevladala ta ograničenja, koriste se posebno dizajnirane klase modela za $f(x)$.

Definirajmo prvo deterministčku vezu između dvije varijable.

Definicija 3.4.1. *Deterministička veza između dvije varijable X i Y je veza zadana pravilom oblika*

$$Y = f(X),$$

gdje je Y zavisna varijabla, X nezavisna varijabla, a f zadana funkcija. Drugim riječima, deterministička veza postoji ako za svaku dopuštenu vrijednost nezavisne varijable X možemo izračunati točnu vrijednost zavisne varijable Y.

U praksi ne možemo očekivati deterministički odnos $Y = f(X)$, stoga se, u regresijskim predviđanjima, koristi metoda u kojoj se pretpostavlja uspostava funkcijske veze, ali uz dodanu grešku. Takve modele nazivamo statistički modeli s aditivnom greškom. Pretpostavimo da su podaci aproksimirani statističkim modelom $Y = f(X) + \varepsilon$, gdje je ε slučajna greška s očekivanjem 0 i nezavisna s X. Za ovaj model je $f(x) = \mathrm{E}(Y \mid X = x)$, štoviše uvjetna distribucija $\mathrm{P}(Y \mid X)$ ovisi o X isključivo kroz uvjetno očekivanje $f(x)$. Takav statistički model uzima u obzir neizmjerene utjecaje na Y, kao i pogreške prilikom mjerenja, na način da ih obuhvaća greškom ε. Kod određenih problema klasifikacije u strojnom učenju može postojati deterministički odnos između ulaznih i izlaznih varijabli. Međutim, u ovom radu se ne fokusiramo na te slučajeve. Pretpostavka da su greške nezavisne i jednako distribuirane nije strogo nužna, ali pojednostavljuje evaluaciju kvadratnih grešaka u EPE. Općenito, uvjetna distribucija $\mathrm{P}(Y \mid X)$ može ovisiti o X na razne načine, ali modeli s aditivnom greškom isključuju takve ovisnosti. Za klasifikacijske probleme se obično ne koriste modeli s aditivnom greškom. Umjesto toga, ciljna funkcija je uvjetna gustoća $\mathrm{P}(G \mid X)$ i modelira se izravno.

Nadzirano učenje pokušava aproksimirati funkciju f na temelju podataka iz trening skupa $\mathcal{T}$. Vrijednosti varijabli ulaza x_i se prosljeđuju u algoritam, obično računalni program, koji generira izlaze $\hat{f}(x_i)$ kao odgovor na ulaze. Algoritmi strojnog učenja imaju svojstvo modifikacije odnosa $\hat{f}$ obzirom na razlike između originalnih i generiranih izlaza $(y_i - \hat{f}(x_i))$. Ovaj proces poznat je kao učenje na temelju primjera te je upravo to motiviralo istraživanja na području strojnog učenja. Matematički i statistički pristup temelji se na aproksimaciji funkcija. Podaci x_i, y_i se tretiraju kao točke u $(p + 1)$-dimenzionalnom Euklidskom prostoru, pri čemu je funkcija $f(x)$ definirana na p-dimenzionalnom prostoru

ulaznih vrijednosti i povezana s podacima putem modela kao što je $y_i = f(x_i) + \varepsilon_i$. Jednostavnosti radi pretpostavljamo da je domena $\mathbb{R}^p$, iako ulazne vrijednosti mogu biti različitih tipova. Cilj je dobiti korisnu aproksimaciju $f(x)$ za sve x unutar određenog područja u $\mathbb{R}^p$, s obzirom na podatke u $\mathcal{T}$. Mnoge aproksimacije imaju pripadajući skup parametara θ koji se može prilagoditi podacima. Na primjer, linearni model $f(x) = x^T\beta$ ima $\theta = \beta$. Još jedna korisna klasa aproksimacija može se izraziti kao linearno proširenje

$$f_\theta(x) = \sum_{k=1}^{K} h_k(x)\theta_k,$$

gdje h_k predstavlja odgovarajući skup funkcija ili transformacija ulaznog vektora x. Česti primjeri su razna polinomna i trigonometrijska proširenja, ali uobičajena su nelinearna proširenja poput

$$h_k(x) = \frac{1}{1 + \exp\left(-x^T\beta_k\right)}.$$

Kao u linearnom modelu, parametri θ u f_θ mogu se procijeniti metodom najmanjih kvadrata na način da minimiziramo sumu kvadrata reziduala (RSS) obzirom na θ,

$$\mathrm{RSS}(\theta) = \sum_{i=1}^{N} (y_i - f_\theta(x_i))^2. \tag{3.3}$$

Ovaj kriterij je razuman za modele s aditivnom greškom. Kada je riječ o aproksimaciji funkcije, parametrizirana funkcija se vizualizira kao ploha u $(p+1)$-dimenzionalnom prostoru, a promatrani podaci se smatraju nesavršenim realizacijama te površine. Na ilustraciji 3.2, preuzetoj iz [4], dan je primjer za $p = 2$. Cilj je pronaći skup parametara θ koji minimizira vertikalne pogreške, u vidu $\mathrm{RSS}(\theta)$, kako bi se ploha što više približila promatranim točkama izlaza.

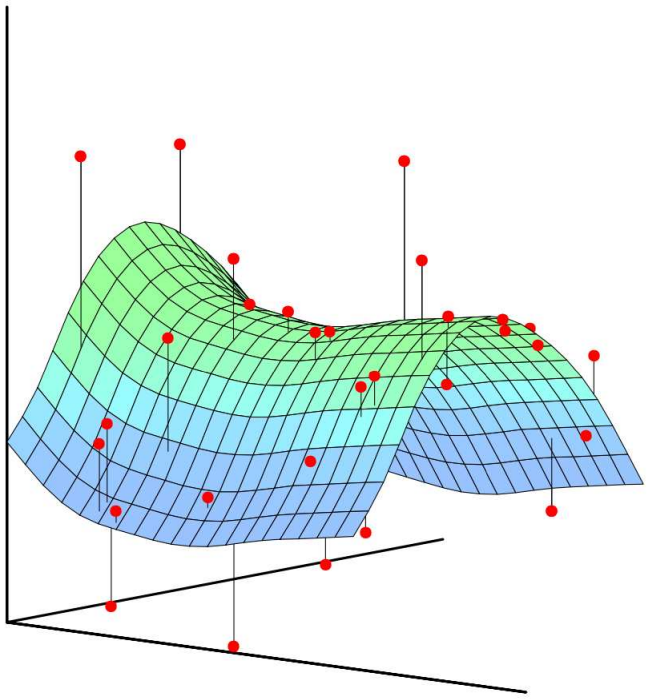
Metoda maksimalne vjerodostojnosti je općenitija metoda procjene od, često vrlo pogodne, metode najmanjih kvadrata. Ona pretpostavlja se da su najrazumnije vrijednosti parametara one koje maksimiziraju log-vjerojatnost promatranog uzorka.

