Statističke metode u planiranju farmaceutskih ispitivanja

Radnić, Marija

Master's thesis / Diplomski rad

2016

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Zagreb, Faculty of Science / Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:217:124786

Rights / Prava: In copyright / Zaštićeno autorskim pravom.

Download date / Datum preuzimanja: **2024-03-28**

Repository / Repozitorij:

Repository of the Faculty of Science - University of Zagreb

SVEUČILIŠTE U ZAGREBU
PRIRODOSLOVNO–MATEMATIČKI FAKULTET
MATEMATIČKI ODSJEK

Marija Radnić

STATISTIČKE METODE U PLANIRANJU FARMACEUTSKIH ISPITIVANJA

Diplomski rad

Voditelji rada:
prof. dr. sc. Siniša Slijepčević

Zagreb, veljača 2016.

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. _____

Sadržaj

Sadržaj **iii**

Uvod **1**

1 Osnovni pojmovi vjerojatnosti i statistike **2**
1.1 Osnovni pojmovi teorije vjerojatnosti 2
1.2 Osnovni pojmovi statistike 6

2 Analiza PRIM metode **13**
2.1 Motivacija 13
2.2 Primjene 16
2.3 Interpretabilnost 18
2.4 Pokrivanje 19
2.5 Indukcija kocaka 20
2.6 *Patient rule induction* 22
2.7 Pravilo zaustavljanja 24
2.8 Ulazne varijable 25
2.9 Kriteriji cijepanja 26
2.10 Primjer 28

3 Usporedba metoda **32**
3.1 Algoritmi pokrivanja 32
3.2 Indukcija stabla odlučivanja 32
3.3 PRIM vs CART 33

4 Primjena PRIM metode - primjer u medicini **36**
4.1 Analiza i rezultati 39

Bibliografija **48**

Uvod

Uspješna istraživanja i razvoj novih lijekova važna su za budućnost farmaceutske industrije, te za ljudsko zdravlje. Trenutno, oko 95% novih lijekova padnu pri kliničkim ispitivanjima, a svako košta oko 800 milijuna dolara ($). Dakle, svota potrebna za testiranje novog lijeka je pozamašna. Kao rezultat toga, industrija je u potrazi za novim inovativnim načinima za poboljšanje vjerojatnosti uspjeha istraživanja.

Jedan poseban aspekt razvoja lijekova koji farmaceutske tvrtke moraju uzeti u obzir je sposobnost prepoznavanja podskupine bolesnika, koji će vjerojatno izvući dodatnu korist od liječenja kako bi preciznije odredila tretmane koji odgovaraju određenim skupinama ljudi. Ti klasifikatori će biti korisni za razvoj medicine te služiti kao temelj za izradu istraživanja u svrhu razvoja lijekova.

S tom motivacijom, u ovom radu je opisana jedna od metoda koja može poslužiti toj svrsi - PRIM metoda (eng. *Patient Rule Induction Method*). PRIM (Friedman, Fisher, 1998) je, kako sami naziv kaže, strpljiva induktivna metoda temeljena na indukciji pravila (kocaka) koja pri svakom koraku omogućuje evaluaciju napravljenog te određenu razinu popravka modela na način da se prvo izvodi eliminacija ulaznih varijabli u particije, te se zatim eliminirane varijable ponovno preispituju i vraćaju u model ukoliko se utvrdi veća značajnost pri takvom postupku. Detaljnije o PRIM metodi i samoj motivaciji u razradi ovog rada.

U prvom poglavlju ovog rada podsjetit ćemo se osnovnih pojmova iz vjerojatnosti i statistike koji su potrebni za bolje razumijevanje daljnjeg teksta.

U drugom poglavlju opisati ćemo PRIM metodu, prikazati primjenu na par manjih primjera te ćemo ju u idućem poglavlju usporediti s drugim sličnim metodama koristeći iste primjere.

U zadnjem poglavlju prikazan je jedan primjer primjene metode na primjeru iz medicine te su opisani rezultati i odabir način odabira rizičnih faktora i parametara metode.

Poglavlje 1

Osnovni pojmovi vjerojatnosti i statistike

1.1 Osnovni pojmovi teorije vjerojatnosti

Klasična definicija vjerojatnosti a posteriori: Ako slučajni pokus zadovoljava uvjet statističke stabilnosti relativnih frekvencija, tada se vjerojatnost a posteriori proizvoljnog događaja A vezanog uz taj pokus definira kao realan broj $P(A)$, $0 \leq P(A) \leq 1$, oko kojeg se grupiraju, odnosno kojemu teže, relativne frekvencije tog događaja.

Klasična definicija vjerojatnosti a priori: Neka imamo slučajni pokus konačno mnogo elementarnih događaja i neka su svi ti elementarni događaji jednako mogući. Tada je vjerojatnost proizvoljnog događaja vezanog uz taj pokus broj elementarnih događaja povoljnih za taj događaj podijeljen s ukupnim brojem elementarnih događaja.

Uvedimo prvo osnovnu definiciju vjerojatnosti na diskretnom prostoru, definiciju slučajne varijable te njenih momenata. Moment je karakteristična funkcija za slučajne varijable i često je predmet aproksimacije u prediktivnoj analizi pa ćemo se zato ovdje prisjetiti definicije momenata.

Definicija 1.1.1. *Neka je Ω neprazan skup i F σ-algebra na ω. Vjerojatnost je funkcija $\mathbb{P} : F \to \mathbb{R}$ sa sljedećim svojstvima:*
(P1) $0 \leq \mathbb{P}(A) \leq 1, \forall A, \forall F$,
(P2) $\mathbb{P}(\Omega) = 1$,

(P3) $A_1, A_2, ... \in F$ *međusobno disjunktni* $\Rightarrow \mathbb{P}(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mathbb{P}(A_n)$.
$(\Omega, F, \mathbb{P})$ *zovemo vjerojatnosni prostor.*

Slučajne varijable

Definicija 1.1.2. *Neka je* $(\Omega, P(\Omega), \mathbb{P})$ *diskretni vjerojatnosti prostor. Funkciju* $X : \Omega \to \mathbb{R}$ *zovemo slučajna varijabla.*

Kod slučajne varijable nas zanima vjerojatnost da ona poprimi neku vrijednost, tj. za $a \in \mathbb{R}$ *zanima nas vjerojatnost događaja*

$$\{X = a\} = X^{-1}(\{a\}) = \{\omega \in \Omega : X(\omega) = a\}.$$

Općenitije, za $B \subseteq^{\mathbb{R}}$ *zanima nas vjerojatnost događaja*

$$\{X \in B\} = X^{-1}(\{B\}) = \{\omega \in \Omega : X(\omega) \in B\}.$$

Napomena 1.1.3. *Ako je* $g : \mathbb{R} \to \mathbb{R}$ *proizvoljna funkcija i* X *slučajna varijabla, onda je i* $g \circ X$ *(kraće* $g(X)$*) slučajna varijabla jer je* $g \circ X : \Omega \to \mathbb{R}$.

Definicija 1.1.4. *Neka je* $(\Omega, P(\Omega), \mathbb{P})$ *diskretni vjerojatnosti prostor i* X *slučajna varijabla na njemu. Ako red*

$$\sum_{\omega \in \Omega} X(\omega)\mathbb{P}(\omega)$$

apsolutno konvergira, tj ako konvergira red $\sum_{\omega \in \Omega} |X(\omega)|\mathbb{P}(\omega)$*, onda kažemo da slučajna varijabla* X *ima matematičko očekivanje*

$$\mathbb{E}X = \sum_{\omega \in \Omega} X(\omega)\mathbb{P}(\omega).$$

Teorem 1.1.5. *Neka je* $(\Omega, P(\Omega), \mathbb{P})$ *diskretni vjerojatnosti prostor i*

$$X \sim \begin{pmatrix} x_1 & x_2 & ... & x_n & ... \\ p_1 & p_2 & ... & p_n & ... \end{pmatrix}$$

slučajna varijabla na njemu. Redovi

$$\sum_{\omega \in \Omega} X(\omega)\mathbb{P}(\omega), \sum_{i \in \mathbb{N}} x_i pi$$

istovremeno ili apsoluno konvergiraju ili apsolutno divergiraju. U slučaju apsolutne konvergencije sume su im jednake i vrijedi

$$\mathbb{E}[X] = \sum_{\omega \in \Omega} X(\omega)\mathbb{P}(\omega) = \sum_{i \in \mathbb{N}} x_i pi$$

Napomena 1.1.6. *Ako je* $g : \mathbb{R} \to \mathbb{R}$ *onda je*

$$\mathbb{E}[g(X)] = \sum_i g(a_i)p_i.$$

Definicija 1.1.7. *Slučajna varijabla* X *na vjerojatnosnom prostoru* $(\Omega, F, \mathbb{P})$ *je neprekidna slučajna varijabla ako postoji funkcija* $f : \mathbb{R} \to \{0, +\infty\rangle$ *takva da je*

$$\mathbb{P}(X \leq a) = \int_{-\infty}^{0} f(t)dt, \text{za sve } a \in \mathbb{R}.$$

Funkcija f *se zove funkcija gustoće slučajne varijable* X*.*

Napomena 1.1.8. *(a) Može se pokazati da je*

$$\mathbb{P}(X \in B) = \int_B f(t)dt, \text{za svaki } B \subseteq \mathbb{R}.$$

Npr.

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f(t)dt, \text{za svea}, b \in \mathbb{R}, a < b.$$

(b) Za $B = \mathbb{R}$ *iz (a) slijedi*

$$1 = \mathbb{P}(X \in \mathbb{R}) = \int_{-\infty}^{\infty} f(t)dt.$$

Dakle, da bi f *bila funkcija gustoće neke neprekidne slučajne varijable, mora vrijediti:*

$$f(x) \geq 0, \text{za sve } x \in \mathbb{R}, \int_{-\infty}^{\infty} f(t)dt = 1$$

(c) Vrijedi

$$\mathbb{P}(X = a) = \int_a^a f(t)dt = 0$$

pa je

$$\mathbb{P}(X \leq a) = \mathbb{P}(X < a) + \mathbb{P}(X = a) = \mathbb{P}(X < a) + 0 = \mathbb{P}(X < a)$$

i slično je

$$\mathbb{P}(a \leq X \leq b) = \mathbb{P}(a \leq X < b) = \mathbb{P}(a < X \leq b) = \mathbb{P}(a < X < b)$$

Definicija 1.1.9. *Funkcija distribucije neprekidne slučajne varijable X je definirana s*

$$F(x) := \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f(t)dt, x \in \mathbb{R}.$$

Napomena 1.1.10. *(a) Ako je f neprekidna, onda je*

$$f = F'.$$

(b) Vrijedi:

$$F(-\infty) := \lim_{x \to -\infty} F(x) = 0, F(+\infty) := \lim_{x \to +\infty} F(x) = 1.$$

Definicija 1.1.11. *Neka je X neprekidna slučajna varijabla s funkcijom gustoće f. Matematičko očekivanje slučajne varijable X je definirano sa*

$$\mathbb{E}(X) := \int_{-\infty}^{\infty} xf(x)dx,$$

ukoliko gornji integral apsolutno konvergira.

Definicija 1.1.12. *Neka je $(\Omega, P(\Omega), \mathbb{P})$ diskretni vjerojatnosti prostor, X slučajna varijabla na njemu i $r > 0$. Definiramo sljedeće momente slučajne varijable X :*

1. r-ti moment:

$$\mathbb{E}[X^r] = \sum_{i \in \mathbb{N}} x_i^r p_i$$

ako $\mathbb{E}[X^r]$ postoji,

2. r-ti centralni moment:

$$m_r = \mathbb{E}[(X - \mathbb{E}[X])^r]$$

ako $\mathbb{E}[(X - \mathbb{E}[X])^r]$ postoji,

3. r-ti apsolutni moment:

$$\mathbb{E}[|X|^r] = \sum_{i \in \mathbb{N}} |x_i|^r p_i$$

ako $\mathbb{E}[|X|^r]$ postoji.

Definicija 1.1.13. *Varijanca slučajne varijable X je definirana kao drugi centralni moment*

$$VarX = \mathbb{E}[(X - \mathbb{E}X)^2].$$

Propozicija 1.1.14. *(Čebiševljeva nejednakost) Neka je X slučajna varijabla koja ima varijancu σ^2 i neka je $k > 0$. Tada vrijedi:*

$$\mathbb{P}(|X - \mu|) \geq k\sigma) \leq \frac{1}{k^2}$$

gdje je μ očekivanje slučajne varijable X.

Za diskretnu slučajnu varijablu iz Bernoulijeve sheme vrijedi sljedeći teorem.

Teorem 1.1.15. *(Integralni Moivre-Laplaceov teorem) Neka je $0 < p < 1$ i $X_n \sim B(n,p)(n \in \mathbb{N})$. Tada za proizvoljne $a, b \in \mathbb{R}, a < b$ vrijedi*

$$\lim_{n\to-\infty} \mathbb{P}(a \leq \frac{X_n - np}{\sqrt{npq}} \leq b) = \frac{1}{2\pi} \int_a^b e^{-\frac{x^2}{2}} dx.$$

Neka je $(X_n, n \in \mathbb{N})$ niz nezavisnih slučajnih varijabli i neka je $S_n = \sum_{k=1}^n X_k, n \in \mathbb{N}$. Opisat ćemo granično ponašanje niza $(S_n, n \in \mathbb{N})$ u smislu konvergencije po distribuciji.

Teorem 1.1.16. *(Levy) Neka je $(X_n, n \in \mathbb{N})$ niz nezavisnih, jednako distribuiranih slučajnih varijabli s očekivanjem m i varijancom σ^2, $0 < \sigma^2 < \infty$ i neka je $S_n = \sum_{k=1}^n X_k, n \in \mathbb{N}$.. Tada vrijedi*

$$\frac{S_n - \mathbb{E}S_n}{\sigma\sqrt{n}} \xrightarrow{D} N(0,1), n \to \infty.$$

1.2 Osnovni pojmovi statistike

Statistika je skup ideja i metoda koje se upotrebljavaju za prikupljanje i interpretaciju podataka u nekom području istraživanja, te za izvođenje zaključaka u situacijama gdje su prisutne nesigurnosti i varijacije.

ELEMENTI ZNANSTVENOG ISTRAŽIVANJA

1. Specifikacija cilja. Ciljevi istraživanja su razni, primjerice istraživanje koliko studenti prosječno potroše novca za slobodne aktivnosti tijekom tjedna, ili istraživanje kemijskih svojstava krutih otpadaka neke tvornice i njihov utjecaj na okolinu.

