

Teorijska pitanja

Strojno učenje 1, UNIZG FER, ak. god. 2023./2024.

Jan Šnajder, v1.1

2 Osnovni koncepti

1. (T) Pogreška modela definirana je kao očekivanje funkcije gubitka na primjerima iz $\mathcal{X} \times \mathcal{Y}$. Međutim, u praksi tu pogrešku aproksimiramo empirijskom pogreškom, koju računamo kao srednju vrijednost funkcije gubitka na skupu označenih primjera $\mathcal{D} \subseteq \mathcal{X} \times \mathcal{Y}$. **Zašto pogrešku modela aproksimiramo empirijskom pogreškom i na kojoj se prepostavci temelji ta aproksimacija?**

- [A] Različitih primjera iz $\mathcal{X} \times \mathcal{Y}$ potencijalno ima beskonačno mnogo, pa pogrešku računamo na uzorku \mathcal{D} za koji prepostavljamo da je reprezentativan
- [B] Ne možemo izračunati očekivanje gubitka jer nam nije poznata distribucija primjera iz $\mathcal{X} \times \mathcal{Y}$, no prepostavljamo da je \mathcal{D} reprezentativan uzorak iz te distribucije
- [C] Očekivanje gubitka ne možemo izračunati jer primjera iz $\mathcal{X} \times \mathcal{Y}$ ima potencijalno beskonačno, stoga pogrešku računamo na temelju skupa \mathcal{D} za koji prepostavljamo da je konačan
- [D] Funkciju gubitka jednostavnije je definirati nego funkciju pogreške, a aproksimacija je točna uz prepostavku i.i.d.

2. (T) Model \mathcal{H} je skup svih parametriziranih funkcija $h(\mathbf{x}; \boldsymbol{\theta})$ indeksiran parametrima $\boldsymbol{\theta}$. To jest:

$$\mathcal{H} = \{h(\mathbf{x}; \boldsymbol{\theta})\}_{\boldsymbol{\theta}}$$

Što to zapravo znači?

- [A] Da model sadrži beskonačno mnogo funkcija h čija konkretna definicija ovisi o vrijednostima parametara $\boldsymbol{\theta}$
- [B] Da različite funkcije h imaju različite parametre $\boldsymbol{\theta}$, i da su sve one sadržane u modelu, to jest za sve njih vrijedi $h \in \mathcal{H}$
- [C] Da za različite parametre $\boldsymbol{\theta}$ dobivamo različite funkcije h , i da su sve one sadržane u modelu, to jest za sve njih vrijedi $h \in \mathcal{H}$
- [D] Da su funkcije h definirane sa slobodnim parametrima $\boldsymbol{\theta}$ i da broj različitih funkcija odgovara broju parametara

3. (T) Modeli strojnog učenja tipično imaju i parametre i hiperparametre. **Koja je razlika između parametara i hiperparametara?**

- [A] Parametre optimira algoritam strojnog učenja, dok optimizacija hiperparametara nije u nadležnosti tog algoritma
- [B] Hiperparametri određuju jačinu regularizacije, a parametri stupanj nelinearnosti hipoteze
- [C] Parametri određuju iznos empirijske pogreške na skupu za učenje, a hiperparametri iznos te pogreške na skupu za provjeru
- [D] Hiperparametri mogu biti diskretni ili kontinuirani, dok su parametri uvijek kontinuirani

4. (T) U strojnom učenju, model je skup funkcija \mathcal{H} indeksiran parametrima θ . **Što to znači?**
- A Svaki θ jednoznačno određuje funkciju koja primjer \mathbf{x} preslikava u oznaku y u ovisnosti o parametrima θ
 - B Svaki skup funkcija \mathcal{H} ima svoj vektor parametara θ i svaki vektor parametara θ određuje skup funkcija \mathcal{H}
 - C Svaki \mathbf{x} određuje parametar θ kojim se oznaka y preslikava u primjer \mathbf{x}
 - D Svaka funkcija koja primjeru \mathbf{x} dodjeljuje oznaku y jednoznačno određuje točku θ u višedimenzijskome prostoru parametra
5. (T) Hipoteza h je funkcija koja primjerima iz \mathcal{X} pridjeljuje oznake iz \mathcal{Y} . Za h kažemo da je definirana “do na parametre θ ”. **Što to znači?**
- A Funkcija h jednoznačno određuje parametre θ iz skupa svih mogućih parametara, koji nazivamo prostor parametara
 - B Svaka vrijednost parametara θ daje jednu konkretnu funkciju h koja se razlikuje od svih drugih funkcija u modelu \mathcal{H}
 - C Funkcija h definirana je bez parametara, i njih treba odrediti naknadno postupkom odabira modela \mathcal{H}
 - D Različite vrijednosti za θ mogu dati različite funkcije h , a skup svih takvih različitih funkcija definira model \mathcal{H}
6. (T) Svaki algoritam strojnog učenja ima neku induktivnu pristranost. Bez induktivne pristranosti, učenje na temelju podataka ne bi imalo smisla, odnosno algoritam bez induktivne pristranosti ne bi mogao ništa naučiti. **Zašto strojno učenje bez induktivne pristranosti nije moguće?**
- A Model bi bio prejednostavan te ne bi postojala hipoteza s empirijskom pogreškom nula
 - B Primjeri ne bi nužno bili linearno odvojivi
 - C Oznaka niti jednog neviđenog primjera ne bi bila jednoznačno određena
 - D Prostor parametara bio bi neograničen, tj. postojalo bi beskonačno mnogo vektora parametara
7. (T) Modeli strojnog učenja općenito su različite složenosti. S porastom složenosti modela raste vjerojatnost da model bude prenaučen. Ta vjerojatnost raste s količinom šuma u podacima. **Zašto šum u podacima za učenje može dovesti do prenaučenosti klasifikacijskog modela?**
- A Zbog šuma granica između klase izgleda nelinearnijom nego što ona to zapravo jest, pa primjeri blizu granice znatno više doprinose pogrešci učenja nego primjeri koji su udaljeni od granice
 - B Efekt šuma je slučajan, pa će hipoteza koja se previše prilagodi šumu na skupu za učenje očekivano imati veliku pogrešku na ispitnom skupu gdje je šum drugaćiji ili ga nema
 - C Povećanjem količine šuma granica između klasa postaje sve nelinearnija, pa raste i složenost modela te dobivena hipoteza očekivano neće odgovarati granici između klasa na ispitnom skupu
 - D Zbog šuma su oznake nekih primjera u skupu za učenje pogrešne, pa sve hipoteze iz modela imaju na tom skupu pogrešku koja je veća od nula, a još veća na ispitnom skupu
8. (T) Svaki model strojnog učenja ima neku induktivnu pristranost. **Što je induktivna pristranost?**
- A Kriterij koji, na temelju modela, jednoznačno određuje hipotezu sa minimalnom empirijskom pogreškom
 - B Svaka pretpostavka koja jednoznačno određuje model na temelju hipoteze i skupa za učenje
 - C Odstupanje procjene parametra na temelju podataka u odnosu na pravu vrijednost parametra u populaciji
 - D Minimalan skup pretpostavki koje, uz skup za učenje, jednoznačno određuju klasifikaciju svakog primjera

9. (T) Model \mathcal{H} je skup hipoteza $h(\cdot; \boldsymbol{\theta})$ koje su indeksirane vektorom parametara $\boldsymbol{\theta}$. Neka $\boldsymbol{\theta} \in \mathbb{R}^{n+1}$, gdje je n dimenzija ulaznog prostora. **Može li skup \mathcal{H} biti beskonačan?**

- A Da, primjerice ako $\mathcal{X} = \mathbb{R}^n$ i $h(\mathbf{x}; \boldsymbol{\theta}) = \boldsymbol{\theta}^T \mathbf{x}$
- B Ne, jer je skup primjera \mathcal{D} uvijek konačan, neovisno o dimenzionalnosti ulaznog prostora n
- C Da, primjerice ako je $\mathcal{X} = \{0, 1\}^n$ i $h(\mathbf{x}; \boldsymbol{\theta}) = \mathbf{1}\{\boldsymbol{\theta}^T \mathbf{x} \geq 0\}$
- D Ne, jer za beskonačan skup \mathcal{H} optimizacijski problem $\operatorname{argmax}_{h \in \mathcal{H}} E(h|\mathcal{D})$ nije definiran

3 Linearna regresija

1. (T) Svaki algoritam strojnog učenja ima neku induktivnu pristranost. Induktivna pristranost sastoji se od pristranosti jezika i pristranosti preferencije. **Kako glasi induktivna pristranost preferencije (neregulariziranog) modela linearne regresije?**

- A Hipoteza h je linearna kombinacija težina \mathbf{w} i značajki \mathbf{x}
- B Težine \mathbf{w} maksimiziraju iznos $\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$
- C Težine \mathbf{w} minimiziraju iznos $\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$
- D Hipoteza h je funkcija iz \mathbb{R}^n u \mathbb{R}

2. (T) Rješenje najmanjih kvadrata za vektor težina \mathbf{w} jest:

$$\mathbf{w} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

Pod kojim uvjetima ćemo težine moći izračunati na ovaj način, i o čemu dominantno ovisi složenost tog postupka?

- A Ako je rang matrice \mathbf{X} jednak $N + 1$, a složenost izračuna dominantno ovisi o N
- B Ako je rang matrice $\mathbf{X}^T \mathbf{X}$ jednak N , a složenost izračuna dominantno ovisi o n
- C Ako je rang matrice \mathbf{X} jednak $n + 1$, a složenost izračuna dominantno ovisi o n
- D Ako je matrica $\mathbf{X}^T \mathbf{X}$ kvadratna i punog ranga, a složenost izračuna dominantno ovisi o N

3. (T) Svaki algoritam strojnog učenja ima neku induktivnu pristranost. Induktivna pristranost sastoji se od pristranosti jezika i pristranosti preferencije. **Koja je razlika između induktivnih pristranosti regularizirane i neregularizirane linearne regresije?**

- A Algoritmi imaju različite pristranosti, i to različitu pristranost preferencije jer regularizirana regresija preferira jednostavnije hipoteze, a onda i različitu pristranost jezika jer je model neregularizirane regresije nadskup modela regularizirane regresije
- B Oba algoritma imaju isti model, definiran kao linearnu kombinaciju značajki i težina, pa dakle imaju istu pristranost jezika, ali se razlikuju u pristranosti preferencije jer imaju različito definiranu empirijsku pogrešku (osim ako je regularizacijski faktor jednak nuli)
- C Algoritmi se ne razlikuju po pristranosti preferencijom budući da koriste istu funkciju gubitka (kvadratni gubitak), međutim regularizirana regresija ima jaču induktivnu pristranost jezika od regularizirane regresije budući da prvi model uključuje drugi model
- D Za razliku od neregularizirane regresije, regularizirana regresija preferira jednostavnije hipoteze, međutim pristranosti su im identične jer su oba algoritma definirana kao linearna kombinacija značajki i težina te oba koriste identičan optimizacijski postupak (pseudoinverz matrice dizajna)

4. (T) Model linearne regresije je poopćeni linearni model i ima probabilističku interpretaciju. Prijetite se, tu smo interpretaciju upotrijebili smo kako bismo opravdali empirijsku funkciju pogreške

definiranu na temelju kvadratnog gubitka. **Kako formalno glasi probabilistička pretpostavka modela linearne regresije?**

- A $p(\mathbf{x}|y) = \mathcal{N}(f(\mathbf{x}), \sigma^2)$
- B $p(y|\mathbf{x}) = \mathcal{N}(0, \sigma^2)$
- C $p(y|\mathbf{x}) = \mathcal{N}(h(\mathbf{x}), \sigma^2)$
- D $p(y) = \mathcal{N}(0, \sigma^2)$

5. (T) Svaki algoritam strojnog učenja ima neku induktivnu pristranost. Induktivna pristranost sastoji se od pristranosti jezika i pristranosti preferencije. **Što je induktivna pristranost preferencije linearног modela regresije?**

- A Pretpostavka $P(\mathbf{y}|\mathbf{w}) = \prod_{i=1}^N P(y^{(i)}|\mathbf{w})$
- B Minimizacija iznosa $\|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2$
- C Odabir linearног modela $h(\mathbf{w}; \mathbf{y}) = \mathbf{w}^T \mathbf{x}$
- D Maksimizacija iznosa $-\ln \mathcal{L}(\mathbf{w}|\mathbf{y})$

6. (T) Optimizacija modela hrbatne regresije (L_2 -regularizirane linearne regresije) ima rješenje u zatvorenoj formi. Neka je λ regularizacijski faktor, n broj značajki u ulaznom prostoru (bez "dummy" jedinice), m broj značajki u prostoru značajki (također bez "dummy" jedinice) te N broj primjera. Glavna komponenta rješenja je izračun inverza matrice izračunate na temelju matrice dizajna Φ . **Koliko redaka odnosno stupaca ima matrica koju invertiramo?**

- A $m + 1$
- B $m + \lambda$
- C $n + \lambda$
- D N

7. (T) Postupak najmanjih kvadrata (OLS) temelji se na izračunu pseudoinverza \mathbf{X}^+ matrice dizajna \mathbf{X} , što je poopćenje običnog inverza \mathbf{X}^{-1} . **U kojoj situaciji je rješenje dobiveno pseudo-inverzom identično rješenju dobivenom običnim inverzom?**

- A Kada je broj primjera veći od broja značajki
- B Kada je broj primjera jednak broju značajki plus jedan i nema multikolinearnosti
- C Kada je broj značajki manji od broja primjera i nema multikolinearnosti
- D Kada nema multikolinearnosti i matrica dizajna je dobro kondicionirana

8. (T) Minimizacija funkcije kvadratne pogreške linearne regresije odgovara maksimizaciji log-izglednosti oznaka pod modelom. **Pod kojim uvjetom vrijedi ova korespondencija?**

- A Primjeri (\mathbf{x}, y) u skupu \mathcal{D} nezavisno su uzorkovani iz zajedničke distribucije $P(\mathbf{x}, y)$
- B Funkcija pogreške $E(\mathbf{w}|\mathcal{D})$ je neprekidna i unimodalna
- C Matrica dizajna \mathbf{X} nije singularna ili je regularizacijski faktor λ veći od 0
- D Oznaka y primjera (\mathbf{x}, y) je normalna varijabla sa srednjom vrijednošću $\mathbf{w}^T \mathbf{x}$

4 Linearna regresija II

1. (T) Rješenje najmanjih kvadrata s L2-regularizacijom (hrbatna regresija) je:

$$\mathbf{w} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}$$

gdje $\lambda \mathbf{I} = \text{diag}(0, \lambda, \dots, \lambda)$. **Koji je efekt regularizacije na Gramovu matricu?**

- A Dodavanje vrijednosti λ na dijagonale Gramove matrice povećava njezin rang
- B Dodavanje vrijednosti λ na dijagonale Gramove matrice povećava normu težina $\|\mathbf{w}\|$
- C Minimizacija norme težina $\|\mathbf{w}\|$ čini Gramovu matricu kvadratnom i singularnom
- D Minimizacija norme težina $\|\mathbf{w}\|$ povećava multikolinearnost Gramove matrice i smanjuje složenost modela

2. (T) Kao regularizacijski faktor kod modela linearne regresije tipično se koristi neka p-norma vektora težina, $\|\mathbf{w}\|_p$. **Na kojoj se činjenici temelji korištenje norme kao regularizacijskog izraza?**
- A Ako su težine hipoteze velike magnitude, model je prenaučen
 - B Ako je model prenaučen, hipoteza će imati velike magnitude težina
 - C Ako je model optimalne složenosti, hipoteza će imati male magnitude težina
 - D Ako su težine hipoteze male magnitude, model je podnaučen
3. (T) L_1 -regularizacija ili LASSO kao regularizacijski izraz koristi prvu normu vektora težina, $\|\mathbf{w}\|_1$. **Što je prednost a što nedostatak L_1 -regularizacije?**
- A Prednost je da L_1 -regulariziranu pogrešku možemo minimizirati gradijentnim spustom, a nedostatak je da rezultira rijetkim modelima
 - B Prednost je da izbacuje značajke iz modela, a nedostatak je da L_1 -regularizirana pogreška nema minimizator u zatvorenoj formi
 - C Prednost je da zadržava sve značajke u modelu, a nedostatak je da Gramova matrica može biti blizu singularne ako u podatcima postoji multikolinearnost
 - D Prednost je da postoji rješenje u zatvorenoj formi (pseudoinverz), a nedostatak da izračun L_1 -regulariziranog pseudoinverza ovisi o broju značajki ali i o broju primjera
4. (T) Optimizacijom regularizirane funkcije pogreške smanjuje se prenaučenost modela. **Kako je definirana L_2 -regularizirana pogreška kod linearne regresije?**
- A Zbroj očekivanja funkcije gubitka unakrsne entropije i druge norme vektora težina
 - B Zbroj prosjeka kvadratnog gubitka na svim primjerima i kvadrata druge norme vektora težina bez težine w_0
 - C Zbroj neregularizirane pogreške i izraza proporcionalnog s kvadratom norme vektora težina bez težine w_0
 - D Zbroj funkcije gubitka po svim primjerima i neregularizirane pogreške bez težine w_0
5. (T) Optimizacijom regularizirane funkcije pogreške smanjuje se prenaučenost modela. **Kakve parametre modela nalazi optimizacija L_2 -regularizirane pogreške?**
- A Parametre koji uz što veću magnitudu vrijednosti daju što manje očekivanje gubitka na skupu za ispitivanje
 - B Parametre koji uz što manju magnitudu vrijednosti daju što manje očekivanje gubitka na skupu za učenje
 - C Parametre koji uz što veću magnitudu vrijednosti daju što veće očekivanje gubitka na skupu za učenje
 - D Parametre koji uz što manju magnitudu vrijednosti daju što veće očekivanje gubitka na skupu za ispitivanje
6. (T) Multikolinearnost značajki jedan je od problema koji može nastupiti kod primjene modela regresije na stvarnim podatcima. Efekt multikolinearnosti i savršene multikolinearnosti dobro je uočljiv kod optimizacijskoga postupka običnih najmanjih kvadrata (OLS) kada se on provodi izračunom pseudoinverza matrice dizajna. Neka je m broj značajki, Φ je matrica dizajna i $\mathbf{G} = \Phi^T\Phi$ je Gramova matrica. **Koji je efekt savršene multikolinearnosti kod postupka OLS?**
- A $\text{rang}(\Phi) < m + 1$, \mathbf{G} nema puni rang i nema inverz, no ima pseudoinverz koji nije numerički stabilan
 - B Φ nema puni rang, $\text{rang}(\mathbf{G}) < m + 1$ i \mathbf{G} nema pseudoinverz
 - C Φ ima puni rang, $\text{rang}(\mathbf{G}) > m$ i \mathbf{G} ima inverz, ali s visokim kondicijskim brojem
 - D $\text{rang}(\Phi) = N$, no $\text{rang}(\mathbf{G}) < N$, pa \mathbf{G} ima pseudoinverz, ali nema numerički stabilan inverz

5 Linearni diskriminativni modeli

1. (T) Na predavanjima smo za klasifikaciju pokušali upotrijebiti algoritam regresije. Zaključili smo da to ne funkciona, tj. da algoritam linearne regresije jednostavno nije klasifikacijski algoritam. **Koje bismo minimalne preinake trebale učiniti u algoritmu linearne regresije, a da dobijemo algoritam koji dobro funkciona kao klasifikacijski algoritam?**
- A Promijeniti funkciju gubitka
 B Promijeniti model i funkciju gubitka
 C Promijeniti funkciju gubitka i optimizacijski postupak
 D Promijeniti model, funkciju gubitka i optimizacijski postupak
2. (T) Algoritam strojnog učenja idealno bi minimizirao gubitak 0-1. Međutim, funkciju gubitka 0-1 u praksi ne možemo koristiti za optimizaciju parametra modela. **Zašto gubitak 0-1 ne možemo koristiti za optimizaciju?**
- A Gradijent gubitka 0-1 svugdje je nula osim za $h(\mathbf{x}) = 0$, pa funkcija pogreške ima zaravni po kojima se gradijentni spust ne može spuštati
 B Gubitak 0-1 pored neispravno klasificiranih primjera kažnjava i ispravno klasificirane primjere, i to tim više što su oni dalje od granice između klasa
 C Funkcija gubitka 0-1 nije diferencijabilna, pa ne postoji rješenje u zatvorenoj formi i ne postoji gradijent
 D Funkcija gubitka 0-1 nije konveksna, pa ni funkcija pogreške nije konveksna već ima lokalne minimume te ne postoji minimizator u zatvorenoj formi
3. (T) Algoritam linearne regresije može se pokušati primijeniti na klasifikacijski problem, kao što smo pokušali na predavanjima, međutim to nije dobro funkcioniralo. Razmotrite tri komponente algoritma linearne regresije: model (M), funkciju gubitka (G) i optimizacijski postupak (O). Također, prisjetite se algoritma logističke regresije, koji je dobar klasifikacijski algoritam. Želimo preinaciti komponente algoritma linearne regresije tako da iz njega dobijemo nov algoritam koji klasifikaciju radi bolje od linearog modela regresije, ali koji je drugačiji od logističke regresije. **Uz koje tri komponente bismo dobili takav algoritam?**
- A M: $h(\mathbf{x}) = \sigma(\mathbf{w}^T \mathbf{x})$; G: $L(y, h(\mathbf{x})) = -y \ln h(\mathbf{x}) - (1-y) \ln(1-h(\mathbf{x}))$, O: $\mathbf{w}^* = \mathbf{X}^+ \mathbf{y}$
 B M: $h(\mathbf{x}) = \mathbf{w}^T \mathbf{x}$; G: $L(y, h(\mathbf{x})) = -y \ln h(\mathbf{x}) - (1-y) \ln(1-h(\mathbf{x}))$ O: \mathbf{w}^* izračunat gradijentnim spustom
 C M: $h(\mathbf{x}) = \mathbf{w}^T \mathbf{x}$; G: $L(y, h(\mathbf{x})) = (y - h(\mathbf{x}))^2$, O: $\mathbf{w}^* = \mathbf{X}^+ \mathbf{y}$
 D M: $h(\mathbf{x}) = \sigma(\mathbf{w}^T \mathbf{x})$; G: $L(y, h(\mathbf{x})) = (y - h(\mathbf{x}))^2$, O: \mathbf{w}^* izračunat gradijentnim spustom
4. (T) Višeklasni problem može se riješiti binarnim klasifikatorom uz primjenu sheme OVO ili sheme OVR. Obje sheme imaju svoje prednosti i nedostatke. Prepostavite da raspolaćemo sa K klase i da svaka klasa ima N/K primjera, gdje je N ukupan broj primjera u skupu za učenje. **Što su prednosti odnosno nedostatci OVO i OVR sheme u takvom slučaju?**
- A OVR treba K puta više klasifikatora nego OVO, ali su kod OVO pozitivne klase $K/2$ puta manje zastupljene nego OVR
 B OVO iziskuje $(K-1)/2$ puta više parametara nego OVR, ali svaki OVR klasifikator ima $K-1$ puta manje pozitivnih primjera nego negativnih
 C OVR iziskuje $K-1$ puta više klasifikatora od sheme OVO, ali kod OVO pozitivne klase imaju $K-1$ puta manje primjera nego kod OVR
 D OVO svaki klasifikator trenira s $K/2$ puta manje primjera nego OVR, ali pozitivne klase kod OVR imaju K puta manje primjera nego kod OVO

5. (T) Jedna od triju komponenta svakog algoritma strojnog učenja je funkcija gubitka. Razmotrite funkcije gubitka perceptronu, logističke regresije (LR) i SVM-a. **Što je specifično funkciji gubitka perceptronu u odnosu na funkcije gubitka LR-a i SVM-a?**

- A Svaki primjer nanosi gubitak, ali manji za točno klasificirane primjere nego za netočno klasificirane primjere
- B Gubitak za sve točno klasificirane primjere je nula, a za netočno klasificirane može biti manji od 1
- C Točno klasificirani primjer nanosi gubitak manji od 1 te gubitak opada što je primjer bliže granici
- D Gubitak netočno klasificiranih primjera raste linearno s udaljenošću od granice

6. (T) Funkcija gubitka perceptronu nalikuje funkciji gubitka SVM-a (funkciji zglobnice). Međutim, postoji ključna razlika između tih dviju funkcija gubitka, koje vode do različitog ponašanja algoritma perceptronu i algoritma SVM-a. **Po čemu se gubitak zglobnice razlikuje od gubitka perceptronu?**

- A Za ispravno klasificirane primjere gubitak zglobnice je manji od gubitka za neispravno klasificirane primjere
- B Gubitak zglobnice je nula za primjere koji su ispravno klasificirani i daleko od granice
- C Za neispravno klasificirane primjere gubitak zglobnice raste linearno s udaljenošću od hiper-ravnine
- D Gubitak zglobnice kažnjava sve primjere koji se nalaze unutar marge, čak i one koji su ispravno klasificirani

6 Logistička regresija

1. (T) Poopćeni linearni model definirali smo kao $h(\mathbf{x}) = f(\mathbf{w}^T \phi(\mathbf{x}))$, gdje je f neka (moguće nelinearna) aktivacijska funkcija, a ϕ je (moguće nelinearna) funkcija preslikavanja u prostor značajki. **Koji od navedenih uvjeta je dovoljan uvjet da granica između klasa u ulaznom prostoru bude linearna?**

- A f je afina funkcija
- B $\phi(\mathbf{x}) = (1, \mathbf{x})$
- C f je afina funkcija i $\phi(\mathbf{x}) = (1, \mathbf{x})$
- D $f(\mathbf{x}) = \mathbf{x}$

2. (T) Kod logističke regresije, pogrešku unakrsne entropije izveli smo modelirajući distribuciju vjerojatnosti oznaka y u skupu označenih primjera. **Na koji smo način modelirali distribuciju vjerojatnosti pojedinačnog primjera y ?**

- A $P(y|\mathbf{x}) = (y - h(\mathbf{x}))\mathbf{x}$
- B $P(y|\mathbf{x}) = h(\mathbf{x})^y(1 - h(\mathbf{x}))^{1-y}$
- C $P(y|\mathbf{x}) = y^{h(\mathbf{x})}(1 - y)^{h(\mathbf{x})}$
- D $P(y|\mathbf{x}) = h(\mathbf{x})(1 - h(\mathbf{x}))$

3. (T) Optimizacija parametara logističke regresije algoritmom grupnog gradijentnog spusta tipično u svakoj iteraciji uključuje i linijsko pretraživanje. **Što nam osigurava uporaba linijskog pretraživanja kod optimizacije logističke regresije?**

- A Da postupak uvijek konvergira, neovisno o odabranoj stopi učenja η i početnim težinama \mathbf{w}
- B Da postupak uvijek konvergira, pod uvjetom da su primjeri linearno neodvojivi ili da regulariziramo sa $\lambda > 0$
- C Da postupak ne može zaglaviti u lokalnome minimumu, pod uvjetom da u skupu \mathcal{D} nema multikolinearnosti
- D Da postupak konvergira brže, pod uvjetom da su primjeri linearno odvojivi i da stopa učenja nije η prevelika

4. (T) Neregularizirani model logističke regresije sklon je prenaučenosti. To je pogotovo slučaj ako se model trenira na linearno odvojivim podatcima, čak i onda kada u podatcima nema nikakvoga šuma i kada se ne koristi nikakvo preslikavanje u prostor značajki. **Zbog čega dolazi do prenaučenosti modela neregularizirane logističke regresije na linearno odvojivim skupovima podataka?**

- A Empirijska pogreška logističke regresije smanjuje se s porastom broja primjera
- B Ispravno klasificirani primjeri koji su vrlo udaljeni od granice nanose malen gubitak
- C S porastom norme vektora težina gubitak na točnim primjerima teži prema nuli
- D Netočno klasificirani primjeri nanose gubitak koji je proporcionalan normi vektora težina

5. (T) Algoritam logističke regresije za optimizaciju može koristiti grupni gradijentni spust s linjskim pretraživanjem. Takav optimizacijski algoritam ima svojstvo globalne konvergencije. Razmotrite neregulariziranu logističku regresiju na linearno neodvojivom problemu. **Što globalna konvergencija konkretno znači u tom slučaju?**

- A Optimizacijski algoritam će konvergirati prema parametrima koji minimiziraju pogrešku na skupu za učenje, ali neće doseći minimum
- B Gradijentni spust neće krivudati u prostoru parametara jer optimizacijski algoritam koristi informaciju o zakriviljenosti površine pogreške
- C Optimizacijski algoritam će konvergirati do minimuma, ali nema garancije da će to biti globalni minimum funkcije pogreške
- D Neovisno o inicijalizaciji, optimizacijski će algoritam pronaći parametre koji minimiziraju pogrešku na skupu za učenje

7 Logistička regresija II

1. (T) Kod logističke regresije za optimizaciju tipično koristimo gradijentni spust ili Newtonov optimizacijski postupak. **Što su prednosti, a što nedostatci gradijentnog spusta u odnosu na Newtonov postupak, i to konkretno kod logističke regresije?**

- A Za razliku od Newtonovog postupka, gradijentni spust može se koristiti za L2-regulariziranu logističku regresiju, no ako je stopa učenja prevelika, postupak može divergirati, dok Newtonov postupak nema taj problem
- B Za razliku od Newtonovog postupka, gradijentni spust može se koristiti za "online" (pojedinačno) učenje, no može krivudati i zato sporije konvergirati od Newtonovog postupka
- C Newtonov postupak brže konvergira, ali se može koristiti samo za konveksnu funkciju pogreške, dok gradijentni spust nema tog ograničenja, ali može zaglaviti u lokalnom optimumu
- D Gradijentni spust znatno je računalno jednostavniji od Newtonovog postupka, no za razliku od Newtonovog postupka kod L2-regularizirane regresije ne konvergira ako primjeri nisu linearno odvojivi

2. (T) Kod Newtonovog postupka optimizacije za logističku regresiju ažuriranje težina provodi se prema sljedećem pravilu:

$$\mathbf{w} \leftarrow \mathbf{w} - \mathbf{H}^{-1} \nabla E(\mathbf{w} | \mathcal{D})$$

Očito, za provođenje ovog postupka potrebno je računati inverz Hesseove matrice, tj. matrice parcijalnih derivacija \mathbf{H} . Općenito, operacija invertiranja matrice nije uvijek izvediva, a čak i kada jest izvediva, rješenje nije uvijek numerički stabilno. **Kod logističke regresije, koji je nužan i dovoljan uvjet za izvedivost i numeričku stabilnost Newtonovog optimizacijskog postupka?**

- A Značajke moraju biti linearno zavisne
- B Funkcija pogreške mora biti konveksna
- C U podatcima ne smije biti multikolinearnosti
- D Broj primjera mora biti veći od broja značajki plus jedan

3. (T) Svi poopćeni linearni modeli mogu se trenirati u “online” (pojedinačnom) načinu, primjernom algoritma LMS. To vrijedi i za algoritam linearne regresije, za koji smo prvotno kao minimizaciju kvadrata provodili računajući pseudoinverz matrice dizajna. Jedna od prednosti algoritma LMS u odnosu na izračun pseudoinverza kod linearne regresije je manja računalna složenost LMS-a. Neka E označava broj epoha, N je broj primjera, a m broj značajki u prostoru značajki. **Koja je (vremenska) računalna složenost algoritma LMS, primjenjenog na linearnu regresiju?**

- A $\mathcal{O}(ENm)$
- B $\mathcal{O}(E(N + m))$
- C $\mathcal{O}(EN^2m)$
- D $\mathcal{O}(ENm^2)$

4. (T) Problem višeklasne ($K > 2$) klasifikacije logističkom regresijom možemo riješiti na više načina. Možemo primijeniti (1) multinomijalnu logističku regresiju (MLR), (2) binarnu logističku regresiju sa shemom OVO (BLR-OVO) ili (3) binarnu logističku regresiju sa shemom OVR (BLR-OVR). **Koja je prednost MLR nad BLR-OVO i BLR-OVR?**

- A MLR ima više parametara od BLR-OVR, ali nije osjetljiva na neuravnoteženost broja primjera po klasama
- B MLR i BLR-OVR imaju manje parametara od BLR-OVO, no jedino za MLR vrijedi $\sum_k P(y = k|\mathbf{x}) = 1$
- C Za razliku od BLR-OVR i BLR-OVO, kod MLR ne postoji područja u ulaznom prostoru za koje klasifikacijska odluka nije određena
- D Za razliku od BLR-OVO i BLR-OVR, MLR koristi funkciju softmax, pa minimizacija L1-regularizirane pogreške ima rješenje u zatvorenoj formi

5. (T) Kod logističke regresije optimizaciju tipično provodimo gradijentnim spustom. Međutim, kod linearne regresije optimizaciju smo provodili izračunom pseudoinverza matrice dizajna. **Zašto optimizaciju kod logističke regresije također ne provodimo izračunom pseudoinverza matrice dizajna?**

- A Optimizaciju parametara linearne regresije također možemo napraviti gradijentnim spustom po empirijskoj pogrešci, ali to ne radimo jer izračun pseudoinverza ima manju računalnu složenost
- B Zbog nelinearnosti logističke funkcije, kod logističke regresije izračun pseudoinverza matrice dizajna nije moguće napraviti u zatvorenoj formi
- C Maksimizacija log-izglednosti oznaka logističke regresije kao rješenje za parametre ne daje izraz u zatvorenoj formi koji sadržava pseudoinverz matrice dizajna
- D Optimizaciju možemo provesti izračunom pseudoinverza matrice dizajna, međutim, za razliku od gradijentnog spusta, taj postupak ne funkcioniра kada je matrica dizajna singularna

6. (T) Poopćeni linearni modeli (linearna regresija, logistička regresija i multinomijalna regresija) probabilistički su algoritmi strojnog učenja. Njihova probabilistička priroda dolazi do izražaja kako kod modela tako i kod optimizacijskog postupka. **Koji je probabilistički princip ugrađen u optimizacijski postupak tih algoritama?**

- A Minimizirati $\sum_{i=1}^N \ln h(\mathbf{x}^{(i)}; \mathbf{w})$, gdje je $h(\mathbf{x}; \mathbf{w}) = f(\mathbf{w}^T \mathbf{x})$
- B Minimizirati $-\sum_{i=1}^N \ln y^{(i)} h(\mathbf{x}^{(i)}; \mathbf{w})$, gdje je $h(\mathbf{x}; \mathbf{w}) = \mathbf{w}^T \mathbf{x}$
- C Maksimizirati $\sum_{i=1}^N \ln p(y^{(i)} | \mathbf{x}^{(i)})$, gdje je $\mathbb{E}[p(y^{(i)} | \mathbf{x}^{(i)})] = f(\mathbf{w}^T \mathbf{x})$
- D Maksimizirati $\prod_{i=1}^N \ln p(y^{(i)} | \mathbf{x}^{(i)})$, gdje je $\mathbb{E}[p(y^{(i)} | \mathbf{x}^{(i)})] = \mathbf{w}^T \mathbf{x}$

7. (T) Postoji poveznica između algoritma logističke regresije (LR) i algoritma neuronske mreže sa sigmoidnim prijenosnim funkcijama (NN). **Koja je točno poveznica između ova dva algoritma?**

- A NN i LR imaju istu funkciju pogreške, ali se samo LR može optimirati Newtonovim postupkom jer funkcija gubitka NN nije konveksna
- B Jezgreni stroj s Gaussovom jezgrenom funkcijom istovjetan je NN sa L_2 -regulariziranom funkcijom pogreške
- C Model LR istovjetan je modelu dvoslojne NN sa sigmoidnim prijenosnim funkcijama i pogreškom unakrsne entropije
- D Model dvoslojne NN istovjetan je modelu LR s poopćenim linearnim modelima sa sigmoidnim funkcijama kao baznim funkcijama

8. (T) Za optimizaciju parametara poopćenih linearnih modela može se koristiti stohastički gradijentni spust, odnosno pravilo LMS. Neka je (\mathbf{x}, y) označeni primjer za koji radimo ažuriranje težina pomoću pravila LMS. **Što možemo reći o razlici između novih (ažuriranih) i starih težina (težina prije ažuriranja)?**

- A Razlika je to veća što je stopa učenja η bliža jedinici
- B Razlika je to manja što je vektor $\phi(\mathbf{x})$ bliži ishodištu
- C Razlika je to veća što je izlaz modela $h(\mathbf{x})$ bliži nuli
- D Razlika je to manja što je oznaka y bliže jedinici

9. (T) Postoji veza između logističke regresije i modela neuronske mreže. **Koja je veza između tva dva modela?**

- A Multinomijalna logistička regresija s aktivacijskom funkcijom softmax istovjetna je dvoslojnoj neuronskoj mreži sa više neurona u izlaznom sloju
- B Logistička regresija koja kao adaptivne bazne funkcije koristi logističku regresiju istovjetna je neuronskoj mreži sa sigmoidnom aktivacijskom funkcijom
- C Neuronska mreža optimirana algoritmom propagacije pogreške unazad istovjetna je logističkoj regresiji optimiranoj stohastičkim gradijentnim spustom
- D Logistička regresija s linearnim jezgrenim funkcijama istovjetna je neuronskoj mreži sa linearnom aktivacijskom funkcijom i kvadratnom funkcijom pogreške

8 Stroj potpornih vektora

1. (T) Kod SVM-a, problem maksimalne margine sveo se na problem minimizacije izraza $\frac{1}{2}\|\mathbf{w}\|^2$ uz određena ograničenja. **Zašto minimizacija ovog izraza daje maksimalnu marginu?**

- A Što je vektor \mathbf{w} kraći, to je manja vrijednost $h(\mathbf{x})$, pa primjeri moraju biti što dalje da bi vrijedilo $h(\mathbf{x}) = \pm 1$, a to znači da je margina to šira
- B Što je vektor \mathbf{w} kraći, to je manja udaljenost d primjera od hiperravnine, a to efektivno znači da je margina to šira jer je margina fiksna a udaljenosti d se smanjuju
- C Što je vektor \mathbf{w} kraći, to je manja vrijednost $h(\mathbf{x})$, ali je težina w_0 konstantna, pa se udaljenosti izmeđi hiperravnine i primjera povećavaju, što znači da se margina širi
- D Što je vektor \mathbf{w} kraći, to je veća udaljenost d primjera od hiperravnine, što znači da se potporni vektori udaljavaju od hiperravnine, a to znači da margina postaje šira

2. (T) Kod optimizacije SVM-a iskoristili smo Lagrangeovu dualnost kako bismo se iz primarnog optimizacijskog problema prebacili u dualni optimizacijski problem. To smo učinili tako da smo na temelju Lagrangeove funkcije L definirali dualnu Lagrangeovu funkciju \tilde{L} i uveli nova ograničenja, što nam je opet dalo kvadratni program. **Kako onda u konačnici glasi optimizacijski problem tvrde margine u dualnoj formulaciji (ako zanemarimo ograničenja)?**

- A $\operatorname{argmin}_{\boldsymbol{\alpha}} \max_{\mathbf{w}, w_0} L(\mathbf{w}, w_0, \boldsymbol{\alpha})$
- B $\operatorname{argmax}_{\mathbf{w}, w_0} \min_{\boldsymbol{\alpha}} L(\mathbf{w}, w_0, \boldsymbol{\alpha})$
- C $\operatorname{argmin}_{\mathbf{w}, w_0} \max_{\boldsymbol{\alpha}} L(\mathbf{w}, w_0, \boldsymbol{\alpha})$
- D $\operatorname{argmax}_{\boldsymbol{\alpha}} \min_{\mathbf{w}, w_0} L(\mathbf{w}, w_0, \boldsymbol{\alpha})$

3. (T) Svaki algoritam strojnog učenja ima neku induktivnu pristranost. Bez induktivne pristranosti nije moguće naučiti model koji bi generalizirao. **Po čemu se induktivna pristranost algoritma**

SVM (tvrdna marga) razlikuje od induktivne pristranosti algoritma perceptron?

- A SVM ima pristranost preferencijom kojom maksimizira marginu, dok perceptron nema induktivnu pristranost preferencijom već samo pristranost jezika
- B Imaju istu pristranost preferencijom, a to je da primjeri moraju biti linearno odvojivi, no SVM ima dodatnu pristranost ograničenjem u vidu optimizacijskih ograničenja
- C Razlikuju se po pristranost preferencijom, jer perceptron ne maksimizira marginu, premda se može dogoditi da pronade rješenje koje maksimizira marginu
- D Imaju istu pristranost jezika, a pristranost preferencijom također će biti ista ako se oba optimiraju gradijentnim spustom s istim početnim težinama i istom stopom učenja

4. (T) Kod izvoda algoritma SVM s tvrdom marginom, pretpostavili smo da za primjere $\mathbf{x} \in \mathbb{R}^n$ vrijedi sljedeći uvjet linearne odvojivosti:

$$\forall (\mathbf{x}^{(i)}, y^{(i)}) \in \mathcal{D}. y^{(i)} h(\mathbf{x}^{(i)}) \geq 0$$

Koliko hipoteza zadovoljava ovaj uvjet, i kako algoritam SVM odabire jednu od njih?

- A Uvjet zadovoljava beskonačno mnogo hipoteza, međutim samo za jednu vrijedi $y h(\mathbf{x}) = 1$ za najbliže primjere, i to je hipoteza koju odabire SVM
- B Uvjet zadovoljava konačan broj hipoteza koje su linearno odvojive, a SVM između njih odabire onu jednu koja minimizira kvadrat vektora težina
- C Uvjet zadovoljava beskonačno mnogo hipoteza, a SVM odabire onu jednu koja minimizira kvadrat vektora težina te koja ispravno klasificira sve primjere, uz uvjet da $h(\mathbf{x})$ nije u intervalu $(-1, +1)$
- D Uvjet zadovoljava konačan broj hipoteza koje su linearno odvojive, no one se razlikuju samo po faktoru koji množi težine (\mathbf{w}, w_0) , pa SVM odabire onu jednu za koju vrijedi $y h(\mathbf{x}) \geq 1$ za sve primjere

5. (T) Model SVM-a može se definirati i optimirati u primarnoj ili dualnoj formulaciji. **Konceptualno, kada će primjer \mathbf{x} u dualnoj formulaciji SVM-a biti klasificiran u pozitivnu klasu?**

- A Ako je linearna kombinacija značajki iz \mathbf{x} s pozitivnim težinama veća ili jednaka linearnej kombinaciji značajki iz \mathbf{x} s negativnim težinama
- B Ako je vektor \mathbf{x} po skalarnom umnošku sličniji potpornim vektorima s pozitivnom oznakom nego potpornim vektorima s negativnom oznakom
- C Ako je skalarni umnožak vektora \mathbf{x} i vektora oznaka \mathbf{y} veća od praga definiranog parametrom w_0
- D Ako većina od ukupno α primjera iz skupa za učenje koji su po euklidskoj udaljenosti najbliži primjeru \mathbf{x} ima pozitivnu oznaku

6. (T) Problem maksimalne marge ima svoju geometrijsku interpretaciju: maksimalna marga odgovara simetrali spojnica konveksnih ljsaka primjera iz dviju klasa. **Što je nužan i dovoljan uvjet da klasifikacijski problem bude rješiv SVM-om s tvrdom marginom?**

- A Primjeri iz obje klase trebaju činiti konveksne skupove u ulaznom prostoru
- B Najbliži primjeri iz suprotnih klasa moraju biti u vrhovima konveksnih ljsaka
- C Barem jedna spojница između primjera jedne klase treba biti unutar konveksne ljske druge klase
- D Konveksne ljske dviju klasa ne smiju se preklapati (trebaju biti disjunktne)

9 Stroj potpornih vektora II

1. (T) Problem meke margine SVM-a formulirali smo kao problema optimizacije uz ograničenja, preciznije kao problem kvadratnog programiranja. Neka je n broj značajki, a N broj primjera. **Koliko primarni optimizacijski problem meke margine ima ukupno ograničenja, a koliko varijabli po kojima optimiramo?**

- A $2N$ ograničenja i $2n$ varijabli
- B N ograničenja i $2N + 1$ varijabli
- C N ograničenja i $n + 1$ varijabli
- D $2N$ ograničenja i $N + n + 1$ varijabli

2. (T) Kod optimizacijskog problema meke margine jedan od uvjeta KKT koji vrijede u točki rješenja je sljedeći uvjet komplementarne labavosti:

$$\alpha_i(y^{(i)}(\mathbf{w}^T \mathbf{x}^{(i)} + w_0) - 1 + \xi_i) = 0$$

Što možemo zaključiti na temelju ovog uvjeta?

- A Da se primjeri koji nisu potporni vektori sigurno nalaze izvan marge
 - B Da se potporni vektori nalaze na margini ili izvan nje, a na pravoj strani granice
 - C Da se primjeri koji nisu potporni vektori nalaze na margini ili unutar marge
 - D Da se potporni vektori ne nalaze izvan marge na pravoj strani granice
3. (T) Problem meke margine SVM-a s u primarnoj se formulaciji svodi na rješavanje sljedećeg optimizacijskog problema:

$$\underset{\mathbf{w}, w_0, \xi}{\operatorname{argmin}} \left\{ \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^N \xi_i \right\}$$

uz određena linearna ograničenja. Ovaj optimizacijski problem odgovara optimizaciji regularizirane pogreške. Kod regularizirane pogreške u opreci su dva cilja: smanjenje vrijednosti funkcije gubitka i smanjenje složenosti modela. **Kako se ta opreka manifestira kod optimizacijskog problema meke margine SVM-a?**

- A Što je veća vrijednost $\|\mathbf{w}\|^2$, to je marga uža i tim manje primjera ulazi u marginu, pa je tim manji zbroj od ξ_i
 - B Što je veća vrijednost $\|\mathbf{w}\|^4$ to je model složeniji, no tim je veća nelinearnost granice i to je veći hiperparametar C
 - C Što je manji zbroj od ξ_i , to više primjera može ući u marginu i tim je veća vrijednost $\|\mathbf{w}\|^2$ te je model manje složenosti
 - D Što je veći zbroj od ξ_i , to više primjera ulazi u marginu i tim je manja vrijednost $\|\mathbf{w}\|^2$ te je model veće složenosti
4. (T) Optimizacijski problem algoritma SVM može se postaviti u formulaciji meke ili tvrde marge te u primarnoj ili dualnoj formulaciji. Ovisno o formulaciji, kvadratni program sadrži različit broj varijabli po kojima optimiramo (optimizacijske varijable). **Ako matrica dizajna ima više redaka nego stupaca, koja formulacija ima najmanje optimizacijskih varijabli?**

- A Primarni problem meke marge
- B Primarni problem tvrde marge
- C Dualni problem meke marge
- D Dualni problem tvrde marge

5. (T) Kod algoritma SVM preporuča se napraviti skaliranje značajki. U protivnom, pri izračunu skalarnog produkta, značajke s većim rasponom (većom skalom) dominirat će nad značajkama s manjim rasponom (manjom skalom). Međutim, skaliranje značajki nije uvijek nužno. **Kada nije potrebno napraviti skaliranje značajki, i zašto?**

- A Kada se koristi RBF jezgra sa Mahalanobisovom udaljenošću, jer ta udaljenost uzima u obzir varijancu značajki
- B Kada se koristi linearna jezgra i značajke su centrirane oko nule, jer se onda ne računa skalarni produkt između značajki
- C Kada se koristi dualna formulacija SVM-a i algoritam SMO, jer se tada implicitno provodi L1-regularizacija
- D Kada se koristi Gaussova jezgrena funkcija, jer ta jezgra implicitno inducira beskonačnodimenzijski prostor značajki

10 Jezgrene metode

1. (T) Treniramo model SVM s nekom jezgrenom funkcijom. Nakon što smo naučili model na skupu primjera, za neki primjer \mathbf{x} želimo izračunati udaljenost tog primjera od hiperravnine u prostoru značajki. **Je li moguće izračunati tu udaljenost?**

- A Ne, jer u dualnoj (neparametarskoj) formulaciji problema maksimalne margine nemamo vektor značajki
- B Ne, jer granica između klase u prostoru značajki općenito ne mora biti hiperravnina, već može biti hiperpovršina
- C Da, ako nismo koristili Gaussovou jezgru ili neku složeniju jezgru koja koristi Gaussovou jezgru kao gradivni blok
- D Da, ako smo koristili linearu jezgru, odnosno ako je ulazni prostor jednak prostoru značajki

2. (T) Neke jezgrene funkcije nazivamo Mercerove jezgre ili pozitivno definitne jezgre. Takve jezgre daju pozitivno definitnu Gramovu matricu. **Zašto je dobro da je jezgrena funkcija Mercerova jezgra?**

- A Zato što takva jezgra odgovara skalarnom produktu u nekom prostoru značajki, a to je nužno da bismo mogli primijeniti jezgreni trik
- B Zato što takva jezgra inducira Hilbertov prostor, tj. prostor beskonačnodimenzijskih značajki, što nam daje potencijalno vrlo složene modele
- C Zato što takva jezgra omogućava da, umjesto da vektoriziramo primjere, klasifikaciju određujemo na temelju sličnosti između primjera i prototipnih primjera
- D Zato što takva jezgra nužno daje nenegativne vrijednosti sličnosti između parova primjera, što je nužno kako gubitak ne bi bio negativan

3. (T) Gaussova jezgrena funkcija $\kappa(\mathbf{x}_1, \mathbf{x}_2)$ nad primjerima \mathbf{x}_1 i \mathbf{x}_2 definirana je s parametrom preciznosti γ , gdje $\gamma = 1/2\sigma^2$. Ovaj parametar ima utjecaj na vrijednost jezgrene funkcije, ali i na složenosnost (nelinearnost) modela jezgrenog stroja s Gaussovom jezgrenom funkcijom. **Kakav je utjecaj parametra γ na vrijednost Gaussove jezgrene funkcije $\kappa(\mathbf{x}_1, \mathbf{x}_2)$, gdje $\mathbf{x}_1 \neq \mathbf{x}_2$, te na nelinearnost modela jezgrenog stroja?**

- A Što je vrijednost γ manja, to je manja vrijednost $\kappa(\mathbf{x}_1, \mathbf{x}_2)$, i to je model manje nelinearan
- B Što je vrijednost γ veća, to je veća vrijednost $\kappa(\mathbf{x}_1, \mathbf{x}_2)$, i to je model manje nelinearan
- C Što je vrijednost γ manja, to je manja vrijednost $\kappa(\mathbf{x}_1, \mathbf{x}_2)$, i to je model više nelinearan
- D Što je vrijednost γ veća, to je manja vrijednost $\kappa(\mathbf{x}_1, \mathbf{x}_2)$, i to je model više nelinearan

4. (T) Može se dokazati da je Gaussova jezgra s hiperparametrom γ Mercerova jezgra. U praktičnom smislu, to znači da Gaussovou jezgru možemo koristiti za jezgreni trik umjesto da eksplizitno koristimo funkciju preslikavanja $\phi : \mathbb{R}^n \rightarrow \mathbb{R}^m$. **Što to znači u matematičkome smislu?**

- A $\forall \mathbf{x}_1 \forall \mathbf{x}_2. \mathbf{x}_1^T \mathbf{x}_2 = \exp(-\gamma \|\phi(\mathbf{x}_1) - \phi(\mathbf{x}_2)\|^2)$
- B $\forall \mathbf{x}_1 \forall \mathbf{x}_2. \phi(\mathbf{x}_1)^T \phi(\mathbf{x}_2) = \exp(-\gamma \Delta^2)$ gdje $\Delta^2 = (\mathbf{x}_1 - \mathbf{x}_2)^T (\mathbf{x}_1 - \mathbf{x}_2)$
- C $\forall \mathbf{x}_1 \forall \mathbf{x}_2. \phi(\mathbf{x}_1^T \mathbf{x}_2) = \exp(-\gamma \Delta^2)$, gdje $\Delta^2 = (\mathbf{x}_1 - \mathbf{x}_2)^T (\mathbf{x}_1 - \mathbf{x}_2)$
- D $\forall \mathbf{x}_1 \forall \mathbf{x}_2. (\mathbf{x}_1 \neq \mathbf{x}_2) \Rightarrow \left(\exp(-\gamma \|\phi(\mathbf{x}_1) - \phi(\mathbf{x}_2)\|^2) = \mathbf{x}_1^T \mathbf{x}_2 \right)$

5. (T) Važna prednost jezgrenih strojeva je mogućnost primjene jezgrenog trika. Ta je prednost pogotovo očita kada prostor ulaznih primjera \mathcal{X} nije euklidski vektorski prostor, odnosno kada primjere nije moguće prikazati kao vektore realnih brojeva. **Koja je prednost primjene jezgrenog trika u takvom slučaju?**

- A Jezgrenim trikom implicitno se ostvaruje nelinearnost prostora značajki, što povećava kapacitet modela i povećava njegovu točnost
- B Umjesto vektorizacije primjera, dovoljno je definirati nelinearnu funkciju preslikavanja iz ulaznog prostora u prostor značajki
- C Jezgrena funkcija mjeri sličnost između primjera, čime se primjeri efektivno preslikavaju u beskonačnodimenzionalni prostor značajki
- D Jezgrena funkcija može biti mjera sličnosti između nevektoriziranih primjera, što implicitno inducira vektorski prostor značajki

6. (T) Stroj potpornih vektora (SVM) jedna je vrsta rijetkoga jezgrenog stroja. Jezgredni stroj za bazne funkcije koriste jezgrene funkcije izračunate u odnosu na odabrane prototipne primjere. **Po čemu je SVM specifičan u odnosu na općeniti algoritam rijetkoga jezgrenog stroja?**

- A Zbog L1-regularizacije, mnoge težine modela bit će pritegnute na nulu
- B Dimenzija prostora značajki ne može biti veća od broja primjera
- C Broj parametara modela jednak je dimenziji prostora značajki
- D Prototipni primjeri odabiru se u okviru optimizacijskog postupka

7. (T) Kažemo da Mercerove jezgrene funkcije implicitno definiraju prostor značajki. **Što to znači?**

- A Klasifikacija primjera definirana je na temelju umnoška jezgredne funkcije za taj primjer i sve druge primjere u skupu za učenje
- B Jezgrena funkcija između primjera u prostoru značajki jednaka je skalarnom produktu tih primjera u ulaznom prostoru
- C Broj dimenzija prostora značajki implicitno ovisi o broju klasa u ulaznom prostoru te može biti beskonačna
- D Vrijednost jezgredne funkcije nad parom vektora jednak je skalarnom produktu tih vektora nakon preslikavanja u prostor značajki

11 Neparametarske metode

1. (T) Algoritam SVM može biti parametarski i neparametarski, ovisno o tome provodimo li optimizaciju u primarnoj ili dualnoj formulaciji. U oba slučaja preferiramo da je model rijedak, tj. da

je nakon treniranja što više parametara postavljeno na nulu. **Kako rijetkost modela ovisi o hiperparametru C ?**

- A Što je C manji, to je neparametarski model rjeđi, ali to nema utjecaja na rijetkost parametarskog modela jer on nema potporne vektore
 - B Što je C veći, to je neparametarski model manje rijedak, dok je parametarski to rjeđi jer λ pada
 - C Što je C manji, to je neparametarski model rjeđi, a također je to rjeđi i parametarski model jer λ raste
 - D Što je C veći, to je neparametarski model manje rijedak, dok parametarski model nije rijedak jer ima L_2 -regularizaciju a ne L_1 -regularizaciju
2. (T) Problem prokletstva dimenzionalnosti (engl. *course of dimensionality*) pojavljuje se kod algoritama koji rade u visokodimenzijskome vektorskem prostoru i manifestira se na različite načine. **Kako se problem prokletstva dimenzionalnosti u visokodimenzijskim prostorima manifestira kod algoritma k**
- A Udaljenosti između primjera se smanjuju i model $k postaje sve složeniji$
 - B Povećava se broj susjeda u okolini svakog primjera i model $k postaje sve jednostavniji$
 - C Svi primjeri su međusobno vrlo udaljeni i gube se razlike u udaljenosti
 - D Broj susjeda k nekog primjera se smanjuje i gube se granice između klase
3. (T) Algoritmi strojnog učenja mogu biti parametarski ili neparametarski. **Što je karakteristika neparametarskih algoritama strojnog učenja?**
- A Prepostavljuju vjerojatnosnu distribuciju podataka
 - B Broj parametara ovisi o broju primjera
 - C Eksplicitno modeliraju granicu između primjera
 - D Svaki primjer ima globalan utjecaj na izgled hipoteze
4. (T) Nalaženje najbližih susjeda kod algoritma $k predstavlja izazov zbog računalne složenosti. Algoritam stabla lopti (engl. *ball tree*) jedan je od pristupa za smanjenje računalne složenosti dohvaćanja susjeda u visokodimenzijskom vektorskem prostoru. **Na koji način funkcioniра algoritam stabla lopti?**$
- A Koristi preslikavanje osjetljivo na lokalne promjene u vektoru kojim se bliske točke pohranjuju u iste pretince
 - B Koristi pretraživanje duž pravca u vektorskome prostoru u smjeru suprotnome od gradijenta funkcije pogreške
 - C Koristi brzo pretraživu binarnu indeksnu strukturu za particioniranje prostora primjera u preklapajuće regije
 - D Koristi jezgreni trik za izračun euklidske udaljenosti između točke upita i svih drugih točaka u skupu primjera

14 Procjena parametara II

1. (T) Funkcija izglednosti $\mathcal{L}(\boldsymbol{\theta}|\mathcal{D})$ nije isto što i vjerojatnost. **Po čemu se izglednost razlikuje od vjerojatnosti?**
- A Funkcija izglednosti $\mathcal{L}(\boldsymbol{\theta}|\mathcal{D})$ jednaka je gustoći vjerojatnosti $p(\boldsymbol{\theta}|\mathcal{D})$, ali, za razliku od gustoće vjerojatnosti, nije odozgo ograničena sa 1
 - B Ako su podaci diskretni (kategoričke značajke), onda je funkcija izglednosti parametara $\boldsymbol{\theta}$ isto što i zajednička vjerojatnost uzorka \mathcal{D} i parametara $\boldsymbol{\theta}$
 - C Funkcija izglednosti $\mathcal{L}(\boldsymbol{\theta}|\mathcal{D})$ jednaka je gustoći vjerojatnosti $p(\mathcal{D}|\boldsymbol{\theta})$, samo što je izglednost funkcija parametara $\boldsymbol{\theta}$, dok je $p(\mathcal{D}|\boldsymbol{\theta})$ funkcija uzorka \mathcal{D}
 - D Za razliku od vjerojatnosti, funkcija izglednosti $\mathcal{L}(\boldsymbol{\theta}|\mathcal{D})$ je simetrična, u smislu da vrijedi $\mathcal{L}(\boldsymbol{\theta}|\mathcal{D}) = p(\mathcal{D}|\boldsymbol{\theta})$

2. (T) Treniranje probabilističkih modela svodi se na procjenu njihovih parametara $\boldsymbol{\theta}$ na temelju funkcije log-izglednosti, $\mathcal{L}(\boldsymbol{\theta}|\mathcal{D})$. **Što je zapravo funkcija log-izglednosti?**
- A Funkcija koja primjeru \mathbf{x} pridjeljuje vjerojatnost $p(y|\mathbf{x})$, uz prepostavku da se ta vjerojatnost pokorava distribuciji definiranoj modelom $h(\mathbf{x}; \boldsymbol{\theta})$
 - B Funkcija koja uzoku \mathcal{D} pridjeljuje vjerojatnost parametra $\boldsymbol{\theta}$, uz prepostavku da se parametri pokoravaju distibuciji definiranoj modelom $h(\mathbf{x}; \boldsymbol{\theta})$
 - C Funkcija koja parametrima $\boldsymbol{\theta}$ pridjeljuje vjerojatnost uzorka \mathcal{D} , uz prepostavku da se uzorak pokorava distribuciji definiranoj modelom $h(\mathbf{x}; \boldsymbol{\theta})$
 - D Funkcija koja primjeru (\mathbf{x}, y) pridjeljuje vjerojatnost pripadanja skupu označenih primjera \mathcal{D} , uz prepostavku da su primjeri nezavisno i identično distribuirani
3. (T) Nepristranost je jedno od svojstava statističkih procjenitelja. Procjenitelj MLE ne mora nužno biti nepristran, tj. može biti pristran. **Za koji od sljedećih parametara distribucije procjena MLE pristrana?**
- A Srednja vrijednost Gaussove distribucije
 - B Varijanca Bernoullijeve distribucije
 - C Srednja vrijednost Bernoullijeve distribucije
 - D Kovarijacijska matrica Gaussove distribucije
4. (T) MAP-procenjenitelj definiramo kao $\hat{\boldsymbol{\theta}}_{\text{MAP}} = \operatorname{argmax}_{\boldsymbol{\theta}} p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$. Pri odabiru apriorne distribucije $p(\boldsymbol{\theta})$, nastojimo da je to neka standardna teorijska distribucija i da je konjugatna distribucija za izglednost $\mathcal{L}(\boldsymbol{\theta}|\mathcal{D})$. **Što to znači i zašto to želimo?**
- A To znači da je apriorna distribucija ista vrsta distribucije kao i vjerojatnost podataka uz dane parametre, tj. izglednost parametara, pa će njihov umnožak biti distribucija koja je proporcionalna aposteriornoj distribuciji i čiji ćemo maksimum moći izračunati Bayesovim pravilom
 - B To znači da je apriorna distribucija upravlјana hiperparametrima kojima možemo ugoditi distribucija parametara koji procjenjujemo, tj. parametri apriorne distribucije i parametri izglednosti su identični, što nam omogućava da te dvije distribucije pomnožimo i zatim nađemo maksimizator
 - C To znači da je aposteriorna distribucija parametara ista kao izglednost parametara, pa primjenom Bayesovog pravila možemo izračunati apriornu vjerojatnost parametara te, nakon zanemarivanja nazivnika koji je za fiksiran skup podataka konstantan, pronaći parametre koji maksimiziraju aposterionu vjerojatnost
 - D To znači da će umnožak izglednosti i apriorne distribucije dati distribuciju koja je iste vrste kao i apriorna distribucija, a ako je riječ o standardnoj teorijskoj distribuciji iz eksponencijalne familije, njezin mod (maksimizator) postoji u zatvorenoj formi, što nam omogućava da procjenitelj izračunamo analitički
5. (T) Kod MAP-procenjenitelja, apriorna distribucija parametra $p(\boldsymbol{\theta})$ tipično se odabire tako da bude konjugatna za funkciju izglednosti $p(\mathcal{D}|\boldsymbol{\theta})$. Prepostavimo da MAP-procenjenitelj izračunavamo heurističkom metodom (npr., gradijentnim usponom). Što se događa ako za apriornu distribuciju parametra upotrijebimo distribuciju koja *nije* konjugatna funkciji izglednosti?
- A Aposteriornu distribuciju $p(\boldsymbol{\theta}|\mathcal{D})$ ne možemo izvesti u zatvorenoj formi, ali MAP možemo izračunati heurističkom optimizacijom
 - B Zajedničku distribuciju $p(\mathcal{D}, \boldsymbol{\theta})$ ne možemo izvesti u zatvorenoj formi, pa MAP nije definiran
 - C Ako je apriorna distribucija $p(\boldsymbol{\theta})$ iz eksponencijalne familije, onda je aposteriona distribucija $p(\boldsymbol{\theta}|\mathcal{D})$ u zatvorenoj formi i MAP je izračunljiv
 - D Neovisno o apriornoj distribuciji parametra $p(\boldsymbol{\theta})$, MAP je izračunljiv optimizacijom drugog reda (npr., Newtonovim postupkom)

6. (T) Parametre modela možemo procijeniti pomoću procjenitelja najveće izglednosti (MLE) ili procjenitelja najveće aposteriorne vjerojatnosti (MAP). Općenito, MLE i MAP daju različite procjene, no u nekim slučajevima mogu dati jednake procjene. **Kada će MLE i MAP dati jednake procjene?**

- A Kada $p(\theta)$ definiramo kao unimodalnu distribuciju
- B Kada broj primjera N teži prema beskonačno
- C Kada je $p(\theta)$ konjugatna izglednost $p(\mathcal{D}|\theta)$
- D Kada se u \mathcal{D} realizirala svaka vrijednost slučajne varijable

7. (T) Za Bayesov klasifikator procjenjujemo parametar μ Bernoullijeve distribucije. Skup primjera za učenje je razmjerno malen. Naš procjenitelj $\hat{\mu}$ parametra μ može biti pristran ili nepristran, dok model čiji je parametar procijenjen s $\hat{\mu}$ može biti prenaučen ili može dobro generalizirati. Razmotrite procjenitelje MLE i MAP (konkretno, Laplaceovo zaglađivanje). **Što od sljedećega općenito vrijedi u ovom slučaju?**

- A - B + C - D - MAP procjenitelj je nepristran i očekujemo da će model dobro generalizirati

8. (T) Procjenitelj MAP kombinira funkciju izglednosti parametara s apriornom distribucijom parametara. **Što i kako maksimizira procjenitelj MAP?**

- A Zajedničku gustoću vjerojatnosti parametara i podataka, u zatvorenoj formi, ako je apriorna distribucija konjugatna za izglednost, inače iterativno
- B Apsteriornu vjerojatnost podataka, u zatvorenoj formi ako je zajednička vjerojatnost parametara i podataka iz eksponencijalne familije, inače iterativno
- C Zajedničku vjerojatnost parametara i podataka, iterativno ako je apriorna distribucija iz eksponencijalne familije, inače u zatvorenoj formi
- D Apsteriornu gustoću vjerojatnosti podataka, u zatvorenoj formi ako su izglednost i apriorna distribucija konjugatne, inače iterativno

15 Bayesov klasifikator

1. (T) Probabilistički modeli mogu biti generativni ili diskriminativni. U praksi su diskriminativni modeli nerijetko veće klasifikacijske točnosti od generativnih modela. **Zašto je tomu tako?**

- A Kod generativnih modela parametri se procjenjuju metodom najveće izglednosti koja je pristrana, dok se diskriminativni modeli uče gradijentnim spustom koji je statistički nepristran
- B Diskriminativni modeli s manje parametara mogu modelirati istu granicu između klasa kao i generativni modeli, pa trebaju manje primjera da ih se nauči, a i teže ih je prenaučiti
- C Generativni modeli mogu modelirati nelinearne zavisnosti između značajki, međutim kada su značajne stohastički nezavisne, to dovodi do prenaučenosti modela
- D Diskriminativni modeli modeliraju zajedničku distribuciju primjera i označaka, pa u slučaju preklapajućih distribucija u ulaznom prostoru ostvaruju veću točnost od generativnih modela

2. (T) Jedan od nedostataka generativnih modela u odnosu na diskriminativne modele jest nepotrebna složenost modeliranja. **Što to zapravo znači?**

- A Zajednička distribucija $p(\mathbf{x}, y)$ može se faktorizirati kao $p(y|\mathbf{x})p(\mathbf{x})$, no takva faktorizacija ima više parametara
- B Generativni modeli modeliraju distribuciju $p(y|\mathbf{x})$, što iziskuje više parametara nego li modeliranje granice između klasa
- C Za razliku od diskriminativnih modela, generativni modeli distribuciju $p(y|\mathbf{x})$ definiraju za sebe za svaku klasu, pa stoga imaju više parametara
- D Za klasifikaciju nam je potrebna samo distribucija $p(y|\mathbf{x})$, i ona se može modelirati sa manje parametara od zajedničke distribucije $p(\mathbf{x}, y)$

3. (T) Bayesov klasifikator definirali smo na sljedeći način: $h_j(\mathbf{x}; \boldsymbol{\theta}) = P(y = j|\mathbf{x}) = \frac{p(\mathbf{x}|y)P(y)}{\sum_{y'} p(\mathbf{x}|y')P(y')}$. Broj klasa neka je dva. Značajke neka su diskretne, i to neke s dvije, a neke s više od dvije moguće vrijednosti. **Koje teorijske distribucije ćemo koristiti za $P(y)$ i $P(\mathbf{x}|y)$?**

- A Bernoullijevu distribuciju za $P(y)$ i multinulijevu distribuciju za $P(\mathbf{x}|y)$
- B Kategoričku distribuciju za $P(y)$ i Gaussovnu distribuciju za $P(\mathbf{x}|y)$
- C Bernoullijevu distribuciju za $P(y)$ i Gaussovnu distribuciju za $P(\mathbf{x}|y)$
- D Gaussovnu distribuciju za $P(y)$ i multinulijevu distribuciju za $P(\mathbf{x}|y)$

4. (T) Bayesov klasifikator definirali smo na sljedeći način:

$$h_j(\mathbf{x}; \boldsymbol{\theta}) = P(y = j|\mathbf{x}) = \frac{p(\mathbf{x}|y)P(y)}{\sum_{y'} p(\mathbf{x}|y')P(y')}$$

Neka je broj klasa veći od dva, $K > 2$, a značajke neka su realni brojevi, $\mathbf{x} \in \mathbb{R}^n$. **Koje teorijske distribucije ćemo koristiti za $P(y)$ i $P(\mathbf{x}|y)$?**

- A Bernoullijevu distribuciju za $P(y)$ i Gaussovnu distribuciju za $P(\mathbf{x}|y)$
- B Kategoričku distribuciju za $P(y)$ i Gaussovnu distribuciju za $P(\mathbf{x}|y)$
- C Gaussovnu distribuciju za $P(y)$ i multinulijevu distribuciju za $P(\mathbf{x}|y)$
- D Kategoričku distribuciju za $P(y)$ i za $P(\mathbf{x}|y)$

5. (T) Bayesov klasifikator definiran je kao

$$h(\mathbf{x}; \boldsymbol{\theta}) = \operatorname{argmax}_y p(\mathbf{x}|y)P(y)$$

Po čemu se vidi da je ovo generativan, a ne diskriminativan model?