Definicija 3.4.2. *Pretpostavimo da je slučajan uzorak y_i, $i = 1, \ldots, N$ iz distribucije s gustoćom $\mathrm{Pr}_\theta(y)$ koja je indeksirana nekim parametrima θ. Tada se log-vjerojatnost promatranog uzorka definira s*

$$L(\theta) = \sum_{i=1}^{N} \log \mathrm{P}_\theta\left(y_i\right). \tag{3.4}$$

Metoda najmanjih kvadrata za model s aditivnom greškom $Y = f_\theta(X) + \varepsilon$, gdje je $\varepsilon \sim N\left(0, \sigma^2\right)$, je ekvivalentna maksimalnoj vjerodostojnosti korištenjem uvjetne vjerodostojnosti

$$\mathrm{P}(Y \mid X, \theta) = N\left(f_\theta(X), \sigma^2\right).$$



Slika 3.2: Prilagodba funkcije s dva ulaza metodom najmanjih kvadrata

Dakle, iako dodatna pretpostavka normalnosti djeluje restriktivnije, rezultati su isti. Log-vjerodostojnost podataka je

$$L(\theta) = -\frac{N}{2}\log(2\pi) - N\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - f_\theta(x_i))^2 .$$

Još jedan zanimljiv primjer je multinomijalna vjerodostojnost za regresijsku funkciju $\mathrm{P}(G \mid X)$ s kvalitativnim izlazom G. Log-vjerodostojnost, također poznata kao unakrsna entropija,

$$L(\theta) = \sum_{i=1}^{N}\log p_{g_i,\theta}(x_i) ,$$

definirana je na temelju modela koji dodjeljuje uvjetne vjerojatnosti svakoj klasi s obzirom na ulaz, odnosno $\mathrm{P}(G = \mathcal{G}k \mid X = x) = pk, \theta(x), k = 1, \ldots, K$.

3.5 Stablo odluke

U svrhu razumijevanja slijedeća dva potpoglavlja dajemo uvod u stabla odluke, prateći [7]. Linearna regresija je globalni model koji koristi jednu formulu za predviđanje svih podataka. Međutim, kada podaci sadrže mnogo značajki koje međusobno djeluju na kompleksne i nelinearne načine, upotreba jednog modela može biti izazovna. Alternativni pristup je particioniranje prostora na manje dijelove kako bi se olakšalo modeliranje interakcija. Metodologija koja se koristi za konstrukciju stabala omogućuje mješavinu realnih i kategoričkih varijabli kao ulaznih varijabli. Proces konstrukcije stabla poznat je kao rekurzivno particioniranje i predstavlja iterativni proces koji podatke dijeli na particije ili grane, a zatim particionira te particije na manje grupe. Cilj je doći do manjih podskupova podataka na kojima se mogu primijeniti jednostavni modeli. Svaki list ili terminalni čvor predstavlja regiju paticije kojoj je pridružen jednostavan model koji se primjenjuje samo na tu regiju. Da bismo odredili u kojoj se regiji nalazimo, započinjemo od korijena stabla i postavljamo niz pitanja o atributima. Unutarnji čvorovi su označeni pitanjima, a grane odgovorima na ta pitanja. Model je u svakoj regiji konstantna procjena izlazne varijable. Na primjer, neka su $(x_1, y_1), (x_2, y_2), \ldots, (x_c, y_c)$ svi uzorci koji pripadaju listu l. Pri klasifikaciji model očitava najmnogobrojniju klasu u regiji, dok pri regresiji daje prosječnu vrijednost $\bar{y}$ regije. Takav model omogućuje brzo predviđanje, a analizom stabla lako je uočiti koje su varijable bitne za predviđanje. Također, čak i ako neki podaci nedostaju, moguće je napraviti predviđanja, a postoje i brzi i pouzdani algoritmi za učenje ovih stabala.

Iako postoji više algoritama za izgradnju stabla odluke, mi ćemo se koncentrirati samo na *CART (Classification and Regression Trees)* metodu koja dozvoljava samo binarno dijeljenje čvorova. Sada opisujemo proces izgradnje stabla odluke. Na početku se svi podaci iz skupa za učenje nalaze u istoj particiji. Zatim, algoritam počinje alocirati podatke u prve dvije particije ili grane koristeći sve moguće binarne rezove na svakom području. Algoritam odabire rez za koji se postiže što veća razina čistoće (eng. *purity*). Postupak se ponavlja za svaku novu granu te se proces nastavlja sve dok svaki čvor ne postane dovoljno mali i postane list. Prilikom predviđanja može doći do problema prenaučenosti (engl. *overfitting*). Prenaučenost ili problem visoke varijance je situacija u kojoj se odlično opisuju trening podatci, no to povlači loša predviđanja novih stvarnih podataka. Budući da se stablo konstruira iz trening skupa, potpuno razvijeno stablo često može pokazati prenaučenost. U svrhu spriječavanja prenaučenosti kosritimo tehniku podrezivanja (eng. *pruning*).

Ansamblom nazivamo postupak kombiniranja predikcija više modela koji su izgrađeni s istim ili različitim algoritmima, koristeći iste ili različite skupove podataka. Cilj ansambla je poboljšati kvalitetu predikcije u usporedbi s pojedinačnim modelom. Primjeri ansambla su boosting, bagging i slučajna šuma. Svi se oni razlikuju u načinu odabira skupa za treniranje, generiranju klasifikatora i kombiniranju njihovih izlaza. Međutim, u sva tri

pristupa se svaki slabi klasifikator uči na cijelom trening skupu.

3.6 Boosting algoritam

Boosting je često korištena metoda ansambla za kombiniranje slabih modela. Metoda pridodaje težine pojedinačnim modelima na temelju njihove preciznosti, što rezultira smanjenjem pristranosti i varijance konačnog modela. Boosting je izvorno razvijen za probleme klasifikacije, no pokazao se uspješnim i za zadatke regresije. Iako boosting dijeli neke sličnosti s drugim ansambl metodama, poput bagginga, posjeduje karakteristike koje ga temeljno drugačijim. Sadržaj ovog poglavlja temeljen je na [4] i [6].

Jedan od prvih i najpopularnijih boosting modela je *AdaBoost.M1* kojega su razvili Freund i Schapire (1996.). AdaBoost se temelji na stablima odluke koja obično imaju jedan čvor i dva lista, tzv. panjevi (eng. *stump*), te stvara šumu takvih stabala. Primjerice, pretpostavimo da je u binarnom klasifikacijskom problemu izlazna varijabla $Y \in \{-1, 1\}$. Tada za svaki vektor prediktorskih varijabli X, klasifikator $G(X)$ generira predikciju koja se nalazi u skupu $\{-1, 1\}$. Greška na trening skupu je dana s

$$\overline{\text{err}} = \frac{1}{N} \sum_{i=1}^{n} I(y_i \neq G(X)),$$

dok je očekivana greška za buduće predikcije $\mathrm{E}_{XY}\, I(Y \neq G(X))$.