2. Sakupljanje informacija (podataka). Informacije se obično sakupljaju u obliku podataka koji numerički mjere neke karakteristike.

3. Analiza podataka. Pažljiva analiza podataka kritična je za potvrdu da je dobiveno novo znanje te za vrednovanje zaključaka.

ULOGA STATISTIKE U ZNANSTVENOM ISTRAŽIVANJU

1. Dizajn eksperimenta (experimental design) je grana statistike koja se bavi planiranjem eksperimenta i sakupljanjem podataka. Primjerice u mnogim područjima znanosti eksperimenti su skupi te je unaprijed potrebno pažljivo odrediti tip i količinu potrebnih podataka.

2. Deskriptivna statistika (descriptive statistics) ili opisna statistika je grana statistike koja se bavi predočavanjem i opisivanjem glavnih karakteristika sakupljenih podataka (tablice, grafikoni, histogrami, srednje vrijednosti,...).

3. Statističko zaključivanje (inferential statistics) je vrednovanje informacija sadržanih u podacima i ocjena novog znanja dobivenog iz tih podataka (procjena parametara promatrane populacije, testiranje statističkih hipoteza,...).

Budući za validaciju modela koristimo χ^2-test, opisati ćemo ga ukratko u idućem potpoglavlju.

Populacija i uzorak. Statistika i parametar

Neka je X varijabla/statističko obilježje koje izučavamo. Cilj statističke analize je na osnovi uzorka izvesti određene zaključke o populacijskoj razdiobi od X.

Definicija 1.2.1. *Statistički uzorak duljine n za X je niz od n nezavisnih, jednako distribuiranih slučajnih varijabli*

$$X_1, X_2, ..., X_n$$

kojima je distribucija jednaka (populacijskoj) razdiobi varijable X. Realizaciju slučajnog uzorka (opažene vrijednosti) od $X_i, i = 1, ..., n$ zovemo uzorkom.

Definicija 1.2.2. *Neka je X statistička varijabla čiju populacijsku distribuciju izučavamo, te neka je $X_1, X_2, ..., X_n$ slučajni uzorak za X iz te populacije. Parametrom razdiobe od X nazivamo onu vrijednost (broj, vektor, graf,...) koja je funkcija populacijske razdiobe od X.*

Statistika je funkcija slučajnog uzorka. Statistike su slučajne varijable. Njihova razdioba se zove uzoračka radioba.

Statistički test

Promatramo statističko obilježje X. Statistička hipoteza je bilo koja pretpostavka o populacijskoj razdiobi od X.

Kažemo da je statistička hipoteza jednostavna ukoliko jednoznačno određuje razdiobu od X. U suprotnom kažemo da je složena.

Želimo na osnovi realizacije slučajno uzorka od X donijeti odluku hoćemo li odbaciti ili ne odbaciti hipotezu. Postupak donošenja odluke zove se testiranje statističkih hipoteza.

Razlikujemo dvije vrste statističkih testova: parametarski - testiramo hipoteze o parametrima poznate razdiobe neke slučajne varijable; neparametarski - testiramo hipoteze o funkciji razdiobe neke slučajne varijable.

Primijetimo da uz osnovnu ili nul-hipotezu (H_0) postoji njoj alternativna hipoteza (H_1). Budući da sve odluke bazirane na uzorcima iz populacije nisu 100% pouzdane, ni zaključak (odluka) statističkog testa nije 100% pouzdan. Dakle, može se dogoditi da je zaključak testa pogrešan. Pogreška koju činimo kada odbacujemo H_0, a ona je istinita je pogreška prve vrste. Pogreška koju činimo kada ne odbacujemo H_0, a istinita je H_1, je pogreška druge vrste.

Test će u potpunosti biti sproveden ako možemo procijeniti vjerojatnosti mogućih pogrešaka u zaključku testa. Razumno je zahtjevati test kojemu se mogu kontrolirati vjerojatnosti obiju grešaka. To nije moguče jer smanjivanjem vjerojatnosti pogreške prve vrste povećava se vjerojatnost pogreške druge vrste i obratno. S druge strane, u velikoj većini slučajeva moguće je zadanu razinu značajnosti testa α ($\alpha \in \langle 0, 1\rangle$) među testovima kojima vjerojatnost pogreške prve vrste ne prelazi broj α naći (konstruirati) test s najmanjom vjerojatnosti pogreške druge vrste.

Neka je $X_1, X_2, ..., X_n$ slučajni uzorak za X i $\mathbf{X} := (X_1, X_2, ..., X_n)$. Tada su realizacije $\mathbf{x} := (x_1, x_2, ..., x_n)$ tog uzorka elementi od $\mathbb{R}_n$.

Definicija 1.2.3. *Test hipoteze H_0 u odnosu na alternativu H_1 je preslikavanje $\tau : \mathbb{R}_n \to \{0, 1\}$.*

Interpretacija. Ako je za realizaciju $\mathbf{x}$ uzorka $\mathbf{X}$ vrijedi $\tau(\mathbf{x}) = 1$, tada odbacujemo H_0 u korist H_1, a ako je $\tau(\mathbf{x}) = 0$, tada ne odbacujemo H_0 u korist H_1. Tada je

$$C := \tau^{-1}(1) = \{\mathbf{x} \in \mathbb{R}_n : \tau(\mathbf{x}) = 1\}$$

područje realizacija uzoraka za koje je H_0 odbacuje u korist H_1. C se naziva kritično područje za test τ.

Neka je $f(x \mid \theta)$ razdioba slučajne varijable X u ovisnosti o parametru θ, $X \sim f(x \mid \theta), \theta \in \Theta$. Vjerodostojnost od θ je preslikavanje $L(\theta \mid x) = \prod_{i=1}^{n} f(x_i \mid \theta)$.

Preslikavanje $\gamma : \Theta \to [0, 1]$ definirano sa

$$\gamma(\theta) := \mathbb{E}_\theta [\tau(\mathbf{X})] = \mathbb{P}(\{\mathbf{X} \in C) = \int_C L(\theta \mid \mathbf{x}) d\mathbf{x}$$

zove se jakost testa τ.

Interpretacija. Ukoliko je θ_1 vrijednost parametra za koju je H_1 istinito, jakost testa $\gamma(\theta_1)$ je sposobnost testa da odbaci H_0 ako je H_0 neistinita hipoteza.

Neka su $H_0 : \theta \in \Theta_0, H_1 : \theta \in \Theta_1$. Preslikavanje $\alpha : \Theta_0 \to [0,1]$ definirano sa:

$$\alpha(\theta) := \gamma(\theta) = \mathbb{P}_\theta(\mathbf{X} \in C)$$

je vjerojatnost pogreške prve vrste.

$$\alpha_\tau = \sup_{\theta \in \Theta_0} \alpha(\theta)$$

je značajnost testa τ. Kažemo da test ima razinu značajnosti α ukoliko mu je značajnost manja ili jednaka α.

Preslikavanje $\beta\Theta_1 \to [0,1]$ definirano sa:

$$\beta(\theta) := 1 - \gamma(\theta) = \mathbb{P}_\theta(\mathbf{X} \notin C)$$

je vjerojatnost pogreške druge vrste.

Definicija 1.2.4. *Kažemo da je test τ uniformno najjači ako za svaki drugi test τ' takav da je $\alpha_{\tau'} \leq \alpha_\tau$, vrijedi da je $\gamma_{\tau'}(\theta) \leq \alpha_\tau(\theta)$ za sve θ.*

Pretpostavimo da želimo testirati hipoteze o dvije vrijednosti parametra θ. Tada vrijedi sljedeći rezultat.

Lema 1.2.5. *(Neyman, Pearson) Neka je $k > 0$ takav broj da za skup*

$$C = \{\mathbf{x} \in \mathbb{R}^n : L(\theta_0 \mid \mathbf{x}) \leq kL(\theta_1 \mid \mathbf{x})\}$$

vrijedi da je $\int_C L(\theta_0 \mid \mathbf{x})d\mathbf{x} = \alpha$ za zadani $\alpha \in \langle 0,1\rangle$.

Ako za neki drugi $B \subset \mathbb{R}_n$ vrijedi da je $\int_B L(\theta_0 \mid \mathbf{x})d\mathbf{x} \leq \alpha$ tada je nužno

$$\int_B L(\theta_1 \mid \mathbf{x})d\mathbf{x} \leq \int_C L(\theta_1 \mid \mathbf{x})d\mathbf{x}$$

Interpretacija. Test $\tau(\mathbf{x}) := \mathbf{1}_C(\mathbf{x})$ za C iz N-P leme je uniformno najjači test za testiranje jednostavne hipoteze H_0 u odnosu na jednostavnu alternativu H_1.

Testovi o parametrima normalne razdiobe $N(\mu, \sigma^2)$

Jednostrani z-test
Neka je $X \sim N(\mu, \sigma^2)$ i neka je varijanca σ^2 poznata. Testiramo $H_0(\mu = \mu_0)$ naspram $H_1(\mu < \mu_0)$.

Pokaže se da je uniformno najjači test za testiranje tih hipoteza na razini značajnosti α, dan testnom statistikom

$$Z = \frac{\overline{X} - \mu_0}{\sigma_0}\sqrt{n}$$

i kritičnim područjem $z \leq -z_\alpha$, pri čemu je μ_0 zadani broj, a z vrijednost od Z.

Analogno, kada je alternativna hipoteza $H_1(\mu > \mu_0)$, kritično područje je $z \geq z_\alpha$.

Dvostrani z-test
Neka je $X \sim N(\mu, \sigma^2)$ i neka je varijanca σ^2 poznata. Testiramo $H_0(\mu = \mu_0)$ naspram $H_1(\mu \neq \mu_0)$.

Pokaže se da je uniformno najjači test za testiranje tih hipoteza na razini značajnosti α, dan testnom statistikom

$$Z = \frac{\overline{X} - \mu_0}{\sigma_0}\sqrt{n}$$

i kritičnim područjem $|z| \geq z_{\frac{\alpha}{2}}$, pri čemu je μ_0 zadani broj, a z vrijednost od Z.

χ^2-test i prilagodba modela podacima

Neka je X statističko obilježje koje izučavamo. Pretpostavimo da je X slučajna varijabla.

Analizom podataka ili iz prirode opažane pojave obično možemo naslutiti kakva je populacijska razdioba od X.

Dakle, želimo testirati hipoteze oblika:

$$H_0 : X \sim \begin{pmatrix} a_1 & a_2 & ... & a_k & ... \\ p_1(\theta) & p_2(\theta) & ... & p_k(\theta) & ... \end{pmatrix}$$

Oznakom $p_i(\theta)$ želimo naglasiti moguću ovisnost razdiobe o parametrima danim vektorom θ.

Ovisno o tome jesu li vrijednosti (svih) parametara određeni s H_0 ili ne, H_0 će biti jednostavna ili složena hipoteza.

Standardni statistički test kojim se testiraju takve nul-hipoteze (u odnosu na alternativu koja je negacija H0) je *Pearsonov* χ^2*-test.*

Neka je $X_1, X_2, ..., X_n$ slučajni uzorak za X.

Opažene frekvencije:

$$N_j := \sum_{i=1}^{n} 1_{\{X_i = a_j\}}, j = 1, 2, ..., k$$

Primijetimo:

$$(N_1, N_2, ..., N_k) \overset{H_0}{\sim} polinomijalna(n; p_1(\theta), ..., p_k(\theta))$$

Očekivane frekvencije:

$$n_j(\theta) := \mathbb{E}_\theta [N_j] = np_j(\theta), j = 1, 2, ..., k$$

Neka je:

$$D(\theta) = \sum_{j=1}^{k} \frac{(N_j - n_j(\theta)^2}{n_j(\theta)}$$

Ako je H_0 jednostavna hipoteza, $\theta = \theta_0$ je poznato (zadano s H_0) i ne treba ga procijeniti iz uzorka.

U tom slučaju je testna statistika:

$$H \equiv D(\theta_0)$$

Broj stupnjeva slobode *df* je:

$$df = k - 1.$$

Ako je H_0 složena hipoteza (tj. θ je d-dimenzionalni nepoznati parametar), procijenimo ga iz uzorka, ili metodom maksimalne vjerodostojnosti ili metodom minimuma χ^2.

Teorem 1.2.6. *Ako je* H_0 *točna hipoteza,*

$$H \overset{D}{\to} \chi^2(df), N \to +\infty.$$

Dakle, za velike uzorke,

$$H \overset{D}{\approx} \chi^2(df)$$

Za zadanu razinu značajnosti α, H_0 odbacujemo ukoliko je opažena vrijednost h od H:

$$h \geq \chi^2_\alpha(df)$$

gdje je $\chi^2_\alpha(df)$ $(1-\alpha)$-kvantil $\chi^2(df)$-distribucije.

Napomena 1.2.7. *Ako je za neki j očekivana frekvencija $n_j < 5$, združimo taj sa susjednim(a) razredom(ima) tako da novodobiveni razred zadovoljava uvjet da mu je očekivana frekvencija barem 5. U tom slučaju se smanji i broj razreda k.*

Poglavlje 2

Analiza PRIM metode

Većina postupaka analize podataka može se formulirati kao problem optimizacije funkcije. Često je, eksplicitno ili implicitno, cilj analize pronaći kombinacije vrijednosti za ulazni set varijabli koji rezultira povećanjem (ili smanjenjem) zavisne (izlazne) varijable. Posebno, traže se podskupovi ulaznih varijabli za koje izlazna varijabla poprima veće/manje vrijednosti od njene srednje vrijednosti nad cijelim prostorom ulaznih varijabli. Dodatno, poželjno je da dobivene rezultate možemo što jednostavnije interpretirati. U ovom poglavlju predstaviti ćemo PRIM metodu (eng. *Patient Rule Induction Method*) koja se koristi upravo u tu svrhu. Kao što sam naziv metode kaže, PRIM je metoda bazirana na indukciji pravila (odnosno kocaka) koristeći strpljivu strategiju, dakle manje korake, u ocjeni relevantnih varijabli za aproksimaciju funkcije cilja. Postupcima cijepanja (eng. *peeling*) i lijepljenja (eng. *pasting*) određuju se one varijable koje imaju značajan utjecaj na ishod te se tako tvore pravila (kocke) koje određuju particije podataka sa svojstvenim ishodima. Prvo ćemo navesti motivaciju za konstrukciju takve metode, a zatim ju i konstruirati. Detaljnu obradu metode napravili su Friedman i Fisher (1998), te dodatno kasnije Polonik i Wang (2007).