- A Zajedničku vjerojatnost primjera i oznaka, $p(\mathbf{x}|y)P(y)$, faktorizira u dva faktora te zanemaruje nazivnik $p(\mathbf{x})$, koji je ionako konstantan za svaku klasu y
- B Parametre distribucija $p(\mathbf{x}|y)$ i $P(y)$, a time indirektno i parametre aposteriorne distribucije $P(y|\mathbf{x})$, računa MAP-procjeniteljem, čime sprječava prenaučenost
- C Modelira vjerojatnost primjera i oznaka, budući da je, na temelju pravila umnoška, umnožak $p(\mathbf{x}|y)P(y)$ jednak zajedničkoj vjerojatnosti $p(\mathbf{x}, y)$
- D Primjer \mathbf{x} klasificira prema MAP-hipotezi, dakle u klasu koja maksimizira aposteriornu vjerojatnost oznake, $p(y|\mathbf{x})$, koja je proporcionalna zajedničkoj vjerojatnosti primjera i oznaka, $p(\mathbf{x}, y)$

6. (T) Jedan od parametara Gaussovog Bayesovog klasifikatora je kovarijacijska matrica $\boldsymbol{\Sigma}$ Gaussove multivariatne distribucije. Broj parametara matrice $\boldsymbol{\Sigma}$ koje treba procijeniti može se smanjiti uvođenjem pretpostavki o distribuciji primjera u ulaznom prostoru. **Uz koje minimalne pretpostavke će broj parametara za $\boldsymbol{\Sigma}$ biti linearan u broju značajki n ?**

- A Šum je isti za sve klase
- B Značajke nisu korelirane i šum ne ovisi o klasi
- C Šum je isti za sve klase i sve značajke
- D Nema linearne zavisnosti između značajki

16 Bayesov klasifikator II

1. (T) Gaussov Bayesov klasifikator s dijeljenom kovarijacijskom matricom (GBC) i logistička regresija (LR) su generativno-diskriminativni par modela. **Što to znači?**
 - A Aposteriorna vjerojatnost klase za GBC može se izraziti kao poopćeni linearni model sa sigmoidnom aktivacijskom funkcijom
 - B GBC i neregularizirana LR modeliraju identične distribucije zajedničke vjerojatnosti primjera i oznaka, ali s različitim brojem parametara
 - C Izlaz modela LR jednak je zajedničkoj vjerojatnosti modela GBC, ali model LR iziskuje manje parametara
 - D Neovisno o optimizacijskom postupku, GBC i LR ostvaruju istu pogrešku na skupu za učenje, ali uz moguće različit broj parametara
2. (T) Gaussov Bayesov klasifikator s dijeljenom kovarijacijskom matricom (GBC) i logistička regresija bez regularizacije (LR) čine generativno-diskriminativni par modela. **Što je od sljedećeg točno za taj generativno-diskriminativan par modela?**
 - A GBC i LR optimiraju različite funkcije pogreške, pa mogu dati različite granice između klasa
 - B Učenje modela GBC općenito je računalno složenije od učenja modela LR
 - C Oba modela daju linearnu granicu između klasa, ali GBC općenito ima više parametara od LR
 - D Vjerojatnosni izlaz modela GBC odgovara komplementu vjerojatnog izlaza modela LR
3. (T) Umjesto naivnog Bayesov klasifikatora, ponekad je bolje koristiti polunaivan Bayesov klasifikator? **Po čemu se polunaivan Bayesov klasifikator (SNBC) razlikuje od naivnog Bayesovog klasifikatora (NBC)?**
 - A SNBC prepostavlja zavisnosti između značajki i klase, ali ima manje parametara od NBC
 - B SNBC ima više parametara nego NBC, ali manje bridova kada ga se prikaže kao Bayesovu mrežu
 - C SNBC zajedničku vjerojatnost faktorizira u manje faktora, pa ima i manje parametara od NBC
 - D SNBC ima više parametara od NBC te modelira zavisnosti između značajki
4. (T) Polunaivan Bayesov klasifikator združuje u jedan faktor varijable kod kojih postoji statistička zavisnost. Za procjenu statističke zavisnosti između varijabli može se upotrijebiti Kullback-Leiblerova divergencija (KL-divergencija). **Kako se pomoću KL-divergencije može izračunati koliko su varijable zavisne?**
 - A Što su varijable manje zavisne, to je manja KL-divergencija između marginalnih vjerojatnosti i uvjetne vjerojatnosti
 - B Što su varijable manje zavisne, to je veća KL-divergencija između faktorizirane vjerojatnosti i marginalne vjerojatnosti
 - C Što su varijable više zavisne, to je veća KL-divergencija između zajedničke vjerojatnosti i faktorizirane vjerojatnosti
 - D Što su varijable više zavisne, to je manja KL-divergencija između zajedničke vjerojatnosti i faktorizirane vjerojatnosti

17 Probabilistički grafički modeli

1. (T) Za Bayesovu mrežu kažemo da je generativni i parametarski model. **Zašto?**
 - A Generativni jer definira zajedničku vjerojatnost svih varijabli, i opaženih i skrivenih, a parametarski jer se parametri modela mogu dobiti MLE-procjenom za svaki čvor Bayesove mreže zasebno, budući da se log-izglednost dekomponira po strukturi mreže
 - B Generativni jer se može koristiti za generiranje skupa primjera na temelju zajedničke distribucije, a parametarski jer su broj čvorova mreže i njihovo povezivanje (dakle graf) definirani parametrima koji se mogu ugađati na skupu za učenje, čime se mogu dobiti različite strukture mreže
 - C Generativni jer svaki čvor odgovara uvjetnoj vjerojatnosti koja je, na temelju Markovljevog uredajnog svojstva, generirana distribucijama čvorova roditelja, a parametarski jer Bayesova mreža zapravo definira zajedničku distribuciju koja je opisana skupom parametara
 - D Generativni jer opisuje postupak kojim se mogu generirati podatci koji se pokoravaju određenoj zajedničkoj vjerojatnosnoj distribuciji, a parametarski jer svaki čvor Bayesove mreže definira uvjetnu vjerojatnost preko teorijske distribucije koja je opisana svojim parametrima
2. (T) Bayesove mreže na sažet način prikazuju zajedničku distribuciju te kodiraju uvjetne stohastičke nezavisnosti između varijabli. No, kao i svaki model strojnog učenja, tako se i Bayesove mreže mogu prenaučiti. **Koja je veza između uvjetnih nezavisnosti varijabli u Bayesovoj mreži i opasnosti od prenaučenosti?**
 - A Uvođenjem pretpostavki o uvjetnoj nezavisnosti povećava se broj čvorova mreže, a time i broj parametara, što model čini složenijim i time sklonijim prenaučenosti
 - B Uvođenje pretpostavki o uvjetnoj nezavisnosti pojednostavljuje strukturu Bayesove mreže i smanjuje broj parametara, čime se smanjuje i mogućnost prenaučenosti
 - C Uvjetne nezavisnosti određuju strukturu mreže na način da definiraju koji su čvorovi mreže međusobno povezani, međutim to nema utjecaja na složenost modela niti na sklonost prenaučenosti
 - D Pretpostavke o uvjetnoj nezavisnosti čine induktivnu pristranost modela, pa što je više uvjetnih nezavisnosti, to je veća pristranost i model je lako prenaučiti
3. (T) Bayesova mreža na sažet način definira zajedničku distribuciju vjerojatnosti uz određene pretpostavke uvjetne nezavisnosti. Neka je jedna takva pretpostavka $x_1 \perp \{x_2, x_3\} | x_4$. **Koji je efekt uvođenja ove pretpostavke na graf Bayesove mreže?**
 - A Dodavanje dva brida
 - B Dodavanje tri brida
 - C Uklanjanje dva brida
 - D Uklanjanje tri brida
4. (T) Skriveni Markovljev model (HMM) posebna je vrsta Bayesove mreže. **Na što se odnosi pridjev "skriveni" u nazivu tog modela?**
 - A Model opisuje prijelaze između stanja, a trenutačno stanje ovisi samo o prethodnom stanju
 - B Neke varijable modela nisu opažene u podatcima, ali zavise o opaženim varijablama
 - C Opažene varijable ovise samo o trenutačnom stanju i prethodno opaženoj varijabli
 - D Stanja modela poredana su u lanac, a izlazi modela vidljivi su samo za posljednje stanje
5. (T) Bayesova mreža može se upotrijebiti za modeliranje kauzalnih odnosa između varijabli, odnosno za zaključivanje o uzrocima i posljedicama događaja. Jedan primjer takvog zaključivanja jest

“efekt objašnjavanja”, koji se primjenjuje kada postoji interakcija uzorka nekog događaja. **Gdje u Bayesovoj mreži nastupa efekt objašnjavanja i kako se on manifestira?**

- A U strukturi $x \rightarrow z \leftarrow y$, gdje su uzroci x i y nezavisni, ali postaju zavisni ako je opažena posljedica z
- B U strukturi $x \rightarrow z \rightarrow y$, gdje su uzrok x i posljedica y zavisni, ali postaju nezavisni ako je opažen posredni uzrok z
- C U strukturi $x \rightarrow z \leftarrow y$, gdje su uzroci x i y zavisni, ali postaju nezavisni ako je opažena posljedica z
- D U strukturi $x \rightarrow z \rightarrow y$, gdje su uzrok x i posljedica y nezavisni, ali postaju zavisni ako je opažen posredni uzrok z

18 Probabilistički grafički modeli II

1. (T) Čest način probabilističkog zaključivanja kod Bayesovih mreža jest izračunavanje “aposteriornog upita”. Kod te vrste upita zanima nas distribucija nekih varijabli (variable upita) na temelju zadanih varijabli (opažene variable). Međutim, Bayesova mreža kodira zajedničku vjerojatnost svih varijabli mreže, uključivo i varijabli koje nisu niti variable upita niti opažene variable (variable smetnje). **Na koji način izračunavamo aposteriorni upit?**
 - A Kao omjer zajedničke vjerojatnosti marginalizirane po varijablama smetnje i zajedničke vjerojatnosti marginalizirane po varijablama smetnje i varijablama upita
 - B Kao omjer zajedničke vjerojatnosti marginalizirane po varijablama upita i zajedničke vjerojatnosti marginalizirane po opaženim varijablama
 - C Kao umnožak zajedničke vjerojatnosti marginalizirane po opaženim varijablama i zajedničke vjerojatnosti marginalizirane po varijablama smetnje
 - D Kao umnožak vjerojatnosti varijable upita uvjetovane na opažene varijable i zajedničke vjerojatnosti marginalizirane po varijablama smetnje
2. (T) Glavna svrha probabilističkih grafičkih modela (PGM) jest provođenje probabilističkih upita. Jedna vrsta upita su MAP-upiti (upiti najvjerojatnijeg objašnjenja). Neka su \mathbf{x}_q , \mathbf{x}_o i \mathbf{x}_n skupovi varijabli upita, opaženih varijabli odnosno varijabli smetnje. **Kako je definiran rezultat MAP-upita?**
 - A $\operatorname{argmax}_{\mathbf{x}_o} \sum_{\mathbf{x}_n} p(\mathbf{x}_q, \mathbf{x}_o, \mathbf{x}_n)$
 - B $\operatorname{argmax}_{\mathbf{x}_o} \sum_{\mathbf{x}_q} p(\mathbf{x}_q, \mathbf{x}_o, \mathbf{x}_n)$
 - C $\operatorname{argmax}_{\mathbf{x}_q} \sum_{\mathbf{x}_o} p(\mathbf{x}_q, \mathbf{x}_o, \mathbf{x}_n)$
 - D $\operatorname{argmax}_{\mathbf{x}_q} \sum_{\mathbf{x}_n} p(\mathbf{x}_q, \mathbf{x}_o, \mathbf{x}_n)$
3. (T) Približno zaključivanje kod Bayesovih mreža može se provesti metodama uzorkovanja. Te metode generiraju slučajan uzorak primjera iz distribucije opisane Bayesovom mrežom, iz kojega se onda mogu procijeniti parametri distribucije. Međutim, da bismo mogli uzorkovati iz Bayesove mreže, već trebamo znati parametre svih uvjetnih distribucija u čvorovima. **Zašto radimo uzorkovanje iz Bayesove mreže da bismo procijenili parametre distribucije, kad te parametre već imamo?**
 - A Ne trebamo znati parametre svih čvorova, već samo čvorova roditelja onih distribucija za koje želimo procijeniti parametre
 - B Vrijednosti parametara svih distribucija su procjene koje inicijaliziramo i zatim iterativno ažuriramo u svakom koraku uzorkovanja
 - C Uzorkovanje koristimo da bismo procijenili parametre bilo koje distribucije, a ne samo uvjetnih distribucija u čvorovima mreže
 - D Uzorkovanjem procjenjujemo parametre latentnih varijabli, čije vrijednosti ne opažamo, pa nam ti parametri nisu poznati prije uzorkovanja

4. (T) Bayesove mreže možemo upotrijebiti za procjenu aposteriorne vjerojatnosti $P(\mathbf{x}_q|\mathbf{x}_o)$, gdje je \mathbf{x}_q vektor varijabli upita a \mathbf{x}_o vektor opaženih varijabli. U tu se svrhu često koriste metode približnog zaključivanja. Najjednostavnija takva metoda je uzorkovanje s odbacivanjem, međutim ta metoda ima nedostatak zbog koje se u praksi ne koristi. **Koji je nedostatak metode uzorkovanja s odbacivanjem?**
- Ako je vjerojatnost $P(\mathbf{x}_o)$ mala, treba generirati mnogo vektora da bi uzorak bio velik i procjena pouzdana
 - Ako je vjerojatnost $P(\mathbf{x}_o)$ velika, mnogo će generiranih vektora biti odbačeno i procjena će biti pristrana
 - Ako je vjerojatnost $P(\mathbf{x}_o)$ mala, generiramo vektore iz apriorne distribucije $P(\mathbf{x}_q)$ i procjena je pristrana
 - Ako je vjerojatnost $P(\mathbf{x}_o)$ velika, generirani vektori nisu iz aposteriorne distribucije i procjena je netočna
5. (T) Unaprijedno uzorkovanje jedna je od metoda za uzorkovanje iz distribucije opisane Bayesovom mrežom. **Koji je problem kod primjene unaprijednog uzorkovanja za izračun aposteriornih upita nad Bayesovom mrežom?**
- Mnogo uzorkovanih vektora morat će biti odbačeni, pa procjena neće biti pouzdana
 - Dobiveni vektori bit će uzorkovani iz apriorne, a ne aposteriorne distribucije
 - Kod marginalizacije varijabli smetnje dolazi do kombinatorne eksplozije
 - Skrivene varijable nisu opažene, pa se log-izglednost ne dekomponira po varijablama mreže
6. (T) Procjena parametara Bayesove mreže temelji se na maksimizaciji log-izglednosti parametara pod modelom. Procjena parametara može biti bitno drugačija za slučaj potpunih podataka, gdje su sve varijable opažene, u odnosu na slučaj nepotpunih podataka, gdje u model trebamo uključiti skrivene ili latentne varijable. **Što je prednost procjene parametara kod potpunih podataka (modela bez skrivenih varijabli) u odnosu na nepotpune podatke (modela sa skrivenim varijablama)?**
- Kod potpunih podataka minimizacija funkcije log-izglednosti ima rješenje u zatvorenoj formi, ali funkcija nije konkavna, pa može imati više lokalnih optimuma, za razliku od modela sa skrivenim varijablama koji ima više parametara, ali konkavnu funkciju log-izglednosti
 - Kod potpunih podataka maksimizacija log-izglednosti ima rješenje u zatvorenoj formi, ali samo ako su opažene varijable na početku niza po topološkom uređaju čvorova, za razliku od modela sa skrivenim varijablama kod kojega MLE procjenitelj ne postoji u zatvorenoj formi
 - Kod potpunih podataka MLE procjena parametara ima rješenje u zatvorenoj formi, dok MAP procjena nema, za razliku od modela sa skrivenim varijablama kod kojeg je situacija obrnuta, a k tome taj model ima još više parametara od modela bez skrivenih varijabli
 - Kod potpunih podataka log-izglednost se dekomponira po strukturi mreže, pa parametre svake uvjetne distribucije možemo procijeniti nezavisno od drugih čvorova i u zatvorenoj formi, međutim parametara može biti više nego kod modela sa skrivenim varijablama
7. (T) Parametre probabilističkih grafičkih modela, uključivo Bayesove mreže, možemo procjenjivati iz potpunih podataka ili nepotpunih podataka. **Zašto i kako Bayesovu mrežu učimo nad nepotpunim podatcima?**
- Jer se log-izglednost dekomponira po čvorovima mreže, pa MLE ili MAP procjenjujemo u zatvorenoj formi
 - Jer MLE nema rješenje u zatvorenoj formi, pa umjesto MLE koristimo gradijentni uspon ili EM-algoritam
 - Jer mreža ima skrivene varijable koje ne opažamo, pa moramo koristiti iterativne metode za MAP ili MLE
 - Jer mreža ima manje čvorova nego što je opaženih varijabli, pa koristimo eliminaciju varijabli

19 Grupiranje

1. (T) Algoritam K-sredina je iterativan algoritam za nalaženje parametara $b_k^{(i)}$ (pripadnosti primjera grupama) i μ_k (centroidi grupe) koji minimiziraju kriterijsku funkciju J za N primjera i K grupa. Algoritam funkciju J optimizira iterativno, jer rješenje u zatvorenoj formi ne postoji. **Zbog čega za problem minimizacije funkcije J ne postoji rješenje u zatvorenoj formi?**
 - A Jer za $b_k^{(i)}$ mora vrijediti ograničenje $\sum_k b_k^{(i)} = 1$ i $b_k^{(i)} \geq 0$
 - B Jer J ovisi o K i inicijalnom odabiru za μ_k , pa rješenje nije jedinstveno
 - C Jer $b_k^{(i)}$ ovisi o vektorima μ_k , a vektor μ_k ovisi o vrijednostima $b_k^{(i)}$
 - D Jer $b_k^{(i)}$ ovisi o N , broj vektora μ_k ovisi o K , a K je odozgo ograničen sa N
2. (T) Konvergencija je poželjno svojstvo algoritma grupiranja. **Je li točno da algoritam k-sredina uvijek konvergira?**
 - A Da, algoritam uvijek konvergira zato što je broj particija N primjera u K skupova ograničen, a optimizacijski postupak definiran je tako da se J u svakoj iteraciji smanjuje
 - B Algoritam konvergira samo ako su početna središta dobro odabrana, inače se može dogoditi da algoritam oscilira između dva rješenja
 - C Kako se radi o algoritmu koji grupira primjere u vektorskom prostoru, broj rješenja je neograničen, stoga algoritam ne mora konvergirati
 - D Algoritam uvijek konvergira zato što je broj primjera N uvijek veći ili jednak broju grupe K , a kao mjera udaljenosti koristi se euklidska udaljenost, koja je nužno nenegativna
3. (T) Algoritam K-means++ proširenje je algoritma K-sredina heurističkim odabirom početnih središta grupe. **Koja je glavna ideja odabira početnih središta grupe kod algoritma K-means++?**
 - A Vjerovatnosc da je neki primjer središte grupe proporcionalna je s brojem primjera u toj grupi i udaljenosti tih primjera od centroida skupa podataka
 - B Središta grupe su srednje vrijednosti grupe projiciranih na pravac u smjeru prve komponente rastava skupa podataka na glavne komponente (PCA)
 - C Središta grupe su mjesta na kojima graf kriterijske funkcije u ovisnosti o broju grupe naglo opada pa stagnira
 - D Najvjerojatnije središte grupe jest primjer koji je najviše udaljen od njemu najbližeg središta
4. (T) Algoritmi grupiranja k-sredina i k-medoida razlikuju se, između ostalog, i po vremenskoj računalnoj složenosti. Naime, algoritam k-medoida računalno je složeniji od algoritma k-sredina. **Zašto je algoritam k-medoida računalno složeniji od algoritma k-sredina?**
 - A Za razliku od algoritma k-sredina koji se zasniva na euklidskoj udaljenosti, čiji je izračun računalno nezahtjevan, algoritam k-medoida koristi funkcije sličnosti čije računanje iziskuje mnogo računalnih operacija
 - B Budući da algoritam k-medoida ne koristi centroide, nego medoide, na kraju svake iteracije mora kombinatoričkom provjerom po primjerima pronaći medoide koje minimiziraju kriterijsku funkciju J
 - C Za razliku od algoritma k-sredina, algoritam k-medoida je algoritam mekog grupiranja, što iziskuje provođenje dodatnih koraka unutar algoritma
 - D Kriterijska funkcija algoritma k-medoida jest mnogo složenija od one k-sredina, upravo zato što algoritam k-medoida koristi medoide, a ne centroide
5. (T) Algoritam K-medoida općenitiji je od algoritma K-sredina budući da se može koristiti za primjere koji nisu prikazani kao vektori. Međutim, razmotrite slučaj kada primjeri jesu prikazani

kao vektori, ali ih želimo grupirati na temelju mjere udaljenosti koja nije euklidska. **Koji bismo algoritam koristili u tom slučaju i zašto?**

- A Algoritam K-medoida, jer kriterijska funkcija algoritma K-sredina koristi euklidsku udaljenost
 - B Algoritam K-sredina, jer za vektorizirane primjere možemo izračunati centroide grupe
 - C Algoritam K-medoida, jer vektorizirani primjeri također mogu biti medoidi
 - D Algoritam K-sredina, jer koristi mjeru udaljenosti, dok algoritam K-medoida koristi mjeru sličnosti
6. (T) Algoritam K-medoida proširenje je algoritma K-sredina na neeuclidske ulazne prostore. Kod algoritma K-medoida kao funkciju različitosti ν možemo u načelu koristiti bilo koju funkciju. Specifično, ako je ulazni prostor vektorski, možemo koristiti euklidsku udaljenost. **Što bi se dogodilo da kao funkciju različitosti koristimo euklidsku udaljenost?**
- A Algoritam bi bio davao grupe koje bi u prosjeku bile više izdužene nego one dobivene algoritmom K-sredina
 - B Algoritam K-medoida davao bi isto grupiranje kao i algoritam K-sredina, ali s većom vremenskom složenošću
 - C Algoritam bi uz veću prostornu složenost davao grupiranje u manji broj grupa nego algoritam K-sredina
 - D Algoritam bi primjere grupirao slično kao i algoritam K-sredina, pogotovo ako u središtima grupe postoje primjeri

20 Grupiranje II

1. (T) Model miješane gustoće sa K komponenti vjerojatnosnu distribuciju neoznačenih podaka definira kao $p(\mathbf{x}) = \sum_{k=1}^K \pi_k p(\mathbf{x}|\boldsymbol{\theta}_k)$. Što modelira uvjetna vjerojatnost $p(\mathbf{x}|\boldsymbol{\theta}_k)$?
- A Gustoću vjerojatnosti primjera \mathbf{x} unutar grupe k
 - B Vjerojatnost da primjer \mathbf{x} pripada grapi k
 - C Stupanj pripadnosti primjera \mathbf{x} grapi k
 - D Gustoću vjerojatnosti grupe k za primjer \mathbf{x}
2. (T) Za meko grupiranje možemo koristiti model miješane gustoće s latentnim varijablama. Kod tog modela, svaki primjer $\mathbf{x}^{(i)}$ ima pridruženu latentnu varijablu $\mathbf{z}^{(i)}$. Obje ove varijable zapravo su slučajne varijable sa svojom pretpostavljenom distribucijom. **Koju distribuciju prepostavljamo za latentnu varijablu $\mathbf{z}^{(i)}$ i zašto?**
- A Beta-distribuciju koja opisuje s kojom vjerojatnošću primjer $\mathbf{x}^{(i)}$ pripada svakoj grapi
 - B Kategoričku distribuciju koja opisuje kojoj grapi primjer $\mathbf{x}^{(i)}$ zapravo pripada
 - C Bernoullijevu distribuciju koja opisuje s kojom vjerojatnošću primjer $\mathbf{x}^{(i)}$ pripada grapi $\mathbf{z}^{(i)}$
 - D Gaussovnu distribuciju koja opisuje vjerojatnost da je grupa generirala primjer $\mathbf{x}^{(i)}$
3. (T) Model Gaussove mješavine često treniramo algoritmom maksimizacije očekivanja. **Na što se odnosi pojam "očekivanje" u nazivu tog algoritma?**
- A Izglednost parametara modela izračunata uz fiksiranu pripadnost svakog primjera njemu najbližoj grapi
 - B Vjerojatnost parametara modela s fiksiranim dodijeljivanjem primjera grupama izračunata na temelju maksimizacije log-izglednosti
 - C Vjerojatnost skupa podataka pod modelom s fiksiranim parametrima izračunata na temelju vjerojatnosti pripadanja primjera svakoj grapi
 - D Vjerojatnost pripadanja primjera svakoj grapi izračunata Bayesovim pravilom na temelju modela Gaussove mješavine

4. (T) Algoritam maksimizacije očekivanja (EM-algoritam) maksimizira očekivanje potpune log-izglednosti, što se pokazuje da dovodi i do maksimizacije nepotpune log-izglednosti. **Koja je razlika između potpune i nepotpune log-izglednosti, i zašto maksimiziramo očekivanje potpune log-izglednosti umjesto izravno log-izglednost?**
- A Potpuna log-izglednost je izglednost izračunata na svim primjerima iz neoznačenog skupa primjera, dok je nepotpuna log-izglednost izračunata samo za označene primjere koji se koriste za evaluaciju modela, a očekivanje računamo zato jer je postupak grupiranja stohastičan
 - B Potpuna log-izglednost je log-izglednost s neopaženim varijablama, a u slučaju GMM-a to su centroidi i kovarijacijske matrice komponenata, koje procjenjujemo metodom MLE, koja maksimizira očekivanje log-izglednosti
 - C Potpuna log-izglednost računa se za označene primjere a nepotpuna log-izglednost za neoznačene primjere, a u oba slučaja kod modela GMM računamo očekivanje log-izglednosti jer postupak za različite početne centroide može dati različite log-izglednosti
 - D Potpuna log-izglednost je log-izglednost modela GMM s latentnim varijablama, koje definiraju koji primjer pripada kojoj grupi, međutim kako to zapravo ne znamo, moramo računati s očekivanjem tih varijabli
5. (T) Za procjenu parametara modela GMM tipično se koristi algoritam maksimizacije očekivanja (EM-algoritam). To je iterativan optimizacijski algoritam. **Pod kojim uvjetima EM-algoritam (primijenjen na model GMM) konvergira, i kamo?**
- A Krenuvši od nekih početnih parametara, algoritam uvijek konvergira do parametara koji maksimiziraju očekivanje log-izglednosti, međutim to ne moraju biti parametri koji maksimiziraju vjerojatnost podataka
 - B Algoritam konvergira samo ako su primjeri u ulaznom prostoru sferični, ako su zavisnosti između značajki linearne, i ako nema multikolinearnosti, jer u protivnom zavisnosti nije moguće modelirati kovarijacijskom matricom
 - C Algoritam uvijek konvergira, i to do točke u prostoru parametara koja maksimizira log-izglednost parametara, no brzina konvergencije ovisi o toma kako su inicijalizirani parametri
 - D Algoritam uvijek konvergira, međutim globalni maksimum log-izglednosti parametara doseže samo ako je broj grupe postavljen na pravi broj grupa ili tako da je broj grupe jednak broju primjera
6. (T) Algoritam GMM, odnosno model Gaussove mješavine s algoritmom maksimizacije očekivanja kao optimizacijskim postupkom, poopćenje je algoritma k-sredina. **Uz koje uvjete algoritam GMM degenerira u algoritam k-sredina?**
- A Umjesto maksimizacije log-izglednosti, minimizira se negativna log-izglednost, a početna središta se odabiru algoritmom k-sredina
 - B Koeficijenti mješavine su jednaki za sve komponente Gaussove mješavine, a kovarijacijske matrice su dijagonalne
 - C Kovarijacijska matrica komponenti Gaussove mješavine je dijeljena i izotropna, a odgovornosti su zaokružene na cijeli broj
 - D Kovarijacijska matrica komponenti Gaussove mješavine je jedinična matrica, a maksimizira se negativna log-izglednost
7. (T) Za grupiranje primjera u K grupe koristimo model Gaussove mješavine (GMM) s dijeljenom kovarijacijskom matricom. Nakon grupiranja, odgovornosti zaokružujemo na cijeli broj, čime dobivamo tvrdo grupiranje. Iste podatke grupiramo algoritmom K-medoida (KM). **Uz koje parametre ovih algoritama očekujemo dobiti najsličnije rezultate grupiranja?**
- A GMM: $\Sigma_k = \text{diag}(\sigma_1^2, \dots, \sigma_n^2)$ i $\pi_k = 1/K$; KM: euklidska udaljenost
 - B GMM: $\Sigma_k = \sigma^2 \mathbf{I}$; KM: euklidska udaljenost
 - C GMM: puna Σ_k i $\pi_k = 1/K$; KM: Mahalanobisova udaljenost
 - D GMM: $\Sigma_k = \sigma^2 \mathbf{I}$ i $\pi_k = 1/K$; KM: Mahalanobisova udaljenost

8. (T) Broj grupa K hiperparametar je mnogih algoritama grupiranja, pa tako i algoritma GMM. Optimalan broj grupa može se odrediti na razne načine, a jedan od njih je Akaikeov kriterij. **Na kojem se principu temelji odabir broja grupa Akaikeovim kriterijem?**

- A Model s optimalnim brojem grupa je onaj koji minimizira log-izglednost nepotpunih podataka, a maksimizira log-izglednost potpunih podataka
- B Optimalan broj grupa je onaj koji maksimizira očekivanje log-izglednost modela, uz prepostavku izotropne kovarijacijske matrice
- C Model s optimalnim brojem grupa je onaj koji podatke čini najvjerojatnijima, ali to čini sa što manje parametara
- D Optimalan broj grupa je onaj kod kojeg, nakon dalnjeg povećanja broja grupa, vrijednost log-izglednosti stagnira ili blago raste

9. (T) Optimizaciju parametara modela Gaussove mješavine (GMM) ne provodimo u zatvorenoj formi. S druge strane, parametre Gaussovog Bayesovog klasifikatora, koji je sličan modelu GMM, optimiramo u zatvorenoj formi. **Zašto parametre GMM-a ne optimiramo u zatvorenoj formi, dok kod Gaussovog Bayesovog klasifikatora to radimo?**

- A Za razliku od Gaussovog Bayesovog klasifikatora, GMM je nenadizirani algoritam, pa log-izglednost podataka nije definirana i nije moguća maksimizacija u zatvorenoj formi
- B Kod GMM-a, pored koeficijenata mješavine i vektora sredina, trebamo procijeniti i kovarijacijske matrice, za što ne postoji procjenitelj u zatvorenoj formi
- C Kod GMM-a ne znamo koji primjer pripada kojoj grupi, pa je gustoća primjera jednaka zbroju gustoći komponenti, za što ne postoji maksimizator u zatvorenoj formi
- D Parametri oba modela mogu se optimirati u zatvorenoj formi, međutim kod modela GMM računalno je jednostavnije koristiti EM-algoritam

10. (T) Algoritam k-medoida proširenje je algoritma k-sredina. **Što algoritam k-medoida i algoritam hijerarhijskog grupiranja HAC imaju zajedničko, a po čemu se razlikuju?**

- A Oba algoritma mogu grupirati na temelju mjere udaljenosti koja nije euklidska, no algoritam k-medoida ima veću vremensku složenost od algoritma HAC
- B Oba algoritma imaju vremenska složenost veću od linearne u broju primjera, no kod k-medoida primjer može pripada u više grupe istovremeno
- C Oba algoritma izvode se onoliko iteracija koliko ima grupe, no algoritam k-medoida može raditi s općenitom mjerom sličnosti ili udaljenosti
- D Oba algoritma mogu grupirati primjere koji nisu vektorizirani, no algoritam HAC daje hijerarhiju dok algoritam k-medoida daje particiju grupe

21 Vrednovanje modela

1. (T) F_1 -mjera računa točnost klasifikatora kao harmonijsku sredinu preciznosti (P) i odziva (R). **Zašto se za F_1 -mjeru koristi harmonijska, a ne aritmetička sredina?**

- A Jer su P i R obrnuto proporcionalni, ali definirani na istom intervalu
- B Jer niti P niti R ne uzimaju u obzir broj stvarno negativnih primjera
- C Jer P i R nisu definirani na istoj skali, ali njihove recipročne vrijednosti jesu
- D Jer P nije definiran za trivijalan klasifikator koji sve primjere klasificira negativno

2. (T) Mjera točnosti nije prikladna za vrednovanje klasifikatora na skupovima podataka s neuravnoteženim brojem primjera po klasama. Jedna alternativa mjeri točnosti je F_1 -mjera, međutim ni ta mjeru nije uvijek prikladna. Prepostavite da vrijednost F_1 -mjere postavljamo na nulu u slučajevima kada je harmonijska sredina preciznosti i odziva nedefinirana. **U kojem slučaju**

F_1 -mjera ne bi bila prikladna mjera za vrednovanje klasifikatora jer bi bila previše optimistična?

- A Ako je većina primjera pozitivna i klasifikator sve primjere klasificira pozitivno
 - B Ako je većina primjera pozitivna, a klasifikator sve primjere klasificira negativno
 - C Ako je većina primjera negativna, a klasifikator sve primjere klasificira pozitivno
 - D Ako je većina primjera negativna i klasifikator sve primjere klasificira negativno
3. (T) Za vrednovanje višeklasnog klasifikatora često se koriste mjere F_1^μ (mikro F_1 -mjera) i F_1^M (makro F_1 -mjera). **Koji je očekivani odnos između vrijednosti tih mjeri, i zašto?**
- A $F_1^M < F_1^\mu$, jer kod makro F_1 -mjere računamo prosjek F_1 -mjere kroz sve klase, a klasifikator na manjim klasama više grijesi
 - B $F_1^M > F_1^\mu$, jer kod mikro F_1 -mjere zbrajamo matrice zabune kroz sve klase, a primjera iz manjih klasa ima manje, pa manje doprinose pogrešci
 - C $F_1^M > F_1^\mu$, jer kod makro F_1 -mjere zbrajamo matrice zabune kroz sve klase, a primjera iz većih klasa ima više, pa više doprinose pogrešci
 - D $F_1^M < F_1^\mu$, jer kod mikro F_1 -mjere računamo prosjek F_1 -mjere kroz sve klase, a klasifikator na većima klasama manje grijesi
4. (T) Za vrednovanje klasifikatora na manjim skupovima podataka može se koristiti metoda unakrsne provjere "izvoji jednog" (engl. *leave-one-out cross-validation*, LOOCV). Prednost te metode je što se procjena dobiva na temelju mnogo uzoraka. No, metoda ima i nekih nedostatka. **Što je nedostatak metode LOOCV?**
- A Procjena je pristrana jer se ispitni skupovi preklapaju u $1/N$ primjera
 - B Varijanca procjene je visoka jer klasifikatori dijele $(N - 2)/N$ primjera za učenje
 - C Procjena može biti pesimistična, jer ispitani model može biti suboptimalne složenosti
 - D Procjena je pristrana jer se svaki primjer u skupu za ispitivanje koristi N puta
5. (T) Procjena pogreške modela metodom unakrsne provjere omogućava nam da procijenimo prediktivnu moć modela, mjerenu kao točnost modela na neviđenom skupu primjera. Daljnja razrada te ideje je ugniježđena višestruka unakrsna provjera, koja se u praksi vrlo često koristi. **Koja je motivacija za korištenje ugniježđene višestruke unakrsne provjere, umjesto obične unakrsne provjere?**
- A Omogućava nam da procijenimo prediktivnu moć modela optimalne složenosti te maksimalno iskoristimo raspoložive podatke za učenje i ispitivanje
 - B Provodi optimizaciju hiperparametra modela na uniji skupa za provjeru i skupa za testiranje, čime postiže bolju točnost modela jer više primjera ostaje za treniranje
 - C Razdvaja skup za učenje od skupa za ispitivanje te time osigurava da doista mjerimo prediktivnu moć modela, odnosno ispitnu pogrešku, a ne pogrešku učenja
 - D Omogućava nam da odredimo točnost modela s klasifikacijskim pragom, na način da u obzir uzimamo preciznost i odziv za različite vrijednosti klasifikacijskog praga
6. (T) Ugniježđena k-struka unakrsna provjera često se koristi za procjenu točnosti modela. **Što je prednost ugniježđene k-struke unakrsne provjere u odnosu na običnu unakrsnu provjeru?**
- A Procjenjujemo pogrešku generalizacije, a ne pogrešku učenja
 - B Model ispitujemo na cijelom raspoloživom skupu primjera
 - C Procjenjujemo ispitnu pogrešku modela s najmanjom pogreškom učenja
 - D Procjenjujemo očekivanu ispitnu pogrešku modela optimalne složenosti

Rješenja

	1	2	3	4	5	6	7	8	9	10
2. Osnovni koncepti	B	B	A	A	D	C	B	D	A	
3. Linearna regresija	C	C	B	C	B	A	B	D		
4. Linearna regresija II	A	B	B	C	B	A				
5. Linearni diskriminativni modeli	C	A	D	B	B	D				
6. Logistička regresija	B	B	B	C	D					
7. Logistička regresija II	B	C	A	B	C	C	D	B	B	
8. Stroj potpornih vektora	A	D	C	C	B	D				
9. Stroj potpornih vektora II	D	D	A	B	A					
10. Jezgrene metode	C	A	D	B	D	D	D			
11. Neparametarske metode	C	C	B	C						
14. Procjena parametara II	C	C	D	D	A	B	B	A		
15. Bayesov klasifikator	B	D	A	B	C	D				
16. Bayesov klasifikator II	A	C	D	C						
17. Probabilistički grafički modeli	D	B	C	B	A					
18. Probabilistički grafički modeli II	A	D	C	A	B	D	C			
19. Grupiranje	C	A	D	B	A	D				
20. Grupiranje II	A	B	C	D	A	C	C	C	C	D
21. Vrednovanje modela	C	A	A	B	A	D				

TEORIJSKA PITNJA

- MI

2 - Osnovni koncepti

(1)

Zašto pogrešni modeli opravdu empirijskom pregrškom i na lojej pretpostavlja se tenuelji ta oprešimacija?

(B)

Ne možemo izračunati očekivanje aubitka jer nam nije poznata distribucija primjera $\mathbb{P}(X \neq Y)$.
Pretpostavljamo da je \mathcal{D} represent. uzorak.

(2)

$$\mathcal{H} = \{h(\vec{x}; \vec{\theta})\}_{\vec{\theta}}$$

Što to znači?

(B)

Različite funkcije h imaju različite parametre $\vec{\theta}$ i da su sve one sadržane u modelu, tj. za sve njih vrijedi $h \in \mathcal{H}$.

Zašto ne:

A = ne mora biti ∞ mnoge fja

C = više različitih $\vec{\theta}$ može definirati isti h

D = broj razl. $h \neq \# \text{param}$

(3)

Razlika između param. i hiperparam?

(A)

parametar optim. alg., hiperparam. ne

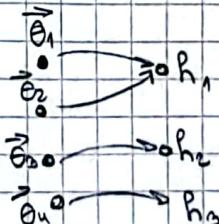
(4)

\mathcal{H} indeksiran $\vec{\theta}$. Što to znači?

(A)

Svaki $\vec{\theta}$ jednoznačno određuje funkciju loja primjer \vec{x} preslikava u iznaku y u okviru o parom. $\vec{\theta}$

! svaki $\vec{\theta}$ jednoznačno određuje h , ali mogu $\vec{\theta}_1$ i $\vec{\theta}_2$ određivati istu h !



(5)

$$h: X \rightarrow Y$$

h definirana do na parametre $\vec{\theta}$

Što to znači?

(D) Različite vrijednosti $\vec{\theta}$ mogu dati različite funkcije h , a skup svih takvih h definira model \mathcal{H}

Zašto ne!

(B) - videti gornju sliku

(C) - ne mora h biti def. bez param.

(A) - h ne određuje jednoznačno $\vec{\theta}$ je prostora

slup min. prepo. pomakaju označa svih
primjera slijedi

DETUXEFTVNO

6) Tačto stvrgno učenje bez indukt. pristranci nije moguće?

(C) Označa niti jednog nevidjenog primjera ne bi bila moguća

Tačto ne!

B - indukt. pristranci nema veze s linear. odgovarajuću

D - indukt. pristranci ne ograničava prostor param.

A - nema veza indukt. pristr. sa složenčcu

7) Tačto šum u podacima za učenje može dovesti do prenauč. klasič. modela?

(B) Efekt šuma je slučajan.

Hipoteza koja uz previše prilogađen šumu na trainicu očekivano imati veliki pogrešku na testu gdje je šum drugačiji

8) Što je induktivna pristranci?

$$D \wedge X \wedge B \rightarrow h_L(\bar{x})$$

(D) Minimum slup predstavkei boje uz D, jednocačno određuju klasič. svakog primjera

A - optimizacija!

B - ne određuje model nego h

C - straight up linija

$$\vec{\theta} \in \mathbb{R}^{n+1}$$

$$\mathcal{H} = \{h(\cdot, \vec{\theta})\}_{\vec{\theta}}$$

Može li \mathcal{H} biti beskonačan?

(A) Da, npr. $X = \mathbb{R}^n$ i $h(\bar{x}, \vec{\theta}) = \vec{\theta}^\top \bar{x}$

C - ovo je končan \mathcal{H}

3 - Linearna regresija

1

Induktivna pristranost preferencije modela linearne regresije?

C) Težine \vec{w} minimiziraju $\|X\vec{w} - \vec{y}\|^2$

A i D - pristranost jezika / ogranič
B - suprotno od C

2

$$\vec{w} = (X^T X)^{-1} X^T \vec{y}$$

Pod ovojim uvjetima \vec{w} možemo izračunati vektor, o čemu dom. ovisi leženost tog protupca?

$$X \rightarrow N \times (n+1)$$

$$X^T X \Rightarrow (n+1) \times N \circ N \times (n+1) \\ \Rightarrow (n+1) \times (n+1)$$

C) Dimenzije matrice X mora biti $n+1$

Složenost dominantno ovisi o n : $O(n^3)$

3

Induktiv. pristranost reg. i nereg. linearne regresije?

B) Oba algoritma ISTI model $\vec{w}^T \vec{x}$ (ista pristranost jezika)
Različito definirana empirijska pogreška (osim ako je $\lambda = 0$)

A } kaže da imaju različite modele
C }
D - kaže da je ista optim.

4

Kako form. glasi probabilistička interpretacija modela linearne regresije?

$$C) p(y | \vec{x}) = \mathcal{N}(h(\vec{x}), \sigma^2)$$

5

Što je indukt. pristranost preferencije linear. modela regresije?

B) minimizacija $\|X\vec{w} - \vec{y}\|^2$

C - ind. pristr. jezika \vec{w}

A - pretp. i.i.d nije ind. pri: prof. (možda jezika ??)

D - da pise minimum. $-\ln(L(\vec{w} | \vec{y}))$ bilo bi točno.

6

Koliko redaka i stupaca ima matrica koju invertiramo u L2-reg. reg?

$$\vec{w} = (\vec{\Phi}^T \vec{\Phi} + \lambda \mathbb{I})^{-1} \vec{\Phi}^T \vec{y}$$
$$\hookrightarrow (m+1) \times N \cdot N \times (m+1) = (m+1) \times (m+1)$$
$$\hookrightarrow (m+1) \times (m+1) + (m+1) \times (m+1)$$

A

$$m+1$$

7) Kada je $X^+ = X^{-1}$?

$$X^+ = (X^T X)^{-1} X^T$$

(B) Kada je broj značajki manji od broja primjera ($n < N$) i nema multikolinarnosti

8) Pod čijim uvjetom vrijedi

$$E(\vec{w} | D) = -C_n P(\vec{y} | X)$$

(D) Označa y primjera (\vec{x}, y) je JF sa $\mu = \vec{w}^T \vec{x}$

A - nije jer nije Poissonova razdoblja

4 - Linearna regresija II

1) $\vec{w} = (X^T X + \lambda I)^{-1} X^T \vec{y}$
Efekt reg. na Gramovu matricu?

(A) Dodavanje λ na dijag. Gramove matrice povećava rang (smanjuje multikolin.)

2) Na kojim se činjenici temelji korist. norme kao reg. izraza?

(B) Ako je model prenaučen \Rightarrow hipoteza će imati veliku magnitudu težine

3) Što je \oplus , a što \ominus L1-reg?

D - nije, te je L2-konda
A - \oplus nije gradj. spust
C - krivo

(B) \oplus : izbacuje značajke iz modela
 \ominus : nema minim. u zatvorenoj formi

4) Kako je dif. L2-reg. pogreška kod lin. regresije?

(C) Uzorci nereg. pogreške i izraz prop. s kvadratom druge norme \vec{w} bez w_0

$$E_2(\vec{w} | D) = \frac{1}{2} \sum (y_i - h(\vec{x}_i))^2 + \frac{\lambda}{2} \|\vec{w}\|_2^2$$

5) Kolike parametre modela načini optimizacija L2-reg. pogreške?

(B) Parametri koji uz što manju magnitudu daju što manje očekiv. gubitke m slupku za učenje

6) $G = \Phi^T \Phi$
 $(m+1) \times (m+1)$

Koji je efekt površine multikolin. kod postupka OLS?

(A) $\text{rang}(\Phi) < m+1$

G ne može biti rang i ne može inverz, ali ima pseudoinverz koji nije numerički stabilan

5 - Linearni diskriminativni modeli

(1) Minimalne preučice u alg. Linearne regresije, a da dobijemo alg. koji je dobar klasifikator?

(C) promjeniti funkciju gubitka i optimizacijski postupak

→ ako promjenimo funkciju gubitka i dalje radi minimizaciju kvadratnog odstupanja!

(2) Zašto gubitak 0-1 ne možemo koristiti za optimizaciju?

(A) gradijent 0-1 gubitka svugdje je nula osim za $h(\vec{x}) = 0$
pa fija pogreške imaju zavrsni po težjima se gradij. spust ne može spustati

(3) Želimo preučiti alg. lin. reg. + log. reg. da bude dobar klasifik.

C - Linearna regresija

A - alg. nema smisla → minimum u obliku gubitka ne možemo dobiti u zadovoljavajućoj formi

B - gubitak još definiran ako je $h(\vec{x}) < 0$ ne može se izračunati gubitak

(D) model: $h(\vec{x}) = \vec{w}^T \vec{x}$

G: $L(y, h(\vec{x})) = (y - h(\vec{x}))^2$

O: \vec{w} putem grad. spusta

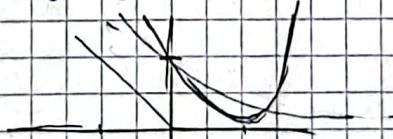
OVA: $\frac{N}{K}$ vs $\frac{N}{K} \cdot (K-1)$, K klasif.

(4) K klasa
 $\frac{N}{K}$ primj. / klasa

OVO: $\binom{K}{2}$ $\frac{N}{2}$ $\frac{N}{2}$

(B) OVO iziskuje $\frac{K-1}{2}$ puta više param. nego OVA, ali svaki OVA klasa ima $(K-1)$ puta manje \oplus nego \ominus

(5) Šta je specif. fil. gubitka perceptrona u odnosu na fil. gubitka LR-a i SVM-a?



(B) Gubitak za sve točno klasificirane primjere je 0, a za netočno klasif. može biti manji od 1

D-granica misli se na hiperplaninu!

(6) Po čemu se gubitak perceptronu razlikuje od gubitka zglobovnice?

(D) Gubitak zglobovnice kažnjava sve primjere koji se nalaze unutar marge čak i one koji su ispravno klasif.

6 - logistička regresija

1) Koji od uvjeta je dovoljan, ujet da granica Izrednog Šosa u ulaznom prostoru bude linearna?

(B) $\phi(\vec{x}) = (1, \vec{x})$

$$h(\vec{x}) = f(\vec{w}^\top \phi(\vec{x}))$$

↓
uvodimo ne-linearnost
u prostor
uznji

2) Na logi smo način modelirali distribuciju vjeroj. pojed primjera y ?

(B) $P(y|\vec{x}) = h(\vec{x})^y (1 - h(\vec{x}))^{1-y}$

3) Što nam osigurava činjstvo pretraživanje kod opt. Cog reg.?

(B) Postupak uvijek konvergira pod uvjetom da su primjeri linearno neodvojivi ili da regulariziramo s $\lambda > 0$

A - ako je linearne odvojivo ne konv.

C - nema lokalnih minimuma jer je $E(\vec{w}|D)$ konveksna!

D - neće konv. ako su lin. odvojivi

4) Zbog čega dolazi do premičnosti modela nereg. Cog. reg. na linearne odvojivim slučajima

(C) S porastom norme vektora težina gubitak na tčenim primjerima će biti null.

5) Što glob. konvergencija vrši u slučaju nereg. log. regresije na linearne neodvojivim problemu?

(D) Navigator o inicijalizaciji, opt. algoritam će pronaći parametre koji minimiziraju pogrešku na sljepu za vjeroj.

7 - Logistička regresija II

1)

Što su \oplus , a što \ominus gradijentnog spusta u odnosu na Newtonov postupak i što je razlika između log. reg?

B)

Gradijentni spust se može koristiti za online učenje no može krivudati i tako spreći konvergenciju od Newtonovog postupka

2)

$$\vec{w} = \vec{w} - H^{-1} \nabla E(\vec{w}) \Delta t$$

Kod log. reg., logi je nužan i dovoljan uvjet za izvedljivost i stabilitet Newtonovog optimizacijskog postupka?

C)

U podacima NE SMJE BITI MULTIKOLINEARNOST!

3)

$$\begin{aligned} F &= \text{epoha} \\ N &= \text{broj primjera} \\ m &= \text{broj inzagača} \end{aligned}$$

$$\nabla L = (h(\vec{x}) - y) \phi(\vec{x})$$

Vremenska računalna složnost algoritma LMS na linear. regresije (online mod)?

A)

$$O(ENm)$$

- po algoritmu

4)

Prednost MLR (softmax) u odnosu na BLR-ONO i BLR-ONA?
BLR = binarna log. reg.

B)

MLR i BLR-ONO imaju manje parametara od BLR-ONA, no jedino za MLR vrijedi $\sum_{\epsilon} P(y=\epsilon | \vec{x}) = 1$

5)

Zašto opt. kod log. reg. takoder ne provodimo izračunom pseudoinverza matrice dizajna?

C)

Maks. log-izglednosti oznaka log. reg. kao rješenje za parametre ne daje izraz u zatv. formi koji sadržava pseudoinverz matrice dizajna

6)

Koji je probab. princip ugrađen u optimizaciju alg. početnih linear. modela (lin., log. i MLR)?

C)

$$\text{Maksimizirati } \sum_{i=1}^n \ln P(y_i | \vec{x}_i), E[\ln P(y_i | \vec{x}_i)] = f(\vec{w}^T \vec{x})$$

7)

Poveznicu između LR i alg. NN sa sigmoid prijenosnim fjama?

D)

Model dvoslojnog NN istovjetan je modelu LR s preć. lin. modelima sa sigmoidalnim fjama kao baznim fjama.

8)

Što možemo reći o razlici izmedu novih i starih težina?
(LMS - pocet. lin. modeli)

$$\vec{w}_n = \vec{w}_s - \eta \nabla L$$

$$\Delta \vec{w} = -\eta \nabla L = -\eta (h(\vec{x}) - y) \phi(\vec{x})$$

(B) Razlika je to manja što je vektor $\phi(\vec{x})$ bliži ishodištu

9)

Veza izmedu LR i NN?

(B) Je logika kao adaptivne bazne funkcije koristi LR. \Leftrightarrow NN s sigmoid. aktv. fjom.

8 - SVM

1)

Zašto minimizirati $\frac{1}{2} \|\vec{w}\|^2$ daje maks. marginu?

$$d = \frac{h(\vec{x})}{\|\vec{w}\|} = \frac{\vec{w}^T \vec{x} + w_0}{\|\vec{w}\|}$$

(A) što je vektor \vec{w} kraći to je manja vrijednost $h(\vec{x})$ pa primjeri moraju biti što daje da bi vrijednost $h(\vec{x}) = \pm 1$, a to znači daje marginu $2d$.

2)

Kako glasi opt. problem tvrde marge u dualnoj formulaciji?

$$\underset{\alpha}{\operatorname{argmax}} \min_{\vec{w}, w_0} L(\vec{w}, w_0, \vec{\alpha})$$

3)

Razlika indukt. pristr. SVM-a i ind. pristranosti perceptronu?

(C) Razlikuju se po pristranosti preferencijom
- perceptron ne može minimizirati marginu (može naci kaštu hipotezu tko)

4)

Koje hipoteze zadovljava uvjet

$$\forall (\vec{x}^i, y^i) \in D \quad y^i h(\vec{x}^i) \geq 0$$

i kako odg. SVM odabere jednu od njih?

(C)

Uvjet zadovljava ∞ mnogo hipoteza
SVM odabire onu jednu koja minimizira kvadrat vektora težina
koja ispravno klasificira sve primjere u ujed. da
 $h(\vec{x}) \in \{-1, 1\}$

5) Kada će primjer \vec{x} u dualnoj formi SVM-a biti svrstan u \oplus klasu?

$$h(\vec{x}) = \sum_i y_i \vec{x}^T \vec{x} + w_0$$

B) ako je vektor \vec{x} po skup. umnošku sličniji pop. vektorima s \oplus oznakom nego potpornim vektorima s \ominus oznakom

6) Što je nužan i dovoljan uvjet da klasif. problem bude rješiv SVM-om s tvrdom marginom?

D) Konveksne čijusc 2 klase ne smiju se preklapati
(trebaju biti disjunktni)

9 - SVM II

1) n = broj znacičajki
 N = broj primjera

Koliko primarni opt. problemima ograničenja, a koliko varijabli po kojima optimiramo?
mala marge

D) Ograničenja: $y_i (\vec{w}^T \vec{x}_i + w_0) \geq 1 - \varepsilon_i$ } $2N$ ogranič.
optimiramo po: $\vec{w}, w_0, \varepsilon \rightarrow N+n+1$

2) $\sum_i (y_i (\vec{w}^T \vec{x}_i + w_0) - 1 + \varepsilon_i) = 0$
komple. rješavost

Što možemo zaključiti na temelju ovog uvjeta?

D) Da se potporni vektori ne nalaze izvan marge na pravoj strani granice

3) Koliko se opreća smanjenju vrij. fje gubitka i smanjenje složenosti modela manifestira kod opt. problema mala marge SVM-a?

$$\underset{\vec{w}, w_0, \varepsilon}{\text{arg min}} \left\{ \frac{1}{2} \|\vec{w}\|^2 + C \sum \varepsilon_i \right\}$$

A) veći $\|\vec{w}\|^2 \rightarrow$ veća marga, manje primjere u margini
 \rightarrow manji izbor $\sum \varepsilon_i$

4) Ako matrica dizajna ima više redaka nego stupaca, koja formulacija ima najmanje opt. varijable?

$$N > n+1$$

prim mala: $\vec{w}, w_0, \varepsilon : N+n+1$

dual mala: $\vec{d} : N$

prim tvrdi: $\vec{w}, w_0 : n+1$

dual tvrdi: $\vec{d} : N$

\Rightarrow B) prim. prob. tvrdi marga

5) Kada nije potrebno skalirati značajke i zašto?

A) Kada se koristi RBF jezgra s Mahalanobisom udaljenosti jer ta udaljenost uima u obliku varijance značajki

10 - Jezgrene metode

1) Je li moguće izračunati udalj. \vec{x} od hiperravn. SVM s nekom jezg. fjom

C) Da, ali nismo koristili Gauss. jezgru ili neku slož. jezgru nego koristili Gaussovnu jezgru kao grad. blok

2) Zašto je dobro da je jezgrena fja Mercerova jezgra?

A) Zato što takva jezgra odg. skalarnom produktu u nekom prostoru značajki \rightarrow nužno za jezreni trik

3) Kakav je utjecaj parametra γ na vrij. Gauss. jezgre $K(\vec{x}_1, \vec{x}_2)$ gdje $\vec{x}_1 \neq \vec{x}_2$ te na relinearnost modela jezg. strucja?

$$K(\vec{x}_1, \vec{x}_2) = \exp(-\gamma \|\vec{x}_1 - \vec{x}_2\|^2) = \exp(-\frac{1}{2\sigma^2} \|\vec{x}_1 - \vec{x}_2\|^2)$$

D) veći $\gamma \Rightarrow$ manji $K(\vec{x}_1, \vec{x}_2)$, veća relinearnost modela

4) Što znači korist. Gauss. jezgre za jezreni trik u mat. smislu?

B) $\forall \vec{x}_1 \forall \vec{x}_2 \quad \Phi(\vec{x}_1)^T \Phi(\vec{x}_2) = \exp(-\gamma \Delta^2)$
 $\Delta = (\vec{x}_1 - \vec{x}_2)^T (\vec{x}_1 - \vec{x}_2)$

5) Koja je prednost jezg. trika u slučaju kada primjere nije moguće prikazati kao vektore realnih brojeva?

D) Jezg. fja može biti mjeru sličnosti između nekategoriziranih primjera, što implicitno inducira vektorski prostor značajki

6) Po čemu je SVM specifičan u odnosu na općeniti alg. rijetkog jezg. strucja?

D) Prototipni primjeri odabiru se u obliku optimizacijskog postupka

7) Što znači da Mercer. jezgre implicitno definiraju prostor značajki?

D) Vrijednost jezrenih fja nad parom vektora jednaka je skalarnom produktu tih vektora način prek. u prostor značajki.

11 Neparametarske metode

1) Vaša rjeđost modela ovisi o hiperparam. C ?
(SVM)

$$C = \frac{1}{\lambda} \quad C \downarrow \rightarrow \lambda \uparrow \rightarrow \text{rjeđi model (parametarski)}$$

C) Što je C manji to je neparametarski model rjeđi
Također je rjeđi i parametarski model jer λ raste

2) Vašo se problem prečekstva dim. u visokodim prostorima manifestira
Pod algoritma L-M?

C) Svi primjeri međusobno vrlo udaljeni i gube se razlike
u udaljenosti

3) Što je karakter. hiperparam. alg. SU ?

B) Broj parametara ovisi o broju primjera

4) Na koji način funkcioniра alg. stabla Lepti ?

C) Koristi brzo pretraživu binarnu strukturu za partidioniranje
prostora primjera u prečekajuće regije.

Teorijska pitanja

V14 Progjena parametara I

1. $L(\vec{\theta} | D) = P(D|\vec{\theta}) \stackrel{def}{=} \prod_{i=1}^N P(\vec{x}^i | \vec{\theta})$

(C) Fja izglednosti $L(\vec{\theta} | D)$ jednaka je gustoći vjeroj. $P(D|\vec{\theta})$ samo što je izglednost fja parametra $\vec{\theta}$, dok je $P(D|\vec{\theta})$ fja uzorka D

2. $C_n L(\vec{\theta} | D) = C_n P(D|\vec{\theta}) \stackrel{def}{=} C_n \prod_{i=1}^N P(\vec{x}^i | \vec{\theta}) = \prod_{i=1}^N C_n p(\vec{x}^i | \vec{\theta})$

(C) Fja log-izgl. = fja koja parametrima $\vec{\theta}$ pridjeljuje vjerojatnost uzorka D uz pretpostavku da se uzorak pokrava dist. def. modelom $p(\vec{x}, \vec{\theta})$

3. Za koji od sljedećih parametara distribucije je progjena MLE pristrana?

(D) - kovarijacijska matrica Gaussove distribucije

Napomena: Varijanca Bernoullijeva i Multinuličeva dist. nije parametar distribucije.

$$\hat{\vec{\theta}}_{MLE} = \underset{\vec{\theta}}{\operatorname{argmax}} p(\vec{\theta} | D) p(\vec{\theta})$$

$p(\vec{\theta})$ = stand. teorijska dist. i da je konjug. dist. za $L(\vec{\theta} | D)$.
Što to znači?

(D) To znači da će umnožak izglednosti i apriorne distribucije dati distribuciju koja je iste vrste kao i apriorna dist.
=> ovo je niz o distribuciji iz ekspon. familije njen mod (maksimizator) postoji u zatv. formi (izračun analitički)

5. MAP progjentici računamo heurističkim metodom.
Što se dogodilo ako za apriornu distribuciju parametara upotrijebimo fju koja NIJE konjugatna fja izglednosti?

(A) aposteriornu dist. $p(\vec{\theta} | D)$ ne možemo izvesti u zatvorenoj formi, ali MAP možemo izračunati heurističkim optum.

6 Kod te MLE i MAP dati jednake progene?

(B) Kada broj primjera N teži u beskonačno.

7 Procjenjujemo parametar μ Bernullijevе distribucije.
Mali D_{train}

MLE i MAP procjena (Laplace, $d = p = 2$, $\hat{\mu}_{MAP} = \frac{d+m-1}{d+p+N-2}$)

$$\hat{\mu}_{MLE} = \frac{m}{N} \quad \hat{\mu}_{MAP} = \frac{m+1}{N+2}$$

Što od sljedećeg općenito vrijedi?

(B) MLE procjenitelj registriran i očekujemo da će model dobro generalizirati

8. Što i kako maksimira procjenitelj MAP?

(A) maksimira zajedničku gustoću vjerojatnosti parametara i podatka
u zatvorenoj formi, ali je apriorna dist. konj. za izglednost, inače iterativna

V15 Bayesov klasifikator

1. Zašto su praksi discriminativni modeli veće klasif. točnosti nego generativni modeli?

(B) Discriminativni modeli s manje parametara mogu modelirati istu granicu između klasa kao i generativni modeli, pa trebaju manje primjera da ih se nauči i teži ih je prenaučiti

2. Nedost generativnih u odnosu na discrimin. modela jest potreba složnost modeliranja. Što to znači?

(D) Za klas. nam je potrebna samo dist. $p(y|\vec{x})$, i ona se može modelirati sa manje param. od ravn. dist. $p(\vec{x}|y)$

(3) Bayesov klas. $\hat{y}_j(\vec{x}; \vec{\theta}) = P(y=j|\vec{x}) = \frac{P(\vec{x}|y=j)P(y=j)}{\sum_j P(\vec{x}|y=j)P(y=j)}$

• diskretne značajke (mogu imati više od 2 vrijednosti)

Teorijske dist. za $P(y)$ i $P(\vec{x}|y)$?

(A) Bernoulli za $P(y)$
multinulli za $P(\vec{x}|y)$

4. $K > 2$
 $\vec{x} \in \mathbb{R}^n$

$$p_j(\vec{x}, \vec{\theta}) = \frac{p(\vec{x}|y=j) P(y=j)}{\sum_j p(\vec{x}|y=j) P(y=j)}$$

Koje teorijske distribucije ćemo koristiti za $P(y)$ i $P(\vec{x}|y)$?

- (B) $P(y) = \text{kategorička}$.
 $P(\vec{x}|y) = \text{Gaussova}$

5. $h(\vec{x}; \vec{\theta}) = \operatorname{argmax}_y p(\vec{x}|y) P(y)$

Po čemu se vidi da je ovo generativan, a ne diskriminativan model?

- (C) modelira vjeroj. primjera i oznaka, budući da je na temelju pravila umnoška, umnožak $p(\vec{x}|y) P(y)$ jednak za svaki vjerjet.

6. Uz ře minimalne pretpostavke, koliko broj parametara za Σ biti linearan u broju značajki n ?

$$K \cdot n(n+1)$$

puno

$$K \cdot n$$

dijagonalna

$$K$$

izotropna

A Šum je isti za sve klase

B Značajke nisu linearno linijski ne ovise o klasi

C Šum je isti za sve klase i svi značajke

(D) Nema linearne zavisnosti između značajki
 $\Leftrightarrow G_i(x_i, x_j) = 0 \Rightarrow \text{diag. } \Sigma$

V16 Bayesov klasifikator III

1 GBC (s dij. Σ) i log. regresija su genera - diskrim. par

\Rightarrow (A) Aposterior vjerojatnost klase za GBC može se izraziti kao
početni linearni model sa sigmoidnom aktivacijom fjom

2 GBC i LR diskri - gener. par

\Rightarrow (C) Oba modela daju lin. granice između klasa, ali GBC
početno ima više parametra od LR

3 polunaivani Bayesov klasifikator (SNBC) vs Naivni Bayesov klasif (NBC)

⇒ (D) SNBC ima više parametara od NBC te modelira zavisnost između značajki

4 $D_{KL}(P(x,y) \parallel P(x)P(y)) = \sum_{x,y} P(x,y) \ln \frac{P(x,y)}{P(x)P(y)}$

⇒ (C) Što su varijable više ZAVISNE, to je veća KL divergencija između zajedničke vjerojatnosti i faktoriizirane vjerojatnosti.

VII PGM

1 Bayesova mreža generativni i parametarski model. Šta?

(D) Generativni = opisuje postupak kolim se mogu generisati podaci
koji se pojavljaju u određeni razdoblju vjerojatnosti
Parametarski = Bayesove mreže uopšte definiraju vjerojatnost
čvorova prema teorijske distribucije koja je opisana svojim parametrima

2 Koja je vez između vjerojatnosti rezavisnosti varijabli u Bayesovoj mreži i oprostosti od prenosičnosti?

(B) uvođenje pretpostavki o vjerojatnosti nezavisnosti predstavlja strukturu Bayesove mreže u smanjen broju parametara, čime se smanjuje mogućnost prenosičnosti

3 $x_1 \perp \{x_2, x_3\} \mid x_4$

Koji je efekt uvođenja ove pretpostavke na graf Bayesove mreže?