Klasifikator koji ima samo malo bolju točnost od slučajnog odabira smatra se slabim. Cilj boostinga je iterativno primijeniti slab klasifikator na modificirane verzije podataka, rezultirajući nizom slabih klasifikatora $G_m(x), m = 1, 2, \ldots, M$. Predikcije ovih slabih klasifikatora kombiniraju se u konačnu predikciju

$$G(x) = \text{sign}\left(\sum_{m=1}^{M} \alpha_m G_m(x)\right).$$

Težine $\alpha_1, \alpha_2, \ldots, \alpha_M$ određene pomoću boosting algoritma, koriste se kako bi se dodijelila veća važnost točnijim klasifikatorima u nizu. Tijekom svakog koraka u algoritmu, podaci se mijenjaju dodjeljivanjem težina $w_1, w_2, \ldots, w_N$ svakoj opservaciji $(x_i, y_i)\,, i = 1, 2, \ldots, N$ iz skupa za treniranje. U prvom koraku su sve težine jednake, dok se u svakoj sljedećoj iteraciji težine opservacija prilagođavaju na temelju njihove prethodne klasifikacije, na način da se povećavaju za pogrešno klasificirane opservacije, a smanjuju za one točno klasificirane. Ovaj proces omogućuje da opservacije koje je teško ispravno klasificirati dobiju veći utjecaj. Dakle, svaki sljedeći klasifikator se usredotočuje na pogrešno klasificirane opservacije prethodnoga.

U nastavku dajemo pseudokod *AdaBoost.M1* algoritma:

1. Inicijaliziraj težine opservacija $w_i = 1/N, i = 1, 2, \ldots, N$.

2. Za $m = 1$ do M:

   a) Primijeni klasifikator $G_m(x)$ na trening podatke koristeći težine w_i.

   b) Izračunaj težinsku grešku
   $$\text{err}_m = \frac{\sum_{i=1}^{N} w_i I\left(y_i \neq G_m\left(x_i\right)\right)}{\sum_{i=1}^{N} w_i}.$$

   c) Izračunaj težinu klasifikatora $G_m(x)$, $\quad \alpha_m = \log\left(\left(1 - \text{err}_m\right) / \text{err}_m\right)$.

   d) Ažuriraj težine opservacija, $\quad w_i = w_i \cdot \exp\left[\alpha_m \cdot I\left(y_i \neq G_m\left(x_i\right)\right)\right], i = 1, 2, \ldots, N.$

3. Izračunaj konačni klasifikator $G(x) = \text{sign}\left[\sum_{m=1}^{M} \alpha_m G_m(x)\right]$.

Algoritam se može lako modificirati za regresijske probleme.

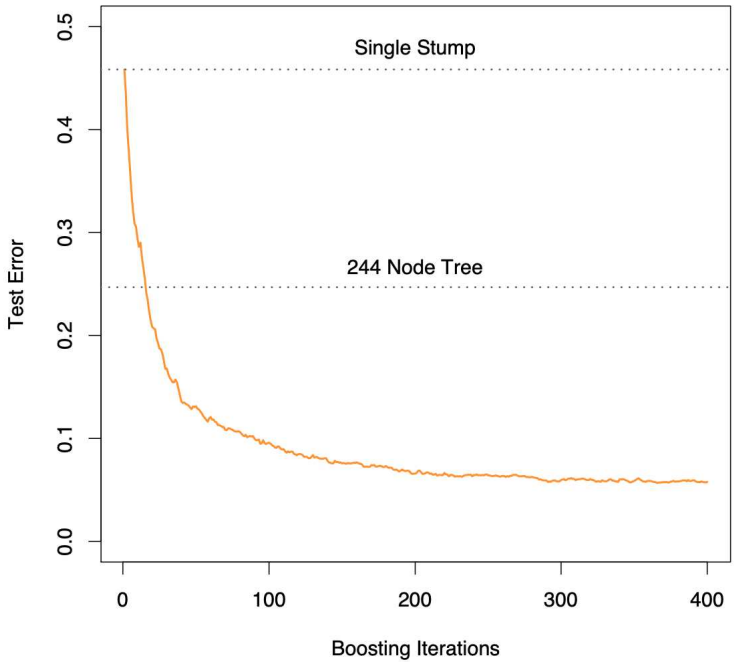
Panjevi sami po sebi ne donose precizne klasifikacije, ali šumom panjeva, u kojoj svaki panj uči na pogreškama svojih prethodnika, dobivamo klasifikator sa zavidnim performansama. To se vidi na slici 3.3, preuzetoj iz [4], stranica 340. Naime, prikazana je greška na testnim podacima u ovisnosti o broju iteracija boosting algoritma. Vodoravnim linijama su istaknute greške za panj i stablo odluke s 244 čvora. U prvoj iteraciji panj ima grešku od 45.8%, što je vrlo blizu slučajnom odabiru. Međutim, kako se broj iteracija u boosting algoritmu povećava, greška se postupno smanjuje, dosežući 5.8% nakon 400 iteracija. Dakle, boosting-om je greška smanjena gotovo osam puta, znatno nadmašujući i poprilično veliko stablo odluke. Modeli su razvijani na simuliranim podacima.

Gore navedeno opravdava Breimanovu izjavu o AdaBoost alogoritmu čiji su klasifikatori stabla odluke kao ”najboljim gotovim klasifikatorom na svijetu”.

3.7 Boosting stabla

Metode temeljene na stablima odluke razdvajaju prostor ulaznih varijabli na skup disjunktnih područja, odnosno regija, R_j, $j = 1, 2, \ldots, J$. Svakoj regiji se prilagođava jednostavan model, poput konstante γ_j, te vrijedi

$$x \in R_j \Rightarrow f(x) = \gamma_j.$$



Slika 3.3: Greška predviđanja na testnim podacima u ovisnosti o broju iteracija boosting algoritma

Stoga se stablo može formalno izraziti kao

$$T(x;\Theta) = \sum_{j=1}^{J} \gamma_j I\left(x \in R_j\right), \tag{3.5}$$

gdje je skup parametara $\Theta = \left\{R_j, \gamma_j\right\}_1^J$. Te parametre dobivamo minimizacijom

$$\hat{\Theta} = \arg\min_{\Theta} \sum_{j=1}^{J} \sum_{x_i \in R_j} L\left(y_i, \gamma_j\right),$$

što je poprilično zahtjevan kombinatorno optimizacijski problem pa su suboptimalna rješenja prihvatljiva. U nastavku navodimo jedno takvo rješenje, ono zahtjeva podjelu problema na dva dijela.

1. Pronalaženje γ_j uz dani R_j.
   Ovo je trivijalni dio, naime $\hat{\gamma}_j$ je prosjek ili većinski glas svih y_i iz regije R_j.