2.1 Motivacija

Brojne analize podataka provode se metodama predikcije (prediktivna analiza). Promatramo bazu podataka koja sadrži ponovljena zapažanja vrijednosti koje poprima izlazna varijabla y uz usporedno praćenje vrijednosti ulaznih varijabli $\mathbf{x} = (x_1, \cdots, x_n)$. Cilj je iskoristiti podatke oblika

$$\{y_i, \mathbf{x}_i\}_1^N \tag{2.1}$$

kako bi odredili ponašanje vrijednosti varijable y u ovisnosti o ulaznim vrijednostima $\mathbf{x}$. Pretpostavimo da su podaci (2.1) slučajan uzorak iz neke nepoznate distribucije

s funkcijom gustoće $p(y, \mathbf{x})$. Tada se problem svodi na određivanje funkcije gustoće varijable y za svaku vrijednost varijable $\mathbf{x}$

$$p(y \mid \mathbf{x}) = \frac{p(y, \mathbf{x})}{\int p(y, \mathbf{x})\, dy}$$

koju možemo gotovo u potpunosti opisati njenim prvim momentom (očekivanjem)

$$f(\mathbf{x}) = E[y, \mathbf{x}] = \int y\, p(y, \mathbf{x})\, dy. \tag{2.2}$$

Jednakost (2.2) minimizira srednju kvadratnu pogrešku predviđanja za svaki $\mathbf{x}$

$$f(\mathbf{x}) = \arg\min_{f} E[(y - f)^2 |\, \mathbf{x}].$$

Uočimo da ako izrazimo izlaznu varijablu u obliku

$$y = E[y \mid \mathbf{x}] + (y - E[y \mid \mathbf{x}]) = f(\mathbf{x}) + \varepsilon \tag{2.3}$$

rješenje problema možemo protumačiti kao aproksimaciju ciljne funkcije $f(\mathbf{x})$ preko niza promatranja pri čemu je njena vrijednost (za svaki $\mathbf{x}$) pomaknuta za slučajan šum (*random noise*) ε. Šum predstavlja slučajnu distribuciju izlazne varijable y oko njene srednje vrijednosti $f(\mathbf{x})$, za svaki $\mathbf{x}$, i objašnjava činjenicu da određivanje niza ulaznih varijabli ne određuje jedinstveno vrijednost y budući i drugi čimbenici, koji nisu mjerljivi, te time nisu sadržani u setu ulaznih varijabli, utječu na konačnu izlaznu vrijednost.

Iako ograničenje modela na promatranje samo prvog momenta (2.2) uvelike pojednostavljuje predikciju, problem analize je i dalje značajan. Značajnost problema nalazi se u ispravnoj aproksimaciji općenite funkcije više varijabli na domeni određenoj ulaznim vrijednostima temeljenim na nekom uzorku (sa ili bez šuma). Set funkcija koje možemo točno aproksimirati s trenutno poznatim metodama i dalje je relativno malen te postoje funkcije s kojima se susrećemo u praksi koje ćemo teško moći aproksimirati u ovom trenutku.

Aproksimacija funkcija se uglavnom koristi u situacijama u kojima je cilj analize saznati (proučiti) samo neka *svojstva* ciljne funkcije. Uobičajeni postupak u takvim situacijama je pokušati procijeniti $f(\mathbf{x})$ na cijelom prostoru ulaznih vrijednosti i ustvrditi značajna svojstva iz dobivenih procjena $\widehat{f}(\mathbf{x})$. Međutim, nerijetko, ovakva strategija može biti kontraproduktivna ukoliko postoji alternativa u kojoj se direktno aproksimiraju značajna svojstva te se time dostiže veća točnost.

Jedan od primjera je (2 - class) klasifikacija. Neka izlazna varijabla y poprima dvije vrijednosti, $y = 0$ što interpretiramo kao prvu klasu i $y = 1$ što interpretiramo kao drugu klasu. Ciljna funkcija je

$$f(\mathbf{x}) = E[y \mid \mathbf{x}] = \Pr(y = 1 \mid \mathbf{x}) \tag{2.4}$$

pri čemu je

$$y = 1(f(\mathbf{x}) > 1/2), \tag{2.5}$$

gdje je $1(\cdot)$ karakteristična funkcija - poprima vrijednost 1 kada je argument funkcije istinit, a u suprotnom poprima vrijednost 0. U ovom slučaju je uobičajeno primijeniti metodu aproksimacije funkcije, koristeći procjenu $\widehat{f}(\mathbf{x})$ u (2.5) u svrhu predviđanja vrijednosti $\widehat{y}$. Vapnik (1995) i Friedman (1997) pokazali su kako ovo pokazuje slabije rezultate u odnosu na procedure koje pokušavaju direktno izračunati granicu $f(\mathbf{x}) = 1/2$, i ponekad vodi ka kontraintuitivnim rezultatima. Još jedan primjer je procjena gustoće gdje je $f(x)$ relativna vjerojatnost promatranja u x. Svojstvo koje želimo promatrati jest kumulativna funkcija distribucije

$$F(x) = \int_{-\infty}^{x} f(x')\, dx'. \tag{2.6}$$

Uvrštavanjem optimalne kernel procjene gustoće $f(x)$ u (2.6) doći ćemo do lošije procjene za $F(x)$ nego pri jednostavnom uvrštavanju *raw* podataka kao procjene gustoće (Hall, 1989)

$$\widehat{F}(x) = \frac{1}{N} \sum_{i-1}^{N} 1(x_i \leq x)$$

Još jedan trivijalni primjer je kad svojstvo koje želimo promatrati jest očekivanje, tj srednja vrijednost (*mean*) ciljne funkcije na cijelom prostoru ulaznih vrijednosti

$$\bar{f} = \int f(\mathbf{x})\, p(\mathbf{x})\, d\mathbf{x}. \tag{2.7}$$

Srednja vrijednost izlazne varijable uzorka

$$\overline{y} = \frac{1}{N} \sum_{i-1}^{N} y_i \tag{2.8}$$

će imati veću točnost nego očekivanje neke procjene $\widehat{f}(\mathbf{x})$ potekle iz podataka.

Optimizacija funkcije

Aproksimacija funkcije često se koristi kad nas zanimaju ekstremne vrijednosti (maksimum i minimum) ciljne funkcije $f(\mathbf{x})$ (2.2). Konkretnije, želimo pronaći podskup domene na kojem ciljna funkcija poprima vrijednosti mnogo veće (ili manje) od prosječne vrijednosti na cijelom prostoru ulaznih vrijednosti (domeni) (2.7). Budući

minimiziranje funkcije odgovara maksimiziranju iste funkcije, ali suprotnog (negativnog) predznaka, bez smanjenja općenitosti, možemo problem promatrati kao problem maksimiziranja funkcije. Neka je S_j niz svih mogućih vrijednosti za ulaznu varijablu x_j

$$\{x_j \in S_j\}_{j-1}^{n}. \tag{2.9}$$

Individualni S_j može predstavljati realne vrijednosti (moguće diskretne), ili kategorijske (neporedane) vrijednosti. Cijela ulazna domena S tada može biti prikazana n-dimenzionalnim (vanjskim) produktnim prostorom

$$S = S_1 \times S_2 \times \cdots \times S_n. \tag{2.10}$$

Cilj je pronaći podskup R domene S, $R \subset S$, za koji vrijedi

$$\overline{f}_R = ave_{\mathbf{x}\in R} f(\mathbf{x}) = \int_{\mathbf{x}\in R} f(\mathbf{x})\, p(\mathbf{x})\, d\mathbf{x} \,/\, \int_{\mathbf{x}\in R} p(\mathbf{x})\, d\mathbf{x} \;\; >> \;\; \overline{f} \tag{2.11}$$

pri čemu je $\overline{f}$ prosječna vrijednost na čitavoj domeni (2.7). Važno svojstvo takvog podskupa je njegova veličina (nosač)

$$\beta_R = \int_{\mathbf{x}\in R} p(\mathbf{x})\, d\mathbf{x}. \tag{2.12}$$

Pokaže se da generalno postoji tzv. trade-off između vrijednosti $\overline{f}_R$ i β_R, odnosno što veće vrijednosti poprima β_R to su manje vrijednosti $\overline{f}_R$.

Neposredne procjene vrijednosti (2.11) i ((2.12)) respektivno će biti korišteni:

$$\widehat{\beta}_R = \frac{1}{N}\sum_{\mathbf{x}_i\in R} 1(\mathbf{x}_i \in R), \;\; \overline{y}_R = \frac{1}{N\cdot\widehat{\beta}_R}\sum_{\mathbf{x}_i\in R} y_i, \tag{2.13}$$

gdje $\{y_i\}_1^N$ predstavljaju promatrane izlazne vrijednosti (2.1) (2.3).

2.2 Primjene

Izravne primjene

Mnogo je okolnosti u kojima je pažnja usmjerena na ekstreme ciljne funkcije $f(\mathbf{x})$ (2.3). U predviđanjima budućih povrata financijskog osiguranja, uglavnom želimo identificirati ona koja osiguravaju najveći povrat. Izlazna varijabla y je u ovom slučaju povrat, a ulazne varijable $\mathbf{x}$ mogu biti neki prošli povrati ili različiti ekonomski čimbenici. Vrijednost nosača β_R (2.12) tada predstavlja udio potencijalnih uloga.

U istraživanju tržišta, y može predstavljati neko ponašanje kupaca pri čemu ulazne vrijednosti $\mathbf{x}$ mogu biti različite demografske varijable. Tada je vrijednost nosača β_R veličina identificiranog tržišnog segmenta. U primjenama u medicini izlazna varijabla bi mogla biti mjera ozbiljnosti bolesti, a ulazne različiti simptomi i liječničke mjere pri čemu je cilj identificirati karakteristike najgore oboljelih pacijenata u svrhu poboljšanja budućih testiranja ili tretmana. Ovdje nosač β_R označava udio pacijenata za koje je takav tretman uopće ostvarljiv. U industriji je cilj maksimizirati neku mjeru kvalitete konačnog proizvoda (izdržljivost, dugotrajnost, itd). Tada y predstavlja promatranu mjeru kvalitete, a ulazne vrijednosti $\mathbf{x}$ su različiti parametri koji kontroliraju proces (sastojci kemikalija, temperatura, vrijeme, itd.). Nekada postoje različite karakteristične veličine $\{z_k\}_1^K$ povezane s proizvodom (voltaža, struja) koje imaju ciljne vrijednosti $\{t_k\}_1^K$. Cilj je pronaći vrijednosti kontrolnih parametara $\mathbf{x}$ koje rezultiraju karakterističnim vrijednostima bliskim ciljnim vrijednostima. Tada y možemo definirati kao

$$y = -\sum_{k-1}^{K}(z_k - t_k)^2. \tag{2.14}$$

U ovakvim procesima vrijednost β_R generalno nije značajna. U svim ovim primjerima, kao i u većini primjera analize podataka, izlazna varijabla y predstavlja mjeru buke ciljne funkcije $f(\mathbf{x})$. Niz promatranih ulaznih vrijednosti u potpunosti karakterizira sve moguće čimbenike koji utječu na izlaznu vrijednost.

Neizravne primjene

Osim primjena koje se direktno fokusiraju na optimizaciju, postoje i mnoge druge često korištene procedure analize podataka. Jedna takva je klasifikacija u kojoj stvarna izlazna vrijednost poprima kategoričke vrijednosti $\{c_k\}_1^K$. Neka postoji K ("dummy") izlaznih varijabli $\{y_k = 1(\text{class} = c_k\}_1^K$. Ulazne vrijednosti $\mathbf{x}$ predstavljaju prediktorske varijable. Cilj je identificirati one podskupove ulaznog prostora u kojima pojedino opažanje potječe iz neke od klasa. Za te podskupove vrijedi da je odgovarajuća vrijednost

$$f_k(\mathbf{x}) = E[y_k \mid \mathbf{x}] = \Pr(y_k = 1 \mid \mathbf{x})$$

veća nego u bilo kojoj drugoj klasi. Dakle, klasifikaciju možemo protumačiti kao pronalazak podskupova u kojima je svaki $f_k(\mathbf{x})$ relativno velik. Vrijednosti (2.12) i (2.13) za svaki podskup mogu biti shvaćene kao meta-parametri zajednički optimiziranih kako bi se maksimizirala točnost klasifikacije u svrhu predviđanja. Druga često korištena metoda u analizi podataka je *clustering*. U ovoj metodi je cilj pronaći ona

područja podataka koja su zastupljenija, to jest vjerojatnosna funkcija gustoće $p(\mathbf{x})$ je velika u usporedbi s nekom referentnom vrijednosti gustoće $p_0(\mathbf{x})$ (uobičajeno se uzme da je to gustoća uniformne distribucije nad cijelim skupom podataka). Područja na kojima je omjer

$$r(\mathbf{x}) = p(\mathbf{x})/p_0(\mathbf{x}) \tag{2.15}$$

velik predstavljaju lokalne koncentracije podataka odnosno tzv. *clustere*. Područja gdje je $r(\mathbf{x})$ malen ($-r(\mathbf{x})$ velik) predstavljaju rupe (*holes*) u podacima koje također mogu biti od interesa.

Omjer $r(\mathbf{x})$ možemo maksimizirati dodijeljivanjem izlazne vrijednosti $y = 1$ za svaki skup opažanja. Tada možemo generirati dovoljno velik Monte Carlo uzorak iz referentne distribucije $p_0(\mathbf{x})$ te mu tada dodijelimo izlaznu vrijednost $y = 0$. Funkcija

$$f(\mathbf{x}) = E[y \mid \mathbf{x}] = r(\mathbf{x})/(1 + r(\mathbf{x})) \tag{2.16}$$

nad sakupljenim podacima je monotona po $r(\mathbf{x})$ te maksimum/minimum predstavljaju *clustere/rupe*. Problem usko povezan s *clusteringom* jest zavisnost podataka. Kao što gornja rasprava ilustrira većina problema analize podataka može se staviti u okvir optimizacije u kojoj je ciljna funkcija promatrana s pretpostavkom šuma. Postupci koji pronalaze područja ulaznog prostora na kojima ciljna funkcija poprima relativno velike vrijednosti predstavljaju (u načelu) potencijalna rješenja za ove probleme.

