Diskriminacijska analiza

Nikolić, Petra

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Petra Nikolić

DISKRIMINACIJSKA ANALIZA

Diplomski rad

Voditelj rada:
doc. dr. sc. Snježana Lubura Strunjak

Zagreb, 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. _____

Hvala mojem Hrvoju na neizmjernoj podršci, ljubavi i motivaciji kad god je to bilo potrebno. Hvala mojim roditeljima što su vjerovali u mene. Hvala dragim prijateljima s PMF-a što ste mi uljepšali i olakšali studentske dane. Hvala mentorici doc. dr. sc. Snježani Luburi Strunjak na savjetima i pomoći prilikom pisanja ovog rada.

Sadržaj

Uvod

Definicije su preuzete iz [5] i [4], a teoremi iz [3].

Definicija 0.0.1. *Kažemo da je k-dimenzionalna slučajna veličina X apsolutno neprekidna ili, kraće, neprekidna ako postoji nenegativna Borelova funkcija f_X definirana na $\mathbb{R}^k$ takva da se funkcija razdiobe F_X može prikazati na sljedeći način:*

$$F_X(x) = \int_{\langle -\infty, x]} f_X(y) d\lambda(y), x \in \mathbb{R}^k \tag{1}$$

*Funkciju f_X zovemo **funkcija gustoće** razdiobe od X ili, kraće, gustoća od X.*

Definicija 0.0.2. *Neka je X slučajna varijabla definirana na vjerojatnosnom prostoru $(\Omega, \mathcal{F}, \mathbb{P})$. Definiramo $X^+ := \max\{0, X\}$ i $X^- := \max\{0, -X\}$. Budući da su X^+ i X^- nenegativne izmjerive funkcije, Lebesgue-Stieltjesov integrali od X^+ i X^- (u odnosu na mjeru $\mathbb{P}$), $\mathbb{E}X^+ := \int_\Omega X^+ d\mathbb{P}$ i $\mathbb{E}X^- := \int_\Omega X^+ d\mathbb{P}$, postoje i vrijednosti su im u skupu $\mathbb{R} \cup \{+\infty\}$. Kažemo da slučajna varijabla X ima matematičko očekivanje ako je barem jedan od integrala $\mathbb{E}X^+$ i $\mathbb{E}X^-$ konačan. U tom slučaju je matematičko očekivanje*

$$\mathbb{E}X := \mathbb{E}X^+ - \mathbb{E}X^-. \tag{2}$$

Definicija 0.0.3. *Neka je $X = (X_1, X_2, \ldots, X_k)$ slučajni vektor. Kažemo da X ima matematičko očekivanje ako svaka komponenta $X_1, X_2, \ldots, X_k$ tog vektora ima matematičko očekivanje, te u tom slučaju definiramo matematičko očekivanje od X kao vektor*

$$\mathbb{E}X := (\mathbb{E}X_1, \mathbb{E}X_2, \ldots, \mathbb{E}X_k) \tag{3}$$

Definicija 0.0.4. *Neka su X i Y slučajne varijable takve da $\mathbb{E}\left(X^2\right) < \infty$ i $\mathbb{E}\left(Y^2\right) < \infty$. Kovarijanca od X i Y definira se kao*

$$\mathrm{Cov}(X, Y) := \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y). \tag{4}$$

Definicija 0.0.5. *Neka je $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ slučajan uzorak duljine n ($n \geq 1$) iz statističkog modela parametriziranog parametrom $\theta \in \Theta$. Pretpostavimo da je $\tau : \Theta \to \mathbb{R}^k$*

izmjeriva funkcija ($k \geq 1$) *te da želimo procijeniti* $\tau(\theta)$*. Procjenitelj* $T = t(\mathbf{X})$ *za* $\tau(\theta)$ *je nepristran ako vrijedi*

$$(\forall \theta \in \Theta) \quad \mathbb{E}_\theta[T] = \tau(\theta) \tag{5}$$

Definicija 0.0.6. *Konfuzijska matrica (engl. "confusion matrix") za klasificiranje u dvije populacije* π_1 *i* π_2 *je tablica koja se koristi za evaluaciju performansi klasifikacijskog modela. Konfuzijska matrica govori koliko je podataka model točno ili pogrešno klasificirao. Konfuzijska matrica se sastoji od četiri različita elementa:*
True Positive (TP): broj podataka koje je model točno klasificirao kao populaciju π_1
False Positive (FP): broj podataka koje je model pogrešno klasificirao kao populaciju π_1
True Negative (TN): broj podataka koje je model točno klasificirao kao populaciju π_2
False Negative (FN): broj podataka koje je model pogrešno klasificirao kao populaciju π_2

Teorem 0.0.7 (Cauchy-Schwarz nejednakost)**.** *Neka su* $\mathbf{b}$ *i* $\mathbf{d}$ *bilo koja dva* $p \times 1$ *vektora. Tada vrijedi*

$$(\mathbf{b}'\mathbf{d})^2 \leq (\mathbf{b}'\mathbf{b})(\mathbf{d}'\mathbf{d}). \tag{6}$$

Jednakost vrijedi ako i samo ako $\mathbf{b} = c\mathbf{d}$ *(ili* $\mathbf{d} = c\mathbf{b}$ *) za neku konstantu c.*

Teorem 0.0.8 (Proširena Cauchy-Schwarz nejednakost)**.** *Neka su* $\underset{(p\times 1)}{\mathbf{b}}$ *i* $\underset{(p\times 1)}{\mathbf{d}}$ *bilo koja dva vektora, i neka je* $\underset{(p\times p)}{\mathbf{B}}$ *pozitivno definitna matrica. Tada vrijedi*

$$(\mathbf{b}'\mathbf{d})^2 \leq (\mathbf{b}'\mathbf{B}\mathbf{b})\left(\mathbf{d}'\mathbf{B}^{-1}\mathbf{d}\right). \tag{7}$$

Jednakost vrijedi ako i samo ako $\mathbf{b} = c\mathbf{B}^{-1}\mathbf{d}$ *(ili* $\mathbf{d} = c\mathbf{B}\mathbf{b}$*) za neku konstantu c.*

Poglavlje 1

Diskriminacija i klasifikacija

1.1 Diskriminacija i klasifikacija za dvije populacije

Motivacija

Diskriminacija i klasifikacija su tehnike koje se bave odvajanjem različitih grupa objekata i dodjeljivanjem novih objekata prethodno definiranim grupama. Diskriminacija se odnosi na postupak pronalaženja razlika između skupova podataka na temelju neke karakteristike te je više istraživačke prirode. Ovaj postupak može biti koristan u identifikaciji razlika između skupova podataka. Klasifikacija je postupak koji se koristi za kategoriziranje podataka u različite grupe ili klase na temelju svojstava koja se proučavaju.

Cilj diskriminacije u statističkoj analizi je identificirati značajke koje najbolje razlikuju različite skupove podataka, dok je cilj klasifikacije razvrstati podatke u različite grupe ili klase na temelju svojstava koja se proučavaju s naglaskom na pronalazak pravila za optimalno razvrstavanje objekata.

Razlika između diskriminacije i klasifikacije je u tome što se diskriminacija koristi za pronalaženje razlika između skupova podataka, dok se klasifikacija koristi za kategoriziranje podataka u različite klase. Međutim, u praksi se ciljevi diskriminacije i klasifikacije često preklapaju pa se oba postupka često koriste zajedno kako bi se identificirale ključne karakteristike koje se razlikuju između skupova podataka i kako bi se kategorizirali podaci u odgovarajuće klase ili kategorije.

Klasifikacija

Pretpostavimo da želimo podijeliti objekte u dvije grupe ili klase: π_1 i π_2. Objekte klasificiramo na temelju opaženih vrijednosti, tj n slučajnih varijabli $X = [X_1, X_2, \ldots, X_n]$. Skup svih opažanja za klasu π_1 možemo promatrati kao populaciju za π_1, slično za π_2. Populacije

možemo opisati vjerojatnosnim funkcijama gustoće $f_1(x)$ i $f_2(x)$. Sada možemo govoriti o dodjeljivanju opažanja populacijama ili objekata u klase.

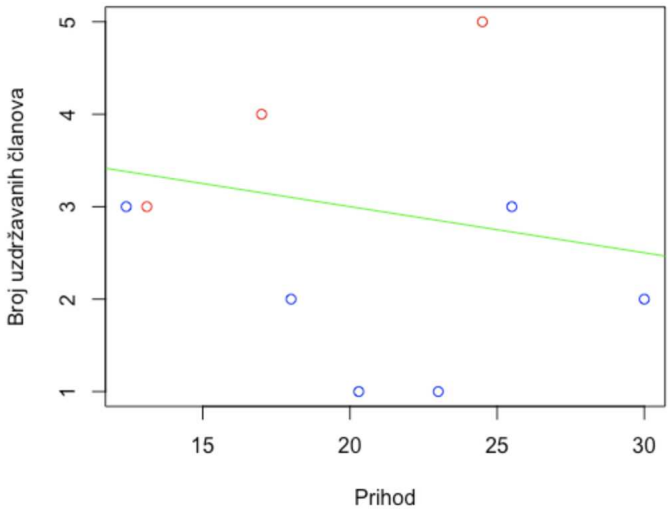
Primjer 1.1.1. *Glavna ideja je da objekt s pripadnim obilježjima klasificiramo kao dio populacije π_1 ili kao populacije π_2. Pravila klasifikacije se obično razvijaju iz postojećih podataka, tj uzoraka koji su prikupljeni i analizirani u određenom kontekstu i za koje se zna iz koje populacije dolaze. Sve moguće kombinacije promatranih vrijednosti podijelimo u dva skupa tako da prvi skup preferira pripadanje populaciji π_1, a drugi skup pripadanje populaciji π_2. Ponekad nije skroz jasno u koju klasu svrstati podatak jer često ne postoji jasna granica između klasa, tj postoji preklapanje između grupa što dovodi do pogrešne klasifikacije.*

Populacije π_1 i π_2	Promatrane varijable $\boldsymbol{X_i}$
Krediti koji se otplate i oni koji se ne otplate.	Prihod, godine, broj kredita, broj uzdržavanih članova

Primjer 1.1.2. *Sljedeća tablica prikazuje nekoliko klijenata kojima je odobren kredit s njihovim obilježjima.*

Prihod (u tisućama eura)	Broj uzdržavanih članova	Otplaćen kredit
18	2	da
20.3	1	da
23	1	da
24.5	5	ne
17	4	ne
13.1	3	ne
12.4	3	da
25.5	3	da
30	2	da

Na slici 1.1 vidimo kako klijenti koji vrate kredit uglavnom imaju veće prihode i manji broj uzdržavanih članova, ali svejedno dolazi do iznimaka. Koristeći neku od metoda koju ćemo spomenuti u narednim poglavljima definiramo pravac koji je zapravo granica koja dijeli prostor na dvije klase R_1 i R_2 tako da se minimizira vjerojatnost pogrešne klasifikacije. Vidimo da imamo jednu crvenu točku koja je pogrešno klasificirana kao plava, tj jednog klijenta koji nije vratio kredit smo pogrešno klasificirani kao nekog tko će vratiti kredit.



Slika 1.1: Grafički prikaz pogrešne klasifikacije

1.2 Kriteriji za optimalnu klasifikaciju

Smatramo da je klasifikacija učinkovita ukoliko rezultira s malo pogrešnih svrstavanja. Pogrešnu klasifikaciju je gotovo nemoguće izbjeći, ali ju je potrebno svesti na minimum, tj. vjerojatnost pogrešne klasifikacije treba biti najmanja moguća. Moguće je da je jedna populacija značajno veća od druge pa je i vjerojatnost pripadanja toj populaciji veća. Stoga je nužno uzeti u obzir apriorne vjerojatnosti. Apriorna vjerojatnost pojavljivanja se odnosi na početnu vjerojatnost koja se dodjeljuje događaju ili ishodu prije nego što se uzmu u obzir bilo kakve nove informacije ili podaci. Konkretno, novi objekt ćemo svrstati u klasu čija je apriorna vjerojatnost veća dokle god podaci značajno ne ukazuju na suprotno.

Neka su $f_1(x)$ i $f_2(x)$ vjerojatnosne funkcije gustoće povezane s vektorom slučajnih varijabli X za populacije π_1 i π_{2^*} Neka je Ω skup svih mogućih vrijednosti za vektor X, R_1 skup svih vrijednosti vektora X čije smo objekte klasificirali u populaciju π_1i slično za R_2. R_1 i R_2 su međusobno disjunktni i u uniji daju cijeli Ω. Vjerojatnost da objekt klasificiramo kao π_2, iako je ustvari π_1 označimo s

$$P(2 \mid 1) = P(X \in R2 \mid \pi 1) = \int_{R2=\Omega \setminus R1} f_1(x)dx. \tag{1.1}$$

Vjerojatnost da objekt klasificiramo kao π_1, iako je ustvari π_2 označimo s

$$P(1 \mid 2) = P(X \in R1 \mid \pi 2) = \int_{R1} f_2(x)dx. \tag{1.2}$$

Sada možemo definirati vjerojatnosti ispravne i pogrešne klasifikacije:

Definicija 1.2.1. *Neka je R_1 skup svih vrijednosti vektora X čije smo objekte klasificirali u populaciju π_1, a R_2 skup svih vrijednosti vektora X čije smo objekte klasificirali u populaciju π_2. Neka je p_1 = apriorna vjerojatnost pripadanja π_1 i s p_2 = apriorna vjerojatnost pripadanja π_2, tako da je $p_1 + p_2 = 1$. Tada definiramo:*
P(x je dobro klasificiran kao π_1) = P(x dolazi iz π_1 i dobro je klasificiran kao π_1)

$$= P(x \in R_1 \mid \pi_1)P(\pi_1) = P(1 \mid 1)p_1$$

P(x je pogrešno klasificiran kao π_1) = P(x dolazi iz π_2, a klasificiran je kao π_1)

$$= P(x \in R_1 \mid \pi_2)P(\pi_2) = P(1 \mid 2)p_2$$

P(x je dobro klasificiran kao π_2) = P(x dolazi iz π_2 i dobro je klasificiran kao π_2)

$$= P(x \in R_2 \mid \pi_2)P(\pi_2) = P(2 \mid 2)p_2$$

P(x je pogrešno klasificiran kao π_2) = P(x dolazi iz π_1, a klasificiran je kao π_2)

$$= P(x \in R_2 \mid \pi_1)P(\pi_1) = P(2 \mid 1)p_1$$

Očekivani trošak pogrešne klasifikacije (ECM)

Još jedan važan aspekt klasifikacije je trošak pogrešne klasifikacije. Ako se vratimo na početni primjer, puno je veći trošak ako nekoga klasificiramo kao klijenta koji će vratiti kredit, a on ga ne vrati, nego kada nekome ne odobrimo kredit iako bi ga vratio. Problem možemo prikazati pomoću matrice:

		klasificirana pripadnost	
		π_1	π_2
stvarna	π_1	0	$c\,(2 \mid 1)$
pripadnost	π_2	$c\,(1 \mid 2)$	0

Trošak je 0 ako je klasifikacija ispravna, a kada objekt klasificiramo kao π_1, iako je π_2 tada trošak iznosi $c\,(1 \mid 2)$, odnosno $c\,(2 \mid 1)$ kada objekt klasificiramo kao π_2, iako je π_1. Očekivani trošak pogrešne klasifikacije (engl. "expected cost of misclassification") iznosi

$$ECM = c\,(2 \mid 1)\,P\,(2 \mid 1)\,p_1 + c\,(1 \mid 2)\,P\,(1 \mid 2)\,p_2. \tag{1.3}$$

Teorem 1.2.2. *Područja R_1 i R_2 koja minimiziraju očekivani trošak pogrešne klasifikacije definirana su vrijednostima od x za koja vrijede sljedeće nejednakosti :*

$$\begin{aligned} R_1 &: \frac{f_1(x)}{f_2(x)} \geq \left(\frac{c(1 \mid 2)}{c(2 \mid 1)}\right)\left(\frac{p_2}{p_1}\right) \\ R_2 &: \frac{f_1(x)}{f_2(x)} < \left(\frac{c(1 \mid 2)}{c(2 \mid 1)}\right)\left(\frac{p_2}{p_1}\right) \end{aligned} \tag{1.4}$$

Dokaz. Uvrštavanjem izraza $P(2 \mid 1) = \int_{R2} f_1(x)dx$ i $P(1 \mid 2) = \int_{R_1} f_2(x)dx$ u izraz 1.3 za očekivani trošak pogrešne klasifikacije dobivamo sljedeće:

$$ECM = c(2 \mid 1)p_1 \int_{R_2} f_1(x)dx + c(1 \mid 2)p_2 \int_{R_1} f_2(x)dx. \tag{1.5}$$

Zbog $\Omega = \mathrm{R}_1 \cup \mathrm{R}_2$ vrijedi $1 = \int_{\Omega} f_1(x)dx = \int_{R_1} f_1(x)dx + \int_{R_2} f_1(x)dx$, pa možemo pisati

$$ECM = c(2 \mid 1)p_1 \left[1 - \int_{R_1} f_1(x)dx\right] + c(1 \mid 2)p_2 \int_{R_1} f_2(x)dx. \tag{1.6}$$

Zbog svojstva aditivnosti integrala imamo:

$$ECM = \int_{R_1} [c(1 \mid 2)p_2 f_2(x) - c(2 \mid 1)p_1 f_1(x)]dx + c(2 \mid 1)p_1. \tag{1.7}$$

Sada, $p_1, p_2, c(1 \mid 2), c(2 \mid 1)$ su nenegativni, a f_1 i f_2 su nenegativne za svaki x_i i jedine ovisne o x. Stoga vrijedi da je izraz za ECM minimiziran ako R_1 sadrži one x za koje vrijedi $c(1 \mid 2)p_2 f_2(x) - c(2 \mid 1)p_1 f_1(x) \leq 0$ i ne sadrži one x za koje je ovaj izraz pozitivan. □

Iz gornjeg rezultata vidimo da nam je za minimiziranje očekivanog troška pogrešne klasifikacije potreban omjer funkcija gustoća izračunate u novoj opaženoj točki, omjer troškova pogrešne klasifikacije i omjer apriornih vjerojatnosti. Nerijetko je jednostavnije odrediti omjere nego individualne vrijednosti. Kada apriorne vjerojatnosti nisu poznate često se uzima da su jednake pa se određivanje minimuma očekivanog troška (ECM) svodi na omjer funkcija gustoće i omjer troškova pogrešne klasifikacije :

$$\frac{p_2}{p_1} = 1 \quad \rightarrow \quad R_1 : \frac{f_1(x)}{f_2(x)} \geq \left(\frac{c(1 \mid 2)}{c(2 \mid 1)}\right), R_2 : \frac{f_1(x)}{f_2(x)} < \left(\frac{c(1 \mid 2)}{c(2 \mid 1)}\right). \tag{1.8}$$

Na sličan način, ukoliko omjer troškova nije poznat, uzima se da je jedan:

$$\frac{c(1 \mid 2)}{c(2 \mid 1)} = 1 \quad \rightarrow \quad \mathrm{R}_1 : \frac{f_1(x)}{f_2(x)} \geq \left(\frac{p2}{p_1}\right), R_2 : \frac{f_1(x)}{f_2(x)} < \left(\frac{p_2}{p_1}\right). \tag{1.9}$$

Zadnji specijalni slučaj se odnosi na to kada su oba omjera jednaka 1. Ako je $\frac{f_1(x)}{f_2(x)} \geq 1$ tada x klasificiramo kao π_1, inače ako je $\frac{f_1(x)}{f_2(x)} < 1$, onda x klasificiramo kao π_2. Postoje i drugi kriteriji za optimiziranu klasifikaciju. Ponekad se može ignorirati trošak pogrešne klasifikacije i birati R_1 i R_2 tako da se minimizira vjerojatnost pogrešne klasifikacije (engl. "total probability of misclassification") - TPM:

$$\begin{aligned} TMP &= P(\text{pogrešna klasifikacija objekta iz } \pi_1 \text{ili pogrešna klasifikacija objekta iz } \pi_2) \\ &= P(\text{pogrešna klasifikacija objekta iz } \pi_1) + P(\text{pogrešna klasifikacija objekta iz } \pi_2) = \\ &p_1 \int_{R_2} f_1(x)dx + p2 \int_{R_1} f_2(x)dx \end{aligned} \tag{1.10}$$

Matematički je ovaj problem jednak problemu minimizacije očekivanog troška pogrešne klasifikacije kada su troškovi pogrešne klasifikacije jednaki ($\frac{c(1|2)}{c(2|1)} = 1$). Još jedan primjer kriterija za optimalnu klasifikaciju je primjena aposteriornih vjerojatnosti na način da se novi podatak svrsta u onu klasu čija je aposteriorna vjerojatnost veća. Koristeći Bayesov teorem slijedi :

$$P(\pi_1 \mid x) = \frac{p_1 f_1(x)}{p_1 f_1(x) + p_2 f_2(x)}, \quad P(\pi_2 \mid x) = 1 - P(\pi_1 \mid x) = \frac{p_2 f_2(x)}{p_1 f_1(x) + p_2 f_2(x)}. \tag{1.11}$$

Poglavlje 2

Klasifikacija s dvije multivarijatne normalne populacije

2.1 Klasifikacija kada vrijedi $\Sigma_1 = \Sigma_2 = \Sigma$

Zbog svoje jednostavnosti i relativno visoke učinkovitosti u raznim populacijskim modelima, algoritmi klasifikacije temeljeni na normalnim populacijama prevladavaju u praksi. Poseban slučaj kada su kovarijacijske matrice jednake rezultira jednostavnom klasifikacijskom statistikom. Pretpostavljamo da su $f_1(x)$ i $f_2(x)$ multivarijatne normalne funkcije gustoće, s vektorima očekivanja μ_1 i μ_2 te kovarijacijskim matricama Σ_1 i Σ_2.