2. Pronalaženje R_j.
   U ovom dijelu problema pronalazimo približna rješenja. Najčešći pristup za pronalaženje regije je pomoću pohlepnog rekurzivnog algoritma particioniranja. Pored toga, za općenitije funkcije gubitka je praktičnije optimizirati R_j ako aproksimiramo $\hat{\Theta}$ s

$$\tilde{\Theta} = \arg\min_{\Theta} \sum_{i=1}^{N} \tilde{L}\left(y_i, T\left(x_i, \Theta\right)\right).$$

Primjerice, u klasifikacijskim stablima odluke čvorovi se biraju, a time se u konačnici određuju i regije, pomoću pohlepnog algoritma koji odabire atribut kojim se postiže najveća čistoća. Mjere čistoće su entropija i Gini indeks. Entropija je mjera nečistoće u skupu podataka. Za slučajnu varijablu V računamo entropiju kao

$$H(V) = -\sum_{k} P(v_k) \log_2 P(v_k),$$

gdje je $P(v_k)$ vjerojatnost ishoda v_k i $\sum_k P(v_k) = 1$. S druge strane, Gini indeks čvora predstavlja vjerojatnost da će nasumično odabrani uzorak biti krivo klasificiran u tom čvoru. Gini indeks se za svaki čvor računa s

$$I_G(n) = 1 - \sum_{i=1}^{J} (p_i)^2,$$

gdje je n trenutni čvor, p_i je vjerojatnost klase i u čvoru n, a J je broj klasa u modelu.

Model boosting stabla je zbroj takvih stabala odluke,

$$f_M(x) = \sum_{m=1}^{M} T\left(x; \Theta_m\right), \tag{3.6}$$

induciran stadijskim (eng. *stagewise*) načinom, odnosno dodavanjem jednog po jednog slabog klasifikatora. Na svakom stadiju postepene izgradnje boosting modela mora se riješiti:

$$\hat{\Theta}_m = \arg\min_{\Theta_m} \sum_{i=1}^{N} L\left(y_i, f_{m-1}\left(x_i\right) + T\left(x_i; \Theta_m\right)\right) \tag{3.7}$$

za skup regija i konstanti $\Theta_m = \left\{R_{jm}, \gamma_{jm}\right\}_1^{J_m}$ sljedećeg stabla, uzimajući u obzir trenutni model $f_{m-1}(x)$. Nakon definiranja regije R_{jm}, cilj je pronaći optimalne konstante γ_{jm} koje

minimiziraju određenu funkciju gubitka. Općenito, optimizacijski problem može se formulirati na sljedeći način:

$$\hat{\gamma}_{jm} = \arg\min_{\gamma_{jm}} \sum_{x_i \in R_{jm}} L\left(y_i, f_{m-1}\left(x_i\right) + \gamma_{jm}\right),$$

gdje je L odabrana funkcija gubitka, y_i opaženi odgovor, a $f_{m-1}(x_i)$ je predikcija iz prethodne faze poboljšanja. S druge strane, pronalaženje regije R_{jm} je izazovan zadatak koji je još teži u usporedbi s konstrukcijom pojedinačnog stabla odlučivanja. Međutim, za određene funkcije gubitka, problem se pojednostavljuje.

Za funkciju gubitka kvadratne greške, rješenje za 3.7 je analogno kao kod pojedinačnog stabla odluke. Dakle, rješenje je regresijsko stablo odluke koje najbolje predviđa trenutne reziduale $y_i - f_{m-1}(x_i)$, te je optimalna konstanta γ_{jm} za ćeliju R_{jm} jednaka

$$\hat{\gamma}_{jm} = \text{Ave}(y_i - f_{m-1}(x_i)), \quad x_i \in R_{jm}.$$

U slučaju binarne klasifikacije i eksponencijalne funkcije gubitka, ovaj stadijski pristup dovodi do AdaBoost metode za boosting klasifikacijska stabala. Konkretno, ograničimo stabla $T(x; \Theta_m)$ iz 3.7 na skalirana stabla klasifikacije, tj. $\beta_m T(x; \Theta_m)$ gdje je $\gamma_{jm} \in \{-1, 1\}$. Tada se gornji optimizacijski problem svodi na

$$\hat{\Theta}_m = \arg\min_{\Theta_m} \sum_{i=1}^{N} w_i^{(m)} I\left(y_i \neq T\left(x_i; \Theta_m\right)\right),$$

gdje su težine dane s $w_i^{(m)} = e^{-y_i f_{m-1}(x_i)}$. Čak i bez ograničavanja stabala, sama eksponencijalna funkcija gubitka pojednostavljuje 3.7

$$\hat{\Theta}_m = \arg\min_{\Theta_m} \sum_{i=1}^{N} w_i^{(m)} \exp\left[-y_i T\left(x_i; \Theta_m\right)\right].$$

Koristeći ovaj težinski eksponencijalni gubitak kao kriterij za razdvajanje, možemo izgraditi pohlepni rekurzivni algoritam za particioniranje. Jednom kada imamo regiju R_{jm}, slijedi da je optimalna konstanta γ_{jm} jednaka

$$\hat{\gamma}_{jm} = \log \frac{\sum_{x_i \in R_{jm}} w_i^{(m)} I\left(y_i = 1\right)}{\sum_{x_i \in R_{jm}} w_i^{(m)} I\left(y_i = -1\right)}.$$

U svrhu veće robustnosti boosting stabala koriste se funkcije gubitka kao što su apsolutna greška ili Huberova greška za regresiju, dok se za klasifikaciju umjesto eksponenci-

jalne funkcije gubitka koristi devijacija

$$\begin{aligned} L(y, p(x)) &= -\sum_{k=1}^{K} I\left(y = \mathcal{G}_k\right) \log p_k(x) \\ &= -\sum_{k=1}^{K} I\left(y = \mathcal{G}_k\right) f_k(x) + \log\left(\sum_{\ell=1}^{K} e^{f_\ell(x)}\right). \end{aligned}$$

Nažalost ove robustne funkcije gubitka ne rezultiraju jednostavnim i brzim boosting algoritmima.

Poglavlje 4

Kreditni skoring u bankarstvu

4.1 Uvod

Tvrdnje u ovom poglavlju temeljene su na [12] i [13]. Kreditni skoring je metodologija koja pruža formalni i automatizirani pristup procjeni kreditnog rizika pojedinog zajmotražitelja. Njegova svrha je olakšati obradu velikog broja zahtjeva za manje kredite. Kroz kreditni skoring, zajmodavac dodjeljuje bodove kako bi dobio numeričku vrijednost koja ukazuje na vjerojatnost da zajmotražitelj doživi određeni događaj ili poduzme određenu radnju, poput kašnjenja u otplati kredita. Ovaj proces omogućuje pojednostavljenje postupka kreditiranja, poboljšanje učinkovitosti kreditnih službenika te dosljedniju ocjenu kreditnog rizika. Također smanjuje ljudsku pristranost u odlukama o kreditiranju, omogućava bankama da prilagode svoju kreditnu politiku prema klasifikaciji rizika te da odobravaju ili prate zajmove nižeg rizika bez fizičke provjere na terenu. Kroz kreditni skoring, moguće je i preciznije kvantificirati očekivane gubitke za različite razine rizika zajmoprimaca te smanjiti vrijeme koje se provodi u procesima naplate nenaplaćenih potraživanja.

