Modeli miješanih distribucija

Zdilar, Gabriela

PRIRODOSLOVNO–MATEMATIČKI FAKULTET
SVEUČILIŠTA U SPLITU

GABRIELA ZDILAR

MODELI MIJEŠANIH DISTRIBUCIJA

DIPLOMSKI RAD

Split, rujan 2024.

PRIRODOSLOVNO–MATEMATIČKI FAKULTET
SVEUČILIŠTA U SPLITU

ODJEL ZA MATEMATIKU

MODELI MIJEŠANIH DISTRIBUCIJA

DIPLOMSKI RAD

Studentica:
Gabriela Zdilar

Mentorica:
izv.prof.dr.sc. Snježana Braić

Split, rujan 2024.

TEMELJNA DOKUMENTACIJSKA KARTICA

PRIRODOSLOVNO–MATEMATIČKI FAKULTET
SVEUČILIŠTA U SPLITU
ODJEL ZA MATEMATIKU

DIPLOMSKI RAD

MODELI MIJEŠANIH DISTRIBUCIJA

Gabriela Zdilar

Sažetak:

Modeli miješanih distribucija dobro su poznat pojam u statistici u kojoj su se kroz povijest koristili za modeliranje heterogenosti populacije i generalizaciju pretpostavki distribucije. U novije vrijeme, ovi modeli također pružaju pogodan i formalan okvir za grupiranje i klasifikaciju, što su pojmovi usko vezani uz strojno učenje. Glavni cilj ovog rada je pregled modela miješanih distribucija, s posebnim naglaskom na Gaussov i Bernoullijev model miješanih distribucija. Čitatelj će biti detaljno upoznat s EM-algoritmom koji je ključan za procjenu parametara modela miješanih distribucija, a primijenit ćemo ga na spomenute Gaussove i Bernoullijeve modele te dati njegov općeniti oblik. Naposlijetku, EM-algoritam ćemo implementirati u RStudiju na raznim primjerima te analizirati dobivene rezultate.

Ključne riječi:

vjerojatnost, statistika, vjerodostojnost, EM-algoritam

Podatci o radu:

63 stranice, 18 slika, 9 literaturnih navoda, izvornik na hrvatskom jeziku

Mentor(ica): *izv.prof.dr.sc. Snježana Braić*

Članovi povjerenstva:

doc.dr.sc. Vesna Gotovac Đogaš

TEMELJNA DOKUMENTACIJSKA KARTICA

dr.sc. Ana Perišić, poslijedoktorand

Povjerenstvo za diplomski rad je prihvatilo ovaj rad *10.rujna 2024.*

BASIC DOCUMENTATION CARD

FACULTY OF SCIENCE, UNIVERSITY OF SPLIT
DEPARTMENT OF MATHEMATICS

MASTER'S THESIS

MIXTURE MODELS

Gabriela Zdilar

Abstract:

Mixture models are a well-known concept in statistics, historically used for modeling population heterogeneity, generalizing distributional assumptions, and more recently, providing a convenient yet formal framework for clustering and classification—concepts closely tied to machine learning.

The main objective of this paper is to review mixture models, with a special emphasis on the Gaussian and Bernoulli mixture models. Furthermore, the reader will be thoroughly introduced to the EM algorithm, which is crucial for estimating the parameters of mixture models. We will apply the EM algorithm to the aforementioned Gaussian and Bernoulli models and present its general form. Finally, the EM algorithm will be implemented in RStudio on various examples, and the obtained results will be analyzed.

Key words:

probability,statistics,likelihood,EM-algorithm

Specifications:

63 pages, 18 figures , 9 references, original in Croatian

Mentor: *Associate Professor, PhD Snježana Braić*

Committee:

professor Vesna Gotovac Đogaš

Ana Perišić, PhD

BASIC DOCUMENTATION CARD

This thesis was approved by a Thesis commettee on *September 10,2024.*

Uvod

U analizi podataka često se susrećemo s kompleksnim skupovima podataka koji se ne mogu prikladno opisati jednostavnim statističkim modelima. Klasične distribucije kao što su normalna, Poissonova ili ekponencijalna prikladne su za određene skupove podataka, ali kad imamo podatke koji potječu iz više različitih populacija ili procesa, ove distribucije postaje neadekvatne. Upravo zbog toga miješane distribucije, odnosno modeli miješanih distribucija predstavljaju snažan alat za modeliranje podataka koji dolaze iz heterogenih izvora. Ovi modeli zapravo kombiniraju više jednostavnih distribucija u jednu složenu distribuciju koja može bolje reprezentirati stvarne podatke. Prvo pojavljivanje konačnih miješanih modela u suvremenoj statističkoj literaturi pojavilo se u 19. stoljeću gdje su se koristili za predviđanje izdvojenica [1]. Nedugo nakon toga koristila se mješavina dviju univarijantnih normalnih distribucija za analizu skupa podataka koji je sadržavao omjere duljina čela i tijela za 1000 rakova koristeći metodu momenata (MOM) za procjenu parametara modela. Sada su već poznate brojne mješavine kao što su: mješavine Poissonovih distribucija, mješavine von Mises-Fisherove distribucije te mješavine koje se sastoje od normalnih (Gaussovih) komponenti.

Budući da je za razumijevanje ovih koncepata bitno poznavanje osnovnih pojmova vjerojatnosti i statistike, u prvom dijelu rada bit ćete upoznati s

[1] outliers

potrebnom teorijom iz tih područja. U drugom dijelu uvest ćemo matematički model miješanih distribucija te se posvetiti EM-algoritmu za procjenu parametara različitih modela, njegovoj konvergenciji i opravdanosti korištenja. U posljednjem dijelu primijenit ćemo EM-algoritam na vlastite generirane podatke iz raznih distribucija te na dobro poznati skup podataka Iris koristeći RStudio.

Sadržaj

Poglavlje 1

Teorija vjerojatnosti i statistika

U ovom poglavlju definirat ćemo osnovne pojmove vezane uz teoriju vjerojatnosti i statistiku koji će nam biti temelj za sve pojmove i algoritme koje uvedemo u nastavku rada.
Svi pojmovi i definicije iz ovog poglavlja preuzeti su iz [1],[2] i [3].

1.1 Vjerojatnosni prostor

Definicija 1.1 *Familiju $\mathcal{F}$ podskupova od Ω za koju vrijede svojstva:*

(F1) $\emptyset \in \mathcal{F}$

(F2) Ako je $A \in \mathcal{F}$*, onda je* $A^c \in \mathcal{F}$

(F3) Ako su $A_n \in \mathcal{F}$*,* $n \in \mathbb{N}$*, onda je* $\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$

nazivamo σ***-algebrom skupova*** *na* Ω*, a uređeni par* $(\Omega, \mathcal{F})$ ***izmjerivim prostorom****.*

Iz definicije odmah slijedi da je $\Omega \in \mathcal{F}$, te da je σ-algebra $\mathcal{F}$ zatvorena na prebrojive presjeke i razlike skupova. Također, svaka σ-algebra je ujedno i algebra te se pojmovi algebre i σ-algebre podudaraju ako je Ω konačan.

Definicija 1.2 *Neka je S proizvoljan neprazan skup i $\mathcal{A} \subseteq \mathcal{P}(\mathcal{S})$ familija podskupova od S. Najmanju σ-algebru podskupova od S koja sadrži $\mathcal{A}$ nazivamo* ***σ-algebrom generiranom s $\mathcal{A}$*** *i označavamo $\sigma(\mathcal{A})$.*

Definicija 1.3 *Neka je $\mathcal{U}$ familija otvorenih skupova u $\mathbb{R}$.* ***Borelova σ-algebra*** *na $\mathbb{R}$ je najmanja σ-algebra na $\mathbb{R}$ generirana familijom $\mathcal{U}$ i označavamo je s $\mathcal{B}$, odnosno $\mathcal{B} = \sigma(\mathcal{U})$.*

Definicija 1.4 *σ-algebru generiranu familijom svih otvorenih podskupova u $\mathbb{R}^n$ nazivamo* ***Borelovom σ-algebrom na $\mathbb{R}^n$*** *i označavamo s $\mathcal{B}^n$.*

Definicija 1.5 *Neka su $(X, \mathcal{A})$ i $(Y, \mathcal{B})$ izmjerivi prostori, gdje je $\mathcal{A} \subseteq 2^X$ i $\mathcal{B} \subseteq 2^Y$ σ-algebre, a $f : X \to Y$ funkcija. Funkcija f je* ***izmjeriva*** *u paru σ-algebri $\mathcal{A}$ i $\mathcal{B}$, ili kraće $\mathcal{A} - \mathcal{B}$ izmjeriva, ako je $f^{-1}(B) \in \mathcal{A}$ za svaki $B \in \mathcal{B}$.*

Definicija 1.6 *Neka je $(\Omega, \mathcal{F})$ izmjerivi prostor. Funkciju $P : \mathcal{F} \to \mathbb{R}$ nazivamo* ***funkcijom vjerojatnosti*** *na $\mathcal{F}$ (ili Ω) ako je*

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

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

(P3) Ako su $A_n \in \mathcal{F}$, $n \in \mathbb{N}$ i $A_i \cap A_j = \emptyset$ za sve $i \neq j$, onda je

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n)$$

(σ-aditivnost)

Uređenu trojku $(\Omega, \mathcal{F}, P)$ nazivamo ***vjerojatnosnim prostorom****.*

Elemente σ-algebre $\mathcal{F}$ vjerojatnosnog prostora $(\Omega, \mathcal{F}, P)$ nazivamo **događajima**, a broj $P(A)$, $A \in \mathcal{F}$, je **vjerojatnost događaja** A. Za skup

Ω kažemo da je **prostor elementarnih događaja**. Vjerojatnosni prostor je **diskretan** ako je prostor elementarnih događaja konačan ili prebrojiv skup. Za diskretni vjerojatnosni prostor vrijedi da se domena funkcije vjerojatnosti može proširiti na cijeli partitivni skup, odnosno da se za σ- algebru može uzeti cijeli partitvni skup, pa se takav prostor označava jednostavno (Ω, P).

Neka je $(\Omega, \mathcal{F}, P)$ proizvoljni vjerojatnosni prostor i $A \in \mathcal{F}$ takav da je $P(A) > 0$. Lako se pokaže da je funkcija $P_A : \mathcal{F} \to [0, 1]$ definirana s

$$P_A(B) = P(B|A) = \frac{P(A \cap B)}{P(A)}$$

vjerojatnost. Tu vjerojatnost nazivamo **uvjetnom vjerojatnošću uz uvjet** A, a broj $P(B|A)$ nazivamo **vjerojatnošću događaja** B **uz uvjet da se dogodio događaj** A.

Definicija 1.7 *Neka je $(\Omega, \mathcal{F}, P)$ proizvoljni vjerojatnosni prostor i $A, B \in \mathcal{F}$. Kažemo da su događaji A i B* ***nezavisni*** *ako je $P(A \cap B) = P(A) \cdot P(B)$.*

Iz prethodne definicije se jasno vidi da ako je $A \in \mathcal{F}$ takav da je $P(A) = 0$, tada su događaji A i B nezavisni događaji za svaki $B \in \mathcal{F}$.

Definicija 1.8 *Neka je $(\Omega, \mathcal{F}, P)$ vjerojatnosni prostor i $S = \{A_i \in \mathcal{F} : i \in I\}$ proizvoljna familija događaja. Kažemo da je S* ***familija nezavisnih događaja*** *ako za svaku konačnu podfamiliju $\{A_{i_1}, \ldots, A_{i_n}\} \subseteq S$ vrijedi*

$$P(A_{i_1} \cap \ldots \cap A_{i_n}) = P(A_{i_1}) \cdot \ldots \cdot P(A_{i_n}).$$

Sljedeći pojam i formule koje uvodimo su nam od suštinske važnosti za uvođenje modela miješanih distribucija jer nam omogućavaju da svi ishodi eksperimenta budu obuhvaćeni i da se izračunaju vjerojatnosti različitih događaja, odnosno omogućavaju stvaranje prikladnih statističkih modela.

Definicija 1.9 *Konačnu ili prebrojivu familiju* $\{H_i \in \mathcal{F} : i \in I \subseteq \mathbb{N}\}$ *nepraznih disjunktnih događaja u vjerojatnosnom prostoru* $(\Omega, \mathcal{F}, P)$ *za koju vrijedi*

$$\bigcup_{i \in I} H_i = \Omega$$

nazivamo ***potpunim sistemom događaja*** *na* Ω*.*

O važnim svojstvima potpunog sistema događaja govori sljedeća propozicija.

Propozicija 1.10 *Neka je* $\{H_i \in \mathcal{F} : i \in I \subseteq \mathbb{N}\}$ *potpuni sistem događaja u vjerojatnosnom prostoru* $(\Omega, \mathcal{F}, P)$*. Tada za svaki* $A \in \mathcal{F}$ *vrijedi:*

1. *Formula totalne vjerojatnosti:*

$$P(A) = \sum_{i \in I} P(H_i) \cdot P(A|H_i)$$

2. *Bayesova formula: ako je* $P(A) \neq 0$*, tada za svaki* $i \in I$ *vrijedi:*

$$P(H_i|A) = \frac{P(H_i) \cdot P(A|H_i)}{P(A)}$$

Dokaz.

1. $P(A) = P(A \cap \Omega) = P(A \cap \bigcup_{i \in I} H_i)$ (distributivnost)

   $= P\left(\bigcup_{i \in I}(A \cap H_i)\right)$ (σ-aditivnost)

   $= \sum_{i \in I} P(A \cap H_i)$

   $= \sum_{i \in I} P(H_i) \cdot P(A|H_i)$

2. $P(H_i|A) = \frac{P(H_i \cap A)}{P(A)} = \frac{P(H_i) \cdot P(A|H_i)}{P(A)}$

■

1.2 Slučajne varijable

U ovom poglavlju upoznat ćemo se sa slučajnim varijablama i njihovom ulogom u teoriji vjerojatnosti i u statistici. Slučajne varijable su ključni pojam u analizi slučajnih eksperimenata i procesa. Proučavanjem slučajnih varijabli, možemo razumjeti i modelirati različite slučajne procese koji se javljaju u stvarnom svijetu.

Definicija 1.11 *Neka je* $(\Omega, \mathcal{F}, P)$ *vjerojatnosni prostor. Preslikavanje* $X : \Omega \to \mathbb{R}$ *je* ***slučajna varijabla*** *ako vrijedi* $X^{-1}(\mathcal{B}) \subseteq \mathcal{F}$.

Intuitivno, slučajna varijabla je veličina koja se dobije kao rezultat mjerenja u nekom slučajnom pokusu.

Definicija 1.12 *Kažemo da su slučajne varijable* $X_i : \Omega \to \mathbb{R},\ i = 1, \ldots, n,$ ***nezavisne*** *ako su događaji* $(X_1 \in E_1), \ldots, (X_n \in E_n)$ *nezavisni za svaki* $E_i \in \mathcal{B},\ i = 1, \ldots, n,$ *tj. ako vrijedi*

$$P((X_1 \in E_1) \cap \cdots \cap (X_n \in E_n)) = P(X_1 \in E_1) \cdot \ldots \cdot P(X_n \in E_n).$$

1.2.1 Diskretne slučajne varijable

Definicija 1.13 *Neka je* (Ω, P) *diskretni vjerojatnosni prostor.*
Funkciju $X : \Omega \to \mathbb{R}$ *nazivamo* ***diskretnom slučajnom varijablom***.