Definicija 2.1.1. *Funkcija gustoće multivarijatne normalne distribucije za vektor slučajnih varijabli* $\mathbf{X} = [X_1, X_2, ..., X_p]'$ *definira se kao*

$$f(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^p |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right) \tag{2.1}$$

gdje je $\boldsymbol{\mu}$ *vektor očekivanja slučajnih varijabli* $\mathbf{X}$*,* Σ *je simetrična kovarijacijska matrica između slučajnih varijabli* $\mathbf{X}$*, a* $|\boldsymbol{\Sigma}|$ *označava determinantu matrice* $\boldsymbol{\Sigma}$*.*

Pretpostavimo da su svi populacijski parametri poznati. Tada se izraz 1.4 svodi na

$$\begin{aligned} R_1 &: \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_1)' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_1) + \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_2)\right] \geq \left(\frac{c(1 \mid 2)}{c(2 \mid 1)}\right)\left(\frac{p_2}{p_1}\right) \\ R_2 &: \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_1)' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_1) + \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_2)\right] < \left(\frac{c(1 \mid 2)}{c(2 \mid 1)}\right)\left(\frac{p_2}{p_1}\right). \end{aligned} \tag{2.2}$$

Teorem 2.1.2. *Neka su populacije π_1 i π_2 opisane multivarijatnim normalnim gustoćama oblika* (2.1)*. Tada pravilo dodjele koje minimizira očekivani trošak pogrešne klasifikacije (ECM) je sljedeće:*
Pridruži x_0 populaciji π_1 ako

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} \mathbf{x}_0 - \frac{1}{2} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \Sigma^{-1} (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \geq \ln \left[\left(\frac{c(1 \mid 2)}{c(2 \mid 1)} \right) \left(\frac{p_2}{p_1} \right) \right] \tag{2.3}$$

Inače, pridruži x_0 populaciji π_2.

Dokaz. Budući da su sve veličine u izrazu nenegativne za svaki $\mathbf{x}$, možemo uzeti njihove prirodne logaritme i sačuvati redoslijed nejednakosti. Također vrijedi,

$$\begin{aligned} -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_1)' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_1) + \frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_2) \\ = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} \mathbf{x} - \frac{1}{2} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \end{aligned} \tag{2.4}$$

zbog čega slijedi,

$$\begin{aligned} R_1 : \quad & (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} \mathbf{x} - \frac{1}{2} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \geq \ln \left[\left(\frac{c(1 \mid 2)}{c(2 \mid 1)} \right) \left(\frac{p_2}{p_1} \right) \right] \\ R_2 : \quad & (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} \mathbf{x} - \frac{1}{2} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) < \ln \left[\left(\frac{c(1 \mid 2)}{c(2 \mid 1)} \right) \left(\frac{p_2}{p_1} \right) \right] \end{aligned} \tag{2.5}$$

□

Međutim, u većini situacija u praksi, populacijski parametri su nepoznati pa je potrebno modificirati pravilo (2.3). Jedna je mogućnost da se populacijski parametri zamijene s uzoračkim parametrima. Neka je $\mathbf{X}' = [X_1, X_2, ..., X_p]$ multivarijatna slučajna varijabla te promatramo n_1 opažanja iz populacije π_1 i n_2 opažanja iz populacije π_2 tako da vrijedi $n_1 + n_2 - 2 \geq p$. Sada imamo slučajne varijable

$$\begin{aligned} \mathbf{X}_{1,(n_1 \times p)} &= \begin{bmatrix} \mathbf{x}'_{11} \\ \mathbf{x}'_{12} \\ \vdots \\ \mathbf{x}'_{1n_1} \end{bmatrix} \\ \mathbf{X}_{2,(n_2 \times p)} &= \begin{bmatrix} \mathbf{x}'_{21} \\ \mathbf{x}'_{22} \\ \vdots \\ \mathbf{x}'_{2n_2} \end{bmatrix} \end{aligned} \tag{2.6}$$

s vektorima očekivanja i kovarijacijskim matricama

$$\begin{aligned}\underset{(p\times 1)}{\overline{\mathbf{x}}_1} &= \frac{1}{n_1}\sum_{j=1}^{n_1}\mathbf{x}_{1j},\ \underset{(p\times p)}{\mathbf{S}_1} = \frac{1}{n_1-1}\sum_{j=1}^{n_1}\left(\mathbf{x}_{1j}-\overline{\mathbf{x}}_1\right)\left(\mathbf{x}_{1j}-\overline{\mathbf{x}}_1\right)' \\ \underset{(p\times 1)}{\overline{\mathbf{x}}_2} &= \frac{1}{n_2}\sum_{j=1}^{n_2}\mathbf{x}_{2j},\ \underset{(p\times p)}{\mathbf{S}_2} = \frac{1}{n_2-1}\sum_{j=1}^{n_2}\left(\mathbf{x}_{2j}-\overline{\mathbf{x}}_2\right)\left(\mathbf{x}_{2j}-\overline{\mathbf{x}}_2\right)'\end{aligned} \tag{2.7}$$

Zbog pretpostavke o jednakosti kovarijacijskih matrica $\Sigma_1 = \Sigma_2 = \Sigma$, možemo kombiniranjem procijenjenih kovarijacijskih matrica S_1 i S_2 nepristrano procijeniti Σ.

Definicija 2.1.3. *Nepristrani procjenitelj za* Σ *definiramo kao*

$$\mathbf{S}_{proc} = \left[\frac{n_1-1}{(n_1-1)+(n_2-1)}\right]S_1 + \left[\frac{n_2-1}{(n_1-1)+(n_2-1)}\right]S_2. \tag{2.8}$$

Zamjenom μ_1 s $\overline{\mathbf{x}}_1$, μ_2 s $\overline{\mathbf{x}}_2$ i na kraju Σ s $\mathbf{S}_{proc}$ dobivamo modifikaciju pravila (2.3):

Korolar 2.1.4 (Linearno pravilo za klasifikaciju)**.** *Pridruži* x_0 *populaciji* π_1 *ako*

$$\left(\overline{\mathbf{x}}_1-\overline{\mathbf{x}}_2\right)'\mathbf{S}_{\boldsymbol{proc}}^{-1}\mathbf{x}_0 - \frac{1}{2}\left(\overline{\mathbf{x}}_1-\overline{\mathbf{x}}_2\right)'\mathbf{S}_{\boldsymbol{proc}}^{-1}\left(\overline{\mathbf{x}}_1+\overline{\mathbf{x}}_2\right) \geq \ln\left[\left(\frac{c(1\mid 2)}{c(2\mid 1)}\right)\left(\frac{p_2}{p_1}\right)\right] \tag{2.9}$$

Inače, pridruži x_0 *populaciji* π_2.

Primjer 2.1.5. *Pogledajmo poseban slučaj kada je* $\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right) = 1$. *Definirajmo*

$$\boldsymbol{y} = \left(\overline{\mathbf{x}}_1-\overline{\mathbf{x}}_2\right)'\mathbf{S}_{\boldsymbol{proc}}^{-1}\mathbf{x} = \mathbf{a}'\mathbf{x}. \tag{2.10}$$

Uvrštavanjem $\boldsymbol{x_o}$ *u* (2.10) *i ubacivanjem u* (2.9) *dobivamo*

$$\boldsymbol{m} = \frac{1}{2}\left(\overline{\mathbf{y}}_1+\overline{\mathbf{y}}_2\right), \tag{2.11}$$

gdje je $\overline{\mathbf{y}}_1 = \mathbf{a}'\overline{\mathbf{x}}_1$ *i* $\overline{\mathbf{y}}_2 = \mathbf{a}'\overline{\mathbf{x}}_2$. *Sada je pravilo za minimizaciju očekivanog troška pogrešne klasifikacije za dvije normalne populacije ekvivalentno stvaranju dvije jednovarijatne populacije za vrijednost y tako da se uzme odgovarajuća linearna kombinacija opažanja iz* π_1 *i* π_2. *Novi podatak* $\mathbf{x}_0$ *se dodjeljuje populaciji na način da se promatra pada li* $\mathbf{y}_0 = \mathbf{a}'\mathbf{x}_0$ *desno ili lijevo od* $\boldsymbol{m}$*, tj. srednje vrijednosti dvaju jednovarijatnih srednjih vrijednosti* $\overline{\mathbf{y}}_1$ *i* $\overline{\mathbf{y}}_2$.

Kada populacijske parametre zamijenimo s procijenjenima, nema garancije da će pravilo zbilja minimizirati očekivani trošak klasifikacije jer je optimalno pravilo (2.3) određeno uz pretpostvku da su $f_1(x)$ i $f_2(x)$ u potpunosti poznate. Međutim, što je veći uzorak, to je veća vjerojatnost bolje procjene.

2.2 Fisherov pristup za klasifikaciju s dvije normalne populacije

Fisher je do izraza (2.10) došao malo drugačijim načinom (vidi [3]). Fisherova ideja bila je transformirati multivarijatna opažanja $\mathbf{x}$ u jednovarijatna opažanja y tako da su y izvedeni iz populacije π_1 i π_2 odvojeni što je više moguće. Uzimao je linearne kombinacije od $\mathbf{x}$ za stvaranje y jer su to bile dovoljno jednostavne funkcije od x za korištenje. Ovaj pristup ne pretpostavlja normalnost populacija, ali implicitno pretpostavlja da su kovarijacijske matrice populacija jednake zbog korištenja zajedničke kovarijacijske matrice. Koliko su dva skupa podataka udaljeni, procijenjuje se na način da se promatra razlika između $\overline{\mathbf{y}}_1$ i $\overline{\mathbf{y}}_2$. Separaciju tada definiramo kao

$$\mathbf{D} = \frac{|\overline{\mathbf{y}}_1 - \overline{\mathbf{y}}_2|}{\mathbf{s}_y}, \quad \text{gdje je} \quad \mathbf{s}_y = \frac{\sum_{j=1}^{n_1} \left(\mathbf{y}_{1j} - \overline{\mathbf{y}}_1\right)^2 + \sum_{j=1}^{n_2} \left(\mathbf{y}_{2j} - \overline{\mathbf{y}}_2\right)^2}{\mathbf{n}_1 + \mathbf{n}_2 - 2} \tag{2.12}$$

procjena zajedničke varijance. Cilj je pronaći linearnu kombinaciju od $\mathbf{x}$ koja maksimizira separaciju $\overline{\mathbf{y}}_1$ i $\overline{\mathbf{y}}_2$.

Lema 2.2.1. *Neka je* $\mathbf{B}$ *pozitivno definitna matrica reda* $p \times p$ *i* $\mathbf{d}$ *dani vektor. Tada za slučajni nenul vektor* $\mathbf{x}$ *vrijedi*

$$\max_{\mathbf{x} \neq 0} \frac{(\mathbf{x}'\mathbf{d})^2}{\mathbf{x}'\mathbf{B}\mathbf{x}} = \mathbf{d}'\mathbf{B}^{-1}\mathbf{d} \tag{2.13}$$

s maksimumom koji se postiže za $\mathbf{x} = c\mathbf{B}^{-1}\mathbf{d}$ *za bilo koju konstantu* $c \neq 0$.

Dokaz. Iz proširene Cauchy-Schwarz nejednakosti (0.0.8) slijedi $(\mathbf{x}'\mathbf{d})^2 \leq (\mathbf{x}'\mathbf{B}\mathbf{x})\left(\mathbf{d}'\mathbf{B}^{-1}\mathbf{d}\right)$. Zato što je $\mathbf{x} \neq \mathbf{0}$ i $\mathbf{B}$ je pozitivno definitna, slijedi $\mathbf{x}'\mathbf{B}\mathbf{x} > 0$. Dijeljenjem obje strane nejednakosti s pozitivnim skalarom $\mathbf{x}'\mathbf{B}\mathbf{x}$ dobivamo gornju granicu

$$\frac{(\mathbf{x}'\mathbf{d})^2}{\mathbf{x}'\mathbf{B}\mathbf{x}} \leq \mathbf{d}'\mathbf{B}^{-1}\mathbf{d} \tag{2.14}$$

Uzimanjem maksimuma po $\mathbf{x}$ dobivamo jednadžbu (2.13) jer se gornja granica postiže za $\mathbf{x} = c\mathbf{B}^{-1}\mathbf{d}$ □

Teorem 2.2.2. *Linearna kombinacija* $\hat{y} = \hat{\mathbf{a}}'\mathbf{x} = (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)\,\mathbf{S}_{proc}^{-1}\mathbf{x}$ *maksimizira omjer*

$$\frac{(\bar{y}_1 - \bar{y}_2)^2}{s_y^2} = \frac{(\hat{\mathbf{a}}'\overline{\mathbf{x}}_1 - \hat{\mathbf{a}}'\overline{\mathbf{x}}_2)^2}{\hat{\mathbf{a}}'\mathbf{S}_{proc}\hat{\mathbf{a}}} = \frac{(\hat{\mathbf{a}}'\mathbf{d})^2}{\hat{\mathbf{a}}'\mathbf{S}_{proc}\hat{\mathbf{a}}} \tag{2.15}$$

preko svih mogućih koeficijenata vektora $\hat{\mathbf{a}}$ *gdje je* $\mathbf{d} = (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)$. *Maksimum omjera* (2.15) *je* $D^2 = (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)'\,\mathbf{S}_{proc}^{-1}\,(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)$.

Dokaz. Maksimum omjera (2.15) slijedi direktnom primjenom (2.13). Označimo $d = (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)$, pa dobivamo

$$\max_{\hat{\mathbf{a}}} \frac{(\hat{\mathbf{a}}'\mathbf{d})^2}{\hat{\mathbf{a}}'\mathbf{S}_{proc}\hat{\mathbf{a}}} = \mathbf{d}'\mathbf{S}_{proc}^{-1}\mathbf{d} = (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)' \mathbf{S}_{proc}^{-1} (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2) = D^2 \tag{2.16}$$

gdje je D^2 uzoračka kvadratna udaljenost između dvije srednje vrijednosti. □

Fisherovo rješenje za separaciju se također može koristiti za klasifikaciju novih opažanja:

Definicija 2.2.3 (Fisherovo pravilo za klasifikaciju)**.** *Dodijeli* $\mathbf{x}_0$ *u* π_1 *ako*

$$\hat{y}_0 = (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)' \mathbf{S}_{proc}^{-1}\mathbf{x}_0 \geq \hat{m} = \frac{1}{2} (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)' \mathbf{S}_{proc}^{-1} (\overline{\mathbf{x}}_1 + \overline{\mathbf{x}}_2) \tag{2.17}$$

Dodijeli $\mathbf{x}_0$ *u* π_2 *ako*

$$\hat{y}_0 < \hat{m} \tag{2.18}$$

Uvjet koji mora vrijediti je $(n_1 + n_2 - 2) \geq p$ kako $\mathbf{S}_{proc}$ ne bi bila singularna i kako bi inverz $\mathbf{S}_{proc}^{-1}$ postojao.

2.3 Klasifikacija kada $\Sigma_1 \neq \Sigma_2$

Pretpostavimo sada da su populacijske kovarijacijske matrice različite, tj $\Sigma_1 \neq \Sigma_2$. Zbog toga su i vektori očekivanja različiti. Funkcije gustoće definiramo kao u (2.1), za $i = 1, 2$. Supstituiranjem funkcija gustoća u izraz (1.4) i uzimanjem prirodnog logaritma dobivamo sljedeće:

$$\begin{aligned} R_1 : \quad & -\frac{1}{2}\mathbf{x}' \left(\Sigma_1^{-1} - \Sigma_2^{-1}\right)\mathbf{x} + \left(\boldsymbol{\mu}_1'\Sigma_1^{-1} - \boldsymbol{\mu}_2'\Sigma_2^{-1}\right)\mathbf{x} - k \geq \ln\left[\left(\frac{c(1 \mid 2)}{c(2 \mid 1)}\right)\left(\frac{p_2}{p_1}\right)\right] \\ R_2 : \quad & -\frac{1}{2}\mathbf{x}' \left(\Sigma_1^{-1} - \Sigma_2^{-1}\right)\mathbf{x} + \left(\boldsymbol{\mu}_1'\Sigma_1^{-1} - \boldsymbol{\mu}_2'\Sigma_2^{-1}\right)\mathbf{x} - k < \ln\left[\left(\frac{c(1 \mid 2)}{c(2 \mid 1)}\right)\left(\frac{p_2}{p_1}\right)\right] \end{aligned} \tag{2.19}$$

gdje je

$$k = \frac{1}{2} \ln\left(\frac{|\Sigma_1|}{|\Sigma_2|}\right) + \frac{1}{2}\left(\mu_1'\Sigma_1^{-1}\mu_1 - \mu_2'\Sigma_2^{-1}\mu_2\right). \tag{2.20}$$

Sada smo za klasifikacijska područja dobili kvadratne funkcije od $\mathbf{x}$. Kvadratni član $-\frac{1}{2}\mathbf{x}' \left(\Sigma_1^{-1} - \Sigma_2^{-1}\right)\mathbf{x}$ nestaje kada $\Sigma_1 = \Sigma_2$.

Teorem 2.3.1. *Neka su π_1 i π_2 populacije s multivarijatnim normalnim funkcijama gustoće, s vektorima očekivanja i kovarijacijskim matricama $\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1$ i $\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2$. Pravilo pridruživanja koje minimizira očekivani trošak pogrešne klasifikacije je sljedeće:*
Dodijeli $\mathbf{x}_0$ u π_1 ako

$$-\frac{1}{2}\mathbf{x}_0'\left(\boldsymbol{\Sigma}_1^{-1}-\boldsymbol{\Sigma}_2^{-1}\right)\mathbf{x}_0+\left(\boldsymbol{\mu}_1'\boldsymbol{\Sigma}_1^{-1}-\boldsymbol{\mu}_2'\boldsymbol{\Sigma}_2^{-1}\right)\mathbf{x}_0-k\geq\ln\left[\left(\frac{c(1\mid 2)}{c(2\mid 1)}\right)\left(\frac{p_2}{p_1}\right)\right] \tag{2.21}$$

Inače, dodijeli $\mathbf{x}_0$ u π_2.
Ovdje je k definiran kao u (2.20).

U praksi se gornje pravilo implementira na način da se populacijska očekivanja i kovarijacijske matrice $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1$, i $\boldsymbol{\Sigma}_2$ zamijene s uzoračkim veličinama $\overline{\mathbf{x}}_1, \overline{\mathbf{x}}_2, \mathbf{S}_1$, and $\mathbf{S}_2$.

Korolar 2.3.2 (Kvadratno pravilo za klasifikaciju)**.** *Neka su $\overline{\mathbf{x}}_1$ i $\overline{\mathbf{x}}_2$ uzoračka očekivanja, a $\mathbf{S}_1$ i $\mathbf{S}_2$ uzoračke kovarijacijske matrice. Tada pravilo za klasifikaciju glasi:*
Dodijeli $\mathbf{x}_0$ u π_1 ako

$$-\frac{1}{2}\mathbf{x}_0'\left(\mathbf{S}_1^{-1}-\mathbf{S}_2^{-1}\right)\mathbf{x}_0+\left(\overline{\mathbf{x}}_1'\mathbf{S}_1^{-1}-\overline{\mathbf{x}}_2'\mathbf{S}_2^{-1}\right)\mathbf{x}_0-k\geq\ln\left[\left(\frac{c(1\mid 2)}{c(2\mid 1)}\right)\left(\frac{p_2}{p_1}\right)\right] \tag{2.22}$$

Inače, dodijeli $\mathbf{x}_0$ u π_2.

Ukoliko podaci nisu normalni, možemo ih transformirati u podatke koji su bliže normalnima, a može se provesti i test za jednakost kovarijacijskih matrica kako bi znali trebamo li iskoristiti linearno (2.9) ili kvadratno (2.22) pravilo za klasifikaciju. Koji god postupak odabrali, potrebno je provjeriti performanse klasifikacijskog postupka. Ukoliko je uzorak podataka dovoljno velik, česta je praksa da se podaci podijele u *train* i *test* poduzorke, gdje se *train* uzorak koristi za razvoj klasifikacijskih funkcija, a *test* uzorak za procjenu performanse i kvalitete postupka.

Poglavlje 3

Procjena klasifikacijskih funkcija

U prijašnjim poglavljima smo spomenuli kako je važno procijeniti performanse klasifikacijskog postupka, a jedan od načina je izračunati razinu pogreške ili vjerojatnost pogrešne klasifikacije. Kada su populacijski parametri u potpunosti poznati tada je relativno jednostavno izračunati vjerojatnost pogrešne klasifikacije. Međutim, u praksi je to rijetkost, pa se tada koncentriramo na izračun razine pogreške povezane s uzoračkom klasifikacijskom funkcijom.

Vjerojatnost pogrešne klasifikacije (engl. "total probability of misclassification") - TPM definiramo kao

$$TPM = p_1 \int_{R_2} f_1(x)dx + p_2 \int_{R_1} f_2(x)dx. \tag{3.1}$$

Pametnim odabirom R_1 i R_2 dobivamo najmanju vrijednost TMP-a koju zovemo *optimalna razina pogreške - OER*.

Definicija 3.0.1. *Optimalnu razinu pogreške definiramo kao*

$$\text{OER} = p_1 \int_{R_2} f_1(\mathbf{x})d\mathbf{x} + p_2 \int_{R_1} f_2(\mathbf{x})d\mathbf{x} \tag{3.2}$$

gdje su R_1 i R_2 definirani kao u (1.9)*:*

$$\text{R}_1 : \frac{f_1(x)}{f_2(x)} \geq \left(\frac{p_2}{p_1}\right), R_2 : \frac{f_1(x)}{f_2(x)} < \left(\frac{p_2}{p_1}\right). \tag{3.3}$$

Ukoliko populacijske parametre moramo procijeniti iz uzorka, tada se performanse klasifikacijskih funkcija mogu procijeniti pomoću *stvarne razine pogrešne*
(engl. "actual error rate") - AER.

Definicija 3.0.2. *Stvarnu razinu pogreške definiramo kao*

$$\text{AER} = p_1 \int_{\hat{R}_2} f_1(\mathbf{x})d\mathbf{x} + p_2 \int_{\hat{R}_1} f_2(\mathbf{x})d\mathbf{x} \tag{3.4}$$

gdje $\hat{R}_1$ i $\hat{R}_2$ predstavljaju područja klasifikacije određena pomoću veličine uzoraka n_1 i n_2.

Iako se AER, slično kao i OER, ne može izračunati zbog nepoznatih funkcija gustoće $f_1(\mathbf{x})$ i $f_2(\mathbf{x})$, može se odrediti procjena AER-a. Postoji veličina koja ne ovisi o populacijskim parametrima te se može izračunati za bilo koji klasifikacijski postupak. Radi se o *očitoj razini pogreške* (engl. "apparent error rate") - APER , koja se definira kao udio opažanja u *train* skupu koji su pogrešno klasificirani korištenjem klasifikacijske funkcije na tom skupu. APER se može jednostavno izračunati iz konfuzijske matrice.

