Modeli informacijske geometrije u analizi medijskog sadržaja

Prpić, Monika

Modeli informacijske geometrije u analizi medijskog sadržaja

Prpić, Monika

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Monika Prpić

MODELI INFORMACIJSKE GEOMETRIJE U ANALIZI MEDIJSKOG SADRŽAJA

Diplomski rad

Voditelj rada:
doc. dr. sc. Mario Bukal

Suvoditelj rada:
izv. prof. dr. sc. Boris Muha

Zagreb, ožujak 2021.

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

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

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Sadržaj

Uvod

Ljudima je vrlo prirodno grupirati različite pojave u skupine koje dijele neka zajednička svojstva kako bi se lakše snalazili te tu intuiciju prenosimo i na podatke. Stoga ne iznenađuje da je velik dio strojnog učenja posvećen upravo klasifikacijskim problemama.

Danas imamo pristup velikim količinama podataka iz raznih područja primjene te svaki tip podataka koji možemo modelirati ima određene karakteristike koje nije moguće obuhvatiti nekim općenitim modelom pa se javlja potreba za kontinuiranim razvojem novih metoda modeliranja. Kako nam je jezik i dalje primarno komunikacijsko sredstvo, očekivano je da se velik broj klasifikacijskih problema bavi baš tekstualnim podacima. Oni su često problematični za modeliranje budući da posjeduju neka otežavajuća svojstva poput visoke dimenzionalonosti i rijetke reprezentacije (engl. *sparsity*), što predstavlja dodatni izazov za izgradnju prikladnih modela.

Cilj ovog rada je prezentirati algoritam za klasifikaciju podataka koji je baziran na konveksnoj kombinaciji von Mises-Fisherovih distribucija. Rad je podijeljen u 4 poglavlja.

U prvom poglavlju iznosimo osnovne rezultate informacijske geometrije pomoću kojih dobivamo alate za proučavanje geometrijske strukture parametarskih familija vjerojatnosnih distribucija. Ovdje nam je od posebnog interesa klasa eksponencijalnih familija budući da ona obuhvaća većinu poznatih parametarskih distribucija. Drugi dio bavi se von Mises-Fisherovom distribucijom koja nam služi kao model za podatke koji imaju smjer, kao što je slučaj i kod tekstualnih podataka. Tu najprije iznosimo njezinu definiciju i osnovna svojstva, nakog čega dajemo procjenu parametara metodom maksimalne vjerodostojnosti. Ponekad podatke nije moguće adekvatno modelirati pomoću jedne von Mises-Fisherove distribucije pa u tu svrhu u trećem poglavlju poopćujemo prethodno dane rezultate te proučavamo konveksnu kombinaciju von Mises-Fisherovih distribucija. Ovdje ćemo predstaviti i algoritam maksimizacije očekivanja pomoću kojega procjenjujemo parametre navedenog modela. Četvrto poglavlje posvećeno je primjerima koji pokazuju kako taj algoritam funkcionira u praksi. Najprije potvrđujemo njegovu korektnost na simuliranim podacima te naposljetku dajemo nekoliko primjera koji se bave klasifikacijom teksta.

Poglavlje 1

Uvod u informacijsku geometriju

Informacijska geometrija je interdisciplinarno područje koje proučava geometrijsku strukturu parametarskih familija vjerojatnosnih distribucija. Naime, familije parametarskih vjerojatnosnih distiribucija $\{p(x;\theta) : \theta \in \Theta\}$ možemo promatrati kao mnogostrukosti te se u analizi takvih modela koriste tehnike diferencijane geometrije. Fundamentalni doprinos dao je Rao [12] koji je prvi predložio takav pristup te opskrbio statistički model strukturom Riemannove mnogostrukosti, uzimajući Fisherovu informacijsku matricu kao Riemannovu metriku. Nadalje, u kontekstu informacijske geometrije posebno je zanimljiva klasa eksponencijalnih familija [8] te stoga njima posvećujemo prvu točku ovog poglavlja.

1.1 Eksponencijalne familije

Za razumjevanje eksponencijalnih familija potreban nam je naprije pojam *dovoljne statistike*, koji je jedan od fundamentalnih statističkih pojmova. Ideja je da bi ta statistika trebala sadržavati sve informacije o nepoznatim parametrima koje pruža promatrani uzorak te nam tako omogućuje redukciju podataka bez gubitka informacije. Iskažimo to formalno.

Definicija 1.1.1. *Neka je $(X_1, \dots X_n)$ slučajni uzorak iz modela $\mathcal{P} = \{f(\cdot;\theta) : \theta \in \Theta\}$ i $T = t(X_1, \dots X_n)$, $t : \mathbb{R}^n \to \mathbb{R}^k$, statistika. Kažemo da je T **dovoljna statistika** za familiju $\mathcal{P}$ ako za svaki $y \in \mathbb{R}^k$ uvjetna distribucija slučajnog vektora $X = (X_1, \dots X_n)$ uz dano $T = y$ ne ovisi o θ.*

Primjetimo da dovoljnu statistiku nije lako odrediti primjenom gornje definicije te nam u tome pomaže sljedeći teorem [10] koji daje dekompoziciju funkcije gustoće iz koje je onda lako iščitavamo.

Teorem 1.1.2 (Neyman-Fisher)**.** *Statistika $T = t(X)$ je dovoljna za familiju $\mathcal{P}$ ako i samo ako postoje nenegativne funkcije g_θ i h takve da se gustoća $f(x;\theta)$ slučajnog uzorka X može*

faktorizirati na sljedeći način

$$f(x;\theta) = g_\theta(t(x))h(x). \tag{1.1}$$

Klasa eksponencijalnih familija je značajna budući da obuhvaća distribucije koje, pod određenim uvjetima, dopuštaju pojednostavljenje informacije dane uzorkom tako da ukupna informacija ostane ista.

Definicija 1.1.3. *Neka je $\mathcal{X}$ promatrani prostor primjera. Model $\mathcal{P} = \{f(\cdot;\theta) : \theta \in \Theta\}$ zovemo* ***eksponencijalna familija*** *ako je odgovarajuća gustoća $f(\cdot;\theta)$ oblika*

$$f(x;\theta) = \exp\left(t(x)\cdot\theta - F(\theta) + k(x)\right), \tag{1.2}$$

gdje je

- $\theta \in \Theta$ *prirodni parametar,*
- $t(x)$ *dovoljna statistika,*
- $F(\cdot)$ *log-normalizator,*
- $k(x)$ *pripadna mjera.*

Kada kažemo pripadna mjera, mislimo da je $k(x)$ funkcija takva da je $dP(x) = k(x)d\mu(x)$ i P mjera koja je apsolutno neprekidna s obzirom na $\mu(x)$, što je obično Lebesgueova (u slučaju neprekidnih distribucija) ili ili brojeća (u diskretnom slučaju) mjera, te je (1.2) funkcija gustoće s obzirom mjeru $dP(x)$.

Eksponencijalne familije su određene svojim log-normalizatorom, koji ime duguje činjenici da osigurava da su sve funkcije gustoće $f(\cdot;\theta)$ normalizirane, tj. da vrijedi

$$\int f(x;\theta)d\mu(x) = 1. \tag{1.3}$$

Prema tome, log-normalizator $F(\cdot)$ je dan s

$$F(\theta) = \log \int \exp\left(t(x)\cdot\theta + k(x)\right)d\mu(x). \tag{1.4}$$

Definicija 1.1.4. *Skup $\mathcal{N}_\theta = \left\{\theta : \int \exp\left(t(x)\cdot\theta + k(x)\right)d\mu(x) < \infty\right\}$ zovemo* ***prostor prirodnih parametara****.*

Red eksponencijalne familije jednak je dimenziji prostora prirodnih parametara $\mathcal{N}_\theta$. Navedimo sada još jednu važnu klasu eksponencijalnih familija čija definicija ovisi o svojstvima prostora $\mathcal{N}_\theta$.

Definicija 1.1.5. *Kažemo da je eksponencijalna familija* ***regularna*** *ako vrijedi* $\mathcal{N}_\theta = \text{Int}\,\mathcal{N}_\theta$.

Klasa eksponencijalnih familija obuhvaća velik broj poznatih parametarskih distribucija poput Gaussove, Poissonove, Bernoullijeve, Gamma pa navedimo kanonske dekompozicije nekih od njih.

Primjer 1.1.6.

(a) Normalna distribucija s parametrom (μ, σ^2)*, čija je gustoća dana sa*

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

je primjer eksponencijalne familije reda 2. Naime, vrijedi

$$f(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2 - \frac{1}{2\sigma^2}\mu^2 - \frac{1}{2}log(\sigma^2)\right\}$$

pa je kanonska dekompozicija sljedeća

- $t(x) = (x, x^2)$ *je dovoljna statistika*
- $\theta = (\theta_1, \theta_2) = (\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2})$ *su prirodni parametari*
- $F(\theta) = -\frac{\theta_1^2}{2\theta_2} + \frac{1}{2}\log(-\frac{\pi}{\theta_2})$ *je log-normalizator*
- $k(x) = 0$ *je pripadna mjera.*

(b) Nadalje, Poissonova razdioba s parametrom λ

$$P(x = k;\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

je primjer eksponencijalne familije reda 1. Njezinu gustoću možemo zapisati na sljedeći način

$$P(x = k;\lambda) = \frac{1}{x!}\exp\{x\log(\lambda) - \lambda\}$$

pa vidimo da je je kanonska dekompozicija dana sa

- $t(x) = x$ *je dovoljna statistika*
- $\theta = \log\lambda$ *je prirodni parametar*
- $F(\theta) = \exp\theta$ *je log-normalizator*
- $k(x) = -\log x!$ *je pripadna mjera.*

Nadalje, navedimo par rezultata koji nam daju korisna svojstva funkcije F [5].

Propozicija 1.1.7. *Prostor prirodnih paramatera $\mathcal{N}_\theta$ je konveksan skup, te je $F(\theta)$ dana izrazom (1.4) konveksna funkcija na $\mathcal{N}_\theta$.*

Dokaz. Neka su $\theta_1, \theta_2 \in \mathcal{N}_\theta$ i $\alpha \in \mathbb{R}$ takav da vrijedi $0 < \alpha < 1$. Trebamo pokazati da vrijedi $\alpha\theta_1 + (1-\alpha)\theta_2 \in \mathcal{N}_\theta$, tj.

$$\int \exp\{(\alpha\theta_1 + (1-\alpha)\theta_2) \cdot t(x)\} h(x)\mu(x) < \infty,$$

pri čemu je $h(x) = \exp k(x)$. Imamo

$$\begin{aligned}
\int \exp\{(\alpha\theta_1 + (1-\alpha)\theta_2) \cdot t(x)\} h(x)\mu(x) &= \int (\exp\{\alpha\theta_1 \cdot t(x)\} \cdot \exp\{(1-\alpha)\theta_2 \cdot t(x)\}) h(x)\mu(x) \\
&= \int (\exp\{\theta_1 \cdot t(x)\})^{\alpha} (\exp\{\theta_2 \cdot t(x)\})^{1-\alpha} h(x)\mu(x) \\
&\leq \left(\int \exp\{\theta_1 \cdot t(x)\} h(x)\mu(x) \right)^{\alpha} \left(\int \exp\{\theta_2 \cdot t(x)\} h(x)\mu(x) \right)^{1-\alpha} \\
&< \infty,
\end{aligned}$$

gdje smo iskoristili Hölderovu nejednakost s parametrima $p = \frac{1}{\alpha}$ i $q = \frac{1}{1-\alpha}$ te prepostavku $\theta_1, \theta_2 \in \mathcal{N}_\theta$ iz koje onda slijedi da su integrali

$$\int \exp\{\theta_1 \cdot t(x)\} h(x)\mu(x), \int \exp\{\theta_2 \cdot t(x)\} h(x)\mu(x)$$

konačni. Time smo dokazali sljedeću nejednakost (koristeći (1.4))

$$\exp\{F(\alpha\theta_1 + (1-\alpha)\theta_2)\} \leq \exp\{\alpha F(\theta_1) + (1-\alpha)F(\theta_2)\},$$

koju možemo zapisati kao

$$F(\alpha\theta_1 + (1-\alpha)\theta_2) \leq \alpha F(\theta_1) + (1-\alpha)F(\theta_2)$$

pa zaključujemo da je $F(\theta)$ konveksna funkcija na $\mathcal{N}_\theta$. □

Propozicija 1.1.8. *[5] Log-normalizator $F(\theta)$ je glatka funkcija na* $\mathrm{Int}\mathcal{N}_\theta$. *Štoviše, F ima derivacije svakog reda za koje vrijedi*

$$\frac{\partial^l}{\partial\theta_1^{l_1} \dots \partial\theta_k^{l_k}} F(\theta) = \int \Big(\prod_{i=1}^{k} x_i^{l_i} \Big) \exp\{\theta \cdot t(x)\} h(x)\mu(x), \tag{1.5}$$

gdje je $\sum_{i=1}^{k} l_i = l$.