Primijetimo da je ova definicija u skladu s definicijom slučajne varijable na općem vjerojatnosnom prostoru. Naime, praslika bio kojeg podskupa od $\mathbb{R}$ će sigurno biti element σ- algebre, odnosno partitivnog skupa od Ω. Nadalje, kako je u diskretnom vjerojatnosnom prostoru prostor elemantarnih događaja Ω konačan ili prebrojiv skup, a slika konačnog ili prebrojivog skupa je opet konačan ili prebrojiv skup, to postoji konačan ili prebrojiv skup $D \subset \mathbb{R}$ takav

da je $P(X \in D) = 1$.

Vrijednosti diskretne slučajne varijable su realni brojevi koji su zapravo rezultati određenih mjerenja, a slika diskretne slučajne varijable je skup svih mogućih vrijednosti takvih mjerenja.

Vjerojatnost da diskretna slučajna varijabla X poprimi vrijednost $a \in \mathbb{R}$ definiramo kao broj

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

Također, vjerojatnost da diskretna slučajna varijabla poprimi vrijednosti iz podskupa $E \subseteq \mathbb{R}$ definiramo kao broj

$$P(X \in E) = P(X^{-1}(E)) = P(\{\omega \in \Omega \mid X(\omega) \in E\}).$$

Opravdanje za takve definicije daje nam sljedeća propozicija.

Propozicija 1.14 *Neka je $X : \Omega \to \mathbb{R}$ diskretna slučajna varijabla. Neka je $\Omega' = X(\Omega)$ i $P' : \mathcal{P}(\Omega') \to [0, 1]$ definirana sa $P'(E) = P(X \in E) = P(X^{-1}(E))$. Tada je P' vjerojatnost.*

Vjerojatnost P' nazivamo **distribucijom** ili **razdiobom** diskretne slučajne varijable X i označavamo s P_X. Svakoj slučajnoj varijabli X na diskretnom vjerojatnosnom prostoru je na jednoznačan način pridružena njezina distribucija, odnosno zakon razdiobe koji se najčešće označava:

$$X \sim \begin{pmatrix} a_1 & a_2 & \cdots & a_n & \cdots \\ p_1 & p_2 & \cdots & p_n & \cdots \end{pmatrix}$$

U prvom retku tablice su sve moguće različite vrijednosti koje slučajna varijabla X poprima, a u drugom vjerojatnosti da slučajna varijabla X poprimi te vrijednosti, to jest $p_i = P(X = a_i)$. Primijetmo da su p_i nenegativni realni

brojevi čija je suma 1. Svi prethodno uvedeni pojmovi su nam potrebni za uvođenje nama važnih pojmova teorije vjerojatnosti: funkcije gustoće i funkcije distribucije.

Definicija 1.15 *Neka je $X : \Omega \to \mathbb{R}$ diskretna slučajna varijabla zadana distribucijom:*

$$X \sim \begin{pmatrix} a_1 & a_2 & \cdots & a_n & \cdots \\ p_1 & p_2 & \cdots & p_n & \cdots \end{pmatrix}$$

Funkcija gustoće *slučajne varijable X (kraće, **gustoća** od X) je funkcija $f_X : \mathbb{R} \to \mathbb{R}$ definirana sa*

$$f_X(x) = P(X = x) = \begin{cases} 0, & \text{ako je } x \neq a_i \text{ za svaki } i \\ p_i, & \text{ako je } x = a_i \text{ za neki } i \end{cases}$$

Neka je $E \subseteq \mathbb{R}$. Tada je

$$P(X \in E) = P(X^{-1}(E)) = P(X^{-1}(E \cap \{a_1, a_2, \ldots\})) = \sum_{a_i \in E} P(X = a_i) = \sum_{x \in E} f_X(x)$$

Definicija 1.16 ***Funkcija distribucije*** *slučajne varijable $X : \Omega \to \mathbb{R}$ je funkcija $F_X : \mathbb{R} \to [0, 1]$ definirana sa*

$$F_X(x) = P(X \leq x) = P(\{\omega \in \Omega : X(\omega) \leq x\}).$$

Funkcija distribucije definirana je za sve realne brojeve i vrijedi

$$F_X(x) = P(X \leq x) = P(\{\omega \in \Omega : X(\omega) \leq x\}) = \sum_{a_i \leq x} p_i = \sum_{y \leq x} f_X(y), \quad x \in \mathbb{R}$$

Često slučajnim pokusom ne mjerimo samo jednu veličinu nego više njih. U tom slučaju vrijednosti koje dobivamo nisu realni brojevi nego uređene n-torke, odnosno elementi prostora $\mathbb{R}^n$. To nas dovodi do definicije slučajnog vektora.

Definicija 1.17 *Neka je (Ω, P) diskretni vjerojatnosni prostor.*
Funkciju $X : \Omega \to \mathbb{R}^n$ nazivamo ***n-dimenzionalnim diskretnim slučajnim vektorom****.*

Važno je napomenuti da su koordinate slučajnog vektora
$X = (X_1, \ldots, X_n) : \Omega \to \mathbb{R}^n$ slučajne varijable $X_i : \Omega \to \mathbb{R}$, $i = 1, \ldots, n$.

1.2.2 Neprekidne slučajne varijable

U ovom poglavlju istražit ćemo neprekidne slučajne varijable i njihovu ulogu u teoriji vjerojatnosti i statistici. Dok smo se u prethodnom dijelu upoznali s diskretnim slučajnim varijablama, sada ćemo se posvetiti neprekidnim slučajnim varijablama koje mogu poprimiti neprebrojivo mnogo vrijednosti iz određenog intervala. Neprekidne slučajne varijable su ključne u analizi različitih kontinuiranih fenomena poput vremena, prostora ili fizičkih svojstava. Proučavanjem ovih varijabli možemo dublje razumjeti distribucije i gustoće vjerojatnosti, očekivanje i varijancu koji se koriste u statističkim analizama.

Definicija 1.18 *Neka je $(\Omega, \mathcal{F}, P)$ vjerojatnosni prostor i $X : \Omega \to \mathbb{R}$ slučajna varijabla. Funkcija distribucije $F_X : \mathbb{R} \to [0, 1]$ slučajne varijable X definirana je kao:*

$$F_X(x) = P(X \leq x) = P\left(\{\omega \in \Omega : X(\omega) \leq x\}\right).$$

Definicija 1.19 *Neka je X slučajna varijabla na vjerojatnosnom prostoru $(\Omega, \mathcal{F}, P)$ i neka je F_X njezina funkcija distribucije. Kažemo da je X* ***neprekidna slučajna varijabla*** *ako postoji funkcija $f_X : \mathbb{R} \to [0, \infty\rangle)$ takva da je $F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt$.*
Za funkciju distribucije F_X neprekidne slučajne varijable X kažemo da je ***apsolutno neprekidna funkcija distribucije****, a nenegativnu realnu*

funkciju f_X nazivamo ***funkcijom gustoće slučajne varijable X*** *ili jednostavno* ***gustoćom*** *od X.*

Teorem 1.20 *Funkcija distribucije F slučajne varijable X je rastuća i neprekidna zdesna na $\mathbb{R}$ i zadovoljava*

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

Neka je $(\Omega, \mathcal{F}, P)$ vjerojatnosni prostor i X slučajna varijabla na Ω. Za $B \in \mathcal{B}$ definiramo funkciju $P_X : \mathcal{B} \to [0,1]$ sa

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

Lako se provjeri da je P_X vjerojatnost. Vjerojatnosni prostor $(\mathbb{R}, \mathcal{B}, P_X)$ nazivamo **vjerojatnosnim prostorom induciran s** X. Na ovaj način, svakoj slučajnoj varijabli X se na prirodan način pridružuje vjerojatnosni prostor $(\mathbb{R}, \mathcal{B}, P_X)$. Za funkciju vjerojatnosti P_X često kažemo da je **vjerojatnost inducirana slučajnom varijablom** X ili da je **zakon razdiobe od** X. Vrijedi

$$P_X((-\infty, x]) = P(X^{-1}((-\infty, x])) = P(\{\omega \in \Omega : X(\omega) \leq x\}) = P(X \leq x)$$
$$= F_X(x), \quad x \in \mathbb{R}$$

gdje je F_X funkcija distribucije slučajne varijable X. Dakle, funkcija distribucije slučajne varijable u potpunosti određuje vjerojatnost induciranu tom slučajnom varijablom.

Propozicija 1.21 *Neka je $f : \mathbb{R} \to \mathbb{R}$ neprekidna funkcija. Funkcija f je gustoća vjerojatnosti neke neprekidne slučajne varijable X ako i samo ako vrijedi:*

1. $f(x) \geq 0, \quad x \in \mathbb{R}$
2. $\int_{-\infty}^{+\infty} f(x)\,dx = 1$

Sve ovo se može poopćiti i na veće dimenzije, to jest na slučajne vektore. Neka je $(\Omega, \mathcal{F}, P)$ vjerojatnosni prostor i $X : \Omega \to \mathbb{R}^n$ slučajni vektor, tj. $X = (X_1, \ldots, X_n)$, gdje su $X_i : \Omega \to \mathbb{R}$, $i = 1, \ldots, n$, slučajne varijable. Za $B \in \mathcal{B}^n$ definiramo funkciju $P_X : \mathcal{B}^n \to [0, 1]$ sa $P_X(B) = P(X^{-1}(B))$. P_X je vjerojatnost na $\mathcal{B}^n$ i zovemo je **zakon razdiobe slučajnog vektora** X. Dakle, svakom n-dimenzionalnom slučajnom vektoru X se na prirodan način pridružuje vjerojatnosni prostor $(\mathbb{R}^n, \mathcal{B}^n, P_X)$ koji zovemo **vjerojatnosni prostor induciran slučajnim vektorom** X.

Definicija 1.22 *Neka je $X : \Omega \to \mathbb{R}^n$ n-dimenzionalni slučajni vektor, $X = (X_1, \ldots, X_n)$.* ***Funkcija distribucije slučajnog vektora*** *X je funkcija $F_X = F : \mathbb{R}^n \to [0, 1]$ definirana sa*

$$F(x) = F(x_1, \ldots, x_n) = P_X((-\infty, x]) = P(X \leq x) = P(X_1 \leq x_1, \ldots, X_n \leq x_n)$$

gdje je $\quad x = (x_1, \ldots, x_n) \in \mathbb{R}^n$.

Definicija 1.23 *Neka su $n, m \in \mathbb{N}$ te $\mathcal{B}^n$ i $\mathcal{B}^m$ σ-algebre Borelovih skupova na $\mathbb{R}^n$ i $\mathbb{R}^m$ respektivno. Za funkciju $g : \mathbb{R}^n \to \mathbb{R}^m$ kažemo da je* ***Borelova funkcija*** *ako je $g^{-1}(B) \in \mathcal{B}^n$ za svaki $B \in \mathcal{B}^m$, tj. ako je $g^{-1}(\mathcal{B}^m) \subseteq \mathcal{B}^n$.*

Neki primjeri Borelovih funkcije su: identiteta, konstanta, proste funkcije (funkcije čija je slika konačan ili prebrojiv skup), stepenaste funkcije, ograničene funkcije...
Nadalje, svaka neprekidna funkcija je Borelova funkcija.

Definicija 1.24 *Neka je $X = (X_1, \ldots, X_n)$ n-dimenzionalni slučajni vektor na vjerojatnosnom prostoru $(\Omega, \mathcal{F}, P)$ i F njegova funkcija distribucije.*

Kažemo da je slučajni vektor X ***neprekidan slučajni vektor*** *ako postoji nenegativna Borelova funkcija*
$f : \mathbb{R}^n \to \mathbb{R}$ *takva da je*

$$F(x) = \int\limits_{\langle -\infty, x \rangle} f(t)\, dt = \int\limits_{-\infty}^{x_1} \dots \int\limits_{-\infty}^{x_n} f(t_1, \dots, t_n)\, dt_1 \dots dt_n, \quad x = (x_1, \dots, x_n) \in \mathbb{R}^n.$$

Funkciju f nazivamo ***funkcijom gustoće slučajnog vektora*** *X.*

1.2.3 Statistički momenti slučajnih varijabli

U ovom poglavlju upoznat ćemo se s osnovnim statističkim momentima koji pružaju važne informacije o razdiobi slučajne varijable. Oni su temeljni alati za razumijevanje i analizu podataka u različitim disciplinama kao što su ekonomija, medicina, psihologija, sociologija, inženjerstvo...

Definicija 1.25 *Neka je X diskretna slučajna varijabla na diskretnom vjerojatnosnom prostoru (Ω, P). Ako red*

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

apsolutno konvergira, onda njegovu sumu zovemo ***matematičkim očekivanjem diskretne slučajne varijable*** *X i označavamo s*

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

Definicija 1.26 *Neka je X apsolutno neprekidna slučajna varijabla na vjerojatnosnom prostoru $(\Omega, \mathcal{F}, P)$ s gustoćom f_X. Tada je njezino* ***očekivanje*** *dano s*

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

ako taj integral postoji.

Za slučajnu varijablu koja ima konačno očekivanje kažemo da je **integrabilna**. Matematičko očekivanje slučajne varijable X možemo interpretirati kao srednju vrijednost slučajne varijable X.

Ako je $g : \mathbb{R} \to \mathbb{R}$ Borelova funkcija i X neprekidna slučajna varijabla s gustoćom f_X, onda je funkcija $g(X) : \Omega \to \mathbb{R}$ također neprekidna slučajna varijabla i vrijedi

$$\mathbb{E}(g(X)) = \int_{-\infty}^{+\infty} g(x) f_X(x)\, dx.$$

Definicija 1.27 *Neka je X slučajna varijabla na vjerojatnosnom prostoru $(\Omega, \mathcal{F}, P)$. $\mathbb{E}(X^r)$ zovemo **r-tim momentom**, a $\mathbb{E}(|X|^r)$ zovemo **r-tim apsolutnim momentom**.*

Dogovorno se uzima $\mathbb{E}(X^0)$=$\mathbb{E}(|X|^0)$=1.

Definicija 1.28 *Neka je X integrabilna slučajna varijabla na vjerojatnosnom prostoru $(\Omega, \mathcal{F}, P)$. Broj $\mathbb{E}((X - \mathbb{E}(X))^r)$ nazivamo **r-tim centralnim momentm** od X, a $\mathbb{E}(|X - \mathbb{E}(X)|^r)$ **r-tim apsolutnim centralnim momentom** od X.*

Definicija 1.29 *Neka je X slučajna varijabla na vjerojatnosnom prostoru $(\Omega, \mathcal{F}, P)$. **Varijanca** od X, koju označavamo s $\mathrm{Var}(X)$ ili σ_X^2, je drugi centralni moment od X. Dakle,*

$$\mathrm{Var}(X) = \mathbb{E}((X - \mathbb{E}(X))^2).$$

Iz definicije se odmah vidi da je varijanca srednje kvadratno odstupanje slučajne varijable X od njenog očekivanja $\mathbb{E}(X)$. Nadalje, vrijedi $\mathrm{Var}(X) \geq 0$. Varijanca zapravo predstavlja mjeru raspršenja vrijednosti slučajne varijable

od njenog očekivanja.

$$\begin{aligned}\mathrm{Var}(X) &= \mathbb{E}((X-\mathbb{E}(X))^2)\\ &= \mathbb{E}(X^2 - 2X\mathbb{E}(X) + (\mathbb{E}(X))^2)\\ &= \mathbb{E}(X^2) - 2\mathbb{E}(X)\cdot\mathbb{E}(X) + (\mathbb{E}(X))^2\\ &= \mathbb{E}(X^2) - 2(\mathbb{E}(X))^2 + (\mathbb{E}(X))^2\\ &= \mathbb{E}(X^2) - (\mathbb{E}(X))^2.\end{aligned}$$

Izraz u posljednjem retku se u teoriji vjerojatnosti često navodi kao definicija varijance.