2.3 Interpretabilnost

U deskriptivnoj analizi podataka interpretabilnost predstavlja bitnu stavku. Želimo restringirati izlazne podatke (rješenja) samo na one koje možemo opisati i interpretirati kroz glavne karakteristike problema, makar to oduzelo na značajnosti modela. U tu svrhu područje R iz (2.11) želimo opisati jasnim *izjavama* (logičkim uvjetima) koje opisuje vrijednosti ulaznih varijabli $\{x_j\}_1^n$. Tada područje rješenja R definiramo kao uniju jednostavno definiranih podskupova $\{B_k\}_1^K$

$$R = \bigcup_{k=1}^{K} B_k \tag{2.17}$$

Neka s_{jk} predstavlja podskup mogućih vijednosti ulazne varijable x_j, odnosno neka je $\{s_{jk} \subseteq S_j\}_1^n$ pri čemu svaki S_j predstavlja sve moguće vrijednosti x_j. Tada svaki B_k (2.17) možemo protumačiti kao kocku

$$B_k = s_{1k} \times s_{2k} \times \cdots \times s_{nk} \tag{2.18}$$

na cijelom ulaznom prostoru (2.10). Svaka kocka B_k (2.18) se može opisati kao presjek podskupova vrijednosti svake ulazne varijable

$$\mathbf{x} \in B_k = \bigcap_{j=1}^{n} (x_j \in s_{jk}) \tag{2.19}$$

Za sve ulazne varijable koje poprimaju realne vrijednosti, podskupovi su reprezentirani neprekidnim podintervalima

$$s_{jk} = [t_{jk}^-, t_{jk}^+] \tag{2.20}$$

Dakle, projekcija kocke B_k na potprostor realnih ulaznih vrijednosti je hiper-pravokutnik. Uočimo da u slučaju kada je podskup vrijednosti zapravo jednak čitavom skupu $s_{jk} \subseteq S_j$, odgovarajući faktor $x_j \in S_j$ može biti izbačen iz definicije (2.19):

$$\mathbf{x} \in B_k = \bigcap_{s_{jk} \neq S_j} (x_j \in s_{jk}) \tag{2.21}$$

Uočimo da varijable x_j za koje vrijedi $s_{jk} \neq S_j$ definiraju B_k. Primjer jedne takve definicije

$$\mathbf{x} \in B_k = \begin{cases} 18 < \text{dob} < 34 \quad \& \\ \text{bračni status} \in \{\text{samac, nevjenčan-žive zajedno}\} \quad \& \\ \text{tip kućanstva} = \text{podstanarstvo} \end{cases} \tag{2.22}$$

2.4 Pokrivanje

Uočimo iz (2.11) (2.17) da je cilj optimizacije ishoditi skup kocaka (2.21) iz podataka (2.1) koji pokrivaju prostor ulaznih varijabli za koje ciljna funkcija $f(\mathbf{x})$ poprima velike vrijednosti. To možemo postići tzv. pokrivanjem (eng. *covering*) ukoliko imamo algoritam za konstrukciju tih kocaka (vidi poglavlje 2.5). Isti algoritam za konstrukciju kocaka koristimo za kreiranje niza kocaka za određene podskupove podataka. Prva kocka B_1 inducirana je cijelim skupom podataka (2.1). Druga B_2 konstruira se nad originalnim podacima iz kojih se izuzmu podaci pokriveni u B_1: $B_2 \sim \{y_i, x_i \mid x_i \notin B_1\}$. U K-toj iteraciji kocka B_K inducirana je podacima preostalih izuzimanjem svih podataka pokrivenih s $K-1$ prethodno induciranih kocaka: $B_K \sim \left\{y_i, x_i \mid x_i \notin \cup_{k=1}^{K-1} B_k\right\}$. Ovaj postupak se nastavlja sve dok srednja vrijednost unutar kocaka

$$\overline{y}_K = ave[y_i \mid \mathbf{x}_i \in B_K \;\&\; \mathbf{x}_i \notin \bigcup_{k=1}^{K-1} B_k] \tag{2.23}$$

ne postane jako mala, primjerice manja neko globalna srednja vrijednost $\bar{y}$ (2.8), ili individualne vrijednosti

$$\beta_K = \frac{1}{N}\sum_{i=1}^{N} 1(\mathbf{x}_i \in B_K \;\&\; \mathbf{x}_i \notin \bigcup_{k=1}^{K-1} B_k) \tag{2.24}$$

ne postanu jako male.

Skup kocaka induciranih na ovaj način može se koristiti pri kreiranju konačnog područja (2.17) s odgovarajućim ciljem analize. Primjerice, možemo odabrati one kocke čija srednja vrijednost (2.23) je veća od nekog unaprijed određenog praga $\bar{y}_0$

$$R = \bigcup_{\overline{y}_k > \overline{y}_0} B_k$$

ili možemo odabrati podskup koji ishoduje najveću srednju vrijednost čitavog područja $\bar{y}_R$ (2.13) za konkretni β_t

$$\beta_R = \sum_{k=1}^{K} \beta_k \simeq \beta_t$$

pri čemu nosač β_k opisuje kocku B_k (2.24). Alternativno, možemo proučavati niz induciranih kocaka i njihove resprektivne srednje vrijednosti kao red

$$\{\overline{y}_k, B_k\}_1^K . \tag{2.25}$$

Ako je ulazna točka $\mathbf{x}$ pokrivena s više kocaka, tada je dodijeljujemo onoj koja se nalazi prva u nizu.

2.5 Indukcija kocaka

Bitan dio optimizacijskog algoritma je konstrukcija individualnih kocaka (grupacija podataka). Za dane podatke (ili podskup podataka) želimo konstruirati kocku B za koju je srednja vrijednost ciljne funkcije

$$\overline{f}_B = \int_{\mathbf{x}\in B} f(\mathbf{x})p(\mathbf{x})dx \;/\; \int_{\mathbf{x}\in B} p(\mathbf{x})dx \tag{2.26}$$

onoliko velika koliko se dozvoljava u (2.21). U tu svrhu opisati ćemo metodu koja prvo izvodi strpljivo odozgo prema dolje (*patient top-down peeling*) uzastopno uglađivanje (cijepanje) podataka te zatim odozdo prema gore (*bottom-up*) rekurzivno proširivanje.

Odozgo prema dolje cijepanje

Prvu fazu započinjemo definiranjem kocke B koji sadrži sve podatke. U svakoj iteraciji manji podskup (podkocka) b unutar trenutne kocke B se "briše" (isključuje) na način da biramo takav b^* koji će doprinijeti da iduća kocka $B - b^*$ ima najveću izlaznu vrijednost

$$b^* = \arg \max_{b \in C(b)} ave[y_i \mid \mathbf{x}_i \in B - b] \tag{2.27}$$

pri čemu $C(b)$ predstavlja klasu potencijalnih podskupova b kvalificiranih za brisanje. Trenutnu kocku B tada možemo zamijeniti sa ažuriranom verzijom

$$B \leftarrow B - b^*. \tag{2.28}$$

Postupak se tada ponavlja na toj novoj ažuriranoj kocki. Ovakav postupak *eliminacije* manjih kocaka nastavlja se dok je nosač unutar trenutnog skupa β_B manji od unaprijed određenog praga β_0 (meta-parametar procesa)

$$\beta_B = \frac{1}{N} \sum_{i=1}^{N} 1(\mathbf{x}_i \in B) \leq \beta_0. \tag{2.29}$$

Klasa $C(b)$ skupova kvalificiranih za eliminaciju ograničena je uvjetom interpretabilnosti (2.27). Svaki element b iz klase $C(b)$ definiran je jednom ulaznom varijablom x_j.

(a) svaka realna ulazna vrijednost ishoduje dva podskupa, b_{j-} i b_{j+}, takve da

$$\begin{array}{rcl} b_{j-} & = & \left\{\mathbf{x} \mid x_j < x_{j(\alpha)}\right\} \\ b_{j+} & = & \left\{\mathbf{x} \mid x_j > x_{j(1-\alpha)}\right\} \end{array} \tag{2.30}$$

pri čemu je $x_{j(\alpha)}$ α-kvantil x_j vrijednosti za podatke unutar trenutnog skupa (kocke), a $x_{j(1-\alpha)}$ je odgovarajući (1-α)-kvantil. Veličina α je još jedan od predefiniranih meta-parametara koji se obično uzima da bude jako malen ($\alpha \leq 0.1$). Uočimo da b_{j-} i b_{j+} predstavljaju donju i gornju granicu trenutne kocke B za j-tu varijablu.

(b) svaka kategorijska varijabla x_j ishoduje set skupova za izbor, po jednu za svaku svoju vrijednost s_{jm} unutar trenutnog skupa

$$b_{jm} \;=\; \left\{\mathbf{x} \mid x_j = s_{jm}\right\}, \quad s_{jm} \in S_j \tag{2.31}$$

Dakle, klasa $C(b)$ sadrži sve odabrane podskupove definirane respektivno iz ulaznih varijabli te se testira ponašanje srednje vrijednosti nad podacima bez podataka iz svakog od podskupova zasebno. Zatim se onaj podskup b_* čije izbacivanje je rezultiralo najvećom srednjom vrijednosti ostatka podataka zaista briše te time dobivamo manji skup podataka (2.28) nad kojim provodimo daljnju analizu.

Odozdo prema gore lijepljenje

Rezultat algoritma cijepanja odozgo prema dolje je kocka koja pokriva podskup ulaznih varijabli takvih da je srednja vrijednost (2.26) relativno velika. Granice rezultantnog skupa određene su vrijednostima onih varijabli koje su ishodile podskupove (2.30) i (2.31) odabrane za eliminaciju (2.27) u različitim stadijima odozgo prema dolje procedure cijepanja podataka. Osim granica određenih u zadnjem koraku, granice nastale u ranijim koracima eliminacijskog procesa nastale su neovisno o kasnijim koracima i eliminacijama. Posljedično, moguće je da konačnu (rezultantnu) kocku možemo poboljšati podešavanjem nekih od granica. To se postiže tzv. odozdo prema gore lijepljenjem (*bottom-up pasting*).

Algoritam odozdo prema gore lijepljenja je zapravo inverz procedure cijepanja (eliminacije). Počevši od rezultata dobivenog eliminacijskim algoritmom, trenutna kocka B je iterativno uvećana dodavanjem malenog podskupa b^*, $B \leftarrow B \cup b^*$. Podskupovi b^* odabiru se iz klase $b \in C(b)$ na način da odabir maksimizira izlaznu srednju vrijednost uvećanog skupa. Klasa $C(b)$ definira se na analogan način kao u eliminacijskom algoritmu.

Odozdo prema gore lijepljenje se iterativno ponavlja, povećavajući trenutnu kocku, sve dok dodavanjem novog podskupa b^* ne dođe do smanjenja srednje vrijednosti $\overline{y}_{B+b^*}$. U tom trenutku proces staje te je dobivena kocka B konačan rezultat. Iako ovaj postupak može rezultirati poboljšanjem, rijetko ima učinak koji je od važnosti. Ipak, postoje slučajevi u kojima rezultira znatnim poboljšanjem eliminacijskog rješenja.

2.6 *Patient rule induction*

Uveli smo meta-parametre koji određuju slijed metode - parametar cijepanja (eng.*peeling fraction*) α (2.30) i nosač β_0 (2.29). U ovom poglavlju diskutirati ćemo statističku važnost i odabir parametra α, a u sljedećem ćemo komentirati parametar β_0.

Da bi mogli govoriti o statističkoj važnosti i svojstvima, prvo je potrebno definirati mjeru cilja kojom možemo ustvrditi u kolikoj mjeri smo postigli željeni cilj. Za danu vrijednost β_0 formalni cilj metode indukcije kocaka možemo definirati kao određivanje maksimalne srednje vrijednosti funkcije $f(\mathbf{x})$ unutar kocke (2.26) s obzirom na parametre s_{jk} (2.19), uz ograničenje $\beta = \beta_0$. Dakle,

$$\overline{f}^* = \max_B ave[f(\mathbf{x}) \mid \mathbf{x} \in B]; \quad \beta_B = \beta_0. \tag{2.32}$$

Neka je $\widehat{f}$ procijenjeno rješenje postignuto eliminacijskim algoritmom nad podacima. Tada je jedna mjera uspješnosti očekivanje srednje-kvadratne pogreške

$$E[\overline{f}^* - \widehat{f}]^2 = (\overline{f}^* - E\widehat{f})^2 + E[\widehat{f} - E\widehat{f}]^2. \tag{2.33}$$

Vrijednost očekivanja (2.33) predstavlja kvadratnu pogrešku od $\widehat{f}$ nad svim skupovima mogućih realizacija podataka veličine N. Stvarni podaci (2.1) odabrani su nasumično između svih potencijalnih setova. Veličina $E\widehat{f}$ predstavlja prosječnu vrijednost procijenjenog rješenja nad svim setovima podataka. Prvi izraz na desnoj strani jednakosti (2.33) je devijacija tog prosjeka od stvarnih vrijednosti te predstavlja kvadratnu pogrešku (*bias-squared*). Drugi dio izraza s desne strane je varijanca rješenja nad svim setovima podataka. Ova mjera karakterizira nestabilnost procesa. Nestabilnost procesa definiramo na način da velika nestabilnost implicira da je rješenje osjetljivo i na najmanje promjene unutar podataka.

Fragmentacija podataka

Uočimo da eliminacijski algoritam nije beskonačan, odnosno da ima ograničen broj koraka budući je temeljen na fragmentaciji podataka. Svaki korak reducira količinu podataka koja ide u sljedeću iteraciju. U nekom trenutku procesa više nema dovoljno podataka i proces staje (slučaj kad vrijednost β padne ispod definirane vrijednosti parametra β_0).

Za dani set podataka veličine N broj koraka određen je količinom fragmentacije dopuštenoj u svakom koraku. U svakoj iteraciji brišu se podaci koji dovode do najvećeg povećanja srednje vrijednosti. Ukoliko iskoristimo pohlepnu (*greedy*) strategiju veći dio podataka će biti reduciran te time u konačnici imamo manji broj iteracija, ali i više prostora za pogrešku u smislu da će podaci biti pristrani (bias). Također, u tom slučaju u svakom idućem koraku iteracije imamo manju šansu popravljanja štete. Iako pohlepnim algoritmom dobivamo na lakšoj interpretaciji budući da manje ulaznih varijabli uđe u definiciju modela (2.21), gubimo na većoj pristranosti zbog manjka uključenih podataka.