⇒ (C) Uklanjanje 2 bridala

4 HMM - Na što se odnosi pridjev "skriveni" u nazivu tog modela?

⇒ (B) Neke varijable modela nisu opažene u podacima, ali zavise o opaženim podacima

5 Gdje u Bayesovoj mreži nastupa efekt objašnjavanja i kako se manifestira?

(A) struktura $x \rightarrow z \leftarrow y$ gdje su x i y nezavisni, ali postaju zavisni ako je opažena posledica z

V18 - PGM II

1.

Na koji način izračunavamo aposteriorni upit?

$$P(\vec{x}_g | \vec{x}_o) = \frac{\sum_{\vec{x}_n} P(\vec{x}_g, \vec{x}_o, \vec{x}_n)}{\sum_{\vec{x}_g, \vec{x}_n} P(\vec{x}_g, \vec{x}_o, \vec{x}_n)}$$

(A)

omjer zajed. vjeroj. margin po varijablama smetnje i zajed. vjeroj. margin po varij. smetnje i upita

2.

Kako je definiran rezultat MAP-upita?

(D)

$$\vec{x}_g^* = \underset{\vec{x}_g}{\operatorname{argmax}} \sum_{\vec{x}_n} P(\vec{x}_g, \vec{x}_o, \vec{x}_n)$$

3.

Točno radimo učvršćivanje iz Bayes. mreže da bismo procjenili parametre dist. Kod te param. već imamo?

(C)

Uzrokovavanje slijestima da bismo procjenili parametre bilo koje distribucije, a ne samo vjetnih distribucija. U čvorovima mreže

4.

Koji je nedostatak metode uzrokovanja s odbacivanjem?

(A)

- ako je $P(\vec{x}_o)$ mala \Rightarrow treba generirati mrežu veličine da bi uzrokovala bio veliki i projekta posudara

5.

Koji problem kod primjene unaprijednog uzrokovanja za izračun aposteriornih upita kod Bayes. mrežom?

(B)

dobiveni vektori bit će uzrokovani iz apriorne, a ne aposteriorne dist.

6.

Što je \oplus procjene parametra kod potpunih podataka u odnosu na nepotpune podatke?

(D)

Kod potpunih podataka $\ln f$ dekomponira po strukturi mreže po parametre svake vjetne dist. možemo procjeniti nezavisno od drugih čvorova u zatvorenoj formi
 \rightarrow parametara može biti više nego kod mreža sa skriv. varijablama

7.

Zašto i kako Bayesiju mrežu učimo kod nepotpunim podatcima?

(C)

Jer mreža ima skrivene varijable, koje ne opazimo pa moramo koristiti iterativne metode za MAP ili MLE

V19 - Grupiranje

1

Zbog čega za problem minimizacije fje J ne postoji rješenje u zadatku?

- (C) Jer bi ovisi o vektorima μ_k , a vektori μ_k ovisi o vrijednostima b_k^t

2

Je li točno da algoritam k -sredina uvek konvergira?

- (A) Da, alg. uvek konvergira zato što je broj particija N primjera u K skupova ograničen, a opt. postupak definiran je tako da se J u svakoj iteraciji smanjuje

3

Koja je glavna ideja odabira početnih središta grupa kod algoritma K-means++?

- (D) najverovatnije središte grupe jest primjer koji je najviše udaljen od njemu najbližeg središta.

4

Zašto je algoritam k -medoida računalno složeniji od alg. k -sredina?

- (B) k -medoida ne koristi centroide nego medioide na kraju svake iteracije mora kada provjeram po primjerima pronaći medoide koji minimum funkcije J

5

Primjeri kada vektori želimo grupirati po NE-EUKLIDSKOJ udalj. Koji alg. bi koristili i zašto?

- (A) Alg. k -medoida jer koristi fju alg. k -sredina koristi euklid. udalj.

6

Što bi se dogodilo da kada fju raz. uzemo euklid. udalj.?

- (D) Alg. k -med. primjene bi grupirao članove kao i alg. k -sredina, pogotovo ako u središtu grupa postoji primjeri

V20 Grupiranje II

1 $p(\vec{x}) = \sum_{e=1}^k \pi_e p(\vec{x} | \theta_e) \quad \text{MM}$

Što modelira ujetra vjerci, $p(\vec{x} | \theta_e)$?

- A). gustoću vjerojatnosti primjera \vec{x} unutar \vec{z}_i

2 Koju distribuciju pretpostavljamo za latentnu varijablu \vec{z}_i i zašto?

- B). Kategoričku distribuciju koja opisuje kojoj grupi primjer \vec{x}_i zapravo pripada

3 Na što se odnosi pojam „očekivanje“ u razini až. maksimizacije očekivanja?

- C). Vjerojatnost skupa podataka pod modelom s fiksiranim parametrima izračunata na temelju vjerojatnosti pripadanja primjera svakoj grupi

4 Koja je razlika između potpune i nepotpune Ent. ($\vec{\theta} | D$) i zašto maksimizir. očekiv. potpuna log-izgled umjesto izračun. log-izgled?

- D). Potpuna log-izglednost je log-izglednost modela GMM s latentnim varijablem koja definiraju koji primjer pripada kojoj grupi, međutim kako to zapravo rezam, moram racurati s očekivanima tih varijabli

Pod logom ujetna EM-až. konvergira li samo?

- A). Krećući od nekih početnih parametara až. uvijek konvergira do param. koji maksim. očekiv. Ent.
- To ne moraju biti parametri koji maks. vjerojatnost podataka

5 Uz koje ujete až. GMM degenerira u až. k-sredine?

- C). Kovarij. matrica komponenti Gaussove mješavine je dijagonalna i izotropna, a odgovornosti su raspoređene na cijeli broj

7. Uz koje parametra modela

- GMM (dijeljena kov. matrica)
- K-medoidsa

očekujemo dobiti najsličnije rezultate grupiranja?

(C) puna Σ i $T_k = \frac{1}{K} \Rightarrow$ GMM
 $KM =$ Mahalanobisova udaljenost

8. Na kojem se principu temelji odabir broja grupa Akaikeovim kriterijem?

$$K^* = \underset{K}{\operatorname{argmin}} (-2\ln L(K) + 2g(K))$$

(C) model s optimalnim brojem grupa je onaj koji podatke čini najvjerojatnije, ali to čini sa što manje parametara

9. Štašto parametre GMM-a ne optimiramo u zadržanoj formi, dok kod Bayesovog klasi. to radimo?

(C) Kad GMM-ju ne znajući koji primjer pripada kojoj grupi, na je gustoća primjera jednaka, zbog toga da su komponenti za što ne postoji maksimum u zadržanoj formi

10. Štašto alg. K-medoidsa i alg. HAC imaju zajedničko, a po čemu se razlikuju?

(D) Oba alg. mogu grupirati primjere koji nisu vektORIZIRANI, no HAC daje hiperbolističku, dok KM daje particijalno grupiranje

V21 Vrednovanje modela

1. Štašto se za mjeru F_1 koristi harmonijska, a ne aritmetička sredina?

(C) jer P i Q nisu definirani na istoj skali, ali njihove recipročne vrijednosti jesu.

2. U kojem slučaju F_1 mjeru ne bi bila priladna mjeru za vrednovanje klasič. jer bi bila previše optimistična?

(A) ako je većina primjera pozitivna i klasič. sve primjere klasičira pozitivno

3.

Voju je očekivani odnos između F_1^M i F_1^N i zašto?

(A)

$$F_1^M < F_1^N$$

- pod mjeru F_1 računom procjče F_1 mjeri kroz sve klase, a klasifikator na manjim klasama više greješi

4.

Što je nedostatak metode LOOCV?

(B)

Varijanca procjene je visoka jer klasifik. dijele $\frac{N-2}{N}$ primjera za učenje

5.

Koja je motivacija za koristiti ugniježđene višestruke ulakrsne projekcije, umjesto obične ulakrsne projekcije?

(A)

omogućava nam da procijenimo prediktivnu moć modela, optimalne skripenosti te maksimalnu iskoristimmo raspoloživih podataka za učenje i ispitivanje

6.

Što je prednost ugniježđene k-struke ulakrsne projekcije u odnosu na običnu ulakrsnu projekciju?

(D)

procjenjujem očekivatu ispitnu pogrešku modela optimalne skripenosti

2. Osnovni koncepti

Strojno učenje 1, UNIZG FER, ak. god. 2023./2024.

Jan Šnajder, vježbe, v3.1

1 Zadatci za učenje

1. [Svrha: Na stvarim problemima razlikovati klasifikaciju od regresije.] Objasnite razliku između klasifikacije i regresije. Koji je od ta dva pristupa prikladan za: (a) filtriranje neželjene e-pošte (*spam*), (b) predviđanje kretanja dionica, (c) rangiranje rezultata tražilice? Kako biste u ovim slučajevima definirali ciljne oznake y ?

2. [Svrha: Razumjeti što je hipoteza, što je model i koja je veza između njih.]

(a) Dopunite praznine:

Hipoteza je funkcija koja preslikava _____ u _____, definirana do na _____. Model je _____ hipoteza, koje su indeksirane _____. Tako parametrizirani skup hipoteza također možemo prikazati kao prostor _____, a dimenzija tog prostora jednaka je _____. Učenje modela odgovara pretraživanju _____ u potrazi za _____ hipotezom. To je ona hipoteza koja _____ klasificira označene primjere, što procjenjujemo pomoću _____ mjerene na _____. Drugim riječima, učenje modela svodi se na _____ parametara modela s _____ kao kriterijskom funkcijom.

(b) Rješavamo problem binarne klasifikacije u prostoru primjera $\mathcal{X} = \{0, 1\}^2$. Definirajte linearan model koji će primjere odvajati pravcem.

(c) Koja je dimenzija prostora parametra? Koliko različitih hipoteza postoji u \mathcal{H} ?

(d) Neka je skup označenih primjera sljedeći:

$$\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\} = \{((0, 0), 0), ((1, 1), 0), ((1, 0), 1), ((0, 1), 1)\}.$$

Odredite konkretnu hipotezu $h \in \mathcal{H}$ koja ima najmanju empirijsku pogrešku.

3. [Svrha: Shvatiti što je to induktivna pristranost i kako ona određuje klasifikaciju neviđenih primjera.] Pročitajte poglavlje 2.3 u skripti (tu temu nismo obradili na predavanju).

(a) Definirajte induktivnu pristranost (neformalno i formalno). Koje su dvije vrste pristranosti koje sačinjavaju induktivnu pristranost?

(b) Raspolažemo skupom označenih primjera u ulaznom prostoru $\mathcal{X} = \{0, 1\}^3$:

$$\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\} = \{((0, 0, 0), 0), ((1, 0, 0), 1), ((1, 0, 1), 1), ((0, 1, 0), 1), ((0, 1, 1), 1)\}.$$

Koja je klasifikacija neviđenih primjera?

(c) Definirajte linearan model \mathcal{H} za $\mathcal{X} = \{0, 1\}^3$. Koja je to vrsta pristranosti?

(d) Možete li odrediti klasifikaciju neviđenih primjera uz odabrani model \mathcal{H} ? Je li pristranost koja proizlazi iz odabira modela dovoljna za jednoznačnu klasifikaciju svih primjera iz \mathcal{X} ?

(e) Definirajte (neformalno) neku dodatnu pristranost takvu da klasifikacija svakog primjera slijedi jednoznačno na temelju skupa primjera \mathcal{D} . Koje je vrste ta dodatna pristranost?

4. [Svrha: Znati nabrojati osnovne komponente algoritma strojnog učenja i povezati ih s induktivnom pristranošću.]

(a) Nabrojite tri osnovne komponente algoritma strojnog učenja.

- (b) Identificirajte uz koje se komponente veže koja vrsta induktivne pristranosti.
5. [Svrha: Razumjeti vezu između funkcije gubitka i empirijske pogreške te mogućnost njihove prilagodbe konkretnom problemu.]
- Pogreška hipoteze je očekivanje funkcije gubitka L . Nad kojom distribucijom je definirano to očekivanje? Koji je problem s takvom definicijom u praksi?
 - Definirajte empirijsku pogrešku preko funkcije gubitka L . Koja je pretpostavka implicitno ugrađena u tu definiciju?
 - Kod asimetričnih gubitaka funkciju L možemo definirati preko matrice gubitka (v. skriptu: poglavlje 2.7 i primjer 2.6). Definirajte takvu matricu za problem klasifikacije neželjene e-pošte te izračunajte funkciju pogreške za slučaj pet pogrešno negativnih i dvije pogrešno pozitivne klasifikacije od ukupno deset ($N = 10$) primjera.
6. [Svrha: Razviti ispravnu intuiciju za odabir modela temeljem unakrsne provjere.]
- Skicirajte krivulje pogreške učenje i ispitne pogreške u ovisnosti o složenosti modela. Naznačite područje prenaučenosti i podnaučenosti.
 - Objasnite zašto pogreška učenja s povećanjem složenosti modela teži k nuli.
 - Raspolažemo modelom \mathcal{H}_α koji ima hiperparametar α kojim se može ugađati složenost modela. Za odabrani α naučili smo hipotezu koja minimizira empirijsku pogrešku. Unakrsnom provjerom utvrdili smo da je ispitna pogreška znatno veća od pogreške učenja. Je li naš odabir hiperparametra α suboptimalan?
 - Raspolažemo modelom \mathcal{H}_α s hiperparametrom α (veći α daje složeniji model). Raspolažemo dvama optimizacijskim algoritmima: L_1 i L_2 . Algoritam L_2 lošiji je od algoritma L_1 , u smislu da L_2 pronalazi parametre $\boldsymbol{\theta}_2$ koji su lošiji od parametara $\boldsymbol{\theta}_1$ koje pronalazi L_1 , tj. $E(\boldsymbol{\theta}_2|\mathcal{D}) > E(\boldsymbol{\theta}_1|\mathcal{D})$. Neka α_1^* označava optimalnu vrijednost hiperparametra za \mathcal{H}_α učenog algoritmom L_1 , a α_2^* optimalnu vrijednost za \mathcal{H}_α učenog algoritmom L_2 . Načinite skicu analognu onoj iz zadatka (a) i naznačite vrijednosti pogrešaka za modele $\mathcal{H}_{\alpha_1^*}$ i $\mathcal{H}_{\alpha_2^*}$.
 - Može li model učen lošijim algoritmom L_2 imati manju ispitnu pogrešku od modela koji je učen boljim algoritmom L_1 , ali nije optimalan? Skicirajte takvu situaciju na prethodnoj skici.

2 Zadaci s ispita

1. (P) U ulaznom prostoru $\mathcal{X} = \{0, 1\}^3$ definiramo sljedeći klasifikacijski model:

$$h(\mathbf{x}; \boldsymbol{\theta}) = \mathbf{1}\{\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 \geq 0\}$$

Koja je dimenzija prostora parametara te koliko različitih hipoteza postoji u ovom modelu?

- A Dimenzija prostora parametara je 4, a hipoteza ima beskonačno mnogo
- B Dimenzija prostora parametara je 4, a hipoteza ima manje od 256
- C Dimenzija prostora parametara i broj hipoteza su beskonačni
- D Dimenzija prostora parametara je 256, a hipoteza ima 14

2. (P) Za ulazni prostor $\mathcal{X} = \{0, 1\}^3$ definiramo klasifikacijski model \mathcal{H} kao skup parametriziranih funkcija definiranih na sljedeći način:

$$h(\mathbf{x}; \boldsymbol{\theta}) = \mathbf{1}\{(\theta_{1,1} \leq x_1 \leq \theta_{1,2}) \wedge (\theta_{2,1} \leq x_2 \leq \theta_{2,2}) \wedge (\theta_{3,1} \leq x_3 \leq \theta_{3,2})\}$$

Parametri su trodimenzijski vektori realnih brojeva, tj. prostor parametara definiran je kao $\boldsymbol{\theta} \in \mathbb{R}^6$. Koliko iznosi $|\mathcal{H}|$?

- A 42 B ∞ C 56 D 28

3. (P) Skup označenih primjera u dvodimenzijском ulaznom prostoru je:

$$\mathcal{D} = \{((0,0),0), ((0,1),0), ((1,1),1)\}$$

Koliko hipoteza ostvaruje empirijsku pogrešku jednaku nuli?

- A 16 B Pitanje nema smisla jer nije definiran model C Beskonačno mnogo D 14

4. (P) Za linearan klasifikator u $\mathcal{X} = \{0,1\}^3$ zadan je sljedeći skup primjera za učenje:

$$\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\} = \{((0,0,0),0), ((1,0,0),1), ((1,0,1),1), ((0,1,0),1), ((0,1,1),1), ((1,1,0),0)\}$$

Razmatramo dva modela:

$$\begin{aligned}\mathcal{H}_a : h_a(\mathbf{x}|\boldsymbol{\theta}) &= \mathbf{1}\{\theta_0 + x_1\theta_1 + x_2\theta_2 + x_3\theta_3 \geq 0\} \\ \mathcal{H}_b : h_b(\mathbf{x}|\boldsymbol{\theta}) &= h_a(\mathbf{x}; \boldsymbol{\theta}_1) \cdot h_a(\mathbf{x}; \boldsymbol{\theta}_2)\end{aligned}$$

Uočite da svaka hipoteza iz modela \mathcal{H}_b kombinira dvije hipoteze iz modela \mathcal{H}_a (operacijom množenja). Neka:

$$\begin{aligned}h_a^* &= \operatorname{argmin}_{h \in \mathcal{H}_a} E(h|\mathcal{D}) \\ h_b^* &= \operatorname{argmin}_{h \in \mathcal{H}_b} E(h|\mathcal{D})\end{aligned}$$

Koja je od navedenih tvrdnji točna?

- A $E(h_a^*|\mathcal{D}) = E(h_b^*|\mathcal{D}) > 0$
 B $E(h_a^*|\mathcal{D}) > E(h_b^*|\mathcal{D}) = 0$
 C $0 < (E(h_a^*|\mathcal{D}) - E(h_b^*|\mathcal{D})) < 1$
 D $E(h_a^*|\mathcal{D}) = E(h_b^*|\mathcal{D}) = 0$

5. (P) Razmatramo klasifikacijski problem u ulaznom prostoru $\mathcal{X} = \{0,1\}^2$. Razmatramo sljedeće modele:

$$\begin{aligned}\mathcal{H}_1 : h_1(\mathbf{x}; \boldsymbol{\theta}) &= \mathbf{1}\{\boldsymbol{\theta}^T \mathbf{x} \geq 0\} & \mathcal{H}_3 : h_3(\mathbf{x}; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) &= h_1(\mathbf{x}; \boldsymbol{\theta}_1) \wedge h_2(\mathbf{x}; \boldsymbol{\theta}_2) \\ \mathcal{H}_2 : h_2(\mathbf{x}; \boldsymbol{\theta}) &= \mathbf{1}\{(x_1 - \theta_1)^2 + (x_2 - \theta_2)^2 \leq \theta_0^2\} & \mathcal{H}_4 &= \mathcal{H}_1 \cup \mathcal{H}_2\end{aligned}$$

Parametri svih modela realni su brojevi, $\boldsymbol{\theta} \in \mathbb{R}^3$. **Koji odnosi vrijede između ovih modela?**

- A $\mathcal{H}_1 = \mathcal{H}_2 \subset \mathcal{H}_3 = \mathcal{H}_4$
 B $\mathcal{H}_1 = \mathcal{H}_2 = \mathcal{H}_4 \subset \mathcal{H}_3$
 C $\mathcal{H}_1 \subset \mathcal{H}_2 \subset \mathcal{H}_3 \subset \mathcal{H}_4$
 D $\mathcal{H}_1 \subset \mathcal{H}_2 = \mathcal{H}_3 \subset \mathcal{H}_4$

6. (P) Za linearan model u $\mathcal{X} = \{0,1\}^3$ zadan je sljedeći skup primjera za učenje:

$$\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\} = \{((0,0,0),0), ((1,0,0),1), ((1,0,1),1), ((0,1,0),1), ((0,1,1),1)\}$$

Optimizacijski postupak klasifikatora funkcioniра tako da minimizira empirijsku pogrešku, definiranu kao očekivanje funkcije gubitka 0-1, i postupak u tome uvijek uspijeva. Želimo znati koju bi klasu ovaj klasifikator dodijelio primjeru $\mathbf{x} = (1,1,1)$. **Možemo li, na temelju iznesenih informacija, odrediti klasifikaciju dotičnog primjera i što nam to govori o induktivnoj pristranosti ovog algoritma?**

- A Ne možemo, jer nije definirana induktivna pristranost preferencijom, pa činjenica da je model linearan nije dovoljan skup pretpostavki da bismo jednoznačno odredili klasifikaciju svih novih primjera
 B Možemo, klasifikacija je $y = 1$, i ovaj klasifikator ima definiranu induktivnu pristranost pomoću koje može jednoznačno odrediti klasifikaciju svakog primjera
 C Možemo, klasifikacija je $y = 1$, premda dane informacije nisu dovoljne za definiciju induktivne pristranosti, pa za ovaj skup primjera više hipoteza savršeno točno klasificira primjere
 D Možemo, $y = 1$, jer klasifikator ima induktivnu pristranost jezikom (linearan model) i preferencijom (primjeri za koje je $h(x) \geq 0$ klasificiraju se pozitivno)

7. (P) Optimizacija parametara modela temelji se na funkciji gubitka $L : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}_0^+$, gdje je $L(y, h(\mathbf{x}))$ gubitak na primjeru (\mathbf{x}, y) . U većini primjena koristimo simetričan gubitak 0-1. Međutim, u nekim primjenama ima više smisla definirati asimetričan gubitak. Jedan takav primjer je zadatak detekcije karcinoma iz medicinskih slika. Taj zadatak možemo formalizirati kao problem binarne klasifikacije s označama $\mathcal{Y} = \{0, 1\}$, gdje $y = 1$ označava postojanje karcinoma, a $y = 0$ nepostojanje karcinoma. **Koje od sljedećih svojstava bi trebala zadovoljiti asimetrična funkcija gubitka za takav zadatak?**

- A $L(0, 1) = 1$ i $L(1, 0) = L(1, 1) = L(0, 0) = 0$
- B $L(0, 1) > L(1, 0)$ i $L(1, 1) = L(0, 0) > 0$
- C $L(1, 0) > L(0, 1)$ i $L(1, 1) = L(0, 0) = 0$
- D $L(0, 1) = L(1, 0) > 0$ i $L(1, 1) = L(0, 0) = 0$

8. (P) Zadan je sljedeći skup sa $N = 6$ označenih primjera iz \mathbb{R}^3 :

$$\begin{aligned}\mathcal{D} &= \{(\mathbf{x}^{(i)}, y^{(i)})\} \\ &= \{((0, 0, 0), 0), ((1, 1, 0), 0), ((1, 0, 0), 1), ((1, 0, 1), 1), ((0, 1, 0), 1), ((0, 1, 1), 1)\}\end{aligned}$$

Razmatramo linearan model i računamo empirijsku pogrešku $E(h|\mathcal{D})$ hipoteza iz tog modela definiranu kao očekivanje asimetričnog gubitka. Gubitak je definiran tako da lažno negativne primjere kažnjava sa 1, a lažno pozitivne primjere sa 0.5. **Koliko iznosi najmanja, a koliko najveća moguća vrijednost tako definirane empirijske pogreške $E(h|\mathcal{D})$?**

- A $0 \leq E(h|\mathcal{D}) \leq 1/4$
- B $1/4 \leq E(h|\mathcal{D}) \leq 2/3$
- C $\frac{1}{48} \leq E(h|\mathcal{D}) \leq 2/3$
- D $1/12 \leq E(h|\mathcal{D}) \leq 3/4$

9. (P) Razmatramo klasifikacijski problem u ulaznom prostoru $\mathcal{X} = \mathbb{Z}^2$. Skup označenih primjera je $\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_i = \{((0, 0), 0), ((0, 2), 0), ((0, -1), 0), ((-1, 0), 1), ((0, 1), 1), ((1, 0), 1)\}$. Razmatramo sljedeće modele, parametrizirane sa $\boldsymbol{\theta} \in \mathbb{R}^{n+1}$:

$$\begin{aligned}\mathcal{H}_1 : h_1(\mathbf{x}; \boldsymbol{\theta}) &= \mathbf{1}\{\boldsymbol{\theta}^T \mathbf{x} \geq 0\} \\ \mathcal{H}_2 : h_2(\mathbf{x}; \boldsymbol{\theta}) &= \mathbf{1}\{(x_1 - \theta_1)^2 + (x_2 - \theta_2)^2 \geq \theta_0^2\}\end{aligned}$$

Pored ova dva modela, razmatramo i njihove kombinacije, modele \mathcal{H}_3 i \mathcal{H}_4 . Neka je $\mathcal{H}_3 = \mathcal{H}_1 \cup \mathcal{H}_2$ te neka je \mathcal{H}_4 skup funkcija definiranih kao $h_4(\mathbf{x}; \boldsymbol{\theta}) = h_1(\mathbf{x}) \cdot h_2(\mathbf{x})$. Neka je E_k minimalna empirijska pogreška koja se modelom \mathcal{H}_k može ostvariti na skupu \mathcal{D} , tj. $E_k = \operatorname{argmin}_{h \in \mathcal{H}_k} E(h|\mathcal{D})$. **Koji odnosi vrijede između minimalnih empirijskih pogrešaka ovih modela?**

- A $E_1 > E_2 = E_3 > E_4$
- B $E_1 = E_2 > E_3 = E_4$
- C $E_1 > E_2 > E_3 = E_4$
- D $E_1 = E_2 = E_3 > E_4$

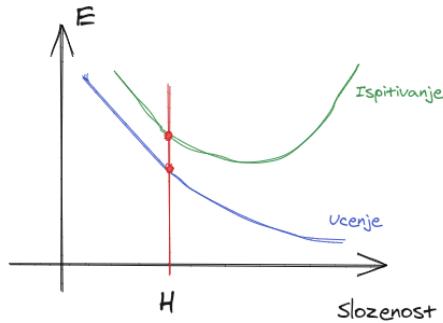
10. (P) Razmatramo klasifikacijski problem u ulaznom prostoru $\mathcal{X} = \mathbb{Z}^2$. Skup označenih primjera je $\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\} = \{((0, 0), 1), ((-1, -1), 0), ((1, 1), 0)\}$. Razmatramo sljedeće modele \mathcal{H} i funkcije preslikavanja $\phi : \mathbb{Z}^2 \rightarrow \mathbb{Z}^2$, kojom primjere iz \mathcal{D} preslikavamo u matricu dizajna Φ :

$$\begin{aligned}\mathcal{H}_1 : h_1(\mathbf{x}; \theta_0, \theta_1) &= \mathbf{1}\{\theta_1 x_1 + \theta_0 \geq 0\} & \phi_1(\mathbf{x}) &= (1, x_2, x_1) \\ \mathcal{H}_2 : h_2(\mathbf{x}; \theta_0, \theta_2) &= \mathbf{1}\{\theta_2 x_2 + \theta_0 \geq 0\} & \phi_2(\mathbf{x}) &= (1, x_1, x_1 x_2) \\ \mathcal{H}_3 : h_3(\mathbf{x}; \boldsymbol{\theta}) &= \mathbf{1}\{\boldsymbol{\theta}^T \mathbf{x} \geq 0\} \\ \mathcal{H}_4 : h_4(\mathbf{x}; \theta_0) &= \mathbf{1}\{x_1^2 + x_2^2 \geq \theta_0\}\end{aligned}$$

U svim modelima parametri su realni brojevi, $\theta_j \in \mathbb{R}$. Razmotrite sve kombinacije modela \mathcal{H} i funkcije preslikavanja ϕ . Za koju kombinaciju modela \mathcal{H} i funkcije preslikavanja ϕ postoji samo jedna hipoteza $h \in \mathcal{H}$ za koju $E(h|\mathcal{D}) = 0$?

- A $h_2 + \phi_2$ B $h_4 + \phi_1$ C $h_3 + \phi_2$ D $h_1 + \phi_1$

11. (P) Na slici ispod prikazan je graf funkcije pogreške učenje i pogreške ispitivanja za neku familiju modela i neki označeni skup primjera:



Crvenom linijom označena je složenost nekog modela \mathcal{H} . Crvene točke odgovaraju ispitnoj pogrešci i pogrešci učenja za hipotezu $h \in \mathcal{H}$ iz tog modela, dobivenoj nekim optimizacijskim algoritmom. Što možemo reći o modelu \mathcal{H} i o hipotezi h ?

- A Model \mathcal{H} nije optimalne složenosti, a čak ni hipoteza h ne mora biti optimalna na skupu za učenje, ako je optimizacijski algoritam loš
 B Model H je podnaučen, ali je barem hipoteza h hipoteza s najmanjom ispitnom pogreškom unutar takvog suboptimalnog modela
 C Model \mathcal{H} je nedovoljne složenosti, ali je barem hipoteza h optimalna u smislu najmanje moguće pogreške na skupu za učenje
 D Model \mathcal{H} je prenaučen, a hipoteza h će loše generalizirati na neviđene primjere
12. (P) Raspolažemo modelom \mathcal{H}_α , koji ima hiperparametar α kojim se može ugađati složenost modela. Isprobavamo dvije vrijednosti hiperparametra: α_1 i α_2 . Treniramo modele \mathcal{H}_{α_1} i \mathcal{H}_{α_2} te dobivamo hipoteze h_{α_1} i h_{α_2} . Zatim računamo empirijske pogreške tih hipoteza na skupu za učenje \mathcal{D}_u i na skupu za ispitivanje \mathcal{D}_i . Utvrđujemo da vrijedi:

$$E(h_{\alpha_1}|\mathcal{D}_i) - E(h_{\alpha_1}|\mathcal{D}_u) < E(h_{\alpha_2}|\mathcal{D}_i) - E(h_{\alpha_2}|\mathcal{D}_u)$$

Što iz toga možemo zaključiti?

- A Model \mathcal{H}_{α_2} je prenaučen
 B Optimalan model je onaj s vrijednošću hiperparametra iz intervala $[\alpha_1, \alpha_2]$
 C Model \mathcal{H}_{α_1} je podnaučen
 D Model \mathcal{H}_{α_1} je manje složenosti od modela \mathcal{H}_{α_2}

V02 - Osnovni koncepti

I Zadaci za učenje

1.1.

Klasifikacija

- na temelju ulaznog primjera model primjer svrstava u 1 od K klasa

Regressija

- na temelju ulaznog primjera model primjeru dodjeljuje numeričku vrijednost

Pr

filtracija neželjene e-pošte - klasifikacija

predviđanje kretanja dionica - regresija

rangiranje rezultata tržilice - klasifikacija

OZNAKA

da/nie

vrijednost dionice

rang

1.2.

a)

Hipoteza je fja koja preslikava skup primjera u skup rješenja, kje su indeksirane parametrima. Tako parametrizirani skup hipoteza također možemo prikazati kao prostor parametara, a dimenzija tog prostora jednaka je broju parametara. Učenje modela odgovara pretraživanju modela u potrazi za optimalnom hipotezom. To je ona hipoteza koja načinje klasificaciju označenih primjera, što procjenjujemo pomoću funkcije pogreške mjerene na skupu za učenje. Drugim riječima, učenje modela sudi se na optimizaciju parametara - modela s empirijskom pogreškom kao kriterijskim fjom.

b)

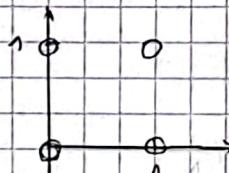
Binarna klasifikacija

$$\chi = \{0, 1\}^2$$

- model primjere odvaja pravcem

$$\vec{w} = [w_0, w_1, w_2]$$

$$w_0 + w_1 x_1 + w_2 x_2 = 0 \leftarrow \text{pravac}$$



MODEL

$$h(\vec{x}; \vec{w}) = 1 \{ w_0 + w_1 x_1 + w_2 x_2 \geq 0 \}$$

↳ Klasificiramo: izlaz modela mora biti 0 ili 1

c)

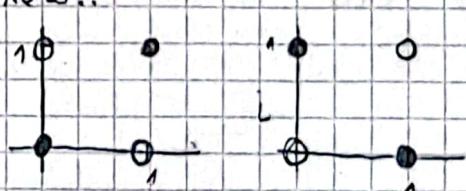
Dimenzija prostora parametara: 3 (w_0, w_1, w_2)

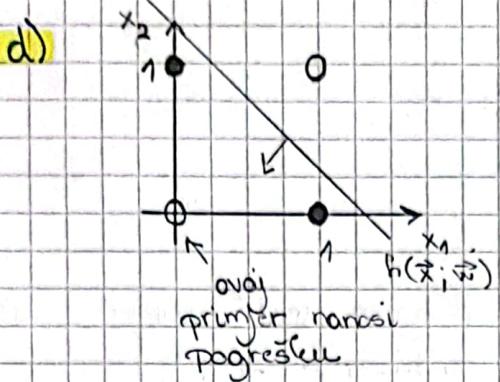
Broj različitih hipoteza: $16 - 2 = 14$,

x_1	x_2	h
0	0	0 ili 1
0	1	0 ili 1
1	0	0 ili 1
1	1	0 ili 1

→ 2 skupa NISU linearno odvojiva
 $X \cup Y$!!

$$2^4 = 16$$





$$E(\vec{w} | D) = \frac{1}{N} \sum_{i=1}^N \mathbb{1}\{y_i + h(\vec{x}_i; \vec{w}) \neq 0\}$$

$$E(\vec{w} | D) = \frac{1}{4}(0+0+0+1) = \frac{1}{4}$$

1.3

a) INDUKTIVNA PRISTRANOST

- neformalno: skup minimalnih pretpostavki pomču kojih označa svih primjera slijedi deduktivno

- formalno:

$$\forall \vec{x} \in X \quad \vec{A} \wedge \vec{B} \vdash h_+(\vec{x})$$

inductive bias

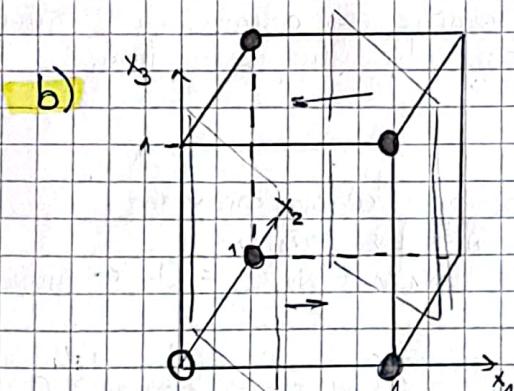
• 2 vrste induktivne pristranosti

• pristranost izuzmom (ograničenjem)

- ograničavamo se na određeni skup hipoteza

• pristranost preferencijom (preferencijom)

- proizlazi iz same implementacije algoritma
- iz više jednako dobroih hipoteza, algoritam pronalazi samo 1



Pitanje: Koja je klasifikacija nevidjenih primjera?

Besmisljeno pitanje
⇒ nema definiranog modela

c) Linearni model H za $X = \{0,1\}^3$

$$h(\vec{x}, \vec{w}) = \mathbb{1}\{w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3 \geq 0\}$$

→ ovo je pristranost jer je

d) Pristranost iz modela nije dovoljna za jednoznačnu klasifikaciju primjera

1.4

a) Osnovne komponente dgl strujnog učenja

- ① model
- ② funkcija pogreške
- ③ optimizacijski postupak

b)

MODEL - pristranost jeruka / ograničenja

FJA POGREŠKE

OPT. POSTUPAK } pristranost pretraživ. / prefer.

1.5

a)

pogreška hipoteze = očekivanje fie aubitka L

⇒ očekivanje definisano nad distribucijskim podatkom

⇒ problem je što u praksi ne znamo kojim se distribucijom podaci priboravaju

b)

empirijska pogreška

nad podacima

$$E(\vec{w} | D) = \frac{1}{N} \sum_{i=1}^N L(\vec{x}^i)$$

$$= \frac{1}{2} \sum_{i=1}^N (y^i - h(\vec{x}^i))^2$$

pretpostavka: podaci modelirani kao I.I.D. varijable

c)

5 - false negative

2 - false positive

Matrica tablina

stvarno

-	x	5	-
	2	y	

\sum - N

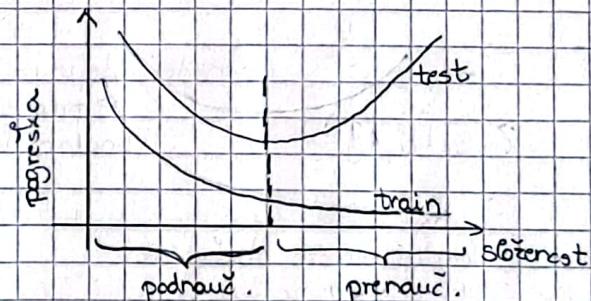
predviđanje

$$\Rightarrow E(\vec{w} | D) = \frac{5+2}{10} = \frac{7}{10}$$

$$x + y = 3$$

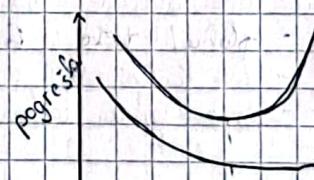
1.6

a)



b) Povećanjem složenosti model se prilagođava šumu u podacima te zbog toga smanjuje empirijsku pogrešku

c) H_k



$$E_{\text{test}} \gg E_{\text{train}}$$

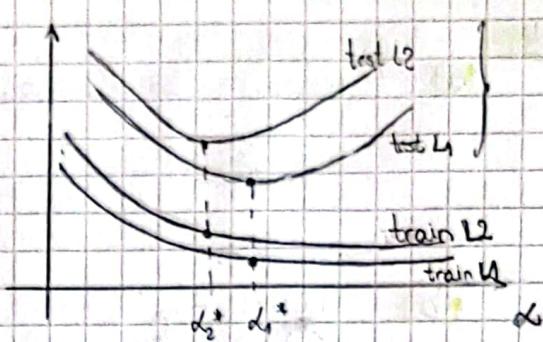
Je li odabir parametra k subopt.?
Ne znamo, premašo informacija

(znamo samo da je progrec velik)

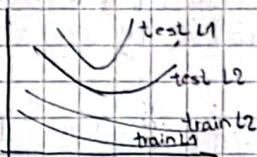
d) veci α - slozeniji model

$$E(\vec{\theta}_2 | D) > E(\vec{\theta}_1 | D)$$

L_2 posjeti model od L_1



e) Može li model učen L_2 imati manju ispitnu pogrešku od modela koji je učen L_1 ? Može jer test krivulje ovise o optimizacijskom postupku



II Zadaci s ispita

2.1

$$h(\vec{x}; \vec{\theta}) = 1 \{ \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 \geq 0 \}$$

Dimenzija prostora parametra: 4

x_1	x_2	x_3	y
0	0	0	0 ili 1
0	0	1	0 ili 1
:			:
1	1	1	0 ili 1

$2^8 = 256$ mogućih D , no neki nisu linearno odgovari prema tome hipoteza ima manje od 256

(B)

2.2. $h(x; \vec{\theta}) = 1 \{ (\theta_{1,1} \leq x_1 \leq \theta_{1,2}) \wedge (\theta_{2,1} \leq x_2 \leq \theta_{2,2}) \wedge (\theta_{3,1} \leq x_3 \leq \theta_{3,2}) \}$

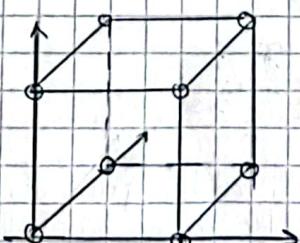
$$|H| = ?$$

$$\vec{\theta} = \begin{bmatrix} \theta_{11} & \theta_{12} \\ \theta_{21} & \theta_{22} \\ \theta_{31} & \theta_{32} \end{bmatrix}$$

$$\vec{x} \in \{0, 1\}^3$$

model definira kvadrat

\Rightarrow primjer ima označen 1 ako je unutar kvadra, 0 inače



• Kvadrat može obuhvaćati

- 1 točku: 8
- brid/2 točke: 12
- polugrid/4 točke: 6
- sve točke: 1
- niti jednu točku: 1

$$|H| = 28$$

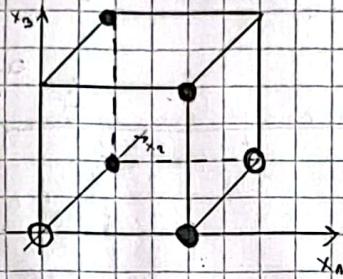
(D)

2.3.

(B)

model nije definiran pa rasprava o empirijskoj pogrešci nema smisla

2.4.



$$h_a(\vec{x} | \vec{\theta}) = 1 \{ \theta_0 + x_1 \theta_1 + x_2 \theta_2 + x_3 \theta_3 \geq 0 \}$$

$$h_b(\vec{x} | \vec{\theta}) = h_a(\vec{x}; \theta_1) \cdot h_a(\vec{x}; \theta_2)$$

Model h_a ne može postići $E(h_a^* | D) = 0$ jer skup nije linearno odvojiv

Model h_b može postići $E(h_b^* | D) = 0$ jer kombinira dve hiperplane pa područje između tih ravnina može preći kroz 0 pa područje klase 1

(B)

$$E(h_a^* | D) > E(h_b^* | D) = 0$$

2.5.

$$\mathcal{H}_1: h_1(\vec{x}; \vec{\theta}) = 1 \{ \vec{\theta}^\top \vec{x} \geq 0 \}$$

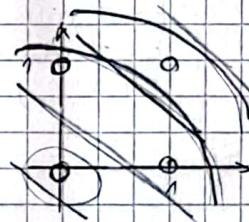
$$\mathcal{H}_2: h_2(\vec{x}; \vec{\theta}) = 1 \{ (x_1 - \theta_1)^2 + (x_2 - \theta_2)^2 \leq \theta_0^2 \}$$

$$\mathcal{H}_3: h_3(\vec{x}; \vec{\theta}_1, \vec{\theta}_2) = h_1(\vec{x}; \vec{\theta}_1) \wedge h_2(\vec{x}, \vec{\theta}_2)$$

$$\mathcal{H}_4 = \mathcal{H}_1 \cup \mathcal{H}_2$$

$$\vec{x} \in \{0, 1\}^2$$

Klasifikacija !!



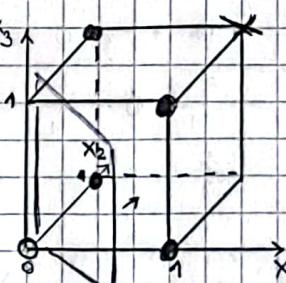
(B)

$$\mathcal{H}_4 = \mathcal{H}_1 = \mathcal{H}_2 \subset \mathcal{H}_3$$

2.6.

$$X = \{0, 1\}^3$$

(C)



fja gubitka 0-1

2.7.

$$L(y, h(\vec{x}))$$

bitnije je kažnjavati ložno negativne od ložno pozit.

$$L(1, 0) > L(0, 1)$$

$$L(1, 1) = L(0, 0) = 0$$

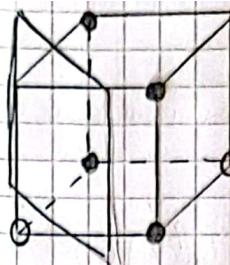
(C)

2.8.

$$FN = \frac{1}{2}$$

$$FP = \frac{1}{2}$$

min i max $E(\vec{w} | D)$



minimalna greška

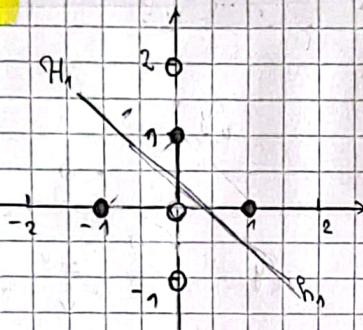
$$E(\vec{w} | D) = \frac{1}{6} (5 \cdot 0 + \frac{1}{2}) = \frac{1}{12}$$

• D

maximalna greška

$$E(\vec{w} | D) = \frac{1}{6} (0 + 4 \cdot 1 + \frac{1}{2}) \\ = \frac{1}{6} \cdot \frac{9}{2} = \frac{3}{4}$$

2.9.



• Linearno neodgovarajući step!

$$H_1: h_1(\vec{x}, \vec{\theta}) = 1 \{ \vec{\theta}^T \vec{x} \geq 0 \} \quad (\text{davnina})$$

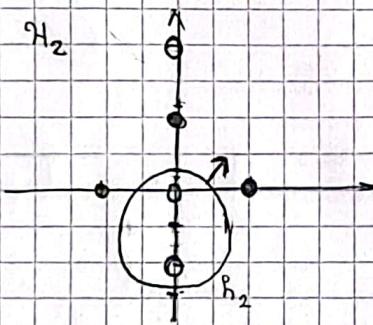
$$H_2: h_2(\vec{x}, \vec{\theta}) = 1 \{ (x_1 - \theta_1)^2 + (x_2 - \theta_2)^2 \geq \theta_0^2 \} \quad (\text{rančriga})$$

$$H_3: H_1 \cup H_2 \quad (111)$$

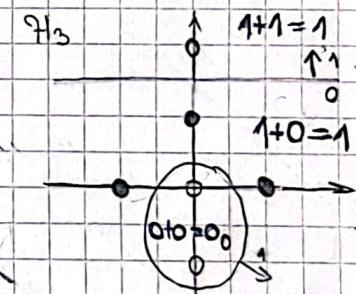
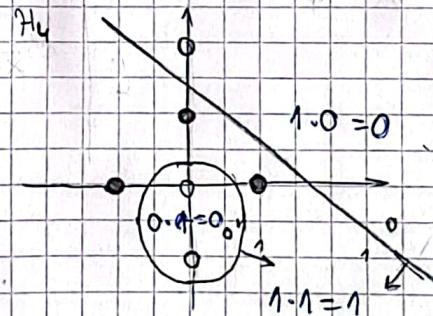
$$H_4: h_4(\vec{x}, \vec{\theta}) = h_1(\vec{x}) \cdot h_2(\vec{x}) \quad (I)$$

H_1 minimalno 2 kriva
 H_2 minimalno 1 kriva

} $E_1 > E_2$

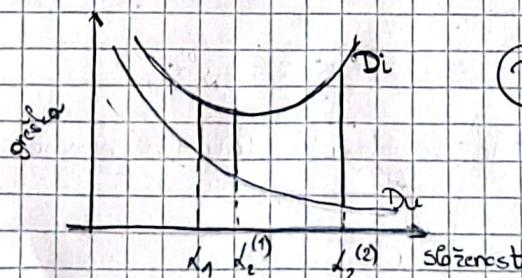


H_3 minimalno 1 kriva
 H_4 0 krivih



A

$$E_1 > E_2 = E_3 = E_4$$

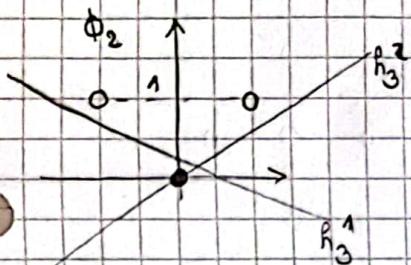


D

H_4 manje složenosti od H_{12}

2.12.

Napomena za 2.10.



Za uvertane h_1^1 i h_2^1

radi se 2 različite hipoteze

\Rightarrow ulazni prostor $X = \mathbb{R}^2$
- ne klasificiraju isto primjer $(-1, 0)$

Za uvertane h_1^2 i h_2^2 radi se o 1 hipotezi!
 \Rightarrow ulazni prostor je \mathbb{R}^2 , a hipoteza sjeci h_1^2 -os negdje na $\langle 0, 1 \rangle$ što isto klasificira sve primjere u \mathbb{R}

2.10.

$$\Phi_1(\vec{x}) = (1, x_1, x_2)$$

$$\Phi_2(\vec{x}) = (1, x_1, x_1 x_2)$$

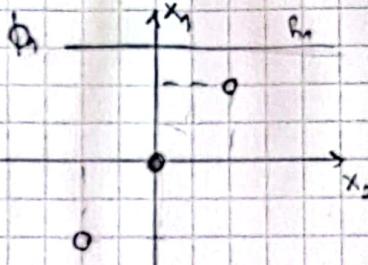
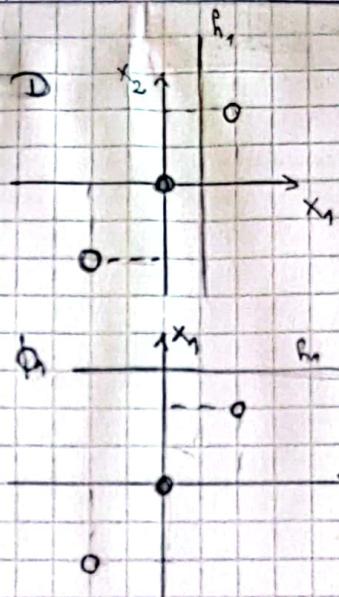
x_1	x_2	y	$x_1 x_2$
0	0	1	0
-1	-1	0	1
1	1	0	1

$$h_1(\vec{x}) = 1 \{ \theta_1 x_1 + \theta_0 \geq 0 \}$$

$$h_2(\vec{x}) = 1 \{ \theta_2 x_2 + \theta_0 \geq 0 \}$$

$$h_3(\vec{x}) = 1 \{ \vec{\theta} \cdot \vec{x} \geq 0 \}$$

$$h_4(\vec{x}) = 1 \{ x_1^2 + x_2^2 \geq \theta_0 \}$$

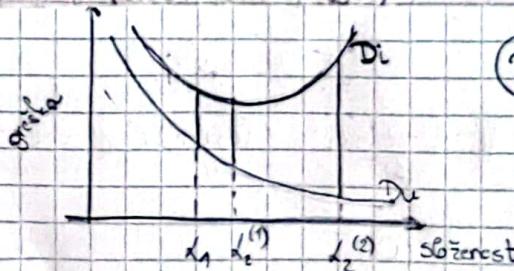


h_1 i h_2 - vertikalni / horizontali pravci u prostoru parametara

$h_4 + \Phi_1$ - kružnica tj. van kružna ishodište $\vec{\theta}$ centar kružnice
- ne može odvojiti dva primjera

$h_3 + \Phi_2$ - primjeri će postati linearno odvojivi moguće ih razdjeliti pravcem, ali postoji više točkih hipoteza (∞ mnogo)

2.12.



D1

Te manje složenosti je H_{D2}

3. Linearna regresija

Strojno učenje 1, UNIZG FER, ak. god. 2023./2024.

Jan Šnajder, vježbe, v2.1

1 Zadatci za učenje

1. [Svrha: Razumjeti osnovne komponente algoritma regresije te motivaciju za kvadratni gubitak i za postupak najmanjih kvadrata.]

- (a) Definirajte tri komponente algoritma linearog modela regresije.
- (b) Objasnite zašto koristimo kvadratnu funkciju gubitka, a ne gubitak 0-1.
- (c) Objasnite zašto težine ne možemo izračunati kao rješenje sustava jednadžbi $\mathbf{Xw} = \mathbf{y}$.

2. [Svrha: Razumjeti matrično rješenje za regulariziranu regresiju i izvježbati potrebnu matematiku. Razumjeti kako je rang matrice povezan sa postojanjem i stabilnošću rješenja. Razumijeti algoritamsku složenost postupka.]

- (a) Izvedite u matričnom obliku rješenje za vektor \mathbf{w} za linearan model regresije uz kvadratnu funkciju gubitka.
- (b) Što minimizira rješenje \mathbf{w} izvedeno pseudoinverzom? Što ako takvih rješenja ima više?
- (c) Raspolažemo sljedećim skupom primjera za učenje:

$$\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^4 = \{(0, 4), (1, 1), (2, 2), (4, 5)\}.$$

Podatke želimo modelirati modelom jednostavne regresije: $h(x) = w_0 + w_1 x$. Napišite kako bi u ovome konkretnom slučaju izgleda jednadžba iz zadatka (a) (Ne morate ju izračunavati, samo ju napišite da se vide konkretni brojevi.)

- (d) Jednadžba iz zadatka (a) daje rješenje u zatvorenoj formi, međutim rješenje nije uvijek izračunljivo na taj način. Što predstavlja problem? Pod kojim uvjetom je rješenje izračunljivo pomoću jednadžbe iz (a)? Možemo li rješenje izračunati i kada taj uvjet nije ispunjen? Kako?
- (e) U situacijama kada je rješenje izračunljivo jednadžbom iz zadatka (a), izračun ponekad može biti računalno zahtjevan. Što predstavlja problem? Je li problem izražen kada imamo mnogo primjera za učenje ili kada imamo mnogo značajki? Obrazložite odgovor.

3. [Svrha: Uvjeriti se da, uz određene prepostavke, funkcija kvadratne pogreške ima probabilističko tumačenje i opravdanje.] Kod postupka najmanjih kvadrata empirijska je pogreška definirana kao:

$$E(\mathbf{w}|\mathcal{D}) = \frac{1}{2} \sum_{i=1}^N (\mathbf{w}^T \mathbf{x}^{(i)} - y^{(i)})^2.$$

Pokažite da je minimizacija gornjeg izraza istovjetna maksimizaciji log-vjerojatnosti $\ln P(\mathbf{y}|\mathbf{X}, \mathbf{w})$ (odnosno minimizaciji negativne log-vjerojatnosti) uz prepostavku normalno distribuiranog šuma $\mathcal{N}(h(\mathbf{x}; \mathbf{w}), \sigma^2)$.

2 Zadaci s ispita

1. (P) Funkcija kvadratne pogreške definirana je kao:

$$E(\mathbf{w}|\mathcal{D}) = \frac{1}{2} \sum_{i=1}^N (\mathbf{w}^T \mathbf{x}^{(i)} - y^{(i)})^2$$

Izvedite matrični zapis ove funkcije. **Kako glasi matrični zapis ove funkcije, nakon sredivanja izraza, a prije deriviranja?**

- A $\frac{1}{2}(\mathbf{w}^T \mathbf{X}^T \mathbf{X} \mathbf{w} - 2\mathbf{y}^T \mathbf{X} + \mathbf{w}^T \mathbf{w})$
- B $\frac{1}{2}(\mathbf{w} \mathbf{X}^T \mathbf{X} \mathbf{w}^T - 2\mathbf{y}^T \mathbf{w} + \mathbf{y}^T \mathbf{y})$
- C $\frac{1}{2}(\mathbf{w}^T \mathbf{X}^T - 2\mathbf{w}^T \mathbf{X} \mathbf{y} + \mathbf{y}^T \mathbf{y})$
- D $\frac{1}{2}(\mathbf{w}^T \mathbf{X}^T \mathbf{X} \mathbf{w} - 2\mathbf{y}^T \mathbf{X} \mathbf{w} + \mathbf{y}^T \mathbf{y})$

2. (P) Razmatramo model jednostavne regresije:

$$h(x; w_0, w_1) = w_0 + w_1 x$$

Model linearne regresije inače koristi funkciju kvadratnog gubitka:

$$L(y, h(\mathbf{x})) = (y - h(\mathbf{x}))^2$$

Međutim, u našoj implementaciji greškom smo funkciju gubitka definirali ovako:

$$L(y, h(\mathbf{x})) = (y + h(\mathbf{x}))^2$$

S tako pogrešno definiranom funkcijom gubitka, postupkom najmanjih kvadrata treniramo naš model na skupu primjera čije su oznake uzorkovane iz distribucije $\mathcal{N}(-1 + 2x, \sigma^2)$, gdje je varijanca σ^2 razmjerno malena (tj. nema mnogo šuma). **Koji vektor težina (w_0, w_1) očekujemo (približno) dobiti kao rezultat najmanjih kvadrata?**

- A $(1, -2)$
- B $(2, -1)$
- C $(-1, 2)$
- D $(0, 0)$

3. (P) Jednostavnom regresijom modeliramo ovisnost nezavisne varijable y o zavisnoj varijabli x . Model treniramo postupkom običnih najmanjih kvadrata (OLS) na skupu podataka $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_i = \{(0, 0), (2, 0), (3, 2), (5, 2)\}$. Neka je h hipoteza koju dobivamo treniranjem modela te neka je L^i gubitak hipoteze h na primjeru $x^{(i)}$, tj. $L^i = L(y^{(i)}, h(x^{(i)}))$. **Što vrijedi za gubitke hipoteze na pojedinim primjerima?**

- A $L^1 = L^2 = 1 < L^3 < L^4$
- B $L^1 = L^4 = 1, L^2 < L^3$
- C $L^1 = L^3 = 0, L^2 = L^4 < 1$
- D $L^1 = L^4 < L^2 = L^3$

4. (N) Model linearne regresije treniramo na skupu označenih primjera iz dvodimenzijskoga ulaznog prostora:

$$\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_i = \{((1, 5), 5), ((2, -3), -2), ((3, -5), 1), ((0, -2), -3), ((0, 0), 0)\}$$

Za preslikavanje iz ulaznog prostora u prostor značajki Φ koristimo funkciju $\phi(\mathbf{x}) = (1, x_1, x_2, x_1 x_2)$. Treniranjem modela na skupu (Φ, \mathbf{y}) dobili smo parametre $\mathbf{w} = (0.28, -0.58, 1.79, -0.75)^T$. Prisjetite se da probabilistički model linearne regresije šum oko $h(\mathbf{x}; \mathbf{w})$ modelira normalnom distribucijom, čija je gustoća vjerojatnosti općenito definirana kao $p(x|\mu, \sigma^2) = (\sqrt{2\pi}\sigma)^{-1} \exp(-\frac{1}{2}\sigma^{-2}(x - \mu)^2)$. Pretpostavite $\sigma^2 = 1$. Uz takav model, zanima nas log-izglednost parametara \mathbf{w} na skupu primjera Φ s oznakama \mathbf{y} . **Koliko iznosi log-izglednost $\ln \mathcal{L}(\mathbf{w}|\Phi, \mathbf{y})$?**

- A -12.63
- B -5.69
- C -4.73
- D -10.64

V03 - Linearna regresija

I Zadaci za učenje

1.1.

a) MODEL

$$h(\vec{x}) = \vec{w}^T \vec{x}$$

POGREŠKA

$$E(\vec{w} | D) = \frac{1}{2} \sum_{i=1}^n (y^i - h(\vec{x}^i))^2$$

OPTIMIZACIJSKI POSTUPAK

$$\vec{w}^* = (X^T X)^{-1} X^T \vec{y} = X^+ \vec{y} = \underset{\vec{w}}{\operatorname{arg\min}} E(\vec{w} | D)$$

- b) Tačto koristimo kvadratni gubitak, a ne gubitak 0-1?
=> kvadratni gubitak derivabilan => zatvorena forma
=> probabilistička interpretacija

- c) Težine re možemo računati iz $X\vec{w} = \vec{y}$
jer ovo rješenje $\vec{w} = X^{-1}\vec{y}$ neće uvijek postojati

X je matrica dimenzije $N \times (n+1)$ t.j. nije kvadratna
i nema inverz. Također u vjetri primjera tipično je
 $N \gg n+1$

1.2.

a)

Izvod (detaljnije vide 4.2.)

$$\begin{aligned} E(\vec{w} | D) &= \frac{1}{2} \sum_{i=1}^n (y^i - \vec{w}^T \vec{x}^i)^2 = \frac{1}{2} (\vec{x}\vec{w} - \vec{y})^T (\vec{x}\vec{w} - \vec{y}) \\ &= \frac{1}{2} (\vec{w}^T X^T \vec{y}^T) (\vec{x}\vec{w} - \vec{y}) \\ &= \frac{1}{2} (\vec{w}^T X^T \vec{x}\vec{w} - \underbrace{\vec{w}^T X^T \vec{y}}_{n \times 1} - \underbrace{\vec{y}^T X \vec{w}}_{1 \times 1} + \vec{y}^T \vec{y}) \\ &\quad \cdot \frac{1}{2} (\vec{w}^T X^T \vec{x}\vec{w} - 2\vec{y}^T X \vec{w} + \vec{y}^T \vec{y}) \end{aligned}$$

$$\nabla_{\vec{w}} E = \frac{1}{2} (\vec{w}^T (X^T X + (X^T X)^T) - 2\vec{y}^T X) = \vec{w}^T (X^T X) - \vec{y}^T X = 0$$

$$\begin{aligned} \vec{w}^T (X^T X) &= \vec{y}^T X \\ \Leftrightarrow X^T X \vec{w} &= X^T \vec{y} \\ \vec{w} &= (X^T X)^{-1} X^T \vec{y} = X^+ \vec{y} \end{aligned}$$

b) Rješenje \vec{w} minimizira kvadratnu pogršku, tj. minimizira $\|X\vec{w} - \vec{y}\|_2$.
 Ako takvih rješenja ima više, račun se rješuje s najmanjom normom težina $\|\vec{w}\|_2$.

c)

$$X = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 4 \end{bmatrix}$$

$$\vec{w} = \left(\begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 4 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 4 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 4 \end{bmatrix} \begin{bmatrix} 4 \\ 1 \\ 2 \\ 5 \end{bmatrix}$$

c) Račun inverza $X^T X$ može biti računalno vrlo zahtjevrav. Problem je kada primjeri imaju puno zračajki.

$$X^T X = [(n+1) \times N \cdot N \times (n+1)] \Rightarrow (n+1) \times (n+1)$$

Moguće izračunati rješenje nekom numeričkom metodom, npr. gradijentni spust.

d) Rješenje $\vec{w} = (X^T X)^{-1} X^T \vec{y}$ nije uveć izračunljivo
 \Rightarrow ako je rang matrice dizajna manji od $n+1$
 (tj. kada je $N < n+1$)

\Rightarrow tada pseudoinverza možemo izračunati drugim metodama, npr. SVD kompozicijom,

1.3.

$$\begin{aligned} P(\vec{y} | X, \vec{w}) &\stackrel{\text{i.i.d.}}{=} \prod_{i=1}^N P(y^i | \vec{x}^i) \\ &= \prod_{i=1}^N \frac{1}{6\sqrt{2\pi}} \exp\left(-\frac{(y^i - h(\vec{x}^i; \vec{w}))^2}{26^2}\right) \\ -\ln P(\vec{y} | X, \vec{w}) &= -\ln \prod_{i=1}^N \frac{1}{6\sqrt{2\pi}} \exp\left(-\frac{(y^i - h(\vec{x}^i; \vec{w}))^2}{26^2}\right) \\ &= -\left(-N \ln(6\sqrt{2\pi}) - \frac{1}{26^2} \sum_{i=1}^N (y^i - h(\vec{x}^i; \vec{w}))^2\right) \\ &= N \ln(6\sqrt{2\pi}) + \frac{1}{26^2} \underbrace{\sum_{i=1}^N (y^i - h(\vec{x}^i; \vec{w}))^2}_{F(\vec{w} | D)}, \end{aligned}$$

II Zadaci s ispita

2.1. $E(\vec{w} | D) = \frac{1}{2} \sum (\vec{w}^\top \vec{x}_i - y_i)^2 = \frac{1}{2} (\vec{x}\vec{w} - \vec{y})^\top (\vec{x}\vec{w} - \vec{y})$

 $= \frac{1}{2} (\vec{w}^\top \vec{x}^\top - \vec{y}^\top) (\vec{x}\vec{w} - \vec{y})$
 $= \frac{1}{2} (\vec{w}^\top \vec{x}^\top \vec{x}\vec{w} - \underbrace{\vec{y}^\top \vec{x}\vec{w}}_{\vec{y}^\top \vec{w}^\top \vec{x}^\top} - \vec{y}^\top \vec{w}^\top \vec{x}^\top + \vec{y}^\top \vec{y})$
 $= \frac{1}{2} (\vec{w}^\top \vec{x}^\top \vec{x}\vec{w} - 2\vec{y}^\top \vec{x}\vec{w} + \vec{y}^\top \vec{y})$

(D)

2.2. $L(y, h(\vec{x})) = (y + h(\vec{x}))^2$

oznake $\mathcal{N}(-1+2x, 6^2)$ 6^2 vrlo mala

$E(\vec{w} | D) = \frac{1}{2} \sum (y_i + \vec{w}^\top \vec{x}_i)^2 = \frac{1}{2} (\vec{x}\vec{w} + \vec{y})^\top (\vec{x}\vec{w} + \vec{y})$
 $= \dots = \frac{1}{2} (\vec{w}^\top \vec{x}^\top \vec{x}\vec{w} + 2\vec{y}^\top \vec{x}\vec{w} + \vec{y}^\top \vec{y})$

$\nabla_{\vec{w}} E = \frac{1}{2} (\vec{w}^\top (\vec{x}^\top \vec{x} + (\vec{x}\vec{x}^\top)) + 2\vec{y}^\top \vec{x})$
 $= \vec{w}^\top (\vec{x}^\top \vec{x}) + \vec{y}^\top \vec{x} = 0$

$\vec{w}^\top (\vec{x}^\top \vec{x}) = -\vec{y}^\top \vec{x} \quad | \top$

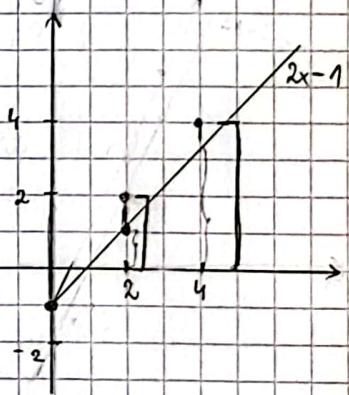
$\vec{x}^\top \vec{x} \vec{w} = -\vec{x}^\top \vec{y}$
 $\vec{w} = (\vec{x}^\top \vec{x})^{-1} \vec{x}^\top \vec{y} \cdot (-1) \quad //$

\downarrow
 $\text{očekujemo } \begin{bmatrix} -1 \\ 2 \end{bmatrix} \cdot (-1) = \begin{bmatrix} 1 \\ -2 \end{bmatrix}$

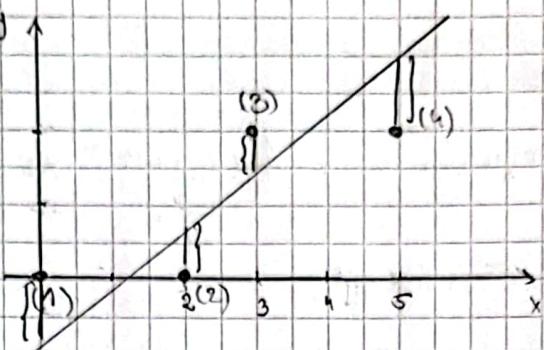
(A)

$\vec{w} = \begin{bmatrix} 1 & -2 \end{bmatrix}$

točno rješenje s $(y - h(x))^2 = L$



2.3.



A - nizgledna opt.

B - $L_1 = L_4 = 1$, neugledna opt.

C - baš neramna smrša

ale je $L_1 = L_3 = 0$, crda
bi model išao tačno kroz

(1) i (3), ali crda L2 i L4
nije mogu biti manje od 1

D) $L_1 = L_4 < L_2 = L_3$

Napomena: Zadatake moguće rješiti i računski

$$\vec{w} = (X^T X)^{-1} X^T \vec{y} = \dots = \begin{bmatrix} -2/13 \\ 6/13 \end{bmatrix}$$

$$h(\vec{x}^i) = -\frac{2}{13} + \frac{6}{13} x$$

$$L(y^i, h(\vec{x}^i)) = (y^i - h(\vec{x}^i))^2 \Rightarrow L_1 = L_4 = \frac{4}{109}, L_2 = L_3 = \frac{100}{109},$$

2.4.