Drugi korijen iz varijance nazivamo **standardnom devijacijom** od X i označavamo je σ_X.

Jedna od osnovnih numeričkih katakteristika dviju slučajnih varijabli je kovarijanca.

Definicija 1.30 *Neka su X, Y i XY integrabilne slučajne varijable. Broj*

$$Cov(X,Y) = \sigma_{X,Y} = \mathbb{E}((X-\mathbb{E}(X))(Y-\mathbb{E}(Y)))$$

nazivamo ***kovarijancom slučajnih varijabli*** *X i Y.*

Primijetimo da je

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

$$\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$$

$$\mathrm{Cov}(X,X) = \mathrm{Var}(X)$$

Kovarijanca mjeri stupanj linearne povezanosti dviju slučajnih varijabli, odnosno koliko promjena jedne slučajne varijable utječe na promjenu druge. Ako su X i Y nezavisne slučajne varijable, onda je $\mathrm{Cov}(X,Y) = 0$. Analogno kao i prije, svi ovi pojmovi se mogu poopćiti na slučajne vektore.

Definicija 1.31 *Neka je $X = (X_1, \ldots, X_n)$ n-dimenzionalni slučajni vektor na vjerojatnosnom prostoru $(\Omega, \mathcal{F}, P)$.* ***Matematičko očekivanje*** *od X je vektor*

$$\mathbb{E}(X) = (\mathbb{E}(X_1), \ldots, \mathbb{E}(X_n)) \in \mathbb{R}^n.$$

Ako $\mathbb{E}(X)$ postoji, kažemo da je X ***integrabilni slučajni vektor****.*

Kovarijacijska matrica Σ slučajnog vektora X definira se kao

$$\Sigma = \mathbb{E}((X - \mathbb{E}X)(X - \mathbb{E}X)^T).$$

To je kvadratna matrica koja na glavnoj dijagonali ima varijance slučajnih varijabli (jer je $\mathrm{Cov}(X_i, X_i) = \mathrm{Var}(X_i)$), a izvan dijagonale na mjestu (i, j) je kovarijanca para slučajnih varijabli X_i i X_j, tj. $[\Sigma]_{ij} = \mathrm{Cov}(X_i, X_j) = \sigma_{ij}$. Nadalje, ona je i simetrična jer vrijedi

$$[\Sigma]_{ij} = \mathrm{Cov}(X_i, X_j) = \mathrm{Cov}(X_j, X_i) = [\Sigma]_{ji}.$$

Kovarijacijska matrica nam je korisna i ako nas za slučajni vektor $X = (X_1, \ldots, X_n)$ zanima međusobni odnos njegovih koordinata, odnosno međusobni odnos slučajnih varijabli X_i, $i = 1, \ldots, n$.

1.3 Normalna razdioba

Definicija 1.32 *Kažemo da neprekidna slučajna varijabla X ima* ***normalnu*** *ili* ***Gaussovu razdiobu*** *s parametrima $\mu \in \mathbb{R}$ i $\sigma^2 > 0$, i pišemo $X \sim \mathcal{N}(\mu, \sigma^2)$, ako joj je funkcija gustoće dana s*

$$\mathcal{N}(x|\mu, \sigma^2) \equiv f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Lako se vidi da funkcija gustoće normalne razdiobe zadovoljava svojstva funkcije gustoće iz Propozicije 1.21.

$$\mathcal{N}(x \mid \mu, \sigma^2) > 0$$

$$\int_{-\infty}^{+\infty} \mathcal{N}(x \mid \mu, \sigma^2)\, dx = 1$$

Nadalje,

$$\mathbb{E}(X) = \int_{-\infty}^{+\infty} \mathcal{N}(x \mid \mu, \sigma^2) \cdot x\, dx = \mu$$

$$\mathbb{E}(X^2) = \int_{-\infty}^{+\infty} \mathcal{N}(x \mid \mu, \sigma^2) \cdot x^2\, dx = \mu^2 + \sigma^2$$

pa je

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2 = \sigma^2$$

Dakle, parametar μ je očekivanje, a σ^2 varijanca normalne razdiobe. U slučaju da je $\mu = 0$ i $\sigma = 1$, slučajnu varijablu $\mathcal{N}(0, 1)$ nazivamo **jediničnom normalnom razdiobom** i ona nam je posebno korisna u praksi.
U nastavku ćemo definirati normalnu razdiobu u višedimenzionalnom prostoru, za što nam treba sljedeća definicija.

Definicija 1.33 *Za matricu $A \in M_n(\mathbb{R})$ kažemo da je* ***pozitivno definitna*** *ako je $x^T Ax > 0$, za svaki $x \in \mathbb{R}^n$, $x \neq 0$. Ako je $x^T Ax \geq 0$ za svaki $x \in \mathbb{R}^n$, kažemo da je A* ***pozitivno semidefinitna****.*

Svaka pozitivno definintna matrica ima inverz jer su joj svojstvene vrijednosti strogo pozitivne. S druge strane, pozitivno semidefinitna matrica ne mora imati inverz (svojstvene vrijednosti su joj nenegativne). Može se pokazati da je kovarijacijska matrica Σ pozitivno semidefinitna pa po prethodnomu ne mora nužno imati inverz. Posljedica toga je da ako matrica ima neke linearno zavisne retke ili je neki element na dijagonali jednak nuli, Σ neće

imati inverz. Znamo da se na dijagonali matrice Σ nalaze varijance svakog obilježja (varijance slučajnih varijabli). Ako je varijanca nekog obilježja nula, znači da mu je mjera raspršenja jednaka nuli pa je obilježje konstantno te nije korisno jer je jednako za svako opažanje. Što se tiče linearne zavisnosti, nju ćemo imati ako se neko obilježje može predvidjeti iz ostalih.

Definicija 1.34 *Neka je $\mu \in \mathbb{R}^n$ i $\Sigma \in \mathbb{R}^{n\times n}$ pozitivno definitna matrica. Kažemo da slučajni vektor X ima* ***višedimenzionalnu normalnu razdiobu*** *s parametrima μ i Σ, te pišemo $X \sim \mathcal{N}(\mu, \Sigma)$, ako je njezina funkcija gustoće f_X dana s*

$$f_X(x) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right), \quad x \in \mathbb{R}^n,$$

gdje je $|\Sigma|$ determinanta matrice Σ, a $x^\top$ je transponirani vektor x.

Ako je $X \sim \mathcal{N}(\mu, \Sigma)$, onda je μ očekivanje slučajnog vektora X, a Σ njegova kovarijacijska matrica.

1.4 Osnovni pojmovi matematičke statistike

Definicija 1.35 *Neka je $(\Omega, \mathcal{F})$ izmjeriv prostor i $\mathcal{P}$ množina funkcija vjerojatnosti na $(\Omega, \mathcal{F})$. Tada je uređena trojka $(\Omega, \mathcal{F}, \mathcal{P})$* ***statistička struktura****.*

Lako se vidi da ako je $\mathcal{P}$ jednočlana množina, tada je statistička struktura vjerojatnosni prostor.
Množina $\mathcal{P}$ je često parametrizirana na način:

$$\mathcal{P} = \{P_\theta : \theta \in \Theta\}$$

pri čemu je Θ skup vrijednosti parametra θ. Pretpostavljamo da je $\Theta \subseteq \mathbb{R}^m$, $m \in \mathbb{N}$, te da je parametrizacija injektivna, tj. da vrijedi

$$(\forall \theta_1, \theta_2 \in \Theta)\ \theta_1 \neq \theta_2 \implies P_{\theta_1} \neq P_{\theta_2}$$

Neka je $X : \Omega \to \mathbb{R}^d$ slučajni vektor i $(\Omega, \mathcal{F}, \mathcal{P})$ statistička struktura, gdje je $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. Za proizvoljni $\theta \in \Theta$ definirajmo

$$F(x; \theta) = P_\theta(X \leq x), \quad x \in \mathbb{R}^d.$$

Tada je $F(\cdot; \theta) : \mathbb{R}^d \to \mathbb{R}, \quad x \mapsto F(x; \theta)$, funkcija distribucije od X uz vjerojatnost $P_\theta \in \mathcal{P}$. Kažemo tada da X pripada statističkom modelu

$$\mathcal{P}' = \{F(\cdot; \theta) : \theta \in \Theta\}.$$

U primjenama će nam od koristi biti statistička struktura koja ima zakon razdiobe X, a indeksirana je parametrom θ. Za takvu strukturu vrijedi

$$F(\cdot; \theta) \overset{1-1}{\longleftrightarrow} P_\theta,$$

pa možemo poistovjetiti $\mathcal{P}$ i $\mathcal{P}'$. Najčešće će X biti neprekidna ili diskretna slučajna varijabla (vektor) s gustoćom $f(\cdot; \theta)$ pa možemo poistovjetiti i $\mathcal{P}$ i $\{f(\cdot; \theta) : \theta \in \Theta\}$.

Definicija 1.36 ***n-dimenzionalni slučajni uzorak*** *na statističkoj strukturi $(\Omega, \mathcal{F}, \mathcal{P})$ je niz $X_1, X_2, \ldots, X_n$ slučajnih varijabli na $(\Omega, \mathcal{F})$ koje su nezavisne i jednako distribuirane u odnosu na svaku vjerojatnost $P \in \mathcal{P}$.*

Ako je x_i realizacija slučajne varijable X_i, $i = 1, \ldots, n$, tada $(x_1, \ldots, x_n)$ nazivamo **vrijednošću** ili **realizacijom uzorka** $(X_1, \ldots, X_n)$. Broj n označava **dimenziju uzorka**.

Definicija 1.37 ***Statistika*** *na statističkoj strukturi* $(\Omega, \mathcal{F}, \mathcal{P})$ *je svaka slučajna varijabla (vektor)* $T : \Omega \to \mathbb{R}^d$ *za kojeg postoji* $n \in \mathbb{N}$ *i* n*-dimenzionalni slučajni uzorak* $(X_1, \ldots, X_n)$ *na* $(\Omega, \mathcal{F}, \mathcal{P})$ *te izmjerivo preslikavanje* $t : \mathbb{R}^n \to \mathbb{R}^d$ *takvo da je* $T = t(X_1, \ldots, X_n)$.

Primjer 1.38 *Neka je* $X_1, \ldots, X_n$ *slučajni uzorak na statističkoj strukturi* $(\Omega, \mathcal{F}, P)$, *pri čemu su* $X_1, \ldots, X_n$ *slučajne varijable. Tada su statistike:*

(i) Uzoračka aritmetička sredina:

$$\bar{x}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

(ii) Uzoračka varijanca:

$$s_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$$

Cilj nam je na osnovu slučajnog uzorka procijeniti vrijednost parametra $\theta \in \Theta \subseteq \mathbb{R}^m$ ili općenito neke funkcije parametra θ, što pišemo $\tau(\theta) \in \mathbb{R}^d$, gdje je $\tau : \Theta \to \mathbb{R}^d$.

Definicija 1.39 *Neka je* $X = (X_1, \ldots, X_n)$ *slučajni uzorak iz modela* $\mathcal{P} = \{f(\cdot; \theta) : \theta \in \Theta\}$, $\Theta \subseteq \mathbb{R}^m$, *i neka je* $\tau : \Theta \to \mathbb{R}^d$. ***(Točkovni) procjenitelj*** *od* $\tau(\theta)$, $\theta \in \Theta$, *je svaka statistika* $T = t(X) = t(X_1, \ldots, X_n)$ *u* $\mathbb{R}^d$ *gdje je* $t : \mathbb{R}^n \to \mathbb{R}^d$ *izmjerivo preslikavanje.*

No, smisleno je promatrati procjenitelje za koje je, za dovoljno veliki n, $T(\Omega) \in \tau(\Theta)$.

Definicija 1.40 *Procjenitelj* $T = t(X)$ *od* $\tau(\theta) \in \mathbb{R}$ *je* ***nepristran*** *za* $\tau(\theta)$ *ako vrijedi*

$$(\forall \theta \in \Theta) \quad \mathbb{E}_\theta(T) = \tau(\theta).$$

Procjenitelj koji nije nepristran je ***pristran*** *procjenitelj za* $\tau(\theta)$.

Definicija 1.41 *Kažemo da je funkcija $\tau(\theta)$* ***procjenjiva*** *ako postoji barem jedan nepristran procjenitelj za nju.*

1.4.1 Metode procjene distribucijskih parametara

Statističke metode koriste se za procjenu parametara vjerojatnosne distribucije. Obično imamo uzorak iz neke populacije na temelju kojeg želimo opisati cijelu populaciju. Zapravo, želimo procijeniti parametar θ i time dobiti opis populacije danog uzorka. Točnije, ako je X slučajna varijabla čija funkcija distribucija F_X dolazi iz familije distribucija $\tau(\theta)$ na skupu parametara θ, mi želimo procijeniti nepoznate parametre distribucije F_X.
Najpoznatije metode za procjenu parametara su metoda maksimalne vjerodostojnosti i Bayesova procjena.

Metoda maksimalne vjerodostojnosti

Definicija 1.42 *Neka je $X = (X_1, \ldots, X_n)$, $n \in \mathbb{N}$, slučajni uzorak iz modela $P = \{f(\cdot;\theta) : \theta \in \Theta\}$ i $x = (x_1, \ldots, x_n)$ neka njegova realizacija. Tada je* ***vjerodostojnost*** *funkcija definirana s*

$$L : \Theta \to \mathbb{R}, \quad L(\theta) = L(\theta|x) = f(x;\theta) = \prod_{k=1}^{n} f(x_k;\theta)$$

Funkcija vjerodostojnosti nam za zadani parametar kaže vjerojatnost uzorka kojeg imamo.
Funkcija f je funkcija gustoće slučajne varijable X pa je, po Propoziciji 1.21, nenegativna. Dakle, možemo izračunati prirodni logaritam funkcije vjerodostojnosti

$$l : \Theta \to \mathbb{R}, \quad l(\theta) = l(\theta|x) = \ln f(x;\theta) = \sum_{k=1}^{n} \ln f(x_k;\theta)$$

Funkciju l nazivamo **log-vjerodostojnošću**.

Definicija 1.43 *Statistika* $\hat{\theta}$ *je* ***procjenitelj maksimalne vjerodostojnosti*** *za* θ *(engl. Maximum Likelihood Estimator - MLE) ako vrijedi*

$$L(\hat{\theta}) = L(\hat{\theta}|X) = \max_{\theta \in \Theta} L(\theta|X).$$

Možemo zamisliti da naš uzorak ima veću vjerojatnost jer se on jedini ostvario, a ostali se nisu realizirali.