Ova propozicija ima važnu posljedicu, a to je da slučajna varijabla ima sve momente konačne te ih računamo pomoću derivacija funkcije $F(\theta)$. Naime, znamo da vrijedi

$$\int f(x;\theta)\mu(x) = \int \exp\{t(x)\cdot\theta - F(\theta)\}h(x)\mu(x) = 1. \tag{1.6}$$

Kada taj izraz deriviramo po parametru θ dobivamo

$$\nabla_\theta \int f(x;\theta)\mu(x) = \nabla_\theta \int \exp\{t(x)\cdot\theta - F(\theta)\}h(x)\mu(x) = 0, \tag{1.7}$$

odakle, diferenciranjem pod znakom integrala[1], slijedi

$$\begin{aligned}
\nabla_\theta \int f(x;\theta)\mu(x) &= \nabla_\theta \int \exp\{t(x)\cdot\theta - F(\theta)\}h(x)\mu(x) \\
&= \int \nabla_\theta \Big[\exp\{t(x)\cdot\theta - F(\theta)\}h(x) \Big]\mu(x) \\
&= \int \underbrace{\exp\{t(x)\cdot\theta - F(\theta)\}h(x)}_{=f(x;\theta)} \Big[t(x) - \nabla_\theta F(\theta) \Big]\mu(x) \\
&= \int f(x;\theta)t(x)\mu(x) - \int_x f(x;\theta)\nabla_\theta F(\theta)\mu(x) \\
&= \underbrace{\int f(x;\theta)t(x)\mu(x)}_{=\mathbb{E}_\theta[t(x)]} - \nabla_\theta F(\theta) \underbrace{\int f(x;\theta)\mu(x)}_{=1} \\
&= \mathbb{E}_\theta[t(x)] - \nabla_\theta F(\theta) \\
&= 0.
\end{aligned} \tag{1.8}$$

Prema tome,

$$\mathbb{E}_\theta[T(x)] = \nabla_\theta F(\theta). \tag{1.9}$$

Analogno se izvodi i izraz za varijancu $\mathrm{Var}(t(x))$ koja je onda jednaka

$$\mathrm{Var}(t(x)) = \mathbb{E}_\theta[t(x)^2] - \mathbb{E}_\theta[t(x)]^2 = \nabla^2 F(\theta). \tag{1.10}$$

Stoga možemo reći da je log-normalizator (1.4) ujedno i funkcija izvodnica eksponencijalne familije. Napomenimo još da je $\nabla^2 F(\theta)$ simetrična i pozitivno definitna matrica [8] koja definira Riemannovu metriku na $\mathcal{N}_\theta$. Svojstvima funkcije F nastavljamo se baviti u sljedećoj točki nakon što uvedemo neke osnovne pojmove iz konveksne analize.

[1] $\frac{d}{dx}\int f(x,t)dt = \int \frac{\partial}{\partial x} f(x,t)dt$

1.2 Dualna parametrizacija

Promotrimo sada još jedan način reprezentacije eksponencijalne familije koji je baziran na dualnoj reprezentaciji log-normalizatora. Fundamentalna dualnost u konveksnoj analizi je Legendre-Fenchelova tranformacija koja, neformalno, govori o tome da strogo konveksne i diferencijabilne funkcije dolaze u parovima. Kako bismo formalizirali navedenu dualnost, potrebno je uvesti neke rezultate iz konveksne analize pa krenimo stoga s definicijom konjugirane funkcije.

Definicija 1.2.1. *Neka je f realna funkcija na $\mathbb{R}^d$. Tada je njezina konjugirana funkcija f^* dana s*

$$f^*(x^*) = \sup_{x \in \mathrm{D}(f)} \{\langle x, x^* \rangle - f(x)\}. \tag{1.11}$$

Transformacija (1.11) je poznata i pod nazivom Legendre-Fenchelova transformacija.

Definicija 1.2.2. *Neka je f prava, zatvorena i konveksna funkcija i $C = \mathrm{Int}(\mathrm{D}(f))$. Uređeni par (C, f) zovemo konveksna funkcija Legendreovog tipa ili Legendreova funkcija ako je zadovoljeno sljedeće:*

(i) $C \neq \emptyset$,

(ii) *f je strogo konveksna i diferencijabilna na C,*

(iii) $(\forall x_b \in \partial C)$, $\lim_{x \to x_b} \|\nabla f(x)\| \to \infty$, *gdje je* $x \in C$.

Nadalje, ako je f zatvorena i konveksna funkcija na skupu $C = \mathrm{Int}(\mathrm{D}(f))$, postoji jedinstvena vrijednost x_0 koja odgovara supremumu definiranom sa relacijom (1.11) te ju dobivamo iz relacije

$$\nabla(\langle x^*, x \rangle - f(x))_{x=x_0} = 0.$$

Dakle, vrijedi $x^* = \nabla f(x_0)$. Stroga konveksnost funkcije f povlači monotonost ∇f pa stoga možemo definirati inverznu funkciju $(\nabla f)^{-1} : C^* \to C$, gdje je $C^* = \mathrm{Int}(\mathrm{D}(f^*))$. Može se pokazati da ako je (C, f) Legendreova funkcija, tada isto vrijedi i za (C^*, f^*) te u tom slučaju kažemo da su oni Legendreovi duali jedan drugoga. Osim toga, njihovi gradijenti su neprekidni i tvore bijekciju između otvorenih skupova C i C^*. Iskažimo sada formalno opisanu vezu u obliku sljedećeg teorema [4].

Teorem 1.2.3. *Neka je f realna, prava, zatvorena i konveksna funkcija te neka je f^* njoj konjugirana funkcija. Nadalje, neka je $C = \mathrm{Int}(\mathrm{D}(f))$ i $C^* = \mathrm{Int}(\mathrm{D}(f^*))$. Ako je (C, f) Legendreova funkcija, onda vrijedi:*

(i) *(C^*, f^*) je također konveksna Legendreova funkcija,*

(ii) (C, f) *i* (C^*, f^*) *su Legendreovi duali, tj.* $(C^{**}, f^{**}) = (C, f)$,

(iii) gradijent $\nabla f : C \to C^*$ *je bijekcija sa otvorenog konveksnog skupa* C *u otvoreni konveksi skup* C^*,

(iv) gradijenti gradijent ∇f, ∇f^* *su neprekidni te je* $\nabla f^* = (\nabla f)^{-1}$.

Sada primjena konveksne dualnosti na eksponencijale familiji proizlazi iz svojstava log-normalizatora. Taj je rezultat iskazan u lemi čiji se dokaz može naći u [13].

Lema 1.2.4. *Neka je* F *log-normalizator regularne eksponencijalne familije sa prostorom prirodnih parametara* $\Theta = \mathrm{D}(F)$. *Tada je* F *prava, zatvorena i konveksna funkcija za koju vrijedi* $\mathrm{Int}(\Theta) = \Theta$ *te je* (Θ, F) *konveksna Legendreova funkcija.*

Promotrimo sada konjugiranu funkciju $G = F^*$ log-normalizatora F. Nju dobivamo iz definicije (1.2.1) tako da izračunamo

$$G(\eta) = \sup_{\theta \in \Theta} \{\langle \theta, \eta \rangle - F(\theta)\}.$$

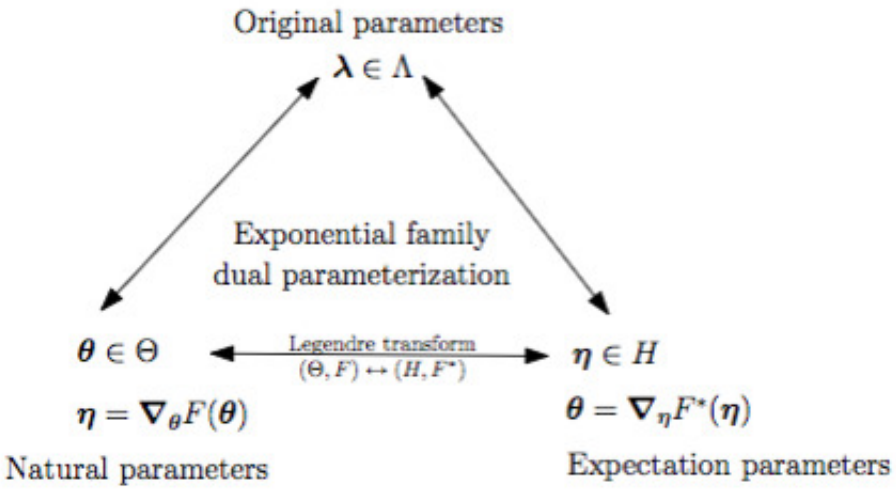
Iz Leme 1.2.4 slijedi da je (Θ, F) strogo konveksna Legendreova funkcija pa primjenom Teorema 1.2.3 dobivamo preslikavanje između prostora prirodnih parametara Θ i $H = \mathrm{Int}(\mathrm{D}(G)) = \Theta^*$ koji nazivamo *prostor očekivanih parametara*. To preslikavanje je dano s

$$\eta(\theta) = \nabla F(\theta) \quad \text{i} \quad \theta(\eta) = \nabla G(\eta) = \nabla F^{-1}(\eta). \tag{1.12}$$

Primjetimo da naziv *očekivani parametar* slijedi iz činjenice da je

$$\eta = \eta(\theta) = \mathbb{E}_\theta[T(x)] = \nabla F(\theta)$$

Sada smo, uz izvorne i prirodne parametre, dobili još jednu parametrizaciju eksponencijalne familiji. Njihova je bijektivna veza prikazana na Slici 1.1.



Slika 1.1: Dualna parametrizacija eksponencijalnih familija

1.3 Geometrija eksponencijalnih familija

Kao što smo naveli ranije, ideja informacijske geometrije je promatrati parametarski statistički model $\{p(x;\theta) : \theta \in \Theta\}$ kao mnogostrukost. Taj pristup nije ograničen samo na eksponencijalne familije stoga navedimo najprije neke općenite pojmove potrebne za razumjevanje osnovnih koncepata ovog poglavlja.

Definicija 1.3.1. *Topološka mnogostrukost M dimenzije n je Hausdorffov prostor sa prebrojivom bazom takav da svaka točka $P \in M$ ima okolinu koja je homeomorfna nekom otvorenom podskupu prostora $\mathbb{R}^n$.*

Uvjet prebrojivosti baze zapravo znači da za svaku točku $P \in M$ postoji okolina $U \subset M$, koju nazivamo koordinatna okolina, takav da postoji izomorfizam između tog skupa i euklidskog prostora. Navedeni izomorfizam definira lokalni koordinatni sustav u svakoj od tih okolina. Nadalje, za mnogostrukost kažemo da je diferencijabilna ako su njezine koordinatne transformacije diferencijabilne. Dakle, u okolini svake točke $P \in M$ uvodimo lokalni koordinatni sustav

$$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n), \tag{1.13}$$

koji se sastoji od n komponenti koje ju jednoznačno određuju. Koordinatni sustav nije jedinstven, čak ni u koordinatnoj okolini. Stoga, kada točku iz M možemo reprezentirati u dva koordinatna sustava između njih mora postojati bijektivna veza. Uvedimo sada pojam

divergencije, koja predstavlja stupanj separacije između dvije točke, koristeći definiciju navedenu u [2].

Definicija 1.3.2. *Neka su $P, Q \in M$. Funkciju $D(P||Q)$ zovemo **divergencija** ako je zadovoljeno sljedeće:*

(i) $D(P||Q) \geq 0$,

(ii) $D(P||Q) = 0$ *ako i samo ako je* $P = Q$,

(iii) Ako su P i Q, dovoljno blizu te njihove koordinate sustave označimo θ_P, $\theta_Q = \theta_P + d\theta$ respektivno, tada je Taylorov razvoj funkcije D dan izrazom

$$D(\theta_P||\theta_P + d\theta) = \frac{1}{2}\sum_{i,j} g_{ij}(\theta_P)d\theta_i d\theta_j + O(|d\theta|^3), \tag{1.14}$$

gdje je $G = [g_{ij}]$ pozitivno definitna matrica koja ovisi o θ_P.

Važno je napomenuti da za funkciju divergrencije općenito ne vrijedi svojstvo simetričnosti stoga ona ne zadovoljava aksiome metrike, no po potrebi ju možemo simetrizirati sljedećim izrazom

$$D_s(\theta_P||\theta_Q) = \frac{1}{2}(D(\theta_P||\theta_Q) + D(\theta_Q||\theta_P)). \tag{1.15}$$

Sljedeće, promotrimo što se događa sa susjednim točkama. Iz izraza (1.14) slijedi da za točke koje su dovoljno blizu možemo definirati infinitezimalnu udaljenost s

$$ds^2 = 2D(\theta||\theta + d\theta) = \sum_{i,j} g_{ij}(\theta_P)d\theta_i d\theta_j. \tag{1.16}$$

Pomoću gornjih rezultata možemo dati pojednostavljenu definiciju Riemannove mnogostrukosti, u kojoj je Riemannova metrika inducirana divergencijom D.

Definicija 1.3.3. *Za mnogostrukost M kažemo da je Riemannova mnogostrukost ako je na njoj definirana pozitivno definitna matrica $G(\theta)$ te je kvadrat udaljenosti dviju susjednih točaka θ i $\theta + d\theta$ dan izrazom (1.16).*

Pod pojmom statističke mnogostrukosti podrazumjevamo mnogostrukost $M = \{p(x;\theta)\}$ koja se sastoji od vjerojatnosnih distribucija s parametrom θ. Tada distribucija $p(x;\theta)$ predstavlja točku mnogostrukosti, a prostor parametara Θ interpretiramo kao pripadni koordinatni sustav pomoću kojeg onda opisujemo članove famlije. Sada, ako želimo statističku mnogostrukost opskrbiti Riemannovom strukturom potrebno je uvesti neku vrstu metrike te je pokazano da je jedina Riemannova metrika koja u tom slučaju ima smisla definirana pomoću Fisherove informacijske matrice, čija definicija slijedi u nastavku.

Definicija 1.3.4. *Fisherova informacijska matrica je pozitivno definitna matrica čiji su koeficijenti dani s*

$$I(\theta) = \mathbb{E}_\theta\left[\frac{\partial \log p(x;\theta)}{\partial \theta_i}\frac{\partial \log p(x;\theta)}{\partial \theta_j}\right] = [g_{ij}(\theta)]. \tag{1.17}$$

Ona nam, naime, služi kao baza za definiranje infinitezimalne udaljenosti između dvije točke θ i $\theta + d\theta$ koje su dovoljnu blizu.

Definicija 1.3.5. *Infinitezimalna udaljenost točaka θ i $\theta + d\theta$ je sljedeća kvadratna forma*

$$ds^2 = ds^2(\theta) = \sum_{i=1}^{d}\sum_{j=1}^{d} g_{ij}d\theta_i d\theta_j = (\nabla\theta)^T I(\theta)\nabla\theta \tag{1.18}$$

Prema tome, Fisherova informacijska matrica definira skalarni produkt na tangencijalnom prostoru u točki θ i na taj način opskrbljuje statističku mnogostrukost strukturom Riemannove mnogostrukosti. Sada možemo računati geodetsku udaljenost između dvije distribucije, tj. najkraći put izmedu između točaka θ_1 i θ_2.