(b)

\vec{x}	y
(1, 5)	5
(2, -3)	-2
(3, -5)	1
(0, -2)	-3
(0, 0)	0

$$\Phi(x) = (1, x_1, x_2, x_1 x_2)$$

$$\vec{w} = [0.28 \ -0.58 \ 1.79 \ -0.75]^T$$

$$\vec{h} = [4.9 \ -1.75 \ 0.84 \ -3.3 \ 0.28]$$

$$\ln P(\vec{w} | \Phi, \vec{y}) = C_n \prod_{i=1}^N \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} (y^i - h(\vec{x}^i, \vec{w}))^2\right)$$

$$= C_n \left(\frac{1}{\sqrt{2\pi}}\right)^N + \sum_{i=1}^N -\frac{1}{2} (y^i - h(\vec{x}^i, \vec{w}))^2 = -4.73,$$

$$= N \cdot C_n \left(\frac{1}{\sqrt{2\pi}}\right) - \underbrace{\frac{1}{2} \left[(5-4.9)^2 + (-2+1.75)^2 + (1-0.84)^2 + (-3+3.3)^2 + (0-0.28)^2 \right]}_{= -4.6 - 0.13325} = -4.73,$$

$$= -4.6 - 0.13325 = -4.73,$$

4. Linearna regresija II

Strojno učenje 1, UNIZG FER, ak. god. 2023./2024.

Jan Šnajder, vježbe, v1.11

1 Zadatci za učenje

1. [Svrha: Shvatiti kako se nelinarna funkcija u ulaznom prostoru funkcija preslikava u linearu funkciju odnosno (hiper)ravninu u prostoru značajki.]

- Regresijom želimo aproksimirati funkciju jedne varijable $y = 3 \cdot (x - 2)^2 + 1$. Skicirajte graf te funkcije. Definirajte linearan model $h(x)$ uz funkciju preslikavanja u prostor značajki $\phi(x) = (1, x, x^2)$. Odredite vektor težina $\mathbf{w} = (w_0, w_1, w_2)$ tog modela.
- Skicirajte u prostoru sa dimenzijama x_1 i x_2 (dakle u prostoru značajki) izokonture funkcije y . Naznačite u tom prostoru točke u koje se preslikavaju primjeri $x^{(1)} = 1$, $x^{(2)} = 2$ i $x^{(3)} = 3$. Koja je vrijednost od $h(x)$ za navedene primjere?

2. [Svrha: Razumjeti matrično rješenje za L2-regulariziranu regresiju. Razumjeti kako regularizacija popravlja lošu kondiciju matrice.]

- Izvedite u matričnom obliku rješenje za vektor \mathbf{w} za hrbatnu (L2-regulariziranu) regresiju.
- Raspolažemo sljedećim skupom primjera za učenje:

$$\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^4 = \{(0, 4), (1, 1), (2, 2), (4, 5)\}.$$

Podatke želimo modelirati polinomijalnom regresijskom funkcijom $h(x) = w_0 + w_1 x + w_2 x^2$. Napišite kako bi u ovome konkretnom slučaju izgleda jednadžba iz zadatka (a), ako se koristi regularizacijski faktor $\lambda = 10$. (Ne morate ju izračunavati, samo ju napišite da se vide konkretni brojevi.)

- Komentirajte na koji način L2-regularizacija rješava problem numeričke nestabilnosti rješenja za \mathbf{w} .
- Koristimo regresiju za predviđanje cijene nekretnine na temelju površine, starosti i udaljenosti od glavne prometnice. Koliko primjera nam je minimalno potrebno a da bi rješenje bilo izračunljivo jednadžbom iz (a), ako pritom ne koristimo preslikavanje. Koliko primjera nam je potrebno ako koristimo preslikavanja s polinomom drugog stupnja i interakcijskim značajkama? Što bi se dogodilo da kao značajku dodamo godinu izgradnje nekretnine? Obrazložite.

3. [Svrha: Isprobati izračun regresijskog modela s različitim funkcijama preslikavanja u prostor značajki te razviti intuiciju kako o tome kako ta funkcija određuje složenost hipoteze u ulaznom prostoru.] Linearnim modelom univarijatne regresije želimo aproksimirati jednu periodu funkcije $f(x) = \sin(\pi x)$. Raspolažemo sljedećim skupom primjera za učenje:

$$\mathcal{D} = \{(0.25, 0.707), (0.5, 1), (1, 0), (1.5, -1), (2, 0)\}.$$

- Izračunajte parametre linearog modela regresije u ulaznom prostoru primjera, tj. s funkcijom preslikavanja u prostor značajki definiranom kao $\phi(x) = (1, x)$. Skicirajte dobivenu regresijsku funkciju.
- Izračunajte parametre modela polinomijalne regresije drugog stupnja, tj. modela koji koristi funkciju preslikavanja u prostor značajki definiranu kao $\phi(x) = (1, x, x^2)$. Skicirajte dobivenu regresijsku funkciju.

- (c) Izračunajte parametre modela polinomijalne regresije četvrтog stupnja, tj. modela koji koristi funkciju preslikavanja u prostor značajki definiranu kao $\phi(x) = (1, x, x^2, x^3, x^4)$, uz L2-regularizaciju ($\lambda = 1$). Skicirajte dobivenu regresijsku funkciju.
- (d) Koji je model u ovom slučaju najprikladniji? Zašto?

Napomena: Izračun možete načiniti u nekom alatu koji podržava izračun matričnih operacija. Skicu također možete načiniti u nekom alatu, ili je možete napraviti ručno, izračunom vrijednosti regresijske funkcije u nekoliko odabralih točaka.

4. [Svrha: Razumjeti vezu između faktora regularizacije i složenosti modela.] Neka $\mathcal{H}_{d,\lambda}$ označava model polinomijalne regresije stupnja d s L2-regularizacijskim faktorom λ . Razmatramo četiri modela: $\mathcal{H}_{2,0}$, $\mathcal{H}_{5,0}$, $\mathcal{H}_{5,100}$, $\mathcal{H}_{5,1000}$ u ulaznom prostoru $\mathcal{X} = \mathbb{R}$. Pretpostavimo da su podatci u stvarnosti generirani funkcijom koja je polinom trećeg stupnja ($d = 3$). Pretpostavite da imamo razmjerno malo podataka i da je šum u podatcima razmjerno velik. Na dva odvojena crteža skicirajte
- (a) regresijsku funkciju $h(x)$ za sva četiri modela te
 - (b) pogrešku učenja i ispitnu pogrešku za sva četiri modela.
5. [Svrha: Shvatiti kako regularizacija utječe na optimizaciju. Shvatiti geometrijski argument zašto L1-regularizacija rezultira rijetkim modelima, a L2-regularizacije ne.]
- (a) Objasnite koja je svrha regularizacije i na kojoj se prepostavci temelji.
 - (b) Koja je prednost regulariziranog modela u odnosu na neregularizirani? Dolazi li ta prednost više do izražaja u slučajevima kada imamo puno primjera za učenje ili kada ih imamo malo?
 - (c) Razmatramo višestruku regresiju, $h(\mathbf{x}; \mathbf{w}) = w_0 + w_1x_1 + w_2x_2$. Skicirajte izokonture neregularizirane funkcije pogreške u ravnini \mathbb{R}^2 koju definiraju parametri w_1 i w_2 (napomena: funkcija pogreške je konveksna). Zatim skicirajte izokonture regularizacijskog izraza definiranog L2-normom vektora težina (i ova je funkcija konveksna). Pomoću ove skice objasnite na koji način regularizacija utječe na izbor optimalnih parametara (w_1^*, w_2^*) . Skicirajte krivulju mogućih rješenja za $\lambda \in [0, \infty)$.
 - (d) Ponovite prethodnu skicu, ali ovog puta sa L1-regularizacijom. Na temelju ove skice pokušajte odgovoriti na pitanje zašto L1-regularizacija daje rjeđe modele od L2-regularizacije.
6. [Svrha: Shvatiti vezu između težine značajki, važnosti značajki i složenosti modela.] Treniramo model regresije uz nelinearno preslikavanje u prostor značajki $\phi : \mathbb{R}^n \rightarrow \mathbb{R}^{m+1}$, gdje $m > n$, uz L2-regularizaciju.
- (a) Kako biste odredili optimalan regularizacijski faktor λ ?
 - (b) Kako, nakon treniranja modela, možemo provjeriti (1) koje su značajke nebitne i (2) je li izvorni (neregularizirani) model presložen?
 - (c) Kako bi se u ovom slučaju ponašao L1-regularizirani model?
 - (d) Pretpostavite da u podatcima postoji skup multikolinearnih značajki koje su, osim što su redundantne, također i irrelevantne, odnosno zavisna varijabla u stvarnosti uopće ne ovisi o tim varijablama. Ako model nije regulariziran, koje su očekivane težine tih značajki?

2 Zadataci s ispita

1. (N) Raspolažemo sljedećim skupom primjera u dvodimensijskome ulaznom prostoru:

$$\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\} = \{((1, 0), 1), ((2, -3), 2), ((3, 5), -1), ((5, 0), -4)\}$$

Na ovom skupu gradijentnim spustom trenirali smo L_1 -regularizirani model linearne regresije sa $\lambda = 1$. Dobili smo težine $\mathbf{w} = (2.12, -0.94, -0.08)$. **Koliko iznosi L_1 -regularizirana pogreška $E(\mathbf{w}|\mathcal{D})$?**

- A 7.10 B 2.69 C 1.58 D 0.29

2. (P) Raspolažemo skupom označenih primjera $\mathcal{D} \subset \mathbb{R}^n \times \mathbb{R}$ koji su u stvarnosti generirani funkcijom koja je polinom trećeg stupnja. Podataka imamo razmjerno malo, a šum u podatcima je velik. Skup \mathcal{D} dijelimo na skup za učenje i skup za ispitivanje. Neka je $\mathcal{H}_{d,\lambda}$ familija modela polinomialne regresije stupnja d s L2-regularizacijskim faktorom λ . Na skupu za učenje postupkom najmanjih kvadrata treniramo četiri modela iz te familije: $\mathcal{H}_{2,0}$, $\mathcal{H}_{5,0}$, $\mathcal{H}_{5,100}$ i $\mathcal{H}_{5,1000}$. Zatim izračunavamo empirijsku pogrešku (očekivanje kvadratnog gubitka) ovih modela na skupu za ispitivanje. **Što možemo zaključiti o ponašanju hipoteza iz ovih modela naučenih na skupu primjera \mathcal{D} ?**

- A Najbolje će generalizirati hipoteza iz $\mathcal{H}_{5,100}$ ili hipoteza iz $\mathcal{H}_{5,1000}$, ovisno o količini šuma u podatcima
- B Hipoteza iz $\mathcal{H}_{2,0}$ imati će veću pogrešku na skupu za učenje od hipoteze $\mathcal{H}_{5,0}$, ali mogu podjednako loše generalizirati
- C Hipoteza iz $\mathcal{H}_{5,1000}$ će generalizirati bolje od hipoteze iz $\mathcal{H}_{5,0}$, ali će imati veću pogrešku na skupu za učenje
- D Hipoteza iz $\mathcal{H}_{5,100}$ će bolje generalizirati od hipoteze iz $\mathcal{H}_{2,0}$ i imat će manju pogrešku na skupu za učenje

3. (P) Koristimo regresiju za predviđanje uspjeha na studiju. Kao značajke možemo koristiti ocjene u četiri razreda srednje škole (značajke x_1 – x_4), prosjek ocjena sva četiri razreda (x_5) te uspjeh iz matematike (x_6) i fizike (x_7) na državnoj maturi (ukupno 7 značajki). Ne moramo iskoristiti sve značajke, ali ih želimo iskoristiti što više. Za preslikavanje u prostor značajki koristimo preslikavanje s kvadratnim, interakcijskim i linearnim značajkama. Od interakcijskih značajki uzimamo samo interakcije parova značajki (npr. x_1x_2) i interakcije trojki (npr. $x_1x_2x_3$) između svih značajki koje koristimo. **Koliko minimalno primjera za učenje trebamo imati, a da bi rješenje bilo stabilno i bez regularizacije?**

- A 75 B 38 C 48 D 63

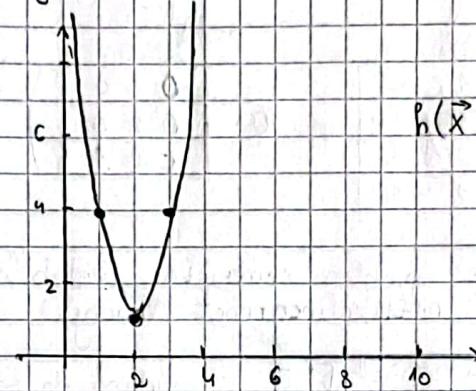
V04 - Linearna regresija II

I Podaci za učenje

1.1.

a)

$$y = 3(x-2)^2 + 1 = 3(x^2 - 4x + 4) + 1 = 3x^2 - 12x + 13$$



$$\Phi(x) = (1, x, x^2)$$

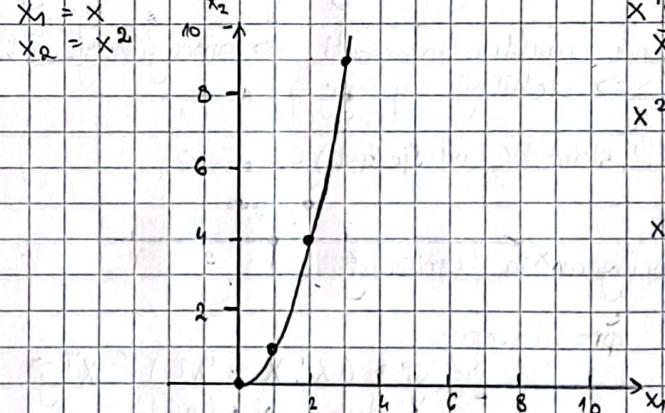
$$h(\vec{x}, \vec{w}) = \vec{w}^T \Phi(\vec{x}) = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \end{bmatrix} \begin{bmatrix} 1 & x & x^2 \end{bmatrix}$$

$$= w_0 + w_1 x + w_2 x^2$$

$$\vec{w} = [13 \quad -12 \quad 3]$$

b)

$$\begin{aligned} x_1 &= x \\ x_2 &= x^2 \end{aligned}$$



$$\begin{aligned} x^1 &= 1 & h(x^1) &= 4 \\ \vec{x} &= (1, 1) \end{aligned}$$

$$\begin{aligned} x^2 &= 2 & h(x^2) &= 1 \\ \vec{x} &= (2, 4) \end{aligned}$$

$$\begin{aligned} x^3 &= 3 & h(x^3) &= 4 \\ \vec{x} &= (3, 9) \end{aligned}$$

1.2.

a)

$$E_a(\vec{w} | D) = \frac{1}{2} \sum_{i=1}^N (y_i - h(\vec{x}_i))^2 + \frac{\lambda}{2} \|\vec{w}\|^2$$

$$= \frac{1}{2} (\Phi \vec{w} - \vec{y})^T (\Phi \vec{w} - \vec{y}) + \frac{\lambda}{2} \vec{w}^T \vec{w}$$

$$= \frac{1}{2} (\vec{w}^T \Phi^T \vec{y}) - \frac{\lambda}{2} \vec{w}^T \vec{w}$$

$$= \frac{1}{2} (\vec{w}^T \Phi^T \Phi \vec{w} - \vec{y}^T \Phi \vec{w} + \vec{y}^T \vec{y}) + \frac{\lambda}{2} \vec{w}^T \vec{w}$$

$$\nabla_{\vec{w}} E_a = \frac{1}{2} (\vec{w}^T ((\underline{\Phi}^T \underline{\Phi})^T + (\underline{\Phi} \underline{\Phi}^T)) - 2 \vec{y}^T \underline{\Phi}) + \frac{\lambda}{2} \vec{w}^T (\underline{I} + \underline{\Phi}^T \underline{\Phi})$$

$$= \vec{w}^T \underline{\Phi}^T \underline{\Phi} - \vec{y}^T \underline{\Phi} + \lambda \vec{w}^T \underline{I} = 0$$

$$\vec{w}^T (\underline{\Phi}^T \underline{\Phi} + \lambda \underline{I}) = \vec{y}^T \underline{\Phi}$$

$$(\underline{\Phi}^T \underline{\Phi} + \lambda \underline{I}) \vec{w} = \underline{\Phi}^T \vec{y}$$

$$\Rightarrow \vec{w} = (\underline{\Phi}^T \underline{\Phi} + \lambda \underline{I})^{-1} \underline{\Phi}^T \vec{y}$$

b)

$$D = \{(x^i, y^i)\} = \{(0, 4), (1, 1), (2, 2), (4, 5)\}$$

$$h(\vec{x}) = w_0 + w_1 x + w_2 x^2$$

$$\lambda = 10$$

$$\vec{w} = (\vec{\Phi}^T \vec{\Phi} + \lambda I)^{-1} \vec{\Phi}^T \vec{y}$$

$$= \left(\begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 4 \\ 0 & 1 & 4 & 16 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 4 & 16 \end{bmatrix} \right) + 10 \cdot$$

w_0 se ne regul.



$$\left(\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 4 \\ 0 & 1 & 4 & 16 \end{bmatrix} \begin{bmatrix} 4 \\ 1 \\ 2 \\ 5 \end{bmatrix}$$

c) L2-regularizacija rješava problem numeričke nestabilnosti rješenja za \vec{w} tako da smanjuje multikolinearnost značajki iz izraza

$$\vec{w} = (\vec{\Phi}^T \vec{\Phi} + \lambda I)^{-1} \vec{\Phi}^T \vec{y}$$

Grammova matrica dodaje vrijednost λ na drugom mjestu.

\Rightarrow smanjenje multikolinearnosti \Rightarrow smanjivanje kondicije matrice \Rightarrow stabilnije rješenje

d)

$$\vec{x} = (\text{površ., starost, udaljenost})$$

$$y = \text{cijena}$$

minimalno primjera za izračunljivost

- bez preslikavanja

$$\vec{w} = (\vec{X}^T \vec{X} + \lambda I)^{-1} \vec{X}^T \vec{y}$$

X dim: $N \times (n+1)$

$N \times 4$

$X^T X$ dim: 4×4

X^T dim: $4 \times N$

\vec{y} dim: $N \times 1$

λI dim: 4×4

• ako je X - nestabilno rješenje

$N > (n+1)$

$| N \geq 4 |$

- kvadratno preslikavanje s interakcijskim značajkama

$$\vec{\Phi}(\vec{x}) = (1, x_1, x_2, x_3, x_1 x_2, x_1 x_3, x_2 x_3, x_1^2, x_2^2, x_3^2)$$

$n+1 = 10 \Rightarrow$ minimalno 10 primjera

1.3

$$f(x) = \sin(\pi x)$$

$$\mathcal{D} = \{(0.25, 0.707), (0.5, 1), (1, 0), (1.5, -1), (2, 0)\}$$

a) $\Phi(\vec{x}) = (1, x)$

$$\underline{\Phi} = \begin{bmatrix} 1 & 0.25 \\ 1 & 0.5 \\ 1 & 1 \\ 1 & 1.5 \\ 1 & 2 \end{bmatrix}$$

$$\begin{aligned} \vec{w} &= (\underline{\Phi}^T \underline{\Phi} + \lambda \mathbb{I})^{-1} \underline{\Phi}^T \vec{y} \\ &= \dots \\ &\approx \begin{bmatrix} 0.9433 \\ -0.7637 \end{bmatrix} \end{aligned}$$

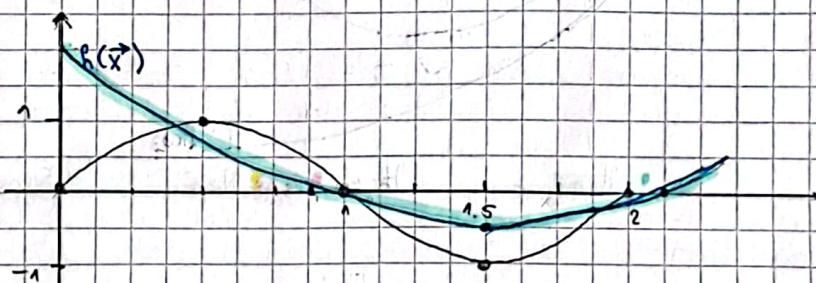
bez reg.



b) $\Phi(\vec{x}) = (1, x, x^2)$

$$\underline{\Phi} = \begin{bmatrix} 1 & 0.25 & 0.0625 \\ 1 & 0.5 & 0.25 \\ 1 & 1 & 1 \\ 1 & 1.5 & 2.25 \\ 1 & 2 & 4 \end{bmatrix}$$

$$\vec{w} = (\underline{\Phi}^T \underline{\Phi})^{-1} \underline{\Phi}^T \vec{y} = \begin{bmatrix} 1.7538 \\ -2.9408 \\ 0.9755 \end{bmatrix}$$



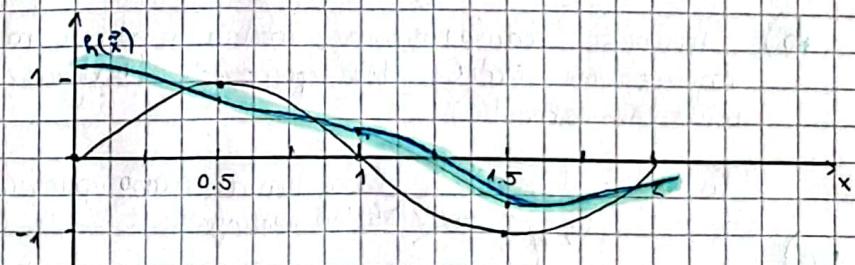
c) $\Phi(x) = (1, x, x^2, x^3, x^4)$

$$\lambda = 1$$

$$\underline{\Phi} = \begin{bmatrix} 1 & 0.25 & 0.0625 & 0.0156 & 3.906 \cdot 10^{-5} \\ 1 & 0.5 & 0.25 & 0.125 & 0.0625 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 1.5 & 2.25 & 3.375 & 5.0625 \\ 1 & 2 & 4 & 8 & 16 \end{bmatrix}$$

$$\vec{w} = (\underline{\Phi}^T \underline{\Phi} + \lambda \mathbb{I})^{-1} \underline{\Phi}^T \vec{y}$$

$$\vec{w} = \begin{bmatrix} 0.8330 \\ -0.2818 \\ -0.4156 \\ -0.3461 \\ 0.2479 \end{bmatrix}$$



d) Najprikladniji je model pod c) jer ima najmanju kvadratnu pogresku

1.4

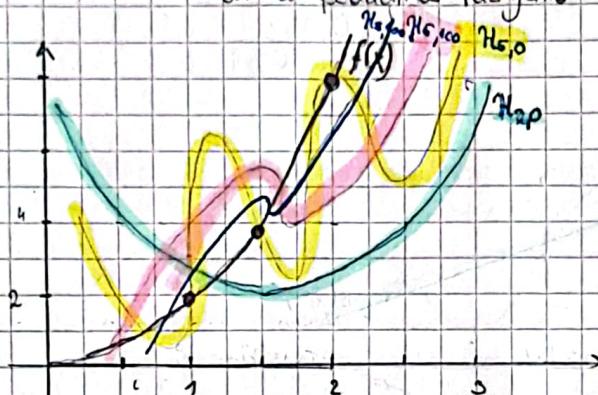
$H_{d, \lambda}$

$$\chi = \Pi$$

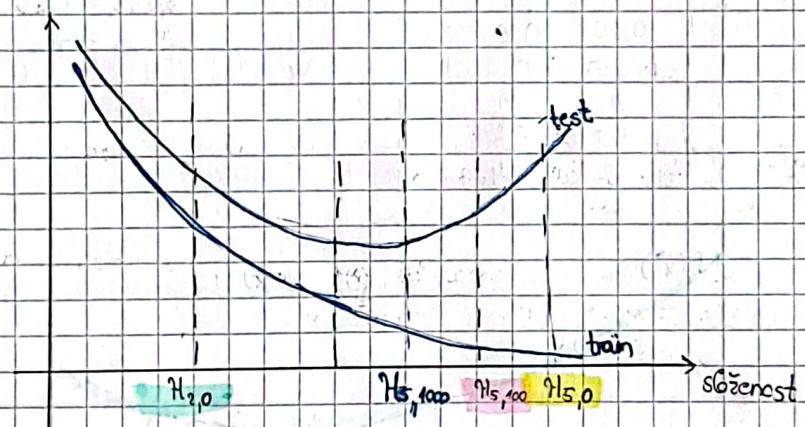
$H_{2,0}, H_{5,0}, H_{5,100}, H_{5,1000}$

a) regresijska fja

- podaci generir. polinomom 3. stupnja
- šum u podacima razmjerno velik



b)



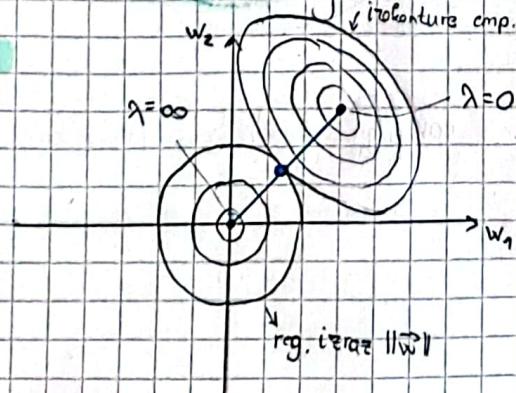
1.5 a) Regularizacija sprječava prenaučenost modela tako što ograničava rast vrijednosti parametara modela.

Temelji se na pretpostavci da je model složeniji što su magnitudo parametara veće.

b) Prednost reguliranih modela u odnosu na neregularizirani jest da je regulirani model teže preraučiti (+ daje stabilnija rješenja smanjivanjem multikolinarnosti)

To dođe do izražaja kada imamo malo primjera za učenje

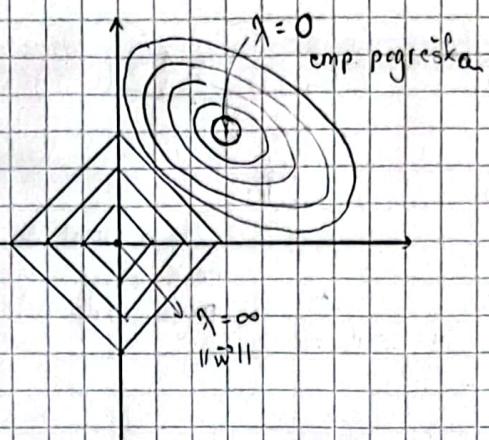
c)



$$\lambda = 0$$

L₂-regularizacijom parametri modela se pritežu prema 0. Optimalni parametri w_1^* i w_2^* su oni u kojima se izkonture dodiruju

d)



- kod L^1 regularizacije, izbor regul. izrata čija su izokonture, oštiri rubovi (ako se dogodi da jedna od težina bude 0, tada \vec{w}^* će biti na koordinatnoj osi)

• posledica \Rightarrow rijetki modeli

1.6

$$\phi: \mathbb{R}^n \rightarrow \mathbb{R}^{m+1}$$

$$m > n$$

L^2 -reg.

- Optimalan regularizacijski faktor λ odrediši bismo unutarsnom projekcijom (cross-validation)
- Nakon treniranja modela uz L^2 -regularizaciju parametri modela koji imaju male magnitudo (≈ 0) stoje už NETBITNE ZNAČAJKE.
Izvorni model će biti PRESLOŽEN ako su parametri uz rađenje značajke blizu 0!
- L^1 regularizirani model bi jače prigušio parametre modela i čak rezultirao rijetkim modelom
- Očekivane težine multikolinearnih značajki u neregulariziranom modelu imat će istu vrijednost za sve multikolinearne značajke.

II Zadaci s ispita

2.1.

$$\lambda = 1$$

$$\vec{w} = [2.12 \quad -0.94 \quad -0.08]$$

$$\|\vec{w}\| = \sum_{j=1}^m |w_j| = 0.94 + 0.08 = 1.02$$

bez w_0

$$E(\vec{w} | D) = \frac{1}{2} \sum_{i=1}^n (y_i - h(\vec{x}_i))^2 + \frac{\lambda}{2} \|\vec{w}\|_1 \rightarrow 0.07$$

$$D = \{(x^i, y^i)\} = \{(1, 0), 1\}, ((2, -3), 2), ((3, 5), -1), ((5, 0), -4)\}$$

$$h(\vec{x}) = w_0 + w_1 x_1 + w_2 x_2 = 2.12 - 0.94 x_1 - 0.08 x_2$$

$$h = [1.18 \quad 0.48 \quad -1.1 \quad -2.58]$$

$$E(\vec{w} | D) = \frac{1}{2} (0.18^2 + (-1.52)^2 + (-0.1)^2 + (1.42)^2) + \frac{1}{2} \cdot 1.02 \\ = 3.6946$$

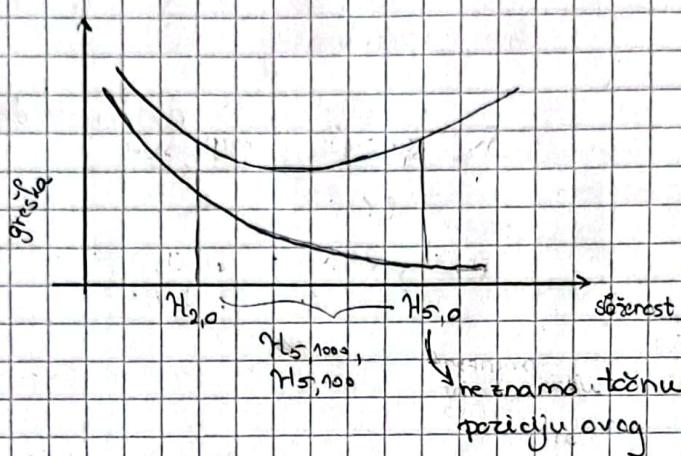
2.2.

$H_{d,n}$

$$H_{2,0} \quad H_{5,0}$$

$$H_{5,100} \quad H_{5,1000}$$

$$E(w|D) = \frac{1}{2} \sum (y_i - h(x_i))^2$$



(b)

$H_{2,0}$ imat će veću pogrešku na skupu za učenje od $H_{5,0}$, ali mogu podjednako loše generalizirati.

2.3.

(c)

$$x_1 - x_4 = \text{ocjena srednje škole}$$

$$\bar{x}_5 = \text{prosjek svih ocjena}$$

$$\bar{x}_6 = \text{mat.}$$

$$x_7 \rightarrow \text{fiz.}$$

\Rightarrow kolinearnost (mičemo iz skupa)

$$\phi(x_1, x_2, x_3, x_4, x_5, x_6, x_7) = \begin{cases} 1 & -1 \\ x_1 & -6 \\ x_2 & -6 \\ x_3 & -15 \\ x_4 & -20 \end{cases} \quad \left. \begin{array}{l} \\ \\ \\ \\ \end{array} \right\} 48 \text{ značajki}$$

već uračunato pred preljev.

• rješenje stabilno i bez regularizacije

$$\begin{aligned} N &\geq m+1 \\ N &\geq 48 \end{aligned}$$

5. Linearni diskriminativni modeli

Strojno učenje 1, UNIZG FER, ak. god. 2023./2024.

Jan Šnajder, vježbe, v2.6

1 Zadatci za učenje

1. [Svrha: Razumjeti geometriju linearног modela.]

- (a) Dokažite da je \mathbf{w} normala (hiper)ravnine.
- (b) Izvedite izraz za predznačenu udaljenost primjera \mathbf{x} od (hiper)ravnine.

2. [Svrha: Isprobati na konkretnom kako se linearна regresija može upotrijebiti za klasifikaciju. Razumjeti kako ostvariti višeklasnu klasifikaciju pomoću više binarnih modela. Razumjeti zašto je korištenje linearne regresije za klasifikaciju loša ideja.] Na predavanjima smo pokazali kako se linearan model regresije može (pokušati) koristiti za klasifikaciju. Pokažite to na sljedećim primjerima iz triju ($K = 3$) klase:

$$\begin{aligned}\mathcal{D} &= \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^6 \\ &= \{((-3, 1), 0), ((-3, 3), 0), ((1, 2), 1), ((2, 1), 1), ((1, -2), 2), ((2, -3), 2)\}.\end{aligned}$$

- (a) Primijenite pristup *jedan-naspram-ostali* (OVR), definirajte matricu dizajna i vektor oznaka \mathbf{y} za svaki od triju modela te izračunajte hipoteze $h_j(\mathbf{x})$ za svaku od triju klase. Izračun možete napraviti ručno ili u nekom alatu.
- (b) Izračunajte diskriminacijske funkcije $h_{01}(\mathbf{x})$, $h_{12}(\mathbf{x})$ i $h_{02}(\mathbf{x})$ između parova susjednih klasa. Skicirajte primjere i dobivene granice u prostoru \mathbb{R}^2 .
- (c) U koju bi klasu bio klasificiran primjer $\mathbf{x} = (-1, 3)$? Obrazložite odgovor.
- (d) Možete li reći koja je vjerojatnost da primjer pripada toj klasi? Obrazložite odgovor.
- (e) Objasnite koja je prednost pristupa OVR nad pristupom *jedan-naspram-jedan* (OVO), a što je nedostatak.
- (f) U praksi linearnu regresiju ne bismo željeli koristiti za klasifikaciju. Zašto? Pokažite na gornjem primjeru u čemu je problem (možete modifcirati primjer).

3. [Svrha: Razumjeti kriterij perceptronu i ograničenja koja proizlaze iz toga što ta funkcija nije derivabilna.]

Algoritam perceptronu minimizira pogrešku $E_p(\mathbf{w}|\mathcal{D})$, koju nazivamo *kriterij perceptronu*. Ta je funkcija aproksimacija udjela pogrešnih klasifikacija (engl. *misclassification ratio*), odnosno očekivanja gubitka 0-1, $E_m(\mathbf{w}|\mathcal{D})$, koju bismo idealno htjeli minimizirati, ali to ne možemo. Pogledajte (u skripti s predavanja) kako izgleda pogreška perceptronu u prostoru parametara.

- (a) Objasnite zašto ne možemo izravno minimizirati $E_m(\mathbf{w}|\mathcal{D})$.
- (b) Je li pogreška perceptronu $E_p(\mathbf{w}|\mathcal{D})$ gornja ograda za pogrešku $E_m(\mathbf{w}|\mathcal{D})$? Objasnite.
- (c) Jedan nedostatak perceptronu jest da rješenje \mathbf{w}^* (a time i položaj granice) ovisi o početnim težinama i redoslijedu predočavanja primjera. Pozivajući se na sliku površine pogreške u prostoru parametara, objasnite zbog čega je to tako.
- (d) Drugi nedostatak perceptronu jest da postupak ne konvergira ako primjeri nisu linearno odvojni. Pozivajući se opet na sliku površine pogreške u prostoru parametara, objasnite zašto je to tako.

4. [Svrha: Razumjeti odnose između funkcija gubitaka različitih modela. Razumjeti kako funkcija gubitka određuje dobra i loša svojstva modela.]

- (a) Skicirajte na jednome grafikonu sljedeće tri funkcije gubitka: (1) kvadratni gubitak regresije, (2) gubitak perceptronu i (3) gubitak 0-1.
- (b) Odgovorite čemu odgovara desna strana grafikona (x -os veća od nule), a čemu lijeva (x -os manja od nule).
- (c) Pozivajući se na skicu, odgovorite za što kvadratni gubitak nije prikladan gubitak u slučajevima kada želimo minimizirati broj pogrešnih klasifikacija.
- (d) Pozivajući se na skicu, odgovorite za koje će modele očekivanje gubitka (empirijska pogreška) biti veće od udjela pogrešnih klasifikacija.

2 Zadatci s ispita

1. (P) Treniramo linearni diskriminativni model u dvodimenziskome ulaznom prostoru. Skup za učenje čine samo dva primjera, $(\mathbf{x}_1, y_1) = ((1, 0), +1)$ i $(\mathbf{x}_2, y_2) = ((0, 1), -1)$. Na tom skupu primjenjujemo algoritam strojnog učenja koji ima induktivnu pristranost takvu da rješenje maksimizira minimalnu udaljenost primjera od hiperravnine. Naučen model ispravno klasificira oba primjera, pri čemu za oba primjera vrijedi $y \cdot h(\mathbf{x}) = 5$. **Koliko iznosi težina w_2 tako naučenog modela?**

A -1 B 5 C -5 D 1

2. (P) Razvijamo sustav za automatsku klasifikaciju novinskih članaka u jednu od pet kategorija. Tih pet kategorija su "sport", "politika", "kriminal", "znanost" i "lifestyle". Najveća razlika u veličini klase je između kategorija "politika" i "znanost". Očekivano, u kategoriji "politika" ima najviše članaka, dok ih u kategoriji "znanost" ima $5 \times$ manje, što je u redu jer to ionako nitko ne čita. Svaki novinski članak prikazujemo kao vektor riječi, gdje su komponente vektora broj pojavljivanja pojedine riječi. Problem rješavamo algoritmom perceptron. Koristimo algoritam perceptron. Budući da je perceptron binaran klasifikator, odlučili smo primijeniti shemu OVR ili shemu OVO za dekompoziciju višeklasnog klasifikacijskog problema u skup binarnih klasifikacijskih problema. **Što možemo očekivati?**

- A OVO će imati $2 \times$ puta manje značajki od OVR, ali bi mogao raditi bolje na člancima iz kategorije "znanost"
- B OVR će imati $2 \times$ manje značajki od OVO, ali bi mogao raditi lošije na člancima iz kategorije "znanost"
- C OVO će imati $5 \times$ puta manje značajki od OVR, ali bi mogao raditi lošije na člancima iz kategorije "znanost"
- D OVR će imati $5 \times$ manje značajki od OVO, ali bi mogao raditi bolje na člancima iz kategorije "znanost"

3. (P) Na skupu od $N = 1000$ primjera sa $n = 555$ značajki rješavamo problem višeklasne klasifikacije. Imamo $K = 4$ klase, s po 400, 300, 200 i 100 primjera. Za klasifikaciju želimo koristiti binarnu logističku regresiju u shemi OVO ili u shemi OVR (ovo nije tipično, ali je moguće). Pretpostavite da ne koristimo nikakvu regularizaciju, $\lambda = 0$. Razmotrite, za obje sheme, za koliko binarnih modela će rješenje optimizacijskog postupka sigurno biti nestabilno zbog loše kondicije matrice dizajna. **Koliko modela će sigurno biti više nestabilno u shemi OVO nego u shemi OVR?**

A 2 B 3 C 4 D 5

4. (N) Raspolažemo sljedećim skupom za učenje u dvodimenziskome ulaznom prostoru:

$$\mathcal{D} = \{(\mathbf{x}(i), y(i))\} = \{((1, 0), +1), ((2, -3), -1), ((2, 5), -1)\}$$

Na ovom skupu treniramo perceptron. Pritom koristimo funkciju preslikavanja u šesterodimenziski prostor značajki, koja je definirana na sljedeći način:

$$\phi(\mathbf{x}) = (1, x_1, x_2, x_1 x_2, x_1^2, x_2^2)$$

Početne težine perceptronu neka su sljedeće:

$$\mathbf{w} = (1, 0, -1, 2, -2, 0)$$

Koliko iznosi empirijska pogreška perceptronu na skupu za učenje prije početka treninga (dakle, s početnim težinama)?

- A 8
- B 9
- C 16
- D 25

5. (P) Razmotrimo sljedeći skup označenih primjera:

$$\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\} = \{((-2, 0), -1), ((-1, 0), +1), ((1, -0), +1), ((2, 0), -1)\}$$

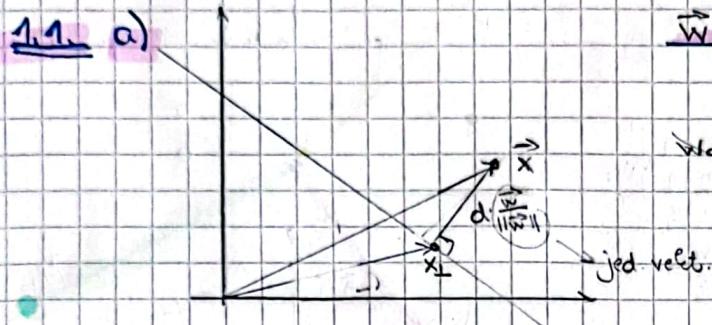
Ovaj skup nije linearno odvojiv i algoritam perceptronu neće konvergirati. Linearna neodvojivost podataka je konceptualni razlog zašto algoritam ne konvergira. **Koji je tehnički razlog zašto algoritam perceptronu na ovom skupu primjera neće konvergirati?**

- A U svakoj točki prostora parametara postoji barem jedan primjer za koji je gradijent gubitka veći od nule
- B Premda je empirijska pogreška na ovom skupu primjera derivabilna, ona je uglavnom konstantna
- C U prostoru parametara ne postoji točka u kojoj je gradijent empirijske pogreške jednak nuli
- D U prostoru parametara postoji više točaka za koje je empirijska pogreška jednaka nuli

VO5 - Linearni diskriminativni modeli

I. Zadaci za vježbe

1.1 a)



\vec{w} normala hiperavnine

\vec{x}_1, \vec{x}_2 na hiperavnini

$$h(\vec{x}_1) = h(\vec{x}_2)$$

$$\vec{w}_0 + \vec{w}^T \vec{x}_1 = \vec{w}^T \vec{x}_2 + \vec{w}_0$$

$$\vec{w}^T (\vec{x}_1 - \vec{x}_2) = 0$$

vektori sujci opst leži na hiperavnini

\Rightarrow skalarni produkt vektora 0

\Rightarrow vektori okomititi

(n)

\vec{w} je normala ravnine (ali bez w_0 !)

b) Iz sljedećeg: $\vec{x} = \vec{x}_\perp + d \cdot \frac{\vec{w}}{\|\vec{w}\|}$ / \vec{w}^T

$$\vec{w}^T \vec{x} = \vec{w}^T \vec{x}_\perp + d \|\vec{w}\| / + w_0$$

$$\vec{w}^T \vec{x} + w_0 = \vec{w}^T \vec{x}_\perp + w_0 + d \|\vec{w}\|$$

$$h(\vec{x}) = h(\vec{x}_\perp) + d \|\vec{w}\|$$

$$d = \frac{h(\vec{x})}{\|\vec{w}\|}$$

1.2

\vec{x}	y
(-3, 1)	0
(-3, 3)	0
(1, 2)	1
(2, 1)	1
(1, -2)	2
(2, -3)	2

a)

OVR - 3 klasifikatora

$$h_e(\vec{x}) =$$

$$\Phi = \begin{bmatrix} 1 & -3 & 1 \\ 1 & -3 & 3 \\ 1 & 1 & 2 \\ 1 & 2 & 1 \\ 1 & 1 & -2 \\ 1 & 2 & -3 \end{bmatrix}$$

$$\vec{y}_0 = [1 \ 1 \ 0 \ 0 \ 0 \ 0]^T$$

$$\vec{y}_1 = [0 \ 0 \ 1 \ 1 \ 0 \ 0]^T$$

$$\vec{y}_2 = [0 \ 0 \ 0 \ 0 \ 1 \ 1]^T$$

$$\vec{w}_e = \Phi^+ \vec{y}_e$$

$$\vec{w}_0 = [0.335 \ -0.217 \ -0.005]^T$$

$$\vec{w}_1 = [0.259 \ 0.234 \ 0.222]^T$$

$$\vec{w}_2 = [0.406 \ -0.017 \ -0.217]^T$$

$$h_0(\vec{x}) = 0.335 - 0.217x_1 - 0.005x_2$$

$$h_1(\vec{x}) = 0.259 + 0.234x_1 + 0.222x_2$$

$$h_2(\vec{x}) = 0.406 - 0.017x_1 - 0.217x_2$$

$$h(\vec{x}) = \operatorname{argmax}_e h_e(\vec{x})$$

b) $h_{01}(\vec{x})$, $h_{02}(\vec{x})$, $h_{12}(\vec{x}) = ?$
 → discriminacijske fje

$$h_{01}(\vec{x}) = w_{01,0} + w_{01,1}x_1 + w_{01,2}x_2; \quad \vec{w}_{01} = \vec{w}_0 - \vec{w}_1$$

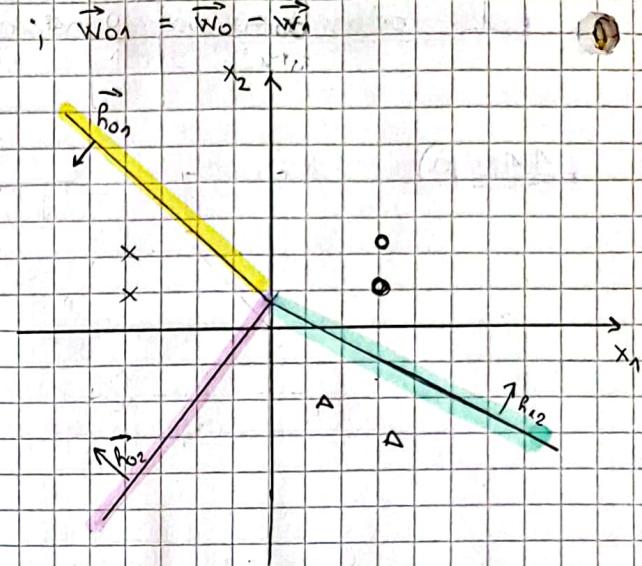
(→ simetrija između h_0 i h_1)

analogno

$$\vec{w}_{02} = \vec{w}_0 - \vec{w}_2$$

$$\vec{w}_{12} = \vec{w}_1 - \vec{w}_2$$

$$\begin{aligned}\vec{w}_{01} &= [0.076 \quad -0.451 \quad -0.227]^T \\ \vec{w}_{02} &= [-0.071 \quad -0.2 \quad 0.212]^T \\ \vec{w}_{12} &= [-0.147 \quad 0.251 \quad 0.439]^T\end{aligned}$$



c) Primjer $\vec{x} = (-1, 3)$

$$\begin{aligned}h_0(\vec{x}) &= 0.537 \\ h_1(\vec{x}) &= 0.691 \\ h_2(\vec{x}) &= 0.228\end{aligned} \quad \left. \begin{array}{l} \text{maksimum: } 0.691 \rightarrow \text{klasa 1} \\ (\text{ki moguće isčitati iz gornje slicice}) \end{array} \right.$$

d) Ne možemo reći koja je vjerojatnost da primjer $(-1, 3)$ pripada klasu 1 jer linearni model koji sruši koristi u skromi OVR nema probabilističku interpretaciju

e) OVR vs OVO

Prednost: u OVR shemi trenira se manje modela nego u OVO shemi (K modela, na prema $\binom{K}{2}$ modela)

Nedostatak: problem neuravnoteženosti klasa

→ za svaki od K klasifikatora postoji mnogo više primjera jedne klase nego druge
 ⇒ čao posljedica optimiz. postupka je smanjuje $E(\vec{w} | D)$ nauštrob primjera iz manje klase

f) Linearna regresija ne koristi se za klasifikaciju zbog
 - nerobustnosti na strane vrijednosti
 - nema probabilističku interpretaciju

Pr. dodavanje primjera $((8, 3), 1)$ u D

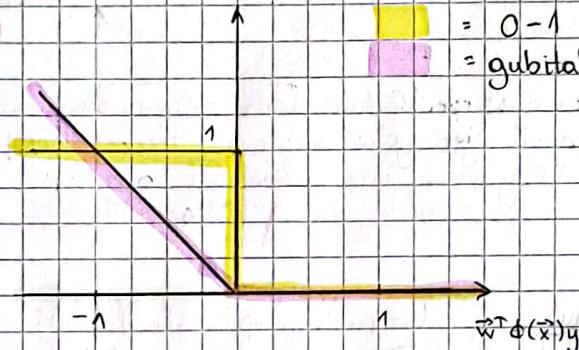
1.3.

a)

Fju očekivanja gubitka $0-1$ $E_m(\vec{w} | D)$ ne možemo izravno optimizirati jer nije derivabilna i na većini domena je konstantna (pa ne možemo koristiti ni metode građljivog spusta).

$$E_m(\vec{w} | D) = \frac{1}{N} \sum_{i=1}^N 1\{|f(\vec{w}^\top \phi(\vec{x}^i))| + y^i|\}$$

b)



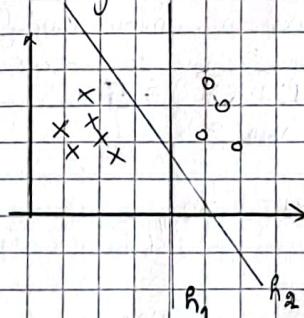
= 0-1 gubitak
= gubitak perceptronra

Pogreška perceptronra nije gornja ograda 0-1 pogrešku

↳ na intervalu $[-1, 0]$
 $E_p(\vec{w} | D) < E_m(\vec{w} | D)$

c)

Perceptron će na temelju početnih težina i redoslijedu predstavljanja primjera odabrati 1 hiperravninu od njih ∞ . Algoritam perceptronra će se zaustaviti čim ispravno klasificira sve primjere za učenje, što ovisno o početnim uvjetima neće dati uvijek hiperravninu koja najbolje klasificira (dabit ćemo h_1 , a htjeli bismo h_2)



d)

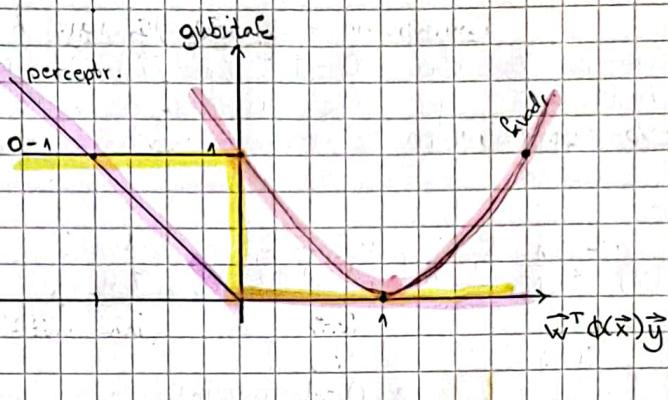
Budući da kod linearne neodvojivih primjera nije moguće postići savršenu klasifikaciju algoritam se ne zaustavlja.
→ tj. ne može se postići $E_p(\vec{w} | D) = 0$
jer će se konstantno naći primjer kažnjivati sa određ. gubitkom

NEDOSTACI PERCEPTRONA

- ① Loša pristranost preferencije (\vec{w} ovisi o poč. uvj.)
- ② algoritam ne konvergira za linearne neodvojive primjere
- ③ izlazi neraju vjerojatn. interpretaciju

1.4.

a) perceptr.



b)

$$x > 0$$

$$\Rightarrow \vec{w}^T \phi(\vec{x}) y > 0$$

- predstavlja
točno klasificirane
primjere

$$x < 0 \Rightarrow \vec{w}^T \phi(\vec{x}) y < 0$$

- predstavlja netočno
klasificir. primjere

c)

Kvadratni gubitak nije prikladan kada želimo minimizirati broj pogrešnih klasifikacija jer jačo kaznjava tečno klasificirane primjere, pa te primjere kaznjava isto kao i pogrešno klasificirane primjere

d)

Očekivanje gubitka (empirijska pogreška) biti će veća od udjela pogrešnih klasifikacija za kvadratni gubitak (i za log).

$$\text{kvadr. gubitak} > 0-1 \text{ GSS}$$

II Zadaci s ispita

2.1.

$$\begin{array}{c|c} \vec{x} & y \\ \hline (1, 0) & 1 \\ (0, 1) & -1 \end{array}$$

maksim. minim. udaljenost primjera od hiperplane

$$y \cdot h(\vec{x}) = 5$$

$$w_2 = ?$$

$$h(\vec{x}) = \frac{w_0 + w_1 x_1 + w_2 x_2}{\sigma} = 5x_1 - 5x_2 //$$

$$y \cdot h(\vec{x}) = 5$$

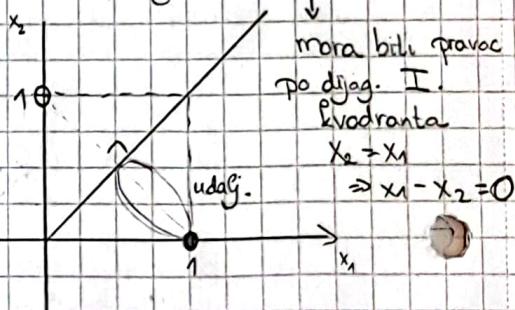
$$1 \cdot (w_1 \cdot 1 + w_2 \cdot 0) = 5$$

$$w_1 = 5$$

$$-1 \cdot (w_1 \cdot 0 + w_2 \cdot 1) = 5$$

$$w_2 = -5$$

(C)



b)

$$x > 0$$

$$\Rightarrow \vec{w}^T \phi(\vec{x}) y > 0$$

- predstavlja
točno klasificirane
primjere

$$x < 0 \Rightarrow \vec{w}^T \phi(\vec{x}) y < 0$$

- predstavlja netočno
klasificir. primjere

2.2. (B)

5 klasa

"sport"

"politika" 5x

"kriminal"

"znanost" x

"lifestyle"

OVO, OVR

perceptron

$$h(\vec{x}) = f(\vec{w}^T \phi(\vec{x})) = \begin{cases} +1 \\ -1 \end{cases}$$

$$\text{OVO} - \binom{5}{2} = 10 \text{ klasif.}$$

$$\text{OVR} - 5 \text{ klasif.}$$

OVR će imati 2x manje znajućih od OVO,
ali mogao bi takođe raditi za klasu znanost,

2.3.

$$N = 1000$$

$$n = 555$$

$$k=4 \quad \{400, 300, 200, 100\}$$

nestabilnost zbog loše kondicije Φ

dummy 1

$$\hookrightarrow \Phi \text{ dimenzija } N \times (n+1)$$

$$\text{stabilno deo } N \geq n+1$$

OVO schema

	poz. prim.	reg. prim.	dim $\bar{\Phi}$	stabihor
h_{01}	400	300	700×556	✓
h_{02}	400	200	600×556	✓
h_{03}	400	100	500×556	✗
h_{12}	300	200	500×556	✗
h_{13}	300	100	400×556	✗
h_{23}	200	100	300×556	✗

(4)

DNA schema $\rightarrow \forall h_j \text{ dim } \bar{\Phi} = 1000 \times 556 - \text{stabilno}$

(C) 4

2.4.

$$\begin{array}{c|ccc|c} \vec{x} & \phi(\vec{x}) & y & \\ \hline (1, 0) & (1, 1, 0, 0, 1, 0) & 1 & \phi(\vec{x}) = (1, x_1, x_2, x_1 x_2, x_1^2, x_2^2) \\ (2, -3) & (1, 2, -3, -6, 4, 9) & -1 & \vec{w} = (1, 0, -1, 2, -2, 0) \\ (2, 5) & (1, 2, 5, 10, 4, 25) & -1 & \end{array}$$

$$E(\vec{w} | D) = \sum_{i=1}^N \max(0, -\vec{w}^T \phi(\vec{x}^i) y^i)$$

$$= \max(0, -1 \cdot (-1)) + \max(0, +16 \cdot (-1)) + \max(0, -1 \cdot (8) \cdot (-1))$$

$$= 1 + 0 + 8 = 9$$

(B)

2.5.

$$\begin{array}{c|cc} \vec{x} & \nabla L & y \\ \hline (-2, 0) & (-1) & -1 \\ (-1, 0) & 1 & 1 \\ (1, 0) & 1 & 1 \\ (2, 0) & -1 & -1 \end{array}$$

Sledeća linearna neodvojivost
-perceptron ne konvergira

Technički razlog nekonvergencije?

$$\nabla_w E = \sum_{i=1}^n -\phi(\vec{x}^i) y^i$$

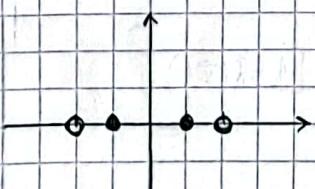
Alg. perceptrona

povlači do konverg.

$$\vec{w} \leftarrow \vec{w} - \eta \nabla_w L$$

$$\nabla_w L = -\phi(\vec{x}^i) y^i$$

(A) U svakoj točki prostora parametara postoji barem 1 primjer za koji je gradient gubiće veći od nule.



VOG - Logistička regresija

I Zadaci za učenje

1.1.

a) Poopteni linearni model

$$h(\vec{x}) = f(\vec{w}^T \phi(\vec{x}))$$

f = aktivacijska funkcija koja izlazi modela ograničava na interval $[0, 1]$ ili $[-1, 1]$

- Ako je f neelinearna fja \Rightarrow model je neelinearan u parametrima
- neelinearna granica \Leftrightarrow neelinearnost ϕ

b) Model logističke regresije:

$$h(\vec{x}) = \sigma(\vec{w}^T \phi(\vec{x})) = \frac{1}{1 + \exp(-\vec{w}^T \phi(\vec{x}))}$$

Sigmoida je prikidan odabir za aktivac. fju jer

- izlaz modela će imati vjerj. interpretaciju
- fja gubitka slabо razvijava TOČNO klasificirane primere
- derivabilna je $\sigma(x)(1-\sigma(x))$ i slična je fji proga

c) $E(\vec{w} | D) = -C_n P(\vec{y} | \vec{w})$

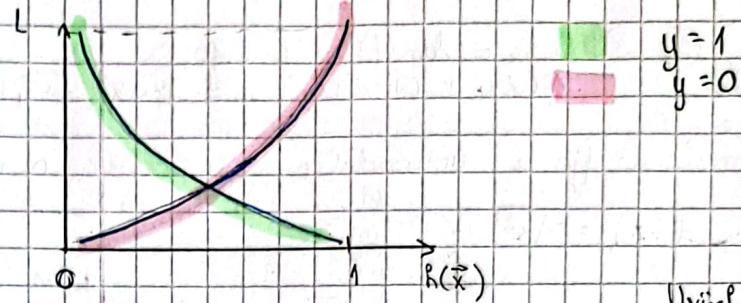
• Izlaz modela - Bernoullijeva jedžiba $P(y| \mu) = \mu^y (1-\mu)^{1-y}$

$$\begin{aligned} P(\vec{y} | \vec{x}, \vec{w}) &= \prod_{i=1}^N P(y^i | \vec{x}^i) \\ \hookrightarrow C_n P(\vec{y} | \vec{x}, \vec{w}) &= \ln \prod_{i=1}^N [h(\vec{x}^i)]^{y^i} [1-h(\vec{x}^i)]^{1-y^i} \\ &= \sum_{i=1}^N \ln [h(\vec{x}^i)]^{y^i} [1-h(\vec{x}^i)]^{1-y^i} \\ &= \sum_{i=1}^N y^i \ln [h(\vec{x}^i)] + (1-y^i) \ln [1-h(\vec{x}^i)] \end{aligned}$$

$$E(\vec{w} | D) \sim -C_n P(\vec{y} | \vec{x}, \vec{w})$$

$$E(\vec{w} | D) = \frac{1}{N} \sum_{i=1}^N -y^i \ln [h(\vec{x}^i)] - (1-y^i) \ln [1-h(\vec{x}^i)]$$

$$d) L(y_i, h(\vec{x}_i)) = -y_i \ln[h(\vec{x}_i)] - (1-y_i) \ln[1-h(\vec{x}_i)]$$



Najveći gubitak : teži ∞
 Najmanji gubitak : ne postoji ; teži 0

Uvijek postoji neki gubitak jer $h(\vec{x})$ lako izlazi nikad ne doje savršeno klasificiran primjer nego je izlaz na intervalu $[0,1]$

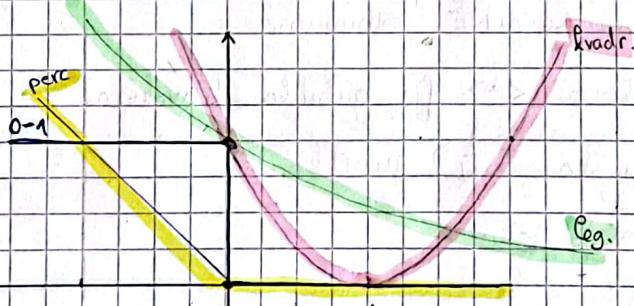
$$e) y \in \{-1, 1\}$$

$$L(y, 0) = 1$$

\Rightarrow vodi na reformulaciju:

$$L(y, h(\vec{x})) = \frac{1}{Cn} \ln(1 + \exp(-\vec{w}^T \phi(\vec{x})y))$$

f)



Tačko je logistički gubitak dokar na klasifikaciju
 Jako ložnjava retko klasificirane primjere, a malo ložnjava tačna klasi.
 primjere. Uvijek raniči neki gubitak!

Je li logistički gubitak konveksni
 surrogat 0-1 gubitka i što to znači?
 Da jest! To znači da je log.
 gubitak konveksan i $\nabla \vec{w}^T \phi(\vec{x})y$ ima
 veću vrijednost od gubitka 0-1

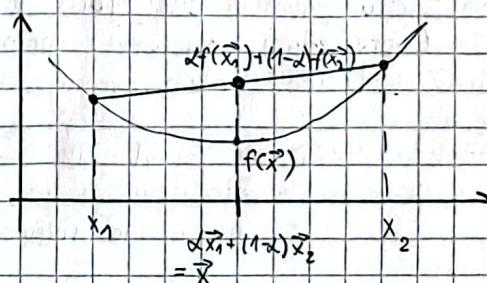
1.2. a) $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$ konveksna

\Leftrightarrow ① domena f je konveksni skup

$$\sum \lambda_i \vec{x}_i \in \text{dom}(f)$$

$$\text{ali } \sum \lambda_i = 1 \quad i \quad \vec{x}_i \in \text{dom}(f)$$

② $\forall x_1, x_2 \in \text{dom}(f) \quad i \quad \forall \lambda \in [0, 1]$
 $\Rightarrow f(\vec{x}) = f(\lambda \vec{x}_1 + (1-\lambda) \vec{x}_2) \leq \lambda f(\vec{x}_1) + (1-\lambda) f(\vec{x}_2)$



b) f je konveksna (unimodala)

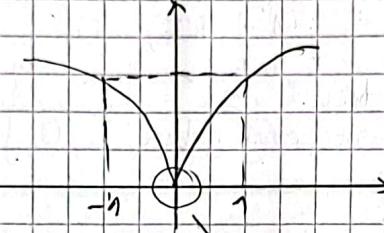
\Leftrightarrow ① $\text{dom}(f)$ je konveksni skup

② $\forall x_1, x_2 \in \text{dom}(f) \quad \forall \lambda \in [0, 1] \quad f(\lambda x_1 + (1-\lambda)x_2) \leq \max\{f(x_1), f(x_2)\}$

Svaka konveksna fja je unimodala, ali obrat ne vrijedi

Pr.

$$f(x) = \sqrt[3]{x^2}$$



$$x_1 = 1, \quad x_2 = 2 \\ \lambda = 0.5$$

Konveksna Ne!

$$f(\lambda x_1 + (1-\lambda)x_2) \\ = f(1.5) = 1.31$$

$$\lambda f(x_1) + (1-\lambda)f(x_2) \\ = 1.29$$

Konvexne DA!

$$f(\lambda x_1 + (1-\lambda)x_2) = 1.31$$

$$\max\{f(x_1), f(x_2)\} = 1.59$$

$$1.31 > 1.29$$

c) U SU preferiramo konveksne fje pogreške jer nam to garantira da ćemo rješavanjem optimizacijskog postupka metodom konvergencije dobiti GLOBALNE, a ne LOKALNE minimume

Fja pogreške je konveksna \Leftrightarrow fja gubitka konveksna

$$\text{Pogreška} = \frac{1}{N} \sum \text{gubitka}$$

1.3.

a) Ideja gradijentnog spusta

- heuristička optimizacija kreće od činjenice da je u tački ekstrema $\nabla f(\vec{x}) = 0$

- u tačkama neekstrema $\nabla f(\vec{x})$ pokazuje u smjeru najbržeg rasta, fje f

- polaskom iz neke početne tačke te pomicanjem u suprotnom smjeru od smjera gradijenta možemo stići u minimum fje

• Linjsko pretraživanje = odabire se m linijski minimizira fju $g(\eta) = f(\vec{x} + \eta \Delta \vec{x})$ u smjeru spusta (suprotno od gradijenta)
→ ubrzava konvergenciju

b) Grupni gradijentni spust radi aržuriranje težina tek nakon završene epohi (pregedari su svi primjeri), a stohastički grad. spust aržuriра težine nakon svakog viđenog primjera.

Prednost SGD-a nad BGD-om jest

- što omogućuje online učenje
- manje je računalno zahtijevan

$$c) \nabla_{\vec{w}} F(\vec{w} | D) = \sum_{i=1}^N \nabla_{\vec{w}} L(y^i, h(\vec{x}^i)) = \sum_{i=1}^N (h(\vec{x}^i) - y^i) \phi(\vec{x}^i)$$

BGD

- 1 $\vec{w} = \vec{0}$
- 2 paravljaj do konvergencije
- 3 $\Delta \vec{w} = \vec{0}$
- 4 za $i = 1, \dots, N$
- 5 $h = g(\vec{w}^T \phi(\vec{x}^i))$
- 6 $\Delta \vec{w} = (h - y^i) \phi(\vec{x}^i)$
- 7 $\vec{w} += m \cdot \Delta \vec{w}$ - Anjelski preračun
- 8 $\vec{w} += m \cdot \Delta \vec{w}$

SGD

- 1 $\vec{w} = \vec{0}$
- 2 paravljaj do konvergencije
- 3 sluč permutiraj D
- 4 za $i = 1, \dots, N$
- 5 $h = g(\vec{w}^T \phi(\vec{x}^i))$
- 6 $\vec{w} -= \eta (h - y^i) \phi(\vec{x}^i)$

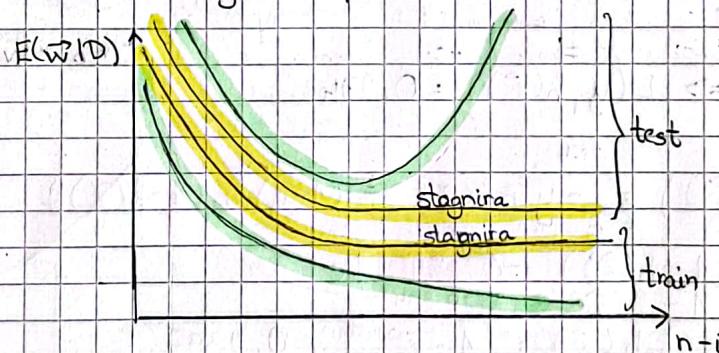
1.4.

$$\lambda = 0 \\ \lambda = 100$$

$$\left[\vec{w} = \vec{w} (1 - \eta \lambda) - \eta \sum_{i=1}^N (h(\vec{x}^i) - y^i) \phi(\vec{x}^i) \right]$$

[!! w_0 se ne regularizira $\lambda = 0$]

a) Linearno odvojiv skup



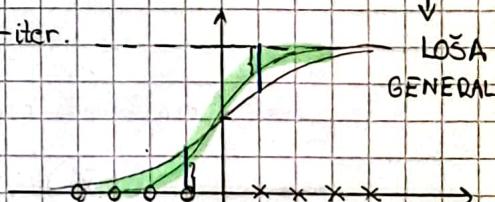
za $\lambda = 0$

- nema regularizacije
- gradijent (pogreške) NIKAD neće biti 0
- => grad. spust neće konvergirati i težine rastu (otvaranje sigmoida)

↓
LOŠA GENERAL

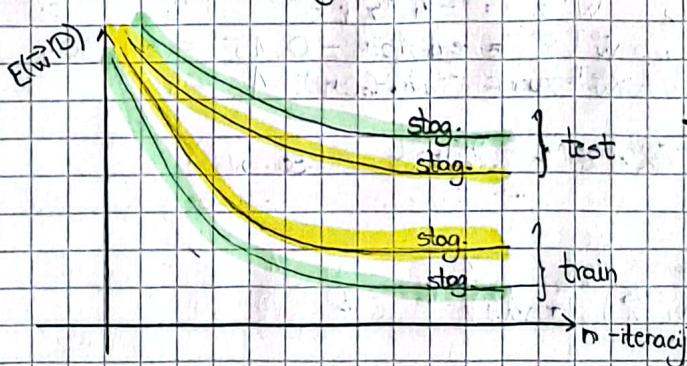
za $\lambda = 100$

- regularizacija ložnjava složene modelce (velike magnitudo težina!)
- sprječava se efekt otvarjanja sigmoida dobiti do konvergencije stagnacije \rightarrow BOJA GENERAL.