Budući da je prirodni logaritam rastuća funkcija, log-vjerodostojnost poprima maksimum u istim točkama kao i vjerodostojnost pa se u praksi ona češće koristi, odnosno tražimo

$$l(\hat{\theta}) = l(\hat{\theta}|X) = \max_{\theta \in \Theta} l(\theta|X) = \max_{\theta \in \Theta} \sum_{k=1}^{n} \log(f(x_k; \theta))$$

Ako radimo s m-parametarskom distribucijom, funkcija vjerodostojnosti ima oblik

$$L(\theta_1, \ldots, \theta_m) \equiv L(\theta_1, \ldots, \theta_m|X) = f(x_1; \theta_1, \ldots, \theta_m) \cdot \ldots \cdot f(x_n; \theta_1, \ldots, \theta_m)$$

te nepoznate parametre dobivamo iz

$$\frac{\partial L(\theta_1, \ldots, \theta_m)}{\partial \theta_i} = 0, \quad i = 1, \ldots, m$$

Bayesova metoda

Bayesova metoda omogućava korištenje prethodnog znanja ili uvjerenja putem a priori vjerojatnosti što je korisno u situacijama s malo podataka ili kada su dostupni različiti izvori informacija. U nastavku ćemo neformalno objasniti pojmove potrebne za razumijevanje Bayesovog teorema. Za formalne definicije i detalje, pogledajte [8].

Neka je X skup svih realizacija opažanja i θ skup parametara koje želimo procijeniti. Funkciju koja opisuje početne vjerojatnosti parametra θ, prije

promatranja podataka, nazivamo **apriorna funkcija gustoće**.

Funkcija apriorne vjerodostojnosti mjeri vjerojatnost promatranih podataka, uz pretpostavku da je parametar θ poznat.

Aposteriorna funkcija gustoće daje ažuriranu vrijednost parametra θ, na osnovu novih podataka X.

Marginalna funkcija gustoće ili samo **funkcija gustoće** je vjerojatnost promatranja podataka X, neovisno o parametru θ. U Bayesovu teoremu osigurava normalizaciju aposteriorne funkcije gustoće.

Apriorno nepoznati skup parametara su elementi o kojima nemamo prethodno znanje ili informacije, ali ih možemo odrediti na osnovu matematičkih pravila, aksioma ili teorema.

Teorem 1.44 ***(Bayesov teorem)*** *Neka je θ apriorno nepoznati skup parametara promatrane distribucije, a X skup mogućih realizacija opažanja. Funkciju gustoće distribucije uvjetno na opažanje računamo formulom*

$$f(\theta|X) = \frac{f(X,\theta)}{f(X)} = \frac{f(X|\theta)f(\theta)}{f(X)}.$$

gdje je $f(\theta)$ apriorna funkcija gustoće, $f(X|\theta)$ funkcija apriorne vjerodostojnosti, $f(X)$ marginalna vjerojatnost, a $f(\theta|X)$ aposteriorna funkcija gustoće.

Vidimo da nazivnik prethodnog izraza ne ovisi o θ pa ga zanemarujemo prilikom analiziranja aposteriorne gustoće, točnije promatramo relaciju proporcionalnosti

$$f(\theta \mid X) \sim f(X \mid \theta)f(\theta),$$

a konstanta proporcionalnosti je marginalna gustoća uzorka koja se računa

$$f(X) = \int_\Theta f(X,\hat{\theta})\, d\hat{\theta} = \int_\Theta f(X \mid \hat{\theta}) f(\hat{\theta})\, d\hat{\theta}$$

Sada ćemo opisati Bayesovu metodu.

Neka je θ nepoznati parametar i $f(\theta)$ njegova apriorna gustoća. Nadalje, označimo s $\hat{\theta}$ procjenitelja od θ.

Funkcija gubitka $\lambda = \lambda(\theta, \hat{\theta})$ mjeri koliko su procjene modela daleko od stvarnih vrijednosti. Očito niža vrijednost funkcije gubitka znači veću točnost procjene. Najčešće korištene funkcije gubitka su apsolutno i srednje kvadratno odstupanje:

$$\lambda(\theta, \hat{\theta}) = \begin{cases} |\theta - \hat{\theta}|, & \text{apsolutno odstupanje} \\ (\theta - \hat{\theta})^2, & \text{srednjekvadratno odstupanje} \end{cases}$$

Bayesova funkcija rizika je očekivanje funkcije gubitka.

Definicija 1.45 *Procjenitelj $\hat{\theta}$ je **Bayesov procjenitelj** ako minimizira Bayesovu funkciju rizika na prostoru parametara Θ.*

1.4.2 Procjena parametara normalne distribucije

U ovom dijelu ćemo procijeniti parametre normalne distribucije jer će nam ta procjena omogućiti precizno modeliranje različitih komponenti mješavine, koje su u središtu zanimanja ovog rada. Koristit ćemo metodu maksimalne vjerodostojnosti za procjenu parametara.

Neka je $X = (X_1, ..., X_n)$ slučajan uzorak iz normalnog modela $\mathcal{N}(\mu, \sigma^2)$, $\theta = (\mu, \sigma^2) \in \mathbb{R} \times [0, \infty\rangle =: \Theta$ te $x = (x_1, \ldots, x_n)$ neka realizacija slučajnog uzorka X. Tada vjerodostojnost ima oblik

$$L(\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2} = \frac{1}{(2\pi)^{n/2} \cdot (\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}.$$

Prirodni logaritam je strogo rastuća diferencijabilna funkcija pa je dovoljno naći globalni ekstrem funkcije log-vjerodostojnosti. Jednostavnim računom

se dobije da je log-vjerodostojnost

$$l(\theta) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 - \frac{n}{2}\ln\sigma^2 - \frac{n}{2}\ln(2\pi)$$

Sada tražimo stacionarne točke.

$$\frac{\partial l}{\partial \mu}(\mu,\sigma^2) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = 0 \Rightarrow \hat{\mu} = \bar{x_n}$$

$$\frac{\partial l}{\partial \sigma^2}(\mu,\sigma^2) = \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i-\mu)^2 - \frac{n}{2\sigma^2} = 0 \Rightarrow \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{n-1}{n}{s_n}^2$$

$$\Rightarrow \hat{\theta}(x) = \left(\bar{x_n}, \frac{n-1}{n}{s_n}^2\right)$$

Vidimo da su procjene za očekivanje i varijancu funkcije od X te se može pokazati da je

$$\mathbb{E}(\hat{\mu}_{\mathrm{MLE}}) = \mu$$

$$\mathbb{E}(\hat{\sigma}^2_{\mathrm{MLE}}) = \frac{n-1}{n}\sigma^2$$

Dakle, vidimo da je procjenitelj μ_{MLE} nepristran procjenitelj srednje vrijednosti populacije dok σ^2_{MLE} nije nepristran procjenitelj varijance. Budući da je $\frac{n-1}{n} < 1$, za svaki $n \in \mathbb{N}$, procjenitelj uvijek podcjenjuje varijancu. Međutim, množenjem σ^2_{MLE} s $\frac{n}{n-1}$ dobivamo nepristranog procjenitelja varijance

$$\sigma^2_{\mathrm{nepr.}} = \frac{n}{n-1}\sigma^2_{\mathrm{MLE}} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\mu_{\mathrm{MLE}})^2.$$

Korekcija se radi samo kad imamo mali uzorak.

Poglavlje 2

Modeli miješanih distribucija i EM-algoritam

2.1 Model konačnih mješavina

Modeli konačnih mješavina su jako važni u statističkoj analizi podataka, procjeni njihove gustoće te grupiranju i klasifikaciji. Koriste se u brojnim područjima, od poljoprivrede, astronomije, bioinformatike, biologije, ekonomije, inženjerstva, genetike pa do medicine, neuroznanosti i raznih alata za obradu slika.

Kao što i sam naziv kaže, ovi modeli se sastoje od konačnog broja distribucija koje ne moraju nužno biti iste.

Zapravo, kao i uvijek u statistici, želimo konstruirati model koji najbolje opisuje podatke (populaciju) koje želimo proučiti. Populacija će često biti podijeljena u manje skupine unutar kojih će podaci biti homogeni, a izvan njih heterogeni. Ponekad ćemo imati informaciju o vjerojatnosnoj distribuciji koju podaci slijede. Cilj ove analize je odrediti komponente, odnosno komponentu kojoj realizacija slučajnog uzorka pripada, te parametre distribucije svake

komponente.

Definicija 2.1 *Neka je $X = (X_1, \ldots, X_r)$ r-dimenzionalni slučajni vektor definiran u vjerojatnosnom prostoru $(\Omega, \mathcal{F}, P)$, $r \in \mathbb{N}$. Uzmimo $N \in \mathbb{N}$ nezavisnih realizacija slučajnog vektora X koje se sastoje od r značajki. Označimo ih s $x = (x_1, \ldots, x_N)$, $x_i = (x_{i_1}, \ldots, x_{i_r}), i = 1, ..., n$. Kažemo da X dolazi iz* ***distribucije konačne mješavine*** *ako njezina vjerojatnosna funkcija gustoće f_X ima formu gustoće mješavine*

$$f_X(x) = \eta_1 f_1(x) + \cdots + \eta_K f_K(x), \quad \text{za svaki } x \in X, \tag{2.1}$$

gdje je $K \in \mathbb{N}$ broj komponenti, a f_k su komponentne gustoće, $k = 1, ..., K$. Vektor $\eta = (\eta_1, \ldots, \eta_K)$ je ***distribucija težina****, gdje je težina komponente K_k označena parametrom $\eta_k, k = 1, ..., K$. Komponente vektora η su nenegativne i normirane, tj. takve da vrijedi*

$$\eta_k \geq 0, \quad \text{za svaki } k \in \{1, \ldots, K\} \tag{2.2}$$

i

$$\eta_1 + \cdots + \eta_K = 1. \tag{2.3}$$

Zbog normiranosti vektora težine dovoljno je procijeniti težinu $K - 1$ komponenti.

Parametar k-te komponente je vjerojatnost da slučajno odabrana jedinka X iz populacije pripada komponenti K_k.

Obično se pretpostavlja da su f_k parametrizirane, odnosno

$$f(x; \vartheta) = \sum_{k=1}^{K} \pi_k f_k(x; \vartheta_k).$$

Ako je već poznat broj komponenti, samo ϑ treba biti procijenjena. Inače, trebamo procijeniti i broj komponenti mješavine.

2.2 Model Gaussove mješavine

U praksi najčešće pretpostavljamo da funkcije gustoće svih komponenti dolaze iz iste familije parametarske distribucije. Ako je u modelu konačne mješavine svaka od komponenti normalno distribuirana, radi se o modelu Gaussove mješavine.
Ovaj model će nam poslužiti kao motivacija za algoritam maksimizacije očekivanja (EM-algoritam).
U nastavku ćemo koristiti oznaku p za funkciju gustoće budući da je to standardna oznaka u literaturi.

Definicija 2.2 *Neka je* $(x_1, \ldots, x_N)$, $N \in \mathbb{N}$, *niz nezavisnih i jednako distribuiranih realizacija slučajnog vektora* X *dimenzije* r, $r \in \mathbb{N}$ *i* $x_i = (x_{i_1}, \ldots, x_{i_r})$, *za svaki* $i \in \{1, \ldots, N\}$. *Kažemo da* X *dolazi iz* ***distribucije modela Gaussove mješavine*** *s* K *komponenti ako je funkcija gustoće od* X *dana sa*

$$p(x|\pi, \mu, \Sigma) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k),$$

gdje je $\pi = (\pi_1, \ldots, \pi_K)$ *takav da je* $\pi_k \geq 0$ *za svaki* $k \in \{1, \ldots, K\}$, $\sum_{k=1}^{K} \pi_k = 1$, $\mu = (\mu_1, \ldots, \mu_K)$, $\Sigma = (\Sigma_1, \ldots, \Sigma_K)$ *te* $\mathcal{N}(x|\mu_k, \Sigma_k)$ *funkcija gustoće normalne razdiobe s očekivanjem* μ_k *i kovarijacijskom matricom* Σ_k *k-te komponente, za svaki* $k = 1, \ldots, K$.

Za $r = 1$ vrijedi $\Sigma_k = \sigma_k^2$ pa model Gaussove mješavine ima oblik

$$p(x \mid \pi, \mu, \sigma^2) = \pi_1 \mathcal{N}(x \mid \mu_1, \sigma_1^2) + \cdots + \pi_K \mathcal{N}(x \mid \mu_K, \sigma_K^2).$$

Neka je $\{x_1, \ldots, x_N\}$, $N \in \mathbb{N}$, skup nezavisnih realizacija slučajnog vektora X te pretpostavljamo da dolaze iz modela Gaussove distribucije. Svaka realizacija ima r obilježja, tj. $x_i = (x_{i_1}, \ldots, x_{i_r})$.

2.2. Model Gaussove mješavine

Uvodimo n-dimenzionalan slučajan vektor $Z = (z_1, \ldots, z_N)$, čija je svaka koordinata $z_i = (z_{i_1}, \ldots, z_{i_K})$. Vektor $z_i = (z_{i_1}, \ldots, z_{i_K})$,$i \in \{1, \ldots, N\}$ ima $1 - K$ prikaz u kojem je jedna koordinata z_{i_k}, $k \in \{1, \ldots, K\}$, jednaka 1, a svi ostali elementi su 0.

Pri modeliranju konačnih mješavina, osim procjene parametara modela, želimo odrediti kojoj komponenti pripada svaka realizacija. Vjerojatnost da podatak x_i pripada komponenti k je koeficijent mješavine k-te komponente π_k, odnosno

$$p(z_{i_k} = 1) = \pi_k, i \in \{1, \ldots, N\}$$

.

Sada ćemo pokazati kako metodom maksimalne vjerodostojnosti procijeniti parametre modela Gaussove mješavine.
Log-vjerodostojnost podataka koji imaju Gaussov model mješavine dana je s

$$\ln p(x \mid \pi, \mu, \Sigma) = \sum_{n=1}^{N} \ln \left(\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n \mid \mu_k, \Sigma_k) \right).$$

Valja primjetiti da ne možemo primijeniti standardnu maksimizaciju funkcije log-vjerodostojnosti zbog prisutnosti singularnosti. Primjerice, to se može dogoditi jer j-ta komponenta ima očekivanje točno jednako jednom od podataka, to jest $\mu_j = x_n$, za neke $j \in \{1, \ldots, K\}$ i $n \in \{1, \ldots, N\}$. Tada vrijedi $N(x_n|x_n, \sigma_j^2 I) = \frac{1}{(2\pi)^{1/2}} \frac{1}{\sigma_j}$ pa u slučaju da $\sigma_j \to 0$, prethodni izraz ide u beskonačnost. Ove poteškoće se ne javljaju kada koristimo Bayesov pristup. Dakle, primjenom maksimalne vjerodostojnosti na Gaussove mješavine, moramo izbjeći prethodnu situaciju i tražiti lokalne maksimume koji se dobro ponašaju. Singularnost bismo trebali izbjeći korištenjem lokalnih heuristika, primjerice možemo primjetiti koja komponenta je problematična, postaviti podatke na nasumične vrijednosti te nastaviti daljnju optimizaciju.

Budući da će mješavina koja se sastoji od K komponenti imati $K!$ ekvivalentnih rješenja za bilo koji slučaj, za svaku točku će postojati $K! - 1$ točaka koje dovode do iste distribucije. Ovaj problem zove se **problemom raspoznatljivosti** jer različiti skupovi parametara neće dati različite distribucije. Formalnu definiciju ovog pojma možete naći u [8].

2.3 EM-algoritam za model Gaussovih mješavina

Latentna ili **skrivena varijabla** je slučajna varijabla čije realizacije ne možemo direktno promatrati ili mjeriti, ali pretpostavljamo njezino postojanje da bismo objasnili manifestne varijable. Model koji koristi latentne varijable za objašnjavanje realizacija naziva se **modelom latentnih varijabli**.