Primjer 3.0.3. *Neka je dana konfuzijska matrica:*

		klasificirana pripadnost π_1	π_2	
stvarna pripadnost	π_1	n_{1C}	$n_{1M} = n_1 - n_{1C}$	n_1
	π_2	$n_{2M} = n_2 - n_{2C}$	n_{2C}	n_2

gdje su

$$\begin{aligned} n_{1C} &= \textit{ broj opažanja iz } \pi_1 \textit{ koja su točno klasificirana} \\ n_{2C} &= \textit{ broj opažanja iz } \pi_2 \textit{ koja su točno klasificirana} \\ n_{1M} &= \textit{ broj opažanja iz } \pi_1 \textit{ koja su pogrešno klasificirana} \\ n_{2M} &= \textit{ broj opažanja iz } \pi_2 \textit{ koja su pogrešno klasificirana} \end{aligned}$$

Tada očitu razinu pogreške računamo kao

$$\text{APER} = \frac{n_{1M} + n_{2M}}{n_1 + n_2} \tag{3.5}$$

što je udio opažanja koja su pogrešno klasificirana u train skupu podataka.

Lachenbruchova metoda odgađanja

Započinjemo proceduru sa skupom podataka iz populacije π_1. Izbacimo iz skupa jedan podatak i razvijemo klasifikacijsku funkciju na temelju preostalih $n_1 - 1$ podataka iz populacije π_1 i n_2 podataka iz populacije π_2. Zatim klasificiramo onaj podatak koji smo na početku izbacili na način da koristimo upravo razvijenu klasifikacijsku funkciju. I sada

ponavljamo postupak izbacivanja i ponovnog klasificiranja dok god ne iskoristimo sve podatke iz populacije π_1. Označimo s $n_{1M}^{(H)}$ broj podataka koji je pogrešno klasificiran za populaciju π_1. Zatim cijeli postupak ponovimo za populaciju π_2. Označimo s $n_{2M}^{(H)}$ broj podataka koji su pogrešno klasificirani za populaciju π_2. Procjene uvjetne vjerojatnosti pogrešne klasifikacije definirane u (1.1) i (1.2) dane su kao

$$\begin{aligned} \hat{P}(2 \mid 1) &= \frac{n_{1M}^{(H)}}{n_1} \\ \hat{P}(1 \mid 2) &= \frac{n_{2M}^{(H)}}{n_2}, \end{aligned} \tag{3.6}$$

a procjena očekivane stvarne razine pogreške dana je kao

$$\hat{E}(\mathrm{AER}) = \frac{n_{1M}^{(H)} + n_{2M}^{(H)}}{n_1 + n_2}. \tag{3.7}$$

Poglavlje 4

Klasifikacija s više populacija

U teoriji je generalizacija klasifikacijskih metoda s dvije populacije na više njih jednostavan proces. Međutim, ne zna se puno o svojstvima odgovarajućih uzoračkih klasifikacijskih funkcija niti o njihovim razinama pogrešaka (vidi [3]). U ovom poglavlju ćemo navesti teorijska pravila te navesti potencijalne modifikacije za primjenu na stvarnim primjerima. U većini slučajeva ćemo uzimati da se radi o normalnim multivarijatnim funkcijama gustoće za pripadne populacije iako to nije nužno za razvijanje opće teorije.

4.1 Minimalni očekivani trošak pogrešne klasifikacije s više populacija

Definicija 4.1.1. *Neka su $f_i(\mathbf{x})$ funkcije gustoće za populacije $\pi_i, i = 1, 2, \ldots, g$. Neka je*

$$\begin{aligned} p_i &= \textit{apriorna vjerojatnost pripadanja populaciji } \pi_i, \quad i = 1, 2, \ldots, g \\ c(k \mid i) &= \textit{ trošak pogrešne klasifikacije objekta kao } \pi_k, \quad k, i = 1, 2, \ldots, g \end{aligned} \tag{4.1}$$

Za $k = i, c(i \mid i) = 0$. Neka je R_k skup x-eva koji su klasificirani kao π_k i

$$\begin{aligned} P(k \mid i) &= P(\textit{ podatak klasificiran kao } \pi_k \mid \pi_i) = \int_{R_k} f_i(\mathbf{x}) d\mathbf{x} \\ &\textit{za } k, i = 1, 2, \ldots, g \textit{ s } P(i \mid i) = 1 - \sum_{\substack{k=1 \\ k \neq i}}^{g} P(k \mid i). \end{aligned} \tag{4.2}$$

Uvjetni očekivani trošak pogrešne klasifikacije x *iz π_1 kao $\pi_2, \ldots$ ili π_g je*

$$\begin{aligned} \mathbf{ECM}(1) &= P(2 \mid 1)c(2 \mid 1) + P(3 \mid 1)c(3 \mid 1) + \cdots + P(g \mid 1)c(g \mid 1) \\ &= \sum_{k=2}^{g} P(k \mid 1)c(k \mid 1). \end{aligned} \tag{4.3}$$

Uvjetni očekivani trošak u gornjoj definiciji se pojavljuje uz apriornu vjerojatnost pripadanja p_1. Na sličan način definiramo i uvjetni očekivani trošak ECM(2), ..., ECM(g).

Definicija 4.1.2. *Množenjem svakog uvjetnog očekivanog troška ECM(i) s pripadnom apriornom vjerojatnosti p_i i sumiranjem dobivamo ukupan očekivani trošak :*

$$\begin{aligned} \text{ECM} &= p_1\,\text{ECM}(1) + p_2\,\text{ECM}(2) + \cdots + p_g\,\text{ECM}(g) \\ &= p_1\left(\sum_{k=2}^{g} P(k \mid 1)c(k \mid 1)\right) + p_2\left(\sum_{\substack{k=1 \\ k\neq 2}}^{g} P(k \mid 2)c(k \mid 2)\right) + \cdots + p_g\left(\sum_{k=1}^{g-1} P(k \mid g)c(k \mid g)\right) \\ &= \sum_{i=1}^{g} p_i\left(\sum_{\substack{k=1 \\ k\neq i}}^{g} P(k \mid i)c(k \mid i)\right) \end{aligned} \tag{4.4}$$

Optimalna klasifikacijska procedura uključuje određivanje područja $R_1, R_2, \ldots, R_g$ koja minimiziraju ukupan trošak (4.4).

Teorem 4.1.3. *Klasifikacijska područja koja minimiziraju ECM* (4.4) *su određena pridruživanjem* x*-a populaciji $\pi_k, k = 1, 2, \ldots, g$, za koje je*

$$\sum_{\substack{i=1 \\ i\neq k}}^{g} p_i f_i(\mathbf{x})c(k \mid i) \tag{4.5}$$

najmanja. Ukoliko imamo dvije populacije za koje je izraz najmanji, **x** *pridružimo bilo kojoj populaciji.*

Dokaz. Neka je **x** točka u $\mathbb{R}^p$ koju klasificiramo u g populacija s apriornim vjerojatnostima $p_1, p_2, \ldots, p_g$. Očekivani trošak pogrešne klasifikacije ove točke je

$$\begin{aligned} \text{ECM}(\mathbf{x}) = \sum_{k=1}^{g} \text{ECM}(\mathbf{x}, \pi_k) &= \sum_{k=1}^{g}\left[c(k \mid k)p_k + \sum_{i\neq k} c(k \mid i)p_i\right] f_k(\mathbf{x}) \\ &= \sum_{k=1}^{g}\left[p_k - c(k \mid k)p_k - \sum_{i\neq k} c(k \mid i)p_i\right] f_k(\mathbf{x}) + \sum_{i=1}^{g} c(i \mid k)p_i f_k(\mathbf{x}) \\ &= \sum_{k=1}^{g}\sum_{i\neq k} p_i f_k(\mathbf{x})c(k \mid i) + \sum_{k=1}^{g} p_k f_k(\mathbf{x})\left[1 - c(k \mid k)\right]. \end{aligned} \tag{4.6}$$

Prva suma u posljednjoj liniji odgovara trošku pogrešne klasifikacije **x** u populaciju π_k i zatim bilo koju drugu populaciju π_i, a druga suma odgovara trošku pogrešne klasifikacije

x u populaciju π_k i zatim pogrešku u klasifikaciji same populacije π_k. Druga suma ne ovisi o tome kojoj populaciji **x** pripada i može se zanemariti u postupku minimizacije.

Stoga, za minimizaciju ECM-a potrebno je pronaći populaciju π_k koja minimizira sljedeći izraz:

$$\sum_{i \neq k} p_i f_k(\mathbf{x}) c(k \mid i). \tag{4.7}$$

Kako su p_i i $f_k(\mathbf{x})$ fiksirani, ovo se može svesti na pronalazak populacije π_k za koju je

$$\sum_{i \neq k} c(k \mid i) < c(j \mid i) \text{ za sve } j \neq k. \tag{4.8}$$

Drugim riječima, populacija π_k mora biti ona za koju je očekivani trošak pogrešne klasifikacije **x** u tu populaciju i bilo koju drugu populaciju najmanji. Ako postoji više populacija za koje je ovaj izraz minimalan, onda je dopušteno **x** pridružiti bilo kojoj od tih populacija. □

Korolar 4.1.4 (Minimalni očekivani trošak s jednakim troškovima pogrešne klasifikacije)**.** *Bez smanjenja općenitosti možemo pretpostaviti da su svi troškovi pogrešne klasifikacije jednaki. Tada bismo x_0 pridružili populaciji za koju je $\sum_{\substack{i=1 \\ i \neq k}}^{g} p_i f_i(\mathbf{x_0})$ minimalan, tj pridruži x_0 populaciji π_k ako*

$$p_k f_k(\mathbf{x}) > p_i f_i(\mathbf{x}) \text{ za sve } i \neq k, \tag{4.9}$$

tj ekvivalentno

$$\ln p_k f_k(\mathbf{x}) > \ln p_i f_i(\mathbf{x}) \text{ za sve } i \neq k \tag{4.10}$$

Primjer 4.1.5. *Klasificirajmo novi podatak x_0 u jednu od $g = 3$ populacija π_1, π_2 ili π_3 koristeći proceduru za minimalni očekivani trošak* (4.5)*. Neka su dani sljedeći podaci*

		Prava populacija		
		π_1	π_2	π_3
Klasificirano kao:	π_1	$c(1 \mid 1) = 0$	$c(1 \mid 2) = 500$	$c(1 \mid 3) = 100$
	π_2	$c(2 \mid 1) = 10$	$c(2 \mid 2) = 0$	$c(2 \mid 3) = 50$
	π_3	$c(3 \mid 1) = 50$	$c(3 \mid 2) = 200$	$c(3 \mid 3) = 0$
Apriorne vjerojatnosti		$p_1 = 0.05$	$p_2 = 0.60$	$p_3 = 0.35$
Gustoća u x_0		$f_1(\mathbf{x}_0) = 0.01$	$f_2(x_0) = 0.85$	$f_3(\mathbf{x}_0) = 2$

Vrijednosti izraza $\sum_{\substack{i=1 \\ i \neq k}}^{3} p_i f_i(\mathbf{x}_0) c(k \mid i)$ su redom

$$k = 1 : p_2 f_2(\mathbf{x}_0) c(1/2) + p_3 f_3(\mathbf{x}_0) c(1 \mid 3) = (0.60)(0.85)(500) + (0.35)(2)(100) = 325$$
$$k = 2 : p_1 f_1(\mathbf{x}_0) c(2 \mid 1) + p_3 f_3(\mathbf{x}_0) c(2 \mid 3) = (0.05)(0.01)(10) + (0.35)(2)(50) = 35.055$$
$$k = 3 : p_1 f_1(\mathbf{x}_0) c(3 \mid 1) + p_2 f_2(\mathbf{x}_0) c(3 \mid 2) = (0.05)(0.01)(50) + (0.60)(0.85)(200) = 102.025$$

Pošto je vrijednost izraza $\sum_{\substack{i=1 \\ i\neq k}}^{3} p_i f_i(\mathbf{x}_0)\, c(k \mid i)$ *najmanja za* $k = 2$*, podatak* $\mathbf{x}_0$ *alociramo u populaciju* π_2*.*

Da su svi troškovi pogrešne klasifikacije bili jednaki, mogli smo koristiti pravilo definirano kao u (4.9)*, što zahtjeva računanje samo sljedećih produkata*

$$\begin{aligned} p_1 f_1(\mathbf{x}_0) &= (0.05)(0.01) = 0.0005 \\ p_2 f_2(\mathbf{x}_0) &= (0.60)(0.85) = 0.510 \\ p_3 f_3(\mathbf{x}_0) &= (0.35)(2) = 0.700 \end{aligned}$$

Pošto je

$$p_3 f_3(\mathbf{x}_0) = 0.700 \geq p_i f_i(\mathbf{x}_0), i = 1, 2$$

novi podatak $\mathbf{x}_0$ *bismo alocirali u populaciju* π_3*.*

4.2 Klasifikacija s više normalnih populacija

Neka su sada funkcije gustoće normalne multivarijatne funkcije, tj

$$f_i(\mathbf{x}) = \frac{1}{(2\pi)^{p/2} |\mathbf{\Sigma}_i|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)' \mathbf{\Sigma}_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i)\right], i = 1, 2, \ldots, g \tag{4.11}$$

gdje su μ_i vektori očekivanja i Σ_i kovarijacijske matrice. Nadalje, neka su troškovi pogrešne klasifikacije jednaki, tj $c(i \mid i) = 0, c(k \mid i) = 1, k \neq i$. Tada se uvjet (4.10) svodi na: Pridruži x populaciji π_k ako

$$\begin{aligned} \ln p_k f_k(\mathbf{x}) = \ln p_k - \left(\frac{p}{2}\right)\ln(2\pi) - \frac{1}{2}\ln|\mathbf{\Sigma}_k| - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)' \mathbf{\Sigma}_k^{-1} (\mathbf{x} - \boldsymbol{\mu}_k) \\ = \max_i \ln p_i f_i(\mathbf{x}) \end{aligned} \tag{4.12}$$

Konstanta $(p/2)\ln(2\pi)$ u (4.12) se može ignorirati pošto je jednaka za sve populacije.

Definicija 4.2.1. *Kvadratni diskriminacijski bod za i-tu populaciju definiramo kao*

$$d_i^Q(\mathbf{x}) = -\frac{1}{2}\ln|\Sigma_i| - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)' \mathbf{\Sigma}_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) + \ln p_i, i = 1, 2, \ldots, g \tag{4.13}$$

Koristeći gornju definiciju i izraz (4.12) dobivamo novo klasifikacijsko pravilo:

Definicija 4.2.2 (Minimalna ukupna vjerojatnost pogrešne klasifikacije (TPM) za normalne populacije s nejednakim kovarijacijskim matricama)**.** *Pridruži* x *populaciji* π_k *ako je* $d_k^Q(\mathbf{x})$ *najveći od* $d_1^Q(\mathbf{x}), d_2^Q(\mathbf{x}), \ldots, d_g^Q(\mathbf{x})$ *gdje je* $d_i^Q(\mathbf{x})$ *dano kao u* (4.13)*.*

U praksi su $\boldsymbol{\mu}_i$ i Σ_i najčešće nepoznati, ali je skup podataka točno klasificiranih podataka dostupan pa on može poslužiti za izračun procjena. Zanimaju nas vektori uzoračkih očekivanja $\overline{\mathbf{x}}_i$ i uzoračka kovarijacijska matrica $\mathbf{S}_i$ te veličina uzorka n_i. Procjena kvadratnog diskriminacijskog boda je tada dana kao

$$\hat{d}_i^Q(\mathbf{x}) = -\frac{1}{2}\ln|\mathbf{S}_i| - \frac{1}{2}(\mathbf{x}-\overline{\mathbf{x}}_i)'\,\mathbf{S}_i^{-1}(\mathbf{x}-\overline{\mathbf{x}}_i) + \ln p_i, \quad i = 1,2,\ldots,g \tag{4.14}$$

Klasifikacijsko pravilo na temelju uzorka glasi:

Definicija 4.2.3 (Procijenjena minimalna ukupna vjerojatnost pogrešne klasifikacije (TPM) za normalne populacije s nejednakim kovarijacijskim matricama)**.** *Pridruži* x *populaciji* π_k *ako je* $\hat{d}_k^Q(\mathbf{x})$ *najveći od* $\hat{d}_1^Q(\mathbf{x}), \hat{d}_2^Q(\mathbf{x}), \ldots, \hat{d}_g^Q(\mathbf{x})$ *gdje je* $\hat{d}_i^Q(\mathbf{x})$ *dano kao u* (4.14)*.*

Definicija 4.2.4. *Ukoliko su kovarijacijske matrice jednake* $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$*, za* $i = 1,2,\ldots,g$*, tada diskriminacijski bod iz* (4.13) *glasi*

$$d_i^Q(\mathbf{x}) = -\frac{1}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x} + \boldsymbol{\mu}_i'\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_i'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \ln p_i \tag{4.15}$$

Definicija 4.2.5. *Linearni diskriminacijski bod za i-tu populaciju definiramo kao*

$$d_i(\mathbf{x}) = \boldsymbol{\mu}_i'\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_i'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \ln p_i \qquad \text{za } i = 1,2,\ldots,g \tag{4.16}$$

Definicija 4.2.6. *Neka je procjenitelj za* $\boldsymbol{\Sigma}$ *dan kao*

$$\mathbf{S}_{proc} = \frac{1}{n_1+n_2+\cdots+n_g-g}\Big((n_1-1)\,\mathbf{S}_1 + (n_2-1)\,\mathbf{S}_2 + \cdots + \Big(n_g-1\Big)\mathbf{S}_g\Big) \tag{4.17}$$

Procijenjeni linearni diskriminacijski bod za i-tu populaciju definiramo kao

$$\hat{d}_i(\mathbf{x}) = \overline{\mathbf{x}}_i'\mathbf{S}_{proc}^{-1}\,\mathbf{x} - \frac{1}{2}\overline{\mathbf{x}}_i'\mathbf{S}_{proc}^{-1}\,\overline{\mathbf{x}}_i + \ln p_i \text{ za } i = 1,2,\ldots,g \tag{4.18}$$

Posljedično, klasifikacijsko pravilo na temelju uzorka glasi:

Definicija 4.2.7 (Procijenjena minimalna ukupna vjerojatnost pogrešne klasifikacije (TPM) za normalne populacije s jednakim kovarijacijskim matricama)**.** *Pridruži* x *populaciji* π_k *ako je* $\hat{d}_k(\mathbf{x})$ *najveći od* $\hat{d}_1(\mathbf{x}), \hat{d}_2(\mathbf{x}), \ldots, \hat{d}_g(\mathbf{x})$ *gdje je* $\hat{d}_i(\mathbf{x})$ *dano kao u* (4.18)*.*

Linearne diskriminacijske bodove definirane u (4.16) možemo uspoređivati iz čega slijedi da je izraz: $d_k(\mathbf{x})$ je najveći među svim diskriminacijskim bodovima $d_1(\mathbf{x}), d_2(\mathbf{x}), \ldots, d_g(\mathbf{x})$ ekvivalentan s

$$0 \le d_k(\mathbf{x}) - d_i(\mathbf{x}) = (\boldsymbol{\mu}_k-\boldsymbol{\mu}_i)'\,\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_k-\boldsymbol{\mu}_i)'\,\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_k+\boldsymbol{\mu}_i) + \ln\left(\frac{p_k}{p_i}\right) \tag{4.19}$$

za sve $i = 1, 2, \ldots, g$. Dobivamo alternativni oblik klasifikacijskog pravila za minimizaciju ukupne vjerojatnosti za pogrešnu klasifikaciju (TPM) koje glasi:
Pridruži $\mathbf{x}$ populaciji π_k ako

$$(\boldsymbol{\mu}_k - \boldsymbol{\mu}_i)' \, \Sigma^{-1} \mathbf{x} - \frac{1}{2} (\boldsymbol{\mu}_k - \boldsymbol{\mu}_i)' \, \Sigma^{-1} (\boldsymbol{\mu}_k + \boldsymbol{\mu}_i) \geq \ln \left(\frac{p_i}{p_k} \right) \tag{4.20}$$

za sve $i = 1, 2, \ldots, g$. Uzoračka verzija klasifikacijskog pravila u (4.20) dobivenog zamjenom $\boldsymbol{\mu}_i$ s $\overline{\mathbf{x}}_\mathrm{i}$ i umetanjem zajedničke uzoračke kovarijacijske matrice S_{proc} umjesto Σ, uz uvjet da je $\sum_{i=1}^{g} (n_i - 1) \geq p$ kako bi inverz S_{proc}^{-1} postojao:
Pridruži x populaciji π_k ako

$$\hat{d}_{ki}(\mathbf{x}) = (\overline{\mathbf{x}}_k - \overline{\mathbf{x}}_i)' \, \mathbf{S}_{\mathrm{proc}}^{-1} \mathbf{x} - \frac{1}{2} (\overline{\mathbf{x}}_k - \overline{\mathbf{x}}_i)' \, \mathbf{S}_{\mathrm{proc}}^{-1} \, (\overline{\mathbf{x}}_k + \overline{\mathbf{x}}_i) \geq \ln \left(\frac{p_i}{p_k} \right) \quad \text{za sve } i \neq k. \tag{4.21}$$

Uzoračka klasifikacijska pravila i nisu najoptimalnija zbog toga što koriste procjene populacijskih parametara. Međutim, njihova se performansa može procijeniti korištenjem Lachenbruchove metode zadržavanja. Ako je $n_{iM}^{(H)}$ broj pogrešno klasificiranih zadržanih podataka u i-toj grupi, $i = 1, 2, \ldots, g$, tada se procjena očekivane razine pogreške $E(\mathrm{AER})$ može izračunati kao:

$$\hat{E}(\mathrm{AER}) = \frac{\sum_{i=1}^{g} n_{iM}^{(H)}}{\sum_{i=1}^{g} n_i}. \tag{4.22}$$

4.3 Fisherova metoda za diskriminaciju s više populacija

Fisher je također predložio generalizaciju klasifikacijskog pravila za više populacija koje je opisano u poglavlju 2.2. Njegov pristup ima nekoliko prednosti ukoliko nas zanima vizualizacija separacije. Primarna svrha ovog pristupa je diskriminacija podataka iako se može koristiti i za klasifikaciju podataka (vidi [3]). Nije nužno da pretpostavimo da su populacije normalne multivarijatne, ali pretpostavljamo da su populacijske kovarijacijske matrice jednake i punog ranga, tj $\Sigma_1 = \Sigma_2 = \cdots = \Sigma_g = \Sigma$.