Strpljivost

Problem fragmentacije u odozgo prema dolje procedurama može se riješiti uvođenjem "strpljive" (*patient*) strategije; u svakom koraku iteracije dopuštena je samo manja redukcija podataka. Ovim smanjujemo utjecaj individualnog koraka na krajnji ishod te omogućavamo da se u idućim koracima može iskoristiti trenutna struktura podataka za kompenzaciju eventualnih krivih prethodnih koraka (napravljenih zbog pristranosti). Također, na ovaj način veći broj ulaznih varijabli ima priliku ući u model.

Stupanj strpljivosti za realne ulazne varijable određen je *peeling* parametrom α (2.30). Broj koraka (*peels*) dan je s

$$L = \log\beta_0 \,/\, \log(1-\alpha)$$

Uočimo, što odaberemo manji α to je L veći, odnosno povećavamo stupanj strpljivosti. Jedan takav α je $\alpha = 1/N_B$ gdje je N_B broj opažanja u trenutnoj kocki, dakle po jedno opažanje je odbačeno u svakom koraku.

No, parametar α ima i drugu ulogu - uglađivanje (tzv. *smoothing* parametri). Cilj svakog koraka cijepanja je maksimizirati prosječnu vrijednost funkcije $f(\mathbf{x})$ na sljedećoj manjoj kocki $B - b$. Varijanca procjene je proporcionalna.

$$\frac{1}{\alpha} var[\varepsilon \mid \mathbf{x} \in b] \tag{2.34}$$

gdje je ε slučajan šum (2.3). Budući je cilj raditi eliminacije na temelju $f(\mathbf{x})$ radije nego na temelju ε, vrijednost α ne smije biti ni premala. Do sada se pokazalo da je uvjet strpljivosti ipak bitniji te da vrijednosti u intervalu $0.05 \leq d \leq 0.1$ funkcioniraju u redu.

Za kategorijske varijable stupanj strpljivosti je teže kontrolirati. Sva opažanja koja imaju jednaku vrijednost ulazne varijable promatramo zajedno. U tom slučaju strpljivost se provodi na način da se reducira po samo jedna takva vrijednost u svakom koraku.

Budući opisana metoda stavlja naglasak na strpljivost (*patience*) otud i ime *patient rule induction method* (PRIM).

2.7 Pravilo zaustavljanja

Eliminacijski proces (cijepanje) završava u trenutku kada je vrijednost nosača β_B za trenutnu kocku B ispod odabranog praga β_0 (2.29). Odabir vrijednosti β_0 ovisi o cilju analize. Primjerice, neka je cilj analize procjena "točke" u kojoj se postiže maksimum. U tom slučaju htjeli bismo da je vrijednost β_0 jako mala u svrhu postizanja što točnije procjene maksimalne vrijednosti od $f(\mathbf{x})$. Budući je funkcija opisana šumom (2.3) može biti kontraproduktivno dopustiti da vrijednost nosača β_B bude jako mala jer u tom slučaju vrijednost $\overline{y}_B$ (2.27) kojom procjenjujemo $\overline{f}_B$ (2.26) postaje manje pouzdana budući ona svakim korakom raste što ne mora biti slučaj s $\overline{f}_B$. Taj scenarij je poznatiji kao "over-fitting" fenomen i karakterističan je za metode optimizacije.

Cross-validacija

Uobičajeni postupak za savladavanje *over-fitting* ishoda je *cross-validation*. Podatke na slučajan način podijelimo u dva dijela, "learning" dio i "test" dio pri čemu je tipično *learning* dio dva puta veći od testnog. Tada eliminacijski postupak provedemo na *learning* podacima s jako malim pragom β_0, dozvoljavajući jako male vrijednosti za nosač β_B. Test podatke tada iskoristimo za procjenu izlazne srednje vrijednosti $\overline{y}_B$

za svaku kocku induciranu nad *training* podacima. Ukoliko je šum povezan s testnim podacima neovisan o onom nad *training* podacima, rezultat je nepristrana procjena od $\widehat{f}_B$ za svaki B. Kocka s najvećim testnim očekivanjem je tada uzeta kao optimalno rješenje.

Ovisnost parametara o namjeri

U nekim slučajevima korisniku odgovara da se uzme veća vrijednost β_B nauštrb manje srednje vrijednosti na kocki B. Primjerice, identificirali smo ona osiguranja koja daju maksimalne povrate, ali je nosač bio premalen da bi dozvolio dovoljno unosne uloge.

U tu svrhu je korisno uključiti krajnjeg korisnika u odabir na način da mu prezentiramo srednju vrijednost i vrijednost nosača β za svaku od L kocaka iz niza $\{B_l\}_1^L$

$$\{\overline{y}_l, \beta_l\}_1^L , \tag{2.35}$$

izračunatih nad testnim podacima. Na taj način osigurali smo da rezultat najbolje odgovara postavljenom cilju analize.

2.8 Ulazne varijable

Suvišne ulazne varijable

Osim srednjom vrijednosti i nosačem, kocka B okarakterizirana je i skupom ulaznih varijabli koji ju definiraju (2.21). Kompleksnost kocke ocijenjena je kardinalitetom tog skupa. Strategija strpljivosti uvela je mogućnost kompleksnijih kocaka, no to također može biti kontraproduktivno u smislu da nebitne varijable na taj način ulaze u definiciju kocke. Takvi suvišni unosi povećavaju kompleksnost problema te time smanjuju interpretabilnost bez da utječu na pristranost modela.

Dva su slučaja koja dovode do suvišnih varijabli, šum i kolinearnost. Šum ε (2.3) inducira varijancu procjene za svaku iteraciju. To može rezultirati odabirom nerelevantne varijable zbog neregularnih promjena unutar podataka. U tu svrhu je bitno ne odabrati premalen parametar α (2.34) kao što je raspravljano u poglavlju o strpljivosti. Iako efikasna, takva strategija ipak ne sprječava problem suvišnih varijabli u potpunosti.

Kolinearnost podataka, odnosno međusobna zavisnost vrijednosti različitih varijabli čest je problem u analizi podataka. U slučaju kolinearnosti samo jedna od kolinearnih varijabli treba ući u definiciju kocke, recimo ona s najvećim restrikcijama na njene vrijednosti.

Bilo da su suvišne varijable ušle u niz ulaznih varijabli zbog šuma ili kolinearnosti, možemo ih ručno izbaciti van. Tada je, također, jako bitna stavka korisnik koji bolje

poznaje materiju podataka te može dati uvid u vrijednost nekog podatka. S druge strane, bitno je i statistički izmjeriti važnost svake varijable u definiciji kocke. Jedna takva mjera je povećanje srednje vrijednosti nad kockom isključivanjem varijable iz nje. Neka je x_j varijabla iz definicije kocke (2.21). Možemo ju isključiti zamjenom odgovarajućeg podskupa vrijednosti s_{jk} setom svih mogućih x_j-vrijednosti S_j

$$s_{jk} \leftarrow S_j. \tag{2.36}$$

Odgovarajući porast u srednjoj vrijednosti se tad zabilježi. Ona ulazna varijabla za koju je taj porast najmanji se smtra trenutno irelevantnom te se privremeno isključuje iz definicije. Taj se postupak ponavlja dok ne ostane samo jedna varijabla te nju označavamo kao najvažnijom. Ostale varijable imaju važnost u ovisnosti o redoslijedu izbacivanja.

U većini situacija je ipak znanje i iskustvo korisnika od neprocjenjive važnosti za odabir varijabli koje ulaze u model.

Nedostatak podataka

Nerijetko se u setu podataka dogodi da nedostaju neke od vrijednosti za određene ulazne varijable. U tom slučaju želimo na neki način nadopuniti ili iskoristiti te praznine. U kontekstu PRIM metode jednostavno uzmemo da nepostojanje vrijednosti kao jednu od vrijednosti koju ta varijabla može poprimiti. Ukoliko je varijabla bila kategorijskog tipa, nema nikakvih promjena u peeling algoritmu. Ukoliko je varijabla bila realna broj podkocaka b (2.30) koje ta varijabla pridonosi u klasu $C(b)$ (2.27) se proširuje na tri

$$\begin{aligned} b_{j-} &= \left\{\mathbf{x} \mid x_j < x_{j(\alpha)}\right\} \\ b_{j+} &= \left\{\mathbf{x} \mid x_j > x_{j(1-\alpha)}\right\} \\ b_{j0} &= \left\{\mathbf{x} \mid x_j = missing\right\} \end{aligned} \tag{2.37}$$

Opažanja unutar trenutne kocke B za koja je $x_j =$ missing naravno ne ulaze u izračun $x_{j(\alpha)}$ i $x_{j(1-\alpha)}$.

Uočimo da, budući je srednja vrijednost y za ta opažanja kojima nedostaju određeni podaci vjerojatno jako blizu globalnoj ciljnoj srednjoj vrijednosti $\hat{y}$, vrijednost *missing* gotovo sigurno neće ući izbor za eliminaciju (cijepanje).

2.9 Kriteriji cijepanja

Vodeći princip eliminacijskog procesa je strpljivost. Za ulazne varijable s više različitih realnih vrijednosti stupanj strpljivosti je efektivno kontroliran meta parametrom α

(2.30). Za realne ulazne varijable s manje različitih vrijednosti uzima se da podskupovi b imaju nosač β_b blizak vrijednosti α. Bliske vrijednosti uvijek uzimamo kao jednu. Za kategorijske varijable nosač β_b ne možemo izravno kontrolirati. U zadnja dva slučaja ipak je poželjno upotrijebiti strpljivost u onoj razini u kojoj je to moguće.

Kriteriji za odabiranje podkocke

Opisani eliminacijski proces koristi srednju vrijednost $\overline{y}_{B-b}$ u sljedećoj manjoj kocki $B-b$ (2.27) kao kriterij za eliminaciju koji treba maksimizirati s obzirom na podkocku b. To je ekvivalentno maksimiziranju funckije poboljšanja definirane na način

$$I(b) = \overline{y}_{B-b} - \overline{y}_B \tag{2.38}$$

nad trenutnom kockom B. Među više podkocaka sa sličnim vrijednostima $I(b)$ biramo onu s najmanjim nosačem β_b te na taj način vipe podataka ostaje slobodno za iduće eliminacijske korake. Tada poboljšanje po isključenom nosaču postaje predmetom promatranja te strpljvost možemo provesti koristeći modificirani *peeling* kriterij

$$J(b) = I(b) \cdot P(\beta_b) \tag{2.39}$$

pri čemu je $P(\beta_b)$ monotono rastuća funkcija. Definirajmo

$$P(\beta_b) = 1/\beta_b \tag{2.40}$$

Na ovaj način kriterij (2.39) mjeri poboljšanje (2.38) prema veličini isključenog nosača β_b. Ako izrazimo (2.40) kao

$$I(b) = \frac{\beta_b}{\beta_B - \beta_b}[\overline{y}_B - \overline{y}_b] \tag{2.41}$$

gdje je $\overline{y}_b$ izlazna srednja vrijednost isključene podkocke b, a β_b je nosač kocke B, kriterij poprima oblik

$$J(b) = \frac{1}{\beta_B - \beta_b}[\overline{y}_B - \overline{y}_b] = (\overline{y}_{B-b} - \overline{y}_b)/\beta_B \tag{2.42}$$

Na ovaj način smo definirali kriterij koristeći razliku između izlazne srednje vrijednosti podataka preostalih isključivanjem podkocke b i srednje vrijednosti podkocke b. Nadalje, strpljivost možemo provesti odabirom

$$P(\beta_b) = \frac{\beta_B - \beta_b}{\beta_b} \tag{2.43}$$

te na analogan način kao gore definirati kriterij.

Svaka od opcija (2.38), (2.40) i (2.43) rezultira drugačijom putanjom eliminacije. Idealno bi bilo ispitati svaku te odabrati onu koja najbolje odgovara cilju koji želimo postići analizom.

Kriterij za ulazne varijable

Dosad navedni kriteriji (2.38), (2.40) i (2.43) ne uklanjaju iz procedure *greedy* komponentu. Podkocka b_* koja optimizira kriterij koristi se za konstrukciju sljedeće kocke $B - b_*$, iako bi isključivanje neke druge kocke moglo rezultirati boljim ishodom u sljedećim koracima cijepanja. Svrha strpljive strategije jest smanjiti taj efekt odabirom većeg broja cijepanja u nadi da će se u kasnijim koracima moći kompenzirati loši (greedy) odabir u prijašnjim koracima. Takvu strategiju možemo okarakterizirati kao pasivnu. Navedeni kriteriji fokusiraju se na podkocke b, a ulazne varijable $\{x_j\}_1^n$ služe samo za definiciju podkocaka koje ulaze u izbor za eliminaciju (2.30)(2.31). Proaktivnija strategija može biti da se fokusiramo izravno na ulazne varijable te pronađemo one koje u danom koraku najviše utječu na $f(\mathbf{x})$ unutar trenutne kocke, $\mathbf{x} \in \beta$. Primjer takvog kriterija je

$$J_j = \max_m \{J(b_{jm})\} - \min_m \{J(b_{jm})\} \tag{2.44}$$

pri čemu je $J(b)$ definiran kao u (2.39), a $\{b_{jm}\}$ su podkocke nastale od J-te varijable. Podkocka $b_{j^*m^*}$ odabrana za cijepanje je tada ona definirana varijablom $j^* = \arg\max_j J_j$ koja maksimizira početni kriterij $m^* = \arg\max_m J(b_{j^*m})$.

Definirajmo dodatnu podkocku povezanu s realnom ulaznom varijablom x_j

$$b_{jc} = \left\{\mathbf{x} \mid x_{j(\alpha)} \leq x_j \leq x_{j(1-\alpha)}\right\}. \tag{2.45}$$

Podkocka b_{jc} služiti će nam kao pomoćna podkocka pri izboru podkocke za cijepanje budući uključivanjem nje u $\{b_{j-}, b_{j+}\}$ algoritam je "osjetljiviji" kad je promatrana ciljna funkcija konveksna. Primjerice, neka je $f(\mathbf{x}) = x_j^2,\ -1 \leq x_j \leq 1$. Tada brisanje (cijepanje) bilo koje od podkocaka $b = b_{j\pm}$ rezultira smanjenjem srednje vrijednost u rezultantnoj kocki $\overline{y}_{B-b} < \overline{y}_B$, te budući nismo uključili b_{jc} u ocjenu varijabla x_j neće biti razmotrena za cijepanje u prilog neke varijable koja nema utjecaja na izlaznu varijablu y ($\overline{y}_{B-b} \simeq \overline{y}_B$). Uočimo da b_{jc} nikada ne brišemo već služi samo za ocjenu.