Manifestna ili **opažena varijabla** je slučajna varijabla čije realizacije možemo promatrati u uzorku i koja ukazuje na prisustvo neke latentne varijable. Latentne varijable često se koriste u modelima kako bi objasnile zajedničke varijacije među manifestnim varijablama.

Algoritam maksimizacije očekivanja ili EM-algoritam [1] je moćna metoda za pronalaženje rješenja maksimalne vjerojatnosti za modele s latentnim varijablama. Prvo ćemo prikazati relativno neformalni prikaz za model Gaussovih mješavina.

Dakle, želimo maksimizirati funkciju log-vjerodostojnosti

$$\ln p(x \mid \pi, \mu, \Sigma) = \sum_{i=1}^{N} \ln \left(\sum_{k=1}^{K} \pi_k N(x_i|\mu_k, \Sigma_k) \right)$$

.

[1] Expectation-Maximization

Deriviranjem $\ln p(X|\pi, \mu, \Sigma)$ po srednjoj vrijednosti μ_k i izjednačavanjem izraza s nulom dobivamo

$$0 = -\sum_{n=1}^{N} \frac{\pi_k N(x_n|\mu_k, \Sigma_k)}{\sum_j \pi_j N(x_n|\mu_j, \Sigma_j)} \Sigma_k (x_n - \mu_k)$$

Množenjem sa Σ_k^{-1} (za koju pretpostavljamo da je nesingularna) te uređivanjem dobivamo

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) x_n$$

pri čemu uvodimo oznake

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk}).$$

i

$$\gamma(z_{nk}) = \frac{\pi_k N(x_n|\mu_k, \Sigma_k)}{\sum_j \pi_j N(x_n|\mu_j, \Sigma_j).}$$

N_k možemo interpretirati kao efektivan broj točaka dodijeljenih klasteru k.

Napomena 2.3 *Neka imamo n podataka $x_1, x_2, \ldots, x_n$ i njihove odgovarajuće težine $w_1, w_2, \ldots, w_n$. Tada se* ***ponderirani prosjek*** *računa kao:*

$$\bar{x}_{ponderirani} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

gdje je x_i vrijednost i-tog podatka, w_i težina ili ponder i-tog podatka, a $\sum_{i=1}^{n} w_i$ suma svih pondera.

Primjećujemo da se srednja vrijednost μ_k za k-tu Gaussovu komponentu dobiva uzimanjem ponderiranog prosjeka svih točaka u skupu podataka pri čemu je faktor ponderiranja za podatak x_n određen vrijednošću $\gamma(z_{nk})$ koja određuje da je komponenta k odgovorna za generiranje x_n. Zbog toga se $\gamma(z_{nk})$ naziva **odgovornošću**.

Zatim deriviramo $\ln p(X|\pi, \mu, \Sigma)$ po Σ_k i izjednačimo izraz s 0. Dobivamo

$$\Sigma_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(x_n - \mu_k)(x_n - \mu_k)^T.$$

Vidimo da smo dobili analogan izraz kao kad imamo samo jednu Gaussovu distribuciju.

Na kraju deriviramo izraz i po π_k. Ovdje moramo uzeti u obzir normiranost koeficijenata mješavine. Korištenjem Lagrangeovih multiplikatora i maksimiziranjem izraza

$$Ł(\pi, \mu, \Sigma) = \ln p(X|\pi, \mu, \Sigma) + \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right)$$

dobivamo

$$0 = \sum_{n=1}^{N} \frac{N(x_n|\mu_k, \Sigma_k)}{\sum_j \pi_j N(x_n|\mu_j, \Sigma_j)} + \lambda.$$

Vidimo ponovno pojavljivanje odgovornosti. Ako sad pomnožimo obje strane s π_k i zbrojimo ih po k te iskoristimo normiranost, dobivamo $\lambda = -N$. Time se eliminira λ. Uređivanjem dobivamo $\pi_k = \frac{N_k}{N}$ i zaključujemo da je mješoviti koeficijent za k-tu komponentu jednak je prosječnoj odgovornosti komponente.

Ovime se dobiva jednostavan iterativni postupak za pronalaženje rješenja maksimalne vjerodostojnosti, odnosno instancu EM-algoritma u posebnom

slučaju modela Gaussovih mješavina.

Prvo odaberemo neke početne vrijednosti za očekivanje, kovarijancu i mješovite koeficijente. Nakon toga ćemo izmjenjivati E i M-korak, pri čemu ćemo u E- koraku koristiti trenutne vrijednosti parametara kako bismo evaluirali odgovornosti definirane izrazom

$$\gamma(z_{nk}) \equiv p(z_{nk} = 1 \mid x_n) = \frac{p(z_{nk} = 1)p(x_n \mid z_{nk} = 1)}{\sum_{j=1}^{K} p(z_{nj} = 1)p(x_n \mid z_{nj} = 1)} = \frac{\pi_k \mathcal{N}(x_n \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n \mid \mu_j, \Sigma_j)}$$

Dobivene vjerojatnosti koristimo u koraku maksimizacije(M-koraku) kako bismo ponovno procijenili očekivanje, kovarijancu i mješovite koeficijente. Vidjet ćemo da svako ažuriranje parametara koje proizlazi iz E-koraka povećava funkciju log-vjerodostojnosti. U nastavku dajemo EM-algoritam.

EM-algoritam za model Gaussovih mješavina

1. Inicijaliziraj μ_k, Σ_k, π_k za svaku od K komponenti modela te izračunaj početnu vrijednost log-vjerodostojnosti.

2. **E-korak:** Izračunaj odgovornosti koristeći trenutne vrijednosti parametara:
$$\gamma(z_{ik}) = \frac{\pi_k \mathcal{N}(x_i \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_i \mid \mu_j, \Sigma_j)}$$
za svaku točku podataka x_i, $k = 1, \ldots, K$.

3. **M-korak:** Izračunaj novu procjenu parametara koristeći izračunate

odgovornosti:

$$\mu_k = \frac{1}{N_k} \sum_{i=1}^{N} \gamma(z_{ik}) x_i,$$
$$\Sigma_k = \frac{1}{N_k} \sum_{i=1}^{N} \gamma(z_{ik})(x_i - \mu_k)(x_i - \mu_k)^T,$$
$$\pi_k = \frac{N_k}{N},$$

pri čemu je $N_k = \sum_{i=1}^{N} \gamma(z_{ik})$.

4. Izračunaj funkciju log-vjerodostojnosti:

$$\ln L(\pi, \mu, \Sigma \mid X) = \sum_{i=1}^{N} \ln \left(\sum_{k=1}^{K} \mathcal{N}(x_i \mid \mu_k, \Sigma_k) \right)$$

te provjeri konvergenciju za log-vjerodostojnost ili za parametre. Ako uvjet zaustavljanja još nije zadovoljen, vrati se na korak 2.

2.4 EM-algoritam za mješavine Bernoullijevih distribucija

U ovom poglavlju upoznat ćemo se s modelom mješavina Bernoullijevih varijabli te ćemo izvesti pripadne izraze za EM-algoritam. Ovaj model je također poznat kao **analiza skrivenih klasa**.

Neka je dan skup $\{x_1, \ldots, x_D\}$, $D \in \mathbb{N}$ nezavisnih Bernoullijevih slučajnih varijabli s parametrima μ_i, $i \in \{1, \ldots, D\}$. . Označimo $x = (x_1, \ldots, x_D)^T$ i $\mu = (\mu_1, \ldots, \mu_D)^T$. Tada vrijedi:

$$p(x \mid \mu) = \prod_{i=1}^{D} \mu_i^{x_i} (1 - \mu_i)^{1 - x_i} \tag{2.4}$$

Srednja vrijednost i matrica kovarijance ove distribucije računaju se kao:

$$\mathbb{E}(x) = \mu$$

$$\Sigma = \text{diag}\left(\mu_1(1-\mu_1), \mu_2(1-\mu_2), \ldots, \mu_D(1-\mu_D)\right)$$

Promotrimo konačnu mješavinu Bernoullijevih distribucija definiranu s

$$p(x \mid \mu, \pi) = \sum_{k=1}^{K} \pi_k p(x \mid \mu_k)$$

pri čemu su $\mu = \{\mu_1, \ldots, \mu_K\}$, $\pi = \{\pi_1, \ldots, \pi_K\}$, i

$$p(x \mid \mu_k) = \prod_{i=1}^{D} \mu_{ki}^{x_i}(1-\mu_{ki})^{1-x_i}.$$

Napomena 2.4 *U prethodnom izrazu, μ_{ki} predstavlja parametar Bernoullijeve distribucije za varijablu x_i u k-toj komponenti modela.*

Može se pokazati da srednja vrijednost i kovarijanca miješane distribucije zadovoljavaju

$$\mathbb{E}(x) = \sum_{k=1}^{K} \pi_k \mu_k$$

$$\Sigma = \sum_{k=1}^{K} \pi_k \left(\Sigma_k + \mu_k \mu_k^T\right) - \mathbb{E}(x)\mathbb{E}(x)^T$$

pri čemu je $\Sigma_k = \text{diag}\left(\mu_1(1-\mu_1), \mu_2(1-\mu_2), \ldots, \mu_D(1-\mu_D)\right)$. Budući da matrica kovarijance Σ više nije dijagonalna, miješana distribucija može obuhvatiti korelacije između varijabli, za razliku od pojedinačne Bernoullijeve distribucije.

Neka imamo skup podataka $X = \{x_1, \ldots, x_N\}$, gdje je svaka komponenta x_i, $i \in \{1, \ldots, N\}$ slučajan vektor dimenzije D s funkcijom gustoće oblika

(2.4).

Funkcija log-vjerodostojnosti za ovaj model dana je s

$$\ln p(X|\mu,\pi) = \sum_{n=1}^{N} \ln \left(\sum_{k=1}^{K} \pi_k p(x_n|\mu_k) \right).$$

Sada izvodimo EM-algoritam za maksimiziranje funkcije vjerodostojnosti za model Bernoullijevih mješavina. Analogno kao u slučaju Gaussovih modela mješavina, uvodimo latentnu(skrivenu) varijablu.

Uvjetna distribuciju komponente x_n, $i \in \{1,\ldots,N\}$ u odnosu na latentnu varijablu z_l, $l \in \{1,\ldots,K\}$ bit će

$$p(x_n|z_l,\mu) = \prod_{k=1}^{K} p(x_n|\mu_k)^{z_{lk}} \tag{2.5}$$

dok je apriorna distribucija latentne varijable z_l

$$p(z_l|\pi) = \prod_{k=1}^{K} \pi_k^{z_{lk}}. \tag{2.6}$$

Koristeći Bayesov teorem te izraze (2.5) i (2.6), dobivamo

$$\ln p(X,Z|\mu,\pi) = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk} \left(\ln \pi_k + \sum_{i=1}^{D} \left(x_{ni} \ln \mu_{ki} + (1-x_{ni}) \ln(1-\mu_{ki}) \right) \right) \tag{2.7}$$

Nakon toga računamo očekivanje log-vjerodostojnosti potpunih podataka s obzirom na aposteriornu distribuciju latentnih varijabli i dobivamo

$$\mathbb{E}_Z(\ln p(X,Z|\mu,\pi)) = \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(z_{nk}) \left(\ln \pi_k + \sum_{i=1}^{D} \left(x_{ni} \ln \mu_{ki} + (1-x_{ni}) \ln(1-\mu_{ki}) \right) \right) \tag{2.8}$$

pri čemu je $\gamma(z_{nk}) = \mathbb{E}(z_{nk})$s obzirom na podatkovnu točku x_n.
U E- koraku se odgovornosti procjenjuju koristeći Bayesov teorem koji nam daje oblik

$$\gamma(z_{nk}) = \mathbb{E}(z_{nk}) = \frac{\sum_{z_{nk}} z_{nk} \left[\pi_k f(x_n|\mu_k)\right]^{z_{nk}}}{\sum_{z_{nj}} \left[\pi_j f(x_n|\mu_j)\right]^{z_{nj}}} = \frac{\pi_k f(x_n|\mu_k)}{\sum_{j=1}^{K} \pi_j f(x_n|\mu_j)}$$

U M-koraku maksimiziramo očekivanu log-vjerodostojnost potpunih podataka s obzirom na parametre μ_k i π. Izjednačavanjem derivacije izraza (2.8) po μ_k s nulom te uređivanjem dobivamo

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) x_n$$

Vidimo da je očekivanje komponente k ponderirana srenja vrijednost podataka, pri čemu su koeficijenti težina odgovarajuće odgovornosti.
Konačno, deriviramo izraz (2.8) po π_k te analogno kao u slučaju Gaussova modela mješavine, korištenjem Lagrangeovih multiplikatora, dobivamo

$$\pi_k = \frac{N_k}{N}$$

Može se primjetiti da je funkcija vjerodostojnosti ograničena odozgo jer vrijedi $0 \leq p(x_n|\mu_k) \leq 1$ pa neće biti singulariteta u kojima funkcija vjerodostojnosti ide u beskonačnost. Postoje točke singulariteta u kojima funkcija vjerodostojnosti ide prema nuli, ali EM-algoritam ih neće pronaći osim ako nije inicijaliziran patološkom početnom točkom(specifična vrijednost koja može dovesti do problema ili neoptimalnih rješenja) jer EM-algoritam uvijek povećava vrijednost funkcije vjerodostojnosti sve dok ne pronađe lokalni maksimum.

2.5 Općeniti EM-algoritam

Algoritam maksimizacije očekivanja(EM-algoritam) je općenita tehnika za pronalaženje rješenja maksimalne vjerodostojnosti za probabilističke modele koji uključuju latentne varijable. U ovom poglavlju vidjet ćemo njegovu općenitu obradu, dokaz da konvergira te da algoritmi izvedeni u prethodnim poglavljima zbilja maksimiziraju funkciju vjerodostojnosti.

Razmotrimo probabilistički model u kojem sve promatrane varijable zajednički označavamo s X, a sve skrivene varijable sa Z. Naš cilj je maksimizirati funkciju vjerodostojnosti danu s

$$p(X|\theta) = \sum_{Z} p(X, Z|\theta).$$

Ovdje zbog jednostavnosti pretpostavljamo da je Z diskretna iako je rasprava identična i u slučaju da Z uključuje neprekidne slučajne varijable ili kombinaciju diskretnih i neprekidnih varijabli (zbroj tada zamijenjenimo odgovarajućom integracijom).

Pretpostavit ćemo da je izravna optimizacija $p(X|\theta)$ teška, ali da je optimizacija funkcije potpune vjerodostojnosti $p(X, Z|\theta)$ znatno lakša.

Neka je q(Z) distribucija definirana nad skrivenim varijablama. Uvodimo oznake

$$L(q, \theta) = \sum_{Z} q(Z) \ln \frac{p(X, Z|\theta)}{q(Z)} \tag{2.9}$$

$$\mathrm{KL}(q\|p) = -\sum_{Z} q(Z) \ln \frac{p(Z|X, \theta)}{q(Z)}. \tag{2.10}$$

Zatim primjenimo Bayesov teorem na izraz $p(Z \mid X, \theta)$ i dobivamo

$$p(Z \mid X, \theta) = \frac{p(X, Z \mid \theta)}{p(X \mid \theta)}. \tag{2.11}$$

Djelovanjem funkcijom ln na izraz (2.11) dobivamo

$$\ln p(X, Z|\theta) = \ln p(Z|X, \theta) + \ln p(X|\theta). \tag{2.12}$$

Konačno, uvrštavanjem izraza (2.12) u (2.9) dobivamo da za bilo koji izbor q(Z) vrijedi sljedeće

$$\ln p(X|\theta) = L(q, \theta) + \mathrm{KL}(q\|p) \tag{2.13}$$

Napomena 2.5 ***Kullback-Lieblerova-divergencija*** *između dvije distribucija q i p općenito se definira kao*

$$KL(q\|p) = \sum_z q(z) \ln \frac{q(z)}{p(z)}$$

pri čemu su q(z) i p(z) vjerojatnosti događaja z prema distribucijama q i p, respektivno.
Očito je $KL(q\|p) \geq 0$ *, a jednakost vrijedi ako i samo ako je* $p(z) = q(z)$.