Linearno neodvojiv skup

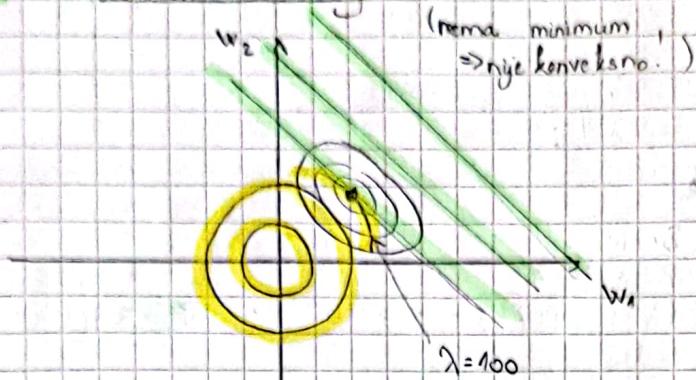
$\lambda = 0, \lambda = 100$



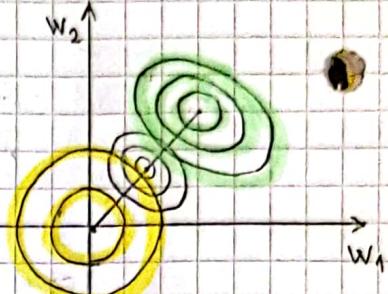
- model za linearno neodvojive primjerice NIKADA SE NEĆE preuciti (neki primjeri će uvijek biti krivo klasificirani)
- \Rightarrow dolazi do stagnacije i konvergencije i za $\lambda = 0$ i $\lambda = 100$
- $\Rightarrow \lambda = 100 \Rightarrow$ bolje gener

b)

Linearno odvojivo



Linearno neodvojivo



c)

$$\textcircled{O} = E_{\lambda}(\vec{w} | D) \text{ za } \lambda = 100$$

$= E(\vec{w} | D)$
= regulariz. izraz $\| \vec{w} \|_2$

$$\textcircled{O} = E_{\lambda}(\vec{w} | D) \text{ za } \lambda = 100$$

$=> E(\vec{w} | D)$
=> reg. izraz

I Zadaci s ispita

2.1. $w_0 = 0.15$

$$y = 0 \Rightarrow L(y, h(\vec{x})) = 0.274$$

$$y = 0 \quad L(y, h(\vec{x})) = -y \ln h(\vec{x}) - (1-y) \ln(1-h(\vec{x}))$$

$$\Rightarrow -\ln(1-h(\vec{x})) = -0.274 \quad |e$$

$$\frac{1}{1-h(\vec{x})} = e^{-0.274}$$

$$h(\vec{x}) = 1 - e^{-0.274} = 0.2397$$

[Ako promijenimo označbu: $L(y, h(\vec{x})) = -\ln h(\vec{x})$]

$$h(\vec{x}) = \sigma(-\vec{w}^T \vec{x}) = \frac{1}{1+e^{-\vec{w}^T \vec{x}}} = h(\vec{x})$$

$$h(\vec{x}) + h(\vec{x}) e^{-\vec{w}^T \vec{x}} = 1$$

$$h(\vec{x}) - \vec{w}^T \vec{x} = \ln \left[\frac{1 - h(\vec{x})}{h(\vec{x})} \right] = -1.154$$

$$\vec{w}^T \vec{x} = -1.154$$

bez w_0

$$\vec{w}^T \vec{x} = -1.154 - 0.15 = -1.304$$

bez w_0

$$\vec{w}^T \vec{x} = -2.604$$

$$\Rightarrow -\vec{w}^T \vec{x} \cdot 2 = -2.454 \quad (\text{sa } w_0)$$

$$y = 1, \text{ značajke } x_2$$

$$h(\vec{x}) = \frac{1}{1 + \exp(-\vec{w}^T \vec{x})} = \frac{1}{1 + e^{-2.454}} = 0.079$$

$$L(y, h(\vec{x})) = -\ln h(\vec{x}) = 2.538$$

(B)

$$2.2. \quad L(y, h(\vec{x})) = 1 \cdot 2 = -y \ln h(\vec{x}) - (1-y) \ln(1-h(\vec{x}))$$

$$\textcircled{1^{\circ}} \quad y=0$$

$$L = -\ln(1-h(\vec{x}))$$

$$-1 \cdot 2 = \ln(1-h(\vec{x})) \mid e$$

$$1-h(\vec{x}) = e^{-1 \cdot 2}$$

$$h(\vec{x}) = 0.698$$

$$\text{Promijeni označen } y=1 \Rightarrow L = -\ln h(\vec{x}) = 0.359$$

$$\textcircled{2^{\circ}} \quad y=1$$

$$L = -\ln(h(\vec{x}))$$

$$h(\vec{x}) = e^{-L} = 0.301$$

$$\text{Promijeni označen } y=0 \Rightarrow L = -\ln(1-h(\vec{x})) = 0.3581$$

\Rightarrow srednji gubitak je 0.36 C

$$2.3. \quad \phi(\vec{x}) = \begin{pmatrix} 1, x_1, x_2, x_1x_2 \\ 0.2, 0.5, -1, 1 \end{pmatrix}$$

L_2 - norma gradijenta gubitka za (\vec{x}, y)

$$\phi(\vec{x}) = (1, -0.5, 2, -1) \quad \vec{w}^T \phi(\vec{x}) = -4.95$$

$$\nabla_w L = (h(\vec{x}) - y) \phi(\vec{x})$$

$$h(\vec{x}) = \sigma(-\vec{w}^T \phi(\vec{x})) = 0.00703$$

$$\nabla_w L = -\underbrace{0.993}_{\approx 1} \phi(\vec{x}) = \phi(\vec{x}) = [1, -0.5, 2, -1]$$

$$\|\nabla_w L\|_2 \approx 2.5 \Rightarrow \textcircled{B}$$

2.4.

$$\begin{aligned} \vec{w}_0 &= \begin{bmatrix} 1 & -4 & 4 \end{bmatrix} \\ \vec{w}_1 &= \begin{bmatrix} 1 & -4 & 6 \end{bmatrix} \\ \vec{w}_2 &= \begin{bmatrix} 1 & -1 & 7 \end{bmatrix} \\ \vec{w}_3 &= \begin{bmatrix} 1 & -7 & 1 \end{bmatrix} \\ \vec{w}_4 &= \begin{bmatrix} 1 & -7 & -3 \end{bmatrix} \end{aligned}$$

$$E(\vec{w}_1 | D) = E(\vec{w}_2 | D) = E(\vec{w}_3 | D) \Rightarrow \text{no istoj izoljuturi!}$$

$$\gamma = 100$$

$$\frac{\gamma}{2} \|\vec{w}\|^2 = 400$$

$$\|\vec{w}\|^2 = 8$$

$$E_R(w | D) = E(w | D) + \frac{\gamma}{2} \|w\|^2$$

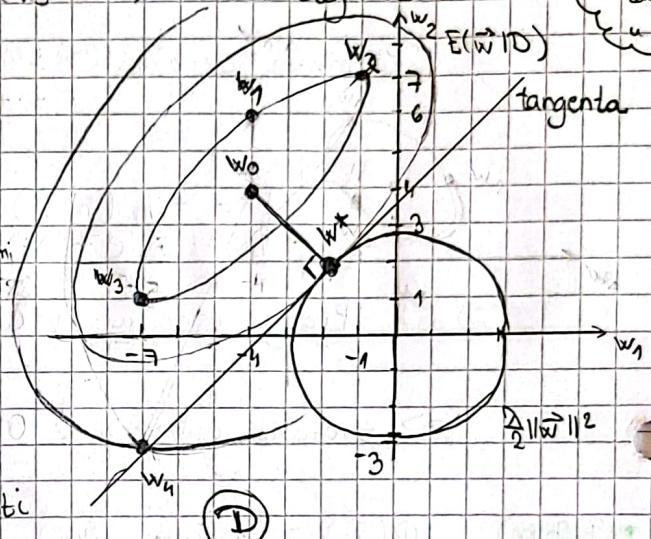
$$\text{kružnica pravim } \sqrt{8} = 2\sqrt{2}$$

w^* - optimum $E(w | D)$

\Rightarrow najbolji inicijalni odabir je onaj koji najmanje krivuda tj. onaj koji leži na tangenti kružnice $\|\vec{w}\|$

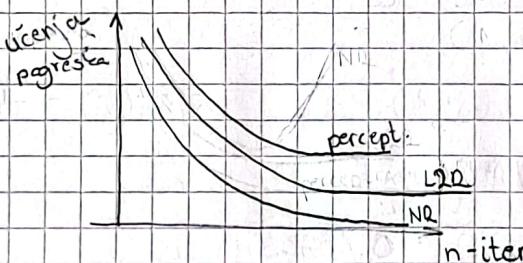
\Rightarrow už skicu vidimo da je w_4 na taj tangenti
 $\Rightarrow w_4$ je najbolji inicijalni odabir

Napomena:
 $w_0 - w^*$ mora biti normala kružnici
 $\frac{1}{2} \|\vec{w}\|^2$
 $E(w | D)$ je konveksna izgledom ista kao $E(\vec{w} | D)$, ali u koord. $w_1 \times w_2$ transformaciji $w_0 - \vec{w}^*$ je tangenčna u w^*



2.5.

• alg. perceptron se razstavio \rightarrow Linearno odgoviv skup



• NR neće konvergirati - otvrdnjavanje sigmoide

(D)

Progresa učenja modela NR asimpt. teži 0 , dok pogreška učenja LL modela na kon određenog broja iteracija stagnira

2.6.

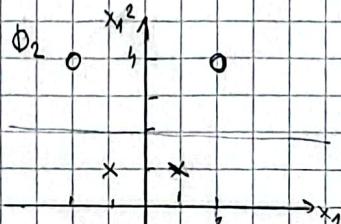
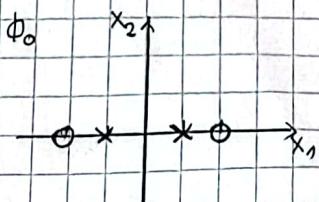
x_1	x_2	y	x_1^2	x_2^2	$x_1 x_2$
-2	0	1	4	0	0
-1	0	1	1	0	0
1	0	1	1	0	0
2	0	-1	4	0	0

konvergencija		LL	Percep.
neodgov.	neodgov.	Da	Ne
konv.	konv.	Φ₁	Da
konv.	konv.	Φ₂	Ne

$$\Phi_0(\vec{x}) = (1, x_1, x_2) \rightarrow \text{Linearno neodgovivo}$$

$$\Phi_1(\vec{x}) = (1, x_1, x_2, x_1^2, x_2^2) \rightarrow \text{Linearno odgovivo}$$

$$\Phi_2(\vec{x}) = (1, x_1, x_2, x_1 x_2) \rightarrow \text{Linearno neodgovivo}$$



• minimizator emp pogreške

(D) $LL + \Phi_2$

2.7.

SGD

$$\lambda = 1000$$
$$\eta = 0.01$$

$$\phi(\vec{x}) = (1, x_1, x_2, x_1 x_2)$$

$$\vec{w} = [0.2 \ 0.5 \ -1.1 \ 2.7]$$
$$\Delta w_1 = ?$$

$$(\vec{x}, y) = ((-1, 2), 1)$$

$$\nabla_w L = (h(\vec{x}) - y) \phi(\vec{x})$$

$$h(\vec{x}) = \sigma(-\vec{w}^T \phi(\vec{x})) = \sigma(7.9) = 3.71 \cdot 10^{-6}$$
$$\phi(\vec{x}) = (1, -1, 2, -2)$$

$$\nabla_w L = \underbrace{(3.71 \cdot 10^{-6} - 1)}_{\approx -1} \phi(\vec{x}) = [-1, 1, -2, 2]$$

$$\begin{aligned} \bar{w}_1 &= w_1 / (1 - \lambda \eta) - \eta \Delta_w L[1] \\ &= 0.5 (1 - 10) - 0.01 \cdot 1 \\ &= -4.51 \end{aligned}$$

$$\Delta w_1 = -4.51 - 0.5 = -5.01 \Rightarrow \textcircled{B}$$

6. Logistička regresija

Strojno učenje 1, UNIZG FER, ak. god. 2023./2024.

Jan Šnajder, vježbe, v2.7

1 Zadatci za učenje

1. [Svrha: Znati definirati model logističke regresije. Razumjeti izvod funkcije pogreške unakrsne entropije i pripadne funkcije gubitka. Shvatiti zašto je ta funkcija gubitka unakrsne entropije prikladna za klasifikaciju, dok funkcija kvadratnog gubitka to nije.]
 - (a) Definirajte poopćeni linearni model. Koja je svrha aktivacijske funkcije?
 - (b) Definirajte model logističke regresije. Zašto je sigmoidna (logistička) funkcija prikladan odabir za aktivacijsku funkciju?
 - (c) Izvedite pogrešku unakrsne entropije $E(\mathbf{w}|\mathcal{D})$ kao negativan logaritam vjerojatnosti oznaka svih primjera iz skupa za učenje prema hipotezi s težinama \mathbf{w} .
 - (d) Napišite funkciju gubitka unakrsne entropije i nacrtajte njezin graf. Koliki je najveći a koliki najmanji mogući gubitak?
 - (e*) Pretpostavimo da su oznake $y \in \{-1, +1\}$ umjesto $y = \{0, 1\}$. Reformulirajte funkciju gubitka unakrsne entropije $L(y, h(\mathbf{x}))$ tako da koristi takve oznake te da vrijedi $L(y, 0) = 1$ (kako bi funkcija bila kompatibilna s ostalim funkcijama gubitka koje smo radili).
 - (f) Nacrtajte graf funkcije gubitka $L(y, h(\mathbf{x}))$ u ovisnosti o udjelu pogrešne klasifikacije $y\mathbf{w}^T\phi(\mathbf{x})$, i to za: gubitak 0-1, kvadratni gubitak i logistički gubitak iz (e). Na temelju skice, odgovorite:
(i) zašto je logistički gubitak dobar za klasifikaciju, a kvadratni gubitak to nije?; (ii) nanose li ispravno klasificirani primjeri ikakav gubitak?; (iii) možemo li reći da je logistički gubitak konveksni surrogat gubitka 0-1, i što to znači?
2. [Svrha: Prisjetiti se definicije konveksnosti funkcije. Razumjeti da konveksnost i unimodalnost nisu jedno te isto.]
 - (a*) Formalno definirajte kada je funkcija $f : \mathbb{R}^n \rightarrow \mathbb{R}$ konveksna.
 - (b*) Funkcija f je kvazikonveksna (ili unimodalna) akko je njezina domena $\text{dom } f$ konveksna te ako za svaki $\mathbf{x}, \mathbf{y} \in \text{dom } f$ i $0 \leq \alpha \leq 1$ vrijedi
$$f(\alpha\mathbf{x} + (1 - \alpha)\mathbf{y}) \leq \max \{f(\mathbf{x}), f(\mathbf{y})\}.$$
Kvazikonveksnost je poopćenje konveksnosti: svaka je konveksna funkcija unimodalna, ali obrat ne vrijedi. Pokažite primjerom da obrat ne vrijedi.
 - (c) Zašto u strojnem učenju preferiramo konkveksne funkcije pogreške? Koja je veza između konveksnosti funkcije pogreške i konveksnosti funkcije gubitka?
3. [Svrha: Razumjeti gradijentni spust i potrebu za linijskim pretraživanjem. Znati izvesti gradijentni spust za logističku regresiju. Demonstrirati upoznatost s prednostima i nedostacima optimizacije drugog reda.]
 - (a) Objasnite ideju gradijentnog spusta i potrebu za linijskim pretraživanjem.
 - (b) Objasnite razliku između grupnog (*batch*) i stohastičkog gradijentnog spusta. Koja je prednost ovog drugog?
 - (c) Izrazite gradijent funkcije pogreške unakrsne entropije $\nabla E(\mathbf{w}|\mathcal{D})$ i napišite pseudokôd algoritma gradijentnog spusta (grupna i stohastička inačica).

4. [Svrha: Razumjeti kako regularizacija i linearna (ne)odvojivost utječu na gradijenti spust i na izgled funkcije pogreške u prostoru parametara.] Koristimo model L2-regularizirane logističke regresije učene algoritmom gradijentnog spusta. Iskušavamo dvije vrijednosti regularizacijskog faktora: $\lambda = 0$ i $\lambda = 100$. Razmatramo posebno linearno odvojiv i linearno neodvojiv problem.

- (a) Skicirajte pogreške učenja i ispitivanja $E(\mathbf{w}|\mathcal{D})$ u ovisnosti o broju iteracija za $\lambda = 0$ i $\lambda = 100$ te za slučaj (i) linearno odvojivih i (ii) linearno neodvojivih primjera (četiri grafikona sa po dvije krivulje).
- (b) Načinite skice izokontura funkcije neregularizirane pogreške $E(\mathbf{w}|\mathcal{D})$ i L2-regularizacijskog izraza u ravnini w_1-w_2 . Napravite dvije odvojene skice: za linearno odvojive i linearne neodvojive primjere.
- (c) Na grafikone iz prethodnoga zadatka dočrtajte izokonture L2-regulariziranih funkcija pogreške za $\lambda = 100$ i naznačite gdje se nalazi točka minimuma (w_1^*, w_2^*) . Gdje bi se nalazila točka minimuma za $\lambda = 0$?

2 Zadatci s ispita

1. (N) Na skupu označenih primjera \mathcal{D} trenirali smo model logističke regresije. Dobili smo neki vektor težina \mathbf{w} i pomak $w_0 = 0.15$. Tako naučenom modelu neki primjer \mathbf{x} , čija je oznaka u skupu primjera $y = 0$, nanosi gubitak unakrsne entropije od $L(0, h(\mathbf{x})) = 0.274$. **Koliki gubitak unakrsne entropije bi nanosio primjer \mathbf{x} kada bismo njegove značajke pomnožili sa dva i promjenili mu oznaku?**

A 4.03 B 2.54 C 7.11 D 1.19

2. (N) Na skupu \mathcal{D} označenih primjera trenirali smo model binarne logističke regresije. Naknadno smo uočili da jedan primjer iz skupa \mathcal{D} modelu nanosi razmjerno velik gubitak. Konkretno, iznos gubitka za dotični primjer je $L(y, h(\mathbf{x})) = 1.20$. Ispostavilo se da je taj primjer pogrešno označen. **Koliko bi iznosio gubitak na istom ovom primjeru, ako bismo sada naknadno promjenili njegovu oznaku, ali model ostavili nepromijenjenim?**

A 0.70 B 0.28 C 0.36 D 0.52

3. (N) Model logističke regresije treniramo stohastičkim gradijentnim spustom. Primjere iz dvodimenzionskog ulaznog prostora preslikali smo u prostor značajki funkcijom

$$\phi(\mathbf{x}) = (1, x_1, x_2, x_1x_2)$$

U jednoj iteraciji treniranja modela vektor parametara jednak je

$$\mathbf{w} = (0.2, 0.5, -1.1, 2.7)$$

Koliko u toj iteraciji iznosi L_2 -norma gradijenta gubitka za primjer $(\mathbf{x}, y) = ((-0.5, 2), 1)$?

A 0.70 B 2.48 C 1.28 D 4.00

4. (P) Na primjerima iz dvodimenzionskoga ulaznog prostora treniramo L_2 -regulariziranu logističku regresiju. Neka su $\mathbf{w}_0 = (1, -4, 4)$, $\mathbf{w}_1 = (1, -4, 6)$, $\mathbf{w}_2 = (1, -1, 7)$, $\mathbf{w}_3 = (1, -7, 1)$ i $\mathbf{w}_4 = (1, -7, -3)$ vektori u prostoru parametara. Neka je $E(\mathbf{w}|\mathcal{D})$ neregularizirana pogreška unakrsne entropije na skupu za učenje \mathcal{D} . Pritom je \mathbf{w}_0 minimizator funkcije $E(\mathbf{w}|\mathcal{D})$ te vrijedi $E(\mathbf{w}_1|\mathcal{D}) = E(\mathbf{w}_2|\mathcal{D}) = E(\mathbf{w}_3|\mathcal{D})$. Napravite skicu izokontura funkcije pogreške u potprostoru $w_1 \times w_2$. Za treniranje modela koristimo gradijentni spust s linijskim pretraživanjem uz regularizacijski faktor $\lambda = 100$. Za tako naučen model vrijednost regularizacijskog izraza $\frac{\lambda}{2}\|\mathbf{w}\|^2$ jednaka je 400. Međutim, broj koraka gradijentnog spusta (broj poziva linijskog pretraživanja) ovisi o tome koliko će spust krivudati, a to ovisi o odabiru inicijalnih parametara. Kao moguće inicijalne parametre razmotrite vektore $\mathbf{w}_1-\mathbf{w}_4$. **S kojim inicijalnim parametarima će algoritam gradijentnog spusta konvergirati u najmanjem broju koraka?**

A \mathbf{w}_1 B \mathbf{w}_2 C \mathbf{w}_3 D \mathbf{w}_4

5. (P) Na skupu označenih primjera treniramo tri modela: (1) model neregularizirane logističke regresije (NR), (2) model L2-regularizirane logističke regresije (L2R) i (3) perceptron. Sva tri modela koriste istu funkciju preslikavanja u prostor značajki. Za sva tri algoritma promatramo iznos empirijske pogreške učenja kroz iteracije optimizacijskog postupka. Nakon određenog broja iteracija, algoritam perceptrona uspješno se zaustavio s rješenjem. **Kako se u ovom slučaju ponaša empirijska pogreška učenja kroz iteracije za dva spomenuta modela logističke regresije, NR i L2R?**

- A Pogreška učenja modela NR nakon određenog broja iteracija doseže nulu, dok pogreška učenja modela L2R najprije pada pa raste
- B Pogreške učenja modela NR i modela L2R dosežu nulu, ali modelu L2R za to treba više iteracija
- C Pogreške učenja modela NR i modela L2R obje stagniraju nakon određenog broja iteracija, ali modelu NR za to treba više iteracija
- D Pogreška učenja modela NR asymptotski teži nuli, dok pogreška učenja modela L2R nakon određenog broja iteracija stagnira

6. (P) Razmotrimo sljedeći skup označenih primjera:

$$\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\} = \{((-2, 0), -1), ((-1, 0), +1), ((1, 0), +1), ((2, 0), -1)\}$$

Nad ovim skupom treniramo dva modela: perceptron (P) i neregulariziranu logističku regresiju (LR). Pored toga, razmatramo tri funkcije preslikavanja:

$$\begin{aligned}\phi_0(\mathbf{x}) &= (1, x_1, x_2) \\ \phi_1(\mathbf{x}) &= (1, x_1, x_2, x_1^2, x_2^2) \\ \phi_2(\mathbf{x}) &= (1, x_1, x_2, x_1 x_2)\end{aligned}$$

Ukupno, dakle, isprobavamo šest kombinacija modela i funkcije preslikavanja. **Za koje će algoritme (model+preslikavanje) optimizacijski postupak pronaći minimizator empirijske pogreške?**

- A P+ ϕ_0 B P+ ϕ_2 C LR+ ϕ_1 D LR+ ϕ_2

7. (N) Model regularizirane logističke regresije treniramo stohastičkim gradijentnim spustom. Koristimo faktor regularizacije $\lambda = 1000$ i stopu učenja $\eta = 0.01$. Primjere iz dvodimenzijskog ulaznog prostora preslikali smo u prostor značajki funkcijom $\phi(\mathbf{x}) = (1, x_1, x_2, x_1 x_2)$. U jednoj iteraciji treniranja modela vektor parametara jednak je $\mathbf{w} = (0.2, 0.5, -1.1, 2.7)$. **Koliko u toj iteraciji iznosi promjena težine w_1 za primjer $(\mathbf{x}, y) = ((-1, 2), 1)$?**

- A -12 B -5 C -2 D +22

7. Logistička regresija II

Strojno učenje 1, UNIZG FER, ak. god. 2023./2024.

Jan Šnajder, vježbe, v1.11

1 Zadatci za učenje

1. [Svrha: Znati izvesti algoritam multinomijalne logističke regresije.]

- Definirajte funkciju $softmax$. Izračunajte $softmax(\alpha)$ za ulazni vektor $\alpha = (2, 8, 1, 5)$. Koja su dva efekta funkcije $softmax$?
- Definirajte model multinomijalne logističke regresije.
- Izvedite pogrešku modela multinomijalne logističke regresije kao negativan logaritam vjerojatnosti oznaka koje model dodjeljuje primjerima iz skupa označenih primjera.

2. [Svrha: Znati izvesti algoritam LMS poopćenih linearnih modela. Razumjeti prednosti tog algoritma.]

- Izvedite pravilo za ažuriranje težina algoritma LMS (engl. *least mean squares*) kao gradijent funkcije gubitka, i to za (i) model linearne regresije i (ii) model logističke regresije.
- Objasnite prednost algoritma LMS (odnosno stohastičkog gradijentnog spusta) nad grupnim (*batch*) gradijentnim spustom.

3. [Svrha: Uočiti zajedničkosti poopćenih linearnih modela.]

- Opišite veze između (i) modela linearne regresije, logističke regresije i multinomijalne logističke regresije, (ii) distribucija zavisne varijable y i (iii) aktivacijskih funkcija f . Što je zajedničko svim distribucijama s kojima smo dosada radili?
- Objasnite riječima ovaj izraz:

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmax}} \ln P(\mathcal{D}|\mathbf{w}) = \underset{\mathbf{w}}{\operatorname{argmin}} E(\mathbf{w}|\mathcal{D})$$

4. [Svrha: Razumjeti motivaciju za adaptivne bazne funkcije i vezu između poopćenih linearnih modela i modela neuronske mreže.]

- Objasnite što su bazne funkcije i koji je problem s fiksним baznim funkcijama.
- Definirajte poopćeni linearnim model s proizvoljnom aktivacijskom funkcijom f koji kao bazne funkcije koristi poopćene linearne modele s istom takvom aktivacijskom funkcijom. Načinite skicu takvog modela, odnosno dvoslojne neuronske mreže. Na skici naznačite komponente ulaznog vektora, težine modela, i bazne funkcije.
- Koja je prednost ovakvog modela u odnosu na (i) poopćeni model bez baznih funkcija i (ii) poopćeni model s fiksnim baznim funkcijama? Koji je nedostatak takvog modela u odnosu na poopćeni model s fiksnim baznim funkcijama?

2 Zadatci s ispita

- (N) Raspolažemo označenim skupom primjera iz triju klasa ($K = 3$) u trodimenzijskome ulaznom prostoru ($n = 3$). Na tom skupu treniramo model multinomijalne logističke regresije. Treniranje

provodimo gradijentnim spustom. U nekoj od iteracija gradijentnog spusta, matrica težina je sljedeća (stupci odgovaraju težinama za pojedine klase):

$$\mathbf{W} = \begin{pmatrix} 1 & 1 & 1 \\ 2 & 0 & -2 \\ 3 & -4 & 6 \\ -3 & 0 & 2 \end{pmatrix}$$

Jedan od primjera u skupu za učenje je primjer $\mathbf{x} = (3, 2, -1)$ s oznakom $\mathbf{y} = (0, 1, 0)$. **Koliko iznosi gubitak unakrsne entropije koji u ovoj iteraciji optimizacijskog postupka nanosi dotični primjer?**

- A 7 B 11 C 23 D 35

2. (P) Poopćeni linearni modeli mogu koristiti adaptivne bazne funkcije. Prednost toga je da ne moramo ručno definirati preslikavanje ϕ u prostor značajki, već se to preslikavanje može naučiti na temelju podataka. Rasplažemo podatcima iz $K = 3$ klase u 10-dimenzijskome ulaznom prostoru. Za taj višeklasni problem koristimo multinomijalnu logističku regresiju, ali s adaptivnim baznim funkcijama. Svaka adaptivna bazna funkcija ϕ_j parametrizirana je kao skalarni produkt vektora značajki i vektora primjera, kao što smo radili na predavanjima. Naš model definirali smo ovako:

$$h_k(\mathbf{x}) = \text{softmax}_k \left(\sum_{j=0}^3 w_{j,k} \phi_j(\mathbf{x}) \right)$$

Ovime je definirana hipoteza za klasu k . Svaka klasa ima svoju hipotezu h_k . Svaka klasa ima i svoje težine $w_{j,k}$. Međutim, bazne funkcije ϕ_j zajedničke su za sve klase (dakle, ti parametri su dijeljeni između klasa). **Koliko ukupno parametara ima ovaj model?**

- A 45 B 49 C 136 D 142

V07 - Logistička regresija II

I Podaci za učenje

1.1.

a) softmax : $\mathbb{R}^n \rightarrow \mathbb{R}^n$

$$\text{softmax}_k(\vec{x}) = \frac{\exp(x_k)}{\sum_j \exp(x_j)}$$

• 2 efekta

① normalizacija t.d.
 $\sum_k \text{softmax}_k(\vec{x}) = 1$

② vrijednosti se gneću na
 $(0, 1)$, manje vrijednosti
 još manje, a veće se
 pojačavaju

$$\begin{aligned} \vec{x} &= [2 \ 8 \ 1 \ 5] \\ \exp(2) &= e^2 \\ \exp(8) &= e^8 \\ \exp(1) &= e^1 \\ \exp(5) &= e^5 \end{aligned}$$

$$\sum_j \exp(x_j) = 3139.48$$

$$\begin{aligned} \text{softmax}_2(\vec{x}) &= 0.0024 \\ \text{softmax}_8(\vec{x}) &= 0.9495 \\ \text{softmax}_1(\vec{x}) &= 0.00087 \\ \text{softmax}_5(\vec{x}) &= 0.0473 \end{aligned}$$

$$\text{softmax}(\vec{x}) = [0.0024 \ 0.9495 \ 0.0009 \ 0.0473]$$

b) Model multinomijalne logist. regresije

$$h_k(\vec{x}; w) = \frac{\exp(\vec{w}_k^\top \Phi(\vec{x}))}{\sum_j \exp(\vec{w}_j^\top \Phi(\vec{x}))} = P(y=k | \vec{x}; w)$$

$$W = [\vec{w}_1 \ \vec{w}_2 \ \dots \ \vec{w}_K]$$

$$\hat{h}(\vec{x}) = \operatorname{argmax}_k h_k(\vec{x}; w)$$

c) $E(\vec{w} | D) = -C_n P(\vec{y} | \vec{x}; w)$

$$\begin{aligned} P(y^i | \vec{x}^i; w) &= \prod_{k=1}^K \mu_k^{y_k^i} \rightarrow \text{multinom}_{K \text{ varijabla}} \text{ označa} \\ &\stackrel{\text{I.I.D.}}{=} -C_n \prod_{i=1}^N P(y^i | \vec{x}^i; w) \\ &= -C_n \prod_{i=1}^N \prod_{k=1}^K h_k(\vec{x}^i; \vec{w}_k)^{y_k^i} \\ &= -\sum_{i=1}^N \sum_{k=1}^K y_k^i \ln[h_k(\vec{x}^i; \vec{w}_k)] \end{aligned}$$

1.2.

a) pravilo za aržur težina LMS početnih lin. modela

$$\boxed{\vec{w} = \vec{w} - \eta H_E(\vec{w})^{-1} \nabla_{\vec{w}} E(\vec{w} | D)}$$

$H_E(\vec{w})$ = Hessova matrica

$$H_{Eij} = \frac{\partial^2 E}{\partial w_i \partial w_j}$$

diag($h(x_i)(1-h(x_i))$)

$$\cdot \text{za logističku regresiju: } H_E = \Phi^T S \Phi$$

b) Prednost SGD (stoh.) nad TGD (batch)

- manje računalno intenzivnije
- pogodno za online učenje

1.3. a)

linearna reg.

model

$$\vec{w}^T \phi(\vec{x})$$

Normalna

Distribucija zavisne var. y

$$N(h(\vec{x}), \sigma^2)$$

Bernoulli

log. reg.

$$\sigma(\vec{w}^T \phi(\vec{x}))$$

$$P(y|\vec{x}, \vec{w}) = h(\vec{x})^y (1-h(\vec{x}))^{1-y}$$

MND

$$\text{softmax} \Rightarrow \frac{\exp(\vec{w}_e^T \phi(\vec{x}))}{\sum_j \exp(\vec{w}_j^T \phi(\vec{x}))}$$

Multinom.

$$P(y|\vec{x}, \vec{w}) = \prod_{i=1}^k P_{y_i}(\vec{x})^{y_i}$$

$$h(\vec{x}; \vec{w}) = f(\vec{w}^T \phi(\vec{x})).$$

distribucije eksponencijalne familije

b)

$$\vec{w}^* = \underbrace{\operatorname{argmax}_{\vec{w}} \mathbb{P}(D|\vec{w})}_{\text{optim. param.}} = \underbrace{\operatorname{argmin}_{\vec{w}} E(\vec{w}|D)}_{\text{maksimizacija cog. izglednosti}}$$

minimizacija emp. pogreške

1.4.

a)

$$\phi: \mathbb{R}^n \rightarrow \mathbb{R}^{m+1}$$

$$\phi(\vec{x}) = (1, \phi_1(\vec{x}), \dots, \phi_m(\vec{x}))$$

$\{\phi_1(\vec{x}), \dots, \phi_m(\vec{x})\}$ = m baznih fja

• problem fiksnih baznih funkcija

= njihov broj i oblik unaprijed određen

= ne znamo unaprijed koje su dobre bazne fje

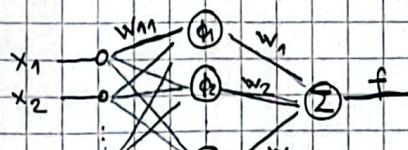
za naš problem

b)

$$h(\vec{x}; \vec{w}) = f(\vec{w}^T \phi(\vec{x})) = f\left(\sum_{j=0}^m w_j^{(2)} f\left(\sum_{i=0}^n w_{ji}^{(1)} x_i\right)\right)$$

$$\cdot f(\vec{w}^T f(\vec{w}^{(1)} \vec{x}))$$

$$\phi_j(\vec{x})$$



c) Prednosti

i) u odnosu na poopćeni model bez baznih fja
 - omogućivanje nelinearne granice

ii) u odnosu na poopćeni model s fiksnim baznim fja
 - bazne fje prilagodive, t.j. ϕ se takođe uči iz podataka

Nedostatak

- složeniji postupak optimizacije \rightarrow NEKONVEKSNOST FJE POGREŠ
- veća mogućnost pretreniranosti modela

II Zadaci s ispita

2.1.

$$K = 3$$

$$3d \text{ ulaz. prost. } n = 3$$

MNR

$$W = \begin{bmatrix} 1 & 1 & 1 \\ 2 & 0 & -2 \\ 3 & -4 & 6 \\ -3 & 0 & 2 \end{bmatrix} \quad \vec{x} = (3, 2, -1) \\ \vec{y} = (0, 1, 0) \quad L(\vec{y}, h(\vec{x})) = ?$$

$$L(\vec{y}^i, h(\vec{x}^i)) = - \sum_{i=1}^k y^i \ln[h_e(\vec{x}^i)] + (1-y^i) \ln[1-h_e(\vec{x}^i)]$$

$$h_e(\vec{x}^i) = \frac{\exp(\vec{w}^T \phi(\vec{x}^i))}{\sum_j \exp(\vec{w}_j^T \phi(\vec{x}^i))} \quad \underbrace{e^{16} + e^{-7} + e^5}_{\sum}$$

$$h(\vec{x}^i) = \text{softmax}(\vec{x}^T W) = \text{softmax}([16 \ -7 \ 5])$$

$$= [0.999 \dots \ 1.03 \cdot 10^{-10} \ 1.67 \cdot 10^{-5}]$$

$$\approx [1 \ 0 \ 0]$$

$$L(\vec{y}^i, h(\vec{x}^i)) = - \left[0 \cdot \ln(1) + \underbrace{1 \cdot \ln(1-1)}_{\stackrel{k=1}{\downarrow}} \right] \\ + \underbrace{1 \cdot \ln(1.03 \cdot 10^{-10})}_{\stackrel{k=2}{\downarrow}} + 0 \cdot \ln(1-1.03 \cdot 10^{-10}) \\ + 0 \cdot \ln(10^{-5}) + \underbrace{1 \cdot \ln(1-10^{-5})}_{\stackrel{k=3}{\downarrow}} \\ = - \ln(1.03 \cdot 10^{-10}) = 22.996 \times 23, \quad \text{(c)}$$

2.2.

$K = 3$

10 dim učizni $\Rightarrow n = 10$

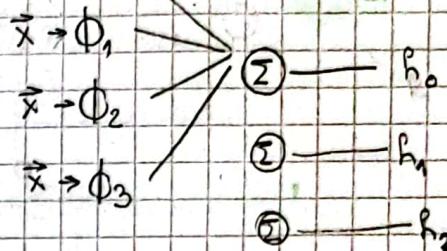
sloborno težina i parametri

$$h_k(\vec{x}) = \text{softmax}_k\left(\sum_{j=0}^2 w_{jk} \phi_j(\vec{x})\right)$$

param.

$$\begin{array}{l} w_{jk} \Rightarrow 4 \cdot 3 = 12 \\ \text{težine } \phi_j \Rightarrow 11 \cdot 3 = 33 \\ \hline \text{(A)} \quad 45 \\ \text{parametara} \end{array}$$

$\vec{x} \rightarrow \phi_0$ ovo nije adaptivno
 \rightarrow to je dummy 1



8. Stroj potpornih vektora

Strojno učenje 1, UNIZG FER, ak. god. 2023./2024.

Jan Šnajder, vježbe, v3.2

1 Zadatci za učenje

1. [Svrha: Razumjeti izvod algoritma stroja potpornih vektora.]

- Definirajte, korak po korak, problem maksimalne margine (tvrdna marge).
- Definirajte problem kvadratnog programiranja, pripadnu Lagrangeovu funkciju te dualnu Lagrangeovu funkciju i pripadne uvjete KKT. Obrazložite svaki uvjet KKT.
- Definirajte, korak po korak, dualni problem maksimalne margine te pripadne uvjete KKT koji vrijede u točki rješenja.
- Koje su prednosti formulacije problema kao dualnoga optimizacijskog problema?
- Napišite primarnu i dualnu formulaciju modela SVM.
- Objasnite što su to potporni vektori i kako znamo da oni sigurno leže na rubu marge.
- Objasnite potrebu za skaliranjem značajki kod dualne formulacije modela SVM.

2. [Svrha: Isprobati izračuna modela potpornih vektora na konkretnom numeričkom primjeru i tako bolje razumjeti formule. Razumjeti povezanost primarne i dualne formulacije problema.] Raspolažemo sljedećim primjerima za učenje:

$$\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_i = \{((0, 0), -1), ((2, 4), -1), ((4, 2), -1), ((6, 4), +1), ((6, 8), +1), ((8, 8), +1)\}$$

- Skicirajte primjere u ulaznom prostoru \mathbb{R}^2 i granicu maksimalne marge. Napišite izraz za linearni model $h(\mathbf{x})$ koji odgovara toj granici.
- Odredite širinu marge.
- U ovom slučaju potporni vektori su $\mathbf{x}^{(3)} = (4, 2)$ i $\mathbf{x}^{(4)} = (6, 4)$. Odredite vektor Lagrangeovih koeficijenata α temeljem izraza za ekspanziju težina \mathbf{w} u potporne vektore.
- Upoznajte se s formulom iz bilješke 20 iz skripte 8 te izračunajte pomak w_0 .
- Odredite klasifikaciju novog primjera $\mathbf{x}^{(7)} = (5, 6)$ na temelju dualne formulacije modela.

2 Zadatci s ispita

1. (P) Raspolažemo sljedećim skupom označenih primjera u dvodimenzijskome ulaznom prostoru:

$$\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\} = \{((-1, 1), -1), ((-2, -1), -1), ((2, -2), -1), ((3, 3), -1), ((3, 4), +1)\}$$

Na ovom skupu treniramo model SVM-a s tvrdom marginom. Međutim, naknadno smo utvrdili da je primjer $(3, 3)$ imao pogrešnu oznaku, pa smo to ispravili te ponovno trenirali SVM. Na ispravljenom skupu primjera dobili smo granicu između klase sa znatno širom marginom nego na početnom skupu primjera. **Koliko je nova margašira od stare?**

- [A] $3\sqrt{2}$ puta [B] $2\sqrt{5}$ puta [C] $\sqrt{26}$ puta [D] $\frac{5}{2}\sqrt{3}$ puta

2. (N) Rješavamo binarni klasifikacijski problem. Raspolažemo označenim skupom primjera. Odgovarajuća matrica dizajna je sljedeća:

$$\mathbf{X} = \begin{pmatrix} 1 & 3 & 16 & -8 & -11 \\ 1 & -5 & 4 & -8 & -7 \\ 1 & 7 & -4 & 11 & 9 \\ 1 & 15 & -20 & 25 & 25 \end{pmatrix}$$

Na ovom skupu treniramo model SVM-a s tvrdom marginom i linearnom jezgrenom funkcijom (tj. bez preslikavanja u prostor značajki). Model treniramo u primarnoj formulaciji. Za rješenje maksimalne margine dobili smo ovaj vektor težina (uključivo s težinom w_0):

$$\mathbf{w} = (+0.1370, -0.0290, +0.0194, -0.0461, -0.0388)$$

Umjesto u primarnoj formulaciji, model smo mogli trenirati u dualnoj formulaciji, pa bismo umjesto vektora težina \mathbf{w} dobili vektor dualnih parametara α , odnosno Lagrangeove multiplikatore. Prijelite se da su vektori čiji su Lagrangeovi multiplikatori veći od nule potporni vektori. Premda to nije uvijek moguće, u ovom konkretnom slučaju dualni parametri modela mogu se izvesti iz rješenja primarnog modela. Izvedite vektor dualnih parametara α . **Koliko iznosi najveća vrijednost parametra u vektoru dualnih parametara α ?** (Rezultate uspoređujte po prve tri decimale.)

- A 0.0013 B 0.0024 C 0.0045 D 0.0089

3. (N) U ulaznome prostoru dimenzije $n = 3$ trenirali smo model SVM-a s linearom jezgrom. Potporne vektore naučenog modela čine označeni primjeri $((2, -5, 15), -1)$, $((1, 8, -305), -1)$ i $((1, -6, 225), +1)$, a njima odgovarajući dualni koeficijenti su $\alpha_1 = 0.5$, $\alpha_2 = 0.8$ i $\alpha_3 = 0.9$. Treniranje smo proveli na skaliranim značajkama: svaku smo značajku x_j standardizirali primjenom transformacije $\frac{x_j - \mu_j}{\sigma_j}$, gdje su μ_j i σ_j srednja vrijednost odnosno varijanca značajke x_j u skupu označenih podataka \mathcal{D} . Parametri skaliranja su $\boldsymbol{\mu} = (15, -2, 100)$ i $\boldsymbol{\sigma} = (4, 1, 12)$. Model SVM-a koristimo za predikciju klase primjera $\mathbf{x} = (1, 2, -30)$. **Koliko će se promijeniti izlaz modela ako kod predikcije propustimo skalirati značajke primjera \mathbf{x} ?**

- A +907.43 B +541.53 C -373.22 D -739.13

4. (P) Skup za učenje čine sljedeći označeni primjeri:

$$\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_i = \{((-2, 3), 0), ((-1, 2), 0), ((0, 1), 0), ((0, 0), 0), ((1, 1), 0), ((1, -1), 1), ((2, 0), 1)\}$$

Na skupu \mathcal{D} treniramo logističku regresiju (LR) i stroj potpornih vektora s tvrdom marginom (SVM). Dodatno, treniramo model linearne regresije (LINR), gdje izlaz tog modela koristimo za klasifikaciju, tj. $h(\mathbf{x}) = \mathbf{1}\{\mathbf{w}^T \mathbf{x} \geq 0\}$. Za modele SVM i LINR umjesto oznake $y = 0$ koristimo oznaku $y = -1$. Za treniranje modela LR koristimo dovoljan broj iteracija tako da možemo pretpostaviti da je dobivena pogreška unakrsne entropije praktički jednaka nuli. Razmotrite primjer $\mathbf{x}^{(7)} = (2, 0)$. Neka je $d(m)$ udaljenost primjera $\mathbf{x}^{(7)}$ od granice između klasa dobivene modelom m . **Što od navedenog vrijedi za tu udaljenost?**

- A $d(\text{SVM}) < d(\text{LR}) < d(\text{LINR})$
 B $d(\text{SVM}) < d(\text{LINR}) < d(\text{LR})$
 C $d(\text{LR}) < d(\text{SVM}) < d(\text{LINR})$
 D $d(\text{LINR}) < d(\text{LR}) < d(\text{SVM})$

V08 - Stroj potpornih vektora

I Podaci za učenje

1.1.

a)

$$\text{MODEL: } h(\vec{x}) = \vec{w}^T \phi(\vec{x}) + w_0$$

$$\text{Oznake: } y \in \{-1, 1\}$$

$$\text{predikcija: } y = \text{sgn}(h(\vec{x}))$$

Tvrda mrgina - primjeri linearne odvojivosti

$$\begin{array}{l} y^i = 1 \Rightarrow h(\vec{x}^i) \geq 0 \\ y^i = -1 \Rightarrow h(\vec{x}^i) < 0 \end{array}$$

$$\cdot \text{predviđena udaj. : } d = \frac{|h(\vec{x})|}{\|\vec{w}\|}$$

$$\cdot \text{nepredviđena udajnost : } d = \frac{y^i |h(\vec{x}^i)|}{\|\vec{w}\|}$$

MARGINA = udajnost do najbližeg primjera

$$\text{bez } w_0 \Rightarrow \frac{1}{\|\vec{w}\|} \min_i \{y^i (\vec{w}^T \phi(\vec{x}^i) + w_0)\}$$

\Rightarrow možemo \vec{w}, w_0 skalirati tako da je $y^i h(\vec{x}) \geq 1$

$$\hookrightarrow \text{Problem: } \boxed{\underset{\vec{w}, w_0}{\text{argmax}} \frac{1}{\|\vec{w}\|} \quad \text{uz } y^i h(\vec{x}) \geq 1}$$

$$\Leftrightarrow \boxed{\underset{\vec{w}, w_0}{\text{argmin}} \frac{1}{2} \|\vec{w}\|^2 \quad \text{uz } y^i h(\vec{x}) \geq 1}$$

b) Problem kvadratnog programiranja
- minimizacija konveksne fje uz ograničenja

$$\left\{ \begin{array}{l} f(\vec{x}) \text{ minimizirati} \\ \text{ogranič. } g_i(\vec{x}) \leq 0 \quad i=1, \dots, m \\ h_i(\vec{x}) = 0 \quad i=1, \dots, p \end{array} \right.$$

$$\text{Lagrangeova fja} \quad L(\vec{x}, \vec{\lambda}, \vec{\beta}) = f(\vec{x}) + \sum_{i=1}^m \lambda_i g_i(\vec{x}) + \sum_{i=1}^p \beta_i h_i(\vec{x})$$

Dualna lagrang. fja

$$\tilde{L}(\vec{\lambda}, \vec{\beta}) = \min_{\vec{x}} L(\vec{x}, \vec{\lambda}, \vec{\beta})$$

KKT uvjeti

$$\left. \begin{array}{l} g_i(x_i) \leq 0 \\ h_i(x_i) = 0 \end{array} \right\} \text{početni uvjeti}$$

$$x_i \geq 0 \quad i=1, \dots, m$$

$$g_i(\vec{x}) = 0 \quad i=1, \dots, m \quad - \text{komplementarna slabost}$$

\rightarrow ograničenje (ne)aktivno

c) Lagrange fija (primarni problem)

$$L(\vec{w}, w_0, \vec{\lambda}) = \frac{1}{2} \|\vec{w}\|^2 - \sum_{i=1}^N \lambda_i [y_i (\vec{w}^T \vec{x}^i + w_0) - 1]$$

→ prelazak u dual

$$\tilde{L}(\vec{\lambda}) = \min_{\vec{w}, w_0} L(\vec{w}, w_0, \vec{\lambda})$$

$$\frac{\partial L}{\partial \vec{w}} = 0 \Rightarrow \vec{w} = \sum_{i=1}^N \lambda_i y_i \vec{x}^i$$

$$\frac{\partial L}{\partial w_0} = 0 \Rightarrow 0 = \sum_{i=1}^N \lambda_i y_i$$

$$\begin{aligned} \tilde{L}(\vec{\lambda}) &= \frac{1}{2} \vec{w}^T \vec{w} - \sum_{i=1}^N \lambda_i y_i \vec{w}^T \vec{x}^i - \underbrace{\sum_{i=1}^N \lambda_i y_i w_0}_{0 \cdot w_0} + \sum_{i=1}^N \lambda_i \\ &= \frac{1}{2} \sum_{i=1}^N \lambda_i y_i (\vec{x}^i)^T \sum_{j=1}^N \lambda_j y_j \vec{x}^j - \sum_{i=1}^N \lambda_i y_i (\vec{x}^i)^T \sum_{j=1}^N \lambda_j y_j \vec{x}^j + \sum_{i=1}^N \lambda_i \end{aligned}$$

$$[\tilde{L}(\vec{\lambda}) = \sum_{i=1}^N \lambda_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \lambda_i \lambda_j y^i y^j (\vec{x}^i)^T \vec{x}^j]$$

ograničenja:

$$\begin{aligned} \lambda_i &\geq 0 & i=1, \dots, N \\ \sum_{i=1}^N \lambda_i y^i &= 0 \end{aligned}$$

• u točku rješenja vrijede uvjeti KKT:

$$\begin{array}{ll} (1) & \lambda_i \geq 0 \\ (2) & y_i (\vec{w}^T \vec{x}^i + w_0) \geq 1 \\ (3) & \lambda_i (y^i h(\vec{x}^i) - 1) = 0 \end{array} \quad i=1, \dots, N$$

→ potparni vektori $y^i h(\vec{x}^i) = 1$

d) Prednosti formulacije dualnog problema.

└ n+1 primarnih varijabli \Rightarrow N dualnih varijabli
- u nekim slučajevima smanjena računalna složnost

└ mogućnost koristenja jednostavnog trika

└ koristenje optimizacijskog postupka SMO
 $\Rightarrow O(N^2)$ umjesto $O(n^3)$

$$h(\vec{x}) = \underbrace{\vec{w}^T \vec{x} + w_0}_{\text{primarno}} = \underbrace{\sum_{i=1}^N \lambda_i y^i (\vec{x}^T (\vec{x}^i))}_{\text{dualno}} + w_0$$

f) Potporni vektori su rprumjeri iz užasnog skupa za koje vrijedi

$$y^i h(\vec{x}^i) = 1$$

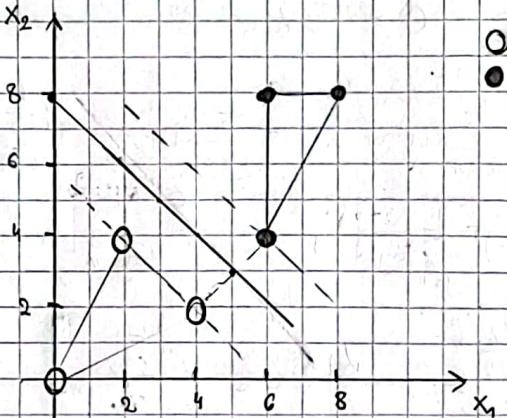
→ oni sigurno leže na marginu jer smo u izvodu problema skalirali težine upravo tako da to vrijedi

g) Skaliranje znacičajki u dualnoj formulaciji pojednostavljuje analizu problema i osigurava da će sve primjere vrijediti

$$y^i h(\vec{x}^i) \geq 1$$

1.2.

a)



$$\textcircled{O} : y = -1 \\ \textcircled{\bullet} : y = 1$$

$$h(\vec{x}) = x_1 + x_2 - 8$$

Za potp. vektorc

$$y^i h(\vec{x}^i) = 1$$

$$(6,4), 1$$

$$1 \cdot (6+4-8) = 2 = 1$$

↳ gornje težine L: 2

$$h(\vec{x}) = \frac{1}{2}x_1 + \frac{1}{2}x_2 - 4$$

b) Širina marge

$$\frac{1}{\|\vec{w}\|} = \frac{1}{\sqrt{\frac{1}{4} + \frac{1}{4}}} = \frac{1}{\sqrt{\frac{1}{2}}} = \sqrt{2}$$

bez w_0 !!!

$$\text{ili } d = \frac{h(\vec{x}^3)}{\|\vec{w}\|} = -\sqrt{2}$$

opet bez w_0

c)

$$\vec{w} = \sum_{i=1}^N \alpha_i y^i \vec{x}^i$$

$$\text{potporni vektori: } \begin{aligned} \vec{x}^3 &= \begin{bmatrix} 4 & 2 \end{bmatrix} & y^3 &= -1 \\ \vec{x}^4 &= \begin{bmatrix} 6 & 4 \end{bmatrix} & y^4 &= 1 \end{aligned}$$

$$\alpha_3 \cdot (-1) \cdot \begin{bmatrix} 4 \\ 2 \end{bmatrix} + \alpha_4 \cdot 1 \cdot \begin{bmatrix} 6 \\ 4 \end{bmatrix} = \begin{bmatrix} 1/2 \\ 1/2 \end{bmatrix}$$

$$\begin{cases} -4\alpha_3 + 6\alpha_4 = 1/2 \\ -2\alpha_3 + 4\alpha_4 = 1/2 \end{cases} / \cdot (-2)$$

$$\begin{cases} -4\alpha_3 + 6\alpha_4 = 1/2 \\ 4\alpha_3 - 8\alpha_4 = -1 \end{cases}$$

$$-2\alpha_4 = -\frac{1}{2} \rightarrow \begin{bmatrix} \alpha_3 = 0.25 \\ \alpha_4 = 0.25 \end{bmatrix}$$

d) $w_0 = ?$

$$y^i (\sum \alpha_j y^j \vec{x}^j + w_0) = 1$$

$$w_0 = y^i - \sum \alpha_j y^j (\vec{x}^j)^T \vec{x}^i$$

$$w_0 = \frac{1}{|S|} \sum_{i \in S} y^i - \sum_{j \in S} \alpha_j y^j (\vec{x}^j)^T \vec{x}^i$$

$$w_0 = \frac{1}{2} \left(-1 - \left(\frac{1}{4}(-1) \begin{bmatrix} 4 \\ 2 \end{bmatrix} \begin{bmatrix} 4 & 2 \end{bmatrix} \right. \right. \\ \left. \left. + \frac{1}{4} \cdot 1 \begin{bmatrix} 6 \\ 4 \end{bmatrix} \begin{bmatrix} 6 & 4 \end{bmatrix} \right) \right)$$

$$+ 1 - \left(\frac{1}{4}(-1) \begin{bmatrix} 6 \\ 4 \end{bmatrix} \begin{bmatrix} 4 & 2 \end{bmatrix} \right)$$

$$+ \frac{1}{4} \cdot 1 \begin{bmatrix} 6 \\ 4 \end{bmatrix} \begin{bmatrix} 6 & 4 \end{bmatrix} \right)$$

$$= \frac{1}{2} (-1 - (-5 + 8) + 1 - (-8 + 16))$$

$$= \frac{1}{2} (-1 - 3 + 1 - 5)$$

$$= -\frac{8}{2} = -4$$

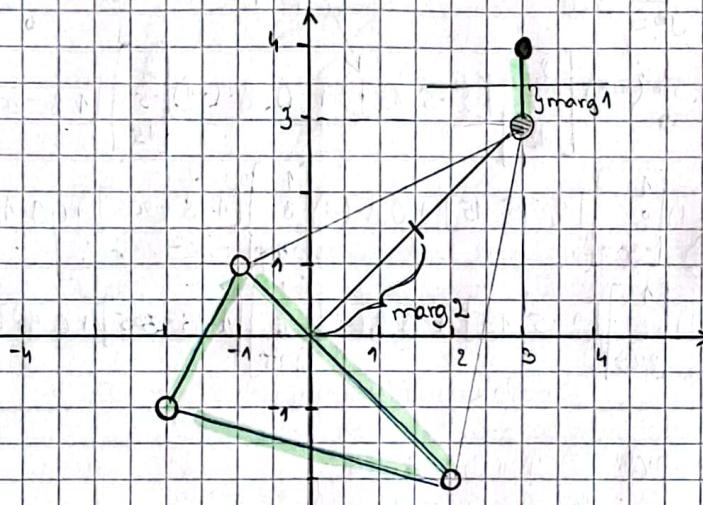
$\boxed{w_0 = -4}$

c) $\vec{x}^* = (5, 6)$

$$\begin{aligned}
 h(\vec{x}^*) &= w_0 + \sum \lambda_i y_i (\vec{x})^\top \vec{x} \\
 &= -4 + \left(\frac{1}{4} \cdot (-1) \begin{bmatrix} 5 \\ 6 \end{bmatrix} [4, 2] + \frac{1}{4} \cdot 1 \begin{bmatrix} 5 \\ 6 \end{bmatrix} [6, 4] \right) \\
 &= -4 + (-8 + 13.5) \\
 &= 1.5 \Rightarrow \text{sgn}(h(\vec{x})) = 1
 \end{aligned}$$

II Zadaci s ispita

2.1.



$$\text{marg 2} = \frac{3\sqrt{2}}{2}$$

$$\text{marg 1} = \frac{1}{2}$$

$$\frac{\text{marg 2}}{\text{marg 1}} = \frac{\frac{3\sqrt{2}}{2}}{\frac{1}{2}} = 3\sqrt{2}$$

L \rightarrow A

$$\begin{aligned}
 \circ : y &= -1 \\
 \bullet : y &= 1 \\
 \circ : y &= \pm 1
 \end{aligned}$$

- konveksne ljuške s $((+3, 3), 1)$

2.2.

$$X = \begin{bmatrix} 1 & 3 & 16 & -8 & 11 \\ 1 & -5 & 4 & -8 & -7 \\ 1 & 7 & -4 & 11 & 9 \\ 1 & 15 & -20 & 25 & 25 \end{bmatrix}$$

$$\vec{w} = [0.137 \quad -0.029 \quad +0.0194 \quad -0.0461 \quad -0.0388]$$

$$h(\vec{x}) = \vec{w}^\top X$$

$$h(\vec{x}) = [1.156 \quad 1 \quad -1 \quad -2.81]$$

ovo će biti potporni vektori $\rightarrow \lambda_2, \lambda_3 > 0$

$$\begin{bmatrix} -0.029 \\ 0.0194 \\ -0.0461 \\ -0.0388 \end{bmatrix} = \lambda_2 \begin{bmatrix} -5 \\ 4 \\ -8 \\ -7 \end{bmatrix} - \lambda_3 \begin{bmatrix} 7 \\ -4 \\ 11 \\ 9 \end{bmatrix}$$

$$\begin{aligned}
 4\lambda_2 + 4\lambda_3 &= 0.0194 / 5 \\
 -5\lambda_2 - 7\lambda_3 &= -0.029 / 4
 \end{aligned}$$

$$\begin{aligned}
 20\lambda_2 + 20\lambda_3 &= 0.007 \\
 -20\lambda_2 - 28\lambda_3 &= -0.116
 \end{aligned}$$

$$\begin{aligned}
 \lambda_3 &= 2.375 \cdot 10^{-3} \\
 \lambda_2 &= 2.475 \cdot 10^{-3}
 \end{aligned}$$

$$\lambda_2 > \lambda_3$$

$$\lambda_2 = 0.0025$$

B

2.3.

 $n = 3$

Potporni vektori

$$\begin{aligned} &((2, -5, 15), -1) \\ &((1, 8, -305), -1) \\ &((1, -6, 225), +1) \end{aligned}$$

$$\begin{aligned} \alpha_1 &= 0.5 \\ \alpha_2 &= 0.8 \\ \alpha_3 &= 0.9 \end{aligned}$$

$$\vec{\mu} = \begin{bmatrix} 15 & -2 & 100 \\ 4 & 1 & 12 \end{bmatrix}$$

$$\vec{x} = (1, 2, -30)$$

$$\Delta h(\vec{x}) = ?$$

$$h(\vec{x}) = w_0 + \sum \alpha_i y^i (\vec{x})^T \vec{x}^i$$

$$w_0 = \frac{1}{|S|} \sum_{i \in S} y^i - \sum_{j \in S} \alpha_j y^j (\vec{x}^j)^T \vec{x}^j$$

potporne vektore! TREBA SKALIRATI!

$$\begin{aligned} &= \frac{1}{3} \left[-1 - (0.5 \cdot (-1)) \begin{bmatrix} 2 \\ -5 \\ 15 \end{bmatrix} [2 -5 15] + 0.8 \cdot (-1) \begin{bmatrix} 2 \\ -5 \\ 15 \end{bmatrix} [1 8 -305] + 0.9 \cdot 1 \begin{bmatrix} 2 \\ -5 \\ 15 \end{bmatrix} [1 -6 225] \right] \\ &= -1 - (0.5 \cdot (-1)) \begin{bmatrix} 1 \\ 8 \\ -305 \end{bmatrix} [2 -5 15] + 0.8 \cdot (-1) \begin{bmatrix} 1 \\ 8 \\ -305 \end{bmatrix} [1 8 -305] + 0.9 \cdot 1 \begin{bmatrix} 1 \\ 8 \\ -305 \end{bmatrix} [1 -6 225] \\ &+ 1 - (0.5 \cdot (-1)) \begin{bmatrix} 1 \\ -6 \\ 225 \end{bmatrix} [2 -5 15] + 0.8 \cdot (-1) \begin{bmatrix} 1 \\ -6 \\ 225 \end{bmatrix} [1 8 -305] + 0.9 \cdot 1 \begin{bmatrix} 1 \\ -6 \\ 225 \end{bmatrix} [1 -6 225] \end{aligned}$$

$$= \dots = 418.826$$

$$\vec{x}_{\text{NESK.}} = [1 \ 2 \ -30]$$

$$\vec{x}_{\text{SKAL.}} = [-3.5 \ 4 \ -10.83]$$

$$\begin{aligned} h(\vec{x}_{\text{NESK.}}) &= w_0 + \sum \alpha_i y^i (\vec{x})^T \vec{x}^i = \dots = 797.605 \\ h(\vec{x}_{\text{SKAL.}}) &= w_0 + \sum \alpha_i y^i (\vec{x})^T \vec{x}^i = \dots = -58.019 \end{aligned}$$

$$\Delta h(\vec{x}) = h(\vec{x}_{\text{SKAL.}}) - h(\vec{x}_{\text{NESK.}}) = -739.13$$

Neki sbrojevima
ne stima

(D)

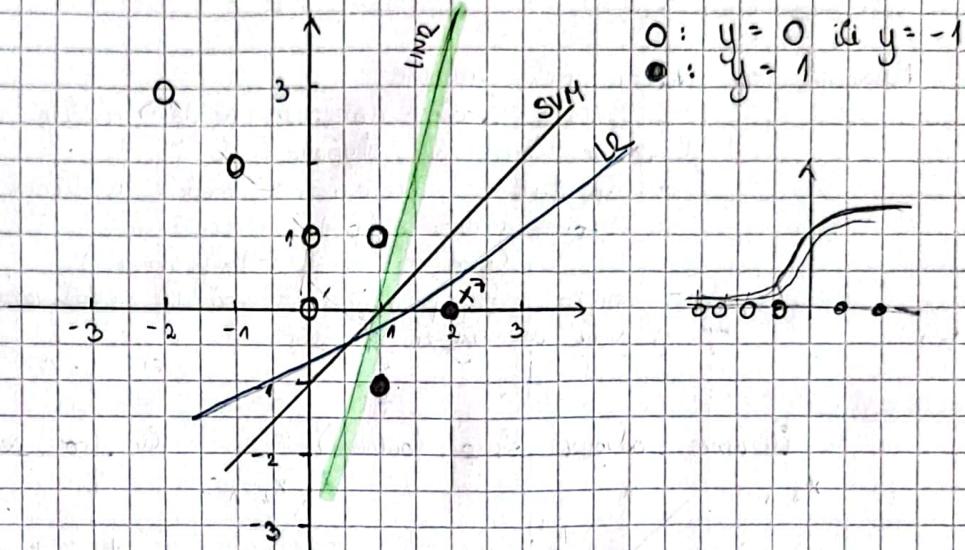
2.4.

LR - Log. reg.

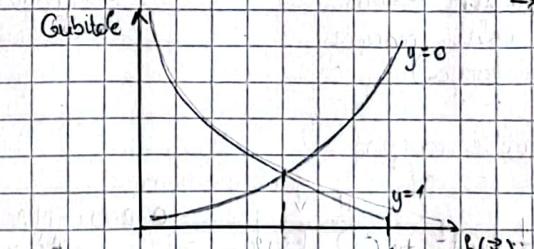
SVM

LINR - $h(\vec{x}) = 1\{\vec{w}^\top \vec{x} \geq 0\}$

$$\vec{x} = (2, 0)$$



- SVM - maksimizira minimalnu udaljenost
- LINR - ložnjava vrlo točno klasificirane primjere \rightarrow prilagođeniji primjerima klase $y=0$
- LR - želi minimizirati gubitak
(udaljuje se od klase gdje je više primjera)
 \rightarrow razšireniji klasi $y=1$



$$d(LR) < d(SVM) < d(LINR)$$

(C)

9. Stroj potpornih vektora II

Strojno učenje 1, UNIZG FER, ak. god. 2023./2024.

Jan Šnajder, vježbe, v2.8

1 Zadatci za učenje

1. [Svrha: Razumjeti potrebu za mekom marginom. Znati izvesti problem meke margine SVM-a preko Lagrangeove dualnosti.]
 - (a) Objasnite motivaciju za uvođenje meke margine. Skicirajte primjer prenaučenosti kod tvrde margine, i to za linearno odvojiv i linearno neodvojiv slučaj.
 - (b) Formulirajte problem optimizacije meke margine.
 - (c) Definirajte dualni kvadratni problem za meku marginu.
 - (d) Krenuvši od uvjeta KKT, dokažite da potporni vektori za koje vrijedi $0 < \alpha_i < C$ leže na margini, a da vektori za koje $\alpha_i = C$ leže na margini ili se nalaze unutar nje.
2. [Svrha: Znati izvesti formulaciju algoritma SVM preko gubitka zglobnice. Razumijeti funkciju pogreške SVM-a.]
 - (a) Krenuvši od problema meke margine, izvedite gubitak zglobnice.
 - (b) Napišite empirijsku pogrešku SVM-a i izrazite vezu između hiperparametara C i regularizacijskog faktora λ .
 - (c) Razmotrite zadatak 2 iz vježbi 8. Pretpostavite da je ispravna klasifikacija primjera $\mathbf{x}^{(7)} = (5, 6)$ iz (e) dijela zadatka negativna. Koliko iznosi gubitak koji primjer $\mathbf{x}^{(7)}$ nanosi SVM modelu iz tog zadatka?
 - (d) Skicirajte pogrešku učenja i pogrešku ispitivanja kao funkciju od C . Kojem području odgovara prenaučenost a kojem podnaučenost?
3. [Svrha: Razumjeti kako se gubitak zglobnice razlikuje od ostalih funkcija gubitaka koje smo razmatrali. Razumjeti kako gubitci određuju robustnost klasifikacijske granice.] Raspolažemo sljedećim primjerima za učenje:
$$\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_i = \{((1, 1), 0), ((0, 2), 0), ((2, 3), 0), ((3, 1), 1), ((4, 3), 1)\}.$$
 - (a) Skicirajte funkcije gubitka L kao funkcije od $y\mathbf{w}^T\phi(\mathbf{x})$ za gubitak (1) linearne regresije, (2) perceptron, (3) logističke regresije i (4) stroja potpornih vektora.
 - (b) Pozivajući se na skice funkcija gubitka, skicirajte predvidive hipoteze ova četiri algoritma.
 - (c) Načinite skicu kao za prethodni zadatak, ali za skup podataka u koji je dodan primjer $((8, 1), 1)$. Komentirajte razliku u odnosu na prethodnu skicu.
 - (d) Pokušajte odgovoriti: zašto algoritam SVM-a često daje rijetke modele, unatoč tome što zapravo koristi L2-regularizaciju, za koju je poznato da ne rezultira rijetkim modelima? (Pomoć: usporedite gubitak zglobnice i gubitak logističke regresije.)