2.10 Primjer

Ilustrirajmo primjenu PRIM metode na kratkom primjeru iz marketinga. Anketu od 502 pitanja popunilo je N= 9409 ispitanika, posjetitelja trgovačkog centra u San Franciscu. Prvih 14 pitanja demografskog karaktera prikazano je u tablici (2.1). Odgovori na ta pitanja predstavljaju ulazne varijable koje su ili realne ili kategorijske. Postoje neodgovorena pitanja (*missing values*). Preostalih 488 pitanja vezana su za ponašanje i doživljaje kupaca. Njih promatramo kao izlazne varijable y kako bi klasificirali ponašanje kupaca ovisno o njihovim demografskim obilježjima (ulazna

Tablica 2.1: Ulazne varijable za podatke iz marketinga

Var.	Demog. obilježja	Broj vrijed.	Kat.
1	Spol	2	*
2	Bračni status	5	*
3	Dob	6	
4	Obrazovanje	6	
5	Zanimanje	9	*
6	Prihodi	9	
7	Godine boravišta u SF	5	
8	Brak / dvojni prihodi	2	*
9	Broj članova kućanstva	9	
10	Broj maloljetnih članova kućanstva	9	
11	Tip kućanstva	3	*
12	Vrsta doma	5	*
13	Etnička pripadnost	8	*
14	Jezik u kućanstvu	3	*

varijabla $\mathbf{x}$). Ilustrirati ćemo klasifikaciju za tri karakteristična ponašanja (obilježja).

Prvo obilježlje je učestalost leta avionom kojeg mjerimo brojem povratnih letova u godini. Globalno očekivanje nad svim podacima je $\overline{y} = 1.7$. Kocke inducirane PRIM-om su:

y = broj letova u godini; $\quad \overline{y} = 1.7$

$B_1 : \overline{y}_1 = 4.2, \quad \beta_1 = 0.08$

obrazovanje $\geq$ 16 godina
zanimanje $\in$ { menađer, prodavač, građevinar}
prihodi $>$ \$50K , & $\neq$ nema podatka
broj maloljetne djece (< 18) u domaćinstvu ≤ 1

$B_2 : \overline{y}_2 = 3.2, \quad \beta_2 = 0.07$

obrazovanje $>$ 12 godina,& $\neq$ nema podatka
prihodi $>$ \$30, & $\neq$ nema podatka
$18 <$ dob < 54
brak/dvojni prihodi $\in$ { samac, vjenčan - 1 prihod}

Prva kocka predstavlja demografska obilježja 8% populacije koja prosječno leti 4.2 puta godišnje. a druga dio od 7% odvojenih od prve grupe, sa srednjom vrijednosti 1.7.

Drugo obilježje koje je ispitivano je posjedovanje kućnog ljubimca. Učestalost posjedovanja kućnog ljubimca je iznosila 52% što u grubo znači da je omjer 1:1 da nasumično odabrana osoba ima kućnog ljubimca. Primjer induciranih PRIM kocaka:

$y = 1$(imaju kućnog ljubimca); $\quad \overline{y} = 0.52$

$B_1 : \overline{y}_1 = 0.80, \quad \beta_1 = 0.17$

dob ≤ 44
obrazovanje ≤ 14 godina
život u području Bay Area ≥ 4 godine
dom $\in$ { kuća, stan}
etnicitet $\in$ { Indijanac, bijelac, nema pod}

$B_2 : \overline{y}_2 = 0.76, \quad \beta_2 = 0.08$

broj maloljetne djece (< 18) u domaćinstvu > 0
tip kućanstva $\in$ {vlastito, kod roditelja, nema pod}
etnicitet $\in$ { Indijanac, bijelac, nema pod}

PRIM metodom smo identificirali dio od 25% testirane populacije koja u omjeru 5:1 ima kućnog ljubimca. Ono što je možda bilo neintuitivno jest da etnička pripadnost također utječe na ishod ovog obilježja.

Treće obilježje su navike slušanja radija.

$$y = \text{slušaju radio} = \begin{cases} 0.0 & \text{nikada} \\ 0.25 & \text{povremeno} \\ 1.0 & \text{redovno} \end{cases}$$

Srednja vrijednost nad svim podacima je $\overline{y} = 0.10$. Primjer PRIM induciranih kocaka

$B_1 : \overline{y}_1 = 0.23, \quad \beta_1 = 0.14$

dob ≥ 35
život u području Bay Area ≥ 10 godina, & $\neq$ nema podatka
broj maloljetne djece (< 18) u domaćinstvu ≤ 0
tip kućanstva $\neq$ podstanarstvo
etnicitet$\notin$ { afroamerikanac, Pacifičko otočje}
vrsta doma $\in$ { kuća, apartman}

$B_2 : \overline{y}_2 = 0.20, \quad \beta_2 = 0.09$

spol = muški
bračni status $\in$ { vjenčan, razveden, nema pod}
zanimanje $\in$ { menađer, prodavač, umirovljenik}

Poglavlje 3

Usporedba metoda

3.1 Algoritmi pokrivanja

Neka izlazna varijabla y poprima samo dvije vrijednosti (npr. $y = \{0, 1\}$). PRIM metodu za takvu izlaznu varijablu možemo shvatiti kao induktivnu metodu u kojoj te dvije vrijednosti predstavljaju negativne i pozitivne dijelove promatranog objekta u strojnom učenju. U ovakvom kontekstu PRIM je jako sličan ostalim algoritmima strojnog učenja koji proučavaju disjunktne skupove pravila kroz nizovno pokrivanje podataka (npr. CN2 (Clark i Niblett, 1989), FOIL (Quinlan, 1990), RIPPER (Cohen, 1995)). Glavna razlika je način na koji ti algoritmi, odnosno PRIM, obrađuju podatke u svrhu izrade pravila tj. odabira kocaka. Ostali algoritmi koriste veoma pohlepne strategije, osobito na kategoričkim varijablama. Kao što smo već spomenuli, takve strategije ograničavaju rezultate te je to glavna motivacija pri izradi PRIM algoritma. Još jedna, manje važna razlika je reprezentacija induciranih pravila (kocaka). Svaka kocka u PRIM metodi (2.21) uključuje implicitne disjunkcije

$$x_j \in s_{jk} = \bigcup_{z_l \in s_{jk}} (x_j = z_l).$$

Većina drugih metoda proizvodi uglavnom konjuktivna pravila. Ostale razlike uključuju način rješavanja problema nedostatka nekih vrijednosti za određene varijable i korištenje višestrukih trajektorija.

3.2 Indukcija stabla odlučivanja

U primjeni, najkonkurentnije metode PRIM metodi su metode indukcije stabla odlučivanja (eng. *decision tree*) poput CART (Breiman, 1984) ili C4.5 (Quinlan, 1994) metoda.

Ove metode kreiraju skup razdvojenih pravila koja zajedno potpuno pokrivaju čitav ulazni prostor kroz rekurzivno particioniranje. Svaka uzastopna particija (*split*) inducirana je uvjetom $x_j \in s_{jk}$ tako da u konačnom pokrivaču svako pravilo ima konjuktivnu formu danu u (2.21). Iako je cilj ovih metoda točnost aproksimacija na ulaznim varijablama, a ne eksplicitna optimizacija, možemo ju postići provjerom pravila povezanih sa najvećim predviđenim izlaznim vrijednostima što se može interpretirati na analogan način kao u PRIM metodi.

Glavna razlika PRIM metode i induktivnih metoda stabla odlučivanja (*decision tree*) je upotreba pokrivanja, a ne particioniranja u kreiranju pravila (kocaka) te strpljivost naspram pohlepne strategije pri odabiru pravila. Pravila nastala pokrivanjem su "izražajnija" od onih nastalih particioniranjem budući su inducirana nezavisno jedno o drugom. Pravila nastala particioniranjem dijele zajedničke konjuktivne izjave $x_j \in s_{jk}$. Pokrivanje ishoduje manje jednostavnijih pravila te time olakšava interpretaciju. Nije dokazano da li pohlepne strategije sa većom izražajnosti rezultiraju većom točnosti - vrlo vjerojatno ovisi o pojedinačnom slučaju.

3.3 PRIM vs CART

Usporedimo PRIM s metodom indukcije stabla odlučivanja koja je najmanje pohlepna - CART (eng. *Classification and Regression Tree Method*). U prosjeku svako konjuktivno ograničenje $x_j \in sjk$ koje definira skup pravila (2.21 uzima jednu polovinu podataka u svakom koraku iteracije, neovisno o tome jesu li varijable realne ili kategorijske. Ovaj postupak možemo okarakterizirati kao polu-pohlepan u odnosu na ostale slične algoritme koji rade pohlepniju (veću) fragmentaciju podataka. Također, CART kreira pravila identična formom onima u PRIM metodi (2.21). U ovom poglavlju usporedit ćemo performanse PRIM i CART metoda iz perspektive maksimizacije funkcija na primjeru nasumičnih podataka iz uniformne distribucije te na marketing primjeru iz poglavlja 2.11.

Za razliku od PRIM metode, CART ne dozvoljava uvid korisnicima u odabir parametra nosača i srednje vrijednosti koje određena podkocka (pravilo) koristi. Usporedba je provedena na sljedeći način. Prvo je nad podacima provedena CART procedura te određeno J pravila s najvećim predviđenim izlaznim vrijednostima. Zatim je provedena PRIM metoda nad istim podacima na način da se inducira J uzastopnih kocaka metodom pokrivanja. Odabir između vrijednosti parametra nosača i srednje vrijednosti napravljen je uzevši u obzir odgovarajuće vrijednosti u odgovarajućem CART pravilu. Na taj način možemo usporediti relativnu jačinu odgovarajućih pravila. Ako su nosači jednaki onda je ona podkocka s većom srednjom vrijednosti bolja, a ako su srednje vrijednosti slične bolja je podkocka s većim nosačem. U oba slučaja

Tablica 3.1: Usporedba rezultata

CART			PRIM		
k	$\bar{y}_k$	$\beta_k(\%)$	k	$\bar{y}_k$	$\beta_k(\%)$
1	0.29	1.4	1	0.28	4.2
2	0.24	1.3	2	0.31	3.3
3	0.19	1.4	3	0.32	3.5
4	0.14	1.9	4	0.34	3.1
		C = 3.9			

podaci su podijeljeni u isti *testni* dio i *learning* dio. Parametar cijepanja (*peeling fraction*) α (2.30) smo uzeli da bude 10% ($\alpha = 0.1$).

Primjer 3.3.1. *Podaci sadrže $N = 1000$ opažanja za $n = 10$ ulaznih varijabli nasumično generiranih iz uniformne razdiobe $\{x_j \sim \mathrm{U}[-1, 1]\}_1^{10}$. Funkcija koju optimiziramo je*

$$y = f(\mathbf{x}) = \prod_{j=1}^{J} x_j. \tag{3.1}$$

Uzmimo $J = 3$. Usporedba rezultata PRIM i CART metode nalazi se u tablici 3.1 U lijevom dijelu tablice prikazani su srednja vrijednost $\bar{y}_k$ i nosač $\beta_k(x100)$ za četiri najoptimalnija CART pravila, a u desnom odgovarajuće vrijednosti iz PRIM metode. Veličina pokrivenost (coverage ratio) je mjera usporedbe performansi pokrivanja PRIM naspram CART definirana na način

$$C = \sum_{k=1}^{4} (\bar{y}_k - \bar{y})\beta_k \tag{3.2}$$

pri čemu je $\bar{y}$ globalna srednja vrijednost (2.8). Uočimo da je na ovom primjeru pokrivenost PRIM metode četiri puta veća nego u CART metodi. Dakle, u ovom primjeru, PRIM nadjačava CART zbog svoje strpljive strategije. U svakom koraku cijepanjem se oduzima samo 10% podataka te se konačna struktura ciljne funkcije generira nad većim brojem podataka, te posljedično, povećanom srednjom vrijednosti.

Primjer 3.3.2. *U ovom primjeru usporediti ćemo metode nad podacima iz primjera iz poglavlja 2.11 Uočimo da su rezultati PRIM metode ili slični ili bolji od onih iz CART metode. Pokrivenost je 45% veća. U svim potprimjerima (vidi tablice 3.2, 3.3, 3.4) CART proizvodi velika stabla (50-200 čvorova). Većina pravila iz CART metode su podosta kompleksnija od onih iz PRIM metode. Primijetimo također da vrijednosti nosača i srednje vrijednosti za PRIM metodu nisu optimalni već su uzeti*

Tablica 3.2: Usporedba rezultata - učestalost letenja

	CART			PRIM	
k	$\bar{y}_k$	$\beta_k(\%)$	k	$\bar{y}_k$	$\beta_k(\%)$
1	5.0	1.9	1	5.1	2.3
2	3.6	2.5	2	3.6	5.1
3	2.5	2.7	3	3.0	5.6

C = 1.87

Tablica 3.3: Usporedba rezultata - kućni ljubimac

	CART			PRIM	
k	$\bar{y}_k$	$\beta_k(\%)$	k	$\bar{y}_k$	$\beta_k(\%)$
1	0.81	4.5	1	0.84	7.8
2	0.75	16.0	2	0.72	22.7
3	0.71	2.4	3	0.69	2.3

C = 1.38

Tablica 3.4: Usporedba rezultata - radio

	CART			PRIM	
k	$\bar{y}_k$	$\beta_k(\%)$	k	$\bar{y}_k$	$\beta_k(\%)$
1	0.25	6.1	1	0.23	14.0
2	0.17	7.9	2	0.17	11.0
3	0.16	5.8	3	0.13	10.0

C = 1.44

da približno odgovaraju onima iz CART metode što znači da je PRIM mogao imati i mnogo bolje performanse od prestavljenih.