Iz (2.10) vidimo da je $KL(q\|p)$ Kullback-Leiblerova divergencija između q(Z) i aposteriorne distribucije $p(Z|X, \theta)$. U slučaju da je $q(Z) = p(Z|X, \theta)$, $\mathrm{KL}(q\|p) = 0$.
Iz (2.13) slijedi da je $L(q, \theta) \leq \ln p(X|\theta)$, odnosno $L(q, \theta)$ je donja granica za $\ln p(X|\theta)$.

Dekompoziciju (2.13) možemo koristiti za definiranje EM-algoritma i dokaz da zbilja maksimizira log-vjerodostojnost.
Pretpostavimo da je trenutna vrijednost vektora parametara θ_{old}.
U E-koraku se donja granica $L(q, \theta_{\mathrm{old}})$ maksimizira s obzirom na $q(Z)$ uz fiksiran θ_{old}. Lako se primjeti da vrijednost $\ln p(X|\theta_{\mathrm{old}})$ ne ovisi o $q(Z)$ pa se najveća vrijednost $L(q, \theta_{\mathrm{old}})$ postiže kada Kullback-Leiblerova divergencija nestane, odnosno kada je $q(Z)$ jednaka aposteriornoj distribuciji $p(Z|X, \theta_{\mathrm{old}})$.

U M-koraku distribucija $q(Z)$ se ostavlja fiksnom, a donja granica $L(q, \theta)$ maksimizira se s obzirom na θ kako bi se dobila nova vrijednost θ_{new}. Ovime se povećava donja granica L (osim ako već nije dostigla maksimum), što će nužno uzrokovati povećanje odgovarajuće funkcije log-vjerodostojnosti. Budući da se distribucija q određuje koristeći stare vrijednosti parametara umjesto novih i drži se fiksnom tijekom M-koraka, ona neće biti jednaka novoj distribuciji $p(Z \mid X, \theta_{\text{new}})$ te će postojati nenul KL- divergencija. Zbog toga je porast funkcije log-vjerodostojnosti veći od porasta donje granice.

Ako supstituiramo $q(Z) = p(Z \mid X, \theta_{\text{old}})$ u izraz (9.71), vidimo da nakon E-koraka donja granica poprima oblik:

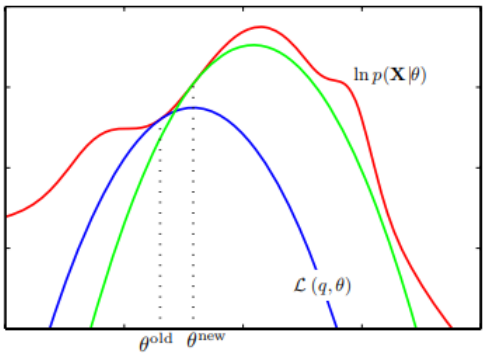
$$L(q, \theta) = \sum_Z p(Z \mid X, \theta_{\text{old}}) \ln p(X, Z \mid \theta) - \sum_Z p(Z \mid X, \theta_{\text{old}}) \ln p(Z \mid X, \theta_{\text{old}})$$

$$= Q(\theta, \theta_{\text{old}}) + \text{const}$$

gdje je konstanta neovisna o θ.

Primijetimo da se varijabla θ, koju optimiziramo u M-koraku, pojavljuje samo unutar logaritma. Ako zajednička distribucija $p(Z, X \mid \theta)$ obuhvaća člana eksponencijalne familije ili produkt takvih članova, tada vidimo da će se logaritam poništiti eksponentom i dovesti do M-koraka koji će obično biti puno jednostavniji od maksimizacije odgovarajuće log-vjerodostojnosti nepotpunih podataka $p(X \mid \theta)$.

U EM-algoritmu krećemo s početnom vrijednošću parametra θ_{old} i u prvom E-koraku procjenjujemo aposteriornu distribuciju latentnih varijabli, što dovodi do donje granice $L(\theta, \theta_{\text{old}})$ čija vrijednost odgovara log-vjerodostojnosti u θ_{old} kako je prikazano plavom krivuljom. Možemo primijetiti da donja granica tangencijalno dodiruje logaritamnu vjerodostojnost u θ_{old} tako da obje krivulje imaju isti gradijent. Ova donja granica je konkavna funkcija



Slika 2.1: Grafički prikaz EM algoritma

koja ima jedinstveni maksimum (za mješavine komponenti iz eksponencijalne familije). U M-koraku, donja granica se maksimizira dajući vrijednost θ_{new} što rezultira većom vrijednošću log-vjerodostojnosti od θ_{old}. Sljedeći E-korak zatim konstruira donju granicu koja je tangencijalna u θ_{new} kako je prikazano zelenom krivuljom.

Za poseban slučaj nezavisnih jednako distribuiranih podataka, X će se sastojati od N podatkovnih točaka x_n, dok će Z obuhvaćati N odgovarajućih latentnih varijabli z_n, $i \in \{1, \ldots, N\}$. Iz pretpostavke o nezavisnosti imamo

$$p(X, Z) = \prod_n p(x_n, z_n)$$

a marginalizacijom(sumiranjem) po $\{z_n\}$ dobivamo

$$p(X) = \prod_n p(x_n)$$

Koristeći pravila o zbroju i produktu, vidimo da a posteriori vjerojatnost koja se procjenjuje u E-koraku ima oblik

$$p(Z|X,\theta) = \frac{p(X,Z|\theta)}{\sum_Z p(X,Z|\theta)} = \frac{\prod_{n=1}^N p(x_n,z_n|\theta)}{\sum_Z \prod_{n=1}^N p(x_n,z_n|\theta)} = \prod_{n=1}^N p(z_n|x_n,\theta).$$

U slučaju Gaussovog modela mješavine, ovo jednostavno znači da odgovornost koju svaka od komponenti mješavine preuzima za određenu podatkovnu točku x_n ovisi samo o vrijednosti x_n i o parametrima θ komponenti mješavine, a ne o vrijednostima drugih podatkovnih točaka.

Vidjeli smo da i E-korak i M-korak EM-algoritma povećavaju vrijednost dobro definirane donje granice na funkciji log-vjerodostojnosti i da će cjelokupan EM-ciklus mijenjati parametre modela da bi se povećala log-vjerodostojnost (osim ako već nije na maksimumu, u tom slučaju parametri ostaju nepromijenjeni).

EM-algoritam dijeli potencijalno težak problem maksimiziranja funkcije vjerodostojnosti na dva koraka, E-korak i M-korak koji obično olakšavaju implementaciju. Ipak, kod složenih modela može se dogoditi da su E-korak, M-korak ili čak oba teško rješiva. To dovodi do dva moguća proširenja EM-algoritma o kojima možete detaljno pročitati u [4], odakle sam i preuzela objašnjenje te opis algoritma.

Poglavlje 3

Implementacija EM-algoritma u RStudiju

U ovom poglavlju istražit ćemo generiranje različitih modela miješanih distribucija u RStudiju. Fokusirat ćemo se na modele miješanih distribucija koje smo u prethodnom poglavlju obradili: Gaussovu(normalnu) miješanu distribuciju i Bernoullijevu miješanu distribuciju. Nakon što generiramo podatke iz ovih distribucija, primijenit ćemo EM-algoritam koristeći gotove funkcije u RStudiju. Nadalje, implementirat ćemo i vlastitu verziju ovog algoritma u RStudiju kako bismo razumjeli pozadinu gotovih funkcija, odnosno izraze koje smo izveli u prethodnom poglavlju.
Na kraju ćemo provesti analogni postupak na skup podataka **Iris**.

3.1 Generiranje miješanih distribucija

U ovom dijelu vidjet ćemo kako se u RStudiju generiraju jednostavni Gaussov i Bernoullijev model miješane distribucije.

Gaussov(normalni) model mješavine

Mješavine distribucija koriste se za modeliranje složenih skupova podataka koji se ne mogu adekvatno opisati jednom jednostavnom distribucijom. U ovom primjeru koristit ćemo mješavinu dviju Gaussovih (normalnih) distribucija kako bismo simulirali podatke koji dolaze iz dviju različitih populacija. Ova tehnika je korisna za testiranje i primjenu EM-algoritma koji je osmišljen za identifikaciju i razdvajanje različitih komponenti unutar mješavine.

Koristit ćemo iduće podatke za generiranje mješavina: prva Gaussova mješavina imat će očekivanje 0 i standardnu devijaciju 1, dok će druga imati očekivanje 5 i standardnu devijaciju 1. Kako bismo osigurali reproducibilnost rezultata, postavit ćemo sjeme (eng.seed) na 123. Koristeći funkciju **rnorm** generiramo slučajne brojeve iz normalne distribucije te će svaka od njih imati 100 točaka.

```
set.seed(123)
gaussova_mjesavina <- c(rnorm(100, mean = 0, sd = 1),
rnorm(100, mean = 5, sd = 1))
head(gaussova_mjesavina)
summary(gaussova_mjesavina)
hist(gaussova_mjesavina, breaks = 30, main = "Histogram
mjesavine Gaussovih distribucija",xlab = "Vrijednosti",
col = "blue", border = "black")
plot(gaussova_mjesavina, main = "Dijagram rasprsenosti
mjesavine Gaussovih distribucija", xlab = "Indeks",
ylab = "Vrijednosti", pch = 19, col = "blue")
```

Naredbom **head** možemo vidjeti nekoliko početnih podataka. Za prikaz osnovne statističke analize koristimo naredbu **summary**.

Vizualizacija podataka je ključna za razumijevanje njihove raspodjele

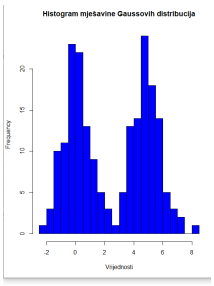
3.1. Generiranje miješanih distribucija

```
[1] -0.56047565 -0.23017749  1.55870831  0.07050839  0.12928774  1.71506499
```

Slika 3.1: Nekoliko početnih podataka

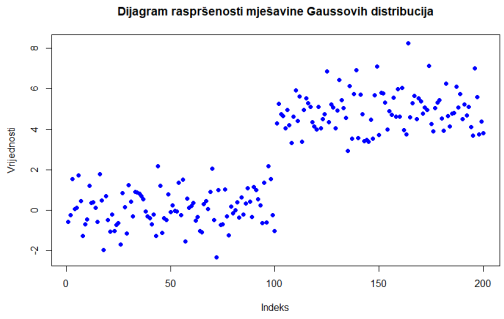
```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-2.30917  0.06613  2.56704  2.49143  4.76895  8.24104
```

i strukture. Koristit ćemo histogram i dijagram raspršenosti kako bismo vizualizirali generirane podatke.



Slika 3.2: Histogram

Iz histograma se jasno mogu uočiti dvije generirane distribucije s različitim očekivanjima.



Slika 3.3: Dijagram raspršenosti

Slično se uočava i iz dijagrama raspršenosti, odnosno jasno se razlikuju dvije nakupine.

Bernoullijev model mješavine

Sad ćemo modelirati Bernoullijev model mješavine, analogno kao u slučaju Gaussovih mješavina. Ponovno ćemo koristiti mješavinu dviju distribucija, samo s različitim parametrima: prva distribucija imat će parametar 0.3, a druga 0.7. Također, svaka od njih će imati po 100 točaka. Dakle, radit ćemo s ukupno 200 točaka.

```
library(flexmix)
set.seed(123)
bernoullijeva_mjesavina <- c(rbinom(100, size = 1, prob = 0.3),
rbinom(100, size = 1, prob = 0.7))
head(bernoullijeva_mjesavina)
summary(bernoullijeva_mjesavina)
hist(bernoullijeva_mjesavina, breaks = 30,
main = "Histogram mjesavine Bernoullijevih distribucija",
xlab = "Vrijednosti", col = "blue", border = "black")
plot(bernoullijeva_mjesavina,
main = "Dijagram rasprsenosti mjesavine Bernoullijevih
distribucija", xlab = "Indeks", ylab = "Vrijednosti", pch = 19,
col = "blue")
```

Paket **flexmix** kojeg učitavamo pogodan je za rad s miješanim modelima i za EM-algoritam. Jedina razlika ovog koda u odnosu na prethodni je naredba **rbinom** koja se koristi za generiranje binomne slučajne varijable. Prvih šest podataka dobivenih naredbom **head** je

```
[1] 0 1 0 1 1 0
```

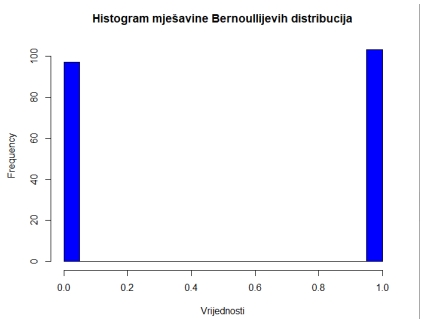
Osnovna statistička analiza dobivena naredbom **summary** dana je na idućoj slici.

3.1. Generiranje miješanih distribucija

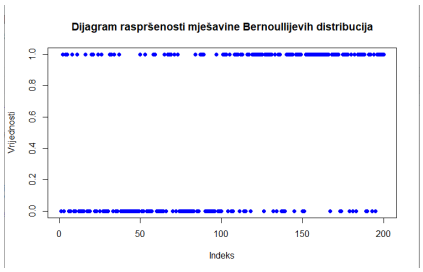
```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.000   0.000   1.000   0.515   1.000   1.000
```

Slika 3.4: Osnovna statistička struktura

Histogram i dijagram raspršenosti nam neće biti od prevelike koristi budući da su nam podaci samo nule i jedinice, ali ćemo ih svakako prikazati.



Slika 3.5: Histogram mješavine Bernoullijevih mješavina



Slika 3.6: Dijagram raspršenosti Bernoullijevih mješavina

3.2 Primjena EM-algoritma na generirane mješavine

U ovom poglavlju primijenit ćemo EM-algoritam na upravo generirane modele mješavina. U prvom dijelu ćemo implementirati vlastite funkcije temeljene na izrazima koje smo dobili u prethodnom poglavlju, a zatim ćemo vidjeti kako se algoritam može jednostavno primijeniti koristeći gotove funkcije u RStudiju.

3.2.1 Vlastite implementacije EM-algoritma

EM-algoritam za Gaussov model mješavine

Funkcija kao argument prima generirane podatke, broj parametara, maksimalni broj iteracija te toleranciju za konvergenciju. Na početku inicijaliziramo parametre: težinu, očekivanje i standardnu devijaciju. Pretpostavljamo da su sve komponente jednako vjerojatne na početku. Nadalje, stvaramo funkciju za izračun log-vjerodostojnosti trenutnih parametara modela. Zatim započinjemo glavnu petlju koja će se izvršavati najviše 100 puta (taj broj smo postavili za maksimalan broj iteracija).