Neka je $\overline{\boldsymbol{\mu}} = \frac{1}{g}\sum_{i=1}^{g} \boldsymbol{\mu}_i$ i $\mathbf{B}_{\boldsymbol{\mu}} = \sum_{i=1}^{g} (\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}})(\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}})'$. Promatramo $Y = \mathbf{a}'\mathbf{X}$ s očekivanjem

$$E(Y) = \mathbf{a}' E(\mathbf{X} \mid \pi_i) = \mathbf{a}'\boldsymbol{\mu}_i \text{ za populaciju } \pi_i$$

i varijancom

$$\mathrm{Var}(Y) = \mathbf{a}' \mathrm{Cov}(\mathbf{X})\mathbf{a} = \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a} \text{ za sve populacije.}$$

Neka je ukupno očekivanje

$$\tilde{\boldsymbol{\mu}}_Y = \frac{1}{g}\sum_{i=1}^{g} \mu_{iY} = \frac{1}{g}\sum_{i=1}^{g} \mathbf{a}'\boldsymbol{\mu}_i = \mathbf{a}'\left(\frac{1}{g}\sum_{i=1}^{g}\boldsymbol{\mu}_i\right) = \mathbf{a}'\overline{\boldsymbol{\mu}} \tag{4.23}$$

i omjer

$$\frac{\sum_{i=1}^{g}(\mu_{iY} - \bar{\mu}_Y)^2}{\sigma_Y^2} = \frac{\sum_{i=1}^{g}(\mathbf{a}'\boldsymbol{\mu}_i - \mathbf{a}'\overline{\boldsymbol{\mu}})^2}{\mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}} = \frac{\mathbf{a}'\left(\sum_{i=1}^{g}(\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}})(\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}})'\right)\mathbf{a}}{\mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}} = \frac{\mathbf{a}'\mathbf{B}_{\mu}\mathbf{a}}{\mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}} \tag{4.24}$$

koji mjeri varijabilnost između grupa vrijednosti Y u odnosu na zajedničku varijabilnost unutar grupa. Pošto su uglavnom $\boldsymbol{\Sigma}$ i $\boldsymbol{\mu}_i$ nepoznate, koristimo skup točno klasificiranih podataka veličine n_i iz populacije π_i, $i = 1, 2, \ldots, g$. Definiramo uzorački vektor očekivanja $\overline{\mathbf{x}}_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\mathbf{x}_{ij}$, kovarijacijske matrice $\mathbf{S}_i$, $i = 1, 2, \ldots, g$ i "ukupan prosjek" $\overline{\mathbf{x}} = \frac{1}{g}\sum_{i=1}^{g}\overline{\mathbf{x}}_i$ koji je $p \times 1$ vektor. Nadalje, slično kao i $\mathbf{B}_{\boldsymbol{\mu}}$ definiramo

$$\mathbf{B} = \sum_{i=1}^{g}(\overline{\mathbf{x}}_i - \overline{\mathbf{x}})(\overline{\mathbf{x}}_i - \overline{\mathbf{x}})' \tag{4.25}$$

te procjenu od Σ

$$\mathbf{W} = \sum_{i=1}^{g}(n_i - 1)\mathbf{S}_i = \sum_{i=1}^{g}\sum_{j=1}^{n_i}\left(\mathbf{x}_{ij} - \overline{\mathbf{x}}_i\right)\left(\mathbf{x}_{ij} - \overline{\mathbf{x}}_i\right)'. \tag{4.26}$$

Posljedično, $\mathbf{W}/\left(n_1 + n_2 + \cdots + n_g - g\right) = \mathbf{S}_{\text{proc}}$ je procjenitelj za $\boldsymbol{\Sigma}$.

Fisherove uzoračke linearne diskriminante

Definicija 4.3.1. *Označimo s* $\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_s > 0$ *svojstvene vrijednosti matrice* $\mathbf{W}^{-1}\mathbf{B}$,*a s* $\hat{\mathbf{e}}_1, \ldots, \hat{\mathbf{e}}_s$ *odgovarajuće svojstvene vektore skalirane tako da vrijedi* $\hat{\mathbf{e}}'\mathbf{S}_{proc}\,\hat{\mathbf{e}} = 1$. *Tada je vektor koeficijenta* $\hat{\mathbf{a}}$ *koji maksimizira*

$$\frac{\hat{\mathbf{a}}'\mathbf{B}\hat{\mathbf{a}}}{\hat{\mathbf{a}}'\mathbf{W}\hat{\mathbf{a}}} = \frac{\hat{\mathbf{a}}'\left(\sum_{i=1}^{g}(\overline{\mathbf{x}}_i - \overline{\mathbf{x}})(\overline{\mathbf{x}}_i - \overline{\mathbf{x}})'\right)\hat{\mathbf{a}}}{\hat{\mathbf{a}}'\left[\sum_{i=1}^{g}\sum_{j=1}^{n_i}\left(\mathbf{x}_{ij} - \overline{\mathbf{x}}_i\right)\left(\mathbf{x}_{ij} - \overline{\mathbf{x}}_i\right)'\right]\hat{\mathbf{a}}} \tag{4.27}$$

dan s $\hat{\mathbf{a}}_1 = \hat{\mathbf{e}}_1$.

Linearna kombinacija $\hat{\mathbf{a}}_1'\mathbf{x}$ se naziva prva uzoračka diskriminanta. Drugu uzoračku diskriminantu dobijemo ako odaberemo $\hat{\mathbf{a}}_2 = \hat{\mathbf{e}}_2$,a k-tu uzoračku diskriminantu ($k \leq s$) dobijemo kao $\hat{\mathbf{a}}_k'\mathbf{x} = \hat{\mathbf{e}}_k'\mathbf{x}$.

Klasifikacija pomoću Fisherovih diskriminanti

Neka je $Y_k = \mathbf{a}_k'\mathbf{X} = k$ -ta diskriminanta, $k \leq s$. Tada vrijedi da $\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_s \end{bmatrix}$ ima vektor očekivanja $\boldsymbol{\mu}_{iY} = \begin{bmatrix} \mu_{iY_1} \\ \vdots \\ \mu_{iY_s} \end{bmatrix} = \begin{bmatrix} \mathbf{a}_1'\boldsymbol{\mu}_i \\ \vdots \\ \mathbf{a}_s'\boldsymbol{\mu}_i \end{bmatrix}$ za populaciju π_i i kovarijacijsku matricu $\mathbf{I}$, za sve populacije. Zato što komponente $\mathbf{Y}$ imaju jedinične varijance s kovarijancama jednakim nula, prikladna mjera kvadratne udaljenosti od $\mathbf{Y} = \mathbf{y}$ do $\boldsymbol{\mu}_{iY}$ je

$$(\mathbf{y} - \boldsymbol{\mu}_{iY})'(\mathbf{y} - \boldsymbol{\mu}_{iY}) = \sum_{j=1}^{s}\left(y_j - \mu_{iY_j}\right)^2. \tag{4.28}$$

Prije nego navedemo Fisherovo klasifikacijsko pravilo, komentirajmo restrikcije na broj diskriminanti. Neka je s = broj diskriminanti = broj nenul svojstvenih vrijednosti od $\Sigma^{-1/2}\mathbf{B}_\mu\Sigma^{-1/2}$. Dimenzija matrice $\Sigma^{-1/2}\mathbf{B}_\mu\Sigma^{-1/2}$ je $p \times p$ pa je nužno $s \leq p$. Nadalje, vrijedi da vektori

$$\boldsymbol{\mu}_1 - \overline{\boldsymbol{\mu}}, \boldsymbol{\mu}_2 - \overline{\boldsymbol{\mu}}, \ldots, \boldsymbol{\mu}_g - \overline{\boldsymbol{\mu}} \tag{4.29}$$

zadovoljavaju $(\boldsymbol{\mu}_1 - \overline{\boldsymbol{\mu}}) + (\boldsymbol{\mu}_2 - \overline{\boldsymbol{\mu}}) + \cdots + \left(\boldsymbol{\mu}_g - \overline{\boldsymbol{\mu}}\right) = g\overline{\boldsymbol{\mu}} - g\overline{\boldsymbol{\mu}} = \mathbf{0}$. Slijedi

$$\mathbf{B}_\mu\mathbf{e} = \sum_{i=1}^{g}(\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}})(\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}})'\,\mathbf{e} = \sum_{i=1}^{g}(\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}})\,0 = \mathbf{0} \tag{4.30}$$

pa je $\mathbf{\Sigma}^{-1}\mathbf{B}_{\mu}\mathbf{e} = 0\mathbf{e}$. Dakle, postoji $p - q$ ortogonalnih svojstvenih vektora koji odgovaraju svojstvenoj vrijednosti nula. Broj nenul svojstvenih vrijednosti s zadovoljava $s \leq \min(p, g-1)$.

Broj varijabli	Broj populacija	Maksimalan broj diskriminanti
za bilo koje p	$g = 2$	1
za bilo koje p	$g = 3$	2
$p = 2$	za bilo koje g	2

Klasifikacijsko pravilo bi trebalo pridruživati $\mathbf{y}$ populaciji π_k ukoliko je kvadratna udaljenost od $\mathbf{y}$ do $\boldsymbol{\mu}_{kY}$ manja nego kvadratna udaljenost od $\mathbf{y}$ do $\boldsymbol{\mu}_{iY}$ za $i \neq k$.

Definicija 4.3.2. *Ako koristimo r diskriminanti za klasifikaciju novog podatka, tada vrijedi: Pridruži* x *populaciji* π_k *ako*

$$\sum_{j=1}^{r}\left(y_j - \mu_{kY_j}\right)^2 = \sum_{j=1}^{r}\left[\mathbf{a}_j'\left(\mathbf{x} - \boldsymbol{\mu}_k\right)\right]^2 \leq \sum_{j=1}^{r}\left[\mathbf{a}_j'\left(\mathbf{x} - \boldsymbol{\mu}_i\right)\right]^2 \quad \text{za sve } i \neq k \tag{4.31}$$

Sljedeći teorem komentira odnos između klasifikacijskog pravila (4.31) i pravila (4.16).

Teorem 4.3.3. *Neka je* $y_j = \mathbf{a}_j'\mathbf{x}$*, gdje je* $\mathbf{a}_j = \mathbf{\Sigma}^{-1/2}\mathbf{e}_j$ *i* $\mathbf{e}_j$ *svojstveni vektor od* $\mathbf{\Sigma}^{-1/2}\mathbf{B}_{\mu}\mathbf{\Sigma}^{-1/2}$*. Tada*

$$\begin{aligned}\sum_{j=1}^{p}\left(y_j - \mu_{iY_j}\right)^2 = \sum_{j=1}^{p}\left[\mathbf{a}_j'\left(\mathbf{x} - \boldsymbol{\mu}_i\right)\right]^2 &= \left(\mathbf{x} - \boldsymbol{\mu}_i\right)'\mathbf{\Sigma}^{-1}\left(\mathbf{x} - \boldsymbol{\mu}_i\right) \\ &= -2d_i(\mathbf{x}) + \mathbf{x}'\mathbf{\Sigma}^{-1}\mathbf{x} + 2\ln p_i.\end{aligned} \tag{4.32}$$

Ako $\lambda_1 \geq \cdots \geq \lambda_s > 0 = \lambda_{s+1} = \cdots = \lambda_p$*, tada* $\sum_{j=s+1}^{p}\left(y_j - \mu_{iY_j}\right)^2$ *je konstanta za sve populacije* $i = 1, 2, \ldots, g$ *pa samo prvih s diskriminanti* y_j *ili* $\sum_{j=1}^{s}\left(y_j - \mu_{iY_j}\right)^2$ *pridonosi klasifikaciji.*
Također, ukoliko su apriorne vjerojatnosti takve da $p_1 = p_2 = \cdots = p_g = 1/g$*, tada je pravilo* (4.31) *s* $r = s$ *ekvivalentno populacijskoj verziji minimalnog TPM pravila* (4.2.7)*.*

Dokaz. Kvadratna udaljenost

$$\begin{aligned}\left(\mathbf{x} - \boldsymbol{\mu}_i\right)'\mathbf{\Sigma}^{-1}\left(\mathbf{x} - \boldsymbol{\mu}_i\right) &= \left(\mathbf{x} - \boldsymbol{\mu}_i\right)'\mathbf{\Sigma}^{-1/2}\mathbf{\Sigma}^{-1/2}\left(\mathbf{x} - \boldsymbol{\mu}_i\right) \\ &= \left(\mathbf{x} - \boldsymbol{\mu}_i\right)'\Sigma^{-1/2}\mathbf{E}\mathbf{E}'\Sigma^{-1/2}\left(\mathbf{x} - \boldsymbol{\mu}_i\right),\end{aligned} \tag{4.33}$$

gdje $\mathbf{E} = \left[\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_p\right]$ je ortogonalna matrica čiji su stupci svojstveni vektori od $\boldsymbol{\Sigma}^{-1/2}\mathbf{B}_\mu\boldsymbol{\Sigma}^{-1/2}$. Budući da je $\boldsymbol{\Sigma}^{-1/2}\mathbf{e}_i = \mathbf{a}_i$ ili $\mathbf{a}_i' = \mathbf{e}_i'\boldsymbol{\Sigma}^{-1/2}$ vrijedi

$$\mathbf{E}'\boldsymbol{\Sigma}^{-1/2}\left(\mathbf{x} - \boldsymbol{\mu}_i\right) = \begin{bmatrix} \mathbf{a}_1'\left(\mathbf{x} - \boldsymbol{\mu}_i\right) \\ \mathbf{a}_2'\left(\mathbf{x} - \boldsymbol{\mu}_i\right) \\ \vdots \\ \mathbf{a}_p'\left(\mathbf{x} - \boldsymbol{\mu}_i\right) \end{bmatrix} \tag{4.34}$$

i

$$\left(\mathbf{x} - \boldsymbol{\mu}_i\right)'\boldsymbol{\Sigma}^{-1/2}\mathbf{E}\mathbf{E}'\boldsymbol{\Sigma}^{-1/2}\left(\mathbf{x} - \boldsymbol{\mu}_i\right) = \sum_{j=1}^{p}\left[\mathbf{a}_j'\left(\mathbf{x} - \boldsymbol{\mu}_i\right)\right]^2. \tag{4.35}$$

Nadalje, svaki $\mathbf{a}_j = \boldsymbol{\Sigma}^{-1/2}\mathbf{e}_j$, $j > s$ je svojstveni vektor od $\boldsymbol{\Sigma}^{-1}\mathbf{B}_\mu$ sa svojstvenom vrijednosti nula. Kao što je objašnjeno u (4.29), a_j je okomit na svaki $\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}}$ i stoga $(\boldsymbol{\mu}_k - \overline{\boldsymbol{\mu}}) - (\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}}) = \boldsymbol{\mu}_k - \boldsymbol{\mu}_i$ za $i, k = 1, 2, \ldots, g$. Uvjet $0 = \mathbf{a}_j'\left(\boldsymbol{\mu}_k - \boldsymbol{\mu}_i\right) = \mu_{kY} - \mu_{iY_j}$ implicira da $y_j - \mu_{kY} = y_j - \mu_{iY_j}$ pa $\sum_{j=s+1}^{p}\left(y_j - \mu_{iY_j}\right)^2$ je konstanta za sve $i = 1, 2, \ldots, g$. Stoga, samo prvih s diskriminanti y_j sudjeluje u klasifikaciji. □

Definicija 4.3.4 (Fisherovo klasifikacijsko pravilo temeljeno na prvih $r \leq s$ uzoračkih diskriminanti)**.** *Pridruži* x *populaciji* π_k *ako*

$$\sum_{j=1}^{r}\left(\hat{y}_j - \bar{y}_{kj}\right)^2 = \sum_{j=1}^{r}\left[\hat{\mathbf{a}}_j'\left(\mathbf{x} - \overline{\mathbf{x}}_k\right)\right]^2 \leq \sum_{j=1}^{r}\left[\hat{\mathbf{a}}_j'\left(\mathbf{x} - \overline{\mathbf{x}}_i\right)\right]^2 \quad \textit{za sve } i \neq k \tag{4.36}$$

gdje je $\hat{\mathbf{a}}_j$ *definirano kao u* (4.27)*,* $\bar{y}_{kj} = \hat{\mathbf{a}}_j'\overline{\mathbf{x}}_k$ *i* $r \leq s$*. Kada su apriorne vjerojatnosti takve da* $p_1 = p_2 = \cdots = p_g = 1/g$ *i* $r = s$*, klasifikacijsko pravilo* (4.36) *je ekvivalentno pravilu* (4.2.7) *koje je temeljeno na najvećem linearnom diskriminacijskom bodu.*

Primjer 4.3.5. *Klasificirajmo novi podatak koristeći Fisherovo klasifikacijsko pravilo definirano u* (4.36)*. Neka su slučajni uzorci za populacije* π_1, π_2 *i* π_3 *zajedno s uzoračkim očekivanjima i kovarijacijskim matricama dani kao:*

$$\pi_1: \quad \mathbf{X}_1 = \begin{bmatrix} -2 & 5 \\ 0 & 3 \\ -1 & 1 \end{bmatrix}, \quad n_1 = 3, \quad \overline{\mathbf{x}}_1 = \begin{bmatrix} -1 \\ 3 \end{bmatrix}, \quad \mathbf{S}_1 = \begin{bmatrix} 1 & -1 \\ -1 & 4 \end{bmatrix}$$

$$\pi_2: \quad \mathbf{X}_2 = \begin{bmatrix} 0 & 6 \\ 2 & 4 \\ 1 & 2 \end{bmatrix}, \quad n_2 = 3, \quad \overline{\mathbf{x}}_2 = \begin{bmatrix} 1 \\ 4 \end{bmatrix}, \quad \mathbf{S}_2 = \begin{bmatrix} 1 & -1 \\ -1 & 4 \end{bmatrix}$$

$$\pi_3: \quad \mathbf{X}_3 = \begin{bmatrix} 1 & -2 \\ 0 & 0 \\ -1 & -4 \end{bmatrix}, \quad n_3 = 3, \quad \overline{\mathbf{x}}_3 = \begin{bmatrix} 0 \\ -2 \end{bmatrix}, \quad \mathbf{S}_3 = \begin{bmatrix} 1 & 1 \\ 1 & 4 \end{bmatrix}$$

Neka su $p_1 = p_2 = 0.25$ *i* $p_3 = 0.50$. *Odredimo prvo Fisherove diskriminante uz pretpostavku da populacije imaju zajedničku kovarijacijsku matricu* $\mathbf{\Sigma}$. *Računamo:*

$$\overline{\mathbf{x}} = \begin{bmatrix} 0 \\ 5 \\ 3 \end{bmatrix}$$

$$\mathbf{B} = \sum_{i=1}^{3} (\overline{\mathbf{x}}_i - \overline{\mathbf{x}})(\overline{\mathbf{x}}_i - \overline{\mathbf{x}})' = \begin{bmatrix} 2 & 1 \\ 1 & 62/3 \end{bmatrix}$$

$$\mathbf{W} = \sum_{i=1}^{3} \sum_{j=1}^{n_i} \left(\mathbf{x}_{ij} - \overline{\mathbf{x}}_i\right)\left(\mathbf{x}_{ij} - \overline{\mathbf{x}}_i\right)' = (n_1 + n_2 + n_3 - 3)\,\mathbf{S}_{proc} = \begin{bmatrix} 6 & -2 \\ -2 & 24 \end{bmatrix}$$

$$\mathbf{W}^{-1} = \frac{1}{140}\begin{bmatrix} 24 & 2 \\ 2 & 6 \end{bmatrix}$$

$$\mathbf{W}^{-1}\mathbf{B} = \begin{bmatrix} 0.3571 & 0.4667 \\ 0.0714 & 0.9000 \end{bmatrix}$$

Nadalje, računamo

$$\left|\mathbf{W}^{-1}\mathbf{B} - \lambda\mathbf{I}\right| = \left\|\begin{bmatrix} 0.3571 - \lambda & 0.4667 \\ 0.0714 & 0.9000 - \lambda \end{bmatrix}\right\| = 0$$

iz čega slijedi

$$(0.3571 - \lambda)(0.9000 - \lambda) - (0.4667)(0.0714) = \lambda^2 - 1.2571\lambda + 0.2881 = 0$$

Koristeći formulu za rješenje kvadratne jednadžbe dobivamo $\hat{\lambda}_1 = 0.9556$ *i* $\hat{\lambda}_2 = 0.3015$. *Rješavajući* $\left(\mathbf{W}^{-1}\mathbf{B} - \hat{\lambda}_i\mathbf{I}\right)\hat{\mathbf{a}}_i = \mathbf{0}$ *za* $i = 1, 2$ *i skalirajući rezultate tako da* $\hat{\mathbf{a}}_i'\mathbf{S}_{proc}\,\hat{\mathbf{a}}_i = 1$ *dobivamo normalizirane svojstvene vektore:*

$$\hat{\mathbf{a}}_1' = \begin{bmatrix} 0.386 & 0.495 \end{bmatrix} \text{ i } \hat{\mathbf{a}}_2' = \begin{bmatrix} 0.938 & -0.112 \end{bmatrix}$$

Fisherove diskriminante su

$$\hat{y}_1 = \hat{\mathbf{a}}_1'\mathbf{x} = \begin{bmatrix} 0.386 & 0.495 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0.386x_1 + 0.495x_2$$

$$\hat{y}_2 = \hat{\mathbf{a}}_2'\mathbf{x} = \begin{bmatrix} 0.938 & -0.112 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0.938x_1 - 0.112x_2$$

Klasificirajmo novi podatak $\mathbf{x}_0' = [x_{01}, x_{02}] = \begin{bmatrix} 1 & 3 \end{bmatrix}$. *Uvrstimo novi podatak* $\mathbf{x}_0'$ *u Fisherove diskriminante*

$$\hat{y}_1 = 0.386x_{01} + 0.495x_{02} = 0.386(1) + 0.495(3) = 1.87$$
$$\hat{y}_2 = 0.938x_{01} - 0.112x_{02} = 0.938(1) - 0.112(3) = 0.60$$

Također vrijedi $\bar{y}_{kj} = \hat{\mathbf{a}}_j' \bar{\mathbf{x}}_k$ *iz čega slijedi*

$$\bar{y}_{11} = \hat{\mathbf{a}}_1' \bar{\mathbf{x}}_1 = \begin{bmatrix} .386 & .495 \end{bmatrix} \begin{bmatrix} -1 \\ 3 \end{bmatrix} = 1.10$$

$$\bar{y}_{12} = \hat{\mathbf{a}}_2' \bar{\mathbf{x}}_1 = \begin{bmatrix} .938 & -.112 \end{bmatrix} \begin{bmatrix} -1 \\ 3 \end{bmatrix} = -1.27$$

Slično,

$$\bar{y}_{21} = \hat{\mathbf{a}}_1' \bar{\mathbf{x}}_2 = 2.37$$
$$\bar{y}_{22} = \hat{\mathbf{a}}_2' \bar{\mathbf{x}}_2 = 0.49$$
$$\bar{y}_{31} = \hat{\mathbf{a}}_1' \bar{\mathbf{x}}_3 = -0.99$$
$$\bar{y}_{32} = \hat{\mathbf{a}}_2' \bar{\mathbf{x}}_3 = 0.22$$

Konačno, tražimo k tako da je $\sum_{j=1}^{2} \left(\hat{y}_j - \bar{y}_{kj}\right)^2 = \sum_{j=1}^{2} \left[\hat{\mathbf{a}}_j \left(\mathbf{x} - \bar{\mathbf{x}}_\mathbf{k}\right)\right]^2$ *najmanji za* $k = 1, 2, 3$. *Računamo*

$$\sum_{j=1}^{2} \left(\hat{y}_j - \bar{y}_{1j}\right)^2 = (1.87 - 1.10)^2 + (0.60 + 1.27)^2 = 4.09$$

$$\sum_{j=1}^{2} \left(\hat{y}_j - \bar{y}_{2j}\right)^2 = (1.87 - 2.37)^2 + (0.60 - 0.49)^2 = 0.26$$

$$\sum_{j=1}^{2} \left(\hat{y}_j - \bar{y}_{3j}\right)^2 = (1.87 + .99)^2 + (0.60 - 0.22)^2 = 8.32$$

Slijedi da se minimum postiže za $k = 2$, *pa novi podatak* $\mathbf{x}_0$ *alociramo u populaciju* π_2.