Kreditni skoring je evoluirao tijekom vremena, a statističke metode su bile i ostale najvažnije u izgradnji kreditnih bodovnih kartica. Jedna od njihovih prednosti je mogućnost korištenja znanja o svojstvima procjenitelja uzoraka te alate za pouzdane intervale i testiranje hipoteza u kontekstu kreditnog skoringa. U početku su se koristile metode temeljene na Fisherovim linearnim diskriminantnim funkcijama, ipak bile su previše restriktivne i time ne održive u praksi. Fisherov pristup se može smatrati kao vrsta linearne regresije, što je potaknulo istraživanje drugih oblika regresije, poput logističke, s manje restriktivnim pretpostavkama kako bi se osigurala njihova optimalnost, ali i dalje rezultirala linearnim pravilima za skoring. Još jedan pristup izračunu kreditnog skoringa je pomoću metoda strojnog učenja kao što su klasifikacijska stabala. Ona ne daju težinu svakom od atributa, kao što daje linearna bodovna kartica, već dijele skup podnositelja zahtjeva u niz različitih podskupova. Neovisno o metodi, rezultat se koristi za odlučivanje hoće li novi podnositelj

zahtjeva biti klasificiran kao zadovoljavajući ili nezadovoljavajući.

Kreditni skoring se temelji na pragmatizmu i empirizmu, s osnovnim ciljem predviđanja rizika umjesto objašnjavanja. Snaga leži u njegovoj solidnoj metodologiji i oslanjanju na empirijski dobivene podatke. Sustavi skoringa se razvijaju na temelju prošlog uspjeha sličnih potrošača, koristeći uzorke prošlih korisnika koji su se prijavili za proizvod. Sve karakteristike koje pomažu u predviđanju, poput stabilnosti, financijske stručnosti ili kapitala potrošača, mogu se koristiti u sustavu skoringa bez potrebe za opravdanjem. Međutim, određene karakteristike poput rase, religije i spola su zakonski zabranjene u kreditnim sustavima skoringa radi sprječavanja diskriminacije. Druge karakteristike, iako nisu zakonski zabranjene, možda se ne koriste zbog kulturne neprihvatljivosti. Upotreba kreditnih agencija za provjeru prijevara i pristup povijesti kreditiranja postala je standardna praksa. U konačnici, kreditni skoring je pokazao značajna poboljšanja u odnosu rizika i povrata za različite kreditne proizvode, iako se rasprave o njegovoj primjeni u određenim kontekstima i dalje nastavljaju.

Rudarenje podataka (eng. *data mining*) podrazumijeva istraživanje i analizu podataka radi otkrivanja važnih uvida i veza. Organizacije, poput banaka, prepoznaju važnost informacija o svojim kupcima koje mogu prikupiti putem elektroničkih kartica i općenitog elektroničkog prijenosa sredstava. Analiza velikih količina podataka postaje moguća zahvaljujući napretku računalne snage. Stoga banke sve više ulažu u razvoj baza podataka za pohranu informacija o kupcima.

Definicija 4.1.1. *Ako dužnik ne podmiri obveze u vremenski ugovorenim rokovima, kažemo da je došlo do neizvršenja novčanih obaveza (eng. default)*

4.2 Linearna regresija u izradi kreditnog skoringa

Prema poglavlju 4 u [12], iskazujemo linearnu regresiju kao pristup kreditnom skoringu. Ovaj pristup proizlazi iz linearne diskriminatne funkcije te se u njemu pokušava naći najbolja linearna kombinacija karakteristika (atributa)

$$w_0 + w_1X_1 + w_2X_2 + \ldots + w_pX_p = w^* \cdot X^{*T},$$

gdje je $w^* = \left(w_0, w_1, w_2, \ldots, w_p\right)$, a $X^* = \left(1, X_1, X_2, \ldots, X_p\right)$ što objašnjava vjerojatnost neispunjavanja obveza podnositelja. Ako je p_i vjerojatnost da podnositelj zahtjeva iz uzorka nije ispunio obveze, želimo pronaći w* koji najbolje aproksimira

$$p_i = w_0 + x_{i1}w_1 + x_{i2}w_2 + \ldots + x_{ip}w_p, \quad \forall i. \tag{4.1}$$

Pretpostavimo da je n_G dio uzorka koji je dobar. Zbog lakše notacije pretpostavljamo da su to prvih n_G u uzorku. Stoga slijedi $p_i = 1$ za $i = 1, \ldots, n_G$. Ostali, $n_G + 1, \ldots, n_G + n_B$ gdje je $n_G + n_B = n$, su loši te za njih vrijedi $p_i = 0$.

U linearnoj regresiji odabiremo koeficijent koji minimizira srednju kvadratnu pogrešku između lijeve i desne strane jednadžbe 4.1. Odnosno, minimiziramo slijedeće

$$\sum_{i=1}^{n_G}\left(1-\sum_{j=0}^{p} w_j x_{ij}\right)^2+\sum_{n_G+1}^{n_G+n_B}\left(\sum_{j=0}^{p} w_j x_{ij}\right)^2. \tag{4.2}$$

Vektorski se 4.1 može zapisati kao

$$\begin{pmatrix} 1 & X_G \\ 1 & X_B \end{pmatrix}\begin{pmatrix} w_0 \\ w \end{pmatrix}=\begin{pmatrix} \mathbf{1}_G \\ 0 \end{pmatrix} \text{ ili } Yw^T=b^T, \tag{4.3}$$

gdje je

$$Y=\begin{pmatrix} \mathbf{1}_G & X_G \\ \mathbf{1}_B & X_B \end{pmatrix},\quad X_G=\begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \vdots \\ x_{n_G 1} & \cdots & x_{n_G p} \end{pmatrix},\quad X_B=\begin{pmatrix} x_{n_G+11} & \cdots & x_{n_G+1 p} \\ \vdots & \vdots & \vdots \\ x_{n_G+n_B 1} & \cdots & x_{n_G+n_B p} \end{pmatrix},$$

a $b^T=\begin{pmatrix} \mathbf{1}_G \\ 0 \end{pmatrix}$ gdje je $\mathbf{1}_G$ ($\mathbf{1}_B$) $1\times n_G(1\times n_B)$ vektor u kojem su sve stavke jedinice. Kao u 4.2, pronalaženje koeficijenata linearne regresije odgovara minimizaciji

$$\left(Yw^T-b^T\right)^T\left(Yw^T-b^T\right) \tag{4.4}$$

To je minimizirano kada je derivacija po w jednaka 0, to jest

$$Y^T\left(Yw^T-b^T\right)=0\Leftrightarrow Y^TYw^T=Y^Tb^T, \tag{4.5}$$

gdje su

$$Y^Tb^T=\begin{pmatrix} 1 & 1 \\ X_G & X_B \end{pmatrix}\begin{pmatrix} \mathbf{1}_G \\ 0 \end{pmatrix}=\begin{pmatrix} n_G \\ n_G m_G \end{pmatrix},$$
$$Y^TY=\begin{pmatrix} 1 & 1 \\ X_G & X_B \end{pmatrix}\begin{pmatrix} 1 & X_G \\ 1 & X_B \end{pmatrix}=\begin{pmatrix} n & n_G m_G+n_B m_B \\ n_G {m_G}^T+n_B {m_B}^T & {X_G}^T X_G+{X_B}^T X_B \end{pmatrix}.$$