Poglavlje 4

Primjena PRIM metode - primjer u medicini

Ishemična bolest srca (IHD) vodeći je uzrok smrtnosti i oboljenja u zapadnom društvu. Razvoj IHD-a izravna je posljedica interakcije između genetskih čimbenika podložnosti oboljenju i čimbenika iz okoliša. Različite kombinacije genetskih i ekoloških čimbenika utječu na kompleksnu etiologiju IHD-a kod različitih podgrupa pojedinaca. U ovom primjeru opisan je postupan provedbe PRIM metode za klasifikaciju rizičnih faktora i odgovarajućih vrijednosti u 16 međusobno isključivih particija s različitim nivoima rizika. Cilj je bio pronaći podgrupe (particije) s visokim rizikom oboljenja i maksimizirati broj slučajeva obuhvaćenih u particijama. Uzastopna PRIM analiza provedena je nad podacima o učestalosti IHD-a prikupljenih kroz 8 godina za 5.455 nezavisnih pojedinaca u sklopu srčane studije u Copenhagenu (CCHS - *Copenhagen City Heart Study*). Nezavisan uzorak od 362 pojedinca zatim je uzet kao test dobivenog modela za svaku particiju. Odabran je opsežan skup faktora definiran iz prethodnih istraživanja etiologije IHD-a. Ova primjena metode provedena je u sklopu članka *An Application of the Patient Rule-Induction Method for Evaluating the Contribution of the Apoliprotein E and Lipoprotein Lipase Genes to Predicting Ischemic Heart Disease*. Svi podaci, tablice i slike uzeti su iz navedenog članka.

Sudionici

Sudionici su regrutirani iz generalne populacije u Copenhagenu kroz tri perioda: 1976-1978, 1981-1983 i 1991-1994. Tijekom svakog perioda, pojedinci iz prijašnjih perioda su ponovno evaluirani. PRIM model proveden je nad 5.455 sudionika iz prvog perioda starijih od 45 godina koji nisu imali IHD u trećem periodu te su praćeni do 1999. Uzorak od 362 sudionika iz drugog perioda također starijih od 45 godina bez oboljenja u trećem iskorišten je kao evaluacija dobivenog modela.

Tablica 4.1: Karakteristike sudionika

Faktor rizika	Oboljeli ($n = 519$)	Neoboljeli($n = 4,936$)
Tradicionalni rizični faktori		
Godine u 3. periodu (god, ±SD)	70.2 (8.8)	65.1 (9.2)
Spol		
Žene	233 (0.45)	2,964 (0.60)
Muškarci	286 (0.55)	1,972 (0.40)
Pušač		
Ne	187 (0.36)	2,157 (0.44)
Da	332 (0.64)	2,779 (0.56)
Dijabetes		
Ne	466 (0.90)	4,704 (0.95)
Da	53 (0.10)	232 (0.05)
Hipertenzija		
Ne	77 (0.15)	1,448 (0.29)
Da	442 (0.85)	3.488 (0.71)
Lipidi i BMI		
Kolesterol		
≤ 200	71 (0.14)	651 (0.13)
(200, 240)	164 (0.32)	1,665 (0.34)
> 240	284 (0.54)	2,620 (0.53)
HDL-C		
< 40	90 (0.17)	536 (0.11)
≥ 40	429 (0.83)	4,400 (0.89)
Trigliceridi		
< 150	231 (0.45)	2,706 (0.55)
≥ 150	288 (0.55)	2,230 (0.45)
BMI		
≤ 25	188 (0.36)	2,179 (0.44)
(25, 30)	225 (0.44)	1,957 (0.40)
> 30	106 (0.20)	800 (0.16)

Varijable

Ulazne varijable (rizični faktori) definirane su kao u tablicama 4.1 i 4.2 gdje je odmah navedena i podjela na oboljele i neoboljele.

Tablica 4.2: Karakteristike sudionika

Faktor rizika	Oboljeli ($n = 519$)	Neoboljeli($n = 4,936$)
Genetski rizični faktori		
APOE -491A>T (E560)		
AA	364 (0.70)	3,526 (0.71)
AT	144 (0.28)	1,290 (0.26)
TT	11 (0.02)	120 (0.03)
APOE -427T>C (E624)		
TT	427 (0.82)	3,946 (0.80)
TC	87 (0.17)	931 (0.19)
CC	5 (0.01)	59 (0.01)
APOE -219G>T (E832)		
GG	135 (0.26)	1,404 (0.28)
GT	274 (0.53)	2,451(0.50)
TT	110 (0.21)	1,081 (0.22)
APOE g.2059T>C (E3937)		
TT	364 (0.70)	3,429(0.69)
TC	141 (0.27)	1,375 (0.28)
CC	14 (0.03)	132 (0.03)
APOE g.2197C>T (E4075)		
CC	446 (0.86)	4,151 (0.84)
CT	68 (0.13)	762 (0.15)
TT	5 (0.01)	23 (0.01)
LPL g.8756G>A (LPL9)		
GG	504 (0.97)	4,803 (0.97)
GA	15 (0.03)	133 (0.03)
LPL g.16577A>G (LPL291)		
AA	489(0.95)	4,689(0.95)
AG	30 (0.05)	245 (0.05)
GG	0 (0.00)	2 (0.00)
LPL g.22772C>G (LPL447)		
CC	424(0.82)	4,016(0.81)
CG	90 (0.17)	868 (0.18)
GG	5 (0.01)	52 (0.01)

4.1 Analiza i rezultati

PRIM

Cilj PRIM analize je pronaći podskupove varijabli, i njihove vrijednosti, koje su optimalni prediktori vjerojatnosti oboljenja u podskupu pojedinaca, producirajući međusobno isključive particije uzorka koristeći korake eliminacije (cijepanja) i lijepljenja. Koraci kreiranja četiri particije prikazani su na slici 4.1.

Proces eliminacije/lijepljenja staje u jednom od dva slučaja: ako je minimalni nosač (β) postignut ili ako nijedna varijabla ne dostiže prag definiran paramentrom kompleksnosti (λ). Nosač particije definiran je kao broj pojedinaca u toj particiji podijeljen s brojem nedodijeljenih pojedinaca koji su potencijalno mogli biti u toj particiji.

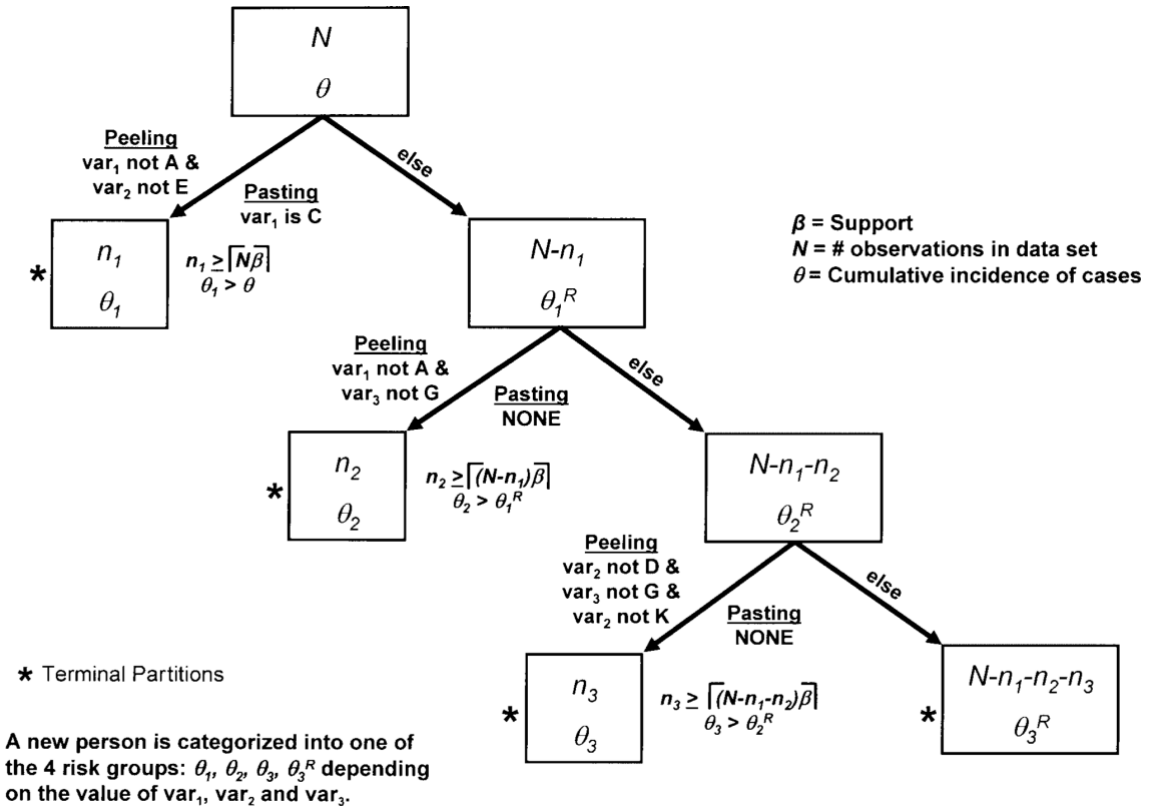
Minimalni nosač određuje broj nedodijeljenih pojedinaca potrebnih za postojanje particije. Patametar kompleksnosti je minimalno povećanje u θ potrebno za profinjenje particije dodavanjem prediktora i odgovarajuće vrijednosti u definiciju particije. Dakle, da bi korak eliminacije/lijepljenja bio valjan potrebno je da postoji dovoljan broj pojedinaca ($= \lceil \beta \times$ broj nedodijeljenih pojedinaca$\rceil$) te povećanje od θ za barem λ.

U ilustraciji PRIM koraka prikazanih na slici 4.1, prva particija nastala je iz dva koraka eliminacije definirana s 'var_1 not A' i 'var_1 not E'. Varijabla 'var_1 not A' je ona među svim mogućim varijablama i odgovarajućim vrijednostima u skupu podataka s najvećom učestalosti takva da postoji dovoljan broj pojedinaca da se zadovolji β i da je θ uvećan za barem λ. U sljedećem koraku eliminacije prediktor 'var_1 not E' dodan je u particiju. Nijedna druga varijabla nije zadovoljavala kriterije nosača i kompleksnosti pa postupak eliminacije staje te se kreće s vraćanjem varijabli u svrhu poboljšanja particije. Pojedinci izbačeni tokom eliminacije koji su imali vrijednost C za 'var_1' vraćeni su u particiju te time završava prvi korak PRIM analize budući više nijedno moguće lijepljenje nije zadovoljavalo kriterije.

Nakon što je kreirana prva particija, ostatak pojedinaca ($N - n_1$) koji nisu u toj particiji, razmatraju se za konstrukciju iduće particije. Postupak se dalje nastavlja dok svi pojedinci nisu dio neke particije. Zatim se provodi testiranje nad particijama da se utvrdi statistička značajnost svake particije. Oni pojedinci koji nisu dio nijedne značajne particije tvore svoju zajedničku particiju.

Stepwise PRIM analiza

Stepwise PRIM analiza iskorištena je da bi izračunali korist od dodavanje varijabli (faktora rizika) u model. U prvom koraku analize tradicionalni faktori rizika iskorišteni su za izgradnju PRIM modela. U drugom koraku, svaka dobivena particija



Slika 4.1: Ilustracija PRIM koraka eliminacije (cijepanja) i lijepljenja

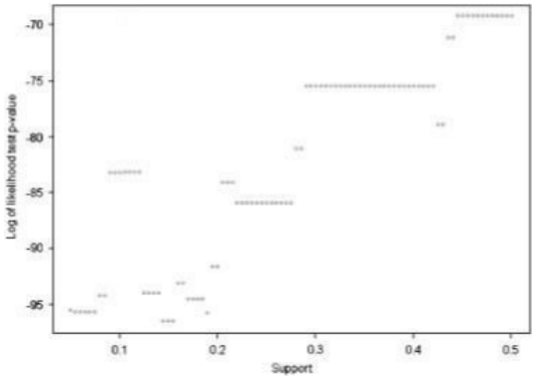
dodatno je podijeljena dodavanjem lipida i BMI faktora u analizu. Zatim, u trećem koraku, su dodani genetski faktori kako bi se ustvrdile značajne particije za svaku od particija dobivenih u drugom koraku. (Vidi sliku)

Rezultati

Statistika za 5,455 pojedinaca sakupljena tokom trećeg perioda pa do kraja 1999. godine dana je u tablicama 4.1 i 4.2. Tokom razdoblja od trećeg perioda do 1999. godine 519 pojedinaca oboljelo je od IHD. Kako bi ustanovili koji prediktori su povezani s IHD-om χ_2 test ili t-test je proveden za svaku od kategorijskih i neprekidnih prediktorskih varijabli. Sudionici koji su oboljeli su svi mahom bili stariji te uglavnom muškog spola, s navikama pušenja, dijabetesom i hipertenzijom. Nije pronađena statistička povezanost između IHD-a i ijednog od 8 genetskih faktora.

Za svaki od PRIM modela parametar kompleksnosti postavljen je na nulu, a

nosač je odabran iz rezultata analize prikazanih na slici 4.2 dobivenih testiranjem različitih vrijednosti nosača nad PRIM particijama i ustvrđujući onog od najveće značajnosti. Slika 4.3 prikazaje dijagram stabla particija dobivenog stepwise PRIM analizom. Detalji o varijablama i odgovarajućim vrijednostima za svaku od particija prikazani su u tablici 4.3.



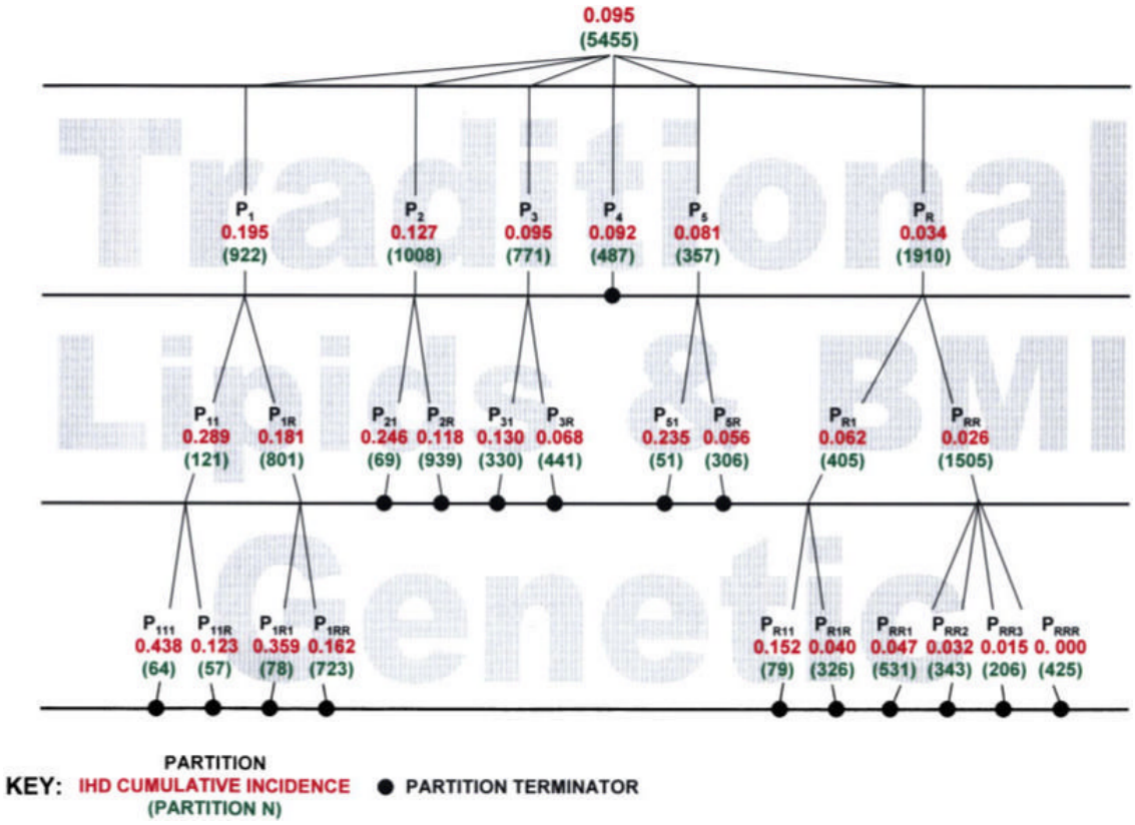
Slika 4.2: Graf korišten pri odabiru parametra nosača za prvi korak PRIM modela. Nosač koji ostvaruje najmanju p-vrijednost odabran je kao parametar modela.