Definicija 1.3.6. *Geodetsku udaljenost između dvije distribucije $p(x;\theta_1)$ i $p(x;\theta_2)$ na nekoj statističkoj mnogostrukosti zovemo Rao-Fisherova udaljenost i definiramo sa*

$$d(p(x;\theta_1), p(x;\theta_2)) = \min_{\theta(t)|\theta(0)=\theta_1,\theta(1)=\theta_2} \int_0^1 \sqrt{ds^2}dt. \tag{1.19}$$

Sa statističkog stajališta posebno je značajna Kullback-Leiberova divergencija, koju zovemo još i relativna entropija, budući da ona mjeri sličnost dviju distribucija pa izrecimo njezinu definiciju.

Definicija 1.3.7. *Neka su $p(x;\theta_1)$ i $p(x;\theta_2)$ vjerojatnosne distribucije. Kullback-Leiberova divergencija D_{KL} je funkcija definirana sa*

$$D_{KL}(p(x;\theta_1)\|p(x;\theta_2)) = \int p(x;\theta_1)\log\frac{p(x;\theta_1)}{p(x;\theta_2)}dx.$$

Kullback-Leiberova divergencija je u uskoj vezi sa Shannonovom entropijom

$$D_{KL}(p(x;\theta_1)\|p(x;\theta_2)) = H^{\times}(p(x;\theta_1)\|p(x;\theta_2)) - H(p(x;\theta_1)), \tag{1.20}$$

gdje je

$$H(p(x;\theta_1)) = \int p(x;\theta_1)\log\frac{1}{p(x;\theta_1)}dx$$

Shannonova, a

$$H^{\times}(p(x;\theta_1)||p(x;\theta_2)) = \int p(x;\theta_1)\log\frac{1}{p(x;\theta_2)}dx$$

unakrsna entropija. Prema tome, možemo reći da Kullback-Leiberova divergencija između $p(x;\theta_1)$ i $p(x;\theta_2)$ izražava gubitak informacije koji nastaje kada aproksimiramo prvu distribuciju drugom. Napomenimo još da Kullback-Leiberova divergencija pripada klasi Bregmanovih divergencija, koje su značajne u kontekstu konveksne optimizacije budući da su parametrizirane konveksnom funkcijom.

Definicija 1.3.8. *Neka je $F : X \to \mathbb{R}$ strogo konveksna funkcija definirana na konveksnom skupu $X \subseteq \mathbb{R}^n$ za koju vrijedi da je diferencijabilna na relativnom interioru od X. Tada je Bregmanova divergencija generirana s F, u oznaci B_F, dana s*

$$B_F(\theta_1||\theta_2) = F(\theta_1) - F(\theta_2) - \langle\theta_1 - \theta_2, \nabla F(\theta_2\rangle. \tag{1.21}$$

Dakle, KL divergencija između dviju distribucija iste eksponencijalne familije jednaka je Bregmanovoj divergenciji između pripadnih parametara pri čemu je generator Bregmanove divergencije log-normalizator $F(\theta)$.

Banerjee i suradnici [4] su dali dualnu vezu između regularne eksponencijalne familije i Bregmanove divergencije preko Legendre-Fenchelove transformacija i o tome govori sljedeća lema.

Lema 1.3.9. *Regularnu eksponencijanu familiju s log-normalizatorom F možemo zapisati u terminina Bregmanove divergencije s generatorom F^*, $\log p_F(x;\theta) = -B_{F^*}(t(x)||\nabla F) + F^*(x) + k(x)$, gdje je F^* Legendreov dual funkcije F.*

Napokon, za eksponencijalne obitelji, Bregmanove divergencije na prirodnim i očekivanim parametrima povezane su s Kullback-Leiblerovom divergencijom na odgovarajućim distribucijama. Time dolazimo do najvažnijeg rezultata koji nam omogućuje određivanje Kullback-Leiblerove divergneciju u zatvorenoj formi.

Propozicija 1.3.10. *Kullback-Leiblerova divergencija između dva člana eksponencijalne familije jednaka je Bregmanovoj divergenciji između njihovih parametara, tj.*

$$D_{KL}(p(x;\theta_1)||p(x;\theta_2)) = B_F(\theta_2||\theta_1) = B_{F^*}(\eta_1||\eta_2) \tag{1.22}$$

Dakle, Kullback-Leiblerova divergencija između dvije vjerojatnosne distribucije na statističkoj mnogostrukosti može se izračunati pomoću Bregmanove divergencije na prirodnim i očekivanim parametrima pomoću konveksne dualnosti.

Poglavlje 2

Von Mises-Fisherova distribucija

Prilikom analize podataka ponekad je potrebno obratiti pozornost na njihov smjer i stoga je tradicionalna statistika često je neprikladna za analizu usmjerenih podataka, koji modeliraju širok spektar pojava u područjima poput meteorologije, geomagmetizma, radiologije i brojnih drugih. Za takav tip podataka potrebni su nam alati područja usmjerene statistike, u kojem je von Mises-Fisherova distribucija najjednostavnija i najčešće korištena vjerojatnosna distribucija pa ćemo ovom poglavlju detaljnije proučiti njezina svojstva.

2.1 Definicija i osnovna svojstva vMF distribucije

Kažemo da d-dimenzionalni jedinični slučajni vektor x ima d-dimenzinalnu von Mises-Fisherovu distribuciju, koju kraće označavamo vMF, ako mu je funkcija gustoće dana s

$$f(x \mid \mu, \kappa) = c_d(\kappa) e^{\kappa \mu^T x}, \qquad x \in \mathbb{S}^{d-1}, \tag{2.1}$$

gdje je $\|\mu\| = 1$, $\kappa \geq 0$, $\mathbb{S}^{d-1}$ označava $(d-1)$-dimenzionalnu jediničnu sferu, a $c_d(\kappa)$ je normalizacijska konstanta dana sa

$$c_d(\kappa) = \frac{\kappa^{d/2-1}}{(2\pi)^{d/2} I_{d/2-1}(\kappa)}. \tag{2.2}$$

Naziv normalizacijska konstanta, naime, proizlazi iz toga što nam ona osigurava da vrijedi

$$\int_{x \in \mathbb{S}^{d-1}} c_d(\kappa) e^{\kappa \mu^T x} dx = 1. \tag{2.3}$$

Detalji izvoda dani su u [7]. Ovdje $I_d(x)$ označava modificiranu Besselovu funkciju prve vrste reda d definiranu s

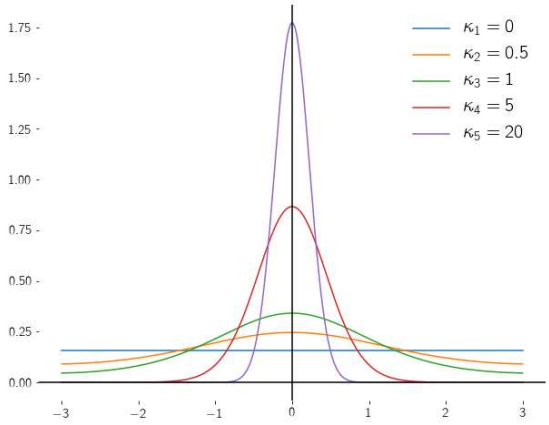
$$I_d(x) = \left(\frac{x}{2}\right)^d \sum_{k=1}^{\infty} \frac{(x/2)^{2k}}{\Gamma(k+1)\Gamma(d+k+1)}, \tag{2.4}$$

pri čemu je $\Gamma(\cdot)$ Gamma funkcija. Gustoća $f(x \mid \mu, \kappa)$ je parametrizirana s usmjerenim očekivanjem μ i koncentracijskim parametrom κ, čiji naziv proizlazi iz činjenice da on karakterizira koliko su jedinični vektori gusto koncentrirani oko μ.

Poseban slučaj vMF distribucije je von Misesova distribucija, koju dobivamo za $d = 2$, a čija je gustoća poprima sljedeći oblik

$$f(x \mid \mu, \kappa) = \frac{e^{\kappa \cos(x-\mu)}}{2\pi I_0(\kappa)}, \qquad 0 \leq x \leq 2\pi. \tag{2.5}$$

Slika (2.1) prikazuje fukciju gustoće von Misesove distribucije za $\mu = 0$ i različite vrijednosti parametra κ. Vidimo da gustoća postaje koncentriranija oko μ kako κ raste te stoga



Slika 2.1: Von Misesova distribucija za različite κ

parametar κ ima smisla nazivati parametarom koncentracije. Također, lako se vidi da je von Misesova distribucija (2.5) 2-parametarska eksponencijalna familija, gdje je

- $T(x) = (\cos x, \sin x)$ dovoljna statistika,
- $(\theta_1, \theta_2) = (\kappa \cos \mu, \kappa \sin \mu)$ prirodni parametar,
- $F(\theta_1, \theta_2) = \log\left(2\pi I_0 \sqrt{\theta_1^2 + \theta_2^2}\right)$ log-normalizator,
- $k(x) = 0$ pripadna mjera.

2.2 Procjena parametara modela metodom maksimalne vjerodostojnosti

Neka je $\mathcal{X} = \{x_1, \ldots, x_n\}$ skup koji se sastoji od n nezavisnih i jednako distribuiranih jedničnih slučajnih vektora čija je funkcija gustoće dana sa (2.1). Cilj nam je procijeniti parametre modela μ i κ koji su nam nepoznati te to činimo metodom maksimalne vjerodostojnosti. Kako su svi x_i, $i = 1, \ldots, n$ nezavisini i jednako distribuirani, funkciju vjerodostojnosti možemo zapisati u sljedećem obliku

$$f(x_1, \ldots, x_n \mid \mu, \kappa) = \prod_{i=1}^{n} f(x_i \mid \mu, \kappa) = \prod_{i=1}^{n} c_d(\kappa) e^{\kappa \mu^T x_i}. \tag{2.6}$$

Prema tome, log-vjerodostojnost je jednaka

$$\mathcal{L}(\mu, \kappa) = n \ln c_d(\kappa) + \kappa \mu^T r, \tag{2.7}$$

gdje je $r = \sum_{i=1}^{n} x_i$. Kako bismo dobili procjenitelje maksimalne vjerodostojnosti, koje kraće označavamo MLE, od μ i κ potrebno je maksimizirati izraz (2.7) uz ograničenja $\mu^T \mu = 1$ i $\kappa \geq 0$. Kombiniranjem tih uvjeta i ciljne funkcije dobivamo Lagrangeovu funkciju

$$L(\mu, \kappa, \lambda; x_1, \ldots, x_n) = n \ln c_d(\kappa) + \kappa \mu^T r + \lambda(1 - \mu^T \mu), \tag{2.8}$$

gdje je λ Lagrangeov multiplikator. Označimo sa $\hat{\mu}$, $\hat{\lambda}$ i $\hat{\kappa}$ procijene parametara μ, λ i κ, respektivno. Deriviramo gornju funkciju s obzirom na μ, λ i κ te dobivene derivacije izjednačimo s 0 pa dobivamo sljedeće jednadžbe koje procjenitelji $\hat{\mu}$, $\hat{\lambda}$ i $\hat{\kappa}$ moraju zadovoljavati

$$\hat{\mu} = \frac{\hat{\kappa}}{2\hat{\lambda}} r, \tag{2.9}$$

$$\hat{\mu}^T \hat{\mu} = 1, \tag{2.10}$$

$$\frac{n c_d'(\hat{\kappa})}{c_d(\hat{\kappa})} = -\hat{\mu}^T r. \tag{2.11}$$

Sada iz (2.9) i (2.10) zaključujemo da vrijedi

$$\hat{\lambda} = \frac{\hat{\kappa}}{2} \|r\| \tag{2.12}$$

te

$$\hat{\mu} = \frac{r}{\|r\|} = \frac{\sum_{i=1}^{n} x_i}{\| \sum_{i=1}^{n} x_i \|}. \tag{2.13}$$

Time smo dobili ML procjenu od μ.
Nadalje, uvrstimo dobiveni $\hat{\mu}$ u (2.11) pa slijedi

$$\frac{c_d'(\hat{\kappa})}{c_d(\hat{\kappa})} = -\frac{\|r\|}{n} = -\bar{r}. \tag{2.14}$$

Označimo sa $s = d/2 - 1$. Sada deriviranjem (2.2) po κ dobivamo

$$c_d'(\hat{\kappa}) = \frac{s\kappa^{s-1}}{\alpha I_s(\kappa)} - \frac{\kappa^s I_s'(\kappa)}{\alpha I_s^2(\kappa)}, \tag{2.15}$$

gdje je $\alpha = (2\pi)^{s+1}$ konstanta. Desnu strane prethodne jednakosti možemo pojednostaviti i zapisati u sljedećem obliku

$$\frac{\kappa^2}{\alpha I_s(\kappa)}\left(\frac{s}{\kappa} - \frac{I_s'(\kappa)}{I_s(\kappa)}\right) = c_d(\kappa)\left(\frac{s}{\kappa} - \frac{I_s'(\kappa)}{I_s(\kappa)}\right). \tag{2.16}$$

Koristeći se rekurzivnom relacijom za Besselove funkcije $\kappa I_{s+1}(\kappa) = \kappa I_s'(\kappa) - s I_s(\kappa)$ dobivamo

$$-\frac{c_d'(\kappa)}{c_d(\kappa)} = \frac{I_{s+1}(\kappa)}{I_s(\kappa)} = \frac{I_{d/2}(\kappa)}{I_{d/2-1}(\kappa)} := A_d(\kappa). \tag{2.17}$$