2 Zadatci s ispita

1. (P) Razmatramo sljedeći skup označenih primjera u dvodimenziskome ulaznom prostoru:

$$\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_i = \{((0, 0), -1), ((-1, -1), -1), ((1, 3), +1), ((2, 2), +1), ((3, -1), +1)\}$$

Na ovom skupu treniramo model SVM-a, i to model s tvrdom marginom te model s mekom marginom sa $C = 1$. Kod modela s mekom marginom za dualne koeficijente vrijedi $\alpha_1 = 1$, $\alpha_2 > 0$, $\alpha_3 > 0$, $\alpha_4 > 0$ i $\alpha_5 = 1$. Skicirajte tvrdu i meku marginu u ulaznomet prostoru. **Koliko je meka margina veća od tvrde margine?**

- [A] $\frac{4}{5}\sqrt{10}$ puta [B] $\frac{3}{5}\sqrt{10}$ puta [C] $\frac{2}{5}\sqrt{10}$ puta [D] $\frac{1}{6}\sqrt{2}$ puta

2. (N) Raspolažemo sljedećim skupom označenih primjera u trodimenziskome ulaznom prostoru:

$$\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\} = \{((-1, 3, 6), -1), ((-4, 4, 4), -1), ((-2, 4, 1), +1)\}$$

Na ovom skupu primjera treniramo model SVM-a s linearom jezgrenom funkcijom i sa $C = 0.01$. Postupak treniranja algoritmom SMO završio je s vektorom Lagrangeovih koeficijenata $\boldsymbol{\alpha} = (0, 0.01, 0.01)$. Iz ovoga se može izračunati da vrijedi $w_0 = -0.8$. Umjesto algoritma SMO, za optimizaciju smo mogli upotrijebiti gradijentni spust i optimirati težine u primarnoj formulaciji problema. U tom slučaju koristili bismo empirijsku pogrešku SVM-a definiranu kao L2-regularizirani gubitak zglobnice. Međutim, tu pogrešku možemo izračunati i naknadno, nakon što smo naučili model. **Koliko iznosi empirijska pogreška ovog SVM-a na skupu primjera \mathcal{D} ?**

- [A] 1.935 [B] 33.935 [C] 1.135 [D] 33.135

Vog - Stroj potpornih vektora II

I Zadaci za učenje

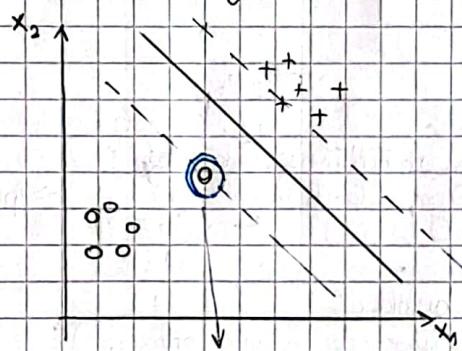
1.1.

a) Motivacija za uvođenje mke marge

- SVM s tvrdom marginom ne daje nikakvo rješenje za linearne neodgovarajuće skupove
- spriječavamo prenaučenost modela u slučaju kada u latnim podacima postoji mnoštvina (šira marga - bolji gener.)
- spriječavamo prenaučenost modela uzrokovano prešlikovanjem u već dimenzijalnom prostoru

Prenaučenost.

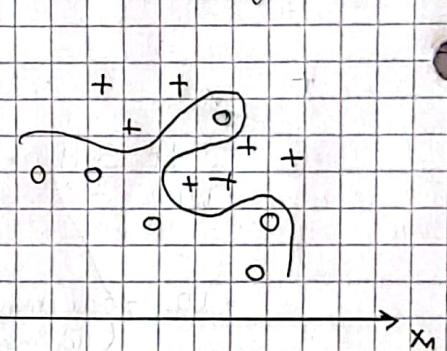
Linearne odgovarajuće skup (outliers)



outlier \Rightarrow šum

\rightarrow uzrokuje usku marginu
(loša generalizacija)

Linearne neodgovarajuće skup



\rightarrow prešložen model (prenaučen!)

b) Problem optimizacije mke marge

$$\underset{\vec{w}, w_0}{\operatorname{argmin}} \left\{ \frac{1}{2} \|\vec{w}\|^2 + C \sum_{i=1}^N \xi_i \right\}$$

užazek u marge
OGRAĐENJA

$$y_i (\vec{w}^\top \vec{x}_i + w_0) \geq 1 - \xi_i \quad \forall i = 1, \dots, N$$

$C = \frac{1}{\lambda} > 0 \Rightarrow$ C određuje kompromis između veličine marge i ložne

\Rightarrow veći C = veća složnost modela i veća ložna

c) Društveni kvadr. problem mke marge:

$$\underset{\vec{w}, w_0}{\operatorname{argmax}} \sum_{i=1}^N \xi_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \xi_i \xi_j y_i y_j (\vec{x}_i)^\top \vec{x}_j$$

isto kao za tvrdu
ALI RAZLIČ. OGRIJČ.

Ograničenja

$$\textcircled{1} \quad 0 \leq \xi_i \leq C$$

$$\textcircled{2} \quad \sum_{i=1}^N \xi_i y_i = 0$$

$$\frac{\partial L}{\partial \xi_i} = 0$$

kompozit.

d)

Potporni vektori
 $0 < \xi_i < C$

$\xi_i = C$ unutar marge

d) BiL. 3, SVM II : Vrijedi $\xi_i = C - \beta_i$ i $\beta_i \xi_i = 0$

$$\xi_i < C \Rightarrow \beta_i > 0 \Rightarrow \xi_i = 0$$

$$\xi_i = C \Rightarrow \beta_i = 0 \Rightarrow \xi_i = 0$$

1.9

a)

- ispravna strana margeine.

$$y^i h(\vec{x}^i) \geq 1 \quad \leftarrow \text{ove primjere ne ložnjavamo}$$

- primjere unutar margeine su sa pogrešne strane margeine
ložnjavamo s $\xi_i = |y_i - h(\vec{x}^i)|$
 $= 1 - y^i h(\vec{x}^i)$

\Rightarrow funkcija gubitka prema tome je

$$L(y, h(\vec{x})) = \max(0, 1 - y^i h(\vec{x}))$$

b)

$$E(\vec{w} | D) = \sum_{i=1}^N \max(0, 1 - y^i h(\vec{x}^i)) + \frac{\lambda}{2} \|\vec{w}\|^2$$

\rightarrow empirijska pogreška SVM-a je L2-regularizirano očekivanje gubitka izobnica

- prim. opt. problem mježe margeine.

$$\underset{\vec{w}, w_0, \xi}{\operatorname{argmin}} \left\{ \frac{1}{2} \|\vec{w}\|^2 + C \sum_{i=1}^N \xi_i \right\}$$

↓
odgovara
regularizacijskom
izrazu

odgovara funkciji
gubitka

$$\lambda = \frac{1}{C}$$

• manji λ daje manje regularizirani model, tj.
složniji model

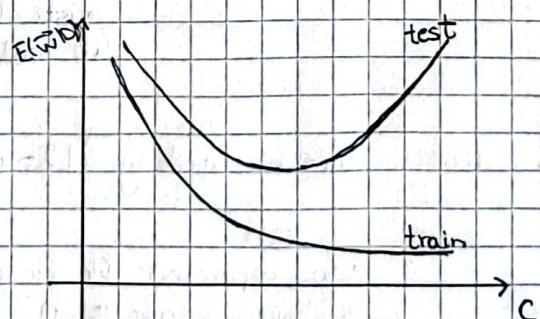
• veći C daje složniji model budući da više
ložnjavamo netočne klasifikacije / učeske u marginu

c)

$$\vec{x}^7 = \begin{bmatrix} 5 & 6 \end{bmatrix} \quad y = -1 \quad (\text{voz, 2. zadatak})$$

$$L(\vec{y}, h(\vec{x}^7)) = \max_{2.5} (0, 1 - \vec{y}^T h(\vec{x})) = \max(0, 1 - (-1) \cdot 1.5)$$

d)

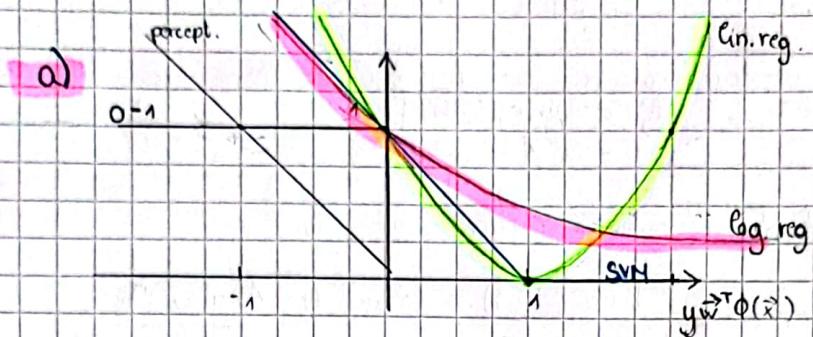


• veći $C \Rightarrow$ složniji model

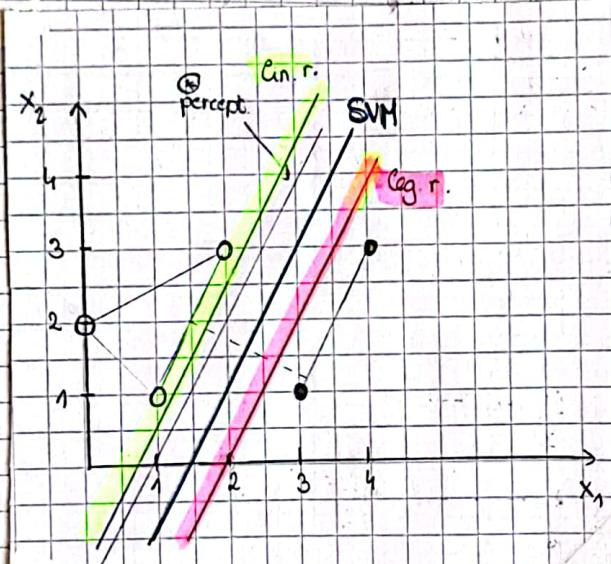
1.3.

→ pogledaj u pogodnoj verziji

$$D = \{(\vec{x}^i, y^i)\} =$$



b)

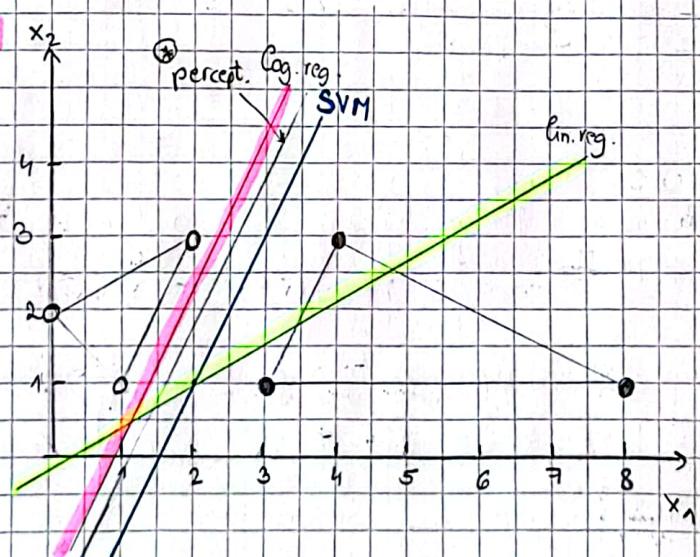


$$\circ : y = 0$$
$$\bullet : y = 1$$

Perceptron

- ovisi o pocetnim tezinama
- moze biti bilo koja hipoteza koja savrsenost klasificira D

c)



SVM = ostaje isti (isti pcp. klt.)

• Log. reg.

- bit je pomaknuta u lijevo (podjednak je broj primjera objekta i klase)
- tci min. gubitak outliera \Rightarrow ako je on daleko od granice

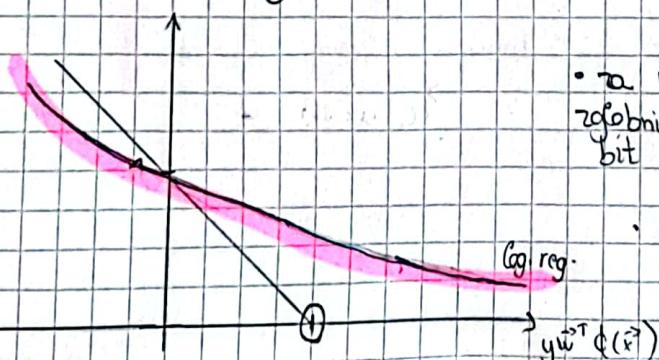
• Lin. reg.

- priklonija granica klasi $y=1$ zbog jake strucice vrijednosti $(8, 1, 1)$

• isti komentar za perceptron kao pod b)

d)

SVM daje rijetke modelce umesto ugradenij L2-regularizaciji



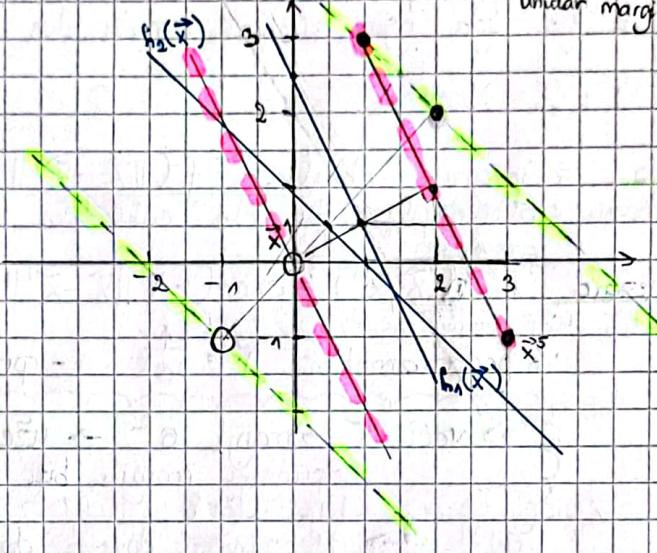
• za ispravno klasificirane primjere fja zelobnica jedra ka je 0 \Rightarrow više tezina bit će pritegnuto na 0

• kod Regist. regresije i ispravno klasificirani primjeri donose grešku \Rightarrow neće rezultirati rijetkim modelom.

II Zadaci s rešenja

2.1.

SVM₁ - tvrda margeira $\Rightarrow h_1(\vec{x})$
 SVM₂ $C=1$ $\omega_1 = \omega_5 = 1$ $\omega_2, \omega_3, \omega_4 > 0$ meka margeira
 $\Rightarrow h_2(\vec{x})$



$$\text{marg}_\text{tvr} = \frac{1}{2} \sqrt{2^2 + 1^2} = \frac{\sqrt{5}}{2}$$

$$\text{marg}_\text{meka} = \frac{1}{2} \sqrt{3^2 + 3^2} = \frac{\sqrt{18}}{2} = \frac{3\sqrt{2}}{2}$$

$$\frac{\text{marg}_\text{meka}}{\text{marg}_\text{tvr}} = \frac{\frac{3\sqrt{2}}{2}}{\frac{\sqrt{5}}{2}} = \frac{3}{5}\sqrt{10} \rightarrow \textcircled{B}$$

2.2.

$$C = 0.01 \rightarrow \lambda = \frac{1}{C} = 100$$

$$\vec{Z} = [0 \ 0.01 \ 0.01]$$

$$w_0 = -0.8$$

$$E(\vec{w} | D) = \sum_{i=1}^n \max(0, 1 - y_i h(\vec{x}_i)) + \frac{\lambda}{2} \|\vec{w}\|^2$$

$$\vec{w} = \sum_i y_i \vec{x}_i = 0.01 \cdot (-1) \begin{bmatrix} -4 \\ 4 \\ 4 \end{bmatrix} + 0.01 \cdot 1 \cdot \begin{bmatrix} -2 \\ 4 \\ 1 \end{bmatrix}$$

$$= \begin{bmatrix} 0.02 \\ 0 \\ -0.03 \end{bmatrix}$$

$$h(\vec{x}) = \vec{w}^T \vec{x} + w_0$$

$$h(\vec{x}) = [-1 \ -1 \ -0.87]$$

$$E(\vec{w} | D) = \max(0, 1 - (-1) \cdot (-1)) + \max(0, 1 - (-1) \cdot (-1)) + \max(0, 1 - 1 \cdot (-0.87)) + 50 \cdot 0.0013$$

$$= 1.87 + 0.065 = 1.935 \rightarrow \textcircled{A}$$

10. Jezgrene metode

Strojno učenje 1, UNIZG FER, ak. god. 2023./2024.

Jan Šnajder, vježbe, v1.11

1 Zadatci za učenje

1. [Svrha: Znati definirati osnovne jezgrene funkcije. Znati definirati jezgreni stroj i razumjeti razliku između jezgrenog stroja i rijetkog jezgrenog stroja.]

- (a) Definirajte jezgenu funkciju, RBF-jezgru i Gaussovou jezgru.
- (b) Je su li RBF-jezgre osjetljive na razlike u skalama značajki? Zašto?
- (c) Definirajte Mahalanobisovu udaljenost i RBF-jezgru koja koristi tu udaljenost. Navedite primjer u kojem biste koristili tu jezgru umjesto Gaussove jezgre.
- (d) Definirajte jezgreni stroj i rijetki jezgredi (vektorski) stroj. Koji od njih je parametarski a koji neparametarski algoritam i što to znači?

2. [Svrha: Isprobati preslikavanje primjera u prostor značajki primjenom Gaussova baznih funkcija. Razumjeti kako preslikavanje utječe na broj parametara i hiperparametra modela.] Raspoložemo skupom primjera:

$$\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^5 = \{((-1, -1), 0), ((0, 0), 0), ((3, -3), 1), ((-2, 1), 1), ((-4, 2), 1)\}.$$

- (a) U ulaznome prostoru skicirajte diskriminacijsku granicu $h(\mathbf{x}) = 0$ koju biste dobili logističkom regresijom uz $\phi(\mathbf{x}) = (1, \mathbf{x})$, tj. bez preslikavanja (izračun nije potreban).
- (b) Na isti skup primjera primijenite jezgredni stroj s baznim funkcijama:

$$\phi_j(\mathbf{x}) = \exp\left(-\frac{\|\mathbf{x} - \boldsymbol{\mu}_j\|^2}{2\sigma_j^2}\right).$$

Konkretno, koristite dvije bazne funkcije s parametrima $\boldsymbol{\mu}_1 = (0, 0)$, $\boldsymbol{\mu}_2 = (-3, 3)$ i $\sigma_1 = \sigma_2 = 1$. Skicirajte primjere u prostoru značajki (dimenzije ϕ_1 i ϕ_2) i granicu koju biste dobili logističkom regresijom (izračun nije potreban).

- (c) Koliko ovaj jezgredni stroj ima parametara a koliko hiperparametara? Kako biste u praksi odredili vrijednosti hiperparametara modela? Određuju li u ovom slučaju hiperparametri složenost modela? Obrazložite odgovor.

3. [Svrha: Razumjeti jezgredni trik kod SVM-a.]

- (a) Za klasifikaciju primjera u ulaznom prostoru $X = \mathbb{R}^2$ koristimo polinomijalnu jezgenu funkciju $\kappa(\mathbf{x}, \mathbf{z}) = (\mathbf{x}^T \mathbf{z} + 1)^2$. Pokažite da je za $n = 2$ jezgra κ Mercerova jezgra. Zašto je to bitno?
- (b) Izvedite pripadno preslikavanje $\phi(\mathbf{x})$ za $n = 2$. U koji će vektor u prostoru značajki efektivno biti preslikan primjer $\mathbf{x} = (2, 3)$ primjenom jezgre $\kappa(\mathbf{x}, \mathbf{z}) = (\mathbf{x}^T \mathbf{z} + 1)^2$?
- (c) Kada će broj parametara neparametarske inačice ovog modela za $n = 2$ biti veći od broja parametara njegove parametarske inačice? (U oba slučaja, parametri su vektori realnih brojeva.)
- (d) Provjerite je li u dobivenom prostoru značajki XOR-problem linearno odvojiv. Objasnite. Vrijedi li isti zaključak za jezgenu funkciju $\kappa(\mathbf{x}, \mathbf{z}) = (\mathbf{x}^T \mathbf{z})^2$?

4. [Svrha: Izvježbati izračun predikcije pomoću jezgrenog trika.] Veza između primarnih i dualnih parametara SVM-a jest $\mathbf{w} = \sum_{i=1}^N \alpha_i y^{(i)} \phi(\mathbf{x}^{(i)})$. Na skupu za učenje trenirali smo SVM s polinomijalnom jezgrenom funkcijom, $\kappa(\mathbf{x}, \mathbf{z}) = (\mathbf{x}^T \mathbf{z} + 1)^3$. Potporni vektori su $\mathbf{x}^{(1)} = (-2, 3, 5)$, $\mathbf{x}^{(2)} = (6, 4, 3)$ i $\mathbf{x}^{(3)} = (8, 8, 2)$. Prvi primjer je negativan, a druga dva su pozitivna. Lagrangeovi koeficijenti su $\alpha_1 = 0.131$, $\alpha_2 = 0.048$ i $\alpha_3 = 0.013$. Pomak je $w_0 = -0.51$. Iskoristite jezgrevi trik te odredite klasifikaciju primjera $\mathbf{x}^{(4)} = (1, 2, 3)$.
5. [Svrha: Razumjeti karakteristike Gaussove jezgre.]
- Primjenom operacija za izgradnju složenijih Mercerovih jezgri iz jednostavnijih Mercerovih jezgri, dokažite da je Gaussova jezgra Mercerova jezgra. (Pomoć: raspišite izraz $\|\mathbf{x} - \mathbf{x}'\|^2$.)
 - Kako parametar $\gamma = 1/2\sigma^2$ Gaussove jezgre utječe na složenost modela? Koji je odnos između hiperparametra C i hiperparametra γ kod SVM-a?
 - Skicirajte očekivana područja prenaučenosti i podnaučenosti modela SVM u prostoru hiperparametara $C \times \gamma$.
 - Koristimo Gaussovou jezgru uz $\gamma = 1$. Možemo li u ovom slučaju odrediti u koji vektor $\phi(\mathbf{x})$ u prostoru značajki će biti preslikan primjer \mathbf{x} ? Možemo li odrediti težine \mathbf{w} . Zašto?
 - (e*) Pročitajte [ovo](#), [ovo](#) i [ovo](#). Odgovorite: jamči li uporaba Gaussove jezgre (1) da će primjeri biti preslikani u beskonačnodimenzionalni prostor značajki, (2) savršenu linearnu odvojivost primjera za učenje u prostoru značajki, (3) empirijsku pogrešku jednaku nuli na skupu za učenje, (4) minimalnu pogrešku na ispitnome skupu? Obrazložite odgovore.
- 6*. [Svrha: Razumjeti na koji se način može kernelizirati algoritam linearne regresije.] Pročitajte poglavlje 14.4.3 iz MLPP (str. 492) te izvedite kerneliziranu inačicu linearne regresije. Koja je prednost takve formulacije algoritma linearne regresije?

2 Zadatci s ispita

1. (P) Na 1000 primjera sa 100 značajki treniramo rijetki jezgrevi stroj s Gaussovim jezgrama. Sve Gaussove jezgre imaju istu varijancu. Nakon treniranja, dobivamo model koji ima 28 prototipa. **Koliko ovaj model ima hiperparametara, koliko parametara moramo optimirati te koliko parametara ima naučeni model?**

- A Model nema hiperparametara, optimiramo 1001 parametara, a naučeni model ima 2857 parametara
- B Model ima 2800 hiperparametara, optimiramo 101 parametar, a naučeni model ima 29 parametara
- C Model ima 1 hiperparametar, optimiramo 1001 parametar, a naučeni model ima 2829 parametara
- D Model 100 hiperparametara, optimiramo 2800 parametara, a naučeni model ima 2801 parametar

2. (P) Raspolažemo sljedećim skupom označenih primjera u dvodimenzionskom ulaznom prostoru:

$$\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\} = \{((1, 1), 1), ((3, 1), 0), ((2, 3), 0), ((3, 4), 0)\}$$

Na ovom skupu treniramo jezgrevi stroj dimenzije $m = 2$ s Gaussovim baznim funkcijama, koje mjere sličnost između primjera. Za model koristimo logističku regresiju. Središta baznih funkcija su primjeri $\mathbf{x}^{(1)}$ i $\mathbf{x}^{(4)}$. Preciznost jezgre odabrana je tako da je primjer $\mathbf{x}^{(3)}$ u prostoru značajki preslikan u vektor $\phi(\mathbf{x}^{(3)}) = (1, 0.1, 0.2)$. Neka je vektor parametara modela \mathbf{w} inicijalno postavljen na $(w_0, w_1, w_2) = (0.2, 1, -1)$. **Koliko iznosi točnost tako inicijaliziranog modela na skupu \mathcal{D} ?**

- A 0
- B 1/4
- C 1/2
- D 3/4

3. (N) Rješavamo problem određivanja podrijetla pojedinih riječi u jeziku: za svaku riječ trebamo odrediti je li engleskog ($y = 1$) ili francuskog ($y = 0$) podrijetla. Problem rješavamo logističkom regresijom izvedenom kao rijetki jezgreni stroj, gdje za bazne funkcije koristimo jezgru κ nad znakovnim nizovima. Funkcija κ definirana je kao $\kappa(\mathbf{x}_1, \mathbf{x}_2) = |\mathbf{x}_1 \cap \mathbf{x}_2| / |\mathbf{x}_1 \cup \mathbf{x}_2|$, gdje su operacije unije i presjeka definirane nad skupovima slova od kojih se riječi sastoje. Na primjer, $\kappa(\text{water}, \text{eau}) = 2/6 = 0.33$. Skup za učenje je sljedeći:

$$\begin{aligned}\mathcal{D} &= \{(\mathbf{x}, y)\}_i \\ &= \{(\text{water}, 1), (\text{eau}, 0), (\text{dog}, 1), (\text{chien}, 0), (\text{paperclip}, 1), (\text{trombone}), 0), (\text{chance}, 1), (\text{hasard}, 0)\}\end{aligned}$$

Treniranjem rijetkoga jezgrenog stroja dobili smo vektor težina $\mathbf{w} = (0.5, 0, 0, 0, -3.5, 0, 1, 0, -1)$. Razmotrite primjer $(\mathbf{x}, y) = (\text{nounours}, 0)$. **Koliko iznosi gubitak modela na primjeru (\mathbf{x}, y) ?**

- A 0.359 B 0.456 C 0.552 D 0.795

4. (N) Treniramo SVM s polinomijalnom jezgrom definiranom kao:

$$\kappa(\mathbf{x}, \mathbf{z}) = (\mathbf{x}^T \mathbf{z})^2$$

Ova jezgra je Mercerova jezgra, što znači da postoji funkcija ϕ takva da $\kappa(\mathbf{x}, \mathbf{z}) = \phi(\mathbf{x})^T \phi(\mathbf{z})$. Konkretno, u slučaju dvodimenzionskoga ulaznog prostora ($n = 2$), ova jezgra odgovara preslikavanju u šesterodimenzionalni prostor. Međutim, postoji odstupanje u konkretnim koeficijentima polinoma. Razmotrite primjer $\mathbf{x} = (1, 0)$ te izračunajte $\phi_\kappa(\mathbf{x})$, koji dobivamo preslikavanjem definiranim implicitno preko jezgre, te $\phi_p(\mathbf{x})$, koji dobivamo preslikavanjem definiranim kao polinom drugog stupnja. **Koliko iznosi euklidska udaljenost između $\phi_\kappa(\mathbf{x})$ i $\phi_p(\mathbf{x})$?**

- A 0 B $2\sqrt{2}$ C 4 D $\sqrt{2}$

5. (N) Na skupu primjera za učenje iz ulaznog prostora $n = 4$ trenirali smo SVM s polinomijalnom jezgrenom funkcijom $\kappa(\mathbf{x}, \mathbf{z}) = (\mathbf{x}^T \mathbf{z} + 2)^3$. Potporni vektori i njihove oznake su sljedeći:

$$\begin{aligned}(\mathbf{x}^{(1)}, y^{(1)}) &= ((9, 30, 21), -1) \\ (\mathbf{x}^{(2)}, y^{(2)}) &= ((-11, -26, -15), -1) \\ (\mathbf{x}^{(3)}, y^{(3)}) &= ((-1, -7, -6), +1)\end{aligned}$$

Lagrangeovi koeficijenti su $\alpha_1 = 2.214 \cdot 10^{-8}$, $\alpha_2 = 3.803 \cdot 10^{-8}$ i $\alpha_3 = 6.017 \cdot 10^{-8}$. **Upotrijebite jezgreni trik da biste odredili vrijednost hipoteze $h(\mathbf{x})$ za primjer $\mathbf{x} = (3, 0, -3)$.**

- A -2.330 B -0.676 C +0.947 D +1.434

6. (N) Treniramo SVM s Gaussovom jezgrenom funkcijom. Model treniramo na skupu od $N = 5$ označenih primjera. Vektor oznaka je $\mathbf{y} = (+1, +1, -1, -1, +1)$. Euklidske udaljenosti između primjera dane su sljedećom matricom udaljenosti:

$$\mathbf{D} = \begin{pmatrix} 0.0 & 7.48 & 6.16 & 13.42 & 12.21 \\ 7.48 & 0.0 & 12.73 & 20.1 & 14.18 \\ 6.16 & 12.73 & 0.0 & 10.49 & 9.95 \\ 13.42 & 20.1 & 10.49 & 0.0 & 20.02 \\ 12.21 & 14.18 & 9.95 & 20.02 & 0.0 \end{pmatrix}$$

Treniranjem uz $C = 10$ i $\gamma = 0.0001$ za vektor dualnih parametara dobili smo $\boldsymbol{\alpha} = (10, 1.052, 10, 10, 8.948)$. **Koliko iznosi gubitak zglovnice ovako naučenog modela SVM za prvi primjer, $L(y^{(1)}, h(\mathbf{x}^{(1)}))$?**

- A 0.03 B 0.24 C 1.18 D 1.64

7. (N) Pomoću SVM-a rješavamo problem binarne klasifikacije grafova. Budući da su primjeri \mathbf{x} grafovi, koristimo SVM s jezgrenom funkcijom nad grafovima. Model treniramo na skupu od $N = 5$ označenih primjera, s vektorom oznaka jednakim $\mathbf{y} = (+1, +1, -1, -1, +1)$ i sa sljedećom jezgrenom matricom:

$$\mathbf{K} = \begin{pmatrix} 1.0 & 0.97 & -0.949 & -0.555 & -0.986 \\ 0.97 & 1.0 & -0.844 & -0.336 & -0.917 \\ -0.949 & -0.844 & 1.0 & 0.789 & 0.988 \\ -0.555 & -0.336 & 0.789 & 1.0 & 0.684 \\ -0.986 & -0.917 & 0.988 & 0.684 & 1.0 \end{pmatrix}$$

Treniranjem uz $C = 1$ za vektor dualnih parametara dobili smo $\alpha = (0, 0.754, 0.754, 1, 1)$. **Koliko iznosi gubitak zglobnice ovako naučenog modela SVM za četvrti primjer, $L(y^{(4)}, h(\mathbf{x}^{(4)}))$?**

- A 2.063 B 0.143 C 0.027 D 1.596

8. (P) Neka je $\mathcal{H}_{C,\gamma}$ model SVM-a s Gaussovom jezgrom. Hiperparametri tog modela su regulacijski faktor C i preciznost jezgre γ . Odabir modela provodimo unakrsnom provjerom i to pretraživanjem po rešetci za sljedeće vrijednosti hiperparametara:

$$C = \{2^{-5}, 2^{-4}, 2^{-3}, 2^{-2}, 2^{-1}, 1, 2^1, 2^2, 2^3, 2^4, 2^5\}$$
$$\gamma = \{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10^1, 10^2, 10^3, 10^4, 10^5\}$$

Za model sa $C = 1$ i $\gamma = 1$ utvrdili smo da je prenaučen. **Koliko modela od ovih koje ćemo još ispitati će sigurno također biti prenaučeni?**

- A 10 B 35 C 65 D 95

9. (P) Na skupu od $N = 1000$ primjera rješavamo problem višeklasne klasifikacije u $K = 4$ klase. Dvije klase imaju svaka po 400 primjera, a dvije svaka po 100 primjera. Razmatramo bismo li koristili SVM u shemi OVO ili SVM u shemi OVR. Model treniramo s jezgrenom funkcijom, no zbog ograničenja na raspoloživu računalnu memoriju moramo pripaziti da Gramova matrica ne postane prevelika. Prisjetite se da je Gramova matrična simetrična, pa je dovoljno pohraniti samo polovicu matrice (bez dijagonale). **Koji je u ovom slučaju najveći omjer veličine Gramove matrice za sheme OVO i OVR?**

- A OVO:OVR $\approx 1:3$ B OVO:OVR $\approx 1:405$ C OVO:OVR $\approx 4:5$ D OVO:OVR $\approx 32:50$

V10 - Jezgrne metode

I Zadaci za učenje

1.1. a)

Jezgrna f_{ja}
= f_{ja} koja mijeri sličnost između dva primjera

$$K(\vec{x}, \vec{x}') : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$$

RBF jezgra = jezgra oblika $f(\|\vec{x} - \vec{x}'\|)$ koja ima omotanu aktivacijsku funkciju euklidske udaljenosti dva primjera

Gaussova jezgra = $K(\vec{x}, \vec{x}') = \exp(-\Gamma \|\vec{x} - \vec{x}'\|^2)$

-jedan tip RBF jezgre

• hiperparametar $\Gamma = \frac{1}{2\sigma^2}$ - preciznost

\Rightarrow veći $\Gamma \rightarrow$ manji $\sigma^2 \rightarrow$ veće Gaussova zvono

• primjeru moraju biti blizu da bi bili slični

\hookrightarrow Često može dovesti do prenaučenosti kod SVM-a \rightarrow pretraživanje po rješetci (odabir parametara Γ i λ)

$\hookrightarrow \Gamma \rightarrow \infty \Rightarrow K(\vec{x}, \vec{x}') \rightarrow 0$

• svaki primjer u svojoj dimenziji

+ primjeri međusobno ortogonalni

\Rightarrow SAVRŠENA LINEARNA

ODVOJIVOST

b) RBF jezgre osjetljive su na razlike u skalamama značajki jer koriste euklidsku udaljenost koja sve značajke tretira jednako važnim

c) Mahalanobisova udaljenost = popravljanje euklidske udaljenosti

$$\text{RBF jezgra} \rightarrow [K(\vec{x}, \vec{x}') = \exp(-\frac{1}{2}(\vec{x} - \vec{x}')^T \Sigma^{-1} (\vec{x} - \vec{x}'))]$$

Σ = kovarijacijska matrica između značajki

\Rightarrow rješava problem osjetljivosti na skale

Pr. Kada bi u ulaznom šuplu imati značajku prihodi i deb
 \Rightarrow oboje su sličnost i deb snažno korelirane u izračunu
će ta razlika imati manju težinu nego razlika primjera u drugim značajkama

a) jezgreni stroj

= poopćeni linearni model logu za preslušavanje koristu jezgrene funkcije

$$\phi(\vec{x}) = (1, K(\vec{x}, \vec{\mu}_1), K(\vec{x}, \vec{\mu}_2), \dots, K(\vec{x}, \vec{\mu}_m))$$

$\vec{\mu}_i$ - uopriyed. odabrani referentni primjeri iz prostora primjera

$$h(\vec{x}; \vec{w}) = f(\vec{w}^T \phi(\vec{x}))$$

• rijetki jezgreni strojevi

→ kor referentne primjere $\vec{\mu}_j$ uzima se podskup učasnih primjera

$$\phi(\vec{x}) = (1, K(\vec{x}, \vec{x}^1), \dots, K(\vec{x}, \vec{x}^N))$$

→ prototipovi

↳ kako se može prenavigati
↳ ako je primjera previse
 $m=N \rightarrow 1$ regle.

⇒ rijetki jezgreni strojevi ovise o broju primjera u učasnom skupu (N)

↳ rijetki jezgreni strojevi su neparometarski algoritam

↳ jezgreni strojevi su paramet. algoritam

• neparometarski algoritam

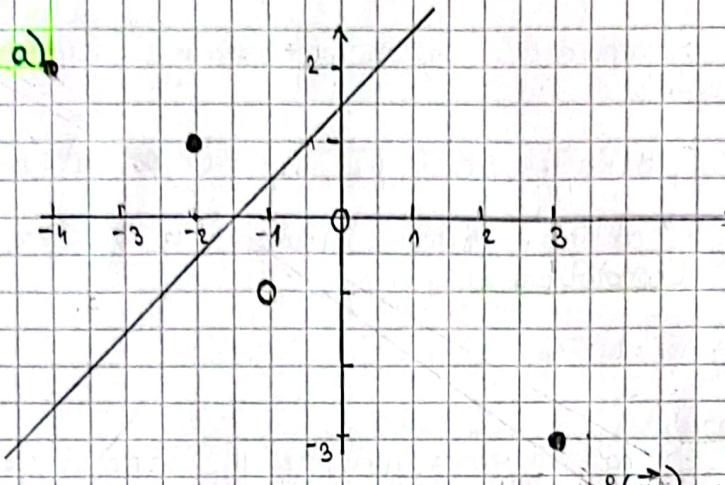
- broj parametara $\sim N$
- ne pretpostavlja se model podataka
- global. apoksim. hipoteze da pohranj. primjera

• parametarski algoritam

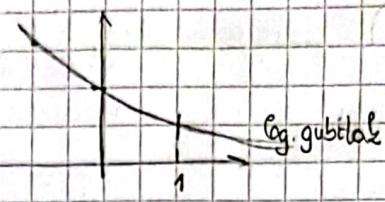
- broj parametra NE OVISI o N
- pretpostavka distribucije podataka
- primjeri \rightarrow globalno utječe na izved. hipoteze

1.2.

a)



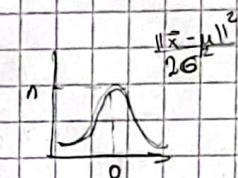
• Šekup je linearan redovjen



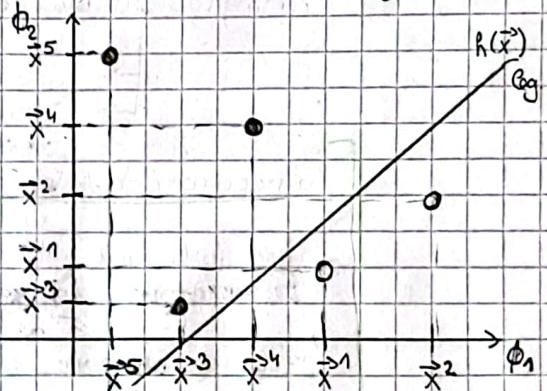
b)

$$\begin{aligned}\vec{\mu}_1 &= (0, 0) \\ \vec{\mu}_2 &= (-3, 3)\end{aligned} \quad \sigma_1 = \sigma_2 = 1$$

$$p_j = \exp\left(-\frac{1}{2\sigma^2} \|\vec{x} - \vec{\mu}_j\|^2\right) = \exp\left(-\frac{1}{2} \|\vec{x} - \vec{\mu}_j\|^2\right)$$



\vec{x}	y	x_1	x_2	ϕ_1	ϕ_2
\vec{x}_1	0	-1	-1	e^{-1}	e^{-10}
\vec{x}_2	0	0	0	e^0	e^{-9}
\vec{x}_3	1	3	-3	e^{-9}	e^{-26}
\vec{x}_4	1	-2	1	$e^{-5/2}$	$e^{-5/2}$
\vec{x}_5	1	-4	2	e^{-10}	e^{-1}



c)

$$h(\vec{x}) = \sigma(\vec{w}^\top \Phi(\vec{x}))$$

• parametri : $\vec{w} = [w_0 \ w_1 \ w_2] \rightarrow 3$

• hiperparametri : $\vec{\mu} = [\mu_1' \ \mu_2'] \quad \{\sigma_1, \sigma_2\} \quad \sigma \quad$ (ako su $\sigma_1 = \sigma_2$ uvijek ordas)

param = 3

hiperparam. = 6

• u praksi vrijednosti hiperparametara određujemo unakrsnim pravljicom

• u ovom slučaju odabir hiperparametra

- $\vec{\mu}$ ne određuje složenost modela
- σ određuje složenost

-> uže zvono \Rightarrow primjeri moraju biti bliže da budu sličniji \Rightarrow složeniji model

1.3.

a) $K(\vec{x}, \vec{z}) = (\vec{x}^T \vec{z} + 1)^2$ $n=2$

$$= (x_1 z_1 + x_2 z_2 + 1)^2$$

$$= x_1^2 z_1^2 + x_2^2 z_2^2 + 1 + 2x_1 x_2 z_1 z_2 + 2x_1 z_1 + 2x_2 z_2 +$$

$$= \begin{bmatrix} 1 \\ x_1^2 \\ x_2^2 \\ x_1 x_2 - 1 \\ x_1 \sqrt{2} \\ x_2 \sqrt{2} \end{bmatrix} \begin{bmatrix} 1 & z_1^2 & z_2^2 & z_1 z_2 \sqrt{2} & z_1 \sqrt{2} & z_2 \sqrt{2} \end{bmatrix}$$

dummy $\rightarrow \phi(\vec{x})^T \phi(\vec{z})$

uz defin. $\phi(\vec{a}) = (1, a_1^2, a_2^2, a_1 a_2 \sqrt{2}, a_1 \sqrt{2}, a_2 \sqrt{2})$

$\Rightarrow K(\vec{x}, \vec{z})$ odgovara skalarnom umnošku u 6-dim. prostoru znacajku $\Rightarrow K$ jest Mercerova jezgra

\Rightarrow ovo svrstav je bitno jer nam omogućuje da upotrijebimo jezgreni SVH t.j. iskoristimo Mercerovu jezgru umjesto skalarne umnoške $\phi(\vec{x})^T \phi(\vec{z})$ pod SVH-a

b) $\vec{x} = [2 \ 3]$

\hookrightarrow primjerenjem jezgre $K(\vec{x}, \vec{z}) = (\vec{x}^T \vec{z} + 1)^2$ \vec{x}
će biti preslikan u vektor

$$\phi(\vec{x}) = [1 \ 4 \ 9 \ 6\sqrt{2} \ 2\sqrt{2} \ 3\sqrt{2}]$$

dummy jed.

c) S obzirom da smo preslikali iz

$n=2$ u 5 dimenzionalni prostor

broj parametara neparametarske inačice ovog modela

biti će veći od broja parametara parametarske inačice za

$N > 5$

param.
neparam.

$$\frac{\vec{w}^T \phi(\vec{x})}{2 \sum_{i=1}^N K(\vec{x}_i, \vec{x}_i)} + w_0$$

d) XOR problem

\vec{x}^i	y	$\phi(\vec{x}^i) = (1, x_1^2, x_2^2, x_1 x_2 \sqrt{2}, x_1 \sqrt{2}, x_2 \sqrt{2})$
(0, 0)	0	(1, 0, 0, 0, 0, 0)
(0, 1)	1	(1, 0, 1, 0, 0, $\sqrt{2}$)
(1, 0)	1	(1, 1, 0, 0, $\sqrt{2}$, 0)
(1, 1)	0	(1, 1, $\sqrt{2}$, $\sqrt{2}$, $\sqrt{2}$, $\sqrt{2}$)

\rightarrow XOR je linearno odvojiv
uparobom pripadnog preslikavanja

\Rightarrow isto vrijedi i za jezgru $K(\vec{x}, \vec{z}) = (\vec{x}^T \vec{z} + 1)^2$ jer je polinomijalna jezgra Mercerova jezgra

pripadnor preslikavanje
 $\phi(x) = (x_1^2, x_2^2, x_1 x_2 \sqrt{2})$,

1.4.

$$K(\vec{x}, \vec{z}) = (\vec{x}^T \vec{z} + 1)^3$$

$$\vec{w} = \sum_{i=1}^N \alpha_i y^i \phi(\vec{x}^i)$$

$$\begin{aligned}\vec{x}^1 &= (-2, 3, 5) \\ \vec{x}^2 &= (6, 4, 3) \\ \vec{x}^3 &= (8, 8, 2)\end{aligned}$$

$$\begin{aligned}\alpha_1 &= 0.131 \\ \alpha_2 &= 0.048 \\ \alpha_3 &= 0.013\end{aligned}$$

$$\begin{aligned}y^1 &= -1 \\ y^2 &= 1 \\ y^3 &= 1\end{aligned}$$

$$w_0 = -0.51$$

$$f(\vec{x}^4 = [1 \ 2 \ 3]) = ?$$

$$\begin{aligned}f(\vec{x}) &= \frac{\sum \alpha_i y^i \phi(\vec{x})^T \phi(\vec{x}^i)}{K(\vec{x}, \vec{x}^i)} + w_0 \\ &= 0.131 \cdot (-1) \cdot (([1 \ 2 \ 3]^T [-2 \ 3 \ 5] + 1)^3 \\ &\quad + 0.048 \cdot (+1) \cdot ([1 \ 2 \ 3]^T [6 \ 4 \ 3] + 1)^3 \\ &\quad + 0.013 \cdot (+1) \cdot ([1 \ 2 \ 3]^T [8 \ 8 \ 2] + 1)^3 \\ &\quad - 0.51\end{aligned}$$

$$\begin{aligned}f(\vec{x}) &= -104.8 + 663.552 + 387.283 - 0.51 \\ &= -2.325 //\end{aligned}$$

1.5.

a) Gaussova jezgra je Mercerova jezgra

Dokaz

$$\begin{aligned}K(\vec{x}, \vec{y}) &= \exp(-\gamma \|\vec{x} - \vec{y}\|^2) \\ &= \exp(-\gamma (\|\vec{x}\|^2 - 2\vec{x}^T \vec{y} + \|\vec{y}\|^2)) \\ &= \underbrace{\exp(-\gamma \|\vec{x}\|^2)}_{\text{vidi str. 12}} \underbrace{\exp(2\gamma \vec{x}^T \vec{y})}_{f(\vec{x})} \underbrace{\exp(-\gamma \|\vec{y}\|^2)}_{f(\vec{y})}\end{aligned}$$

Treba vidjeti je li $K_\gamma(\vec{x}, \vec{y}) = 2\gamma \vec{x}^T \vec{y}$ Mercerova jezgra

$$K_\gamma(\vec{x}, \vec{y}) = 2\gamma \vec{x}^T \vec{y} = \vec{x}^T \sqrt{2\gamma} \cdot \vec{y} \cdot \sqrt{2\gamma} = \phi(\vec{x})^T \phi(\vec{y})$$

b) $\gamma = \frac{1}{2\sigma^2} \Rightarrow$ veća preciznost (γ) učvršćuje manju σ^2 , tj. veće Gaussovo zvono
 \Rightarrow primjeri će onda biti različitiji, tj. $K(\vec{x}, \vec{x}') \rightarrow 0$
 što odgovara povećanju složenosti modela

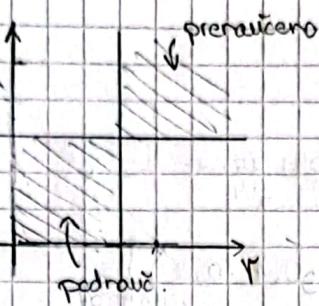
γ raste \Rightarrow raste složenost

$C = \frac{1}{\lambda} \Rightarrow$ porastom kazne (C) manje regulariziramo model, tj.
 složenost raste

C raste \Rightarrow raste složenost

• Ako odaberemo veliku preciznost moramo odabrati manji C kako bi smanjili složenost modela.

c)



d)

$T = 1$

$$\mathcal{K}(\vec{x}, \vec{x}') = \exp(-T\|\vec{x} - \vec{x}'\|^2)$$

$$= \exp(-\|\vec{x} - \vec{x}'\|^2)$$

Ne možemo odrediti preslikavanje $\phi(\vec{x})$ jer je ono definisano implicitno.

Podjedнако ne možemo odrediti ni težine jer su one dane s $\vec{w} = \sum_i \mathcal{L}_{iy} \phi(\vec{x}_i)$,

c)

Gaussova jezgra jamoči visoku dimenzionalnost.

- Gramova matica K za Gauss. jezgru je punog ranga

$\Rightarrow \phi(\vec{x})$ su linearno nezavisni \Rightarrow prostor značajki je beskonačno dim. (dim. koliko u primjera)

- jamoči i savršenu linearnu odgovarjanost
- $T \rightarrow \infty \Rightarrow K(\vec{x}, \vec{x}') \rightarrow 0 \rightarrow$ primjeri ortogonalni i normirani
- svaki primjer = 1. vrh (os) u prostoru značajki
- empirijska pogreška = 0 na skupu za učenje
 \rightarrow zbog savrš. lin odgovarjnosti

II Zadaci s ispita

2.1.

$N = 1000$

$n = 100$

nakon treniranja 28 prototipa

$$\# \text{param. nauč. mod} = 28 \text{ prototipa}, \alpha, w_0 = 28 \cdot 100 + 28 \cdot \alpha + 1 = 2829$$

$$\# \text{hiperparam} = ? = 1$$

$$\# \text{celic. param. opt.} = ? = 1000 \cdot \alpha + w_0 = 1001 \quad (\text{tražimo potp. vektore})$$

$$h(\vec{x}) = w_0 + \sum_{i=1}^N \alpha_i \mathcal{L}_{iy} \mathcal{K}(\vec{x}, \vec{x}_i)$$

$$\downarrow \exp(-T\|\vec{x} - \vec{x}_i\|^2)$$

\rightarrow sve Gauss imaju isti $T/2$ \Rightarrow 1 hiperpar.

(C)



2.2

	\vec{x}^i	y^i
1	(1, 1)	1
2	(3, 1)	0
3	(2, 3)	0
4	(3, 4)	0

$m = 2$
Gauss. jezgre
 \rightarrow log. regresija

prototipovi: $\vec{x}^1 \vec{x}^4$

$$\left\{ \phi(\vec{x}^3) = (1 \ 0.1 \ 0.2) \right\}$$

Točnost = ?

$$\vec{w} = [0.2 \ 1 \ -1]$$

$$h(\vec{x}) = \sigma(-\vec{w}^\top \phi(\vec{x}))$$

$$\begin{aligned} \phi(\vec{x}) &= (1 \ \phi_1(\vec{x}) \ \phi_2(\vec{x})) \\ \phi_1(\vec{x}) &= \exp(\vec{x}, \vec{x}^1) = \exp(-\gamma_1 \|\vec{x} - \vec{x}^1\|^2) \\ \phi_2(\vec{x}) &= \exp(\vec{x}, \vec{x}^2) = \exp(-\gamma_2 \|\vec{x} - \vec{x}^2\|^2) \end{aligned}$$

$$\phi_1(\vec{x}^3) = \exp(-\gamma_1 \|\begin{bmatrix} 2 \\ 3 \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \end{bmatrix}\|^2) = 0.1 / c_n$$

$$\begin{aligned} -\gamma_1 \|\begin{bmatrix} 1 \\ 2 \end{bmatrix}\|^2 &= c_n 0.1 \\ -\gamma_1 \cdot 5 &= c_n 0.1 \Rightarrow \underline{\gamma_1 = 0.46} \end{aligned}$$

$$\begin{aligned} \phi_2(\vec{x}^3) &= \exp(-\gamma_2 \|\begin{bmatrix} 2 \\ 3 \end{bmatrix} - \begin{bmatrix} 3 \\ 4 \end{bmatrix}\|^2) = 0.2 / c_n \\ -\gamma_2 \cdot 2 &= c_n 0.2 \Rightarrow \underline{\gamma_2 = 0.805} \end{aligned}$$

$$\begin{aligned} \phi_1(\vec{x}) &= \exp(-0.46 \|\vec{x} - \begin{bmatrix} 1 \\ 1 \end{bmatrix}\|^2) \\ \phi_2(\vec{x}) &= \exp(-0.805 \|\vec{x} - \begin{bmatrix} 3 \\ 4 \end{bmatrix}\|^2) \end{aligned}$$

$$h(\vec{x}^1) = \frac{1}{1 + \exp(-[\begin{smallmatrix} 0.2 \\ -1 \end{smallmatrix}] [\begin{smallmatrix} 1 & 1 \\ 1 & 0 \end{smallmatrix}] \begin{smallmatrix} 1 \\ 1 \end{smallmatrix})} = 0.7058 \geq 0.5 \Rightarrow 1 \checkmark$$

$$h(\vec{x}^2) = \dots = 0.5789 \Rightarrow 1 \times$$

$$h(\vec{x}^3) = \dots = 0.5249 \Rightarrow 1 \times$$

$$h(\vec{x}^4) = \dots = 0.31 \Rightarrow 0$$

Točnost: $\frac{1}{2}$

c)

2.3.

$$\text{erg. } y=1 \\ \text{fronc. } y=0$$

$$K(\vec{x}_1, \vec{x}_2) = \frac{|\vec{x}_1 \cap \vec{x}_2|}{|\vec{x}_1 \cup \vec{x}_2|} \rightarrow \text{repon sloba}$$

$$\vec{w} = [0.5 \ 0 \ 0 \ 0 \ -3.5 \ 0 \ 1 \ 0 \ -1]$$

 w_0

\vec{x}^i	y_i^i	$K(\vec{x}, \vec{x}^i)$	$h(\vec{x}) = G(-\vec{w}^T \phi(\vec{x}))$
WATER	1	$1/9$	$\{w, a, t, e, r\}$
EAU	0	$1/7$	$\{e, a, u\}$
DOG	1	$1/7$	$\{d, o, g\}$
CHIEN	0	$1/9$	$\{c, h, i, e, n\}$
PAPERCLIP	1	$1/13$	$\{p, a, p, e, r, c, l, i, p\}$
TROMBONE	0	$3/10$	$\{t, r, o, m, b, n, o, n, c\}$
CHANCE	1	$1/10$	$\{c, h, a, n, c\}$
HASARD	0	$2/9$	$\{h, a, s, r, d\}$

$$\phi(\vec{x}) = [1 \ \underline{\frac{1}{9}} \ \underline{\frac{1}{7}} \ \underline{\frac{1}{9}} \ \underline{\frac{1}{13}} \ \underline{\frac{3}{10}} \ \underline{\frac{1}{10}} \ \underline{\frac{2}{9}}]$$

logistička regres. s korn.

$$\vec{x} = \text{nourours} = \{n, o, u, r, s\} \\ y = 0$$

$$h(\vec{x}) = G(-\vec{w}^T \phi(\vec{x})) = \\ = G(-\frac{11}{90}) \\ = 0.547$$

$$L(y, h(\vec{x})) = -\ln(1 - h(\vec{x})) = 0.792$$

UNIJA - sve (sporavljajem) - presječ	⇒ (D)
--------------------------------------	-------

2.4.

D

$$K(\vec{x}, \vec{z}) = (\vec{x}^T \vec{z})^2 \\ = (x_1 z_1 + x_2 z_2)^2 \\ = x_1^2 z_1^2 + x_2^2 z_2^2 + 2x_1 x_2 z_1 z_2$$

$$\phi_K(\vec{x}) = (x_1^2, x_2^2, x_1 x_2 \sqrt{2})$$

$$\phi_P(\vec{x}) = (1, x_1, x_2, x_1 x_2, x_1^2, x_2^2)$$

$$\vec{x} = (1, 0) \quad \phi_K(\vec{x}) = (1, 0, 0) \Rightarrow (0, 0, 0, 0, 1, 0) \\ \phi_P(\vec{x}) = (1, 1, 0, 0, 1, 0)$$

prije da se
zvodiće
polapaju!

$$\|\phi_K(\vec{x}) - \phi_P(\vec{x})\| = \sqrt{1^2 + 1^2} = \sqrt{2}$$

2.5dim. ulaz. $n=4$

$$K(\vec{x}, \vec{z}) = (\vec{x}^T \vec{z} + 2)^3$$

$$\begin{bmatrix} 9 & 20 & 21 \\ -1 & -20 & -15 \\ -1 & -7 & -6 \end{bmatrix} \begin{bmatrix} y_1 \\ -1 \\ 1 \end{bmatrix} \begin{bmatrix} 2.214 \cdot 10^{-8} \\ 3.803 \cdot 10^{-8} \\ 6.017 \cdot 10^{-8} \end{bmatrix}$$

$$\vec{x} = [3, 0, -3]$$

$$h(\vec{x}) = w_0 + \sum_{i=1}^N x_i y_i K(\vec{x}, \vec{x}^i) = \dots = 0.9461$$

(C)

$$K([3, 0, -3], \vec{x}^i) \\ (-3+2)^4 = 1326.320 \\ (12+2)^4 = 38416 \\ (15+2)^4 = 83521$$

$$w_0 = ? \quad h(\vec{x}^1) = \vec{w}^T \vec{x}^1 + w_0 = -1$$

$$\vec{w} = \sum x_i y_i \vec{x}^i \quad w_0 = -1 - \vec{w}^T \vec{x}^1$$

$$\vec{w} = [1.589 \cdot 10^{-3} \\ -9.661 \cdot 10^{-8} \\ -2.555 \cdot 10^{-7}]$$

$$w_0 = -1$$

2.6.

$N=5$

SVM s Gauss.

$$\vec{y} = [+1 \quad +1 \quad -1 \quad -1 \quad +1]$$

$$K(\vec{x}, \vec{y}) = \exp(-\gamma d^2) \\ = \exp(-d^2 \cdot 10^{-4})$$

Udaljenosti

$$D = \begin{bmatrix} x^1 & x^2 \\ \vdots & \vdots \end{bmatrix} \left[\begin{array}{ccccc} 0 & 7.48 & 6.16 & 13.42 & 12.21 \\ 0 & 0 & 12.73 & 20.1 & 14.81 \\ 0 & 0 & 10.49 & 9.95 & \\ 0 & 0 & 0 & 20.02 & \\ 0 & 0 & 0 & 0 & \end{array} \right]$$

→ simetrična matrica

$$C = 10 \\ \gamma = 10^{-4}$$

$$\vec{L} = [10 \quad \underline{1.052} \quad 10 \quad 10 \quad \underline{8.948}]$$

$0 < \alpha_i \leq C \rightarrow$ potporni vektori
 $\alpha_i = C \rightarrow$ unutar marge!

$$L(y^1, h(\vec{x}^1)) = \max(0, 1 - y^1 h(\vec{x}^1)) = ? = 0.241$$

$$h(\vec{x}^1) = w_0 + \sum_{i=1}^N \alpha_i y^i K(\vec{x}, \vec{x}^i) = \dots = 0.759$$

svi primjeri unutar
i na marge

$$y^i h(\vec{x}^i) = 1 \\ w^T \vec{x}^i + w_0 = y^i \\ w_0 = y^i - \vec{w}^T \vec{x}^i$$

$$w_0 = \frac{1}{|S|} \sum_{i \in S} y^i - \sum_{j=1}^N \alpha_j y^j K(\vec{x}^j, \vec{x}^i) = \frac{1}{5} = 0.68$$

tu samo
oni koji su
na marge

tu svih primjeri

2.7. SVM, bin. Klas.

$N = 5$

$$\vec{y} = [+1, +1, -1, -1, +1]$$

$$K = \begin{bmatrix} 1 & 0.97 & -0.949 & -0.555 & -0.986 \\ 1 & 1 & -0.844 & -0.326 & -0.917 \\ 1 & 1 & 1 & 0.789 & 0.988 \\ 1 & 1 & 1 & 1 & 0.684 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

\Leftrightarrow symmetrische!

$$C = 1 \Rightarrow \lambda = \frac{1}{C} = 1$$

$$\vec{x} = [0 \underbrace{0.754}_{\text{na marginu}} \underbrace{0.754}_{\text{na marginu}} \underbrace{1}_{\text{unter margin}} \underbrace{1}_{\text{unter margin}}] \quad 0 < \lambda_i \leq C$$

$$L(y^4, h(\vec{x}^4)) = \max(0, 1 - y^4 h(\vec{x}^4)) = ? \\ = \max(0, 1 - (-1) \cdot (-0.973)) = 0.027 \Rightarrow \textcircled{C}$$

$$h(\vec{x}) = w_0 + \sum_{i=1}^N \lambda_i y^i K(\vec{x}, \vec{x}^i)$$

$$y h(\vec{x}) = 1 \\ y \left(\sum_{i=1}^N \lambda_i y^i K(\vec{x}, \vec{x}^i) + w_0 \right) = 1 \\ w_0 = y - \sum_{i=1}^N \lambda_i y^i K(\vec{x}, \vec{x}^i)$$

$$w_0 = ? \\ w_0 = \frac{1}{|S|} \sum_{i \in S} y^i - \sum_{j=1}^N \lambda_j y^j K(\vec{x}^i, \vec{x}^j)$$

$$|S| = 2 \quad i \in \{2, 3\}$$

$$\underline{i=2} \quad y^2 - \sum_{j=1}^N \lambda_j y^j K(\vec{x}^2, \vec{x}^j) = \\ = 1 - \left(0 \cdot \frac{0.754 \cdot 1 \cdot 1 + 0.754 \cdot (-1) \cdot (-0.844)}{1 \cdot (-1) \cdot (-0.326) + 1 \cdot 1 \cdot (-0.917)} + \right. \\ \left. = 1 - 0.8098 = \right. \\ \left. = 0.1906 \right.$$

$$\underline{i=3} \quad y^3 - \sum_{j=1}^N \lambda_j y^j K(\vec{x}^3, \vec{x}^j) = \\ = -1 - \left(0 \cdot \frac{0.754 \cdot 1 \cdot (-0.844) + 0.754 \cdot (-1) \cdot 1 + 1 \cdot (-1) \cdot 0.789}{1 \cdot 1 \cdot 0.988} \right. \\ \left. = -1 + 1.1914 = 0.1914 \right.$$

$$\Rightarrow w_0 = \frac{1}{2} (0.1906 + 0.1914) = 0.191$$

$$h(\vec{x}^4) = w_0 + (-1.16425) \\ = -0.9731$$

$$h(\vec{x}^4) = w_0 + \sum \lambda_i y^i K(\vec{x}^4, \vec{x}^i) = w_0 + \left(0 \cdot \frac{0.754 \cdot (-1) \cdot 0.789 + 1 \cdot (-1) \cdot 1 + 1 \cdot 1 \cdot 0.684}{0.754 \cdot (-1) \cdot 0.789 + 1 \cdot (-1) \cdot 1 + 1 \cdot 1 \cdot 0.684} \right)$$

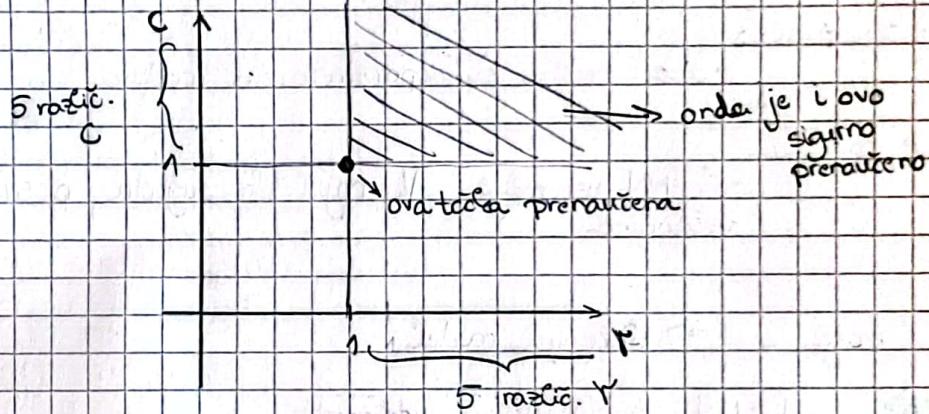
2.7. • isti kao 2.6. samo je zadana direktna jezgrena matrica, a ne euklidova udaljenost.

TDEN. POSTUPAK RJEŠAVANJA

1.8.

$$\left. \begin{array}{l} C = 1 \\ T = 1 \end{array} \right\} \text{prenaučení}$$

$$\left. \begin{array}{l} |\mathbf{C}| = 11 \\ |\mathbf{Y}| = 11 \end{array} \right\} 121 \text{ mode}$$



$$\Rightarrow 6 \cdot 6 = 36 \text{ modeCa prenaučeno}$$

$$36 - 1 = 35$$

$$\Rightarrow (1,1) \\ C=1 = \gamma$$

2.9. $N = 1000$ primitiva

$$V = 4 \text{ pcasc}$$

{ 400, 400, 100, 100 }

- SVM u OVR ili OVR shemici

maximalni omjer $\frac{OQ}{QR} = ?$

OVO Grammova matrica jezgre
dim. $N = 800 \quad 800 \times 800$

$$N = 800 \quad 800 \times 800$$

N = 500 500 x 500

$N > 900$ 200×200

0V D dim: 1000 x 1000

$$\frac{OVO}{OVR} = \frac{800 \times 900}{1000 \times 1000} - \frac{64}{100} = \frac{16}{25} = \frac{32}{50} \Rightarrow D$$

Napomena

Mogu smo računati s pola dim., ali bi se u vrijeme skratilo

11. Neparametarske metode

Strojno učenje 1, UNIZG FER, ak. god. 2023./2024.

Jan Šnajder, vježbe, v1.5

1 Zadatci za učenje

1. [Svrha: Razumjeti sličnosti i različitosti algoritma k -NN i SVM.] Napišite dualni model algoritma SVM te model algoritma k -NN. Što ova dva modela imaju zajedničko? Po čemu se algoritmi razlikuju?
2. [Svrha: Isprobati klasifikator k -NN na konkretnom primjeru. Razumjeti kako hiperparametar k i broj primjera N utječe na složenost modela.]
 - (a) Klasifikator 4-NN s euklidskom udaljenošću učen je na sljedećim primjerima iz $\mathbb{R}^3 \times \{0, 1\}$:
$$\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^6 = \{(4, 4, 0), 1), (4, 3, 1), 1), (6, 0, 2), 1), ((5, 2, 2), 0), ((5, 1, 1), 0), ((7, 2, 0), 0)\}.$$
Odredite klasifikaciju primjera $\mathbf{x}^{(1)} = (4, 2, 1)$ i $\mathbf{x}^{(2)} = (0, 3, 3)$.
 - (b) Ponovite klasifikaciju s težinskim modelom 4-NN, primjenom inverzne kvadratne jezgre.
 - (c) Skicirajte (za općenit slučaj) pogrešku učenja i ispitnu pogrešku kao funkcije od k .
 - (d) Skicirajte (za općenit slučaj) pogrešku učenja i ispitnu pogrešku kao funkcije broja primjera N za $k = 1$ i $k = 3$ (nacrtajte dva zasebna grafikona).

3. [Svrha: Shvatiti uzročne veze između nevezanih veličina.] Obrazložite u kakvim su odnosima sljedeći pojmovi: (a) složenost modela, (b) broj parametara modela, (c) dimenzija ulaznog prostora n i (d) broj primjera N . Analizirajte odnose između svih parova pojmljiva, posebno za parametarske, a posebno za neparametarske metode.