1. KORAK PRIM ANALIZE Tradicionalni faktori rizika iskorišteni su za izgradnju PRIM modela nad cijelim uzorkom od $5,455$ pojedinaca. Koristeći logističku regresiju parametar nosača procijenjen je na 0.145 (vidi 4.2). U drugom stupcu tablice 4.3 nabrojane su vrijednosti varijabli koje definiraju svaku od pet značajnih particija. Kumulativna učestalost oboljenja od IHD-a za navedene particije je u rangu između 0.034 i 0.195 (vidi 4.3).

2. KORAK PRIM ANALIZE Analiza je dalje provedena nad svakom od dobivenih particija (P_1, P_2, P_3, P_4, P_5 i P_R) koristeći faktore za lipide i BMI. Pet od šest PRIM modela (P_1, P_2, P_3, P_5 i P_R) rezultiralo je daljnjim particioniranjem. U četvrtom stupcu tablice 4.3 nabrojane su vrijednosti varijabli koje definiraju svaku od novih značajnih particija. Odabrane vrijednosti parametra nosača su redom 0.130, 0.055, 0.350, 0.085 i 0.145. Kumulativna učestalost oboljenja od IHD-a za navedene particije je u rangu između 0.026 i 0.289.

3. KORAK PRIM ANALIZE Profinjenje dobivenih 11 particija (P_{11}, P_{1R}, P_{21}, P_{2R}, P_{31}, P_{3R}, P_4, P_{51}, P_{5R}, P_{R1} i P_{RR}) napravili smo koristeći genetske faktore. Samo

četiri su rezultirale novim profinjenim particijama. U zadnjem stupcu tablice 4.3 nabrojane su vrijednosti varijabli koje definiraju svaku od novih značajnih particija. Odabrane vrijednosti parametra nosača su redom 0.455, 0.085, 0.180 i 0.325. Primijetimo da se svaki od genetskih faktora pojavio barem jednom u novim particijama. Dodavanje genetskih faktora identificiralo je grupu od 425 pojedinaca koji nemaju nikakvih obilježja IHD-a. Najveća kumulativna učestalost opažena nakon trećeg koraka analize (0.438) dobivena je iz daljnjeg particioniranja grupe P_{11}, djelomično određena s E3937 i E4075 koji definiraju poznate $\epsilon 2$, $\epsilon 3$ i $\epsilon 4$ alele *APOE* gena.



Slika 4.3: Dijagram particija dobivenih stepwise PRIM analizom. Crni tekst oznaka je particije, crvena brojka je kumulativna učestalost oboljenja unutar particije, a zelena brojka je količina uzorka unutar particije.

Tablica 4.3: Particije dobivene stepwise PRIM analizom

Particija	Tradicionalni faktori rizika	Particija	Lipidi i BMI	Particija	Genetski faktori rizika
P_1	>65, muškarac, hipertenzija	P_{11}	HDL <40, TRIG>150, CHOL >200	P_{111}	(E3937≠TC, E4075≠CT, E624≠TC, LPL291≠AG) ili E560=TT
				P_{11R}	ne P_{111}
		P_{1R}	ne P_{11}	P_{1R1}	(E560≠AT, E4075≠CT, E624=TC, LPL9≠GA) ili E4075=TT
				P_{1RR}	ne P_{1R1}
P_2	(>65, žena, pušač) ili dijabetes	P_{21}	HDL <40, CHOL $\neq(200,240)$		
		P_{2R}	ne P_{21}		
P_3	>65, hipertenzija, žena, nepušač, nema dijabetes	P_{31}	(TRIG>150, CHOL $\neq(200,240)$) ili CHOL <200		
		P_{3R}	ne P_{31}		
P_4	<65, muškarac, pušač, hipertenzija, nema dijabetes				
P_5	muškarac, ne hipertenzija, pušač, nema dijabetes	P_{51}	(HDL <40, CHOL <240) ili BMI >30		
		P_{5R}	ne P_{51}		
P_R	ne (P_1, P_2, P_3, P_4, P_5)	P_{R1}	BMI >25, TRIG>150, HDL >40 CHOL >200	P_{R11}	(E3937=TC, E624≠TC, LPL291≠AG, LPL447≠CG) ili E624=CC
				P_{R1R}	ne P_{R11}
		P_{RR}	ne P_{R1}	P_{RR1}	(E3937=TT, E560≠AT, E624=TT, LPL291≠AG) ili LPL9=GA
				P_{RR2}	E3937≠TT, E560≠TT, E624≠CC, E832≠GG, LPL447=CC
				P_{RR3}	E3937≠CC, E4075≠CT, E624=TT, E832≠GG, LPL291≠AG, LPL447≠GG
				P_{RRR}	ne (P_{RR1}, P_{RR2}, P_{RR3})

Tablica 4.4: Procjena IHD rizika za sudionike regrutirane u drugom periodu ispitivanja

Particije	Procjena rizika	Veličina particije	Procjenjeni broj oboljelih	Realizacija - broj oboljelih
P_{111}	0.438	0	0	0
P_{11R}	0.123	3	0	0
P_{1R1}	0.359	2	1	0
P_{1RR}	0.163	44	7	7
P_{21}	0.246	8	2	4
P_{2R}	0.118	51	6	9
P_{31}	0.130	4	1	1
P_{3R}	0.068	16	1	0
P_4	0.092	45	4	6
P_{51}	0.235	9	2	0
P_{5R}	0.056	38	2	2
P_{R11}	0.152	5	1	0
P_{R1R}	0.040	33	1	1
P_{RR1}	0.047	46	2	3
P_{RR2}	0.032	17	1	1
P_{RR3}	0.015	9	0	0
P_{RR4}	0.000	32	0	2

Validacija modela

Validacija modela provedena je nad 362 pojedinca uvedena u ispitivanje tokom drugog regrutacijskog perioda. Karakteristike svakog pojedinca provedene su kroz dobivene particije te je on/a na taj način svrstan/a u neku od dobivenih particija. Tablica 4.4 predstavlja predviđanja za tu grupu. Iako je veličina ove grupe mala da bi sa značajnom sigurnošću ustvrdili valjanost modela, zanimljivo je da je dobiveno predviđanje jako blizu stvarnoj realizaciji. Iznimke se događaju samo na nedodijeljenim participama, primjerice P_{RRR} koja ima predviđanje od 0.000, a u stvarnosti se u toj grupi nalazi dvoje oboljelih.

Tablica 4.5: ANOVA za značajne varijable i interakcije koristeći logističku regresiju za predviđanje IHD-a

Kovarijanta	*df*	statistika χ_2	*p*-vrijednost
BMI	2	13.4	0.001
Spol	1	39.9	0.000
Dijabetes	1	14.7	0.000
Ovisnost o pušenju	1	9.3	0.002
Hipertenzija	1	42.3	0.000
Kolesterol	2	2.5	0.299
Trigliceridi	1	5.1	0.024
Dob	1	95.7	0.000
BMI - ovisnost o pušenju	2	8.1	0.017
Spol - kolesterol	2	6.9	0.033
Dijabetes - dob	1	6.9	0.008
Ovisnost o pušenju - hipertenzija	1	5.4	0.020

Tablica 4.6: Test značajnosti za genetske faktore

Genetski faktor	*df*	statistika χ_2	*p*-vrijednost
LPL9	1	0.264	0.607
LPL291	2	0.497	0.780
LPL447	2	0.011	0.995
E560	2	0.412	0.814
E624	2	2.075	0.354
E832	2	3.202	0.202
E3937	2	0.172	0.918
E4075	2	4.684	0.096
$\epsilon 2$, $\epsilon 3$, $\epsilon 4$	5	5.247	0.386

Sažetak

Uspješna istraživanja i razvoj novih lijekova važna su za budućnost farmaceutske industrije, te za ljudsko zdravlje. Jedan poseban aspekt razvoja lijekova koji farmaceutske tvrtke moraju uzeti u obzir je sposobnost prepoznavanja podskupine bolesnika, koji će vjerojatno izvući dodatnu korist od liječenja kako bi se preciznije odredio tretman tih pojedinaca unutar zdravstvenog sustava.

Jedna od metoda koje se mogu koristiti je PRIM (eng. *Patient Rule Induction Method*) metoda kao što je demonstrirano na primjeru u poglavlju 4.

PRIM je namijenjen kao alat analitičara podataka (statističara), koji će se koristiti kada je cilj bilo eksplicitna ili implicitna optimizacija (poglavlje 2.2) funkcije cilja. Njegovo glavno obilježje je strpljiva metoda eliminacije varijabli, te ponovno lijepljenje natrag onih koje su naknadno ocjenjene kao značajne u kombinaciji s višestrukim putanjama razvoja (poglavlje 2.5) kako bi se poboljšala snaga i stabilnost modela, te sudjelovanje (uvid) korisnika u proces odabira modela (varijabli i parametara). Pokazali smo na primjerima da PRIM ima učinak superioran u odnosu na usporedive metode, kao što je CART (eng. *Classification and Regression Tree Method*), koje su namijenjene, i popularnije korištene, za aproksimaciju funkcija. Mjera u kojoj ovi primjeri i performansa PRIM metode generaliziraju druge situacije tek treba biti uspostavljena.

U posljednjem poglavlju prikazan je primjer primjene PRIM metode za analizu rizika od oboljenja od IHD-a.

Summary

The successful research and development of new drugs is critical for the future of the pharmaceutical industry and for human health. One particular aspect of drug development that pharmaceutical companies are required to consider is the ability to identify subgroups of patients that are likely to derive additional benefit from treatment and to help quantify the burden of those patients on the healthcare system.

One of the methods that can be used is PRIM method (*Patient Rule Induction Method*) as shown on the example in section 4.

PRIM is intended as an addition to the data analyst's tool kit, to be used when the goal is either explicit or implicit optimization (section 2.2). Its distinguishing characteristics include patient peeling/pasting coupled with multiple trajectories (section 2.5) to enhance power and stability, and intimate user involvement in the model selection process. It tends to produce parsimonious interpretable descriptions of the structure it uncovers, and on the problems considered here exhibits performance superior to comparable procedures such as CART (*Classification and Regression Tree Method*) that are intended for function approximation. The extent to which this performance gain generalizes other situations remains to be established. As with all learning procedures, relative performance will likely be problem dependent.

The final chapter gives an example of application PRIM method for analyzing the risk of IHD.

Bibliografija

[1] Dyson, G.; Frikke-Schmidt, R.; Sing, C.F. *An Application of the Patient Rule-Induction Method for Evaluating the Contribution of the Apolipoprotein E and Lipoprotein Lipase Genes to Predicting IHD*. 2007. Genet Epidemiol.

[2] Durrett, R. *Probability. Theory and Examples* 2010. Cambridge University Press.

[3] Friedman, J.H. *Data Mining and Knowledge Discovery*. 1997. [55-77]

[4] Friedman, J.H.; Fisher, N.I. *Bump hunting in high-dimensional data*. 1998. Stat Comput.

[5] Huzak, M. *Predavanja iz statistike*. PMF Zagreb.

[6] Polonik, W.; Wang, Z. *Prim Analysis*. 2007. NFS.

[7] Sarapa, N. *Teorija vjerojatnosti*. 2002. Školska knjiga.

[8] Vapnik, V. *The Nature of Statistical Learning Theory*. (1995). Springer.

[9] Wu, L.; Chipman, H. *Bayesian model-assisted PRIM algorithm*. 2003. Technical Report.

[10] Wagner, V. *Materijali iz statistike*. PMF Zagreb.

Životopis

Marija Radnić rođena je 03. veljače 1991. godine u Splitu, te od tad do polaska na fakultet živi u Kaštel Kambelovcu gdje pohađa Osnovnu školu Kneza Trpimira. 2009. godine završava Prirodoslovno-matematičku gimnaziju u Splitu i seli se u Zagreb gdje pri Matematičkom odsjeku Prirodoslovno-matematičkog fakulteta upisuje preddiplomski studij inženjerske matematike. Po završetku preddiplomskog studija Matematike, 2013. godine, upisuje Diplomski studij Matematičke statistike na PMF-u u Zagrebu.

Tokom svog srednjoškolskog i fakultetskog školovanja redovito održava instrukcije iz matematike, fizike, kemije i pripreme za državnu maturu iz matematike i fizike, te po završetku preddiplomskog studija i instrukcije iz matematičkih kolegija na tehničkim fakultetima. Također, tokom fakulteta ljeti radi u turističkoj agenciji na otoku Braču. Tokom zadnje godine studiranja zapošljava se kao student u konzultantskom društvu Texo Management d.o.o. gdje pomaže pri administrativnim poslovima te sudjeluje u izradi financijskih modela za poslovanje.

U studenom 2015. godine zapošljava se kao pripravnik u Službi za podršku BI i CRM sustavima u tvrtki mStart d.o.o. (poslovno-tehnološka tvrtka orijentirana na razvoj, implementaciju, integraciju i podršku poslovnih informacijskih sustava i rješenja).