Sada iz (2.15) i (2.17) vidimo da MLE za parameter κ možemo dobiti rješavanjem

$$A_d(\kappa) = \bar{r}. \tag{2.18}$$

Funkcija $A_d(\kappa)$ ima inverz budući da je strogo rastuća [14] pa slijedi da je

$$\hat{\kappa} = A_d^{-1}(\bar{r}). \tag{2.19}$$

Procjene za parametar κ

Budući da je $A_d(\kappa)$ omjer Besselovih funkcija, rješenje gornje jednakosti ne postoji u zatvorenoj formi. Prema tome, moramo se poslužiti numeričkim ili asimptotskim metodama kako bismo procijenili κ. Najjednostavnije procjene parametra κ dali su Mardia i Jupp [11]. Oni, naime, navode sljedeća dva slučaja za $\hat{\kappa}$

$$\begin{aligned} \hat{\kappa} &\approx \frac{d-1}{2(1-\bar{r})} \quad \text{za velike } \bar{r}, \\ \hat{\kappa} &\approx d\bar{r}\Big(\frac{d}{d+2}\bar{r}^2 + \frac{d^2(d+8)}{(d+2)^2(d+4)}\bar{r}^4\Big) \quad \text{za male } \bar{r}. \end{aligned} \tag{2.20}$$

Ove aproksimacije dodatno pretpostavljaju $\kappa \gg d$, a to često ne vrijedi kod visoko dimenzionalnih podataka poput teksta pa je u tu svrhu potrebno naći pogodnije procjene parametra

κ. Jedva takva je opisana u [3], a ideju njezinog izvoda navodimo ovdje. Najprije primjetimo da je $A_d(\kappa)$ omjer Besselovih funkcija koje se razlikuju u redu samo za jedan pa se možemo poslužitu reprezetacijom u obliku verižnog razlomka [16] danom sa

$$A_d(\kappa) = \frac{I_{d/2}}{I_{d/2-1}} = \cfrac{1}{\cfrac{d}{\kappa} + \cfrac{1}{\cfrac{d+2}{\kappa} + \dots}}. \tag{2.21}$$

Iskoristimo $A_d(\kappa) = \overline{r}$ pa (2.21) možemo zapisati kao

$$\frac{1}{\overline{r}} \approx \frac{d}{\kappa} + \overline{r} \tag{2.22}$$

odakle slijedi aproksimacija

$$\kappa \approx \frac{d\,\overline{r}}{1-\overline{r}^2}. \tag{2.23}$$

Banerjee i suradnici su empirijski odredili korektivni član $-\overline{r}^3/(1-\overline{r}^2)$ koji je potrebno dodati gornjoj aproksimaciji kako bi se povećala njezina točnost [3]. Time konačno dobivamo

$$\hat{\kappa} = \frac{\overline{r}d - \overline{r}^3}{1-\overline{r}^2}. \tag{2.24}$$

Ova procjena je vrlo praktična za upotrebu zbog svoje brzine i jednostavnosti implementacije, no po pitanju točnosti su se neke druge procjene pokazale kao bolji izbor. Nadalje, procjena (2.24) se može dodatno poboljšati korištenjem nekoliko iteracija Newtonowe metode primijenjene na $A_d(\kappa) - \overline{r} = 0$ te pri tome koristimo sljedeći rezultat.

Lema 2.2.1. *Vrijedi* $A_d'(\kappa) = 1 - A_d(\kappa)^2 - \frac{d-1}{\kappa}A_d(\kappa)$.

Dokaz. Stavimo $s = d/2 - 1$. Budući da je $A_d(\kappa) = \frac{I_{s+1}(\kappa)}{I_s(\kappa)}$ vrijedi

$$A_d'(\kappa) = \frac{I_{s+1}'(\kappa)}{I_s(\kappa)} - \frac{I_{s+1}(\kappa)I_s'(\kappa)}{I_s^2(\kappa)}$$

Koristeći se rekurzijom za derivacije Besselovih funkcija danom sa $\kappa I_d'(\kappa) = \kappa I_{d-1}(\kappa) - dI_d(\kappa)$ dobivamo

$$\frac{I_{s+1}'(\kappa)}{I_s(\kappa)} = 1 - \frac{s+1}{\kappa}\frac{I_{s+1}(\kappa)}{I_s}$$

Sada pak upotrijebimo rekurzivnu relaciju $\kappa I_d'(\kappa) = pI_d(\kappa) + \kappa I_{p+1}(\kappa)$ pa slijedi

$$\frac{I_s'(\kappa)}{I_s(\kappa)} = \frac{s}{\kappa} + \frac{I_{s+1}(\kappa)}{I_s}$$

Koristeći gornje jednakosti dobivamo

$$A_d'(\kappa) = 1 - A_d(\kappa)^2 - \left(\frac{s}{\kappa} + \frac{s+1}{\kappa}\right)A_d(\kappa)$$

Uvrstimo još $s = p/2 - 1$ pa slijedi tražena jednakost. □

Dakle, u slučaju korištenja Newtonove metode za poboljšanje procjene, iskoristimo relaciju (2.24) kao početnu iteraciju, tj. stavimo $\kappa_0 = \kappa = \frac{\bar{r}d - \bar{r}^3}{1-\bar{r}^2}$. Sljedeće iteracije Newtonove metode računamo koristeći

$$\kappa_{i+1} = \kappa_i - \frac{A_d(\kappa_i) - \bar{r}}{1 - A_d(\kappa_i)^2 - \frac{d-1}{\kappa_i}A_d(\kappa_i)}, \quad i = 1, 2, \ldots$$

Bitno je napomenuti da ova je ova metoda dosta spora, no Sra [15] je eksperimentalno pokazao da je dovoljno ograničiti se na dvije iteracije te tako dobili procjenu koja je u prosjeku daje bolje rezultate od prethodne dvije, a nije računski puno zahtjevnija. Također, eksperimentalno se lako pokaže da nije potrebno uvoditi više od dvije iteracije Newtonove metode budući da svaka dodatna iteracija uključuje poziv funkcije $A_d(\kappa)$, koja je računski zahtjevna, a konačni rezultat nije mnogo bolji. U vidu tog rezultata, procjene parametra κ u ovom radu biti će temeljena baš na toj metodi.

Poglavlje 3

Algoritam maksimizacije očekivanja za konveksnu kombinaciju vMF distribucija

U praksi se često susrećemo s kompleksnim podacima koje nije moguće modelirati pomoću samo jedne von Mises-Fisherove distribucije te su nam u tom slučaju od velike koristi konveksne kombinacije vMF distribucija koje kraće označavamo s movMF (engl. *mixture of vMF distributions*). Takvi modeli pokazali su se vrlo efikansnima pri analizi visokodimenzionalih podataka poput teksta, što je i fokus ovoga rada.

3.1 Konveksna kombinacija vMF distribucija

Označimo sa $f(x|\theta_k)$ funkciju gustoće von Mises-Fisherove distribucije sa parametrom $\theta_k = (\mu_k, \kappa_k)$, $1 \leq k \leq K$. Tada konveksna kombinacija K vMF distribucija (movMF) ima gustoću danu sa

$$f(x|\boldsymbol{\theta}) = \sum_{k=1}^{K} \alpha_k f(x|\theta_k), \tag{3.1}$$

gdje je $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ te vrijedi $\sum_k \alpha_k = 1$ i $\alpha_k \geq 0$.
Nadalje, $f(x|\theta_k)$ zovemo komponente konveksne kombinacije (engl. *mixture components*), a parametre α_k koeficijenti (engl. *mixture coefficients*).
Neka je $\mathcal{X} = \{x_1, \ldots x_n\}$ dani skup primjera generiran s (3.1). Naš je cilj odrediti parametre modela te tako ustanoviti s kojom vjerojatnoću primjeri pripadaju pojedinim komponentama. Kao i kod von Mises-Fisherove distribucije, parametre procijenjujemo metodom maksimalne vjerodostojnosti. No sada, budući da ne znamo kojoj komponenti pripada koji primjer, pronalažanje optimalnih parametara modela je puno kompleksnije nego u prethod-

nom slučaju.

Naime, log-vjerodostojnost na $\mathcal{X}$ dana je s

$$\mathcal{L}(\boldsymbol{\theta}|\mathcal{X}) = \ln \prod_{i=1}^{n} f(x_i|\boldsymbol{\theta}) = \ln \prod_{i=1}^{n} \sum_{k=i}^{K} \alpha_k f(x_i|\theta_k) = \sum_{i=1}^{n} \ln\Big(\sum_{k=i}^{K} \alpha_k f(x_i|\theta_k)\Big). \tag{3.2}$$

Maksimizacija ove log-vjerodostojnosti nema rješenje u zatvorenoj formi zbog logaritamske funkcije koja djeluje na sumu te zato pri rješavanju koristimo iterativne metode. Jedna od takvih metoda je i algoritam maksimizacije očekivanja, koji je opisan u nastavku.

3.2 Općenita formulacija algoritma maksimizacije očekivanja

Algoritam maksimizacije očekivanja (engl. *expectation maximization*) ili, kraće, EM- algoritam je iterativni optimizacijski postupak za maksimizaciju vjerodostojnosti koji se često koristi kod modela sa skrivenim varijablama. Skrivene varijable su one slučajne varijable koje ne opažamo direktno, već su neizravno procijenjene na temelju drugih opaženih varijabli. Dakle, cilj je pronalazak parametara $\boldsymbol{\theta}$ i koeficijenata α_k tako da je log-vjerodostojnost $\mathcal{L}(\boldsymbol{\theta}|\mathcal{X})$ maksimalna, gdje je $\mathcal{X}$ dani skup primjera. Najprije ćemo dati općenitu formulaciju algoritma, a zatim ćemo ga primjenti na konveksnu kombinaciju vMF distribucija. Često, kao što je i slučaj kod maksimizacije (3.2), rješenje nije moguće naći analitički. Tada, prije korištenja iterativne optimizacije, dani model proširujemo skupom skrivenih varijabli $\mathcal{Z}$ na način da svaka skrivena varijabla opisuje vezu između primjera i komponente. Skup $\{\mathcal{X}, \mathcal{Z}\}$ zovemo potpuni, dok je skup $\mathcal{Z}$ nepotpuni skup primjera. U nastavku radimo sa zajedničkom gustoćom $f(\mathcal{X}, \mathcal{Z}|\boldsymbol{\theta})$ te pomoću nje izrazimo novu log-vjerodostojnost

$$\mathcal{L}(\boldsymbol{\theta}|\mathcal{X}, \mathcal{Z}) = \ln f(\mathcal{X}, \mathcal{Z}|\boldsymbol{\theta}) \tag{3.3}$$

koju zovemo potpuna log-vjerodostojnost. Analognu, polaznu funkciju zovemo nepotpuna log-vjerodostojnost. Nju možemo napisati u obliku

$$\mathcal{L}(\boldsymbol{\theta}|\mathcal{X}) = \ln f(\mathcal{X}|\boldsymbol{\theta}) = \ln \sum_{\mathcal{Z}} f(\mathcal{X}, \mathcal{Z}|\boldsymbol{\theta}), \tag{3.4}$$

jer za marginalnu gustoću vrijedi $f(\mathcal{X}|\boldsymbol{\theta}) = \sum_{\mathcal{Z}} f(\mathcal{X}, \mathcal{Z}|\boldsymbol{\theta})$. Primjetimo da u izrazu (3.4), gdje optimiramo marginalne gustoće, logaritamska funkcija djeluje na zbroj, a u drugom slučaju (3.3) djeluje direktno na zajedničku gustoću te nam to omogućava pronalazak analitičkog rješenja. Nažalost, vrijednosti skrivenih varijabli su nam nepoznate i (3.3) je zapravo slučajna varijabla koja ovisi o distribuciji od $\mathcal{Z}$ pa ne možemo direktno raditi

s potpunom log-vjerodostojnošću. Umjesto toga moramo raditi s njezinim očekivanjem $\mathbb{E}[\mathcal{L}(\boldsymbol{\theta}|\mathcal{X}, \mathcal{Z})]$. Njega ćemu u nastavku označavati sa $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$. Iteracije algoritma alterniraju između dva koraka:

- E-korak (korak procjene očekivanja),
- M-koraka (korak maksimizacije).

U E-koraku algoritma računamo očekivanje potpune log-vjerodostojnosti s obzirom na trenutnu uvjetnu distribuciju od $\mathcal{Z}$ uz dane $\mathcal{X}$ i fiksne trenutne vrijednost parametara $\boldsymbol{\theta}^{(t)}$. Dakle, računamo

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &= \mathbb{E}_{\mathcal{Z}|(\mathcal{X},\boldsymbol{\theta}^{(t)})}[\mathcal{L}(\boldsymbol{\theta}|\mathcal{X}, \mathcal{Z})] = \mathbb{E}_{\mathcal{Z}|(\mathcal{X},\boldsymbol{\theta}^{(t)})}[\ln f(\mathcal{X}, \mathcal{Z}|\boldsymbol{\theta})] \\ &= \sum_{\mathcal{Z}} P(\mathcal{Z}|(\mathcal{X}, \boldsymbol{\theta}^{(t)}) \ln f(\mathcal{X}, \mathcal{Z}|\boldsymbol{\theta}) \end{aligned} \tag{3.5}$$

Funkciju $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ zovemo očekivana log-vjerodostojnost. U izrazu (3.5) $P(\mathcal{Z}|(\mathcal{X}, \boldsymbol{\theta}^{(t)})$ označava aposteriornu vjerojatnost od $\mathcal{Z}$ uz dani $\mathcal{X}$ i trenutnu vrijednost parametra $\boldsymbol{\theta}^{(t)}$ koju ćemo izračunati primjenom Bayesovog teorema.

U M-koraku maksimiziramo očekivanje dobiveno u prethodnom koraku, tj. procijenjujemo nove parameter $\boldsymbol{\theta}^{(t+1)}$ koji maksimiziraju (3.5), tj.

$$\boldsymbol{\theta}^{(t+1)} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}). \tag{3.6}$$

Pokažimo najprije da je postupak zamjene nepotpune log-vjerodostojnosti s očekivanjem potpune log-vjerodostojnosti opravdan u obliku sljedeće propozicije.