Poglavlje 5

Logistička regresija

Do sada smo razmatrali klasifikacijske probleme temeljene na kvantitativnim varijablama, a u ovom poglavlju ćemo komentirati pristup klasifikaciji u kojem su neke ili sve varijable kvalitativne - logističkoj regresiji. U najjednostavnijem slučaju, varijabla odaziva je ograničena na dvije vrijednosti (binarna varijabla) pa je u svakom slučaju možemo kodirati kao 0 i 1. U tom slučaju vjerojatnost p od 1 je parametar koji nas zanima. Ta vjerojatnost predstavlja udio u populaciji koji je kodiran kao 1. Udio nula je $1 - p$ što ponekad označavamo s q. Srednju vrijednost računamo kao $0 \times (1 - p) + 1 \times p = p$, a varijancu kao $0^2 \times (1 - p) + 1^2 \times p - p^2 = p(1 - p)$.

Neka je Y varijabla odaziva koja poprima vrijednosti 0 ili 1. Modelirajmo vjerojatnost da varijabla Y poprimi vrijednost 1 pomoću jednog linearnog modela kao

$$p = E(Y|z) = \beta_0 + \beta_1 z \tag{5.1}$$

dodajući grešku ϵ na kraju. Problem koji se javlja je to što bi predviđene vrijednosti mogle biti veće od 1 i manje od 0. Drugi problem je što regresijska analiza pretpostavlja da je varijanca od Y konstanta što ovdje nije slučaj.

Logit model

Umjesto da modeliramo vjerojatnost p koristeći linearni model, promotrit ćemo prvo omjer šansi (engl. "odds ratio")

$$\text{odds ratio} = \frac{p}{1 - p}, \tag{5.2}$$

što je zapravo omjer vjerojatnosti da varijabla odaziva poprimi vrijednost 1 i vjerojatnosti da poprimi vrijednost 0. Za razliku od vjerojatnosti, omjer šansi može poprimiti i vrijednosti veće od 1. Primjenom prirodnog logaritma dozvoljavamo i poprimanje negativnih

vrijednosti. Kada su ishodi 0 ili 1 podjednako vjerojatni, vrijednost logaritma omjera šansi iznosi nula.

U logističkoj regresiji modeliramo prirodnim logaritmom omjer šansi, što se naziva logit model:

$$\text{logit}(p) = \ln(\text{ odds }) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 z \tag{5.3}$$

Eksponenciranjem dobivamo

$$\theta(z) = \frac{p(z)}{1-p(z)} = \exp(\beta_0 + \beta_1 z) \tag{5.4}$$

iz čega slijedi

$$p(z) = \frac{\exp(\beta_0 + \beta_1 z)}{1 + \exp(\beta_0 + \beta_1 z)} = \frac{1}{1 + \exp(-\beta_0 - \beta_1 z)} \tag{5.5}$$

Parametar β_0 daje vrijednost $\exp(\beta_0) / (1 + \exp(\beta_0))$ za p u $z = 0$, dok parametar β_1 određuje koliko se brzo mijenja vjerojatnost p sa z iako je njegova interpretacija malo kompliciranija nego kod linearne regresije.

5.1 Analiza logističke regresije

Promatramo logistički model s više prediktorskih varijabli. Neka su $\left(z_{j1}, z_{j2}, \ldots, z_{jr}\right)$ vrijednosti od r prediktora za j-to opažanje. Označimo $\mathbf{z}_j = \left[1, z_{j1}, z_{j2}, \ldots, z_{jr}\right]'$ i pretpostavimo da je Y_j Bernoullijeva varijabla s parametrom uspjeha $p\left(\mathbf{z_j}\right)$. Tada vrijedi

$$P\left(Y_j = y_j\right) = p^{y_j}\left(\mathbf{z}_j\right)\left(1 - p\left(\mathbf{z}_j\right)\right)^{1-y_j} \qquad \text{za } y_j = 0, 1 \tag{5.6}$$

tako da

$$E\left(Y_j\right) = p\left(\mathbf{z}_j\right) \quad \text{ i } \quad \text{Var}\left(Y_j\right) = p\left(\mathbf{z}_j\right)\left(1 - p\left(\mathbf{z}_j\right)\right). \tag{5.7}$$

Prirodni logaritam omjera šansi prati linearni model, tj dobivamo model

$$\ln\left(\frac{p(\mathbf{z})}{1-p(\mathbf{z})}\right) = \beta_0 + \beta_1 z_1 + \cdots + \beta_r z_r = \beta' \mathbf{z}_j \tag{5.8}$$

gdje je $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_r]'$.

Metoda maksimalne vjerodostojnosti

Parametri β u logističkoj regresiji se procijenjuju metodom maksimalne vjerodostojnosti (engl. "Maximum Likelihood Estimation", MLE). Funkcija vjerodostojnosti (engl. "likelihood function") L dana je kao

$$L(b_0, b_1, \ldots, b_r) = \prod_{j=1}^{n} p^{y_j}\left(\mathbf{z}_j\right)\left(1 - p\left(\mathbf{z}_j\right)\right)^{1-y_j} = \frac{\prod_{j=1}^{n} e^{y_j\left(b_0 + b_1 z_{j1} + \ldots + b_r z_r\right)}}{\prod_{j=1}^{n}\left(1 + e^{b_0 + b_1 z_{j1} + \ldots + b_r z_r}\right)} \tag{5.9}$$

Ideja metode je da se procijene vrijednosti parametara koji maksimiziraju funkciju vjerodostojnosti L. Međutim, vrijednosti tih parametara nije moguće izraziti u zatvorenoj formi pa se moraju odrediti numerički, počevši od inicijalnog pogađanja i zatim iteriranjem do maksimuma funkcije vjerodostojnosti L. Opisana metoda se naziva IRLS (engl. "iteratively re-weighted least squares") (vidi [3]). Procijenjene vrijednosti parametra označimo s vektorom $\hat{\boldsymbol{\beta}}$.

Pouzdani intervali za parametre

Ako je uzorak dovoljno velik, procjenitelj $\hat{\boldsymbol{\beta}}$ je asimptotski normalan s očekivanjem $\boldsymbol{\beta}$ i procijenjenom kovarijacijskom matricom

$$\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}) \approx \left[\sum_{j=1}^{n} \hat{p}\left(\mathbf{z}_j\right)\left(1 - \hat{p}\left(\mathbf{z}_j\right)\right)\mathbf{z}_j\mathbf{z}_j'\right]^{-1} \tag{5.10}$$

95% pouzdani interval za β_k je dan kao

$$\hat{\beta}_k \pm 1.96 S E\left(\hat{\beta}_k\right) \quad k = 0, 1, \ldots, r \tag{5.11}$$

Test omjera vjerojatnosti (engl. "Likelihood Ratio Test")

Za model s r prediktorskih varijabli, maksimalna vjerodostojnost dana je kao

$$L_{\max} = L\left(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_r\right) \tag{5.12}$$

Neka je nulta hipoteza $H_0 : \beta_k = 0$, tada je maksimalna vjerodostojnost za reducirani model dana kao

$$L_{\max,\ \text{reducirani}} = L\left(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_{k-1}, \hat{\beta}_{k+1}, \ldots, \hat{\beta}_r\right) \tag{5.13}$$

Nultu hipotezu testiramo koristeći statistiku

$$D = -2 \ln\left(\frac{L_{\max,\ \text{reducirani}}}{L_{\max}}\right), \tag{5.14}$$

koja se u ovom kontekstu naziva devijanca i u uvjetima nulte hipoteze prati χ^2 distribuciju s jednim stupnjem slobode. H_0 odbacujemo za velike vrijednosti devijance D. Alternativni način za testiranje značajnosti parametra u modelu je Wald test koji za testiranje nulte hipoteze $H_0 : \beta_k = 0$ koristi testnu statistiku $Z = \hat{\beta}_k / SE\left(\hat{\beta}_k\right)$ koja prati standardnu normalnu distribuciju ili njezinu χ^2 verziju Z^2 s jednim stupnjem slobode. Općenito, ako nulta hipoteza tvrdi da je m parametara istovremeno jednako nula, tada devijanca prati χ^2 distribuciju s m stupnjeva slobode.

5.2 Klasifikacija

Neka je varijabla odaziva Y jednaka 1 ukoliko promatrani podatak pripada populaciji 1 i neka je 0 ukoliko pripada populaciji 2. Jednom kad je logistička funkcija određena, korištenjem skupa podataka za treniranje možemo početi s klasifikacijom. Klasifikacijsko pravilo sada glasi:
Pridruži $\mathbf{z}$ populaciji 1 ukoliko je procijenjeni omjer šansi veći od 1, tj

$$\frac{\hat{p}(\mathbf{z})}{1-\hat{p}(\mathbf{z})} = \exp\left(\hat{\beta}_0 + \hat{\beta}_1 z_1 + \cdots + \hat{\beta}_r z_r\right) > 1. \tag{5.15}$$

Ekvivalentno, imamo i jednostavno linearno diskriminacijsko pravilo:
Pridruži $\mathbf{z}$ populaciji 1 ukoliko je linearna diskriminanta veća od 0, tj

$$\ln \frac{\hat{p}(\mathbf{z})}{1-\hat{p}(\mathbf{z})} = \hat{\beta}_0 + \hat{\beta}_1 z_1 + \cdots + \hat{\beta}_r z_r > 0. \tag{5.16}$$

5.3 Logistička regresija s binomnom odazivom

Sada promatramo općenitiji slučaj u kojem imamo nekoliko ponavljanja s istim vrijednostima nezavisnih varijabli $\mathbf{z}_j$ i neka postoji m različitih skupova u kojima su prediktorske varijable konstante. Kada se izvrši n_j nezavisnih pokušaja s prediktorskim varijablama z_j, odziv Y_j se modelira kao binomna distribucija s vjerojatnošću $p\left(\mathbf{z}_j\right) = P\left(\text{Uspjeh} \mid \mathbf{z}_j\right)$. Zbog pretpostavke o nezavisnosti Y_j, funkcija vjerodostojnosti glasi :

$$L\left(\beta_0, \beta_1, \ldots, \beta_r\right) = \prod_{j=1}^{m} \left(\begin{array}{c} n_j \\ y_j \end{array} \right) p^{y_j}\left(\mathbf{z}_j\right)\left(1 - p\left(\mathbf{z}_j\right)\right)^{n_j - y_j} \tag{5.17}$$

gdje vjerojatnost $p\left(\mathbf{z}_j\right)$ prati logit model (5.8). Kada je uzorak dovoljno velik, procijenjena kovarijacijska matrica je

$$\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}) \approx \left[\sum_{j=1}^{m} n_j \hat{p}\left(\mathbf{z}_j\right)\left(1 - \hat{p}\left(\mathbf{z}_j\right)\right)\mathbf{z}_j \mathbf{z}_j'\right]^{-1} \tag{5.18}$$

i i-ti dijagonalni element je procjena varijance za $\hat{\beta}_{i+1}$. Procjena varijance vjerojatnosti $\hat{p}\left(\mathbf{z}_j\right)$ je dana kao

$$\widehat{\text{Var}}\left(\hat{p}\left(\mathbf{z}_k\right)\right) \approx \left(\hat{p}\left(\mathbf{z}_k\right)\left(1-\hat{p}\left(\mathbf{z}_k\right)\right)^2 \mathbf{z}_j' \left[\sum_{j=1}^{m} n_j \hat{p}\left(\mathbf{z}_j\right)\left(1-\hat{p}\left(\mathbf{z}_j\right)\right)\mathbf{z}_j\mathbf{z}_j'\right]^{-1} \mathbf{z}_k \right. \tag{5.19}$$

Provjera modela

Jednom kada odredimo model, važno je ispitati koliko je dobro model prilagođen podacima, postoje li sistematična odstupanja od prilagođenog logističkog modela ili neki podaci koji odskaču (outlieri) ili značajno utječu na zaključke u analizi kada ih se izbaci ili ubaci u model. Ukoliko ne znamo parametre potrebne za određivanje vjerojatnosti $p\left(\mathbf{z}_j\right) = P(\text{Uspjeh} \mid \mathbf{z}_j)$, spomenute vjerojatnosti ćemo procijeniti pomoću opaženih vrijednosti y_i (broja uspjeha) u n_i pokušaja. Neparametarska vjerodostojnost za j-ti slučaj sada izgleda

$$\left(\begin{array}{c} n_j \\ y_j \end{array}\right) p^{y_j}\left(\mathbf{z}_j\right)\left(1-p\left(\mathbf{z}_j\right)\right)^{n_j-y_j} \tag{5.20}$$

koja se maksimizira biranjem $\hat{p}\left(\mathbf{z}_j\right) = y_j/n_j$ za $j = 1, 2, \ldots, n$. Sada je $m = \mathbf{\Sigma} n_j$. Neparametarska funkcija vjerodostojnosti dana je kao

$$-2\ln L_{\max,NP} = -2\sum_{j=1}^{m}\left[y_j \ln\left(\frac{y_j}{n_j}\right) + \left(n_j - y_j\right)\ln\left(1-\frac{y_j}{n_j}\right)\right] + 2\ln\left(\prod_{j=1}^{m}\left(\begin{array}{c} n_j \\ y_j \end{array}\right)\right) \tag{5.21}$$

Definirajmo i devijancu između neparametarskog modela i modela koji ima konstantu i $r-1$ prediktora prilagođenog podacima kao

$$G^2 = 2\sum_{j=1}^{m}\left[y_j \ln\left(\frac{y_j}{\hat{y}_j}\right) + \left(n_j - y_j\right)\ln\left(\frac{n_j - y_j}{n_j - \hat{y}_j}\right)\right] \tag{5.22}$$

gdje je $\hat{y}_j = n_j\hat{p}\left(\mathbf{z}_j\right)$ broj uspjeha dobiven regresijom. Za velike uzorke, G^2 ima približno χ^2 distribuciju s df stupnjeva slobode, gdje je df broj podataka m, umanjen za broj procijenjenih parametara β. Primijetite da devijanca za puni model, $G^2_{\text{puni,}}$, i devijanca za reducirani model, $G^2_{\text{reducirani}}$, govore o doprinosu dodatnih prediktora

$$G^2_{\text{reducirani}} - G^2_{puni} = -2\ln\left(\frac{L_{\max,\text{ reducirani}}}{L_{\max}}\right). \tag{5.23}$$

Ova razlika je približno distribuirana kao χ^2 s stupnjevima slobode $df = df_{\text{reducirani}} - df_{\text{puni}}$. Velika vrijednost razlike govori da je ipak potreban puni model. Kada je m velik, postoji previše vjerojatnosti koje je potrebno procijeniti prema neparametarskom modelu,a i χ^2 aproksimacija ne može se dokazati postojećim metodama pa se bolje osloniti na testove omjera vjerodostojnosti i logističkih modela koji ne uključuju nekoliko prediktora.

Reziduali i Goodness-of-Fit testovi

Reziduali su važni jer omogućavaju procjenu razlike između stvarnih vrijednosti promatrane varijable i vrijednosti predviđene modelom. Ako model dobro opisuje podatke, tada će reziduali biti slučajno raspoređeni oko nule i neće se uočiti nikakvi uzorci u njihovoj distribuciji. Međutim, ako postoje pravilnosti u rezidualima, to ukazuje na to da model nije dovoljno dobar u objašnjavanju varijabilnosti u podacima ili da postoji neki drugi nedostatak u modelu. U tom slučaju, potrebno je ponovno razmotriti model i provjeriti je li moguće poboljšati njegovu preciznost ili napraviti izmjene u varijablama koje se koriste kao prediktori. Spomenut ćemo tri definicije reziduala.

Definicija 5.3.1. *1. Reziduali devijance* $\left(d_j\right)$ *se definiraju kao:*

$$d_j = \pm \sqrt{2\left[y_j \ln\left(\frac{y_j}{n_j \hat{p}\left(\mathbf{z}_j\right)}\right) + \left(n_j - y_j\right) \ln\left(\frac{n_j - y_j}{n_j\left(1 - \hat{p}\left(\mathbf{z}_j\right)\right)}\right)\right]} \tag{5.24}$$

gdje je predznak od d_j *jednak kao kod* $y_j - n_j \hat{p}\left(\mathbf{z}_j\right)$ *i*

$$\begin{aligned} &\text{ako } y_j = 0, \text{onda } d_j = \sqrt{2n_j \left|\ln\left(1 - \hat{p}\left(\mathbf{z}_j\right)\right)\right|} \\ &\text{ako } y_j = n_j, \text{onda } d_j = -\sqrt{2n_j \mid \ln \hat{p}\left(\mathbf{z}_j\right) \mid} \end{aligned} \tag{5.25}$$

2. Pearsonovi reziduali $\left(r_j\right)$ *definiraju se kao:*

$$r_j = \frac{y_j - n_j \hat{p}\left(\mathbf{z}_j\right)}{\sqrt{n_j \hat{p}\left(\mathbf{z}_j\right)\left(1 - \hat{p}\left(\mathbf{z}_j\right)\right)}} \tag{5.26}$$

3. Standardizirani Pearsonovi reziduali $\left(r_{sj}\right)$ *definiraju se kao:*

$$r_{sj} = \frac{r_j}{\sqrt{1 - h_{jj}}} \tag{5.27}$$

gdje h_{jj} *je* (j, j)*-ti element u matrici* $\mathbf{H}$ *definiranoj kao u* (5.29).

Test dobrog prilagođenja (engl."Goodness of fit test") mjeri razliku između promatranog i očekivanog podatka prema nultoj hipotezi. Nulta hipoteza pretpostavlja da promatrani podatak slijedi teorijsku distribuciju. χ^2- test dobrog prilagođenja je često korišten, posebno za male uzorke:

$$X^2 = \sum_{j=1}^{m} r_j^2 = \sum_{j=1}^{n} \frac{\left(y_j - n_j \hat{p}\left(\mathbf{z}_j\right)\right)^2}{n_j \hat{p}\left(\mathbf{z}_j\right)\left(1 - \hat{p}\left(\mathbf{z}_j\right)\right)} \tag{5.28}$$

Primijetimo da je χ^2 statistika (5.28) zapravo zbroj kvadrata reziduala, tj. jedan broj koji sažima cijelu kvalitetu prilagodbe.

Točke visoke poluge i utjecajne točke

Ekvivalent "hat" matrice $\mathbf{H}$ u logističkoj regresiji sadrži procijenjene vjerojatnosti $\hat{p}_k\left(\mathbf{z}_j\right)$. Utjecajnost se mjeri pomoću dijagonalnih elemenata h_{jj} matrice:

$$\mathbf{H} = \mathbf{V}^{-1/2}\mathbf{Z}\left(\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{V}^{-1/2} \tag{5.29}$$

gdje je $\mathbf{V}^{-1}$ dijagonalna matrica s (j, j) elementom $n_j\hat{p}\left(\mathbf{z}_j\right)\left(1-\hat{p}\left(\mathbf{z}_j\right)\right)$, a $\mathbf{V}^{-1/2}$ je dijagonalna matrica s (j, j) elementom $\sqrt{n_j\hat{p}\left(\mathbf{z}_j\right)\left(1-\hat{p}\left(\mathbf{z}_j\right)\right)}$.