Ako u svrhu obrazloženja označimo procijenjeno očekivanje kao stvarno očekivanje, tada dobijemo

$${X_G}^T X_G+{X_B}^T X_B=nE\left[X_iX_j\right]=nS+n_G m_G m_G^T+n_B m_B m_B^T, \tag{4.6}$$

gdje je S uzoračka kovarijacijska matrica. Koristeći 4.6, možemo proširiti 4.5 do

$$\begin{aligned} nw_0+\left(n_G m_G+n_B m_B\right)w^T&=n_G \\ \left(n_G {m_G}^T+n_B {m_B}^T\right)w_0+\left(nS+n_G m_G {m_G}^T+n_B m_B {m_B}^T\right)w^T&=n_G {m_G}^T. \end{aligned} \tag{4.7}$$

Ako supstituiramo prvu jednakost iz 4.7 u drugu jednakost, dobijemo

$$\begin{aligned}&\left(\left(n_G m_G^T + n_B m_B^T\right)\left(n_G - \left(n_G m_G + n_B m_B\right) w^T\right)/n\right)\\&+\left(n_G m_G {m_G}^T + n_B m_B {m_B}^T\right) w^T + nS\, w^T = n_G {m_G}^T\end{aligned} \tag{4.8}$$

Iz čega konačno slijedi

$$S\, w^T = c\left(m_G - m_B\right)^T. \tag{4.9}$$

Jednakost 4.8 nam daje najbolji odabir $\mathrm{w} = \left(\mathrm{w}_1, \mathrm{w}_2, \ldots, \mathrm{w}_\mathrm{p}\right)$ za koeficijente linearne regresije. U [12] se pokazuje da je dobiveni w isti kao u linearnoj diskriminantnoj funkciji. Stoga, ovaj pristup pokazuje da možemo dobiti koeficijente kreditne bodovne kartice pomoću metode najmanjih kvadrata.

4.3 Strojno učenje u izradi kreditnog skoringa

Jedan drugačiji statistički pristup klasifikaciji i diskriminaciji je koncept klasifikacijskih stabala, također poznat kao algoritmi rekurzivnog particioniranja (eng. *RPA*). Ideja iza ovog pristupa je podijeliti skup odgovora na aplikaciju u različite regije i odrediti većinsku klasifikaciju, dobra ili loša, unutar svake. Ovaj pristup je prvotno razvijen za opće probleme klasifikacije od strane Breimana i Friedmana 1973. godine, a njegova primjena u ocjenjivanju kreditne sposobnosti brzo je slijedila. Slične ideje su također prihvaćene u području umjetne inteligencije, te je razvijen specijalizirani softver za implementaciju.

Postupak klasifikacijskog stabla uključuje nekoliko odluka. Prvo, pravilo podjele koje određuje kako podijeliti skup u dva podskupa. Drugo, pravilo zaustavljanja koje određuje kada podskup postaje završni čvor, odnosno list, u stablu. Na kraju, dodjela završnih čvorova u dobre i loše kategorije temeljem većinske klasifikacije ili razmatranjem minimizacije troškova. Za kontinuirane varijable, najbolja podjela se određuje evaluacijom različitih pragova, dok se za kategoričke varijable razmatraju sve moguće podjele. Različite mjere mogu se koristiti za procjenu kvalitete podjele, kao što su Kolmogorov-Smirnov statistika, indeksi nečistoće (poput Gini indeksa i entropije) ili polusuma kvadrata. Ove mjere pomažu identificirati optimalnu podjelu u klasifikacijskom stablu [12].

4.4 Praktična primjena prediktivnih modela u izradi kreditnog skoringa

Uvod

U članku [14] se opisuje projekt realiziran za jednu veliku europsku banku. Pokazalo se da je isključivo koristeći se podacima iz *PSD2* moguće izvesti izuzetno snažan prediktor

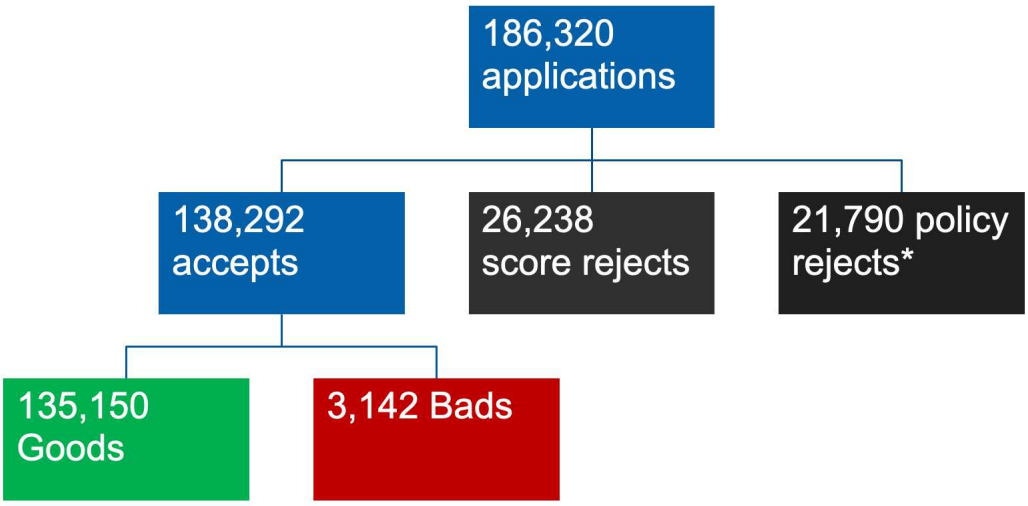
zaduženja potrošača, koji omogućuje donošenje profitabilnih odluka o odobravanju kredita. Revidirana direktiva o platnim uslugama, ili skraćeno PSD2 (eng. *Payment Services Directive*), je direktiva Europske unije koja je stupila na snagu 2018. godine s ciljem promicanja transparentnosti, sigurnosti, inovacija i konkurencije u financijskom sektoru. Njom se omogućuje korisnicima bankarskih usluga da koriste usluge trećih strana, za obavljanje različitih aktivnosti temeljem njihovih financijskih podataka. Direktiva zahtijeva od banaka da omoguće, pružateljima usluga trećih strana, pristup računima svojih korisnika, naravno uz njihovu suglasnost. Kao rezultat, korisnici mogu imati koristi od pristupa novim i inovativnim proizvodima te poboljšanih razina usluga. U [14] se pokazuje kako PSD2 podaci imaju potencijal da transformiraju ocjenjivanje kreditne sposobnosti i donošenje odluka o kreditiranju u financijskom sektoru na način da, poboljšavajući procjenu rizika, omogućuju precizne odluke o kreditiranju i personalizirano iskustvo za korisnike. Također, ističe se prednosti kombiniranja PSD2 podataka i strojnog učenja za dobivanje vrijednih uvida.

Podaci

Banke obično koriste razumljive varijable, poput kreditne povijesti, kako bi procijenile rizik kreditiranja zajmotražitelja. Kreditna povijest uključuje faktore poput prethodnih kašnjenja u plaćanju, broj puta i količina uzastopnih mjeseci kašnjenja. Iako neki modeli mogu uzeti u obzir plaću i određene transakcije tekućeg računa, ove varijable obično nisu glavni pokretači modela zbog dodatnog napora potrebnog za njihovo izvlačenje.