Propozicija 3.2.1. *Maksimizacija očekivanja potpune log-vjerodostojnosti implicira povećanje nepotpune log-vjerodostojnosti.*

Dokaz. Za svaki skup skrivenih varijabli $\mathcal{Z}$ takvih da vrijedi da je vjerojatnost $P(\mathcal{Z}|\mathcal{X}, \boldsymbol{\theta})$ nepotpunu log-vjerodostojnost možemo zapisati kao

$$\ln f(\mathcal{X}|\boldsymbol{\theta}) = \ln f(\mathcal{X}, \mathcal{Z}|\boldsymbol{\theta}) - \ln P(\mathcal{Z}|\mathcal{X}, \boldsymbol{\theta}). \tag{3.7}$$

Gornju jednakost najprije pomnožimo sa aposteriornom vjerojatnosti $P(\mathcal{Z}|\mathcal{X}, \boldsymbol{\theta}^{(t)})$ i sumiramo po $\mathcal{Z}$, a zatim uzmemo njegovo očekivanje po svim skrivenim varijablama iz $\mathcal{Z}$ uz dani trenutni parametar $\boldsymbol{\theta}^{(t)}$ pa imamo

$$\begin{aligned} \ln f(\mathcal{X}|\boldsymbol{\theta}) &= \sum_{\mathcal{Z}} P(\mathcal{Z}|\mathcal{X}, \boldsymbol{\theta}^{(t)}) \ln f(\mathcal{X}, \mathcal{Z}|\boldsymbol{\theta}) - \sum_{\mathcal{Z}} P(\mathcal{Z}|\mathcal{X}, \boldsymbol{\theta}^{(t)}) \ln P(\mathcal{Z}|\mathcal{X}, \boldsymbol{\theta}) \\ &= Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) + H^{\times}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}), \end{aligned} \tag{3.8}$$

gdje je $H^{\times}$ unakrasna entropija dana s $H^{\times}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \sum_{\mathcal{Z}} P(\mathcal{Z}|\mathcal{X}, \boldsymbol{\theta}^{(t)}) \ln P(\mathcal{Z}|\mathcal{X}, \boldsymbol{\theta})$ Izraz (3.8) vrijedi za svaku vrijednost $\boldsymbol{\theta}$ pa i za $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}$. Dakle, vrijedi

$$\ln f(\mathcal{X}|\boldsymbol{\theta}^{(t)}) = Q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}) + H^{\times}(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}). \tag{3.9}$$

Oduzimanjem (3.9) od (3.8) dobijemo

$$\ln f(\mathcal{X}|\boldsymbol{\theta}) - \ln f(\mathcal{X}|\boldsymbol{\theta}^{(t)}) = Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) - Q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}) + H^{\times}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) - H^{\times}(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}) \tag{3.10}$$

Iskoristimo sada Gibbsovu nejednakost prema kojoj za dvije vjerojatnosne distribucije p i q vrijedi

$$\sum_{x \in \mathcal{X}} p(x) \log p(x) \geq \sum_{x \in \mathcal{X}} p(x) \log q(x).$$

Ona nam daje $H^{\times}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) \geq H^{\times}(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)})$ pa možemo zaključiti da vrijedi

$$\ln f(\mathcal{X}|\boldsymbol{\theta}) - \ln f(\mathcal{X}|\boldsymbol{\theta}^{(t)}) \geq Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) - Q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}) \tag{3.11}$$

Dakle, odabirom parametra $\boldsymbol{\theta}$ koji poboljšava $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$, nepotpuna log-vjerodostojnost će se barem jednako poboljšati. □

Nadalje, primjetimo da EM algoritam svoj naziv duguje činjenici da njegov $i-$ti korak maksimizira očekivanu log-vjerodostojnost $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t+i-1)})$ koju smo prethodno dobili za $\boldsymbol{\theta}^{(t+i-1)}$. Neka je $\{\boldsymbol{\theta}^{(t+i)}\}_{i\geq 1}$ niz parametara koji zadovoljavaju relaciju (3.6). Važno svojstvo tog niza daje nam teorem [10], koji navodimo bez dokaza budući da je on analogan dokazu prethodne propozicije.

Teorem 3.2.2. *Niz* $\{\boldsymbol{\theta}^{(t+i)}\}_{i\geq 1}$ *zadovoljava*

$$\mathcal{L}(\boldsymbol{\theta}^{(t+i+1)}|\mathcal{X}) \geq \mathcal{L}(\boldsymbol{\theta}^{(t+i)}|\mathcal{X}), \tag{3.12}$$

gdje jednakost vrijedi ako i samo ako je $Q(\boldsymbol{\theta}^{(t+i+1)}|\boldsymbol{\theta}^{(t+i)}) = Q(\boldsymbol{\theta}^{(t+i)}|\boldsymbol{\theta}^{(t+i)})$.

Iako nam Teorem 3.2.2 garantira da se log-vjerodostojnost povećava svakom iteracijom, na temelju navedenih rezultata još uvijek ne možemo zaključiti da niz $\{\boldsymbol{\theta}^{(t+i)}\}_{i\geq 1}$ konvergira točki maksimuma. Ipak, uz neke dodatne pretpostavke, osigurana je konvergencija prema stacionarnoj točki, koja je ili lokalni maksimum ili sedlasta točka. Taj rezultat je iskazan u sljedećem teoremu kojeg navodimo bez dokaza [10].

Teorem 3.2.3. *Ako je očekivana log-vjerodostjnost* $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ *neprekidna u* $\boldsymbol{\theta}$ *i* $\boldsymbol{\theta}^{(t)}$, *onda su svi limesi niza* $\{\boldsymbol{\theta}^{(t+i)}\}_{i\geq 1}$ *stacionarne točke nepotpune log-vjerodostojnost* $\mathcal{L}(\boldsymbol{\theta}|\mathcal{X})$ *te* $\mathcal{L}(\boldsymbol{\theta}^{(t+i)}|\mathcal{X})$ *monotono konvergira prema* $\mathcal{L}(\hat{\boldsymbol{\theta}}|\mathcal{X})$, *za neku stacionrnu točku* $\hat{\boldsymbol{\theta}}$.

Dakle, prilikom primjene najprije incijaliziramo parametre, a zatim alterniramo E-korak i M-korak sve dok algoritam ne konvergira.
Prethodno opisani algoritam maksimizacija očekivanja predstavlja probabilistički pristup grupiranju budući da primjeri pripadaju grupama s određenom vjerojatnošću te pojedini primjer može pripadati više grupa. Takav tip klasifikacije zovemo meko grupiranje (engl. *soft clustering*). Drugi tip klasifikacije kod koje svaki primjer pripada isključivo jednom grupi zovemo čvrsto grupiranje (engl. *hard clustering*).

3.3 Primjena algoritma maksimizacije očekivanja na movMF

Pokažimo sada kako prethodno opisani EM algoritam možemo primijeniti na konveksnu kombinaciju vMF distribucija. Neka je $\mathcal{X} = \{x_1, \ldots, x_n\}$ skup primjera, a $\mathcal{Z} = \{z_1, \ldots, z_n\}$ skup skrivenih varijabli koje određuju kojoj komponenti pripada pojedini primjer iz $\mathcal{X}$. Dakle, vrijedi da je $z_i = k,\ i = 1, \ldots, n,\ k = 1, \ldots, K$ ako $i-$ti primjer x_i dolazi iz $k-$te komponente. Nadalje, za svaku latentnu varijablu vrijedi $P(z_i = k) = \alpha_k$ pa imamo

$$P(\mathcal{X}|\mathcal{Z}, \boldsymbol{\theta}) = \prod_{i=1}^{n} P(z_i) = \prod_{i=1}^{n} \prod_{k=1}^{K} \alpha_k^{\mathbb{I}_{(z_i=k)}}, \tag{3.13}$$

gdje je

$$\mathbb{I}_{(z_i=k)} = \begin{cases} 1, & \text{za } z_i = k \\ 0, & \text{inače.} \end{cases} \tag{3.14}$$

Osim toga vrijedi

$$P(\mathcal{Z}) = \prod_{i=1}^{n} P(z_i) = \prod_{i=1}^{n} \prod_{k=1}^{K} f(x_i|\theta_k)^{\mathbb{I}_{(z_i=k)}}. \tag{3.15}$$

Sada zajedničku gustoću možemo izraziti preko (3.13) i (3.15) i ona je jednaka

$$\begin{aligned} f(\mathcal{X}, \mathcal{Z}|\boldsymbol{\theta}) &= P(\mathcal{Z}) p(\mathcal{X}|\mathcal{Z}) \\ &= \prod_{i=1}^{n} \prod_{k=1}^{K} \alpha_k^{\mathbb{I}_{(z_i=k)}} \prod_{i=1}^{n} \prod_{k=1}^{K} f(\mathcal{X}|\theta_k)^{\mathbb{I}_{(z_i=k)}} \\ &= \prod_{i=1}^{n} \prod_{k=1}^{K} (\alpha_k f(\mathcal{X}|\theta_k))^{\mathbb{I}_{(z_i=k)}}. \end{aligned} \tag{3.16}$$

Iskoristimo dobiveni rezultat kako bismo izveli izraz za potpunu log-vjerodostojnost danog modela. Ona glasi

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}|\mathcal{X},\mathcal{Z}) &= \ln f(\mathcal{X},\mathcal{Z}|\boldsymbol{\theta}) \\ &= \ln \prod_{i=1}^{n}\prod_{k=1}^{K}(\alpha_k f(\mathcal{X}|\theta_k))^{\mathbb{I}(z_i=k)} \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K}\mathbb{I}_{(z_i=k)}(\ln\alpha_k + \ln f(x_i|\theta_k)).\end{aligned} \tag{3.17}$$

Ako usporedimo dobivemo s polaznim izrazom (3.4) za nepotpunu log-vjerodostojnost, uočavamo da smo se riješili logaritma sume koji nam je predstavljao problem te je time rješavanje optimizacijskog problema znatno pojednostavljeno. Provedimo sada E-korak i M-korak algoritma.

E-korak

Kao što je ranije navedeno, u E-koraku tražimo očekivanje $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$. Imamo

$$\begin{aligned}Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &= \mathbb{E}_{\mathcal{Z}|(\mathcal{X},\boldsymbol{\theta}^{(t)})}[\mathcal{L}(\boldsymbol{\theta}|\mathcal{X},\mathcal{Z})] \\ &= \mathbb{E}_{\mathcal{Z}|(\mathcal{X},\boldsymbol{\theta}^{(t)})}\Big[\sum_{i=1}^{n}\sum_{k=1}^{K}\mathbb{I}_{(z_i=k)}(\ln\alpha_k + \ln f(x_i|\theta_k))\Big] \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K}\mathbb{E}_{z_i|(\mathcal{X},\boldsymbol{\theta}^{(t)})}[\mathbb{I}_{(z_i=k)}(\ln\alpha_k + \ln f(x_i|\theta_k))] \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K}p(k|x_i,\boldsymbol{\theta})(\ln\alpha_k + \ln f(x_i|\theta_k)) \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K}p(k|x_i,\boldsymbol{\theta})(\ln\alpha_k) + \sum_{i=1}^{n}\sum_{k=1}^{K}p(k|x_i,\boldsymbol{\theta})(\ln f(x_i|\theta_k)),\end{aligned} \tag{3.18}$$

gdje koristimo linearnost očekivanja, nezavisnost i jednaku distribuiranost primjera iz $\mathcal{X}$ te činjenicu da je očekivanje skrivene varijable jednako njezinoj aposteriornoj vjerojatnosti koju zatim možemo izračunati primjenom Bayesovog teorema. Dakle, vrijedi

$$p(k|x_i,\boldsymbol{\theta}) = \frac{f(x_i|k;\boldsymbol{\theta})P(k|\boldsymbol{\theta})}{\sum_{i=1}^{n} f(x_i|k;\boldsymbol{\theta})P(k|\boldsymbol{\theta})} = \frac{f(x_i|\theta_k)\alpha_k}{\sum_{k=1}^{K} f(x_i|\theta_k)\alpha_l}. \tag{3.19}$$

M-korak

U M-koraku algoritma računamo (3.18), tj. maksimiziramo gornji izraz kako bismo dobili nove procjene parametara $\boldsymbol{\theta}$. Primjetimo da u tom postupku maksimizacije očekivanja dane log-vjerodostojnosti možemo posebno maksimizirati članove koji sadrže α_k te one koji sadrže θ_k budući da su nezavisni.