Poglavlje 6

Primjeri

Primjer 6.0.1. *Godišnji financijski podaci za tvrtke koje su bankrotirale prikupljeni su otprilike dvije godine prije njihova bankrota, a za financijski zdrave tvrtke otprilike u isto vrijeme. Podaci o četiri varijable, X_1 = CF/TD = (novčani tok)/(ukupni dug), X_2 = NI/TA = (neto dobit)/(ukupna imovina), X_3 = CA/CL = (trenutna imovina)/(trenutne obveze), i X_4 = CA/NS = (trenutna imovina)/(neto prodaja), dani su u Table 11.4 Bankruptcy Data u [3].*

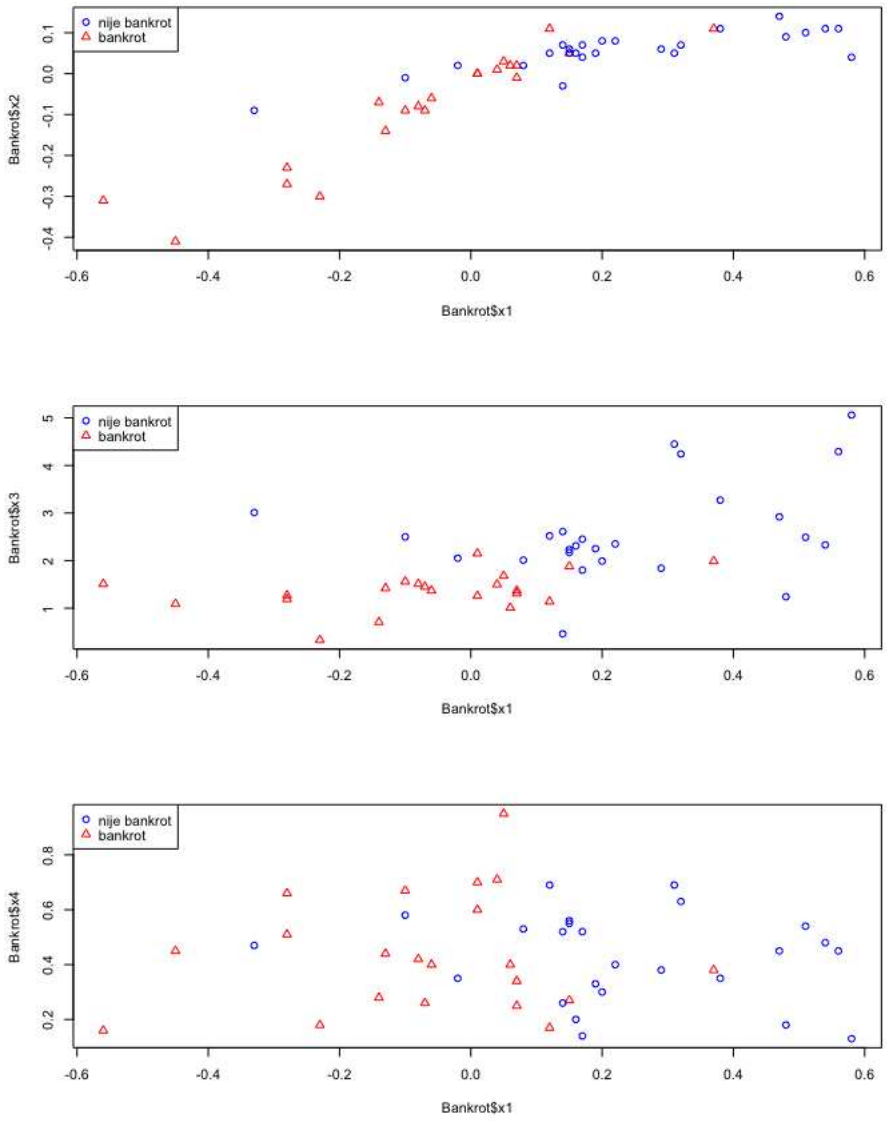
Pokažimo prvo scatterplot za parove varijabli. Na slici (6.1) *crvenim trokutićima su prikazani podaci koji odgovaraju tvrtkama koje su otišle u bankrot, a plavim kružićima one koje nisu. Prema izgledu podataka koji se čine eliptični možemo naslutiti kako se možda radi o podacima koji dolaze iz bivarijatne normalne razdiobe.*

Imamo $n_1 = 21$ podataka za tvrtke koje su otišle u bankrot i $n_2 = 25$ podataka o tvrtkama koje nisu otišle u bankrot. Izračunajmo očekivanja i kovarijacijske matrice za razne parove varijabli.
Za par (x_1, x_2):

$$\begin{aligned} \bar{x}_1 &= \begin{bmatrix} -0.068 \\ -0.081 \end{bmatrix}, \quad S_1 = \begin{bmatrix} 0.044 & 0.028 \\ 0.028 & 0.021 \end{bmatrix} \\ \bar{x}_2 &= \begin{bmatrix} 0.265 \\ 0.056 \end{bmatrix}, \quad S_2 = \begin{bmatrix} 0.047 & 0.009 \\ 0.009 & 0.002 \end{bmatrix} \end{aligned} \tag{6.1}$$

Za par (x_1, x_3):

$$\begin{aligned} \bar{x}_1 &= \begin{bmatrix} -0.068 \\ 1.367 \end{bmatrix}, \quad S_1 = \begin{bmatrix} 0.044 & 0.035 \\ 0.035 & 0.164 \end{bmatrix} \\ \bar{x}_2 &= \begin{bmatrix} 0.235 \\ 2.594 \end{bmatrix}, \quad S_2 = \begin{bmatrix} 0.047 & 0.075 \\ 0.075 & 1.047 \end{bmatrix} \end{aligned} \tag{6.2}$$



Slika 6.1: Grafički prikaz podataka

Za par (x_1, x_4)*:*

$$\begin{aligned} \bar{x}_1 &= \begin{bmatrix} -0.068 \\ 0.438 \end{bmatrix}, \quad S_1 = \begin{bmatrix} 0.044 & 0.004 \\ 0.004 & 0.045 \end{bmatrix} \\ \bar{x}_2 &= \begin{bmatrix} 0.235 \\ 0.427 \end{bmatrix}, \quad S_2 = \begin{bmatrix} 0.047 & -0.007 \\ -0.007 & 0.026 \end{bmatrix} \end{aligned} \tag{6.3}$$

Koristeći pretpostavku da podaci dolaze iz bivarijatne normalne razdiobe odredimo kvadratna klasifikacijska pravila po uzoru na (2.22)*. U tablici ispod su prikazane kvadratne klasifikacijske funkcije za razne kombinacije apriornih vjerojatnosti* p_1 *i* p_2*. Uvijek pretpostavljamo da je* $c\,(1 \mid 2) = c\,(2 \mid 1)$*.*

Varijable	apriorne vjerojatnost	klasifikacijska funkcija
(x_1, x_2)	$p_1 = p_2$	$-60.65x_1^2 + 30.15x_1x_2 - 406.52x_2^2 + 5.49x_1 - 29.88x_2 - 0.17$
(x_1, x_3)	$p_1 = p_2$	$-1.59x_1^2 + 3.99x_1x_2 - 3.10x_2^2 - 10.84x_1 + 7.95x_2 - 3.17$
(x_1, x_4)	$p_1 = p_2$	$-0.44x_1^2 + 7.58x_1x_2 + 8.31x_2^2 - 10.00x_1 + 8.00x_2 + 2.21$
(x_1, x_2)	$p_1 = 0.05, p_2 = 0.95$	$-60.65x_1^2 + 30.15x_1x_2 - 406.52x_2^2 + 5.49x_1 - 29.88x_2 - 3.11$
(x_1, x_3)	$p_1 = 0.05, p_2 = 0.95$	$-1.59x_1^2 + 3.99x_1x_2 - 3.10x_2^2 - 10.84x_1 + 7.95x_2 - 6.11$
(x_1, x_4)	$p_1 = 0.05, p_2 = 0.95$	$-0.44x_1^2 + 7.58x_1x_2 + 8.31x_2^2 - 10.00x_1 + 8.00x_2 - 0.73$

Tablica 6.1: Kvadratne klasifikacijske funkcije

Klasifikacijsko pravilo: ako je vrijednost kvadratne klasifikacijske funkcije veća od nule, klasificiraj podatak u populaciju π_1 *(tvrtke koje su otišle u bankrot), a inače klasificiraj u populaciju* π_2 *(tvrtke koje nisu otišle u bankrot). Vidimo kako se kvadratne funkcije za iste parove varijabli mijenjanju samo u slobodnom članu kako mijenjamo apriorne vjerojatnosti.*

Procijenimo sada performanse ovih klasifikacijskih pravila pomoću razina grešaka APER koje je definirano u (3.5) *i* $\hat{E}\,(AER)$ *definirano u* (3.7)*.*

Varijable	apriorne vjerojatnost	$APER$	$\hat{E}\,(AER)$
(x_1, x_2)	$p_1 = p_2$	19.57%	21.74%
(x_1, x_3)	$p_1 = p_2$	10.87%	13.04%
(x_1, x_4)	$p_1 = p_2$	17.39%	21.74%
(x_1, x_2)	$p_1 = 0.05, p_2 = 0.95$	26.09%	26.09%
(x_1, x_3)	$p_1 = 0.05, p_2 = 0.95$	36.96%	39.13%
(x_1, x_4)	$p_1 = 0.05, p_2 = 0.95$	39.13%	45.65%

Tablica 6.2: Razine grešaka

Za $p_1 = p_2$ najbolje rezultate za greške dobivamo za par varijabli (x_1, x_3) jer su obje greške male i približno jednake. Kod apriorne vjerojatnosti $p_1 = 0.05$ i $p_2 = 0.95$ najbolja opcija je za par varijabli (x_1, x_2) jer je tu greška najmanja.

Probajmo sada podatke klasificirati pomoću Fisherovog linearnog pravila za klasifikaciju. Prvo odredimo S_{proc} za par varijabli (x_1, x_2)

$$S_{proc} = \begin{bmatrix} 0.046 & 0.018 \\ 0.018 & 0.011 \end{bmatrix} \tag{6.4}$$

Da bismo definirali klasifikacijsko pravilo treba nam funkcija kao u (2.10) *i granica definirana kao u* (2.11) *:*

$$\hat{y} = -4.71x_1 - 5.00x_2, \quad m = -0.33 \tag{6.5}$$

Klasifikacijsko pravilo: Ako je $\hat{y} = ax_0 - m$ veće od nule, onda pridruži populaciji π_1 (tvrtke u bankrotu), inače pridruži π_2. Za ovo klasifikacijsko pravilo APER iznosi 19.57%. *Korištenje Fisherovog linearnog pravila nije prikladno zbog toga što kovarijacijske matrice nisu jednake, ali prema APER kriteriju, performansa mu je jednako dobra kao kod korištenja kvadratnog klasifikacijskog pravila.*

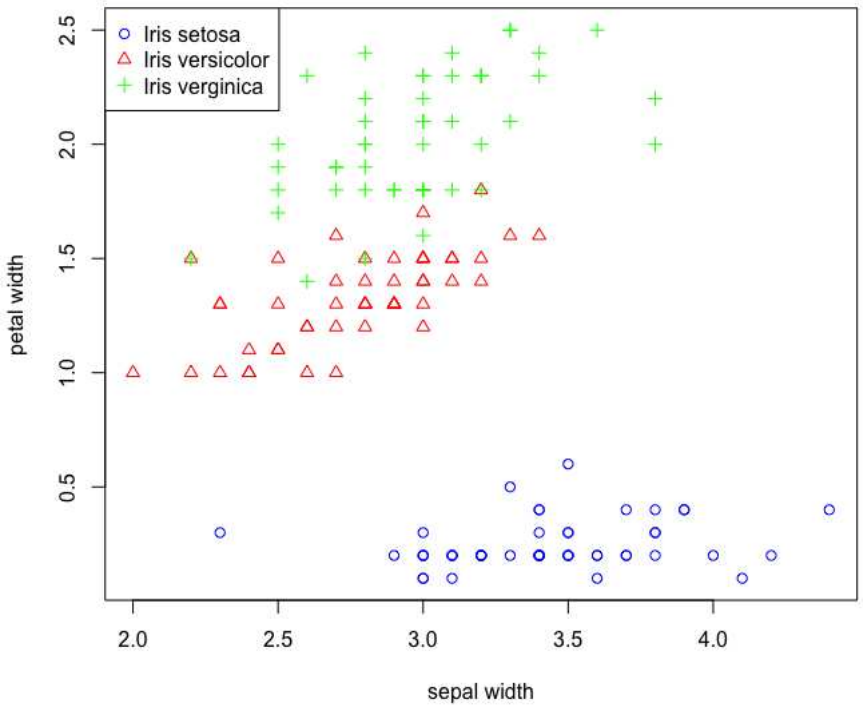
Primjer 6.0.2. *Promatramo podatke cvjetova Iris i pripadne varijable X_2 =* ***sepal width*** *i X_4 =* ***petal width*** *za uzorke iz tri različite populacije: iris setosa, iris versicolor i iris verginica. Svi uzorci su veličine $n_1 = n_2 = n_3 = 50$. Podaci su dani u [3] Table 11.5 Data on Irises. Na slici* (6.2) *možemo vidjeti da su podaci eliptično raspoređeni što upućuje da se možda radi o bivarijatnoj normalnoj razdiobi, međutim vidimo i kako je smjer podatak za populaciju iris setosa drugačiji od smjera ostale dvije populacije pa naslućujemo kako kovarijacijske matrice neće biti jednake.*

Pretpostavimo da podaci zbilja dolaze iz bivarijatne normalne razdiobe s istim kovarijacijskim matricama. Testirajmo jednakost srednjih vrijednosti na razini značajnosti od 95%. *Hipoteze su sljedeće:*

$$\begin{gathered} H_0 : \mu_1 = \mu_2 = \mu_3 \\ H_a : \text{barem jedan } \mu_i \text{ se razlikuje od ostalih} \end{gathered} \tag{6.6}$$

Koristeći MANOVA test dobivamo jako malu p vrijednost pa odbacujemo H_0 u korist alternative što smo i naslutili. Konstruirajmo klasifikacijsko pravilo po uzoru na (4.2.3) *i klasificirajmo novi podatak $x_0 =$* [3.5 1.75] *koristeći* (4.14). *Neka su $p_1 = p_2 = p_3 = \frac{1}{3}$.*

$$\begin{aligned} d_1^Q(\mathbf{x_0}) &= -104.8719 \\ d_2^Q(\mathbf{x_0}) &= -1.055387 \\ d_3^Q(\mathbf{x_0}) &= -2.325402 \end{aligned} \tag{6.7}$$



Slika 6.2: Grafički prikaz podataka Iris

Klasifikacijsko pravilo kaže da se novi podatak alocira u onu poplaciju čiji je diskriminacijski bod najveći, a to je u ovom slučaju populacija ***versicolor****. Ukoliko pretpostavimo da su kovarijacijske matrice jednake, možemo koristiti i linearni diskriminacijski bod definiran kao* (4.18) *i zajedničkom procijenjenom kovarijacijskom matricom* (4.17)*.*

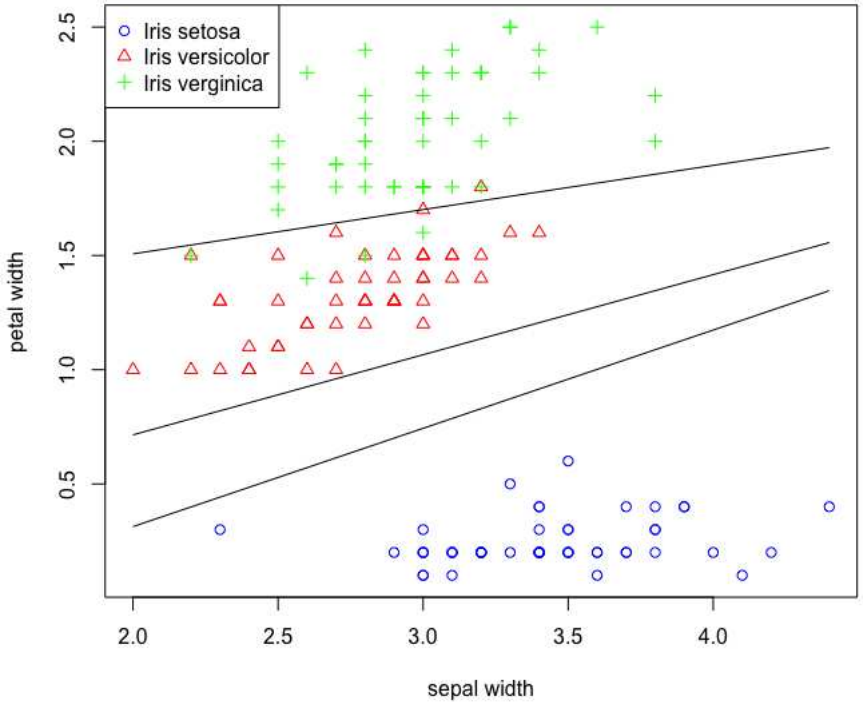
$$\begin{aligned} \hat{d}_1(\mathbf{x_0}) &= 27.01748 \\ \hat{d}_2(\mathbf{x_0}) &= 57.75736 \\ \hat{d}_3(\mathbf{x_0}) &= 56.81905 \end{aligned} \tag{6.8}$$

Dobivamo isti zaključak kao kod korištenja kvadratnog diskriminacijskog boda: novi podatak pridružujemo populaciji ***versicolor****. Pošto metoda koja koristi linearni diskriminacijski bod koristi pretpostavku o jednakosti kovarijacijskih matrica, preferiramo metodu koja koristi kvadratni diskriminacijski bod jer nam tu pretpostavka o jednakosti kovarijacijskih matrica nije potrebna.*

Klasificirajmo sada novi podatak $x_0 = [3.5 \quad 1.75]$ *koristeći* (4.20).

$\hat{d}_{ki}$	π_1 : Setosa	π_2 : Versicolor	π_3 : Virginica
π_1 : Setosa	0	-30.74	-29.80
π_2 : Versicolor	30.74	0	0.94
π_3 : Virginica	29.80	-0.94	0

Vidimo kako su svi $\hat{d}_{ki} \geq 0$ *za sve* $i \neq k$ *kod populacije* $\pi_2 : Versicolor$ *pa ponovno zaključujemo da novi podatak alociramo populaciji versicolor. Na slici* (6.3) *možemo vidjeti i klasifikacijska područja* $\hat{R}_1, \hat{R}_2$, *i* $\hat{R}_3$.



Slika 6.3: Grafički prikaz klasifikacijskih područja $\hat{R}_1, \hat{R}_2$, i $\hat{R}_3$

Za kraj izračunajmo razine grešaka:

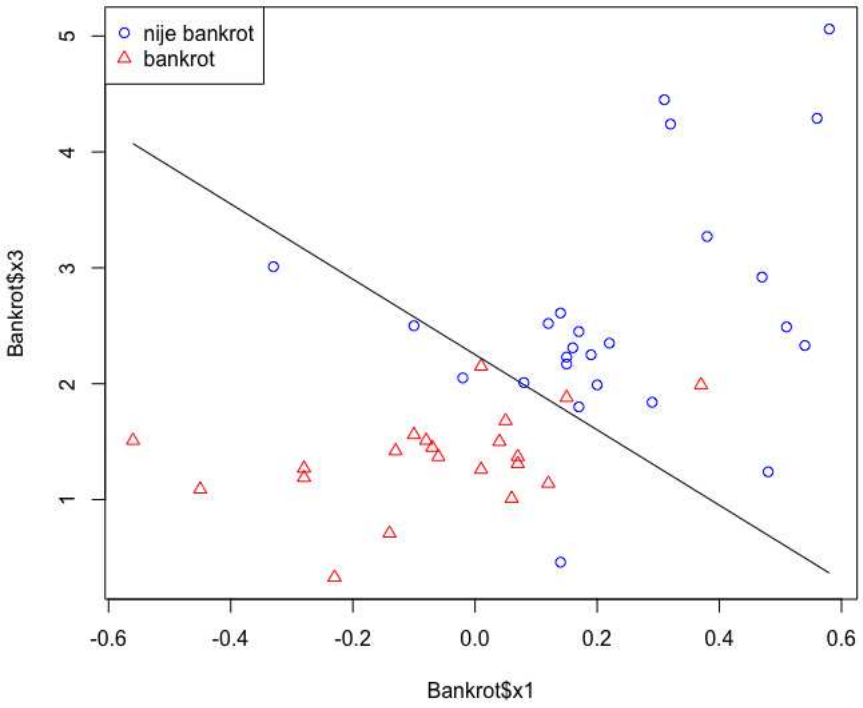
$$\begin{aligned} APER &= 3.3\% \\ \hat{E}\,(AER) &= 4\% \end{aligned} \tag{6.9}$$

Za računanje $\hat{E}$ (AER) je korištena Lachenbruchova metoda odgađanja objašnjena u (3).

Primjer 6.0.3. *Iskoristimo sada ponovo podatke iz [3] Table 11.4 vezane za bankrot. Pogledajmo kako uklanjanje nekih točaka iz skupa podataka za klasifikaciju utječe na razinu greške i je li utjecaj značajan. Konstruirajmo prvo Fisherovu linearnu diskriminantu za varijable x_1 i x_3:*

$$\hat{y} = -4.81x_1 - 1.47x_3 + 3.33 \tag{6.10}$$

Klasifikacijsko pravilo glasi: ako je $\hat{y} \geq 0$, onda pridruži populaciji π_1 (bankrot), inače pridruži π_2. Pogledajmo i grafički prikaz podataka i diskriminacijske funkcije te izračunajmo APER.



Slika 6.4: Grafički prikaz podataka i diskriminacijske funkcije $\hat{y} = -4.81x_1 - 1.47x_3 + 3.33$

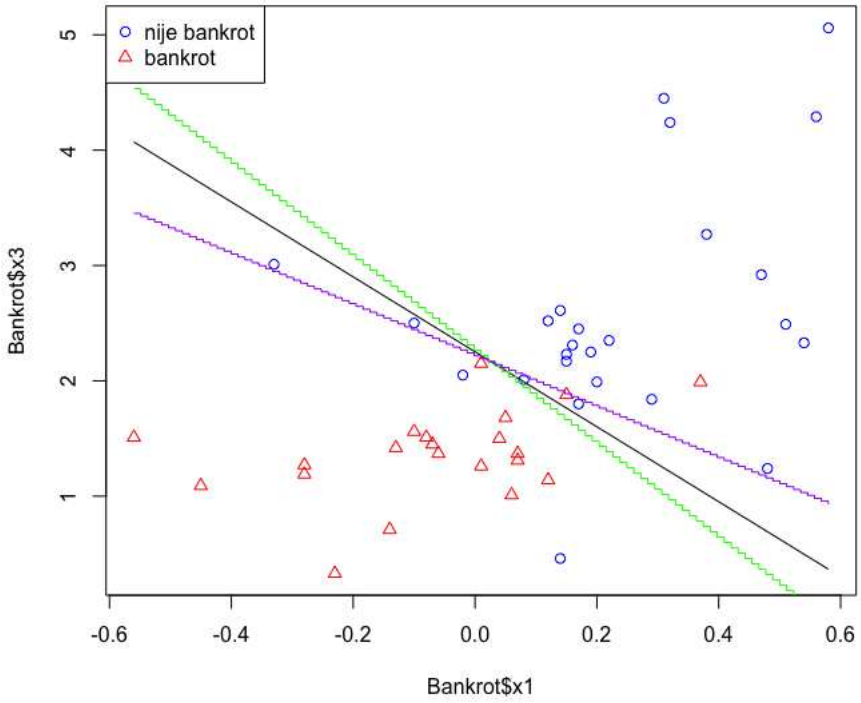
APER iznosi 13.04%. *Promatrajući podatke, vidimo da postoje podaci koji potencijalno možda mogu biti utjecajne točke, npr. za tvrtke koje su otišle u bankrot je to možda 16. podatak $(x_1, x_3) = (0.37, 1.99)$, i 34. podatak za drugu populaciju $(x_1, x_3) = (0.14, 0.46)$.*

Konstruirajmo ponovno Fisherove diskriminante , ali prvo bez jedne točke pa bez druge. Fisherova linearna diskriminanta bez točke $(x_1, x_3) = (0.37, 1.99)$ *glasi:*

$$\hat{y} = -5.95x_1 - 1.46x_3 + 3.31, \ \textit{gdje je APER} \ = 11\%. \tag{6.11}$$

Fisherova linearna diskriminanta bez točke $(x_1, x_3) = (0.14, 0.46)$ *glasi:*

$$\hat{y} = -4.36x_1 - 1.97x_3 + 4.36, \ \textit{gdje je APER} \ = 8.9\%. \tag{6.12}$$



Slika 6.5: Grafički prikaz podataka i tri različite diskriminacijske funkcije

Na slici (6.5) *zelenom bojom je označena diskriminacijska funkcija bez 16. podatka, a ljubičastom bojom je označena funkcija bez 34. podatka. Vidimo kako je izbacivanje tih točaka značajno utjecalo na promjenu diskriminacijske funkcije, a i na smanjenje razine greške (originalni APER* = 13.04% *).*

Primjer 6.0.4. *Podaci o dvije vrste lososa (aljaški i kanadski) i nekim njihovim karakteristikama kao što su spol ("Gender") te promjer prstena slatkovodnog ("Freshwater") i morskog ("Marine") rasta preuzeti su iz [3] Table 11.2 Salmon Data i primjera Example 11.8. Uzorak se sastoji od $n_1 = 50$ aljaških i $n_2 = 50$ kanadskih lososa. Odredimo puni model definiran kao u* (5.8)*. Dobiveni model je oblika:*

$$\ln\left(\frac{p(\mathbf{z})}{1-p(\mathbf{z})}\right) = 3.50501 + 0.28156 * Gender + 0.12642 * Freshwater - 0.04865 * Marine \tag{6.13}$$

Rezultat Likelihood Ratio testa, koji testira značajnost modela obzirom na reducirani model (model koji sadrži samo β_0 koeficijent), je statistički značajan na razini značajnosti od 5%, što nam govori kako je potrebna barem jedna prediktorska varijabla u modelu. Ako posebno promatramo značajnost varijabli u modelu, vidimo kako varijabla spol ("Gender") nije statistički značajna. Zapravo provodimo test značajnosti koeficijenta i testiramo hipotezu $H_0 : \beta_{Gender} = 0$ i pošto dobivamo p vrijednost od 0.73, na razini značajnosti od 5% ne možemo odbaciti nul hipotezu i zaključujemo kako varijabla Gender nije statistički značajna u modelu. S druge strane, varijable "Freshwater" i "Marine" poprimaju male p vrijednosti, manje od 0.05 pa zaključujemo kako su statistički značajne u modelu. Nakon izbacivanja varijable spol ("Gender") iz modela dobivamo sljedeće:

$$\ln\left(\frac{p(\mathbf{z})}{1-p(\mathbf{z})}\right) = 3.92484 + 0.12605 * Freshwater - 0.04854 * Marine \tag{6.14}$$

Promatrajući p vrijednosti u testu za značajnost koeficijenata zaključujemo kako su sve manje od 0.05 i da su sada u modelu sve varijable statistički značajne.