U E-koraku računamo odgovornosti, odnosno vjerojatnosti da svaki podatkovni uzorak pripada pojedinoj komponenti mješavine. Rezultate pohranjujemo u matricu **gamma**. Na kraju, normaliziramo vrijednosti koje se u njoj nalaze.

U M-koraku ažuriramo parametre na temelju odgovornosti izračunatih u E-koraku, a za to ćemo koristiti efektivni broj točaka pridruženih svakoj komponenti.

Konačno, provjeravamo konvergenciju algoritma uspoređivanjem promjene

log-vjerodostojnosti između iteracija.

Implementacija funkcije u RStudiju dana je u nastavku:

```
em_algoritam_gaussov<- function(data, k = 2, max_iter = 100,
tol = 1e-6) {
  n <- length(data)
  # Inicijalizacija parametara
  pi <- rep(1/k, k)
  mu <- sample(data, k)
  sigma <- rep(sd(data), k)
  # Funkcija za izracunavanje log-vjerodostojnosti
  log_likelihood <- function(data, pi, mu, sigma) {
    sum(log(rowSums(sapply(1:k, function(j)
    pi[j] * dnorm(data, mean = mu[j], sd = sigma[j])))))
  }
  log_likelihood_history <- c(log_likelihood(data, pi, mu, sigma))
  for (iter in 1:max_iter) {
    # E-korak: Izracunavanje odgovornosti
    gamma <- sapply(1:k, function(j)
    pi[j] * dnorm(data, mean = mu[j], sd = sigma[j]))
    gamma <- gamma / rowSums(gamma)
    # M-korak: Azuriranje parametara
    N_k <- colSums(gamma)
    pi_new <- N_k / n
    mu_new <- colSums(gamma * data) / N_k
    sigma_new <- sqrt(colSums(gamma * (data - mu_new)^2) / N_k)
    # Provjera konvergencije
    log_likelihood_new <- log_likelihood(data, pi_new,mu_new,
    sigma_new)
    log_likelihood_history <- c(log_likelihood_history,
    log_likelihood_new)
    if (abs(log_likelihood_new - tail(log_likelihood_history, 2)[1])
```

3.2. Primjena EM-algoritma na generirane mješavine

```
    < tol) {
      break
    }
    # Azuriranje parametara za sljedecu iteraciju
    pi <- pi_new
    mu <- mu_new
    sigma <- sigma_new
  }
  list(pi = pi, mu = mu, sigma = sigma,
  log_likelihood_history = log_likelihood_history)
}
# Pokretanje EM algoritma
em_rezultati <- em_algoritam_gaussov(gaussova_mjesavina)
# Ispis rezultata
print(em_rezultati$pi)
print(em_rezultati$mu)
print(em_rezultati$sigma)
print(em_rezultati$log_likelihood_history)
```

Dobiveni rezultati su:

```
> print(em_rezultati$pi)
[1] 0.580604 0.419396
> print(em_rezultati$mu)
[1] 2.636925 2.290008
> print(em_rezultati$sigma)
[1] 2.587523 2.574047
> print(em_rezultati$log_likelihood_history)
  [1] -498.9227 -492.6774 -477.2558 -474.9933 -474.1937 -473.8153 -473.6057 -473.4771 -473.
 [16] -473.1887 -473.1786 -473.1701 -473.1630 -473.1570 -473.1518 -473.1473 -473.1434 -473.
 [31] -473.1248 -473.1234 -473.1222 -473.1210 -473.1199 -473.1190 -473.1181 -473.1173 -473.
 [46] -473.1125 -473.1121 -473.1117 -473.1113 -473.1110 -473.1106 -473.1103 -473.1100 -473.
 [61] -473.1081 -473.1079 -473.1078 -473.1076 -473.1074 -473.1073 -473.1071 -473.1070 -473.
 [76] -473.1061 -473.1060 -473.1059 -473.1058 -473.1057 -473.1056 -473.1055 -473.1055 -473.
 [91] -473.1049 -473.1049 -473.1048 -473.1048 -473.1047 -473.1047 -473.1046 -473.1046 -473.
```

Vidimo da, prema algoritmu, prva komponenta ima otprilike 58% udjela u mješavini, a druga 42% . Nadalje, procijenjeno očekivanje za prvu komponentu je otprilike 2.64, a za drugu 2.29. Također, procjene standardnih devijacija su otprilike 2.59 i 2.57 redom. Vidimo da log-vjerodostojnost postaje stabilna nakon otprilike 20-30 iteracija, što ukazuje na konvergenciju algoritma.

EM-algoritam za Bernoullijev model mješavine

Sada radimo s generiranim Bernoullijevim modelom miješanih distribucija. Funkcija prima iste parametre kao u Gaussovom modelu: generirane podatke, broj komponenti, maksimalni broj iteracija i kriterij konvergencije. Ponovno inicijaliziramo početne težine komponenti i parametara Bernoullijeve distribucije te započinjemo iteracije (ponovno maksimalno 100).

U E-koraku stvaramo matricu s n redaka i k stupaca koja će sadržavati podatke o vjerojatnosti da svaki podatak pripada određenoj komponenti. Funkcijom **dbinom** računamo vjerojatnost za svaki podatak po trenutnim parametrima, te će se u i-tom retku i j-tom stupcu nalaziti ukupna vjerojatnost za i-ti podatak i j-tu komponentu. Na kraju ovog koraka normaliziramo vjerojatnosti.

U M-koraku ponovno ažuriramo parametre distribucije, to jest vektor broja podataka dodijeljenih svakoj komponenti, nove težine komponenti i nove parametre Bernoullijevih distribucija.

Nakon obavljenih koraka, računamo log-vjerodostojnost i provjeravamo konvergenciju, kao i u slučaju Gaussovih modela mješavina. Funkcija kao rezultat vraća listu s procijenjenim težinama komponenti, parametrima Bernoullijeve distribucije i log-vjerodostojnostima.

R-kod je sljedeći:

3.2. Primjena EM-algoritma na generirane mješavine

```
em_algoritam_bernoullijev <- function(data, k, max_iter = 100,
tol = 1e-6) {
  n <- nrow(data)
  d <- ncol(data)
  # Inicijalizacija parametara
  pi <- rep(1/k, k)  # Tezine komponenti
  theta <- matrix(runif(k * d, min = 0.25, max = 0.75), nrow = k,
  ncol = d)  # Parametri Bernoullijevih distribucija
  log_likelihood <- numeric(max_iter)
  for (iter in 1:max_iter) {
    # E-korak: Izracunavanje odgovarajuvih vjerojatnosti
    gamma <- matrix(0, n, k)
    for (j in 1:k) {
      for (i in 1:n) {
        prob <- dbinom(data[i, ], size = 1, prob = theta[j, ])
        gamma[i, j] <- pi[j] * prod(prob)
      }
    }
    gamma <- gamma / rowSums(gamma)
    # M-korak: Azuriranje parametara
    N_k <- colSums(gamma)
    pi <- N_k / n
    for (j in 1:k) {
      theta[j, ] <- colSums(gamma[, j] * data) / N_k[j]
    }
    # Izracunavanje log-vjerodostojnosti
    log_likelihood[iter] <- sum(log(rowSums(gamma)))
    # Provjera konvergencije
    if (iter > 1 && abs(log_likelihood[iter] -
    log_likelihood[iter - 1]) < tol) {
      break
    }
```

```
  }
  list(pi = pi, theta = theta,
  log_likelihood = log_likelihood[1:iter])
} }
```

Sada primijenimo algoritam na naše podatke:

```
bernoullijeva_em_vlastiti <- em_algoritam_bernoullijev
(bernoullijeva_mjesavina_prilagodba, k = 2)
print(bernoullijeva_em_vlastiti)
```

i dobijemo rezultate:

```
$pi
[1] 0.4993893 0.5006107

$theta
          [,1]
[1,] 0.5366643
[2,] 0.4933885

$log_likelihood
[1]  0.000000e+00 -1.076916e-14
```

Iz rezultata vidimo da je model procijenio da oba distribucijska oblika poprilično ravnomjerno doprinose ukupnom modelu, što znamo da je točno. Također, vidimo da je procijenjena vrijednost za parametar θ prve distribucije malo ispod 0.54, a druge skoro 0.5. Konačno, primjećujemo da log-vjerodostojnost započinje praktički s nulom te se smanjuje na vrlo malu negativnu vrijednost, što znači da je model dosegnuo stabilnost, odnosno konvergenciju.

3.2.2 Gotove funkcije za primjenu EM-algoritma

Dosad smo već mogli primijetiti koliko je EM-algoritam važan alat i da se koristi u brojnim područjima znanosti. Upravo zato i ne čudi što većina alata i programskih jezika sadrži gotove funkcije za njegovu primjenu. Mi smo u prethodnom dijelu implementirali vlastite funkcije kako bismo vidjeli pozadinu rada i opravdali izraze koje smo uveli u prethodnom poglavlju, dok ćemo u ovom poglavlju primijeniti gotove funkcije na naše dvije generirane mješavine.

Gotova funkcija za Gaussov model mješavine

Prvo se treba instalirati paket **mclust** koji sadrži funkcije za Gaussov model mješavine. Nakon toga, jednostavno pozovemo funkciju **Mclust** koja kao argument prima našu generiranu Gaussovu mješavinu. Ponovno naredbama **summary** i argumentom **parameters** možemo vidjeti glavne statističke karakteristike modela. R-kod slijedi u nastavku:

```
library(mclust)
gaussian_model <- Mclust(gaussova_mjesavina)
gaussian_model
summary(gaussian_model)
gaussian_model$parameters
```

Prije nego analiziramo rezultate, upoznat ćemo najvažnije mjere koje se koriste za vrednovanje kvalitete modela.

AIC [1]zasniva se na konceptu entropije i mjeri kompromis između točnosti modela i njegove složenosti. Formula mu je

$$\mathrm{AIC} = 2k - 2\ln(L) \tag{3.1}$$

[1]Akaike Information Criterion

pri čemu je k broj parametara modela i L funkcija maksimalne vjerodostojnosti.

BIC [2] je sličan AIC-u, ali dodaje penalizaciju na broj parametara modela, što ga čini strožim kriterijem(pogotovo kad imamo veći skup podataka). Formula je dana s

$$\mathrm{BIC} = \ln(n)k - 2\ln(L) \tag{3.2}$$

pri čemu je n broj podataka kojima raspolažemo, k broj parametara modela i L MLE.

ICL [3] je kriterij posebno koristan kod klaster analize i latentnih promjenjivih modela. On u obzir uzima i vjerojatnost cjelokupnih podataka, uključujući i latentno promjenjive. Formula mu je

$$\mathrm{ICL} = \mathrm{BIC} + \sum_{i=1}^{n}\sum_{j=1}^{k}\gamma_{ij}\log(\gamma_{ij}) \tag{3.3}$$

Ovdje je γ_{ij} a posteri vjerojatnost da objekt i pripada klasteru j.

Za sva tri kriterija vrijedi da su uspješniji što im je vrijednost niža. Naredbom summary dobijemo:

```
----------------------------------------------------
Gaussian finite mixture model fitted by EM algorithm
----------------------------------------------------

Mclust E (univariate, equal variance) model with 2 components:

 log-likelihood   n df       BIC       ICL
      -408.4705 200  4 -838.1342 -839.4361

Clustering table:
  1   2
100 100
```

Vidimo da je vrijednost log-vjerodostojnosti otprilike -404.47, da imamo 200 točaka u podacima(što smo i znali) i 4 stupnja slobode.

[2] Bayesian Information Criterion

[3] Integrated Completed Likelihood

```
$pro
[1] 0.4982444 0.5017556

$mean
          1           2
0.08485736 4.88116089

$variance
$variance$modelName
[1] "E"

$variance$d
[1] 1

$variance$G
[1] 2

$variance$sigmasq
[1] 0.8891601


$vinv
NULL
```

Nadalje, otprilike 50% pripada jednoj i 50% drugoj distribuciji, što znamo da je stvarno stanje stvari. Također, očekivanje prve komponente je otprilike 0.084, a druge 4.881. Možemo primijetiti da za varijancu imamo više parametara. Parametar **modelName** daje vrijednost E, što znači da imamo jednoličnu varijancu. Zbog toga je i parametar **Vinv** NULL. Konačno, broj dimenzija je 1(parametar d), broj komponenti modela 2(parametar G), a procijenjena varijanca je 0.88.

Gotova funkcija za Bernoullijev model mješavine

Ovdje ćemo koristiti već spomenuti paket **flexmix** koji omogućava modeliranje miješanih modela koristeći EM-algoritam. Morat ćemo pretvoriti naše podatke u odgovarajući tip kako bi bio valjan argument funkcije flexmix. Prvo stavljamo formulu koja specificira model kojem se prilagođava y, odnosno odgovor. Vidimo da model prilagođavamo konstanti, što znači da se samo procijenjuje srednja vrijednost ili vjerojatnost uspjeha za svaku komponentu mješavine. Zatim stavljamo prilagođene podatke, broj komponenti te specificiramo model **FLXMCmvbinary()** koji se koristi za binarne podatke u mješavinama.
R kod slijedi u nastavku.

3.2. Primjena EM-algoritma na generirane mješavine

```
data <- data.frame(y = bernoullijeva_mjesavina)
fit <- flexmix(y ~ 1, data = data, k = 2, model = FLXMCmvbinary())
parameters(fit)
summary(fit)
```

Rezultati prethodnih naredbi su:

```
> parameters(fit)
Comp.1.center Comp.2.center
    0.5253759     0.5032051
> summary(fit)

Call:
flexmix(formula = y ~ 1, data = data, k = 2, model = FLXMCmvbinary())

       prior size post>0 ratio
Comp.1 0.532  200    200     1
Comp.2 0.468    0    200     0

'log Lik.' -138.5394 (df=3)
AIC: 283.0788   BIC: 292.9738
```

Vidimo da je procijenjena središnja vrijednost uspjeha prve komponente 0.525, a druge 0.503. Nadalje, a priori vjerojatnosti za komponente su 0.532 i 0.468 redom. Model je zapravo svaki podatak dodijelio objema komponentama, odnosno postoje značajne vjerojatnosti da svaki podatak pripada jednoj ili drugoj komponenti. Ovo ukazuje na loše postavke modela, podatci su zapravo jako slični da bi ih ova funkcija uspješno razdvojila. Za bolje rezultate trebalo bi pokušati primijeniti neku drugu funkciju. Vrijednost log-vjerodostojnosti promatranog modela je -138.5394, uz 3 stupnja slobode. Konačno, model ima visoke vrijednosti AIC i BIC kriterija, što je znak loše prilagođenosti podatcima.

3.3 Primjena EM-algoritma na skup podataka Iris

U prethodnom odlomku smo primijenili gotove i vlastite EM-algoritme na jednostavne modele mješavina koje smo sami generirali. U ovom odlomku ćemo primijeniti EM-algoritam na dobro poznati skup podataka Iris koji je dostupan u RStudiju.

Skup podataka Iris jedan je od najpoznatijih skupova podataka u strojnom učenju i statistici te je često korišten za testiranje i demonstraciju algoritama klasifikacije. Prvi put ga je objavio R.A.Fisher 1936, a sadrži informacije o morfološkim karakteristikama triju različitih vrsta cvijeta iris (setosa, versicolor i virginica). Karakteristike koje sadrži su: dužina čašične liske, širina čašične liske, dužina krunične liske, širina krunične liske i vrsta. Dostupno je 50 podataka svake vrste, odnosno skup podataka Iris sadrži ukupno 150 podataka.