Najprije nam je cilj maksimizirali (3.18) s obzirom na α_k, uz ograničenje $\sum_k \alpha_k = 1$. Pri tome drugu sumu u izrazu (3.18) zanemarujemo budući da ne ovisi o parametrima α_k. Dakle, u tu svrhu formiramo sljedeću Lagrangeovu funkciju

$$L(\alpha_k, \lambda) = \sum_{k=1}^{K}\sum_{i=1}^{n}(\ln \alpha_k)p(k|x_i, \boldsymbol{\theta}) - \lambda\left(\sum_{k=1}^{K}\alpha_k - 1\right) \tag{3.20}$$

gdje λ označava Lagrangeov multiplikator.
Deriviramo izraz (3.20) po parametrima α_k i izjednačimo s nulom pa dobivamo

$$\sum_{i=1}^{n} p(k|x_i, \boldsymbol{\theta}) + \lambda\alpha_k = 0. \tag{3.21}$$

Dalje, sumiranjem relacije (3.21) po k nalazimo $\lambda = -n$, iz čega slijedi

$$\hat{\alpha_k} = \frac{1}{n}\sum_{i=1}^{n} p(k|x_i, \boldsymbol{\theta}). \tag{3.22}$$

Promotrimo sada maksimizaciju člana koji sadrži parametre $\theta_k = (\mu_k, \kappa_k)$ uz uvjete $\mu_k^T\mu_k = 1$ i $\kappa_k \geq 0$. U tom slučaju promatramo Lagrangeovu funkciju danu sa

$$\begin{aligned} L(\{\mu_k, \kappa_k, \lambda_k\}_{k=1}^K) &= \sum_{k=1}^{K}\sum_{i=1}^{n}(\ln f_k(x_i|\theta_k))p(h|x_i, \boldsymbol{\theta}) + \sum_{k=1}^{K}\lambda_h(1 - \mu_k^T\mu_k) \\ &= \sum_{k=1}^{K}\left[\sum_{i=1}^{n}(\ln c_d(\kappa_k))p(k|x_i, \boldsymbol{\theta}) + \sum_{i=1}^{n}\kappa_k\mu_k^T x_i + p(k|x_i, \boldsymbol{\theta}) + \lambda_k(1 - \mu_k^T\mu_k)\right]. \end{aligned} \tag{3.23}$$

Slično kao u prethodnom slučaju, deriviramo (3.23) po parametrima $\{\mu_k, \kappa_k, \lambda_k\}_{k=1}^K$ te izjednačimo s nulom. Tada za svaki k slijedi

$$\mu_k = \frac{\kappa_k}{2\lambda_k}\sum_{i=1}^{n} x_i p(k|x_i, \boldsymbol{\theta}), \tag{3.24}$$

$$\mu_k^T\mu_k = 1, \tag{3.25}$$

$$\frac{c_d'(\kappa_k)}{c_d(\kappa_k)} \sum_{i=1}^{n} p(k|x_i, \boldsymbol{\theta}) = -\mu_k^T \sum_{i=1}^{n} x_i p(k|x_i, \boldsymbol{\theta}). \tag{3.26}$$

Konačno, iz (3.24) i (3.25) dobivamo

$$\lambda_k = \frac{\kappa_k}{2} \left\| \sum_{i=1}^{n} x_i p(k|x_i, \boldsymbol{\theta}) \right\|, \tag{3.27}$$

$$\mu_k = \frac{\sum_{i=1}^{n} x_i p(k|x_i, \boldsymbol{\theta})}{\| \sum_{i=1}^{n} x_i p(k|x_i, \boldsymbol{\theta}) \|} = \frac{r_k}{\|r_{\boldsymbol{\theta}}\|}. \tag{3.28}$$

Nadalje, uvršavanje (3.28) u (3.26) nam daje

$$\frac{c_d'(\kappa_k)}{c_d(\kappa_k)} = -\frac{\| \sum_{i=1}^{n} x_i p(k|x_i, \boldsymbol{\theta}) \|}{\sum_{i=1}^{n} p(k|x_i, \boldsymbol{\theta})} \tag{3.29}$$

što možemo pisati u sljedećem obliku

$$A_d(\kappa_k) = \frac{\| \sum_{i=1}^{n} x_i p(k|x_i, \boldsymbol{\theta}) \|}{\sum_{i=1}^{n} p(k|x_i, \boldsymbol{\theta})}, \tag{3.30}$$

gdje je $A_d(\kappa_k) = \frac{I_{d/2}(\kappa_k)}{I_{d/2-1}(\kappa_k)}$. Primijetimo da su formule dane s (3.28) i (3.30) poopćenja izraza (2.13) i (2.18) te zbog toga svaki κ_k možemo procijeniti postupkom analognom onome koji je opisan u Poglavlju 2.2. Pseudo-kod opisanog EM algoritma je dan u Algoritmu 1.

Algorithm 1: EM algoritam

Input : Skup primjera $\mathcal{X}$
Output: Optimalni parametri α_k, μ_k, κ_k, $k = 1, \ldots, K$
Inicijalizacija parametara α_k, μ_k, κ_k, $k = 1, \ldots, K$
repeat
  (E-korak)
  for *i=1 to n* **do**
    for *k=1 to K* **do**
      $f(x_i|\theta_k) \leftarrow c_d(\kappa_k)e^{\kappa_k \mu_k^T x_i}$
    end
    for *k=1 to K* **do**
      $P(k|x_i, \theta_k) \leftarrow \frac{\alpha_k f(x_i|\theta_k)}{\sum_{l=1}^{K} \alpha_l f(x_i|\theta_l)}$
    end
  end
  (M-korak)
  for *k=1 to K* **do**
    $\alpha_k \leftarrow \frac{1}{n} \sum_{i=1}^{n} P(k|x_i, \Theta)$
    $\mu_k \leftarrow \sum_{i=1}^{n} x_i P(k|x_i, \Theta)$
    $\overline{r} \leftarrow \frac{\|\mu_k\|}{n\alpha_k}$
    $\mu_k \leftarrow \frac{\mu_k}{\|\mu_k\|}$
    $\kappa_k \leftarrow \frac{\overline{r}d - \overline{r}^3}{1 - \overline{r}^2}$
  end
until *konvergencija*;

Promotrimo još dva važna koraka algoritma, a to su inicijalizacija parametara i uvjet konvergencije. U prethodnom poglavlju smo pokazali da EM algoritam nužno konvergira, ali ne nužno i u globalni optimum i stoga nam je inicijalizacija paramatera vrlo važan korak budući da će krajnji rezultat ovisiti o početno izabranim parametrima.
Navedimo stoga najprije neke standardne metode inicijalizacije.

(i) Najjednostavnija metoda je nasumičan odabir početnih vrijednosti parametara modela

(ii) Sljedeća najčešće korištena metoda je algoritam *k*-sredina (engl. *k-means*). Cilj ovog algoritma je minimizacija srednje kvadratne greške

$$E = \frac{1}{n} \sum_{x} \|x - \mu_{h(x)}\|^2,$$

gdje je $h(x) = \underset{h \in \{1,\ldots,k\}}{\operatorname{argmin}} \|x - \mu_h\|$ indeks grupe koja je najbliža primjeru x.

(iii) Zadnja metoda koju ćemo navesti je algoritam sfernih k-sredina, koji je najpogodniji za ovakav klasifikacijski problem budući koristi kosinusnu sličnost, a ta mjera se pokazala superiornom u odnosu na euklidsku metriku. Dakle, osnovna ideja ovog algoritma je slična kao kod algoritma k-sredina, samo što u ovo slučaju maksimizirano kosinusnu sličnost

$$L = \sum_{x} x^T \mu_{h(x)},$$

gdje je $h(x) = \underset{h \in \{1,\dots,k\}}{\operatorname{argmax}} x^T \mu_h$.

Nadalje, običajeni kriteriji za provjeru konvergencije EM algoritma su sljedeći:

(i) konvergencija log-vjerodostojnosti, tj. ako je apsolutna razlika vrijednost log-vjerodostojnosti manja od nekog odabranog ε_1, koji može biti proizvoljno malen,

(ii) konvergencija parametara, tj. apsolutna razlika vrijednost parametara mora biti manja od nekog ε_2, koji je također proizvoljno malen.

Poglavlje 4

Klasifikacija teksta primjenom vMF distribucije

Općenito, pod pojmom klasifikacije podrazumjevamo onaj tip problema kod kojeg se podaci grupiraju u skupine na osnovi njihovih zajedničkih karakteristika. Kao što smo i ranije naveli, korištenje von Mises-Fisherove distribucije, koja služi kao model za podatke koji imaju smjer, pokazalo se pogodnim za klasifikaciju teksta budući da kod takvog tipa problema veliku uloga igra i orijentacija podataka, a ne samo njihov opseg. Ilustrativni primjeri koje ćemo prikazati u ovom poglavlju su rekonstrukcije ekperimenata opisanih u [3] te ćemo se pritom baviti s dva tipa problema. Prvi je simulacija podataka čija je primarna svrha razviti intuiciju o ponašanju algoritama baziranih na vMF distribuciji te utvrditi njihovu korektnost, a zatim ćemo pokazati kako oni funkcioniraju u praksi prilikom klasifikacije velike količine tekstualnih podataka ovisno njihovoj tematici. Pritom koristimo programski jezik Python, koji je danas jedan od najzastupljenijih programskih jezika, posebno u području strojnog učenja. Njegova popularnost, između ostalog, leži u velikom broju biblioteka koje korisnicima znatno pojednostavljuju modeliranje.

4.1 Simulacija podataka

4.1.1 Simulacija podataka iz vMF distribucije

Prvi eksperiment sastoji se od generiranja podataka iz von Mises-Fisherove distribucije i procjene parametara. Ovaj korak nam je bitan za daljnju izgradnju klasifikacijskog algoritma budući da je EM algoritam temeljen na procjenama parametara vMF distribucije, koje dobivamo metodom maksimalne vjerodostojnosti opisanom u Poglavlju 2.2. Simulacije su izvedene na temelju postupka iz [7], čiji je psedudokod dan u Algoritmu 2.

Algorithm 2: Simulacija podataka iz vMF distribucije

Input : n - veličina uzorka; μ, κ - parametri vMF distribucije
Output: Uzorak $X = \{x_1, \ldots, x_n\}$

$d \leftarrow dim(\mu)$
$t_1 \leftarrow \sqrt{4\kappa^2 + (d-1)^2}$
$b \leftarrow \frac{-2\kappa + t_1}{d-1}$
$x_0 \leftarrow \frac{b-1}{b+1}$
$X \leftarrow zeros(n, d)$
$m \leftarrow \frac{d-1}{2}$
$c \leftarrow \kappa x_0 + (d-1)\log(1 - x_0^2)$
for *h=1 to n* **do**
    $t \leftarrow -1000$
    $u \leftarrow 1$
    while $t < \log u$ **do**
        $z \leftarrow Beta(m, m)$
        $u \leftarrow Unif([0, 1])$
        $w \leftarrow \frac{1-(1+b)z}{1-(1-b)z}$
        $t \leftarrow \kappa w + (d-1)\log(1 - x_0 w)$
    end
    $v \leftarrow U(d-1)$
    $v \leftarrow \frac{v}{\|v\|}$
    $X(i, 1 : d-1) \leftarrow \sqrt{1-w}v^T$
    $X(i, d) \leftarrow w$
end
Provedi ortogonalnu transformaciju na svakom uzorku iz X, gdje je uzorak pohranjen u matricu X po retcima
Vrati X

Primjer

Dakle, u ovom primjeru generirali smo nekoliko skupova podataka iz 3-dimenzionalne vMF distrubucije, a zatim na njima proveli MLE procjenu parametara modela. Dobiveni rezultati su prikazanu u Tablici 4.1, gdje n označava duljinu simuliranog uzorka te μ i κ stvarne vrijednosti parametra, dok su $\hat{\mu}$ i $\hat{\kappa}$ njihovi procjenitelji.

Sada vidimo da su razlike između stvarnih vrijednosti i njihovih procjena male, čime zaključujemo da dani procjenitelji uspješno aproksimiraju parametre modela.

n	μ	$\hat{\mu}$	$\mu^T\mu$	κ	$\hat{\kappa}$
100	[0.3660, 0.8091, 0.4597]	[0.3522, 0.8408, 0.4112]	0.9982	5	4.88
1000	[0.2910, 0.9383, 0.1868]	[0.2894 , 0.9500, 0.1172]	0.9975	2	2.08
1000	[0.9013, 0.1886, 0.3899]	[0.8997, 0.1836, 0.3961]	0.9999	10	10.04

Tablica 4.1: MLE za 3-dimenzionalnu vMF distribuciju

4.1.2 Simulacija podataka iz konveksne kombinacije vMF distribucija

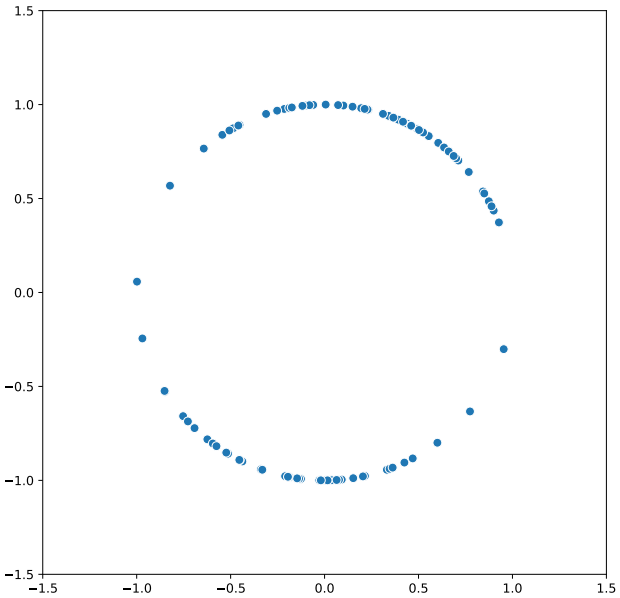
Mali skup podataka

Sljedeći korak je promotriti kako se ponaša EM algoritam iz Poglavlja 3.3 na konveksnoj kombinaciji vMF distribucija. Ovdje smo simulirali 100 točaka iz dvije vMF distribucije, koristeći Algoritam 2 na sljedeći način: najprije odabremo oznaku grupe k i pripadni koeficijent k-te komponente α_k, a zatim simuliramo podatke iz odgovarajuće vMF distribucije s parametrom (μ_k, κ_k). Cilj ovog primjera je dočarati kako funkcionira meko grupiranje podataka i pritom se ograničavamo na 2-dimenzionalni slučaj radi lakše vizualizacije. Dakle, iz grafičkog prikaza simuliranih podataka (Slika (4.1)) vidimo da se oni prirodno grupiraju te je za većinu točaka vrlo jasno kojoj od dvije komponente pripada, dok za njih par nije na prvu vidljivo iz koje distrubucije potječu. U tom slučaju, EM algoriram detektira takve točke te iz djelomično svrsta u svaku od komponenti.

Promotrimo sada Sliku 4.2 koja prikazuje grupirane podatke. Primjetimo da točke koje vizualnom inspekcijom nismo mogli sa sigurnošću klasificirati zaista djelomično pripadaju svakoj grupi. U praksi, posebno kod grupiranja teksta, često nam je bitno imati alat koji omogućuje meko grupiranje budući da se kategorije često mogu preklapati te ovim zapravo postižemo točniju klasifikaciju, jer se ne opredjeljujemo samo za jednu grupu.
Kao i u prethodnom primjeru, stvarne i procjenjene vrijednost parametara prikazujemo u Tablici 4.2 te na temelju dobivenih vrijednosti zaključujemo da je implementacija EM algoritma korektna.