Iskoristimo pravilo za klasifikaciju definirano u (5.16) *koje glasi:*
Pridruži novi podatak u populaciju 2 ("Canadian salmon") ako je $\ln\left(\frac{p(\mathbf{z})}{1-p(\mathbf{z})}\right) \geq 0$*, inače pridruži populaciji 1 ("Alaskan salmon").*
Konfuzijska matrica izgleda:

		Stvarna populacija	
		Alaskan	Canadian
Klasificirano kao:	Alaskan	46	3
	Canadian	4	47

Razina pogreške APER iznosi $= \frac{4+3}{50+50} = \frac{7}{100} = 7\%$.

Pogledajmo sada vektore očekivanja i kovarijacijske matrice za varijable "Freshwater" i "Marine".

$$\begin{aligned} \bar{\mathbf{x}}_1 &= \begin{bmatrix} 98.38 \\ 429.66 \end{bmatrix}, \quad S_1 = \begin{bmatrix} 260.61 & -188.09 \\ -188.09 & 1399.09 \end{bmatrix} \\ \bar{\mathbf{x}}_2 &= \begin{bmatrix} 137.46 \\ 365.62 \end{bmatrix}, \quad S_2 = \begin{bmatrix} 326.09 & 133.50 \\ 133.50 & 893.26 \end{bmatrix} \end{aligned} \tag{6.15}$$

Pretpostavimo da su troškovi pogrešne klasifikacije jednaki i da su apriorne vjerojatnosti jednake. Iskoristimo linearno klasifikacijsko pravilo (vidi (2.10) *i* (2.11)*):*

$$\hat{w} = \hat{y} - m = -0.12838726 * Freshwater + 0.05194311 * Marine - 5.541204 \tag{6.16}$$

Klasifikacijsko pravilo: Ako je $\hat{y} = a\mathbf{z} - m$ *veće od nule, onda pridruži populaciji* π_1 *("Alaskan salmon"), inače pridruži* π_2 *("Canadian salmon"). Sada dobivamo ovakvu konfuzijsku matricu:*

		Stvarna populacija	
		Alaskan	Canadian
Klasificirano kao:	Alaskan	44	6
	Canadian	1	49

Greška APER je ponovno 7% *kao i kod logističke regresije.*

Poglavlje 7

Dodatak R-kod

Listing 7.1: Klasifikacija tvrtki pomoću kvadratne i linearne klasfikacijske funckije i Fisherovog linearnog pravila

```
par(mfrow=c(3,1))
plot(Bankrot$x1,Bankrot$x2,pch = ifelse(Bankrot$populacija ==
"1", 1, 2), col = ifelse(Bankrot$populacija == "1", "blue",
"red"))
legend("topleft", legend = c("nije bankrot", "bankrot"), pch =
c(1,2), col = c("blue", "red"))
plot(Bankrot$x1,Bankrot$x3,pch = ifelse(Bankrot$populacija ==
"1", 1, 2), col = ifelse(Bankrot$populacija == "1", "blue",
"red"))
legend("topleft", legend = c("nije bankrot", "bankrot"), pch =
c(1,2), col = c("blue", "red"))
plot(Bankrot$x1,Bankrot$x4,pch = ifelse(Bankrot$populacija ==
"1", 1, 2), col = ifelse(Bankrot$populacija == "1", "blue",
"red"))
legend("topleft", legend = c("nije bankrot", "bankrot"), pch =
c(1,2), col = c("blue", "red"))
par(mfrow=c(1,1))

n1<-length(Bankrot[Bankrot$populacija==0,]$x1)
n2<-length(Bankrot[Bankrot$populacija==1,]$x1)

#vektori ocekivanje
ocekivanje<-function(i,j){
  mean1<-c(mean(Bankrot[Bankrot$populacija==0,i]),
```

```
   mean(Bankrot[Bankrot$populacija==0,j]))
   mean2<-c(mean(Bankrot[Bankrot$populacija==1,i]),
   mean(Bankrot[Bankrot$populacija==1,j]))
   m<-matrix(c(mean1,mean2),2,2)
   return(m)}
ocekivanje(1,2)
ocekivanje(1,3)
ocekivanje(1,4)

#kovarijacijske matrice
kov_matrice<-function(i,j){
   x11_1<-cov(Bankrot[Bankrot$populacija==0,i],
   Bankrot[Bankrot$populacija==0,i])
   x12_1<-cov(Bankrot[Bankrot$populacija==0,i],
   Bankrot[Bankrot$populacija==0,j])
   x22_1<-cov(Bankrot[Bankrot$populacija==0,j],
   Bankrot[Bankrot$populacija==0,j])
   s1<-matrix(c(x11_1,x12_1,x12_1,x22_1),2,2)
   x11_2<-cov(Bankrot[Bankrot$populacija==1,i],
   Bankrot[Bankrot$populacija==1,i])
   x12_2<-cov(Bankrot[Bankrot$populacija==1,i],
   Bankrot[Bankrot$populacija==1,j])
   x22_2<-cov(Bankrot[Bankrot$populacija==1,j],
   Bankrot[Bankrot$populacija==1,j])
   s2<-matrix(c(x11_2,x12_2,x12_2,x22_2),2,2)
   l<-list(s1,s2)
   return(l)}
kov_matrice(1,2)
kov_matrice(1,3)
kov_matrice(1,4)

#klas_funkcije
m<-ocekivanje(1,2)
mean1<-m[,1]
mean2<-m[,2]
c<-kov_matrice(1,2)
s1<-c[[1]]
s2<-c[[2]]
s1_inverz <- solve(s1)
```

```
s2_inverz <- solve(s2)
k<-1/2*log(det(s1)/det(s2))+1/2*(mean1%*%s1_inverz%*%t(
t(mean1))-mean2%*%s2_inverz%*%t(t(mean2)))
k1<-1/2*log(det(s1)/det(s2))+1/2*(mean1%*%s1_inverz%*%t(
t(mean1))-mean2%*%s2_inverz%*%t(t(mean2)))+log(0.95/0.05)
-1/2*(s1_inverz-s2_inverz)
((mean1%*%s1_inverz-mean2%*%s2_inverz))
m<-ocekivanje(1,3)
mean1<-m[,1]
mean2<-m[,2]
c<-kov_matrice(1,3)
s1<-c[[1]]
s2<-c[[2]]
s1_inverz <- solve(s1)
s2_inverz <- solve(s2)
k<-1/2*log(det(s1)/det(s2))+1/2*(mean1%*%s1_inverz%*%t(
t(mean1))-mean2%*%s2_inverz%*%t(t(mean2)))
k1<-1/2*log(det(s1)/det(s2))+1/2*(mean1%*%s1_inverz%*%t(
t(mean1))-mean2%*%s2_inverz%*%t(t(mean2)))+log(0.95/0.05)
-1/2*(s1_inverz-s2_inverz)
((mean1%*%s1_inverz-mean2%*%s2_inverz))
m<-ocekivanje(1,4)
mean1<-m[,1]
mean2<-m[,2]
c<-kov_matrice(1,4)
s1<-c[[1]]
s2<-c[[2]]
s1_inverz <- solve(s1)
s2_inverz <- solve(s2)
k<-1/2*log(det(s1)/det(s2))+1/2*(mean1%*%s1_inverz%*%t(
t(mean1))-mean2%*%s2_inverz%*%t(t(mean2)))
k1<-1/2*log(det(s1)/det(s2))+1/2*(mean1%*%s1_inverz%*%t(
t(mean1))-mean2%*%s2_inverz%*%t(t(mean2)))+log(0.95/0.05)
-1/2*(s1_inverz-s2_inverz)
((mean1%*%s1_inverz-mean2%*%s2_inverz))

#APER i E(AER) za p1=p2
aper<-c()
e_aer<-c()
```

```
for(j in 2:4){
m<-ocekivanje(1,j)
mean1<-m[,1]
mean2<-m[,2]
c<-kov_matrice(1,j)
s1<-c[[1]]
s2<-c[[2]]
s1_inverz <- solve(s1)
s2_inverz <- solve(s2)
k<-1/2*log(det(s1)/det(s2))+1/2*(mean1%*%s1_inverz%*%t(
t(mean1))-mean2%*%s2_inverz%*%t(t(mean2)))
a<--1/2*(s1_inverz-s2_inverz)
b<-mean1%*%s1_inverz-mean2%*%s2_inverz

klas_pravilo1<-function(x,y){
a[1,1]*x^2+2*a[1,2]*x*y+a[2,2]*y^2+b[1]*x+b[2]*y-k}

n1<-length(Bankrot[Bankrot$populacija==0,]$x1)
n2<-length(Bankrot[Bankrot$populacija==1,]$x1)

n1c=sum(klas_pravilo1(Bankrot[Bankrot$populacija==0,]$x1,
Bankrot[Bankrot$populacija==0,j])>=0)
n2c=sum(klas_pravilo1(Bankrot[Bankrot$populacija==1,]$x1,
Bankrot[Bankrot$populacija==1,j])<0)
n1m=sum(klas_pravilo1(Bankrot[Bankrot$populacija==0,]$x1,
Bankrot[Bankrot$populacija==0,j])<0)
n2m=sum(klas_pravilo1(Bankrot[Bankrot$populacija==1,]$x1,
Bankrot[Bankrot$populacija==1,j])>=0)
aper[j-1]=(n1m+n2m)/(n1+n2)

klas_pravilo<-function(x,y,a,b,c,d,e,k){
a*x^2+b*x*y+c*y^2+d*x+e*y-k}

n1m_h=0
n2m_h=0
for (i in 1:(n1+n2)) {
holdout_data<-Bankrot[-i,]
holdout_x<-Bankrot[i,]
```

```
mean1<-c(mean(holdout_data[holdout_data$populacija==0,]$x1),
mean(holdout_data[holdout_data$populacija==0,j]))
mean2<-c(mean(holdout_data[holdout_data$populacija==1,]$x1),
mean(holdout_data[holdout_data$populacija==1,j]))

x11_1<-cov(holdout_data[holdout_data$populacija==0,]$x1,
holdout_data[holdout_data$populacija==0,]$x1)
x12_1<-cov(holdout_data[holdout_data$populacija==0,]$x1,
holdout_data[holdout_data$populacija==0,j])
x22_1<-cov(holdout_data[holdout_data$populacija==0,j],
holdout_data[holdout_data$populacija==0,j])
s1<-matrix(c(x11_1,x12_1,x12_1,x22_1),2,2)

x11_2<-cov(holdout_data[holdout_data$populacija==1,]$x1,
holdout_data[holdout_data$populacija==1,]$x1)
x12_2<-cov(holdout_data[holdout_data$populacija==1,]$x1,
holdout_data[holdout_data$populacija==1,j])
x22_2<-cov(holdout_data[holdout_data$populacija==1,j],
holdout_data[holdout_data$populacija==1,j])
s2<-matrix(c(x11_2,x12_2,x12_2,x22_2),2,2)

s1_inverz <- solve(s1)
s2_inverz <- solve(s2)
k<-1/2*log(det(s1)/det(s2))+1/2*(mean1%*%s1_inverz%*%t(
t(mean1)) - mean2%*%s2_inverz%*%t(t(mean2)))

a<--1/2*(s1_inverz-s2_inverz)
b<-mean1%*%s1_inverz-mean2%*%s2_inverz

if (klas_pravilo(holdout_x[1,1],holdout_x[1,j],a[1,1],2*a[1,2],
a[2,2],b[1],b[2],k)<0 && holdout_x[1,5]==0 ){n1m_h=n1m_h+1}
if (klas_pravilo(holdout_x[1,1],holdout_x[1,j],a[1,1],2*a[1,2],
a[2,2],b[1],b[2],k)>0 && holdout_x[1,5]==1 ){n2m_h=n2m_h+1}}

e_aer[j-1]=(n1m_h+n2m_h)/(n1+n2)}

#APER i E(AER) za p1=0.05, p2=0.95
for(j in 2:4){
m<-ocekivanje(1,j)
```

```
mean1<-m[,1]
mean2<-m[,2]
c<-kov_matrice(1,j)
s1<-c[[1]]
s2<-c[[2]]
s1_inverz <- solve(s1)
s2_inverz <- solve(s2)
k<-1/2*log(det(s1)/det(s2))+1/2*(mean1%*%s1_inverz%*%t
(t(mean1))-mean2%*%s2_inverz%*%t(t(mean2)))+log(0.95/0.05)
a<--1/2*(s1_inverz-s2_inverz)
b<-mean1%*%s1_inverz-mean2%*%s2_inverz

klas_pravilo1<-function(x,y){
a[1,1]*x^2+2*a[1,2]*x*y+a[2,2]*y^2+b[1]*x+b[2]*y-k}

n1<-length(Bankrot[Bankrot$populacija==0,]$x1)
n2<-length(Bankrot[Bankrot$populacija==1,]$x1)
n1c=sum(klas_pravilo1(Bankrot[Bankrot$populacija==0,]$x1,
Bankrot[Bankrot$populacija==0,j])>=0)
n2c=sum(klas_pravilo1(Bankrot[Bankrot$populacija==1,]$x1,
Bankrot[Bankrot$populacija==1,j])<0)
n1m=sum(klas_pravilo1(Bankrot[Bankrot$populacija==0,]$x1,
Bankrot[Bankrot$populacija==0,j])<0)
n2m=sum(klas_pravilo1(Bankrot[Bankrot$populacija==1,]$x1,
Bankrot[Bankrot$populacija==1,j])>=0)
aper[j-1]=(n1m+n2m)/(n1+n2)

klas_pravilo<-function(x,y,a,b,c,d,e,k){
a*x^2+b*x*y+c*y^2+d*x+e*y-k
}

n1m_h=0
n2m_h=0
for (i in 1:(n1+n2)) {
holdout_data<-Bankrot[-i,]
holdout_x<-Bankrot[i,]

mean1<-c(mean(holdout_data[holdout_data$populacija==0,]$x1),
mean(holdout_data[holdout_data$populacija==0,j]))
```

```
mean2<-c(mean(holdout_data[holdout_data$populacija==1,]$x1),
mean(holdout_data[holdout_data$populacija==1,j]))
x11_1<-cov(holdout_data[holdout_data$populacija==0,]$x1,
holdout_data[holdout_data$populacija==0,]$x1)
x12_1<-cov(holdout_data[holdout_data$populacija==0,]$x1,
holdout_data[holdout_data$populacija==0,j])
x22_1<-cov(holdout_data[holdout_data$populacija==0,j],
holdout_data[holdout_data$populacija==0,j])
s1<-matrix(c(x11_1,x12_1,x12_1,x22_1),2,2)
x11_2<-cov(holdout_data[holdout_data$populacija==1,]$x1,
holdout_data[holdout_data$populacija==1,]$x1)
x12_2<-cov(holdout_data[holdout_data$populacija==1,]$x1,
holdout_data[holdout_data$populacija==1,j])
x22_2<-cov(holdout_data[holdout_data$populacija==1,j],
holdout_data[holdout_data$populacija==1,j])
s2<-matrix(c(x11_2,x12_2,x12_2,x22_2),2,2)
s1_inverz <- solve(s1)
s2_inverz <- solve(s2)
k<-1/2*log(det(s1)/det(s2))+1/2*(mean1%*%s1_inverz%*%t(
t(mean1)) - mean2%*%s2_inverz%*%t(t(mean2)))+log(0.95/0.05)
a<--1/2*(s1_inverz-s2_inverz)
b<-mean1%*%s1_inverz-mean2%*%s2_inverz

if (klas_pravilo(holdout_x[1,1],holdout_x[1,j],a[1,1],2*a[1,2]
,a[2,2],b[1],b[2],k)<0 && holdout_x[1,5]==0 ) {n1m_h=n1m_h+1}
if (klas_pravilo(holdout_x[1,1],holdout_x[1,j],a[1,1],2*a[1,2]
,a[2,2],b[1],b[2],k)>0 && holdout_x[1,5]==1 ) {n2m_h=n2m_h+1}}
e_aer[j-1]=(n1m_h+n2m_h)/(n1+n2)}

#FISHER
m<-ocekivanje(1,2)
mean1<-m[,1]
mean2<-m[,2]
c<-kov_matrice(1,2)
s1<-c[[1]]
s2<-c[[2]]
Sproc=(n1-1)/(n1-1+n2-1)*s1+(n2-1)/(n1-1+n2-1)*s2
Sproc
y<-(mean1-mean2)*solve(Sproc)
```

```
a<-c(y[1,1]+y[2,1],y[1,2]+y[2,2])
m<-1/2*(a%*%t(t(mean1))+a%*%t(t(mean2)))
klas_prav_fisher<-function(x,y){
   a[1]*x+a[2]*y-m}
n1c=sum(klas_prav_fisher(Bankrot[Bankrot$populacija==0,]$x1,
Bankrot[Bankrot$populacija==0,]$x2)>=0)
n2c=sum(klas_prav_fisher(Bankrot[Bankrot$populacija==1,]$x1,
Bankrot[Bankrot$populacija==1,]$x2)<0)
n1m=sum(klas_prav_fisher(Bankrot[Bankrot$populacija==0,]$x1,
Bankrot[Bankrot$populacija==0,]$x2)<0)
n2m=sum(klas_prav_fisher(Bankrot[Bankrot$populacija==1,]$x1,
Bankrot[Bankrot$populacija==1,]$x2)>=0)
aper=(n1m+n2m)/(n1+n2)
```