Autori su imali pristup podacima jedne od vodećih banaka u središnjoj Europi. Na raspolaganju su imali sve zahtjeve za potrošačke kredite u razdoblju od svibnja 2016. do svibnja 2017. godine, šest mjeseci transakcijske povijesti prije podnošenja zahtjeva za kredit, tzv. prozor promatranja, te jednogodišnje razdoblje nakon podnošenja zahtjeva koje je korišteno za utvrđivanje defaulta. U podacima je bilo više od 100 000 zahtjeva za kredit, sa stopom neizvršenja većom od 7%, što znači da su imali dovoljno opažanja za provedbu detaljnih analiza i značajne uzorke za validacijske svrhe. Ilustracija 4.1 pokazuje distribuciju zahtjeva za kredit s obzirom na njihov ishod.

Ipak, autori su se suočili s nekoliko izazova u učinkovitom korištenju ovih podataka. Prvi izazov je bio stvaranje značajki (eng. *features*) iz neprocesiranih podataka o transakcijama, budući da se ti podaci značajno razlikuju od tradicionalnih podataka korištenih za izračun kreditnog skoringa. Tradicionalni modeli se oslanjaju na agregirane podatke, dok PSD2 podaci pružaju pojedinačne detalje o transakcijama. Drugi izazov bio je suočavanje sa šumom i nepravilnostima u podacima transakcija tekućeg računa, što otežava izdvajanje specifičnih informacija poput plaće. Osim toga, izvučene informacije iz transakcija često su visoko korelirane, što zahtijeva inovativne pristupe kako bi se održala stabilnost i diskriminatorna moć u modelima kreditnog skoringa.



Slika 4.1: Ilustracija preuzeta iz [15].

Značajke (Features)

Jedna od ključnih snaga pristupa bila je metodologija ekstrakcije informacija iz transakcijskih podataka za ulazak u modele strojnog učenja. Kreirano je otprilike 3000 značajki u različitim vremenskim razdobljima i vremenskim rasponima unutar prozora promatranja, koji hvataju različite trendove i događaje. Na primjer,

- Jednostavne varijable kao minimalna, maksimalna, prosječna transakcija
- Agregirane varijable poput maksimalnog broja dana bez prihoda
- Statističke mjere kao medijan, prosjek, standardna devijacija transakcija određene vrste
- Složene varijable koje su generirane pomoću specijaliziranih paketa (npr. Fourierove transformacije vremenskih nizova)

Također, neke značajke su izračunate pomoću prilagođenih sekundarnih algoritama. Primjerice, za procjenu plaće su razvili relativno jednostavno stablo odluke koje je aproksimiralo plaću klijenta na temelju analize kreditnih transakcija tijekom vremena. Zatim, su definirali novu značajku zvanu stabilnost plaće koja predstavlja interpretaciju rezultata algoritma za plaću i daje razinu pouzdanosti procjene plaće. Kasnije je pokazano kako je

stabilnost plaće bila važnija od same plaće. Drugi primjer je pristup mjerenju pravilnosti u potrošnji. Iznos i vrsta potrošnje svakako su bile važne značajke, ali trendovi i varijabilnost potrošnje tijekom vremena pružili su dodatne informacije koje su se pokazale kao snažan prediktor kreditnog rizika.

Takve složene značajke unesene su u model nadziranog strojnog učenja. U nastavku navodimo nekolicinu najvažnijih značajki za jedan od najbolje izvedenih modela:

- *6months_Current_balance_std_min* - Najmanje mjesečno standardno odstupanje trenutnog stanja u posljednjih 6 mjeseci prije prijave

- *6months_Current_balance_min_max* - Najveći mjesečni minimum trenutnog stanja u posljednjih 6 mjeseci prije prijave

- *6months_I_mean_min* - Najmanja mjesečna prosječna odlazna uplata u posljednjih 6 mjeseci prije prijave

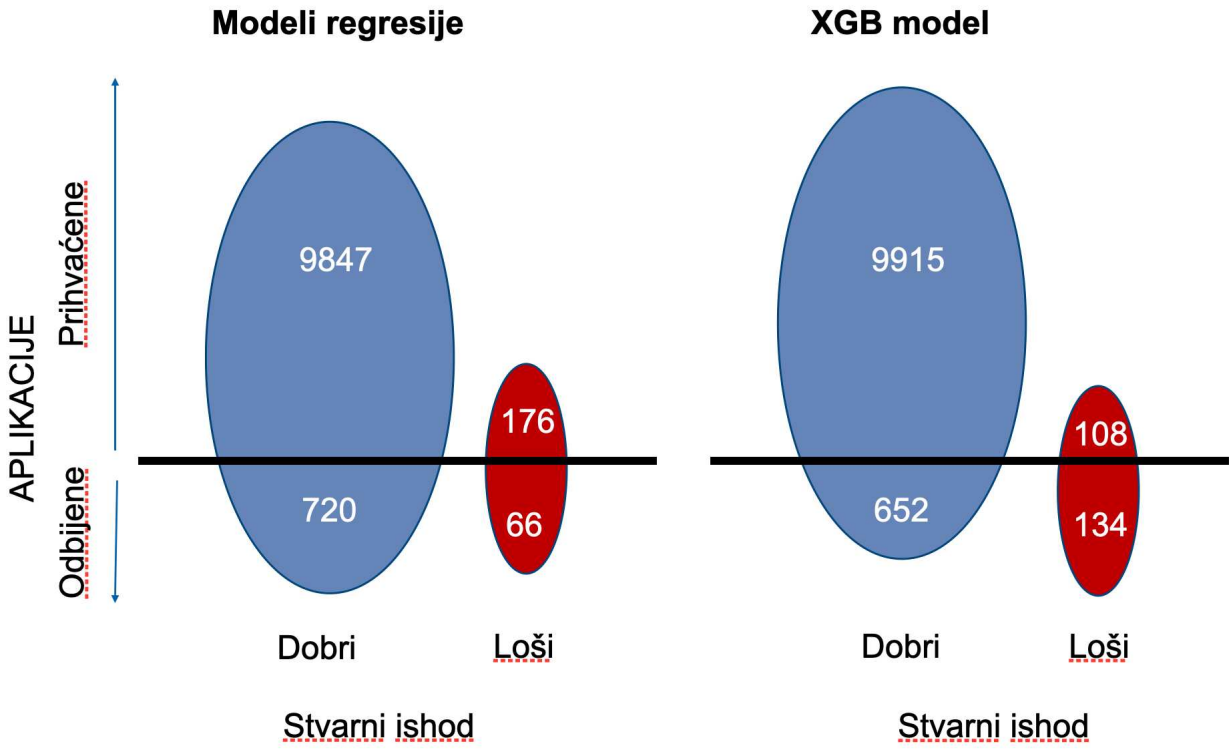
- *4week_size_of_transactions2week* - Broj transakcija u 2 tjedna prije prijave

- *6months_avg_CTO_minus_DTO_ratio* - Razlika između zbroja svih dolaznih i odlaznih transakcija po mjesecima (prosjek zadnja 3 mjeseca / prosjek 3 mjeseca prije)