Grupa	μ	$\hat{\mu}$	$\mu^T\mu$	κ	$\hat{\kappa}$	α	$\hat{\alpha}$
1	[-0.251, -0.968]	[-0.1912 , -0.9816]	0.9981	4	4.06	0.48	0.4647
2	[0.399, 0.917]	[0.2917, 0.9565]	0.9935	4	3.85	0.52	0.5352

Tablica 4.2: EM procjene parametara za konveksnu kombinaciju dviju vMF distribucija



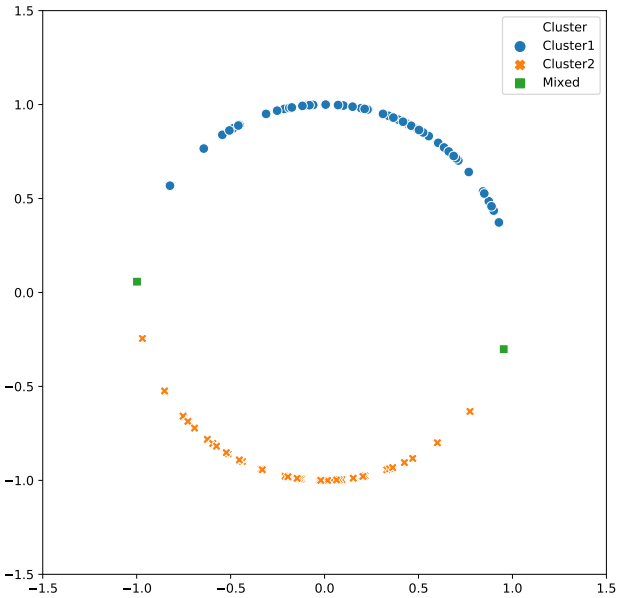
Slika 4.1: Orginalni skup simuliranih podataka

Veliki skup podataka

Za kraj još navodimo jedan kratak primjer kako bismo se uvjerili u korektnost EM algoritma na većoj količini podataka. Ovdje simuliramo 3000 točaka dimenzije 20 iz konvenksne kombinacije četiriju vMF distribucija. Iz rezultata prikazanih u Tablici 4.3 vidimo da je točnost procjenjnih parametara nešto slabija nego u prethodnim primjerima, ali je i dalje zadovoljavajuća.

Grupa	κ	$\hat{\kappa}$	α	$\hat{\alpha}$	$\mu^T\mu$
1	14	13.96	0.16	0.1968	0.9707
2	9	9.29	0.28	0.2751	0.9522
3	11	11.35	0.31	0.3186	0.8913
4	6	6.27	0.25	0.2657	0.9395

Tablica 4.3: EM procjene parametara za konveksnu kombinaciju četiri vMF distribucije



Slika 4.2: Grupirani skup simuliranih podataka

4.2 Klasifikacija tekstualnih podataka

4.2.1 Predprocesiranje teksta

Općenito, jedna od glavnih karakteristika tesktualnih podataka je njegova visoka dimenzionalost koja je često otežavajući faktor u analizi. Stoga prije samog modeliranja teksta nastojimo provesti određene transformacije s ciljem reduciranja dimenzionalnosti i taj korak nazivamo predprocesiranje. Navedimo neke uobičajane korake predprocesiranja teksta. Mnogi programski jezici, među kojima je i Python, ne razlikuju velika i mala slova pa je potrebno riječi svesti na isti oblik prebacivanjem svih velikih slova u mala, ili obrnuto, kako računalo ne bi tretiralo, npr. riječi ”diplomski” i ”Diplomski” kao dvije različite. Sljedeći korak je uklanjanje neinformativnih riječi. Prvi takav tip su tzv. stop-riječi (engl. *stop-words*), pod kojima podrazumjevavo riječi poput veznika koji sami po sebi nemaju neko posebno značenje već pomažu pri izgradnji rečenica. Drugi tip neinformativnih riječi u kontekstu klasifikacije teksta su one koje se pojavljuju premali ili prevelik broj puta te time ne daju nikakvu konkretnu informaciju koju bismo mogli povezati sa tematikom teksta.

Također, dobra je praksa ukloniti interpunkcijske i ostale posebne znakove, osim u slučaju kada unaprijed znamo da nam određeni znakovi nose neku relevantnu informaciju.

4.2.2 Vektorska reprezentacija teksta

Kako bismo uopće mogli modelirati tekst potrebno je odabrati prigodnu reprezentaciju koja nam omogućuje očuvanje informacija koje su njime dane, ali i efektivno računanje. Tekstualim podacima smatramo kolekciju dokumenata (to mogu biti članci, knjige i slično) te ćemo riječi koje se u njemu pojavljuju nazivati pojmovi.

Uobičajena je praksa koristiti vektorsku reprezentaciju, čija je ideja prezentirati svaki dokument kao vektor, pri čemu elementi vektora predstavljaju težine pojmova koji se u njemu pojavljuju. Sada se, naravno, postavlja pitanje kako odrediti težine pojmova kako bi optimirali njihovu informativnost. Najintuitivniji pristup bio bi svakom pojmu dodijeliti težinu ovisno o frekvenciji njegovog pojavljivanja u dokumentu, no takav način reprezenatcije često nije dobar izbor budući da nam ne daje stvarnu informaciju o važnosti pojma za određeni dokument pa ga je potrebno adaptirati i to tako što ćemo uzeti u obzir relevantnost pojma. Jedan od takvih modela reprezentacije opisan je u [6] te je poznat kao tf_idf (engl. *term frequency-inverse document frequency*).

Neka $\mathrm{tf}_{t,d}$ predstavlja broj pojavljivanja pojma t u dokumentu d (engl. *term frequency*). Dakle, što je ova vrijednost veća, pojam je relevantniji za promatrani dokument i time potencijalno dolazimo do problema budući da se može dogoditi da se taj pojam, koji smatrano relevantnim za dani dokument, često se pojavljuje i u drugima, čime je znatno smanjena njegova informativnost i zato je potrebno skalirati njegovu težinu. Dalje, označimo sa df_t broj dokumenata u kojima se pojavljuje pojam t. Sada možemo definirati inverznu dokumentnu frekvenciju (eng *inverse document frequency*) sljedećom formulom

$$\mathrm{idf}_t = \log \frac{n}{\mathrm{df}_t}, \tag{4.1}$$

gdje je n broj dokumenata u kolekciji.

Prema tome, idf_t vrijednost pojma t će biti mala ako se on pojavljuje u mnogo dokumenata, što smo i željeli postići budući da takvi pojmovi ne nose u sebi dovoljno informacija za diferenciranje između tema. S druge strane, ta će vrijednost biti velika ako je pojam prisutan samo u nekim dokumentima, što indicira da je karakterističan za određeno područje i time nam olakšava klasificiranje. Sada, kako nam je cilj da obje vrijednost $\mathrm{tf}_{t,d}$ i idf_t pojma t budu visoke, dolazimo da sljedećeg izraza za težinu

$$\mathrm{tf\_idf}_{t,d} = \mathrm{tf}_{t,d} \cdot \mathrm{idf}_t. \tag{4.2}$$

4.2.3 Evaluacija modela

Banerjee i suradnici [3] sugeriraju evaluaciju modela mjerom koju zovemo *uzajamna informacija* (engl. *mutual information*) te je i u ovom radu korišten isti pristup. Najveća prednost ove metrike je što je nezavisna od odabira oznaka grupa, tj. permutacije unutar skupa oznaka ne utječu na njezinu vrijednost što nam je za sljedeće primjere posebno bitno budući stvarne i predviđene grupe nisu označane jednako. Navodimo njezinu matematički formulaciju.

Pretpostavimo da imamo dva skupa oznaka X i Y jednake duljine n. Njihova entropija je definirana sljedećim izrazima

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i, \tag{4.3}$$

gdje je $p_i = |X_i|/n$. vjerojatnost da proizvoljan primjer iz X pripada grupi X_i, te analogno za Y,

$$H(Y) = -\sum_{j=1}^{n} q_j \log q_j, \tag{4.4}$$

uz $q_j = |Y_j|/n$.
Sada uzajamnu informaciju MI između X i Y računamo pomoću formule

$$I(X, Y) = \sum_{i=1}^{n} \sum_{j=1}^{n} p_{i,j} \log \left(\frac{p_{i,j}}{p_i q_j} \right). \tag{4.5}$$

Ovdje $p_{i,j} = |X_i \cap Y_j|/n$ označava vjerojatnost da proizvoljno odabran primjer pada u obje grupe X_i i Y_j. Napomenimo još da je za evaluaciju modela u ovom radu korištena funkcija `mutual_info_score`[1] iz Pythonovog modula `sklearn.metrics`.

4.2.4 Primjeri

Kolekcija dokumenata koju ćemo modelirati poznata je pod nazivom News20 te je jedan od standardnih primjera u domeni strojnog učenja za probleme klasifikacije i obrade prirodnog jezika. Ona se sastoji od 18846 poruka iz 20 različitih tematskih grupa na USENET-u te posjeduje sve nezgodne karakteristike tekstualnih podataka: visoku dimenzionalost, rijetku reprezentaciju te preklapanje tema.

[1] `https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mutual_info_score.html`

Zbog svoje popularnosti sadržana je i u Pythonovoj biblioteci `sklearn.datasets`[2], što nam znatno olakšava njezino učitavanje. Transformacija u vektorizirani tf_idf oblik napravljena je pomoću funkcije `TfidfVectorizer`[3]. Navedena funkcija obavlja i velik dio predprocesiranja. Naime, definiranjem posebnih parametara moguće je provesti pretvorbu malih slova u velika, ukloniti specijalne znakove, te riječi koje se pojavljuju nedovoljan ili prevelik broj puta. Također, budući da su tekstovi koje klasificiramo na engleskom jeziku, uklonili smo i engleske stop-riječi.

Zatim ćemo podijeliti dani skup podataka na dva podskupa kako bismo dio podataka mogli iskoristiti za učenje parametara modela, tj. izgradnju movMF modela za svaku grupu, a ostatak za testiranje kako bismo odredili vjerojatnost da novi primjer pripada pojedinoj grupi. Dakle, formiramo skup za treniranje i skup za testiranje tako da prvi skup sadrži 70% podataka, a drugi 30% i pritom koristimo standardu funkciju Pythonovu funkciju `train_test_split` [4] iz modula `sklearn.model_selection`.

Nadalje, koristimo EM algoritam iz paketa `sphereclustering`[5], čija je implementacija također napravljena po uzoru na [3], zbog njegove optimizacije za podatke visokih dimenzija. Budući da EM algoritam daje meko grupiranje, tj. vjerojatnost da primjer pripada pojedinoj grupi, rezultat je potrebno adaptirati kako bismo mogli provesti evaluaciju modela pomoći MI vrijednosti. To činimo tako da meko grupiranje transformiramo u tvrdo na način da za oznaku grupe primjera uzmemu onu s najvećom aposteriornom vjerojatnosti.

Za kraj naglasimo još da smo kao konačan rezultat MI vrijednosti prezentirali prosječan rezultat dobiven nakon 5 izvođenja modela s proizvoljnom inicijalizacijom parametara.

4.2.4.1 News20 skup podataka

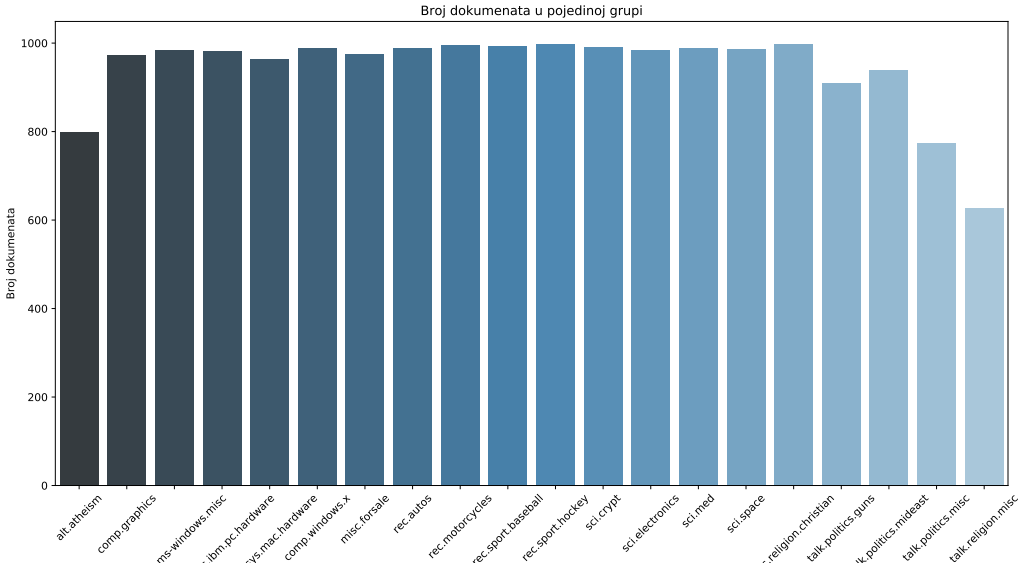
Prvi primjer odnosi se na klasifikaciju orginalnog News20 skupa podataka, koji, nakon provođenja navedenih transformacija daje vektorski prostor dimenzije (18846, 21362). Kao što vidimo na Slici 4.3, grupe sadrže podjednaki broj dokumenata, što je važan podatak jer je podjednaka zastupljenost tekstova između grupa bitan zahtjev za točnost klasifikacije. Promotrimo na Slici 4.4 kako se ponašaju MI vrijednosti ovisno o broju grupa. Nju računamo između skupa X koji se sastoji od stvarnih, unaprijed poznatih oznaka grupa i skupa Y, čiji su elementi predviđene oznake grupa koje dobivamo iz izgrađenog modela. Naime, vidimo da MI vrijednost za oba skupa raste dok ne dođemo do stvarnog broja grupa nakon čega se rast usporava pa možemo zaključujemo da dani model uspješno kategorizira