2 Zadatci s ispita

1. (N) Bavimo se zadatkom određivanja etimologije riječi. Zanima nas je li neka nama nepoznata riječ latinskog ili slavenskog porijeka. Zadatak rješavamo kao binarnu klasifikaciju. Prikupili smo označeni skup primjera, koji se sastoji od latinskih riječi i riječi iz svih dvanaest živućih slavenskih jezika. Npr., u našem skupu imamo (*stroj*, 1), (*strues*, 0), (*tracto*, 0) i (*trasa*, 1), gdje 1 označava da je to slavenska riječ, a 0 da je latinska. Na ovom skupu primjera treniramo algoritam k -NN (k najbližih susjeda). Kao funkciju udaljenosti koristimo Levenshteinovu udaljenost. Levenshteinova udaljenost L između dviju riječi najmanji je broj umetanja, brisanja i zamjena jednog znaka potrebnih da se jedna riječ pretvori u drugu. Npr., $L(stroj, straja) = 2$. Razmatramo dva modela. Model h_1 je 3-NN. Model h_2 je težinski k -NN s jezgrenom funkcijom definiranom kao $\kappa(\mathbf{x}, \mathbf{x}') = 1/(1+L(\mathbf{x}, \mathbf{x}'))$. Koja je klasifikacija riječi $\mathbf{x} = \text{straja}$ prema modelima h_1 i h_2 ?
 A $h_1 = h_2 = 0$ B $h_1 = h_2 = 1$ C $h_1 = 1, h_2 = 0$ D $h_1 = 0, h_2 = 1$

2. (N) Algoritam k -NN koristimo za višeklasnu klasifikaciju riječi prema jeziku kojemu pripadaju. Skup za učenje sastoji se od sljedećih riječi i oznaka klasa:

$$\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\} = \{(\text{"water"}, 0), (\text{"voda"}, 1), (\text{"zrak"}, 1), (\text{"luft"}, 2), (\text{"feuer"}, 2)\}$$

Kao mjeru sličnosti između primjera koristimo jezgrenu funkciju nad znakovnim nizovima, definiranu kao $\kappa(\mathbf{x}_1, \mathbf{x}_2) = |\mathbf{x}_1 \cap \mathbf{x}_2|/|\mathbf{x}_1 \cup \mathbf{x}_2|$, gdje je su operacije unije i presjeka definirane nad skupovima slova od kojih se riječi sastoje. Npr., $\kappa(\text{"water"}, \text{"voda"}) = 1/8 = 0.125$. Razmatramo dvije varijante algoritma: 3-NN i težinski k -NN. Kod potonjeg u obzir uzimamo sve primjere, tj. $k = N$. Odredite klasifikaciju primjera $\mathbf{x} = \text{"zemlja"}$ pomoću ova dva algoritma. U slučaju izjednačenja glasova između klase, prednost se daje klasi s numerički manjom oznakom y . **U koju će klasu biti klasificiran primjer \mathbf{x} algoritmom 3-NN, a u koju algoritmom težinski k -NN?**

- A $y = 0$ i $y = 0$ B $y = 0$ i $y = 1$ C $y = 0$ i $y = 2$ D $y = 1$ i $y = 1$

VM - Neparametarske metode

I Zadaci za vježbe

1.1.

Dual. SVM: $h(\vec{x}) = \sum_{i=1}^N \alpha_i y^i \phi(\vec{x})^\top \phi(\vec{x}^i) + w_0$

Rg. k-NN $h(\vec{x}) = \operatorname{argmax}_{j \in \{0, \dots, k-1\}} \sum_{i \in N_k(\vec{x})} 1[y^i = j]$

- zajedničko

- oba modela su neparametarska modela

- razlika

- k-NN uspoređuje k najbljužih susjeda, a SVM samo s potpornim vektorma

1.2.

manji $k \Rightarrow$ složeniji model

a) 4-NN s euklidiskom udaljenosti

	\vec{x}^i	y^i	$\ \vec{x}^i - \vec{x}^1\ $	$\ \vec{x}^i - \vec{x}^2\ $
1	[4 4 0]	1	$\sqrt{5} = 2.24$	$\sqrt{26} = 5.1$
2	[4 3 1]	1	1	$\sqrt{25} \approx 4.47$
3	[6 0 2]	1	$\sqrt{3}$	$\sqrt{48} \approx 6.78$
4	[5 2 2]	0	$\sqrt{2} = 1.41$	$\sqrt{33} \approx 5.196$
5	[5 1 1]	0	$\sqrt{2} = 1.41$	$\sqrt{33} = 5.745$
6	[7 2 0]	0	$\sqrt{10} = 3.16$	$\sqrt{59} \approx 7.68$

$$\begin{aligned} \vec{x}^1 &= [4 \ 2 \ 1] \\ \vec{x}^2 &= [0 \ 3 \ 3] \end{aligned} \rightarrow \text{najbljuži: } \vec{x}^1, \vec{x}^2, \vec{x}^3, \vec{x}^4, \vec{x}^5 \Rightarrow \text{najbljuži: } \vec{x}^1, \vec{x}^2, \vec{x}^4, \vec{x}^5$$

$$h(\vec{x}^1) = \operatorname{argmax}_{y \in \{0, 1\}} (2, 2) \Rightarrow \text{nema odluke}$$

$$h(\vec{x}^2) = \operatorname{argmax}_y (2, 2) \rightarrow \text{nema odluke}$$

b)

$$\text{Inverzna kvadr. jezgra } K(\vec{x}, \vec{x}^i) = \frac{1}{1 + \|\vec{x} - \vec{x}^i\|^2}$$

	\vec{x}^i	y^i	$K(\vec{x}, \vec{x}^i)$	$K(\vec{x}, \vec{x}^i)$
1	(4, 4, 0)	1	$1/6 \approx 0.17$	$1/27 = 0.037$
2	(4, 3, 1)	1	$1/2 = 0.5$	$1/21 \approx 0.0476$
3	(6, 0, 2)	1	$1/10 = 0.1$	$1/47 \approx 0.0213$
4	(5, 2, 2)	0	$1/3 = 0.33$	$1/28 \approx 0.036$
5	(5, 1, 1)	0	$1/3 = 0.33$	$1/34 \approx 0.029$
6	(7, 2, 0)	0	$1/11 = 0.91$	$1/60 \approx 0.017$

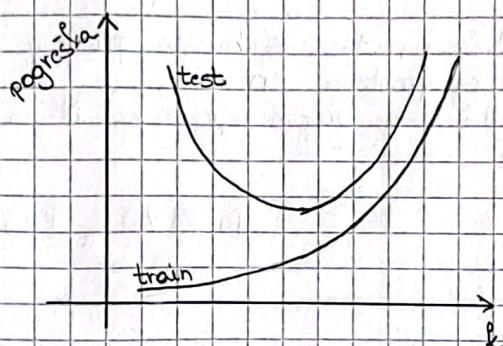
tečinski model: $h(\vec{x}) = \operatorname{argmax}_j \sum_{i=1}^N K(\vec{x}, \vec{x}^i) \cdot 1[y^i = j]$

$$h(\vec{x}^1) = \operatorname{argmax}_j \left(\underbrace{\frac{1}{6} + \frac{1}{2}}_{y=1}, \underbrace{\frac{1}{3} + \frac{1}{3}}_{y=0} \right) = \operatorname{argmax} \left(\frac{2}{3}, \frac{2}{3} \right) \rightarrow \text{nerra odluka}$$

$$h(\vec{x}^2) = \operatorname{argmax}_j \left(\underbrace{\frac{1}{27} + \frac{1}{21}}_{y=1}, \underbrace{\frac{1}{28} + \frac{1}{34}}_{y=0} \right) = \operatorname{argmax} (0.085, 0.065) = 1$$

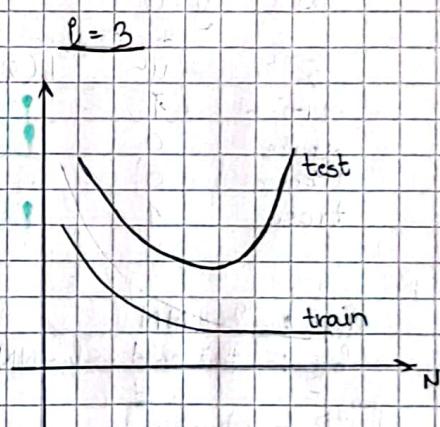
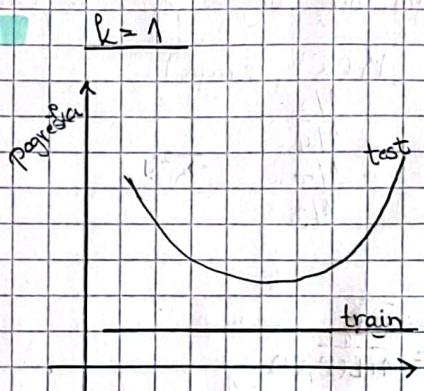
$h(\vec{x}^2)$ bit će klasificiran kao $y=1$,

c)



• veći $k \rightarrow$ jednostavniji model

d)



mali N i mali $k \rightarrow$ složen model

veliki N i mali $k \rightarrow$ loša general.

mali N i velik $k \rightarrow$ podnaučenost

veliki N i veliki $k \rightarrow$ najbolja komb. (poziv na prenačerst!)

1.3.

Parametarski modeli

- broj parametara modela raste s dimenzijom ulaznog prostora (n)
- broj parametara modela ne ovisi o broju primjera za učenje N
- složenost modela raste s povećanjem broja parametara, a ne ovisi o broju primjera

Neparametarski modeli

- broj parametara modela raste s brojem primjera N , a ne ovisi o dimenziji ulaznog prostora n
- složenost modela raste s brojem primjera N , a ne ovisi o dim. ulaznog prostora n

II Zadaci s ispita

2.1.

šav.
Cat.
 $y=1$
 $y=0$

• fja udaf. = Levenshtcinaova udaljenost : $L(\vec{x}, \vec{x}')$
→ najm. broj zamjena, brisanja, umetanja

\vec{x}^i	y^i	$L(\vec{x}^i, \text{stroja})$	$K(\vec{x}^i, \text{stroja})$
stroj	1	2	1/3
strues	0	3	1/4
tractor	0	4	1/5
trasa	1	2	1/3

$$h_1 : 3 - \text{NN}$$

$$h_2 : \text{težinski } k - \text{NN} \quad K(\vec{x}, \vec{x}') = \frac{1}{1 + L(\vec{x}, \vec{x}')}}$$

$$[\vec{x} = \text{stroja}]$$

$$L(\text{strues}, \text{stroja}) = 3$$

strues \rightarrow strues \rightarrow straes \rightarrow strojs \rightarrow stroja

$$L(\text{tracto}, \text{stroja}) = 4$$

tracto \rightarrow stracto \rightarrow strayto \rightarrow strajao \rightarrow stroja

$$h_1(\vec{x}) = \underset{\substack{j \in \{0, 1\} \\ \text{NN}_3}}{\operatorname{argmax}} \sum_{i=1}^N \mathbb{1}\{y^i = j\} = \underset{\substack{y=1 \\ y=0}}{\operatorname{argmax}} (2, 1) = 1 //$$

$$\begin{aligned} h_2(\vec{x}) &= \operatorname{argmax} \sum_i K(\vec{x}^i, \vec{x}) \mathbb{1}\{y^i = j\} = \operatorname{argmax} \left(\underbrace{\frac{1}{3} + \frac{1}{3}}_{y=1}, \underbrace{\frac{1}{4} + \frac{1}{5}}_{y=0} \right) \\ &= \operatorname{argmax} \left(\frac{2}{3}, \frac{9}{20} \right) \\ &= 1 // \end{aligned}$$

$\rightarrow \textcircled{b}$

2.2.

x-NN

\vec{x}^i	y^i	$K(\vec{x}^i, \text{zemlja})$
water	0	• $\frac{2}{8}$
voda	1	• $\frac{1}{8}$
zral	1	• $\frac{2}{7}$
luft	2	• $\frac{0}{9} = 0$
feuer	2	• $\frac{1}{8}$

acl/zemlja/wtr
a/zemlja/vod
za/zemlja/k
ø/luft/zemlja
e/feuer/zemlja

$$K(\vec{x}_1, \vec{x}_2) = \frac{|\vec{x}_1 \cap \vec{x}_2|}{|\vec{x}_1 \cup \vec{x}_2|} \rightarrow \text{slučnost}$$

najslučniji \Rightarrow najveća vrg.

h_1 : 3-NN višeklasni

h_2 : težinski x-NN

$$h_2 = \operatorname{argmax} \sum_{i=1}^N K(\vec{x}^i, \vec{x}) \cdot 1\{y^i = j\}$$

$$\vec{x} = \text{zemlja}$$

$$h_1(\vec{x}) = \operatorname{argmax}_{y \in \{0, 1, 2\}} [1 \ 2 \ 0] = 1$$

$$h_2(\vec{x}) = \operatorname{argmax} \left(\frac{2}{8}, \frac{1}{8} + \frac{2}{7}, \frac{1}{8} \right) = 1$$

D)

$$y_1 = 1, y_2 = 1 //$$

13. Procjena parametara

Strojno učenje 1, UNIZG FER, ak. god. 2022./2023.

Jan Šnajder, vježbe, v2.1

1 Zadatci za učenje

1. [Svrha: Prisjetiti se očekivanja, varijacije, kovarijacije i korelacije varijabli.] Neka je zajednička vjerojatnost $P(X, Y)$ varijabli X i Y sljedeća: $P(1, 1) = 0.2$, $P(1, 2) = 0.05$, $P(1, 3) = 0.3$, $P(2, 1) = 0.05$, $P(2, 2) = 0.3$, $P(2, 3) = 0.1$.
 - (a) Izračunajte marginalne vjerojatnosti $P(X)$ i $P(Y)$ te uvjetne vjerojatnosti $P(X|Y)$ i $P(Y|X)$. Uvjerite se da Bayesov teorem daje isti rezultat.
 - (b) Izračunajte očekivanje $\mathbb{E}[X]$, varijancu $\text{Var}(X)$, kovarijancu $\text{Cov}(X, Y)$, koeficijent korelacije $\rho_{X,Y}$ i kovarijacijsku matricu Σ .
 - (c) Dokažite:
 - i. $\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$
 - ii. $\text{Var}(aX) = a^2\text{Var}(X)$
 - iii. $\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$
2. [Svrha: Razumjeti nezavisnost slučajnih varijabli i shvatiti da linearna nekoreliranost ne znači nezavisnost.]
 - (a) Definirajte nezavisnost slučajnih varijabli (preko zajedničke vjerojatnosti i preko uvjetne vjerojatnosti).
 - (b) Sudeći po iznosu koeficijenta korelacije $\rho_{X,Y}$, jesu li varijable iz zadatka 1 linearno zavisne? Jesu li nezavisne?
 - (c) Za koje od sljedećih varijabli očekujete da su zavisne, a za koje da je ta zavisnost linearna:
 - (i) dob i veličina cipela,
 - (ii) dob i sati spavanja,
 - (iii) razina buke i udaljenost od izvora buke,
 - (iv) dob i prihodi?
 - (d) Dokažite da su nezavisne varijable linearno nekorelirane.

V13 - Procjena parametara

$$p(X|Y) = \frac{p(X \cap Y)}{p(Y)}$$

I Podaci za učenje

1.1.

$$\begin{aligned} P(1,1) &= 0.2 \\ P(1,2) &= 0.05 \\ P(1,3) &= 0.3 \end{aligned}$$

$$\begin{aligned} P(2,1) &= 0.05 \\ P(2,2) &= 0.3 \\ P(2,3) &= 0.1 \end{aligned}$$

$$X \in \{1, 2\} \quad Y \in \{1, 2, 3\}$$

a) $P(X) \quad P(Y) = ?$
 $P(X|Y), P(Y|X) = ?$

$$P(X|Y) = \frac{P(X \cap Y)}{P(Y)}$$

Bayesov teorem
 $P(X|Y) = \frac{P(Y|X) \cdot P(X)}{P(Y)}$

$$P(X=1) = 0.2 + 0.05 + 0.3 = 0.55$$

$$P(X=2) = 0.05 + 0.3 + 0.1 = 0.45$$

$$\begin{aligned} P(Y=1) &= 0.2 + 0.05 = 0.25 \\ P(Y=2) &= 0.05 + 0.3 = 0.35 \\ P(Y=3) &= 0.1 + 0.3 = 0.4 \end{aligned}$$

$p_{0,1}$
 \downarrow
 $P(X|Y)$

$$P(X=1 | Y=1) = \frac{P(1,1)}{P(Y=1)} = \frac{0.2}{0.25} = \frac{4}{5} = 0.8$$

$$P(X=2 | Y=1) = \frac{P(2,1)}{P(Y=1)} = \frac{0.05}{0.25} = \frac{1}{5} = 0.2$$

$$\begin{cases} P(X=1 | Y=2) = \frac{0.05}{0.35} = \frac{1}{7} \\ P(X=2 | Y=2) = \frac{0.3}{0.35} = \frac{6}{7} \end{cases} \quad \begin{cases} P(X=1 | Y=3) = \frac{0.3}{0.4} = \frac{3}{4} \\ P(X=2 | Y=3) = \frac{0.1}{0.4} = \frac{1}{4} \end{cases}$$

\downarrow
 $P(Y|X)$

$$\begin{cases} P(Y=1 | X=1) = \frac{P(1,1)}{P(X=1)} = \frac{0.2}{0.55} = \frac{4}{11} \\ P(Y=2 | X=1) = \frac{P(1,2)}{P(X=1)} = \frac{0.05}{0.55} = \frac{1}{11} \\ P(Y=3 | X=1) = \frac{P(1,3)}{P(X=1)} = \frac{0.3}{0.55} = \frac{6}{11} \end{cases}$$

$$\begin{cases} P(Y=1 | X=2) = \frac{0.05}{0.45} = \frac{1}{9} \\ P(Y=2 | X=2) = \frac{0.3}{0.45} = \frac{2}{3} = \frac{6}{9} \\ P(Y=3 | X=2) = \frac{0.1}{0.45} = \frac{2}{9} \end{cases}$$

Provjera Bayesovog TEOREMA

$$P(X|Y) = \frac{P(Y|X) \cdot P(X)}{P(Y)} \Leftrightarrow P(X|Y)P(Y) = P(Y|X)P(X)$$

• samo za neke primjere

$$\underline{1} \quad P(X=1|Y=1) \cdot P(Y=1) \stackrel{?}{=} P(Y=1|X=1) \cdot P(X=1)$$

$$\frac{4}{5} \cdot \frac{1}{4} = \frac{4}{11} \cdot \frac{11}{20}$$

$$\frac{1}{5} = \frac{1}{5} \quad \checkmark$$

$$\underline{2} \quad P(X=2|Y=3) \cdot P(Y=3) \stackrel{?}{=} P(Y=3|X=2) \cdot P(X=2)$$

$$\frac{1}{4} \cdot \frac{4}{10} = \frac{2}{3} \cdot \frac{9}{20}$$

$$\frac{1}{10} = \frac{1}{10} \quad \checkmark$$

Jednakošt sljedi direktno
iz definicije
 $P(A, B) = P(A) \cdot P(B|A)$
 $= P(B) \cdot P(A|B)$

b)

$$P(X=1) = 0.55$$

$$P(X=2) = 0.45$$

$$P(Y=1) = 0.25$$

$$P(Y=2) = 0.35$$

$$P(Y=3) = 0.4$$

$$\mathbb{E}[X] = \sum x \cdot P(X=x) = 1 \cdot 0.55 + 2 \cdot 0.45 = 1.45 = \mu_X$$

$$\mathbb{E}[Y] = \sum y \cdot P(Y=y) = 1 \cdot 0.25 + 2 \cdot 0.35 + 3 \cdot 0.4 = 2.15 = \mu_Y$$

$$\text{Var}(X) = \sigma^2_X = \mathbb{E}[(X - \mathbb{E}[X])^2] = \sum (x - \mu_X)^2 p(x)$$

$$= \sum x^2 p(x) - 2\mu_X \underbrace{\sum x p(x)}_{\mu_X} + \mu_X^2 \underbrace{\sum p(x)}_1$$

$$= \sum x^2 p(x) - \mu_X^2 = 1 \cdot 0.55 + 4 \cdot 0.45 - 1.45^2$$

$$= 0.2475$$

$$\text{Var}(Y) = \sigma^2_Y = \mathbb{E}[(Y - \mathbb{E}[Y])^2] = \dots = \mathbb{E}[Y^2] - \mathbb{E}[Y]^2$$

$$= \sum y^2 p(y) - \mu_Y^2$$

$$= 1^2 \cdot 0.25 + 2^2 \cdot 0.35 + 3^2 \cdot 0.4 - 2.15^2$$

$$= 0.6275$$

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

$$= \sum_x \sum_y (x - \mu_X)(y - \mu_Y) P(x, y)$$

$$= P(1,1)(1-1.45)(1-2.15) + P(1,2)(1-1.45)(2-2.15) + P(1,3)(1-1.45)(3-2.15)$$

$$+ P(2,1)(2-1.45)(1-2.15) + P(2,2)(2-1.45)(2-2.15) + P(2,3)(2-1.45)(3-2.15)$$

$$= \dots = -0.45 \cdot 0.0175 + 0.55 \cdot (-0.0175) = -0.0175$$

$$S_{X,Y} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} \approx -0.04441$$

$$\Sigma = \begin{bmatrix} 6^2 x & 6x, y \\ 6y, x & 6^2 y \end{bmatrix} = \begin{bmatrix} 0.2475 & -0.0175 \\ -0.0175 & 0.0275 \end{bmatrix}$$

c) i) $\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$

$$\begin{aligned} \text{Var}(X) &= \mathbb{E}[(X - \mathbb{E}[X])^2] = [\text{označimo radi jednost. } \mu_X = \mathbb{E}[X]] \\ &= \mathbb{E}[(X - \mu_X)^2] = \\ &= \sum_x (x - \mu_X)^2 P(X=x) = \sum_x (x^2 - 2x\mu_X + \mu_X^2) P(x) \\ &= \underbrace{\sum_x x^2 P(x)}_{\mathbb{E}[X^2]} - 2\mu_X \underbrace{\sum_x x P(x)}_{\mu_X} + \mu_X^2 \underbrace{\sum_x P(x)}_1 \\ &= \mathbb{E}[X^2] - 2\mu_X^2 + \mu_X^2 \\ &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 \end{aligned}$$

ii)

$$\begin{aligned} \text{Var}(aX) &= \mathbb{E}[(aX - \mathbb{E}[aX])^2] \stackrel{(1)}{=} \\ &= \mathbb{E}[(aX - a\mathbb{E}[X])^2] = \\ &= \mathbb{E}[a^2(X - \mathbb{E}[X])^2] \stackrel{(1)}{=} \\ &= a^2 \mathbb{E}[(X - \mathbb{E}[X])^2] = \\ &= a^2 \text{Var}(X) \end{aligned}$$

$$\left. \begin{aligned} \mathbb{E}[aX] &= \sum_x a \cdot x P(x) \\ &= a \sum_x x P(x) \\ &= a \mathbb{E}[X] \end{aligned} \right\}$$

iii) $\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X]) \cdot (Y - \mathbb{E}[Y])]$

$$\begin{aligned} &= \mathbb{E}[XY - X\mathbb{E}[Y] - Y\mathbb{E}[X] + \mathbb{E}[X]\mathbb{E}[Y]] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[Y\mathbb{E}[X]] + \mathbb{E}[\mathbb{E}[X]\mathbb{E}[Y]] \\ &= \left\{ \begin{aligned} \mathbb{E}[X \cdot \mathbb{E}[Y]] &= \sum_x x \mu_Y P(x) = \mu_X \mu_Y = \mathbb{E}[X]\mathbb{E}[Y] \\ \mathbb{E}[\mathbb{E}[X]\mathbb{E}[Y]] &= \mathbb{E}[X]\mathbb{E}[Y] \end{aligned} \right. \\ &= \mathbb{E}[XY] - 2\mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[Y] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y], \end{aligned}$$

1.2.

a) Nezavisnost slučajnih varijabli X i Y \Leftrightarrow preko zajedničke vjerojatnosti : $P(X, Y) = P(X)P(Y)$ \Leftrightarrow preko uvjetne vjerojatnosti : $P(Y|X) = P(Y)$
 $P(X|Y) = P(X)$

$$\begin{aligned} P(X, Y) &= P(X|Y)P(Y) \\ &= P(Y|X)P(X) \\ &= P(X)P(Y) \end{aligned}$$

b) Sudeći po $S_{X,Y}$ iz 1. redatelja jesu li X i Y linearno zavisne?

$S_{X,Y} = -0.04441 \approx 0$

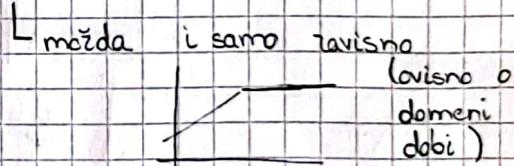
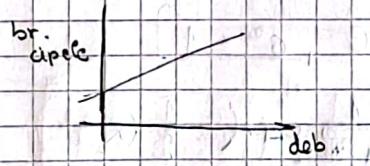
 $\Rightarrow X$ i Y nisu linearno zavisne

Jesu li nezavisne?

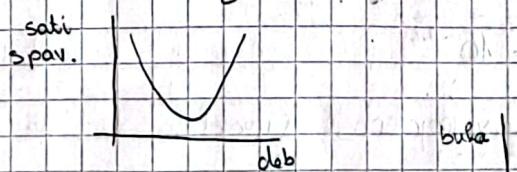
X i Y su linearno nezavisne, međutim ne možemo reći da su čvrsto X i Y , nezavisne. Razlog tome je što Pearsonov koeficijent korelacije mjeri uključivo LINEARNU ZAVISNOST. Drugim rečima moguće je da su X i Y nelinearno zavisne (npr. kvadratno), a da je $S_{X,Y} \approx 0$.

c)

I DOB i VELIČINA CIPELA = - linearno zavisno



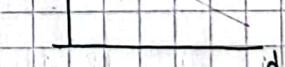
II DOB i SATI SPAVANJA = zavisno



III RAZINA BUKE i UDALJ. OD IZVORIA

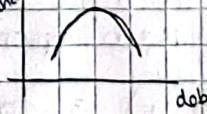
- linearno zavisno

(prijevišenje)



IV DOB i PRIHODI

- zavisno

d) X i Y nezavisne varijable

$\Rightarrow P(X, Y) = P(X)P(Y)$

$$\begin{aligned} \Rightarrow \text{Cov}(X, Y) &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = \sum \sum xyP(x,y) - \mu_x\mu_y \\ &= \sum \sum xyP(X)P(Y) - \mu_x\mu_y \\ &= \sum xP(x) \sum yP(y) - \mu_x\mu_y \\ &= \mu_x\mu_y - \mu_x\mu_y \\ &= 0 \end{aligned}$$

vidi 1.1c

$\Rightarrow \rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} = 0 \Rightarrow X$ i Y linearno nezavisne!

14. Procjena parametara II

Strojno učenje 1, UNIZG FER, ak. god. 2023./2024.

Jan Šnajder, vježbe, v2.2

1 Zadatci za učenje

1. [Svrha: Razumjeti kako podatci određuju izglednost parametara putem funkcije izglednosti.]
 - (a) Definirajte funkciju izglednosti $\mathcal{L}(\boldsymbol{\theta}|\mathcal{D})$. Na kojoj se prepostavci o skupu \mathcal{D} temelji ta definicija?
 - (b) Raspolažemo skupom (neoznačenih) primjera $\mathcal{D} = \{x^{(i)}\}_i = \{-2, -1, 1, 3, 5, 7\}$. Prepostavljamo da se primjeri pokoravaju Gaussovoj distribuciji, $x^{(i)} \sim \mathcal{N}(\mu, \sigma^2)$. Napišite funkciju izglednosti $\mathcal{L}(\mu, \sigma^2|\mathcal{D})$. Koliko iznosi izglednost parametara $\mu = 0$ i $\sigma^2 = 1$, a koliko vjerojatnost uzorka \mathcal{D} uz te parametre?
 - (c) Novčić bacamo N puta, pri čemu smo m puta dobili glavu, a $N - m$ puta pismo. Ishodi bacanja novčića sačinavaju naš uzorak \mathcal{D} . Napišite izraz za funkciju izglednosti parametra μ Bernoullijeve distribucije, parametrizirane s N i m , tj. $\mathcal{L}(\mu|N, m)$.
 - (d) Skicirajte funkciju izglednosti za slučaj $N = 10$ i $m = 1$. Koja je vrijednost parametra μ najizglednija? Uz koju je vrijednost μ skup \mathcal{D} najvjerojatniji?
2. [Svrha: Osvježiti znanje matematike potrebno za izvođenje MLE-procenitela dviju osnovnih univarijatnih razdioba.]
 - (a) Definirajte MLE-procenitelj $\hat{\boldsymbol{\theta}}_{ML}$.
 - (b) Izvedite MLE-procenitelj $\hat{\mu}_{ML}$ za parametar μ Bernoullijeve razdiobe $P(x|\mu)$.
 - (c*) Izvedite MLE-procenitelj $\hat{\mu}_{k,ML}$ za parametar μ_k kategorijske (“multinulijeve”) razdiobe $P(\mathbf{x}|\boldsymbol{\mu})$. Ovdje je kod optimizacije potrebno osigurati da vrijedi ograničenje $\sum_{k=1}^K \mu_k = 1$; za to upotrijebite metodu Lagrangeovih multiplikatora.
 - (c) Izvedite MLE-procenitelje $\hat{\mu}_{ML}$ i $\hat{\sigma}^2$ za parametre μ odnosno σ^2 univarijatne Gaussove razdiobe $p(x|\mu, \sigma^2)$.
3. [Svrha: Isprobati izračun pristranost procjenitelja i shvatiti da MLE-procenitelj može biti pristran, tj. da najveća izglednost ne jamči nepristranost.]
 - (a) Dokažite da je $\hat{\mu}_{ML}$ nepristran, a $\hat{\sigma}_{ML}^2$ pristran. Koliko iznosi pristranost $b(\hat{\sigma}^2)$?
 - (b) Je li ta pristranost u praksi problematična? Obrazložite.
4. [Svrha: Izvježbati izračun procjene parametara multivarijatne Gaussove razdiobe (v. primjer 3.5 u skripti). Uočiti da multikolinearnost značajki dovodi do problema.] Raspolažemo uzorkom $\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i=1}^6$ za koji prepostavljamo da potječe iz multivarijatne normalne razdiobe:
$$\begin{aligned}\mathbf{x}^{(1)} &= (9.5, -0.7, -2.8) & \mathbf{x}^{(4)} &= (2.3, 0.3, 1.2) \\ \mathbf{x}^{(2)} &= (8.8, -0.8, -3.2) & \mathbf{x}^{(5)} &= (2.2, 0, 0) \\ \mathbf{x}^{(3)} &= (6.5, -0.2, -0.8) & \mathbf{x}^{(6)} &= (3.6, 0.3, 1.2)\end{aligned}$$
 - (a) Izračunajte MLE-procjenu vektora srednje vrijednosti i MLE-procjenu kovarijacijske matrice.
 - (b) Izračunajte gustoću vjerojatnosti za primjer $\mathbf{x} = (-2, 1, 0)$. Je li ta gustoća dobro definirana? Zašto?

- (c) Matrica kovarijacije Σ mora biti pozitivno definitna a da bi imala pozitivnu determinantu i inverz. Multikolinearnost značajki jedan je od mogućih razloga zašto matrica nije pozitivno definitna. Izračunajte Pearsonov koeficijent korelacije ρ između svih parova varijabli te izbacite varijablu koja je najviše korelirana s nekom drugom varijablom. Zatim u tako smanjenome ulaznom prostoru pokušajte ponovno izračunati funkciju gustoće za primjer \mathbf{x} .
5. [Svrha: Razumjeti MAP-procenitelj i način njegovog izračuna za Bernoullijevu distribuciju (beta-Bernoullijev model). Uočiti kako svojstvo konjugatnosti olakšava izračun aposteriorne distribucije.]
- Definirajte MAP-procenitelj $\hat{\theta}_{\text{MAP}}$ i objasnite zašto je on bolji od MLE-procenitelja $\hat{\theta}_{\text{ML}}$.
 - Objasnite što je to (1) konjugatna distribucija i (2) konjugatna apriorna distribucija. Zašto nam je svojstvo konjugatnosti bitno?
 - Apriornu distribuciju parametra μ Bernoullijeve distribucije modeliramo beta-distribucijom $p(\mu|\alpha, \beta)$. Beta-distribucija konjugatna je apriorna distribucija za Bernoullijevu funkciju izglednosti $\mathcal{L}(\mu|N, m)$. Skicirajte beta-distribuciju za (1) $\alpha = \beta = 1$, (2) $\alpha = \beta = 2$, (3) $\alpha = 2$, $\beta = 4$ i (4) $\alpha = 4$, $\beta = 2$.
 - Izvedite izraz za aposteriornu distribuciju parametra, $p(\mu|N, m, \alpha, \beta)$.
 - Recimo da vjerujemo da je novčić pravedan, ali da u to nismo baš u potpunosti uvjereni. To možemo modelirati beta-distribucijom $p(\mu|\alpha = 2, \beta = 2)$. Zatim smo u $N = 10$ bacanja novčića samo $m = 1$ puta dobili glavu. Skicirajte apriornu gustoću $p(\mu|\alpha = 2, \beta = 2)$, funkciju izglednosti $\mathcal{L}(\mu|N = 10, m = 1)$ te njihov umnožak. Iskoristite činjenicu da je maksimizator (mod) beta-distribucije jednak $\frac{\alpha-1}{\alpha+\beta-2}$.
 - Izračunajte $\hat{\mu}_{\text{MAP}}$ i $\hat{\mu}_{\text{ML}}$ te komentirajte razliku. Kako bi porast broja primjera N utjecao na ovu razliku?
 - Pokažite da se MAP-procenitelj za parametar μ Bernoullijeve distribucije svodi na Laplaceov procenitelj, ako se apriorna distribucija parametra modelira beta-distribucijom te ako se odaberu odgovarajući (koji?) parametri α i β .
6. [Svrha: Razumjeti MAP-procenitelj i način njegovog izračuna za kategorisku (multinulijsku) varijablu (Dirichlet-kategorijski model).]
- Definirajte Dirichletovu distribuciju.
 - Definirajte Dirichlet-kategorijski model i izvedite MAP procenitelj za $\alpha_k = 2$.
7. [Svrha: Razumjeti vezu između probabilističkih modela i poopćenih linearnih modela preko veze između MLE-procenitelja i minimizacije empirijske pogreške. Razumjeti vezu između MAP-procenitelja i minimizacije L2-regularizirane empirijske pogreške.]
- Pokažite da je MLE-procjena za parametre \mathbf{w} kod linearne regresije (uz pretpostavku normalno distribuiranog šuma) ekvivalentna postupku najmanjih kvadrata.
 - Pokažite da je MLE-procjena za parametre \mathbf{w} kod logističke regresije (uz pretpostavku Bernoullijeve distribucije oznaka) ekvivalentna minimizacija pogreške unakrsne entropije.
 - (c*) Gornja dva zadatka demonstriraju vezu između MLE-procenitelja i minimizacije empirijske pogreške. Postoji analogna veza između MAP-procenitelja i minimizacije L2-regularizirane empirijske pogreške. Razmotrimo konkretno linearnu regresiju. Ako se apriorna gustoća vjerojatnosti težina \mathbf{w} definira kao:
- $$p(\mathbf{w}) = \mathcal{N}(0, \alpha^{-1} \mathbf{I})$$
- t.j. kao multivariatna normalna razdioba sa središtem u ishodištu prostora parametara i s izotropnom kovarijacijskom matricom pomnoženom nekim hiperparametrom α^{-1} , onda je MAP-procenitelj ekvivalentan L2-regulariziranoj kvadratnoj pogrešci. Dokažite to. (Pomoć: slajdovi 30–31 [ovdje](#) i poglavljje 3.3.1 u PRML.)
- (d*) Je li u prethodnom zadatku bilo ključno to što je Gaussova distribucija samokonjugatna? Možemo li isti princip primijeniti i kod modela gdje izglednost nije Gaussova, npr. kod logističke regresije (i drugih poopćenih linearnih modela)? Zašto?

2 Zadaci s ispita

1. (N) Raspolažemo sljedećim skupom označenih primjera:

$$\mathcal{D} = \{x^{(i)}, y^{(i)}\} = \{(-2, 1), (-2, 1), (-1, 0), (0, 0), (1, 1), (3, 1)\}$$

Na ovom skupu treniramo univarijatni Bayesov klasifikator, za što trebamo procijeniti izglednosti klase $p(x|y)$. Te su izglednosti definirane Gaussovom gustoćom vjerojatnosti:

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

Parametre μ i σ^2 gustoće vjerojatnosti $p(x|y)$ procjenjujemo MLE-om. Neka su μ_1 i σ_1^2 parametri gustoće vjerojatnosti $p(x|y=1)$ dobiveni MLE-om na podskupu primjera $\mathcal{D}_{y=1}$. **Koliko iznosi log-izglednost $\mathcal{L}(\mu_1, \sigma_1^2 | \mathcal{D}_{y=1})$?**

- A -22.60 B -8.68 C -8.76 D +0.48

2. (N) Zadan je uzorak kontinuirane slučajne varijable, $\mathcal{D} = \{-3, -2, 1, 4\}$. Pretpostavljamo razdiobu $\mathcal{N}(\mu, \sigma^2)$, čija je gustoća vjerojatnosti:

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

Koristimo MLE kako bismo na uzorku \mathcal{D} procijenili parametre $\hat{\mu}_{MLE}$ i $\hat{\sigma}_{MLE}^2$. Usporedbe radi, parametar σ^2 dodatno procjenjujemo nepristranim procjeniteljem $\hat{\sigma}_{UB}^2$. Izračunajte log-izglednost parametara $(\hat{\mu}_{MLE}, \hat{\sigma}_{MLE}^2)$ te log-izglednost parametara $(\hat{\mu}_{MLE}, \hat{\sigma}_{UB}^2)$. **Koliko je prva log-izglednost veća od druge log-izglednosti?**

- A 0.018 B 0.058 C 0.075 D 0.095

3. (P) Neka je $\mathcal{L}(\mu, \sigma^2 | \mathcal{D})$ log-izglednost parametara μ i σ^2 normalne distribucije izračunata nad uzorkom \mathcal{D} koji sadrži ukupno N opažanja normalne varijable x . Nadalje, neka su $(\mu_{MLE}, \sigma_{MLE}^2)$ parametri distribucije procijenjeni MLE-om nad uzorkom \mathcal{D} , te neka je σ_{UB}^2 nepristrana procjena varijance, izračunata kao $\sigma_{UB}^2 = \frac{N}{N-1} \sigma_{MLE}^2$. Konačno, neka je \mathcal{D}' slučajno uzorkovan podskup uzorka \mathcal{D} , tj. $\mathcal{D}' \subset \mathcal{D}$, pri čemu je poduzorkovanje načinjeno nakon procjene parametara. Razmotrite sljedeće četiri vrijednosti funkcije log-izglednosti $\mathcal{L}(\mu, \sigma^2 | \mathcal{D})$:

$$\begin{aligned}\mathcal{L}_0 &= \mathcal{L}(\mu_{MLE}, \sigma_{MLE}^2 | \mathcal{D}) \\ \mathcal{L}_1 &= \mathcal{L}(0, 1 | \mathcal{D}) \\ \mathcal{L}_2 &= \mathcal{L}(\mu_{MLE}, \sigma_{UB}^2 | \mathcal{D}) \\ \mathcal{L}_3 &= \mathcal{L}(\mu_{MLE}, \sigma_{UB}^2 | \mathcal{D}')\end{aligned}$$

Što možemo zaključiti o odnosima između ovih vrijednosti funkcije log-izglednosti?

- A $\mathcal{L}_0 > \mathcal{L}_1$, $\mathcal{L}_1 \neq \mathcal{L}_2$, $\mathcal{L}_2 \geq \mathcal{L}_3$
 B $\mathcal{L}_1 \geq \mathcal{L}_0$, $\mathcal{L}_2 \geq \mathcal{L}_3$
 C $\mathcal{L}_0 < \mathcal{L}_3$, $\mathcal{L}_0 \leq \mathcal{L}_2$
 D $\mathcal{L}_0 \geq \mathcal{L}_1$, $\mathcal{L}_0 > \mathcal{L}_2$, $\mathcal{L}_2 \neq \mathcal{L}_3$

4. (N) U beta-Bernoullijevom modelu, apriornu vjerojatnost parametra μ modeliramo beta-distribucijom, čija je gustoća vjerojatnosti definirana kao:

$$p(\mu|\alpha, \beta) = \frac{1}{B(\alpha, \beta)} \mu^{\alpha-1} (1-\mu)^{\beta-1}$$

Mod (maksimizator) te distribucije jest:

$$\mu^* = \frac{\alpha-1}{\alpha+\beta-2}$$

Aposteriorna distribucija parametra definirana je kao:

$$p(\mu|\mathcal{D}, \alpha, \beta) = \mu^{m+\alpha-1} (1-\mu)^{N-m+\beta-1} \frac{1}{B(\alpha, \beta)p(\mathcal{D})}$$

Neka $\alpha = \beta = 2$. Računamo MAP-procjenu za parametar μ Bernoulijeve varijable. To radimo na dva uzorka, $\mathcal{D}_1 = (N_1, m_1)$ i $\mathcal{D}_2 = (N_2, m_2)$, koji nam pristižu jedan za drugim. Pritom koristimo svojstvo konjugatnosti, na način da aposterioru gustoću vjerojatnosti izračunatu na temelju prvog uzorka koristimo kao apriornu gustoću vjerojatnosti pri procjeni na temelju drugog uzorka. U prvom uzorku, veličine $N_1 = 50$, Bernoullijeva varijabla realizirana je s vrijednošću $y = 1$ ukupno $m_1 = 42$ puta. U drugom uzorku, veličine $N_2 = 15$, Bernoullijeva varijabla realizirana je s vrijednošću $y = 1$ ukupno $m_2 = 3$ puta. Izračunajte MAP-procjene za parametar μ na temelju ova dva uzorka. **Koliko iznosi promjena u procjeni za μ između prve i druge procjene?**

- A -0.59 B +0.45 C -0.14 D -0.64

5. (P) Koristimo MAP-procenitelj kako bismo procjenili parametre distribucije kategoričke (multinulijeve) varijable X . Varijabla može poprimiti tri vrijednosti, x_1 , x_2 i x_3 , pa dakle trebamo procjeniti vektor parametara (μ_1, μ_2, μ_3) . Budući da se ovdje radi o kategoričkoj varijabli, za MAP-procjenu koristimo Dirichlet-kategorički model. Na temelju stručnog znanja o problemu koji rješavamo, u procjenu smo ugradili naše pretpostavke. To znači da smo na prikladan način definiiali Dirichletovu apriornu gustoću vjerojatnosti, $p(\mu_1, \mu_2, \mu_3 | \alpha_1, \alpha_2, \alpha_3)$, gdje je $(\alpha_1, \alpha_2, \alpha_3)$ vektor hiperparametara (parametri Dirichletove distribucije). Konkretno, te smo hiperparametre definiiali kao $(\alpha_1, \alpha_2, \alpha_3) = (2, 2, 1)$. Međutim, skup podataka \mathcal{D} ne odgovara našoj pretpostavci. U tom skupu, varijabla X je u pola slučajeva realizirana s vrijednošću x_2 , u pola slučajeva s vrijednošću x_3 , no baš niti jednom s vrijednošću x_1 . **Kakva će biti naša MAP-procjena parametara (μ_1, μ_2, μ_3) ?**

- A $\mu_1 = 0, \frac{1}{2} < \mu_2 < 1, 0 < \mu_2 < \mu_3 < 1$
 B $0 < \mu_1 < \frac{1}{3}, \frac{1}{2} < \mu_2 < 1, \mu_3 = 0$
 C $0 < \mu_1 < \mu_3 < 1, \frac{1}{3} < \mu_2 < \frac{2}{3}$
 D $0 < \mu_1 < \frac{1}{3}, \frac{1}{3} < \mu_2 < 1, 0 < \mu_3 < \mu_2 < 1$

6. (P) Bacanje igrače kocke modeliramo kategoričkom varijablom \mathbf{x} , gdje indikatorske varijable x_1, \dots, x_6 odgovaraju vrijednosti koju dobivamo bacanjem kocke. Za procjenu parametara $\boldsymbol{\mu}$ kategoričke distribucije koristimo MAP-procenitelj s Dirichletovom distribucijom za apriornu gustoću vjerojatnosti. U stvarnosti, kocka je modificirana tako da će nešto češće davati šesticu, odnosno realizaciju $x_6 = 1$, međutim mi to ne znamo. Naprotiv, na temelju manjeg broja opažanja ranijih bacanja kocke utvrdili smo da je kocka najčešće davala peticu, no svjesni smo da je naša procjena temeljena na manjem broju opažanja. **Uz koje parametre Dirichletove distribucije će naša procjena za $\boldsymbol{\mu}$ biti najbliža stvarnoj vrijednosti tih parametara?**

- A $\boldsymbol{\alpha} = (1, 1, 1, 1, 1, 1)$
 B $\boldsymbol{\alpha} = (\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6})$
 C $\boldsymbol{\alpha} = (2, 2, 2, 2, 2, 2)$
 D $\boldsymbol{\alpha} = (1, 1, 1, 1, 3, 1)$

7. (P) MAP-proceniteljem na skupu \mathcal{D} procjenjujemo parametre $\boldsymbol{\mu}$ kategoričke (multinulijeve) varijable \mathbf{x} , tj. procjenjujemo parametre multinulijeve distribucije $P(\mathbf{x}|\boldsymbol{\mu})$. Varijabla \mathbf{x} može propnimiti tri moguće vrijednosti ($K = 3$). Apriorna Dirichletova distribucija $P(\boldsymbol{\mu}|\boldsymbol{\alpha})$ definirana je parametrima $\boldsymbol{\alpha}$ definiranima kao $\boldsymbol{\alpha} = \alpha \cdot (1, 4, 2)$, gdje faktor α određuje vršnost Dirichletove distribucije, $\alpha \geq 1$. U skupu podataka \mathcal{D} realizirane su sve tri vrijednosti kategoričke varijable \mathbf{x} , i to sa sljedećim brojem realizacija: $N_1 = 50$, $N_2 = 29$, $N_3 = 40$. **Koliko mora iznositi faktor α , a da bi vrijednost x_2 bila najvjerojatnija vrijednost kategoričke varijable \mathbf{x} prema procjenjenoj distribuciji $P(\mathbf{x}|\boldsymbol{\mu})$?**

- A $\alpha > 2$ B $\alpha > 5$ C $\alpha > 7$ D $\alpha > 8$

8. (P) U beta-Bernoullijevom modelu, apriornu vjerojatnost parametra μ modeliramo beta-distribucijom. Gustoća vjerojatnosti i mod (maksimizator) beta-distribucije su:

$$p(\mu|\alpha, \beta) = \frac{1}{B(\alpha, \beta)} \mu^{\alpha-1} (1-\mu)^{\beta-1} \quad \mu^* = \frac{\alpha-1}{\alpha+\beta-2}$$

Na temelju beta-Bernoullijevog modela na skupu \mathcal{D} računamo MAP procjenu parametra μ Bernoullijeve distribucije. MLE procjena za isti parametar na skupu \mathcal{D} iznosi 0.2. MAP i MLE procjene mogu se poklopiti i onda kada ne koristimo uniformnu apriornu razdiobu. **Uz koje parametre neuniformne beta-distribucije će MLE i MAP procjene biti identične?**

- A $\alpha = 2, \beta = 5$ B $\alpha = 2, \beta = 10$ C $\alpha = 4, \beta = 8$ D $\alpha = 5, \beta = 7$

V14 - Procjena parametara II

I Zadaci za vježbe

1.1.

a) fja igrčednosti $L(\vec{\theta} | D) : \vec{\theta} \rightarrow p(D | \vec{\theta})$

$$L(\vec{\theta} | D) = p(D | \vec{\theta}) = p(x^1, x^2, \dots, x^N | \vec{\theta}) \\ = \prod_{i=1}^N p(x^i | \vec{\theta})$$

- pretpostavka na skupu $D = \{x^i\}_{i=1}^N$
 - podaci nezavisno i identično distribuirani
 (I.I.D.)

b) $D = \{x^i\} = \{-2, -1, 1, 3, 5, 7\} \sim N(\mu, \sigma^2)$

$$L(\mu, \sigma^2 | D) = p(D | \mu, \sigma^2) \\ = \prod_{i=1}^6 p(x^i | \mu, \sigma^2) \\ = \prod_{i=1}^6 \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2} (x^i - \mu)^2\right) \\ = \left[\mu = 0, \sigma^2 = 1 \right] \\ = \frac{1}{\prod_{i=1}^6 \sqrt{2\pi}} \exp\left(-\frac{1}{2} (x^i)^2\right) \\ = \left(\frac{1}{\sqrt{2\pi}}\right)^6 \cdot \frac{1}{\prod_{i=1}^6 c^{-\frac{1}{2}(x^i)^2}} \\ \approx 1 \cdot 10^{-22}$$

→ igrčednost i vjerojatnost uzorka D uz parametre $\mu = 0$ i $\sigma^2 = 1$ iznosi 10^{-22} .

c)

Inovčić N puta

↳ m gava

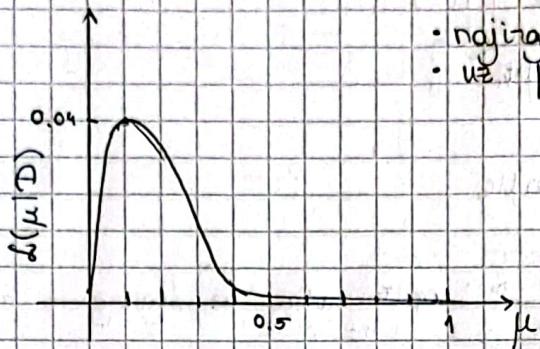
↳ $N-m$ pismo

Bernoulli

$$L(\mu | N, m) = p(D | N, m, \mu) = \prod_{i=1}^N p(x^i | N, m, \mu) \\ = \prod_{i=1}^N \mu^{x^i} (1-\mu)^{1-x^i} \\ = \mu^m (1-\mu)^{N-m}$$

d)

$$\left. \begin{array}{l} N=10 \\ m=1 \end{array} \right\} L(\mu | D) = \mu^m (1-\mu)^{N-m} = \mu (1-\mu)^9$$



- najjača vrednost parametra $\mu = 0.1$
- uz $\mu = 0.1$ slup D je najvjerojatniji

$$L(\mu | D) = P(D | \mu)$$

1.2.

prosječna MLE

a) definicija $\vec{\theta}_{ML} = \underset{\vec{\theta}}{\operatorname{argmax}} L(\vec{\theta} | D)$

↳ također zbog matematičke jednostavnosti i monotonosti logaritamske funkcije

$\vec{\theta}_{ML} = \underset{\vec{\theta}}{\operatorname{argmax}} (\ln [L(\vec{\theta} | D)])$ ↳ log-izglednost

b) $\hat{\mu}_{ML} = ?$ (Bernoullijska razdioba)

log-izglednost

$$\begin{aligned} \ln [L(\hat{\mu}_{ML} | D)] &= \ln [P(D | \hat{\mu}_{ML})] = \ln \prod_{i=1}^N P(x^i | \hat{\mu}_{ML}) = \ln \prod_i \mu^{x^i} (1-\mu)^{1-x^i} \\ &= \sum x^i \ln(\mu) + (1-x^i) \ln(1-\mu) \\ &= m \ln(\mu) + (N-m) \ln(1-\mu) \end{aligned}$$

• maksimizacija L

$$\frac{dL}{d\mu} = \frac{m}{\mu} + \frac{N-m}{1-\mu} \cdot (-1) = 0$$

$$m(1-\mu) - (N-m)\mu = 0$$

~~$m - m\mu - N\mu + m\mu = 0$~~

$$\left[\hat{\mu}_{ML} = \frac{m}{N} = \frac{1}{N} \sum_{i=1}^N x^i \right]$$

c*) $\hat{\mu}_{k,ML} = ?$

(Katgorička „multinomijalna“ razdiobba)

$$\begin{aligned}\ln L(\vec{\mu} | \mathcal{D}) &= \ln \left[\prod_{i=1}^N P(x_i | \vec{\mu}) \right] \\ &= \ln \left[\prod_{i=1}^N \frac{1}{\Gamma(k)} \prod_{k=1}^K \mu_k^{x_k^i} \right] \\ &= \sum_{i=1}^N \sum_{k=1}^K x_k^i \ln \mu_k\end{aligned}$$

• ograničenje : $\sum_{k=1}^K \mu_k = 1 \rightarrow$ optimizacija uz Lagrangeove multipl.

Lagrange. fja: $f(\vec{\mu}, \lambda) = \sum_{i=1}^N \sum_{k=1}^K x_k^i \ln \mu_k + \lambda \left(\sum_{k=1}^K \mu_k - 1 \right)$

$$\frac{\partial f}{\partial \mu_k} = 0 \Rightarrow \mu_{k0} = -\frac{1}{\lambda} \sum_{i=1}^N x_k^i$$

ograničenje.

$$\sum_{k=1}^K \mu_k = -\frac{1}{\lambda} \underbrace{\sum_{k=1}^K \sum_{i=1}^N x_k^i}_{= N} = 1$$

$$\Rightarrow \lambda = -N$$

$$\hookrightarrow \left[\hat{\mu}_{k,ML} = \frac{1}{N} \sum_{i=1}^N x_k^i = \frac{N_k}{N} \right]$$

d) $\hat{\mu}_{ML}, \hat{\sigma}^2 = ?$

(Univarijantna Gaussova razdioba)

$$\begin{aligned}f(\mu, \sigma^2) &= \ln [L(\mu, \sigma^2 | \mathcal{D})] = \ln [P(\mathcal{D} | \mu, \sigma^2)] = \ln \left[\prod_{i=1}^N P(x_i | \mu, \sigma^2) \right] \\ &= \ln \left[\prod_{i=1}^N \frac{1}{\sqrt{2\pi}} \exp \left(-\frac{1}{2\sigma^2} (x_i - \mu)^2 \right) \right] \\ &= \sum_{i=1}^N \left[\ln \left(\sqrt{2\pi} \right)^{-1} + \left(-\frac{1}{2\sigma^2} (x_i - \mu)^2 \right) \right] \\ &= \sum_{i=1}^N -\ln \sigma - \frac{1}{2} \ln (2\pi) - \frac{1}{2\sigma^2} (x_i - \mu)^2 \\ &= -N \ln \sigma - \underbrace{\frac{N}{2} \ln (2\pi)}_{\text{konstanta}} - \frac{1}{2\sigma^2} \sum_{i=1}^N (x_i - \mu)^2 \rightarrow \nabla \ln [L] = 0\end{aligned}$$

$$\frac{\partial f}{\partial \mu} = -\frac{1}{\sigma^2} \sum_{i=1}^N -2x_i + 2\mu = 0$$

$$\frac{\partial f}{\partial \sigma^2} = 0$$

$$-2 \sum_{i=1}^N x_i - \mu = 0$$

$$-N\mu + \sum_{i=1}^N x_i = 0 \Rightarrow \left[\hat{\mu}_{ML} = \frac{1}{N} \sum_{i=1}^N x_i \right]$$

$$\hookrightarrow \left[\hat{\sigma}^2_{ML} = \frac{1}{N} \sum_{i=1}^N (x_i - \hat{\mu}_{ML})^2 \right]$$

pristram!

1.3.

a)

$$\hat{\mu}_{ML} = \frac{1}{N} \sum_{i=1}^N x_i \quad \hat{\sigma}^2_{ML} = \frac{1}{N} \sum_{i=1}^N (x_i - \hat{\mu})^2$$

$$\begin{aligned} \mathbb{E}[\hat{\mu}] &= \mathbb{E}\left[\frac{1}{N} \sum x_i\right] = \frac{1}{N} \mathbb{E}\left[\sum x_i\right] = \frac{1}{N} \sum_{i=1}^N \mathbb{E}[x_i] \\ &= \frac{1}{N} \sum \mu = \frac{1}{N} \cdot N \cdot \mu = \mu \end{aligned}$$

$\Rightarrow \hat{\mu}_{ML}$ je nepristran prognozator

$$\begin{aligned} \mathbb{E}[\hat{\sigma}^2] &= \frac{1}{N} \mathbb{E}\left[\sum x_i^2 - 2\hat{\mu} \sum x_i + \hat{\mu}^2\right] \\ &= \frac{1}{N} \mathbb{E}\left[\sum x_i^2 - 2\hat{\mu} \sum x_i + \sum \mu^2\right] \\ &= \frac{1}{N} \mathbb{E}\left[\sum (x_i - \hat{\mu})^2 + N\hat{\mu}^2\right] \\ &= \frac{1}{N} \mathbb{E}\left[\sum (x_i - \hat{\mu})^2 - N\hat{\mu}^2\right] \\ &= \frac{1}{N} \sum_{i=1}^N \mathbb{E}[x_i^2] - \frac{1}{N} N \mathbb{E}[\hat{\mu}^2] \\ &= \mathbb{E}[x^4] - \mathbb{E}[\hat{\mu}^2] \quad \left\{ \begin{array}{l} \mathbb{E}(\hat{\mu}^2) = \frac{\sigma^2}{n} + \mu^2 \\ \mathbb{E}[x^4] = \sigma^2 + \mu^2 \end{array} \right. \\ &= \dots \\ &= \sigma^2 + \mu^2 - \left(\frac{\sigma^2}{n} + \mu^2\right) \\ &= \frac{n-1}{n} \sigma^2 + \sigma^2 \end{aligned}$$

$\Rightarrow \hat{\sigma}^2_{ML}$ je pristran prognozator

$$b(\hat{\sigma}^2_{ML}) = \frac{N-1}{N} \sigma^2 - \sigma^2 = -\frac{\sigma^2}{N}$$

b)

Priistranost u praksi nije problematična

- za malu uzorak \Rightarrow korekcija $\hat{\sigma}^2_{ML}$ s faktorom $\frac{N-1}{N}$

- za veliki uzorak ($N \rightarrow \infty$) $\Rightarrow b(\hat{\sigma}^2) \rightarrow 0$,

1.4

$$\mathcal{D} = \{\vec{x}^i\}_{i=1}^6$$

$$\begin{aligned}\vec{x}^1 &= (9.5, -0.7, -2.8) \\ \vec{x}^2 &= (8.8, -0.8, +3.2) \\ \vec{x}^3 &= (6.5, -0.2, -0.8) \\ \vec{x}^4 &= (2.3, 0.3, 1.2) \\ \vec{x}^5 &= (2.2, 0, 0) \\ \vec{x}^6 &= (3.6, 0.3, 1.2)\end{aligned}$$

a)

$$\hat{\mu}_{MLE} = \frac{1}{N} \sum_{i=1}^N \vec{x}^i = \left(\frac{329}{60}, \frac{-11}{60}, \frac{-11}{15} \right)$$

Multivarijatna Gaussova razdioba

$$\hat{\Sigma}_{MLE} = ?$$

$$\begin{aligned}\hat{\Sigma}_{MLE} &= \frac{1}{N} \sum_{i=1}^N (\vec{x}^i - \hat{\mu}_{MLE}) (\vec{x}^i - \hat{\mu}_{MLE})^T \\ &= \dots \\ &= \begin{bmatrix} 8.771 & -1.198 & -4.792 \\ -1.198 & 0.181 & 0.766 \\ -4.792 & 0.766 & 3.06 \end{bmatrix}\end{aligned}$$

b) $\vec{x} = (-2, 1, 0)$

$$P(\vec{x} | \hat{\mu}, \hat{\Sigma}) = \frac{1}{(2\pi)^{\frac{n}{2}} |\hat{\Sigma}|^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2} (\vec{x} - \hat{\mu})^T \hat{\Sigma}^{-1} (\vec{x} - \hat{\mu}) \right\}$$

$$|\hat{\Sigma}| = \dots = 0$$

→ Gustota nije dobro definirana jer je determinanta kovarijacijske matrice 0
 → dijeljenje s 0 + nepostojivi inverz

↳ raznaka multikolinearnosti ili konstantnih značajki

c)

$$S_{x,y} = \frac{\text{Cov}(X, Y)}{6x_1 \cdot 6y}$$

$$\hat{\Sigma}_{MLE} = \begin{bmatrix} x_1 & x_2 & x_3 \\ 8.771 & -1.198 & -4.792 \\ -1.198 & 0.191 & 0.766 \\ -4.792 & 0.766 & 3.06 \end{bmatrix} \begin{matrix} x_1 \\ x_2 \\ x_3 \end{matrix}$$

- na diagonali - varijancu
- van diagonale - kovarijancu

$$6x_1 = \sqrt{8.771}$$

$$6x_2 = \sqrt{0.191}$$

$$6x_3 = \sqrt{3.06}$$

$$\text{Cov}(X_1, X_2) = -1.198$$

$$\text{Cov}(X_1, X_3) = -4.792$$

$$\text{Cov}(X_2, X_3) = 0.766$$

$$S_{x_1, x_2} = \frac{\text{Cov}(X_1, X_2)}{6x_1 \cdot 6x_2} \approx -0.9256$$

$$S_{x_1, x_3} \approx -0.92498$$

$$S_{x_2, x_3} \approx 1.002$$

→ iz učasnog prostora izbacujemo značajku X_2
jer ima najveću korelaciju s ostalim značajkama

→ nakon izbacivanja značajke

$$\hat{\mu}_{ML} = (5.48, -0.73)$$

$$\hat{\Sigma}_{ML} = \frac{1}{N} \sum_{i=1}^N (\vec{x}_i - \hat{\mu}_{ML})(\vec{x}_i - \hat{\mu}_{ML})^T = \begin{bmatrix} 8.771 & -4.792 \\ -4.792 & 3.06 \end{bmatrix}$$

$$|\hat{\Sigma}_{ML}| = 3.8945$$

$$\hat{\Sigma}_{ML}^{-1} = \begin{bmatrix} 0.7863 & 1.2304 \\ 1.2304 & 0.2522 \end{bmatrix}$$

$$\vec{x} = (-2, 1, 0)$$

ignoriramo

$$P(\vec{x} | \hat{\mu}_{ML}, \hat{\Sigma}_{ML}) = \frac{1}{(2\pi)^{\frac{n}{2}} |\hat{\Sigma}_{ML}|^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2} (\vec{x} - \hat{\mu})^T \hat{\Sigma}_{ML}^{-1} (\vec{x} - \hat{\mu}) \right\}$$

$$\cdot [n = 2]$$

$$= \dots$$

$$= 1.0346 \cdot 10^{-8}$$

15

a)

MAP procjenitelj

$$\vec{\theta}_{MAP} = \underset{\vec{\theta}}{\operatorname{argmax}} L(\vec{\theta} | D) p(\vec{\theta})$$

$$= \underset{\vec{\theta}}{\operatorname{argmax}} p(\vec{\theta} | D)$$

Procjenitelj MAP procjenjuje maksimalnu aposteriornu distribuciju parametara. Kombinira izglednost parametra $\vec{\theta}$ (informacije iz podataka) s apriornom distribucijom parametra (prijašnje znanje).

U odnosu na MLE procjenitelji MAP

- omogućava ugradnjivanje prijašnjeg znanja u model
- omogućuje online učenje

b) $p(\vec{\theta} | D) \propto p(D | \vec{\theta}) p(\vec{\theta})$

[Bayesovo pravilo]

Ako su distribucije $p(\vec{\theta} | D)$ i $p(\vec{\theta})$ odabране tako da su to distribucije iste vrste; nazivamo ih konjugatnim distribucijama.

Konjugatna apriorna distribucija je apriorna distribucija $p(\vec{\theta})$ koja pomnožena s izglednošću daje distribuciju iste vrste kao aposteriorna.