Listing 7.2: Klasifikacija cvjetova Iris pomoću kvadratnog i linearnog boda te pripadne razine grešaka

```
data(iris)
setosa <- iris[iris$Species == "setosa", c("Sepal.Width",
"Petal.Width")]
versicolor <- iris[iris$Species == "versicolor",
c("Sepal.Width", "Petal.Width")]
virginica <- iris[iris$Species == "virginica", c("Sepal.Width",
"Petal.Width")]

plot(iris$Sepal.Width,iris$Petal.Width,pch = ifelse(
iris$Species == "setosa", 1, ifelse(iris$Species ==
"versicolor",2,3)), col = ifelse(iris$Species == "setosa",
"blue", ifelse(iris$Species == "versicolor", "red","green"
)), xlab = "sepal width", ylab = "petal width" )
legend("topleft", legend = c("Iris setosa", "Iris versicolor",
"Iris verginica"), pch = c(1,2,3), col = c("blue", "red",
"green"))

#test jednakosti ocekivanja
manova_data <- iris[,c("Sepal.Width","Petal.Width","Species")]
model <- manova(cbind(Sepal.Width, Petal.Width) ~ Species,
data = manova_data)
summary(model)
# p vrijednost manja od svih razina znacajnosti
```

```
#vektori ocekivanja
mean1<-c(mean(iris[iris$Species== "setosa",]$Sepal.Width),
mean(iris[iris$Species== "setosa",]$Petal.Width))
mean2<-c(mean(iris[iris$Species=="versicolor",]$Sepal.Width)
,mean(iris[iris$Species=="versicolor",]$Petal.Width))
mean3<-c(mean(iris[iris$Species=="virginica",]$Sepal.Width)
,mean(iris[iris$Species=="virginica",]$Petal.Width))

#kovarijacijske matrice
cov1<-cov(setosa)
cov2<-cov(versicolor)
cov3<-cov(virginica)

klasifikacijska_fja_kvadratna<-function(x,mean,cov){
-1/2*log(det(cov))-1/2*(x-mean)%*%solve(cov)%*%t(t(
x-mean))+log(1/3)}

x0<-c(3.5,1.75)
klasifikacijska_fja_kvadratna(x0,mean1,cov1) #-104.8719
klasifikacijska_fja_kvadratna(x0,mean2,cov2) #-1.055387
klasifikacijska_fja_kvadratna(x0,mean3,cov3) #-2.325402
#najvecu vrijednost dobivamo za populaciju="versicolor"


#procijenjena zajednicka kovarijacijska matrica
n1=n2=n3=50
Sproc<-1/(n1+n2+n3-3)*((n1-1)*cov1+(n2-1)*cov2+(n3-1)*cov3)

klasifikacijska_fja_linearna<-function(x,mean){
  mean%*%solve(Sproc)%*%t(t(x))-1/2*(mean)%*%solve(Sproc)%*%
  t(t(mean))+log(1/3)}

klasifikacijska_fja_linearna(x0,mean1) #27.01748
klasifikacijska_fja_linearna(x0,mean2) #57.75736
klasifikacijska_fja_linearna(x0,mean3) #56.81905
#ponovno najvecu vrijednost dobivamo za populaciju=
"versicolor"
```

```
#alternativna klas lin fja
klasifikacijska_fja<-function(x,m1,m2){
(m1-m2)%*%solve(Sproc)%*%t(t(x))-1/2*(m1-m2)%*%solve
(Sproc)%*%t(t(m1+m2))}

klasifikacijska_fja(x0,mean1,mean2)   #-30.73988
klasifikacijska_fja(x0,mean2,mean1)   # 30.73988
klasifikacijska_fja(x0,mean1,mean3)   #-29.80157
klasifikacijska_fja(x0,mean3,mean1)   # 29.80157
klasifikacijska_fja(x0,mean2,mean3)   # 0.938316
klasifikacijska_fja(x0,mean3,mean2)   #-0.938316

(mean1-mean2)%*%solve(Sproc)
(mean1-mean3)%*%solve(Sproc)
(mean2-mean3)%*%solve(Sproc)
-1/2*(mean1-mean2)%*%solve(Sproc)%*%t(t(mean1+mean2))
-1/2*(mean1-mean3)%*%solve(Sproc)%*%t(t(mean1+mean3))
-1/2*(mean2-mean3)%*%solve(Sproc)%*%t(t(mean2+mean3))
f1<-function(x){
  return((-16.7129*x+21.26505)/-38.84)}
f2<-function(x){
  return((20.52756*x+0.7848593)/58.53307)}
f3<-function(x){
  return((3.814655*x+22.0499)/19.69308)}
plot(iris$Sepal.Width,iris$Petal.Width,pch = ifelse(iris$
Species == "setosa", 1, ifelse(iris$Species == "versicolor"
,2,3)), col = ifelse(iris$Species == "setosa","blue", ifelse(
iris$Species == "versicolor", "red","green" )), xlab =
"sepal width",ylab = "petal width" )
legend("topleft", legend = c("Iris setosa", "Iris versicolor",
"Iris verginica"), pch = c(1,2,3), col = c("blue", "red",
"green"))
curve(f1(x), add = T)
curve(f2(x), add = T)
curve(f3(x), add = T)

#APER i E(AER)
klas_fj<-function(x,mean,S){
mean%*%solve(S)%*%t(x)-1/2*(mean)%*%solve(S)%*%t(t(mean))
```

```
+log(1/3)}
klasifikacijska_fja_linearna<-function(x,mean){
mean%*%solve(Sproc)%*%t(x)-1/2*(mean)%*%solve(Sproc)%*%t(
t(mean))+log(1/3)}

for(i in 1:150){
if(klasifikacijska_fja_linearna(iris[i,c(2,4)],mean1)>
klasifikacijska_fja_linearna(iris[i,c(2,4)],mean2)){
max<-klasifikacijska_fja_linearna(iris[i,c(2,4)],mean1)
maxi<-"setosa"
} else if(klasifikacijska_fja_linearna(iris[i,c(2,4)],mean2)
>klasifikacijska_fja_linearna(iris[i,c(2,4)],mean3)){
max<-klasifikacijska_fja_linearna(iris[i,c(2,4)],mean2)
maxi<-"versicolor"
} else {
maxi<-"virginica"
}
iris$klas[i]<-maxi
}
n1c=sum(iris$Species==iris$klas)
aper=(150-n1c)/150
aper

n1m_h=0
n2m_h=0
n3m_h=0
for (i in 1:(n1+n2+n3)) {
holdout_data<-iris[-i,c(2,4,5)]
holdout_x<-iris[i,c(2,4,5)]

mean1<-c(mean(holdout_data[holdout_data$Species== "setosa",]
$Sepal.Width),mean(holdout_data[holdout_data$Species==
"setosa",]$Petal.Width))
mean2<-c(mean(holdout_data[holdout_data$Species=="versicolor"
,]$Sepal.Width),mean(holdout_data[holdout_data$Species==
"versicolor",]$Petal.Width))
mean3<-c(mean(holdout_data[holdout_data$Species=="virginica"
,]$Sepal.Width),mean(holdout_data[holdout_data$Species==
"virginica",]$Petal.Width))
```

```
setosa <- holdout_data[holdout_data$Species == "setosa",
c("Sepal.Width", "Petal.Width")]
versicolor <- holdout_data[holdout_data$Species ==
"versicolor",c("Sepal.Width", "Petal.Width")]
virginica <- holdout_data[holdout_data$Species ==
"virginica", c("Sepal.Width", "Petal.Width")]
cov1<-cov(setosa)
cov2<-cov(versicolor)
cov3<-cov(virginica)

if(i<=50){Sproc<-1/(49+n2+n3-3)*((49-1)*cov1+(n2-1)*cov2+
(n3-1)*cov3)}
if(i>50 && i<=100){Sproc<-1/(n1+49+n3-3)*((n1-1)*cov1+(49-1)
*cov2+(n3-1)*cov3)}
if(i>100){Sproc<-1/(n1+n2+49-3)*((n1-1)*cov1+(n2-1)*cov2
+(49-1)*cov3)}
max<-0

if(klas_fj(holdout_x[1,c(1,2)],mean1,Sproc)>klas_fj(
holdout_x[1,c(1,2)],mean2,Sproc)){
max<-klas_fj(holdout_x[1,c(1,2)],mean1,Sproc)
maxi<-"setosa"
} else if(klas_fj(holdout_x[1,c(1,2)],mean2,Sproc)>klas_fj(
holdout_x[1,c(1,2)],mean3,Sproc)){
max<-klas_fj(holdout_x[1,c(1,2)],mean2,Sproc)
maxi<-"versicolor"
} else {
maxi<-"virginica"}

if(maxi!=holdout_x[1,3] && i<=50) {n1m_h=n1m_h+1}
if(maxi!=holdout_x[1,3] && i>50 && i<=100) {n2m_h=n2m_h+1}
if(maxi!=holdout_x[1,3] && i>100) {n3m_h=n3m_h+1}}
eaer<-(n1m_h+n2m_h+n3m_h)/(n1+n2+n3)
```

Listing 7.3: Utjecajne točke kod klasifikacije tvrtki za bankrot

```
n1<-length(Bankrot[Bankrot$populacija==0,]$x1)
n2<-length(Bankrot[Bankrot$populacija==1,]$x1)

ocekivanje<-function(i,j){
```

```
mean1<-c(mean(Bankrot[Bankrot$populacija==0,i]),mean(
Bankrot[Bankrot$populacija==0,j]))
mean2<-c(mean(Bankrot[Bankrot$populacija==1,i]),mean(
Bankrot[Bankrot$populacija==1,j]))
m<-matrix(c(mean1,mean2),2,2)
return(m)}

kov_matrice<-function(i,j){
x11_1<-cov(Bankrot[Bankrot$populacija==0,i],Bankrot[
Bankrot$populacija==0,i])
x12_1<-cov(Bankrot[Bankrot$populacija==0,i],Bankrot[
Bankrot$populacija==0,j])
x22_1<-cov(Bankrot[Bankrot$populacija==0,j],Bankrot[
Bankrot$populacija==0,j])
s1<-matrix(c(x11_1,x12_1,x12_1,x22_1),2,2)
x11_2<-cov(Bankrot[Bankrot$populacija==1,i],Bankrot[
Bankrot$populacija==1,i])
x12_2<-cov(Bankrot[Bankrot$populacija==1,i],Bankrot[
Bankrot$populacija==1,j])
x22_2<-cov(Bankrot[Bankrot$populacija==1,j],Bankrot[
Bankrot$populacija==1,j])
s2<-matrix(c(x11_2,x12_2,x12_2,x22_2),2,2)
l<-list(s1,s2)
return(l)}

m<-ocekivanje(1,3)
mean1<-m[,1]
mean2<-m[,2]
c<-kov_matrice(1,3)
s1<-c[[1]]
s2<-c[[2]]
Sproc=(n1-1)/(n1-1+n2-1)*s1+(n2-1)/(n1-1+n2-1)*s2
Sproc
y<-(mean1-mean2)*solve(Sproc)
a<-c(y[1,1]+y[2,1],y[1,2]+y[2,2])
m<-1/2*(a%*%t(t(mean1))+a%*%t(t(mean2)))

klas_prav_fisher<-function(x,y){
a[1]*x+a[2]*y-m}
```

```
n1c=sum(klas_prav_fisher(Bankrot[Bankrot$populacija==0,]$x1,
Bankrot[Bankrot$populacija==0,]$x3)>=0)
n2c=sum(klas_prav_fisher(Bankrot[Bankrot$populacija==1,]$x1,
Bankrot[Bankrot$populacija==1,]$x3)<0)
n1m=sum(klas_prav_fisher(Bankrot[Bankrot$populacija==0,]$x1,
Bankrot[Bankrot$populacija==0,]$x3)<0)
n2m=sum(klas_prav_fisher(Bankrot[Bankrot$populacija==1,]$x1,
Bankrot[Bankrot$populacija==1,]$x3)>=0)
aper=(n1m+n2m)/(n1+n2)

f<-function(x){(m-a[1]*x)/a[2]}

plot(Bankrot$x1,Bankrot$x3,pch = ifelse(Bankrot$populacija ==
"1", 1, 2), col = ifelse(Bankrot$populacija == "1", "blue",
"red"))
legend("topleft", legend = c("nije bankrot", "bankrot"), pch =
c(1,2), col = c("blue", "red"))
curve(f(x), add = T)

##bez 16.tocke
podaci1<-Bankrot[-16,]
n1<-length(podaci1[podaci1$populacija==0,-16]$x1)
n2<-length(podaci1[podaci1$populacija==1,]$x1)

ocekivanje<-function(i,j){
mean1<-c(mean(podaci1[podaci1$populacija==0,i]),mean(
podaci1[podaci1$populacija==0,j]))
mean2<-c(mean(podaci1[podaci1$populacija==1,i]),mean(
podaci1[podaci1$populacija==1,j]))
m<-matrix(c(mean1,mean2),2,2)
return(m)}

kov_matrice<-function(i,j){
x11_1<-cov(podaci1[podaci1$populacija==0,i],podaci1[
podaci1$populacija==0,i])
x12_1<-cov(podaci1[podaci1$populacija==0,i],podaci1[
podaci1$populacija==0,j])
x22_1<-cov(podaci1[podaci1$populacija==0,j],podaci1[
```

```
podaci1$populacija==0,j])
s1<-matrix(c(x11_1,x12_1,x12_1,x22_1),2,2)
x11_2<-cov(podaci1[podaci1$populacija==1,i],podaci1[
podaci1$populacija==1,i])
x12_2<-cov(podaci1[podaci1$populacija==1,i],podaci1[
podaci1$populacija==1,j])
x22_2<-cov(podaci1[podaci1$populacija==1,j],podaci1[
podaci1$populacija==1,j])
s2<-matrix(c(x11_2,x12_2,x12_2,x22_2),2,2)
l<-list(s1,s2)
return(l)}

m<-ocekivanje(1,3)
mean1<-m[,1]
mean2<-m[,2]
c<-kov_matrice(1,3)
s1<-c[[1]]
s2<-c[[2]]
Sproc=(n1-1)/(n1-1+n2-1)*s1+(n2-1)/(n1-1+n2-1)*s2
Sproc
y<-(mean1-mean2)*solve(Sproc)
a1<-c(y[1,1]+y[2,1],y[1,2]+y[2,2])
m1<-1/2*(a1%*%t(t(mean1))+a1%*%t(t(mean2)))

klas_prav_fisher<-function(x,y){
a1[1]*x+a1[2]*y-m1}

n1c=sum(klas_prav_fisher(podaci1[podaci1$populacija==0,]$x1,
podaci1[podaci1$populacija==0,]$x3)>=0)
n2c=sum(klas_prav_fisher(podaci1[podaci1$populacija==1,]$x1,
podaci1[podaci1$populacija==1,]$x3)<0)
n1m=sum(klas_prav_fisher(podaci1[podaci1$populacija==0,]$x1,
podaci1[podaci1$populacija==0,]$x3)<0)
n2m=sum(klas_prav_fisher(podaci1[podaci1$populacija==1,]$x1,
podaci1[podaci1$populacija==1,]$x3)>=0)
aper=(n1m+n2m)/(n1+n2)

##bez 13.tocke
podaci2<-Bankrot[-34,]
```

```
n1<-length(podaci2[podaci2$populacija==0,-16]$x1)
n2<-length(podaci2[podaci2$populacija==1,]$x1)

ocekivanje<-function(i,j){
mean1<-c(mean(podaci2[podaci2$populacija==0,i]),mean(
podaci2[podaci2$populacija==0,j]))
mean2<-c(mean(podaci2[podaci2$populacija==1,i]),mean(
podaci2[podaci2$populacija==1,j]))
m<-matrix(c(mean1,mean2),2,2)
return(m)}

kov_matrice<-function(i,j){
x11_1<-cov(podaci2[podaci2$populacija==0,i],podaci2[
podaci2$populacija==0,i])
x12_1<-cov(podaci2[podaci2$populacija==0,i],podaci2[
podaci2$populacija==0,j])
x22_1<-cov(podaci2[podaci2$populacija==0,j],podaci2[
podaci2$populacija==0,j])
s1<-matrix(c(x11_1,x12_1,x12_1,x22_1),2,2)
x11_2<-cov(podaci2[podaci2$populacija==1,i],podaci2[
podaci2$populacija==1,i])
x12_2<-cov(podaci2[podaci2$populacija==1,i],podaci2[
podaci2$populacija==1,j])
x22_2<-cov(podaci2[podaci2$populacija==1,j],podaci2[
podaci2$populacija==1,j])
s2<-matrix(c(x11_2,x12_2,x12_2,x22_2),2,2)
l<-list(s1,s2)
return(l)}

m<-ocekivanje(1,3)
mean1<-m[,1]
mean2<-m[,2]
c<-kov_matrice(1,3)
s1<-c[[1]]
s2<-c[[2]]
Sproc=(n1-1)/(n1-1+n2-1)*s1+(n2-1)/(n1-1+n2-1)*s2
Sproc
y<-(mean1-mean2)*solve(Sproc)
a2<-c(y[1,1]+y[2,1],y[1,2]+y[2,2])
```

```
m2<-1/2*(a2%*%t(t(mean1))+a2%*%t(t(mean2)))

klas_prav_fisher<-function(x,y){
  a2[1]*x+a2[2]*y-m2}

n1c=sum(klas_prav_fisher(podaci2[podaci2$populacija==0,]$x1,
podaci2[podaci2$populacija==0,]$x3)>=0)
n2c=sum(klas_prav_fisher(podaci2[podaci2$populacija==1,]$x1,
podaci2[podaci2$populacija==1,]$x3)<0)
n1m=sum(klas_prav_fisher(podaci2[podaci2$populacija==0,]$x1,
podaci2[podaci2$populacija==0,]$x3)<0)
n2m=sum(klas_prav_fisher(podaci2[podaci2$populacija==1,]$x1,
podaci2[podaci2$populacija==1,]$x3)>=0)
aper=(n1m+n2m)/(n1+n2)

f1<-function(x){(m1-a1[1]*x)/a1[2]}
f2<-function(x){(m2-a2[1]*x)/a2[2]}

plot(Bankrot$x1,Bankrot$x3,pch = ifelse(Bankrot$populacija ==
"1", 1, 2), col = ifelse(Bankrot$populacija == "1", "blue",
"red"))
legend("topleft", legend = c("nije_bankrot", "bankrot"), pch =
c(1,2), col = c("blue", "red"))
curve(f(x), add = T)
curve(f1(x), add = T, col="green", type = "s")
curve(f2(x), add = T, col="purple",type = "s")
```

Listing 7.4: Klasifikacija lososa pomoću logističke regresije

```
library(lmtest)
#puni model
model<-glm(Y~Gender+Freshwater+Marine,family=binomial(link
= "logit"),data=salmon)
summary(model)
#testiranje znacajnosti modela - Likelihood Ratio
lrtest(model) #mala p_vrijednost
#model bez varijable Gender
model2<-glm(Y~Freshwater+Marine,family=binomial(link =
"logit"),data=salmon)
summary(model2)
```

```
#testiranje znacajnosti modela - Likelihood Ratio
lrtest(model2, model)
#puni model nije bolji, tj ne odbacujemo nul hipotezu o
neznacajnosti dodatne varijable Gender
#klasificiranje podataka pomocu klasifikacijskog pravila.
predicted <- predict(model2, salmon, type="response")
library(caret)
salmon$Y_predicted <- ifelse(log(predicted/(1-predicted))
>=0, 1, 0)
confusionMatrix( as.factor(salmon$Y_predicted),as.factor
(salmon$Y))

#ocekivanja
mean1<-c(mean(salmon[salmon$Y==0,2]),mean(salmon
[salmon$Y==0,3]))
mean2<-c(mean(salmon[salmon$Y==1,2]),mean(salmon
[salmon$Y==1,3]))

#kovarijacijske matrice
x11_1<-cov(salmon[salmon$Y==0,2],salmon[salmon$Y==0,2])
x12_1<-cov(salmon[salmon$Y==0,2],salmon[salmon$Y==0,3])
x22_1<-cov(salmon[salmon$Y==0,3],salmon[salmon$Y==0,3])
s1<-matrix(c(x11_1,x12_1,x12_1,x22_1),2,2)
x11_2<-cov(salmon[salmon$Y==1,2],salmon[salmon$Y==1,2])
x12_2<-cov(salmon[salmon$Y==1,2],salmon[salmon$Y==1,3])
x22_2<-cov(salmon[salmon$Y==1,3],salmon[salmon$Y==1,3])
s2<-matrix(c(x11_2,x12_2,x12_2,x22_2),2,2)
#linearno klasifikacijsko pravilo
n1=50
n2=50
Sproc=(n1-1)/(n1-1+n2-1)*s1+(n2-1)/(n1-1+n2-1)*s2
Sproc
y<-(mean1-mean2)*solve(Sproc)
a<-c(y[1,1]+y[2,1],y[1,2]+y[2,2])
m<-1/2*(a%*%t(t(mean1))+a%*%t(t(mean2)))
klas_prav_fisher<-function(x,y){
  a[1]*x+a[2]*y-m}
n1c=sum(klas_prav_fisher(salmon[salmon$Y==0,]$Freshwater,
salmon[salmon$Y==0,]$Marine)>=0)
```

```
n2c=sum(klas_prav_fisher(salmon[salmon$Y==1,]$Freshwater,
salmon[salmon$Y==1,]$Marine)<0)
n1m=sum(klas_prav_fisher(salmon[salmon$Y==0,]$Freshwater,
salmon[salmon$Y==0,]$Marine)<0)
n2m=sum(klas_prav_fisher(salmon[salmon$Y==1,]$Freshwater,
salmon[salmon$Y==1,]$Marine)>=0)
aper=(n1m+n2m)/(n1+n2)
```

Bibliografija

[1] T.W. Anderson, *An lntroduction to Multivariate Statistical Analysis (3rd ed.)*, New York: John Wiley i Sons, Inc, 2003.

[2] Miljenko Huzak, *Vjerojatnost i matematička statistika*, 2006, `http://aktuari.math.pmf.unizg.hr/docs/vms.pdf`.

[3] R.A. Johnson i D.W. Wichern, *Applied Multivariate Statistical Analysis, 6th ed.*, New Jersey: Pearson Prentice Hall, 2007.

[4] Max Kuhn i Kjell Johnson, *Applied Predictive Modeling*, Springer, 2013.

[5] Nikola Sarapa, *Teorija vjerojatnosti*, Školska Knjiga, 2002.

Sažetak

U ovom diplomskom radu navest ćemo neke metode za diskriminaciju i klasifikaciju podataka. Na početku ćemo definirati kriterije za optimalnu klasifikaciju podataka, a kasnije i metode za procjenu klasifikacijskih funkcija te objasniti kako izračunati razinu pogreške prilikom klasifikacije.

U drugom poglavlju ćemo navesti nekoliko metoda za klasifikaciju s dvije multivarijatne normalne populacije za slučaj kada su kovarijacijske matrice jednake i za slučaj kada nisu. U četvrtom poglavlju ćemo se baviti klasifikacijom s više populacija. Spomenut ćemo i Fisherov pristup za klasifikaciju. U petom poglavlju ćemo opisati logističku regresiju kao još jednu metodu za klasifikaciju i diskriminaciju podataka. U šestom poglavlju ćemo riješiti i objasniti nekoliko primjera na raznim podacima kako bismo bolje razumijeli primjenu navedenih klasifikacijskih pravila u praksi. Vidjet ćemo kako nema jedinstevnog odgovora koji je način najbolji za klasifikaciju, a svoje ćemo zaključke temeljiti na pretpostavkama klasifikacijskih pravila i na razinama pogrešaka.

Summary

In this thesis, we will present some methods for data discrimination and classification. At the beginning, we will define criteria for optimal data classification, and later on, we will explain the methods for evaluating classification functions and how to calculate the error rates in classification.

In the second chapter, we will list several methods for classification with two multivariate normal populations in the case when covariance matrices are equal and in the case when they are not. In the fourth chapter, we will deal with classification with several populations. We will also mention Fisher's approach to classification. In the fifth chapter, we will describe logistic regression as another method for data classification and discrimination. In the sixth chapter, we will solve and explain several examples on various data to better understand the application of the mentioned classification rules in practice. We will see that there is no unique answer to which method is best for classification, and we will base our conclusions on the assumptions of classification rules and error rates.

Životopis

Petra Nikolić rođena je 26.10.1997. u Zagrebu. Nakon što je završila osnovnu školu u Svetoj Nedelji i osnovnu glazbenu školu Fedro Livadić u Samoboru, smjer: violina, upisuje opću XI. gimnaziju u Zagrebu i srednju glazbenu školu Fredo Livadić, smjer: teorija. Akademske godine 2016./17. upisuje Preddiplomski studij matematike na matematičkom odsjeku Prirodoslovno-matematičkog fakulteta u Zagrebu, a potom 2020./21. upisuje i Diplomski studij: Matematička statistika na istom odsjeku.

Od veljače 2020. uz studij kreće raditi u hrvatskom *startupu Photomath*, a u ožujku 2023. se u spomenutoj tvrtci zapošljava kao *Business Insights Specialist*.