3.3.1 Vlastita implementacija EM-algoritma

Pretpostavit ćemo da podaci potječu iz Gaussovih modela miješanih distribucija, što je samo jedan od mogućih modela za klasteriranje podataka. Prirodno je koristiti Gaussov model mješavine jer Iris sadrži četiri neprekidne varijable koje imaju neku vrstu normalne raspodjele unutar svojih klastera. Algoritam se sastoji od nekoliko temeljnih funkcija koje ćemo detaljno objasniti u nastavku.

Funkcija za inicijalizaciju parametara Gaussovog modela mješavine

Na početku učitavamo dva paketa koja ćemo trebati: paket **MASS** koji sadrži funkciju **mvnorm** za generiranje višedimenzionalnih normalnih distribucija

te paket **scales** koji ćemo koristiti za normalizaciju podataka. Standardno, učitavamo skup podataka Iris i pretvaramo ga u oblik matrice koji će nam biti potreban za daljnje operacije. Zatim funkcijom **scale** normaliziramo podatke. Funkcija **dvnorm** izračunava vrijednost multivarijantne normalne gustoće za dane podatke, srednju vrijednost i kovarijacijsku matricu, odnosno koristi se za procjenu odgovornosti svakog uzorka za svaki klaster.

```
dmvnorm <- function(x, mean, sigma) {
  k <- ncol(sigma)
  sqrt_det_sigma <- sqrt(det(sigma))
  inv_sigma <- solve(sigma)
  const <- 1 / ((2 * pi)^(k / 2) * sqrt_det_sigma)
  exp_term <- exp(-0.5 * rowSums((x - mean) %*% inv_sigma *
  (x - mean))) #Mahalonobisova udaljenost
  const * exp_term
}
```

Sada kreće inicijalizacija parametara, odnosno postavljamo 3 klastera te koristimo **kmeans** algoritam za pronalaženje srednjih vrijednosti klastera. Mogli smo koristiti drugačije metode inicijalizacije, kao što su slučajan odabir parametara, hijerarhijsko klasteriranje ili spektralno klasteriranje, ali smo zbog jednostavnosti i brzine odabrali kmeans. Također, inicijaliziramo kovarijacijsku matricu, težine i odgovornosti.

```
# Inicijalizacija parametara
set.seed(123)
G <- 3  # Povecavamo broj klastera na 3
n <- nrow(X)
d <- ncol(X)
# K-means inicijalizacija
kmeans_result <- kmeans(X, G)
```

```
means <- kmeans_result$centers
# Inicijalne matrice kovarijance
covariances <- array(0, dim = c(d, d, G))
for (g in 1:G) {
  covariances[,,g] <- diag(d)
}
# Inicijalne tezine (proporcije klastera)
weights <- rep(1 / G, G)
# Inicijalne odgovornosti
responsibilities <- matrix(0, n, G)
```

Sada krećemo na implementaciju EM-algoritma. Sljedećom funkcijom računamo vrijednost log-vjerodostojnosti te inicijaliziramo toleranciju za zaustavljanje, maksimalni broj iteracija, brojač iteracija i razliku log-vjerodostojnosti. Iteracije će se izvršavati sve dok ne napravimo maksimalan broj iteracija ili dok algoritam ne konvergira. U E-koraku računamo odgovornosti za svaki uzorak i klaster na temelju trenutnih parametara, dok u M-koraku ažuriramo parametre modela na temelju izračunatih odgovornosti. Također, pratimo promjenu log-vjerodostojnosti kroz svaku iteraciju kako bi se procijenila konvergencija algoritma. Na temelju najveće odgovornosti zaključujemo kojem klasteru svaki podatak pripada. Iz tablice vidimo u koju skupinu, prema algoritmu, pripada pojedini podatak. U RStudiju to izgleda ovako:

```
log_likelihood <- function() {
  sum(log(rowSums(sapply(1:G, function(g) {
    weights[g] * dmvnorm(X, means[g,], covariances[,,g])
  }))))
}
log_likelihoods <- c()
```

```
tol <- 1e-6
max_iter <- 100
iter <- 1
diff <- Inf
while (iter <= max_iter && diff > tol) {
  # E-korak
  for (g in 1:G) {
    responsibilities[, g] <- weights[g] *
    dmvnorm(X, means[g,], covariances[,,g])
  }
  responsibilities <- responsibilities / rowSums(responsibilities)
  # M-korak
  N_g <- colSums(responsibilities)
  weights <- N_g / n
  for (g in 1:G) {
    means[g, ] <- colSums(responsibilities[, g] * X) / N_g[g]
    diffs <- X - means[g, ]
    covariances[,,g] <- t(diffs) %*% (responsibilities[, g] *
    diffs) / N_g[g]
  }
  # Log-vjerodostojnost
  log_likelihoods[iter] <- log_likelihood()
  if (iter > 1) {
    diff <- log_likelihoods[iter] - log_likelihoods[iter - 1]
  }
  iter <- iter + 1
}
# Rezultati klasteriranja
clusters <- apply(responsibilities, 1, which.max)
# Prikaz rezultata
table(iris$Species, clusters)
```

EM-algoritam

Funkcija kao argumente prima ulazne podatke, broj komponenti u modelu Gaussovih mješavina, maksimalan broj iteracija koje će algoritam obaviti (u našem slučaju 100) te toleranciju za provjeru konvergencije.
Prvo inicijaliziramo parametre funkcijom za inicijalizaciju parametara te stavljamo početnu vrijednost log-vjerodostojnosti na nulu. Zatim započinjemo petlju koja ima maksimalno 100 iteracija. U svakoj iteraciji se pozivaju funkcije za E-korak i M-korak te provjeravamo je li zadovoljen uvjet konvergencije. Ako jest, petlja se prekida. U protivnom dodajemo ažuriranu log-vjerodostojnost i nastavljamo postupak. Kao rezultat dobivamo konačne parametre modela i vjerojatnosti pripadanja klasterima. R-kod slijedi u nastavku:

```
em_algoritam_iris <- function(X, K, max_iter = 100, tol = 1e-6) {
  params <- initialize_params(X, K)
  log_likelihood <- 0
  for (i in 1:max_iter) {
    gamma <- expectation_step(X, params)
    params <- maximization_step(X, gamma)
    new_log_likelihood <- sum(log(rowSums(gamma)))
    if (abs(new_log_likelihood - log_likelihood) < tol) {
      break
    }
    log_likelihood <- new_log_likelihood
  }
  return(list(params = params, gamma = gamma))
}
```

Vidimo rezultate na idućoj slici. Svi uzorci vrste Setosa ispravno su klasificirani

```
           clusters
             2  3
setosa      50  0
versicolor  45  5
virginica   35 15
```

u klaster 2. Većina uzoraka Versicolor je klasificiran u klaster 2. Također je većina Virginca uzoraka u drugom klasteru. Versicolor i Virginica su sličnije jedna drugoj pa je razumljivo da bi njihovo razlikovanje moglo biti teže. Kod EM-algoritma klasteri se temelje na gustoći podataka, što može rezultirati različitim brojem klastera od očekivanog. Gotova funkcija također razdvaja podatke u dvije klase, osim ako joj ne pošaljemo parametar kojim naglašavamo drugačije. Razlog zašto je algoritam odvojio podatke u dvije umjesto u tri klase je vjerojatno pretpostavka da je ovo Gaussov model mješavine. Za bolje rezultate trebale bi se provesti detaljnije analize podataka.

3.3.2 Korištenje gotove funkcije za EM-algoritam

U ovom dijelu ćemo ponovno primijeniti paket **mclust** i njegove ugrađene funkcije. Učitavamo Iris podatke i odbacujemo zadnji stupac, budući da su u njemu podaci koje trebamo predvidjeti. Funkcija koju ćemo koristiti je **Mclust** koja automatski odabire najbolji model. Znamo da postoje 3 klastera pa smo to poslali kao parametar funkciji. Rezultat funkcije je objekt koji sadrži rezultate klasteriranja.

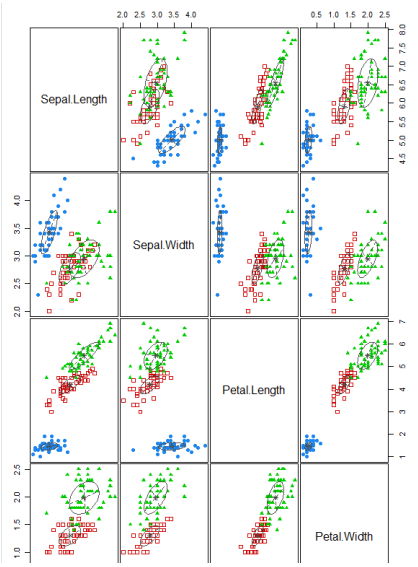
```
1 library(mclust)
2 data(iris)
3 X <- iris[, 1:4]
4 model <- Mclust(X,G=3)
5 summary(model)
6 plot(model, what = "classification")
7 plot(model, what = "uncertainty")
```

Osnovni statistički podaci dobiveni naredbom **summary** su:

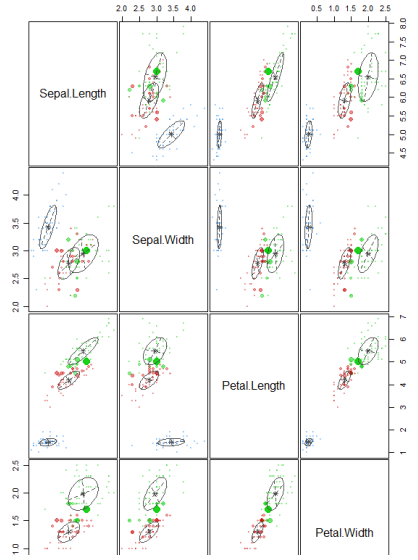
```
----------------------------------------------------
Gaussian finite mixture model fitted by EM algorithm
----------------------------------------------------

Mclust VEV (ellipsoidal, equal shape) model with 3 components:

 log-likelihood   n df       BIC       ICL
       -186.074 150 38 -562.5522 -566.4673

Clustering table:
 1  2  3
50 45 55
```

Iz rezultata se može vidjeti da imamo Model VEV, odnosno svaki klaster ima jednak oblik, s različitom veličinom (volumenom) i orijentacijom. Dobili smo negativnu vrijednost log-vjerodostojnosti, što je očekivano jer je logaritamska funkcija za vjerojatnosti manje od 1 uvijek negativna. BIC i ICL kriteriji nam daju dobre i slične rezultate pri čemu BIC koristi više parametara, a ICL penalizira složenost modela radi boljeg razdvajanja klastera. Model je odabran na temelju ICL kriterija (ima nešto nižu vrijednost od BIC-a). Vidimo da je model identificirao 3 klastera, kao što smo mu zadali. Nadalje, prvom klasteru dodijeljeno je 50 elemenata, drugom 45 i trećem 55.
Klasifikacija podataka prema pronađenim klasterima je sljedeća:

3.3. Primjena EM-algoritma na skup podataka Iris



Također, možemo prikazati i nesigurnost klasifikacije za svaki podatak:



Zaključak

U ovom radu cilj je bio shvatiti potrebu definiranja modela miješanih distribucija te detaljno upoznati EM-algoritam koji je najvažniji algoritam za procjenu parametara modela.

Kako bi čitatelj bolje razumio općeniti EM-algoritam, prvo se demonstrirala njegova primjena na jednostavne modele Bernoullijevih i Gaussovih mješavina. Zatim smo izveli općeniti EM algoritam, dokazali njegovu konvergenciju i opravdanost korištenja. U zadnjem dijelu smo algoritam primijenili na razne slučajeve i primjere iz RStudija. Tu smo vidjeli da su rezultati varirali jer na uspjeh procjene parametara može utjecati mnogo čimbenika, najviše incijalizacija. Također, pokazane su gotove funkcije za EM-algoritam koje se u praksi koriste i koje u obzir uzimaju sve parametre te izabiru one koji daju najbolje rezultate.

Time je čitatelj dobio podlogu za daljnje proučavanje ove kompleksne teme i mogućnost eksperimentiranja s mnogim faktorima kako bi se u konkretnim problemima dobilo najbolje rješenje.

Zahvala

Još uvijek nerealno zvuči da sam nakon pet godina na Prirodoslovno-matematičkom fakultetu na kraju svog studentskog putovanja. Teško se ne sjetiti svih kolegija, rada, učenja i ispita. Toliko prepreka, od kojih su mi se mnoge u tom trenutku činile nepremostivima, sada je iza mene. Da mi je netko tada rekao da će mi sve te brige faliti, iako još službeno nisu ni završile, ne bih vjerovala.

Prije svega, želim zahvaliti svojoj dragoj mentorici na velikoj pomoći i potpori u cijelom ovom procesu. Bili ste jedan od prvih profesora s kojima sam se susrela na fakultetu te ste mi od samih početaka pokazali da je moguće učiti u opuštenoj i poticajnoj atmosferi, uvijek dostupni za sva pitanja. Vi ste jedan od razloga zašto studenti vole matematiku. Hvala Vam na svemu!

Ova diploma jednako je moja kao i cijele moje obitelji. Dragi mama, tata i brate, hvala vam na potpori koju mi pružate cijeli život, u svakom izazovu i problemu. Vi ste oni kojima prvima javljam sve dobre i loše vijesti, koji imate strpljenja i razumijevanja za mene i kad ih nitko drugi nema te mi olakšavate svaku životnu situaciju.

Hvala mojim dugogodišnjim prijateljima koji su uz mene bili u raznim situacijama, s kojima sam skupa rasla i razvijala se te postala osoba koja sam danas.

Posebna zahvala svim kolegicama i kolegama koje sam upoznala na ovom

fakultetu koji su mi, s ponosom mogu reći, postali prijatelji za cijeli život. Uz vas je svaka prepreka bila lakša i jedan ste od razloga što sam toliko voljela ići na fakultet.

Na kraju, hvala svim zaposlenicima i zaposlenicama Prirodoslovno-matematičkog fakulteta u Splitu, od profesora do teta u referadi. Ovaj fakultet bio mi je dom pet godina i da opet biram, ponovno bih ga izabrala.

Sada, s diplomom u ruci i uspomenama koje će mi zauvijek ostati, spremna sam za neke nove životne i poslovne stranice, ali studentsko razdoblje i PMF će uvijek imati posebno mjesto u mom srcu.

Hvala!

Literatura

[1] Sarapa, N., Teorija vjerojatnosti, Školska knjiga, Zagreb, 2002.

[2] Braić, S., Uvod u vjerojatnost i statistiku, 2020.

[3] Elezović, N., Matematička statistika, Stohastički procesi, Element, Zagreb, 2007.

[4] Bishop, C. M., Nasrabadi, N. M. (2006). Pattern recognition and mac hine learning (Vol. 4, No. 4, p. 738). New York: springer.

[5] McLachlan, G. and Peel, D. (2000). Finite Mixture Models. John Wiley and Sons, Inc., New York.

[6] Melnykov, V., Maitra R., Finite mixture models and model-based clustering, Statist. Surv. 4 80 - 116, 2010.

[7] Gotovac Dogaš, V.,Statistika-Skripta, 2024.

[8] Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. 3rd ed. Boca Raton: Chapman and Hall/CRC, 2013.

[9] [12] McLachlan, G. J., Lee, S. X., Rathnayake, S. I. (2019). Finite mixture models. Annual review of statistics and its application, 6, 355-378.