[2] `https://scikit-learn.org/0.19/datasets/twenty_newsgroups.html`

[3] `https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html`

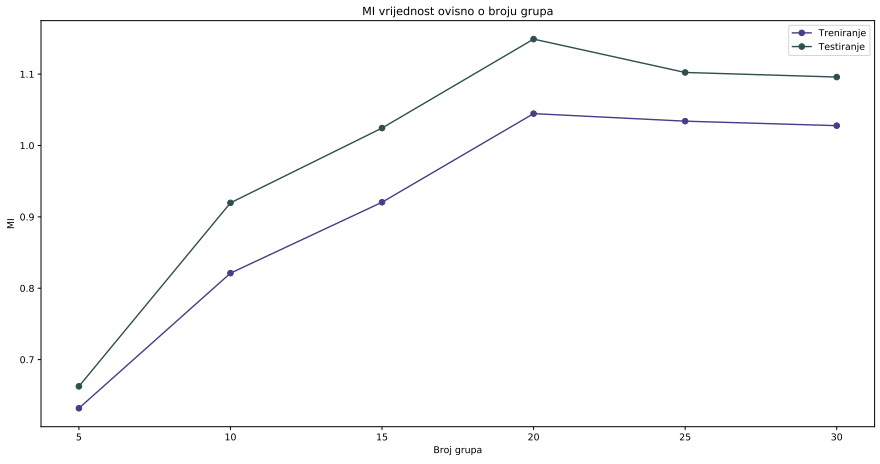
[4] `https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html`

[5] `https://github.com/jasonlaska/spherecluster`



Slika 4.3: Broj tekstova iz pojedine teme u News20 datasetu

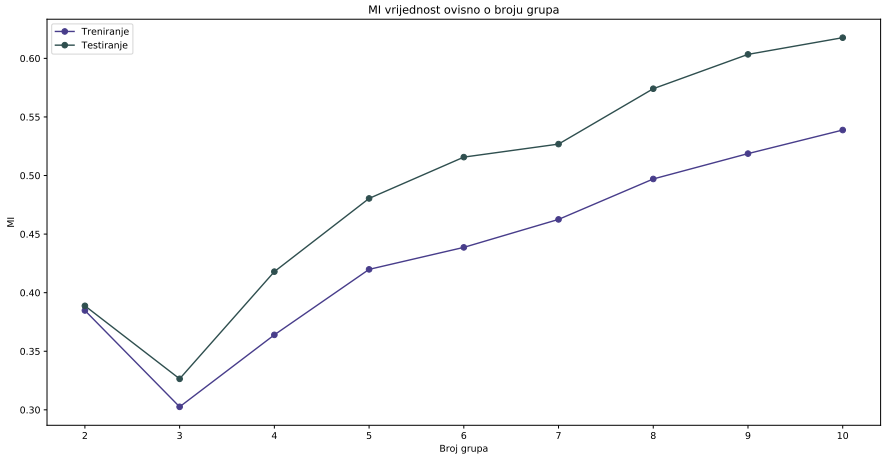
tekstove u predviđene grupe te da će novi primjeri biti svrstanu u odgovarajuću grupu čak s većom točnošću.



Slika 4.4: MI vrijednosti ovisno o broju grupa

4.2.4.2 Podskupovi News20 skup podataka

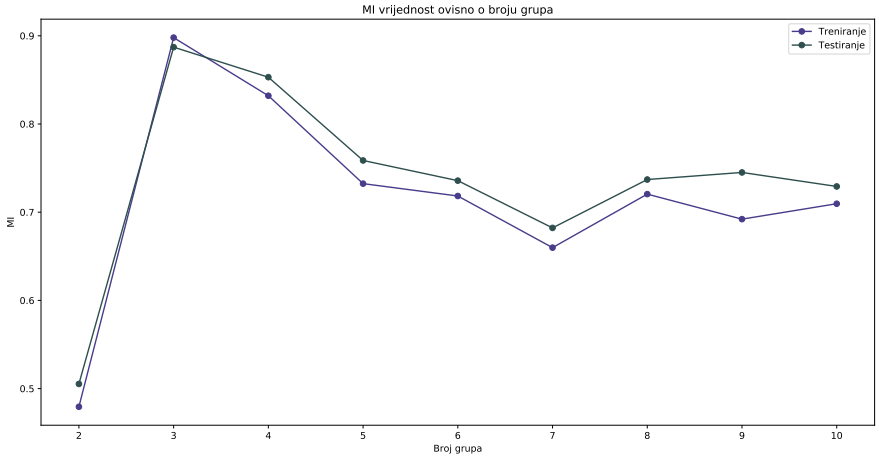
Sljedeće što ćemo napraviti je iz orginalnog seta podataka kreirati dva podskupa od kojih će prvi sadržavati tekstove iz 3 vrlo srodne teme, dok će se drugi sastoji od tekstova iz 3 različite teme. Dokumenti srodne tematike podijeljeni su u sljedeće kategorije: talk.politics.guns, talk.politics.mideast, talk.politics.misc te, nakon transformacija, daju vektorski prostor dimenzija (2625, 2920). Slika 4.5 prikazuje evaluaciju modela pomoću MI vrijednosti ovisno o broju grupa. Vidimo da model dosta loše prepoznaje pravi broj grupa, kako prilikom treniranja tako i prilikom testiranja, no kako ih povećavamo MI vrijednost se također povećava, što sugerira da bi bilo bolje podatke klasificirati u više skupina nego što je inicijalno predviđeno. To zapravo i ima smisla jer su teme jako slične i preklapaju se pa bi ih bilo bi bolje staviti ih u što manje kategorija ili ipak kreirati više grupa kako bi se naglasila razlika među njima. Baš u ovakvim slučajevima posebno dolaze do izražaja prednosti mekog grupiranja budući da ćemo tekstove srodne tematike vrlo često moći svrstati u više sličnih grupa.



Slika 4.5: MI vrijednosti ovisno o broju grupa za dokumente slične tematike

Drugi podskup podataka je dimenzije (2972, 1962) te obuhvaća dokumente iz idućih kategorija: rec.sport.baseball, sci.med, comp.windows.x. Promotrimo ponovo MI vrijednosti za testiranje i treniranje ovisno o broju grupa (Slika 4.5). Kod klasifikacije tekstova različite tematike uočavamo da modela dobro odrađuje svoj posao te je MI vrijednost najviša za stvaran broj grupa, nakon čega će opadati što zapravo i ne iznenađuje jer smo uzeli nesrodne teme među kojima ne bi trebalo biti previše preklapanja te očekujemo da ih

model može jasno razlikovati. Točnost klasifikacije novih primjera prati isti trend uočen na evaluaciji skupa za treniranje.



Slika 4.6: MI vrijednosti ovisno o broju grupa za dokumente različite tematike

Dodatak A

Kodovi korišteni u praktičnom dijelu rada

Simulacija uzorka iz vMF distribucije

```
def vMF(mi, kappa, n):
    """Funkcija za simulaciju uzorka iz vMF distribucije

    Input:
    - mi: Parametar mi vMF distribucije
    - kappa: Parametar kappa vMF distribucije
    - n: Duljina uzorka


    Output:
    - X: Uzorak iz vMF distribucije
    """

    d = len(mi)
    t1 = np.sqrt(4*(kappa**2) + (d-1)**2)
    b = (d-1)/(2*kappa + t1)
    x0 = (1-b)/(1+b)
    X = np.zeros((n,d))
    m = (d-1)/2
    c = kappa*x0 +(d-1)*np.log(1-x0**2)
    for i in range (0,n):
        t = -1000
        u = 1
```

```
        while (t - c < np.log(u)):
            z = np.random.beta(m,m)
            u = np.random.uniform(0,1)
            w = (1-(1+b)*z)/(1-(1-b)*z)
            t = kappa*w+(d-1)*np.log(1-x0*w)


        v = np.random.randn(d)
        proj_mu_v = mi * np.dot(mi, v) / np.linalg.norm(mi)
        orthto = v - proj_mu_v
        v = orthto / np.linalg.norm(orthto)

        X[i, :] = v * np.sqrt(1. - w**2) + w * mi

    return X
```

Simulacija uzorka iz konveksne kombinacije vMF distribucija

```
def movMF(mi, kappa, n):
    """Funkcija za simulaciju uzorka iz movMF distribucije

    Input:
    - mi: Parametar mi=(mi_1, ..., mi_K) movMF distribucije
    - kappa: Parametar kappa=(kappa_1, ..., kappa_K) movMF distribucije
    - n: Duljina uzorka


    Output:
    - X: Uzorak iz movMF distribucije
    """

    K = mi.shape[0]
    d = mi.shape[1]
    X = np.zeros((n,d))
    for i in range(n):
        label = random.randint(0, K-1)
        X[i,:] = vMF(mi[label], kappa[label], 1)

    return X
```

MLE procjena parametara vMF distribucije

```
def MLE(X):
    """Procjena parametara vMF distribucije metodom maksimalne vjerodostojnosti

    Input:
    - X: Uzorak iz vMF distribucije

    Output:
    - mi_kapica: MLE procjena parametra mi
    - kappa_kapica: MLE procjena parametra kappa
    """


    n = X.shape[0]
    d = X.shape[1]
    r = sum(X)
    r_norm = np.linalg.norm(r)
    mi_kapica = (r / r_norm)
    R = r_norm / n
    kappa_kapica_tmp = (R*d - R**3) / (1 - R**2)

    ad0 = iv(1/2*d, kappa_kapica_tmp)/iv((1/2*d)-1, kappa_kapica_tmp) -
    kappa_kapica_tmp_2 = kappa_kapica_tmp
     (ad0-R)/(1-ad0**2-(d-1)*(ad0/kappa_kapica_tmp))
    ad1 = iv(1/2*3, kappa_kapica_tmp_2)/iv(1/2*d-1, kappa_kapica_tmp_2)
    kappa_kapica = kappa_kapica_tmp_2 -
     (ad1-R)/(1-ad1**2-(d-1)*(ad1/kappa_kapica_tmp_2))

    return mi_kapica, kappa_kapica
```

Bibliografija

[1] S. Amari, *Divergence function, information monotonicity and information geometry*, In Workshop on information theoretic methods in science and engineering (WITMSE). Citeseer, 2009.

[2] S. Amari, *Information geometry and its applications*, Vol. 194. Springer, 2016.

[3] A. Banerjee, I. S. Dhillon, J. Ghosh, S. Sra, G. Ridgeway, *Clustering on the Unit Hypersphere using von Mises-Fisher Distributions*, Journal of Machine Learning Research 6, no. 9 (2005).

[4] A. Banerjee, S. Merugu, I. S. Dhillon, J. Ghosh, J. Lafferty, *Clustering with Bregman divergences*, Journal of machine learning research 6, no. 10 (2005).

[5] L. D. Brown, *Fundamentals of statistical exponential families: with applications in statistical decision theory*, Ims, 1986.

[6] G. Salton, C. Buckley, *Term-weighting approaches in automatic text retrieval*, Information processing & management 24, no. 5 (1988), str. 513-523.

[7] I. S. Dhillon, S. Sra, *Modeling data using directional distributions*, Technical Report TR-03-06, Department of Computer Sciences, The University of Texas at Austin. URL `https://www.cs.utexas.edu/users/inderjit/public_papers/tr03-06.pdf`, 2003.

[8] F. Nielsen, V. Garcia, *Statistical exponential families: A digest with flash cards*, arXiv preprint arXiv:0911.4863 (2009).

[9] K. Hornik, B. Grün, *movMF: an R package for fitting mixtures of von Mises-Fisher distributions*, Journal of Statistical Software 58, no. 10 (2014), str. 1-31.

[10] E. L. Lehmann, G. Casella, *Theory of point estimation*, Springer Science & Business Media, 2006.

[11] K. V. Mardia, P. E. Jupp, *Directional statistics*, Vol. 494. John Wiley & Sons, 2009.

[12] C. R. Rao, *Information and the accuracy attainable in the estimation of statistical parameters*, In Breakthroughs in statistics, Springer, New York, NY, 1992., str. 235-247

[13] R. T. Rockafellar, *Convex analysis*, Vol. 36. Princeton university press, 1970.

[14] G. Schou, *Estimation of the concentration parameter in von Mises–Fisher distributions*, Biometrika 65, no. 2 (1978), str. 369-377.

[15] S. Sra, *A short note on parameter approximation for von Mises-Fisher distributions: and a fast implementation of $I_s(x)$* Computational Statistics 27, no. 1 (2012), str. 177-190.

[16] G. N. Watson, *A treatise on the theory of Bessel functions*, Cambridge University Press, 2nd edition, 1995.

[17] S. Zhong, *Efficient online spherical k-means clustering*, In Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005., vol. 5, IEEE, 2005., str. 3180-3185

Sažetak

U ovom rada predstavljena je izgradnja klasifikacijskog algoritma baziranog na idejama informacijske geometrije. Stoga najprije uvodimo neke osnovne koncepte iz područja informacijske geometrije, pri čemo posebni naglasak stavljamo na eksponencijalne familije koje su u tom kontekstu najzanimljivija klasa vjerojatnosnih distribucija. Zatim se bavimo von Mises-Fisherovom distribucijom i njezinim svojstvima, budući da je upravo ona najčešće korištena distribucija za modeliranje usmjerenih podataka. Nakon toga, rezultate poopćujemo na konveksnu kombinaciju von Mises-Fisherovih distribucija te proučavamo procjenu njezinih parametara pomoću algoritma maksimizacija očekivanja. Naposljetku, primijenjujemo prethodno uvedene rezultate kako bismo pokazali kako u praksi funkcionira klasifikacija veće količine tekstualnih podataka.

Summary

In this work we present the development of the classification algorithm that arises from ideas gathered in information geometry. Therefore, we first introduce some basic concepts from information geometry, with emphasis on exponential families, which are in this context the most interesting class of probability distributions. Next, we address von Mises-Fisher distribution and its properties, as it's the most frequently used distribution for modelling directional data. After that, we generalised the results for the vMF mixture model and study estimation of its parameters using the expectation-maximization algorithm. Finally, we apply previously introduced methods to show how the classification of a large sample of text data performs in practice.

Životopis

Rođena sam 06.06.1993. u Zagrebu, gdje završavam osnovnu školu i potom VII. gimnaziju. Preddiplomski studij matematike završila sam 2017. godine na Odjelu za matematiku Sveučilišta u Rijeci. Iste godine upisujem diplomski studij Matematičke statistike na Prirodoslovno-matematičkom fakultetu u Zagrebu, koji završavam ovim radom. Od 2020. godine zaposlena sam kao podatkovni znanstvenik.