Modeliranje kreditnog skoringa

Autori su proveli istraživanje koristeći različite tehnike strojnog učenja kako bi razvili prediktivne modele za procjenu kreditnog rizika. Počeli su s jednostavnim modelima poput linearnog i logističkog regresijskog modela kako bi postavili referentnu točku, postižući Gini indekse između 55% i 59%. Na temelju zadovoljavajućih rezultata tih jednostavnijih modela, zaključili su da stvorene varijable uspješno prikazuju ponašanje klijenata. Također su testirali duboko učenje, poput neuronskih mreža i metode potpornih vektora, ali su utvrdili da su manje učinkoviti zbog njihove osjetljivosti na korelaciju. Čak i nakon primjene tehnika smanjenja dimenzionalnosti, neuronske mreže nisu nadmašile modele temeljene na slučajnim šumama. Kao konačni rezultat, ansambli su se pokazali kao najbolji prediktori vjerojatnosti defaulta. Koristeći boosting ansamble, uspjeli su sustavno razmotriti kombinirani utjecaj više značajki, što je rezultiralo poboljšanom predvidljivošću u usporedbi s tradicionalnim metodama ocjenjivanja kreditnog skoringa koje se oslanjaju na segmentaciju portfelja.

Na slici 4.2 vidimo usporedbu regresijskog modela s boosting modelom na testnom skupu podataka.



Slika 4.2: Stvarni ishod je da *Extreme Gradient Boosting* model pokazuje značajno bolje rezultate: razlika u odlukama rezultira 39% manje loših klijenata i gotovo 1% više prihvaćenih dobrih klijenata.

Zaključak

Regresija nije uspješna jer se u ovom slučaju dvije značajke, koje se obično povezuju s dobrim klijentima, neovisno od ostalih varijabli povećavaju. XGB model može ”umiriti” tipično dobre značajke ovisno o drugima, što rezultira točnijim rezultatima.

Zaključujemo da bi banke, tj. tvrtke, trebale prihvatiti tehnologije poput strojnog učenja i znanosti o podacima kako bi ostale konkurentne na ovom tržištu. Pokazano je da strojno učenje nadmašuje tradicionalne tehnike regresije u pogledu prediktivnosti kredinog skoringa. Preciznija procjena rizika, za korisnika rezultira povećanom mogućnošću dobivanja

kredita i smanjenjem stope defaulta kredita, dok banke dobivaju veći udio tržišta, veće stope konverzije zahtjeva za kreditom, kraće vrijeme odobrenja i niže troškove obrade.

Bibliografija

[1] K. M. Ramachandran, C. P. Tsokos, *Mathematical Statistics with Applications*, Elsevier Academic Press, USA, 2009.

[2] M. Huzak, *Vjerojatnost i matematička statistika, predavanja*, dostupno na `http://aktuari.math.pmf.unizg.hr/docs/vms.pdf` (lipanj 2023.).

[3] T. A. Horvat, *Uvod u Bayesovu statistiku*, Diplomski rad, Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet (Matematički odsjek), 2020.

[4] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning*, Springer, USA, 2008.

[5] T.M. Mitchell, *Machine learning*, McGraw-Hill Science, Portland, 1997.

[6] R. E. Schapire, *A brief introduction to boosting*, Citeseer, USA, 1999.

[7] S. Hartshorn, *Machine Learning With Random Forests And Decision Trees*, Kindle edition, 2016.

[8] N. Sandrić, Z. Vondraček, *Vjerojatnost, predavanja*, dostupno na `https://www.pmf.unizg.hr/images/50025978/skripta_nikola_vondra.pdf` (lipanj 2023.).

[9] D. Halcoussis, *Understanding Econometrics*, South-Western, Ohio, 2005.

[10] P. Škiljan, *Linearna regresija u aktuarstvu*, Diplomski rad, Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet (Matematički odsjek), 2019.

[11] Ž. Pauše, *Uvod u matematičku statistiku*, Školska knjiga, Zagreb, 1993.

[12] L. Thomas, D. Edelman, J. Crook, *Credit Scoring and Its Applications*, SIAM, Philadelphia, 2002.

[13] J. Márquez, *An Introduction to Credit Scoring For Small and Medium Size Enterprises*, dostupno na http://siteresources.worldbank.org/EXTLACOFFICEOFCE/Resources/870892-1206537144004/MarquezIntroductionCreditScoring.pdf (lipanj 2023.).

[14] D. Barić, M. Gaudart, S. Slijepčević, T. Vlaić, *PSD2 information has the power to transform credit scoring and lending decisions*, preprint, 2018.

[15] M. Gaudart, S. Slijepčević, *Open Banking current account information has the power to transform credit scoring and lending decisions*, Credit Scoring and Credit Control Conference XVIII, Edinburgh, 2019., dostupno na https://www.crc.business-school.ed.ac.uk/sites/crc/files/2020-10/V81-Open-Banking-Current-Account-Information-Gaudart2-1.pdf (lipanj 2023.).

Sažetak

Ovaj diplomski rad proučava problematiku izračuna kreditnog skoringa pomoću prediktivnih modela strojnog učenja. U početku uvodimo osnovne koncepte vjerojatnosti i statistike nužne za daljnje razumijevanje rada. U nastavku rada upoznajemo prediktivne modele, poput jednostavnih regresijskih modele i kompleksnijih modela strojnog učenja. Za kraj rada pokazujemo primjenu navedenih modela u bankarstvu, točnije pri izračunu kreditne sposobnosti klijenta, te navodimo primjer iz prakse u kojem se proučava prediktivna moć transakcijskih podataka tekućeg računa za izračun kreditnog rizika.

Summary

This thesis examines the issue of credit scoring calculation using predictive models in machine learning. We begin by introducing the basic concepts of probability and statistics, which are necessary for further understanding of the thesis. In the subsequent sections, we explore predictive models such as simple regression models and more complex machine learning models. Finally, we demonstrate the application of these models in the banking sector, specifically in assessing a client's creditworthiness. We also provide a practical example where the predictive power of transactional data from a current account is studied for credit risk calculation.

Životopis

Rođen sam u Puli, 16. svibnja 1995. godine. Školovanje započinjem u Osnovnoj školi Jurja Dobrile u Rovinju, nakon koje upisujem Prirodoslovno-matematičku gimnaziju u Srednjoj školi Zvane Črnje, također u Rovinju. Po završetku srednjoškolskog obrazovanja 2014. godine, odlazim u Zagreb gdje upisujem preddiplomski studij matematike na Prirodoslovno-matematičkom fakultetu. Zvanje sveučilišnog prvostupnika matematike stječem 2018. godine, kada upisujem i diplomski studij Matematičke statistike na istom fakultetu. Dvije godine kasnije odlazim u Njemačku, gdje u sklopu Erasmus+ programa upisujem zimski semestar na Tehničkom sveučilištu u Dresdenu. Tamo razvijam poseban interes za Data Science područje pa se, po povratku u Zagreb, zapošljavam u branši kao podatkovni analitičar, inženjer i znanstvenik.

U slobodno vrijeme se bavim raznim fizičkim aktivnostima, najrađe u prirodi i uz more. Posebno volim provoditi vrijeme uz svoju obitelj i prijatelje.