Svojstvo konjugatnosti je bitno jer omogućava online učenje.

c)

apriorna dist: $p(\mu | \alpha, \beta) = \frac{1}{B(\alpha, \beta)} \mu^{\alpha-1} (1-\mu)^{\beta-1}$

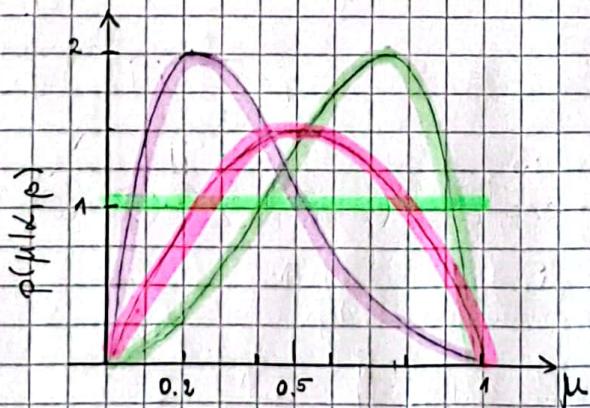
• skica za

$$\begin{aligned}\mu &= \frac{1}{2} \\ \mu &= \frac{1}{4} \\ \mu &= \frac{3}{4}\end{aligned}$$

$$\begin{aligned}\alpha &= \beta = 1 \\ \alpha &= \beta = 2 \\ \alpha &= 2, \beta = 4 \\ \alpha &= 4, \beta = 2\end{aligned}$$

• mod beta-distribucije:

$$\boxed{\mu = \frac{\alpha - 1}{\alpha + \beta - 2}}$$



d) aposteriorna distribucija
 $p(\mu | D) = ?$

$$p(\mu | D) = \frac{p(D | \mu) p(\mu)}{p(D)}$$

$p(\mu)$ = apriorna dist. (Beta)

$p(D | \mu)$ = vjerojatnost uzorka (Bernoulli) (izglednost)

$$p(\mu) = \frac{1}{B(\alpha, \beta)} \mu^{\alpha-1} (1-\mu)^{\beta-1}$$

$$p(D | \mu) = \mu^m (1-\mu)^{N-m}$$

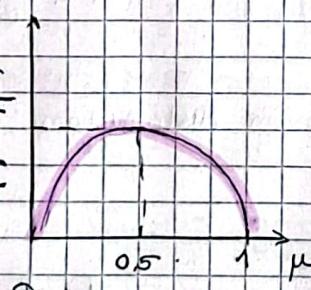
$$\begin{aligned} p(\mu | N, m, \alpha, \beta) &= \frac{1}{p(D)} \cdot \frac{1}{B(\alpha, \beta)} \mu^{\alpha-1} (1-\mu)^{\beta-1} \cdot \mu^m (1-\mu)^{N-m} \\ &= \frac{1}{p(D) B(\alpha, \beta)} \mu^{\alpha+m-1} (1-\mu)^{N-m+\beta-1} \\ &= \frac{1}{B(\alpha', \beta')} \mu^{\alpha'-1} (1-\mu)^{\beta'-1} \end{aligned}$$

$$\boxed{\begin{array}{l} \alpha' = \alpha + m \\ \beta' = N - m + \beta \end{array}}$$

c) apriorna d. $p(\mu | \alpha=2, \beta=2)$. $\rightarrow \text{mod} = \frac{\alpha-1}{\alpha+\beta-2} = \frac{1}{2}$
 $N=10, m=1$

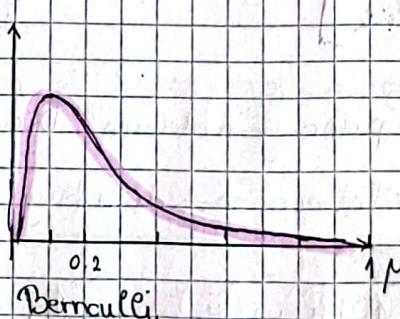
$$\alpha(\mu | N, m) = \mu(1-\mu)^9 \rightarrow \text{mod} = \frac{1}{N} = 0.1$$

$$p(\mu | \alpha, \beta)$$



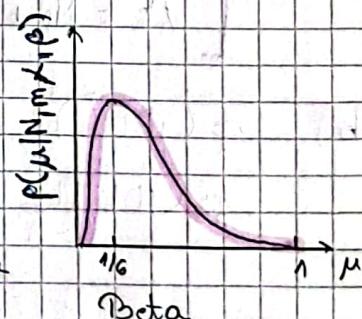
Beta

$$f(\mu | N, m)$$



Bernoulli

$$p(\mu | N, m, \alpha, \beta)$$



Beta

aposteriorna dist.

$$\alpha' = m + \alpha = 2 + 1 = 3$$

$$\beta' = N - m + \beta = 9 + 2 = 11$$

$$\text{mod} = \frac{\alpha'-1}{\alpha'+\beta'-2} = \frac{2}{12} = \frac{1}{6}$$

f)

$$\hat{\mu}_{MAP} = \frac{\lambda' - 1}{\lambda' + \beta - 2} = \frac{m + \lambda - 1}{m + \lambda + N - m + \beta - 2} = \frac{m + \lambda - 1}{\lambda + \beta + N - 2} = \frac{\lambda + \beta - 1}{2\lambda + 2 + 10 - 2}$$

$$= \frac{2}{12} = \frac{1}{6} \approx 0.1667$$

$$\hat{\mu}_{MLE} = \frac{m}{N} = \frac{1}{10} = 0.1$$

Porastom broja primjera N razlika između $\hat{\mu}_{MAP}$ i $\hat{\mu}_{MLE}$ bi se smanjila jer vekt λ utječe na $\hat{\mu}_{MAP}$ i ima N u razinama - više se vjeruje podacima, a ne apriornom znanju

g) $\lambda = \beta = 2 \rightarrow \hat{\mu}_{MAP}$ se svodi na Laplaceov projekcijelj

$$\hat{\mu}_{MAP} = \frac{m + \lambda - 1}{\lambda + \beta + N - 2} = \frac{m + 1}{N + 2}$$

1.6. Dirichlet - kategorički model

a) Dirichletova distribucija

$$P(\vec{\mu} | \vec{\lambda}) = \frac{1}{B(\vec{\lambda})} \prod_{k=1}^K \mu_k^{\lambda_k - 1} \quad \sum_{k=1}^K \mu_k = 1 \quad \mu_k \geq 0$$

$$\left[\hat{\mu}_{k,MAP} = \frac{\lambda_k - 1}{\sum_{k=1}^K \lambda_k - K} \right]$$

$$\left[\lambda'_k = N_k + \lambda_k \right]$$

b) $P(\vec{\mu} | \vec{\lambda}, D)$ - prior je modeliran Dirichletovom distribucijom
 $P(D | \vec{\mu})$ - kategorička distribucija.

$$\begin{aligned} P(\vec{\mu} | \vec{\lambda}, D) &= \frac{P(D | \vec{\mu}) P(\vec{\mu})}{P(D)} = \frac{1}{P(D)} \prod_{i=1}^N \prod_{k=1}^K \mu_k^{x_{k,i}} \cdot \frac{1}{B(\vec{\lambda})} \prod_{k=1}^K \mu_k^{\lambda_k - 1} \\ &= \frac{1}{B(\vec{\lambda}) P(D)} \prod_{k=1}^K \mu_k^{N_k} \cdot \prod_{k=1}^K \mu_k^{\lambda_k - 1} \\ &= \frac{1}{B(\vec{\lambda}) P(D)} \prod_{k=1}^K \mu_k^{\lambda_k + N_k - 1} = \frac{1}{B(\vec{\lambda})} \prod_{k=1}^K \mu_k^{\lambda_k + N_k - 1} \end{aligned}$$

$$\left[\lambda'_k = N_k + \lambda_k \right]$$

$$\text{mod } \left[\hat{\mu}_{k,MAP} = \frac{\lambda'_k - 1}{\sum_{k=1}^K \lambda'_k - K} = \frac{\lambda_k + N_k - 1}{N + \sum_{k=1}^K \lambda_k - K} \right]$$

$$d_k = 2$$

$$\hat{\mu}_{k, \text{MAP}} = \frac{d_k + N_k - 1}{\sum d_k + N - K} = \frac{N_k + 1}{2K + N - K} = \frac{N_k + 1}{N + K}$$

1.7.

a) MLE procjena za \vec{w} (kod linearne regresije)

$$y \sim \mathcal{N}(\mu, \sigma^2)$$

$$\hat{\vec{w}}_{\text{MLE}} = \underset{\vec{w}}{\operatorname{argmax}} [C_n \cdot L(\vec{w} | D)]$$

$$\begin{aligned} C_n [L(\vec{w} | D)] &= C_n [\ln p(D | \vec{w})] = C_n \left[\prod_{i=1}^N p(y^i | \vec{x}^i) \right] \\ &= C_n \left[\prod_{i=1}^N \mathcal{N}(y^i; h(\vec{x}^i; \vec{w}), \sigma^2) \right] \end{aligned}$$

$$= C_n \left[\prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left(-\frac{1}{2\sigma^2} (y^i - h(\vec{x}^i))^2 \right) \right]$$

$$= \sum_{i=1}^N -C_n \ln \sqrt{2\pi\sigma^2} + -\frac{1}{2\sigma^2} (y^i - h(\vec{x}^i))^2$$

$$= -\underbrace{N \ln \sigma - \frac{N}{2} \ln 2\pi}_{\text{konstanta}} - \frac{1}{2\sigma^2} \sum_{i=1}^N (y^i - h(\vec{x}^i; \vec{w}))^2$$

$$\propto \underbrace{-\frac{1}{2} \sum_{i=1}^N (y^i - h(\vec{x}^i; \vec{w}))^2}_{E(\vec{w} | D)}$$

• maksimizacija log-izglednosti \Leftrightarrow minimizacija kvadratne pogreške

b) MLE procjena za \vec{w} (kod logističke regresije)

$$y \sim \text{Bin}(\mu)$$

$$\hat{\vec{w}}_{\text{MLE}} = \underset{\vec{w}}{\operatorname{argmax}} [C_n L(\vec{w} | D)]$$

$$\begin{aligned} C_n L(\vec{w} | D) &= C_n [\ln p(\vec{y} | \vec{w})] = C_n \left[\prod_{i=1}^N p(y^i | \vec{w}) \right] \\ &= C_n \left[\prod_{i=1}^N \mu^{y^i} (1-\mu)^{1-y^i} \right] \\ &= \sum_{i=1}^N y^i \ln(\mu) + (1-y^i) \ln(1-\mu) \end{aligned}$$

• maksim. log izglednosti \Leftrightarrow minimizacija unakrsne pogreške

c) i d) - skip.

II zadaci s ispita

2.1.

$$\begin{array}{c} (x^i, y^i) \\ \cdot (-2, 1) \\ \cdot (-2, 1) \\ (-1, 0) \\ \cdot (0, 0) \\ \cdot (1, 1) \\ \cdot (3, 1) \end{array}$$

$$p(x|y) \sim N(\mu, \sigma^2)$$

$$\begin{aligned} L(\mu_1, \sigma_1^2 | D_{y=1}) &= p(D_{y=1} | \mu_1, \sigma_1^2) \\ &= \prod_{i=1}^4 p(x^i | \mu_1, \sigma_1^2) \\ &= \prod_{i=1}^4 \frac{1}{\sigma_1 \sqrt{2\pi}} \exp \left\{ -\frac{1}{2\sigma_1^2} (x^i - \mu_1)^2 \right\} \\ &= \frac{1}{\prod_{i=1}^4 \sqrt{2\pi}} \exp \left\{ -\frac{1}{2\sigma_1^2} \sum_{i=1}^4 (x^i - \mu_1)^2 \right\} \\ &= \frac{1}{3\sqrt{\pi}} \exp \frac{-\sum_{i=1}^4 (x^i)^2}{9}. \end{aligned}$$

$$\text{podesup} \quad \hat{\mu}_1 = \frac{1}{N} \sum_{i=1}^N x^i = \frac{1}{4} (-2-2+1+3) = 0$$

$$\begin{cases} D_{y=1} \\ N=4 \end{cases} \quad \hat{\sigma}_1^2 = \frac{1}{N} \sum_{i=1}^N (x^i - \hat{\mu}_1)^2 = \frac{9}{2}$$

log - izglednost

$$\begin{aligned} \ln L(\mu_1, \sigma_1^2 | D) &= \ln \prod_{i=1}^4 \frac{1}{3\sqrt{\pi}} \exp \frac{-1}{9} x^i \\ &= \sum_{i=1}^4 \ln \frac{1}{3\sqrt{\pi}} + -\frac{1}{9} \sum x^i \\ &= -4 \ln \frac{1}{3\sqrt{\pi}} - \frac{1}{9} \sum x^i = -8.684 \Rightarrow \textcircled{B} \end{aligned}$$

2.2.

$$D = \{-3, -2, 1, 4\}$$

kontinuir. slučajna varijabla

$$\Rightarrow N(\mu, \sigma^2)$$

$$\hat{\mu}_{MLE} = \frac{1}{N} \sum x^i = 0$$

$$\hat{\sigma}_{MLE}^2 = \frac{1}{N} \sum (x^i - \hat{\mu}_{MLE})^2 = \frac{15}{2}$$

$$\hat{\sigma}_{US}^2 = \frac{1}{N-1} \sum (x^i - \hat{\mu}_{MLE})^2 = 10$$

$$\begin{aligned} \ln [L(\mu, \sigma^2 | D)] &= \ln \left[\prod_{i=1}^N p(x^i | \mu, \sigma^2) \right] \\ &= \ln \left[\prod_{i=1}^N \frac{1}{\sigma \sqrt{2\pi}} \exp \left(-\frac{1}{2\sigma^2} (x^i - \mu)^2 \right) \right] \\ &= \sum_{i=1}^N \left[-\ln \sigma - \frac{1}{2} \ln 2\pi - \frac{1}{2\sigma^2} (x^i - \mu)^2 \right] \\ &= -N \ln \sigma - \underbrace{\frac{N}{2} \ln 2\pi}_{\text{konst}} - \frac{1}{2\sigma^2} \sum_{i=1}^N (x^i - \mu)^2 \end{aligned}$$

$$\begin{aligned} \ln L_1 &= -9.7055 \\ \ln L_2 &= -9.7809 \end{aligned}$$

$$\Delta L = \ln L_1 - \ln L_2 = 0.075 \Rightarrow \textcircled{C}$$

2.3.

$$L(\mu, \sigma^2 | D) = \prod_{i=1}^N \frac{1}{\sigma \sqrt{2\pi}} \exp\left(-\frac{1}{2}(x_i - \mu)^2\right)$$

= ...

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

$$\hat{\mu}_{MLE} = \frac{1}{N} \sum x_i$$

$D' \subset D$

$$\hat{\sigma}^2_{MLE} = \frac{1}{N} \sum (x_i - \hat{\mu}_{MLE})^2$$

$$\hat{\sigma}^2_{UB} = \frac{N}{N-1} \hat{\sigma}^2_{MLE}$$

$$L_0 = L(\hat{\mu}_{MLE}, \hat{\sigma}^2_{MLE} | D)$$

$$L_1 = L(0, 1 | D)$$

$$L_2 = L(\hat{\mu}_{MLE}, \hat{\sigma}^2_{UB} | D)$$

$$L_3 = L(\hat{\mu}_{MLE}, \hat{\sigma}^2_{UB} | D')$$

$\hat{\sigma}^2_{MLE}$ podjednjuje $\hat{\sigma}^2$ ($b = -\frac{\hat{\sigma}^2}{N}$)

• MLE = most likelihood estimation (najveći je za MLE prognoziraju)
→ $L_0 \geq L_1 \quad L_0 > L_2$
 $(\hat{\sigma}^2_{UB} \neq \hat{\sigma}^2_{MLE})$

→ otpada B ($L_1 \neq L_0$) i C ($L_0 < L_3, L_0 \leq L_2$)

A odgovor netačan

$L_1 \neq L_2 \Rightarrow$ može se desiti da je skup D' baš takav

da $\hat{\mu}_{MLE} = 0$ i $\hat{\sigma}^2_{UB} = 1$

- u principu o L_1 ne možemo ništa zaključiti osim da je

$$L_0 \geq L_1$$

D

$L_0 \geq L_1$ } zbroj pravila MLE

$L_0 > L_2$ - ovise tako da D' uzrokuju može se dogoditi

$L_2 > L_3$ i $L_2 < L_3$, ali ne i $L_2 = L_3$ jer je

D' podskup od D

2.4.

$$\alpha = \beta = 2$$

$$\hat{\mu}_{MAP} = ?$$

① D_1

$$N_1 = 50 \\ m_1 = 42$$

② D_2

$$N_2 = 15 \\ m_2 = 3$$

- "online" učenje

• učenje na skupu D_1

$$\alpha' = \beta' = 2$$

$$\alpha' = \alpha + m_1 = 2 + 42 = 44 \\ \beta' = \beta + N_1 - m_1 = 2 + 50 - 42 = 2 + 8 = 10$$

$$\hat{\mu}_{MAP,1} = \frac{\alpha' - 1}{\alpha' + \beta' - 2} = \frac{43}{54 - 2} = \frac{43}{52}$$

• učenje na skupu D_2

$$\alpha = 44$$

$$\beta = 10$$

$$\alpha' = \alpha + m_2 = 44 + 3 = 47$$

$$\beta' = \beta + N_2 - m_2 = 10 + 15 - 3 = 22$$

$$\hat{\mu}_{MAP,2} = \frac{\alpha' - 1}{\alpha' + \beta' - 2} = \frac{46}{67}$$

$$\Delta \mu = \hat{\mu}_{MAP,1} - \hat{\mu}_{MAP,2} = 0.1404$$

$$\Rightarrow \textcircled{C} \quad \Delta \mu = \hat{\mu}_2 - \hat{\mu}_1 = -0.14$$

2.5.

$$\hat{\mu}_{e,MAP}$$

$$K = 3$$

• Dirichlet - katg. model
 $(\alpha_1, \alpha_2, \alpha_3) = (2, 2, 1)$ \leftarrow apriorno

• realizacija $\cdot (0, \frac{1}{2}N, \frac{1}{2}N)$

$$\overrightarrow{\mu}_{MAP} = ?$$

$$\alpha'_e = \alpha_e + N_e$$

$$\hat{\mu}_{e,MAP} = \frac{\alpha'_e - 1}{\sum \alpha'_e - K}$$

$$\begin{cases} \alpha'_1 = 2 + 0 = 2 \\ \alpha'_2 = 2 + \frac{1}{2}N = \frac{N+4}{2} \\ \alpha'_3 = 1 + \frac{1}{2}N = \frac{N+2}{2} \end{cases}$$

$$\boxed{\sum \alpha'_e - K = \sum_{e=1}^3 \alpha'_e - 3 = \frac{2N+6}{2} + 2 - 3 = N + 3 + 2 - 3 = N + 2}$$

$$\hat{\mu}_1 = \frac{2-1}{N+2} = \frac{1}{N+2}$$

$$\hat{\mu}_2 = \frac{\frac{N+4}{2} - 1}{N+2} = \frac{\frac{N+2}{2}}{N+2} = \frac{1}{2}$$

$$\hat{\mu}_3 = \frac{\frac{N+2}{2} - 1}{N+2} = \frac{N}{2N+4}$$

- A ne može $\hat{\mu}_1 \neq 0$
 B ne može $\hat{\mu}_2 = 1/2$
 C ne može

(nije nujno da $\hat{\mu}_1 < \hat{\mu}_2$)

D $0 < \hat{\mu}_1 < \frac{1}{3}$

$\frac{1}{3} < \hat{\mu}_2 < 1$

$0 < \hat{\mu}_3, \hat{\mu}_2 < 1$

$$N=2$$

$$\hat{\mu}_1 = \frac{1}{4} = 0.25$$

$$\hat{\mu}_2 = \frac{1}{2} = 0.5$$

$$\hat{\mu}_3 = \frac{1}{8} = \frac{1}{4} = 0.25$$

2.7

$\vec{\mu}_{MAP}$ \rightarrow Dirichlet - kateg. model

$$K=3$$

$$\text{prior: } P(\vec{\mu} | \vec{\alpha})$$

$$\vec{\alpha} = \alpha(1, 4, 2), \alpha = 1$$

$$\sum \alpha_i = 7\alpha$$

$$N_1 = 50$$

$$N_2 = 29$$

$$N_3 = 40$$

$$N = 119$$

$$P(\vec{\mu} | \vec{\alpha}) = \frac{K}{\prod_{i=1}^K \alpha_i^{x_i + \alpha_i - 1}} \cdot \frac{1}{B(K)}$$

MAP proc.

$$\begin{aligned} \vec{\mu}_{MAP} &= \operatorname{argmax}_{\vec{\mu}} P(D | \vec{\mu}) P(\vec{\mu} | \vec{\alpha}) = \operatorname{argmax}_{\vec{\mu}} P(\vec{\mu} | D) \\ &= \operatorname{argmax}_{\vec{\mu}} P(\vec{x} | \vec{\mu}) P(\vec{\mu} | \vec{\alpha}) \\ &= \operatorname{argmax}_{\vec{\mu}} L(\vec{\mu} | \vec{x}) P(\vec{\mu}) \end{aligned}$$

$$\vec{\mu}_{MAP} = \frac{\vec{\alpha} + 1}{\sum \alpha_i + K} = \frac{N_1 + N_2 + 1}{\sum N_i + N - K} = \frac{N_1 + N_2 - 1}{7\alpha + 119}$$

$$\vec{\mu}_{MAP} = \left(\frac{N_1 + 49}{7\alpha + 119}, \frac{N_2 + 28}{7\alpha + 119}, \frac{N_3 + 38}{7\alpha + 119} \right)$$

$$4\alpha + 28 > \alpha + 49$$

$$3\alpha > 21$$

$$\alpha > 7$$

$$2\alpha > 11$$

$$\alpha > 5.5$$

$$\Leftrightarrow \boxed{\alpha > 7}$$

C

2.6.

$$\vec{x} = (x_1, \dots, x_6)$$

μ_{MAP}

• apriorna PDF \rightarrow Dirichlet

• realizacija \rightarrow češće 6-ica (u stvarnosti)

• opažanje \rightarrow najčešće 5-ica, ali malo broj opažanja

$\xrightarrow{\text{L}}$ određuje vjerojatnost $\vec{\mu}$, už $\sum_{k=1}^6 \mu_k = 1$

Uz $\vec{\mu}$ je \vec{L} naša prognoza za $\vec{\mu}$ biti najbliža stvarnoj vrijednosti tih parametara?

(a)

$$\vec{L} = (1, 1, 1, 1, 1, 1)$$

\rightarrow svačka kombinacija $\vec{\mu}$ jednako vjerojatna (ne uzima apriorna znanje, vjerujemo podacima)

b)

$$\vec{L} = (\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6})$$

\rightarrow najvjerojatniji $\vec{\mu}$ na rubovima simpleksa

c)

$$\vec{L} = (2, 2, 2, 2, 2, 2)$$

\rightarrow $\vec{\mu}$ najvjerojatnije svr medusobno jednake ($\mu_i = \mu_j$, jer je $\lambda_i > 1$)

d)

$$\vec{L} = (1, 1, 1, 1, 3, 1)$$

\rightarrow osimetrne kombinacije daju veću vjerojatnost kombinacijama kod kojih ne nisu jednaki (najčešći je 5. ishod)

2.8.

Beta-Bernoulli model

$$\mu_{MAP} = ?$$

$$\mu_{MLE} = 0.2 = \frac{m}{N} \rightarrow m = 0.2N$$

$$\mu_{MAP} = \frac{\lambda - 1}{\lambda + \beta - 2} = \frac{\lambda + m - 1}{\lambda + \beta + N - 2} = 0.2$$

$$\lambda + m - 1 = 0.2\lambda + 0.2\beta + 0.2N - 0.4$$

$$0.8\lambda - 0.2\beta = 0.6$$

$$8\lambda - 2\beta = 6$$

$$4\lambda - \beta = 3$$

$$\xrightarrow{\text{L}} \beta = 4\lambda - 3$$

$$\begin{aligned} \lambda &= 2 \\ \beta &= 5 \end{aligned}$$

\rightarrow A

15. Bayesov klasifikator

Strojno učenje 1, UNIZG FER, ak. god. 2023./2024.

Jan Šnajder, vježbe, v3.2

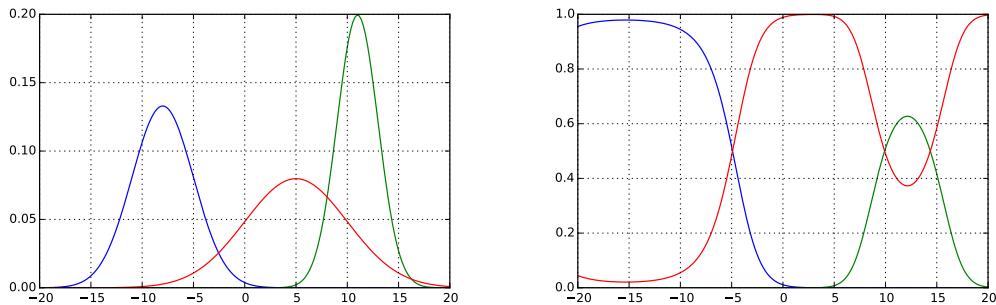
1 Zadatci za učenje

1. [Svrha: Razumjeti model Bayesovog klasifikatora i njegove komponente. Razumjeti što su to generativni modeli, kako se razlikuju od diskriminativnih te koje su njihove prednosti i njihovi nedostatci.]
 - (a) Definirajte model Bayesovog klasifikatora i navedite sve veličine koje se pojavljuju u definiciji modela. Objasnite zašto faktoriziramo brojnik. Objasnite ulogu nazivnika i objasnite kada ga možemo zanemariti.
 - (b) Je li taj model parametarski ili neparametarski? Obrazložite odgovor.
 - (c) Objasnite zašto Bayesov klasifikator nazivamo generativnim i opišite generativnu priču Bayesovog klasifikatora.
 - (d) Objasnite razliku između generativnih i diskriminativnih modela te navedite prednosti jednih i drugih.
2. [Svrha: Isprobati izračun maksimalne aposteriorne hipoteze i najvjerojatnije hipoteze uz minimizaciju rizika.] Razmotrimo problem klasifikacije neželjene el. pošte u klase *spam* ($y = 1$), *important* ($y = 2$) i *normal* ($y = 3$). Neka su apriorne vjerojatnosti tih klasa $P(y = 1) = 0.2$, $P(y = 2) = 0.05$ i $P(y = 3) = 0.75$. Za neku poruku el. pošte \mathbf{x} izglednosti iznose $p(\mathbf{x}|y = 1) = 0.8$ i $p(\mathbf{x}|y = 2) = p(\mathbf{x}|y = 3) = 0.5$. Izračunajte aposteriorne vjerojatnosti za svaku od klasa te maksimalnu aposteriornu hipotezu za primjer \mathbf{x} .
3. [Svrha: Razviti intuiciju za model kontinuiranog Bayesovog klasifikatora.] Izrađujemo Bayesov model za klasifikaciju primjera iz $\mathcal{X} = \mathbb{R}$ u tri klase. Učenjem na skupu primjera dobili smo sljedeće parametre modela: $P(y = 1) = 0.3$, $P(y = 2) = 0.2$, $\mu_1 = -5$, $\mu_2 = 0$, $\mu_3 = 5$, $\sigma_1^2 = 5$, $\sigma_2^2 = 1$, $\sigma_3^2 = 10$. Skicirajte funkcije gustoće vjerojatnosti $p(x|y)$, $p(x,y)$, $p(x)$ i $p(y|x)$.
4. [Svrha: Razumjeti izvod modela kontinuiranog Bayesovog klasifikatora i osvježiti potrebno znanje matematike.]
 - (a) Krenuvši od izraza (4.29) iz skripte, izvedite model višedimenzijskog Bayesovog klasifikatora s kontinuiranim ulazima s dijeljenom i diagonalnom kovarijacijskom matricom.
 - (b) Napišite broj parametara ovog modela.
 - (c) Objasnite zašto je izglednost faktorizirana u produkt univariatnih razdioba, što odgovara pretpostavci o uvjetnoj nezavisnosti, premda značajke mogu biti nelinearno uvjetno zavisne.
5. [Svrha: Razviti intuiciju za složenost modela kontinuiranog Bayesovog klasifikatora i shvatiti kako se problem u konačnici svodi na odabir optimalnog modela.] Želimo izgraditi klasifikator za klasifikaciju brucoša u jednu od dvije klase: $y = 1 \Rightarrow$ "Završava FER u roku" i $y = 2 \Rightarrow$ "Produljuje studij". Svaki je primjer opisan sa šest ulaznih varijabli: prosjek ocjena 1.–4. razreda (četiri varijable), bodovi državne mature iz matematike te bodovi državne mature iz fizike. Raspolažemo trima modelima: modelom \mathcal{H}_1 s dijeljenom kovarijacijskom matricom, modelom \mathcal{H}_2 s diagonalnom (i dijeljenom) kovarijacijskom matricom i modelom \mathcal{H}_3 s izotropnom kovarijacijskom matricom.

- (a) Koliko svaki od ova tri modela ima parametara?
- (b) Za koji od ova tri modela očekujete da će najbolje generalizirati u ovom konkretnom slučaju (uzmite u obzir prirodu problema i očekivane odnose između značajki)? Zašto?
- (c) Nacrtajte skicu funkcije empirijske pogreške i pogreške generalizacije i naznačite na njoj točke koje označavaju navedenim trima modelima.
- (d) Kako biste u praksi odredili koji će model upotrijebiti?

2 Zadaci s ispita

1. (P) Koristimo Gaussov Bayesov klasifikator kako bismo riješili troklasni klasifikacijski problem. Procijenjene gustoće vjerojatnosti za izglednosti klasa su $p(x|y = 1) = \mathcal{N}(-8, 3)$, $p(x|y = 2) = \mathcal{N}(5, 5)$ i $p(x|y = 3) = \mathcal{N}(11, 2)$. Na slikama ispod prikazane su izglednosti klasa (lijeva slika) i aposteriorne vjerojatnosti dobivene Bayesovim pravilom (desna slika):



S obzirom na ova dva grafikona, što su najizglednije vrijednosti za apriorne vjerojatnosti klasa?

- A $P(y = 1) = 0.1, P(y = 2) = 0.7, P(y = 3) = 0.2$
- B $P(y = 1) = P(y = 2) = P(y = 3) = \frac{1}{3}$
- C $P(y = 1) = P(y = 2) = 0.4, P(y = 3) = 0.2$
- D $P(y = 1) = P(y = 2) = 0.1, P(y = 3) = 0.8$

2. (P) Gaussovim Bayesovim klasifikatorom rješavamo problem klasifikacije u $K = 10$ klasa sa $n = 5$ značajki. Prisjetite se da kod Gaussovog Bayesovog klasifikatora uvođenjem odgovarajućih pretpostavki na kovarijacijsku matricu Σ možemo utjecati na broj parametara modela a time onda i na složenost modela. Razmatramo tri modela s kovarijacijskim matricama u koje smo ugradili sljedeće pretpostavke:

\mathcal{H}_1 : Značajke nisu korelirane, no imaju različite varijance unutar klase i između klasa

\mathcal{H}_2 : Značajke nisu korelirane, imaju jednaku varijancu unutar svake klase, no različitu za svaku klasu

\mathcal{H}_3 : Između značajki postoje korelacije, ali se one ne razlikuju između klasa

Neka ' \supset ' označava relaciju "složeniji od", a neka ' $>$ ' označava relaciju "ima više parametara od". **Što možemo zaključiti o složenosti i broju parametara za gornja četiri modela?**

- A $\mathcal{H}_1 > \mathcal{H}_3 > \mathcal{H}_2, \mathcal{H}_1 \supset \mathcal{H}_2$
- B $\mathcal{H}_1 > \mathcal{H}_2 > \mathcal{H}_3, \mathcal{H}_1 \supset \mathcal{H}_2 \supset \mathcal{H}_3$
- C $\mathcal{H}_3 > \mathcal{H}_1 > \mathcal{H}_2, \mathcal{H}_1 \supset \mathcal{H}_2$
- D $\mathcal{H}_3 > \mathcal{H}_1 > \mathcal{H}_2, \mathcal{H}_3 \supset \mathcal{H}_2 \supset \mathcal{H}_1$

3. (N) Na skupu označenih primjera u ulaznome prostoru dimenzije $n = 3$ treniramo Gaussov Bayesov klasifikator za klasifikaciju primjera u $K = 2$ klase, uz pretpostavku dijeljene kovarijacijske matrice. Model je definiran kao

$$h_j(\mathbf{x}) = \ln p(\mathbf{x}, y)$$

Prisjetimo se da je izglednost klase s oznakom $y = j$ kod Gaussovog Bayesovog klasifikatora definirana multivariantnom Gaussovom gustoćom vjerojatnosti:

$$p(\mathbf{x}|y = j) = \frac{1}{(2\pi)^{n/2}|\Sigma_j|^{1/2}} \exp \left\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1} (\mathbf{x} - \boldsymbol{\mu}_j) \right\}$$

gdje je Σ_j matrica kovarijacije za klasu j . Treniranjem modela dobili smo sljedeće procjene za parametre:

$$\begin{aligned} \hat{\mu}_1 &= 0.2 & \hat{\boldsymbol{\mu}}_1 &= (1, 0, -2) & \hat{\Sigma}_1 &= \begin{pmatrix} 5 & 2 & 4 \\ 2 & 5 & 3 \\ 4 & 3 & 6 \end{pmatrix} \\ \hat{\mu}_2 &= 0.8 & \hat{\boldsymbol{\mu}}_2 &= (2, -1, 5) & \hat{\Sigma}_2 &= \begin{pmatrix} 6.25 & -0.5 & -1 \\ -0.5 & 1.25 & -0.75 \\ -1 & -0.75 & 3.5 \end{pmatrix} \end{aligned}$$

Iz ovoga smo zatim procijenili dijeljenu kovarijacijsku matricu $\hat{\Sigma}$ definiranu kao težinski prosjek kovarijacijskih matrica $\hat{\Sigma}_j$, $j = 1, 2$. Zanima nas klasifikacija modela za primjer $\mathbf{x} = (0, 0, 0)$. **Koliko iznosi predikcija modela za klasu $y = 1$ za taj primjer, $h_1(\mathbf{x})$?**

- A] -6.885 B] +0.002 C] -4.819 D] -6.429

4. (P) Gaussov Bayesov klasifikator koristimo za klasifikaciju jednodimenzionalnih podataka u tri klase. Procijenjene izglednosti klase su $p(x|y = 1) = \mathcal{N}(-10, 2)$, $p(x|y = 2) = \mathcal{N}(2, 2)$ i $p(x|y = 3) = \mathcal{N}(8, 2)$, a procijenjene apriorne vjerojatnosti klase su $P(y = 1) = P(y = 2) = 2/5$ i $P(y = 3) = 1/5$. Međutim, nakon što smo naučili ovaj model, zaključili smo da na ispitnom skupu postoji pomak u distribuciji podataka u odnosu na skup za učenje te da zbog toga model ne generalizira dobro. Zaključili smo da se ovo može ispraviti tako da se naučeni model malo izmjeni, i to tako da se varijanca izglednosti klase $y = 1$ postavi na 5 i da se apriorne vjerojatnosti klase ujednače, $P(y = 1) = P(y = 2) = P(y = 3) = 1/3$. Skicirajte gustoće zajedničke vjerojatnosti naučenog i izmijenjenog modela. Neka su h_1 i h_2 MAP-hipoteze prvog i drugog modela, te neka su a i b pozitivne konstante. Razmotrite segment ulaznog prostora za koji vrijedi $-10 \leq x \leq 10$. **Na kojim se dijelovima tog segmenta ulaznog prostora MAP-hipoteze prvog i drugog modela razlikuju?**

- A] $[-4 - a, 5 + b]$
 B] $[-4 - a, -4 + b]$
 C] $[-4 - a, -4] \cup [5 - b, 5]$
 D] $[-4, -4 + a] \cup [5, 5 + b]$

5. (P) Gaussov Bayesov klasifikator koristimo za klasifikaciju u dvije klase ($y = 1$ i $y = 2$) u dvodimenzionalnom ulaznom prostoru ($\mathcal{X} = \mathbb{R}^2$). Apriorne vjerojatnosti klase su jednake, dok su izglednosti klase modelirane bivarijatnim Gaussovim distribucijama sa sljedećim parametrima:

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 3 \\ 6 \end{pmatrix} \quad \boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} \quad \boldsymbol{\mu}_2 = \begin{pmatrix} 6 \\ 3 \end{pmatrix} \quad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}$$

Skicirajte gustoću zajedničke vjerojatnosti u ulaznom prostoru i granicu između klasa definiranu jednadžbom $h(x_1, x_2) = 0$. **Koje su od sljedećih točaka (x_1, x_2) najbliže točkama kroz koje prolazi ta granica?**

- A] (1, 2), (3, 4), (6, 7) B] (1, 3), (5, 4), (7, 9) C] (1, 6), (3, 3), (5, 0) D] (3, 2), (4, 5), (9, 7)

6. (N) Na skupu označenih primjera u ulaznome prostoru dimenzije $n = 2$ treniramo Gaussov Bayesov klasifikator za klasifikaciju primjera u $K = 2$ klase, uz pretpostavku dijeljene i dijagonalne kovarijacijske matrice. Izglednost klase s oznakom $y = j$ definirana je multivarijantnom Gaussovom gustoćom vjerojatnosti:

$$p(\mathbf{x}|y=j) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}_j|^{1/2}} \exp \left\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1} (\mathbf{x} - \boldsymbol{\mu}_j) \right\}$$

Model treniramo na skup podataka od $N = 7$ primjera:

$$\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_i = \{((-1, -2), 0), ((0, 0), 0), ((1, 2), 0), ((3, -1), 1), ((4, -1), 1), ((4, 1), 1), ((5, 1), 1)\}$$

Procijenite parametre modela na ovom skupu primjera. Budući da je skup primjera malen, za procjenu kovarijacijske matrice koristite nepristran procjenitelj. Izlaz modela za klasu $y = j$ neka je zajednička gustoća vjerojatnosti, $h_j(\mathbf{x}) = \ln p(\mathbf{x}, y = j)$. **Koliko iznosi $h_0(\mathbf{x})$ za primjer $\mathbf{x} = (0, 0)$?**

- A -4.13 B -3.84 C -3.03 D -2.75

V15 - Bayesov klasifikator

I Zadaci za učenje

1.1.

a) model Bayesovog klasifikatora $\vec{\theta}$ = vektor parametara distrib.

$$f_j(\vec{x}; \vec{\theta}) = P(y=j | \vec{x}) = \frac{P(\vec{x}|y) P(y)}{\sum_{y'} P(\vec{x}|y') P(y')} = \frac{P(\vec{x}|y) P(y)}{P(\vec{x})}$$

$P(\vec{x}|y)$ = izglednost klase

- gustoća vjerojatnosti primjera uz zadani klasi

"vjerojatnost primjera za zadani klasi"

$P(y)$ = apriorna vjerojatnost klase

$P(\vec{x})$ = gustoća vjerojatnosti primjera nevisno o klasi

• brojnik faktoriziramo kako bismo tako modelirali složenu zajednicku distribuciju $P(\vec{x}, y)$

• $P(\vec{x})$ dolazi iz Bayesovog pravila i modelira gustoću vjerojatnosti primjera

- slično kao i PDF $P(\vec{x}, y)$ može biti složena distribucija pa ju faktoriziramo na $\sum P(x|y') P(y')$

• ako ne značimo vjerojatnost klasifikacije primjera u klasi j razinom možemo zanemariti da model postaje

$$P(\vec{x}; \vec{\theta}) = \operatorname{argmax}_y P(\vec{x}|y) P(y)$$

\rightarrow MAP HIPOTEZA

b) Bayesov klasifikator jest parametarski model

\hookrightarrow model pretpostavlja da se primjer \vec{x} i oznake y poskoravaju nekoj teorijskoj distribuciji

c) Bayesov klasifikator nazivamo generativnim jer modelira zajednicku vjerojatnost $P(\vec{x}, y)$

\rightarrow modelira postupak generiranja podataka

GENERATIVNA PRICA BAY. KLAS.

Primjeri $D = \{(\vec{x}^i, y^i)\}$ su nastaju u 2 koraka

① odabrana je oznaka y po distribuciji $P(y)$

② Za odabranu označku y odabran je primjer \vec{x} po distribuciji $P(\vec{x}|y)$

d)

Generativni modeli

- modeliraju zajed. distribuciju $P(\vec{x}, y)$

Discriminativni modeli

- izravno modeliraju aposteriornu vjerojatnost $P(y|\vec{x})$
- dobiva se decision granica između klasa

Prednosti gener. modela

- 1) Čak ugraditi apriorna vjerojatnost o problemu
→ apriorna distribucija
- 2) intuitivna interpretabilnost rezultata i mogućnost različitih analiza
 - objena pouzdanosti klasifikacije
→ izlaz modela može tumačiti kao pouzdanost
 - odabijanje klasifikacija
↳ pouzdanost manja od proga
 - Čak detektiranje stršćih vrijednosti

Nedostaci generat. modela

- 1) velik broj parametara
- 2) repetitivna složenost modeliranja

1.2.

apriorne vjeroj.

$$\begin{aligned}P(y=1) &= 0.2 \\P(y=2) &= 0.05 \\P(y=3) &= 0.75\end{aligned}$$

$$P(y|\vec{x}) = ? = h_j(\vec{x})$$

$$h_j(\vec{x}) = \frac{P(\vec{x}|y)P(y)}{P(\vec{x})}$$

izg

$$\begin{aligned}P(\vec{x}|y=1) &= 0.8 \\P(\vec{x}|y=2) &= 0.5 \\P(\vec{x}|y=3) &= 0.5\end{aligned}$$

$$\begin{aligned}P(\vec{x}) &= \sum_{y=1}^3 P(\vec{x}|y)P(y) \\&= 0.8 \cdot 0.2 + 0.5 \cdot 0.05 + \\&\quad 0.5 \cdot 0.75 \\&= 0.56\end{aligned}$$

$$h_1(\vec{x}) = P(y=1|\vec{x}) = \frac{P(\vec{x}|y=1)P(y=1)}{P(\vec{x})} = \frac{0.8}{0.56} \approx 0.286$$

$$h_2(\vec{x}) = P(y=2|\vec{x}) = \frac{0.5}{0.56} \approx 0.089$$

$$h_3(\vec{x}) = P(y=3|\vec{x}) = \frac{0.75}{0.56} \approx 0.67$$

MAP hipoteza : $h(\vec{x}) = \operatorname{argmax}_j P(\vec{x}|y=j)P(y=j) = \operatorname{argmax} \{ 0.16, 0.025, 0.375 \}$

$$= 3 \quad (\text{klasa "normal"})$$

1.3.

$$K=3$$

0.0534	0.08	0.063
$P(y=1) = 0.3$	$P(y=2) = 0.2$	$P(y=3) = 0.5$

$$X = \mathbb{R}$$

$\mu_1 = -5$	$\sigma_1^2 = 5$
$\mu_2 = 0$	$\sigma_2^2 = 1$
$\mu_3 = 5$	$\sigma_3^2 = 10$

$$P(\vec{x}|y)$$

$$\mathcal{N}(-5, 5)$$

$$0.178$$

$$\mathcal{N}(0, 1)$$

$$0.4$$

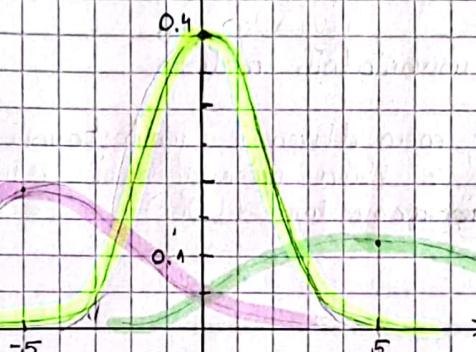
$$\mathcal{N}(5, 10)$$

$$0.196$$

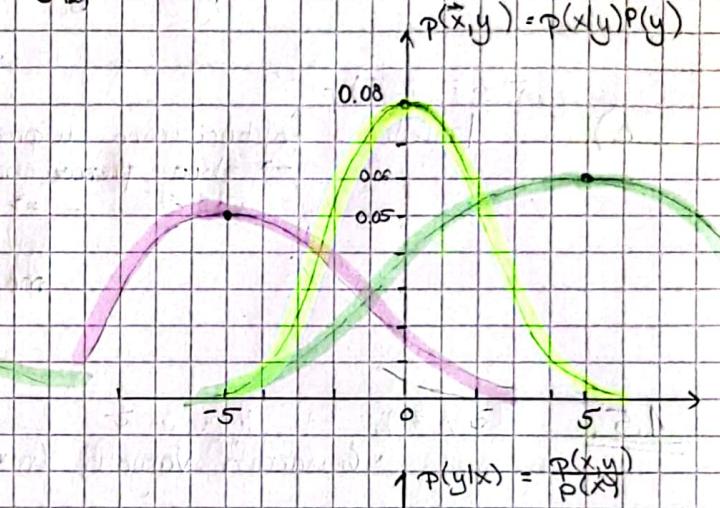
$$P(x|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

\rightarrow za mod. Gausse za x univisti $x = \mu$
 $\rightarrow \frac{1}{\sigma\sqrt{2\pi}}$ mod.

$$P(\vec{x}|y)$$



$$P(\vec{x}) = \sum P(\vec{x}|y)$$



$$P(\vec{x}, y) = P(x|y)P(y)$$

$$P(y|x) = \frac{P(x,y)}{P(x)}$$

$$0.08$$

$$0.06$$

$$0.05$$

$$0.04$$

$$0.03$$

$$0.02$$

$$0.01$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$0.00$$

$$\begin{aligned}
 h_j(\vec{x}) &= c_n p(\vec{x} | y=j) + c_n P(y=j) \\
 &= \dots \\
 &= -\frac{1}{2} \sum_{i=1}^n \left(\frac{x_i - \mu_j}{\sigma_i} \right)^2 + c_n P(y=j)
 \end{aligned}$$

b) #param = $\underbrace{k \cdot n}_{\mu_j} + \underbrace{n}_{\sigma_i^2} + \underbrace{k-1}_{P(y=j)}$

c) Izglednost faktorizirana u produkt univarijatnih razdioba
 \rightarrow zbroj prenaučenosti
 \Rightarrow na ovaj račun dobiven je jednostavniji model
 Isto je manje sklon prenaučenosti i ima
 manje broj parametara ($O(n)$)

15. $K=2$
 • 6 ulaznih varijabli ($n=6$)

3 modela

H_1 = dijagonalna kovarij. matrica

H_2 = dijag. i dijek. kovarij. matrica

H_3 = izotropna kovarij. matrica

$$p(\vec{x} | y=j) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma_j|^{\frac{1}{2}}} \exp \left(-\frac{1}{2} (\vec{x} - \vec{\mu}_j)^T \Sigma_j^{-1} (\vec{x} - \vec{\mu}_j) \right)$$

$$\begin{aligned}
 h_j(\vec{x}) &= c_n [p(\vec{x} | y=j) P(y=j)] \\
 &= c_n p(\vec{x} | y=j) + c_n P(y=j) \\
 &= -\frac{1}{2} c_n 2\pi - \frac{1}{2} c_n |\Sigma_j| - \frac{1}{2} (\vec{x} - \vec{\mu}_j)^T \Sigma_j^{-1} (\vec{x} - \vec{\mu}_j) + c_n P(y=j)
 \end{aligned}$$

a) H_1
 #param: $\underbrace{\frac{n(n+1)}{2}}_{\text{gorj. } \Delta \Sigma_j} + \underbrace{k \cdot n}_{\vec{\mu}_j \text{ (centroidi)}} + \underbrace{k-1}_{\text{pričri}} - O(n^2)$

H_2
 #param: $n + K \cdot n + K - 1$

H_3
 #param: $1 + K \cdot n + K - 1$

$K=2$
 $n=6$

H_1	#param	34
H_2	#param	19
H_3	#param	14

b) Očekujemo da će najbolje generalizirati model H_2

• model H_1

- sklon prenajčenosti - modeliraju se porodice svih značajki - u ovoj primjeru lće

• model H_3

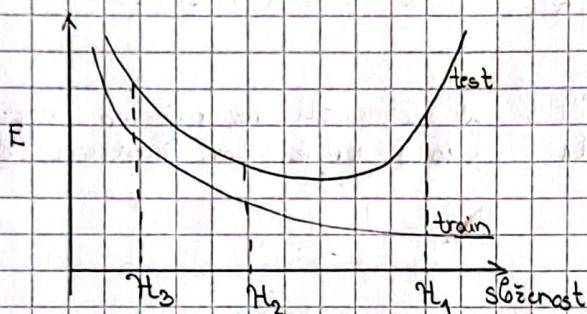
- podražen model

- razlike u skloama značajki

• model H_2

- rasjet. na skale

c)



d)

U praksi odabir modela

bismo napravili unakrsnom provjerom

II Zadaci s ispita

2.1.

$$P(x|y=1) = \mathcal{N}(-8, 3)$$

$$P(x|y=2) = \mathcal{N}(5, 5)$$

$$P(x|y=3) = \mathcal{N}(11, 2)$$

odnosi apriornih vrijednosti

$$P(y|x) = \frac{P(x|y)P(y=j)}{P(x)}$$

→ sjećista aposteriornih distribucija \Leftrightarrow sjećista zajedničkih vjerojatnosti

Plava ($y=1$) i crvena ($y=2$) klasa

- aposteriorno sjećiste ≈ -5
 - sjećiste izglednosti ≈ -3
- \uparrow pomak u lijevo \Rightarrow crvena krivulja širiča više od plave
 "lijevi rep"

$$\Rightarrow P(y=2) > P(y=1)$$

Crvena ($y=2$) i zelena ($y=3$) klasa

- ① opst. sjećiste ≈ 10 \uparrow pomak desno
 sjećiste izgled. ≈ 7 "lijevi rep"

crvena krivulja širiča se više od zelene

- ② opst. sjećiste ≈ 14 \uparrow pomak u lijevo
 sjećiste izgled. ≈ 16 "desni rep"

$$\Rightarrow P(y=2) > P(y=3)$$

⇒ A)

2.2.

$$K = 10$$

$$n = 5$$

- H_1 = nedijagonalna + dijagonalna Σ_j
 H_2 = nedijagonalna + tzotropna Σ_j
 H_3 = pura, dijagonalna Σ_j

$$\varphi(x|y) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2} (\vec{x} - \vec{\mu}_j)^T \Sigma^{-1} (\vec{x} - \vec{\mu}_j) \right\}$$

$$p_{yj}(\vec{x}) = -\frac{n}{2} \ln(2\pi) - \frac{1}{2} \ln |\Sigma_j| - \frac{1}{2} (\vec{x} - \vec{\mu}_j)^T \Sigma^{-1} (\vec{x} - \vec{\mu}_j) + \ln P(y=j)$$

Složnost: $H_1 > H_2$

$\Rightarrow H_3$ po složnosti ne možemo usporoditi s H_1 (zato jer je kov. matrica dijagonalna)

H_1

#param

$$K \cdot n + K \cdot n + K - 1 = 10 \cdot 5 + 10 \cdot 5 + 4 = 104$$

H_2

$$K + K \cdot n + K - 1 = 10 + 50 + 4 = 64$$

H_3

$$\frac{n}{2}(n+1) + K \cdot n + K - 1 = \frac{5 \cdot 6^2}{2} + 10 \cdot 5 + 4 = 69$$

$$\begin{array}{l} H_1 > H_3 > H_2 \\ H_1 > H_2 \end{array} \quad \boxed{A}$$

2.3.

$$n = 3$$

$$K = 2$$

dijagonalna kovarijacijska matrica

$$\begin{aligned} \varphi(y=1) &= \hat{\mu}_1 = 0.2 & \hat{\mu}_1 &= (1, 0, -2) \\ & \hat{\mu}_2 = 0.8 & \hat{\mu}_2 &= (2, -1, 5) \end{aligned}$$

$$\hat{\Sigma}_1 = \begin{bmatrix} 5 & 2 & 4 \\ 2 & 5 & 3 \\ 4 & 3 & 6 \end{bmatrix}$$

$$\hat{\Sigma}_2 = \begin{bmatrix} 6.25 & -0.5 & -1 \\ -0.5 & 1.25 & -0.75 \\ -1 & -0.75 & 0.5 \end{bmatrix}$$

$$\hat{\Sigma} = \sum_j p_{yj} \hat{\Sigma}_j = \begin{bmatrix} 6 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 4 \end{bmatrix}$$

$$|\hat{\Sigma}| = 6 \cdot 2 \cdot 4 = 48$$

$$\hat{\Sigma}^{-1} = \begin{bmatrix} 1/6 & 0 & 0 \\ 0 & 1/2 & 0 \\ 0 & 0 & 1/4 \end{bmatrix}$$

$$\vec{x} = (0, 0, 0)$$

$$h_1(\vec{x}) = -\frac{1}{2} \ln |\hat{\Sigma}| - \frac{1}{2} (\vec{x} - \hat{\mu}_1)^T \hat{\Sigma}^{-1} (\vec{x} - \hat{\mu}_1) + \ln P(y=j) - \frac{n}{2} \ln(2\pi)$$

$$= -\frac{1}{2} \ln |\hat{\Sigma}| - \frac{1}{2} \cdot \frac{7}{6} + \ln 0.2 - \frac{3}{2} \ln(2\pi)$$

$$= -\frac{1}{2} \ln 48 - \frac{7}{12} + \ln 0.2 - \frac{3}{2} \ln(2\pi)$$

$$= -6.885$$

"
A

2.4.

$$K=3$$

$$n=1$$

proj

$$P(x|y=1) = \text{JF}(-10, 2)$$

$$P(x|y=2) = \text{JF}(2, 2)$$

$$P(x|y=3) = \text{JF}(8, 2)$$

$$P(y=1) = \frac{2}{5}$$

$$P(y=2) = \frac{2}{5}$$

$$P(y=3) = \frac{1}{5}$$

lepravci

$$6^2 = 36$$

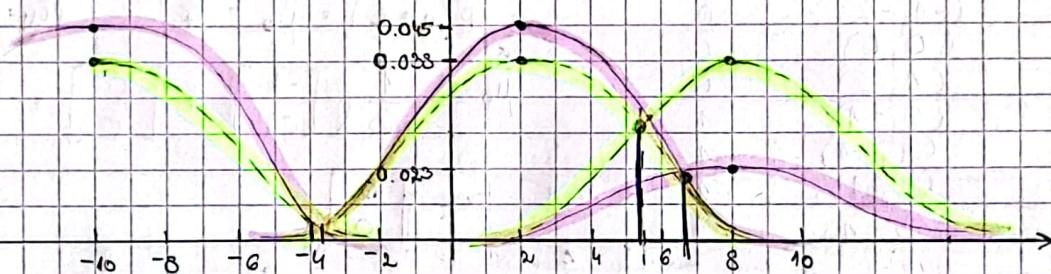
$$P(y=1) = 1/3$$

$$P(y=2) = 1/3$$

$$P(y=3) = 1/3$$

$$P(x|y) = \frac{1}{6\sqrt{2\pi}} \exp\left\{-\frac{1}{26^2}(x-\mu)^2\right\}$$

↑ $P(x|y) = P(x|y)P(y)$



— = prav lepravci

- - - = nalen lepravci

$$\begin{bmatrix} -4 \\ 5 \end{bmatrix}, \begin{bmatrix} -4+a \\ 5+b \end{bmatrix} \leftarrow \text{razlika } h_1$$

$$\begin{bmatrix} -4 \\ 5 \end{bmatrix}, \begin{bmatrix} -4+b \\ 5+a \end{bmatrix} \leftarrow \text{razlika } h_2$$

D

$$h_1(\vec{x}) = P(x|y=1)P(y=1)$$

$$h_2(\vec{x}) = P(x|y=2)P(y=2)$$

2.5.

$$K=2$$

$$n=2$$

$$P(y=1) = P(y=2) = 0.5$$

pozit korel.

Parametri izglednosti

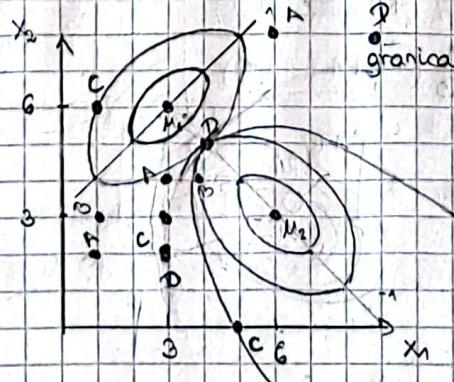
$$\vec{\mu}_1 = (3, 6)$$

$$\Sigma_1 = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$$

D

$$\vec{\mu}_2 = (6, 3)$$

$$\Sigma_2 = \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix} \leftarrow \text{neg. korel}$$



$$\vec{x}^T (\Sigma_1^{-1} - \Sigma_2^{-1}) \vec{x} = \vec{x}^T \begin{bmatrix} 1 & -2 \\ -2 & -1 \end{bmatrix} \vec{x}$$

D (3, 2), (4, 5), (9, 7)

2.6.

$$\begin{matrix} n=2 \\ k=2 \end{matrix}$$

dijeljena i dijagonalna Σ

$$h_0(\vec{x}) = C_n p(\vec{x}, y=0)$$

$$h_0(\vec{x}) = ?$$

$$\vec{x} = (0, 0)$$

$$h_0(\vec{x}) = C_n p(\vec{x}|y=0) + C_n p(y=0)$$

$$= -\frac{n}{2} \ln(2\pi) - \frac{1}{2} \ln|\Sigma| - \frac{1}{2} (\vec{x} - \vec{\mu}_0)^T \Sigma^{-1} (\vec{x} - \vec{\mu}_0) + C_n p(y=0)$$

D

\vec{x}^i	y^i
(-1, -2)	0
(0, 0)	0
(1, 2)	0
(3, -1)	1
(4, -1)	1
(4, 1)	1
(5, 1)	1

$$N=7$$

$$P(y=0) = \hat{\mu}_0 = \frac{3}{7}$$

$$P(y=1) = \hat{\mu}_1 = \frac{4}{7}$$

$$\vec{\mu}_0 = \frac{1}{3} \left(\begin{bmatrix} -1 \\ -2 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 1 \\ 2 \end{bmatrix} \right) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$(\vec{x} - \vec{\mu}_0)(\vec{x} - \vec{\mu}_0)^T$$

$$\hat{\Sigma}_0 = \frac{1}{N_0 - 1} \left(\begin{bmatrix} -1 \\ -2 \end{bmatrix} \begin{bmatrix} -1 & -2 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \end{bmatrix} \begin{bmatrix} 0 & 0 \end{bmatrix} + \begin{bmatrix} 1 \\ 2 \end{bmatrix} \begin{bmatrix} 1 & 2 \end{bmatrix} \right)$$

$$= \frac{1}{2} \cdot \left(\begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix} + \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix} \right) = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$$

Napomena
Svejedno koristiš li

$\hat{\Sigma}_j = \frac{1}{N} \sum (x - \mu)(x - \mu)^T$
li procjenjujes Gv i Var za

sveku tekuću rasobru!
Samo paži na faktor pristranosti

$$\hat{\mu}_1 = \frac{1}{4} \cdot \begin{bmatrix} 3+4+4+5 \\ -1-1+1+1 \end{bmatrix} = \begin{bmatrix} 4 \\ 0 \end{bmatrix}$$

$$\hat{\Sigma}_1 = \begin{bmatrix} \text{Var}(x_1) & \text{Cov}(x_1, x_2) \\ \text{Cov}(x_2, x_1) & \text{Var}(x_2) \end{bmatrix} = \begin{bmatrix} \frac{1}{N_1-1} \sum (x_{1i} - \mu_{1i})^2 & \frac{1}{N_1-1} \sum (x_{1i} - \mu_{1i})(x_{2i} - \mu_{2i}) \\ \frac{1}{N_1-1} \sum (x_{2i} - \mu_{2i})^2 & \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{3} (1+0+0+1) & \frac{1}{3} (+1+0+0+1) \\ \frac{1}{3} (+1+1+1+1) & \end{bmatrix} = \begin{bmatrix} \frac{2}{3} & \frac{2}{3} \\ \frac{2}{3} & \frac{4}{3} \end{bmatrix}$$

$$\hat{\Sigma} = \sum \mu_j \hat{\Sigma}_j = \frac{3}{7} \hat{\Sigma}_0 + \frac{4}{7} \hat{\Sigma}_1 = \begin{bmatrix} \frac{17}{21} & 0 \\ 0 & \frac{52}{21} \end{bmatrix}$$

piše u red. da je dijag.!

$$|\Sigma| = \frac{17}{21} \cdot \frac{52}{21}$$

$$h_0(\vec{x}) = -\frac{n}{2} \ln(2\pi) - \frac{1}{2} \ln|\Sigma| - \frac{1}{2} (\vec{x} - \vec{\mu}_0)^T \Sigma^{-1} (\vec{x} - \vec{\mu}_0) + C_n p(y=0)$$

$$= -\ln(2\pi) - \frac{1}{2} \ln \frac{17 \cdot 52}{21^2} + \ln \frac{6}{7} = -3.03 \Rightarrow \textcircled{C}$$

16. Bayesov klasifikator II

Strojno učenje 1, UNIZG FER, ak. god. 2023./2024.

Jan Šnajder, vježbe, v3.2

1 Zadatci za učenje

1. [Svrha: Razumjeti vezu između Bayesovog klasifikatora i logističke regresije, odnosno probabilitičku interpretaciju logističke regresije. Razumjeti razliku u broju parametara između diskriminativnog i generativnog modela te utjecaj broja klasa i broja primjera na taj odnos.]

- Izvedite model logističke regresije krenuvši od generativne definicije za $P(y = 1|\mathbf{x})$. Izvod napravite korak po korak te se uvjerite da možete obrazložiti svaki korak u izvodu. Napišite sve pretpostavke koje ste ugradili u izvod.
 - Model logističke regresije koristimo za binarnu klasifikaciju primjera s $n = 100$ značajki. Odredite broj parametara modela logističke regresije te njemu odgovarajućeg generativnog modela.
 - Izračunajte broj parametara za isti slučaj, ali sa $K = 5$ klasa.
 - Prepostavite da klasificiramo u $K = 10$ klasa. Izračunajte koliko velika mora biti dimenzija prostora značajki n , a da bi se logistička regresija isplatila jer ima manje parametara od odgovarajućega generativnog modela.
2. [Svrha: Isprobati na konkretnom primjeru procjenu parametara naivnog Bayesovog klasifikatora.] Naivan Bayesov klasifikator želimo upotrijebiti za binarnu klasifikaciju "Skupo ljetovanje na Jadranu". Skup primjera za učenje je sljedeći:

i	Mjesto	Otok	Smještaj	Prijevoz	$y^{(i)}$
1	Istra	da	privatni	auto	da
2	Kvarner	ne	kamp	bus	ne
3	Dalmacija	da	hotel	avion	da
4	Dalmacija	ne	privatni	avion	ne
5	Istra	ne	privatni	auto	da
6	Kvarner	ne	kamp	bus	ne
7	Dalmacija	da	hotel	auto	da

- Izračunajte MLE procjene svih parametara modela te klasificirajte primjere (Istra, ne, kamp, bus) i (Dalmacija, da, hotel, bus).
 - Izračunajte Laplaceove (zaglađene) procjene za sve parametre modela te klasificirajte nanovo iste primjere.
3. [Svrha: Razviti intuiciju o uvjetnoj nezavisnosti i odnosu između nezavisnosti i uvjetne nezavisnosti.]

- Definirajte uvjetnu nezavisnost slučajnih varijabli. Pokažite da je definicija pomoću zajedničke vjerojatnosti istovjetna definiciji pomoću uvjetne vjerojatnosti.
- Za sljedeće primjere razmotrite sve parove varijabli i odredite za koje parove možemo pretpostaviti nezavisnost odnosno uvjetnu nezavisnost:
 - $P \equiv$ danas je ponedjeljak, $S =$ danas je subota, $L \equiv$ danas je listopad.
 - $S \equiv$ sunčano je; $V =$ vruće je; $K =$ ljudi se kupaju.

- iii. $L \equiv$ dokument sadrži riječ "lopta"; $N \equiv$ dokument sadrži riječ "nogomet";
 $S \equiv$ dokument je o sportu.
- iv. $K \equiv$ pada kiša; $C =$ pukla je cijev; $M \equiv$ ulica je mokra.
- (c) Temeljem prethodnih primjera, odgovorite implicira li nezavisnost dviju varijabli njihovu uvjetnu nezavisnost, $A \perp B \Rightarrow A \perp B|C$? Vrijedi li obrnut slučaj, $A \perp B \Rightarrow A \perp B|C$?
4. [Svrha: Razumjeti definiciju uzajamne informacije i način njezina izračuna. Razumjeti razliku između zavisnosti i liinearne zavisnosti.]
- (a) Krenuvši od definicija za entropiju i relativnu entropiju, izvedite mjeru uzajamne informacije $I(X, Y)$ kao Kullback-Leiblerovu divergenciju između zajedničke razdiobe, $P(X, Y)$, i zajedničke razdiobe uz pretpostavku nezavisnosti, $P(X)P(Y)$.
- (b) Neka je zajednička vjerojatnost $P(X, Y)$ varijabli X i Y sljedeća: $P(1, 1) = 0.2$, $P(1, 2) = 0.05$, $P(1, 3) = 0.3$, $P(2, 1) = 0.05$, $P(2, 2) = 0.3$, $P(2, 3) = 0.1$. Izračunajte mjeru uzajamne informacije $I(X, Y)$ za varijable X i Y . Biste li, temeljem vrijednosti uzajamne informacije, rekli da su varijable X i Y nezavisne? Jesu li varijable linearno zavisne?
- (c*) Uzajamna informacija nije odozgo ograničena, ali je ograničena odozdo. Primjenom Jensenove nejednakosti, dokažite da vrijedi $I(X, Y) \geq 0$.
5. [Svrha: Shvatiti kako uvjetna nezavisnost varijabli određuje optimalnu strukturu polunaivnog Bayesovog modela te kako to onda određuje broj parametara.] Želimo naučiti model za klasifikaciju pacijenata s obzirom na rizik oboljenja od kardiovaskularnih bolesti. Ciljne klase su $C_1 = VisokRizik$, $C_2 = UmjerenRizik$, $C_3 = NizakRizik$. Koristimo sedam diskretiziranih ulaznih varijabli: spol, dob, težina, visina, indeks tjelesne mase (BMI), indikacija je li osoba pušač (binarna varijabla) i indikacija bavi li se osoba sportom (binarna varijabla).
- (a) Bi li naivan Bayesov model u ovom slučaju bio dobar odabir? Zašto? Predložite polunaivni model.
- (b) Izračunajte broj parametara predloženog polunaivnog modela i usporedite ga s brojem parametara naivnog modela.
- (c) Razmatramo familiju modela polunaivnog Bayesovog klasifikatora \mathcal{H}_α kod kojeg se združivanje varijabli provodi za sve parove varijabli (x_i, x_j) za koje $I(x_i, x_j) \geq \alpha$. Skicirajte pogreške učenja i ispitivanja modela \mathcal{H}_α kao funkcije praga α (dvije krivulje na istoj skici).

2 Zadatci s ispita

1. (P) Gaussov Bayesov klasifikator i logistička regresija su generativno-diskriminativni par modela, što znači da, uz prikladan odabir parametara, oba modela mogu ostvariti identičnu granicu u ulaznom prostoru. Međutim, Gaussov Bayesov klasifikator je generativni model, dok je logistička regresija diskriminativan model, pa ta dva modela općenito imaju različit broj parametara. U pravilu, logistička regresija imat će manje parametara od njoj odgovarajućeg modela Gaussovog Bayesovog klasifikatora. Razmotrite slučaj binarne klasifikacije u ulaznom prostoru dimenzije $n = 100$ pomoću modela logističke regresije i njoj odgovarajućeg modela Gaussovog Bayesovog klasifikatora. **Koliko će model Gaussovog Bayesovog klasifikatora imati više parametara od modela logističke regresije?**

A 200 B 5049 C 5150 D 10200

2. (N) Treniramo naivan Bayesov klasifikator za binarnu klasifikaciju "Skupo ljetovanje na Jadranu". Skup primjera za učenje je sljedeći:

i	Mjesto	Otok	Smještaj	Prijevoz	$y^{(i)}$
1	Kvarner	da	privatni	auto	1
2	Kvarner	ne	kamp	bus	1
3	Dalmacija	da	hotel	avion	1
4	Dalmacija	ne	privatni	avion	0
5	Istra	da	kamp	auto	0
6	Istra	ne	kamp	bus	0
7	Dalmacija	da	hotel	auto	0

Procjene parametara radimo Laplaceovim MAP-procjenniteljem. Zanima nas klasifikacija sljedećeg primjera:

$$\mathbf{x} = (\text{Istra, ne, kamp, bus})$$

Koliko iznosi aposteriorna vjerojatnost $P(y = 1|\mathbf{x})$?

- A 0.1747 B 0.0032 C 0.6856 D 0.3144

3. (P) Naivan Bayesov klasifikator pretpostavlja uvjetnu nezavisnost značajki unutar klase, to jest $x_j \perp x_k | y$. Međutim, u stvarnosti ta pretpostavka rijetko kada vrijedi. Kao primjer, razmotrite model za klasifikaciju novinskih članaka, čija je zadaća odrediti je li tema članka pandemija koronavirusa ($y = 1$) ili ne ($y = 0$). Model koristi binarne značajke koje indiciraju pojavljivanje određene riječi u novinskom članku. Na primjer, izglednost $P(\text{stožer}|y = 1)$ jest vjerojatnost da se u članku koji je na temu pandemije koronavirusa pojavi riječ "stožer". Razmotrite sljedeće četiri riječi koje se općenito mogu pojaviti u novinskim člancima: "stožer", "pandemija", "koronavirus" i "general". **Za koju od sljedećih jednakosti općenito očekujemo da ne vrijedi i da se time onda narušava pretpostavka naivnog Bayesovog klasifikatora?**

- A $P(\text{stožer}|y = 1) = P(\text{stožer}|pandemija, y = 1)$
 B $P(\text{general}|y = 0) = P(\text{general}|\text{stožer}, y = 0)$
 C $P(\text{koronavirus}|y = 0) = P(\text{koronavirus}|\text{general}, y = 0)$
 D $P(\text{pandemija}|y = 1) = P(\text{stožer}|y = 1)$

4. (N) Treniramo Bayesov klasifikator za odluku o dobroj destinaciji za Erasmus+ studijski boravak. Skup primjera za učenje, izgrađen na temelju iskustava prijatelja i prijatelja prijatelja, je sljedeći:

i	Država	Stipendija	Semestar	Studij	Gовори Језик	$y^{(i)}$
1	Njemačka	da	ljetni	dipl	da	1
2	Poljska	ne	zimski	preddipl	ne	1
3	Italija	da	ljetni	dipl	da	1
4	Njemačka	ne	zimski	preddipl	ne	0
5	Austrija	da	ljetni	dipl	da	1
6	Poljska	ne	zimski	dipl	ne	1
7	Austrija	da	zimski	dipl	ne	1
8	Njemačka	ne	zimski	dipl	ne	0

Očekujemo zavisnost između varijabli *Država* i *Stipendija*, pa koristimo polunaivan Bayesov klasifikator u kojemu su te dvije varijable združene. Procjene izglednosti klasa radimo Laplaceovim MAP-procjenniteljem. Zanima nas klasifikacija za $\mathbf{x} = (\text{Italija, ne, zimski, dipl, ne})$. **Koliko iznosi aposteriorna vjerojatnost $P(y = 1|\mathbf{x})$?**

- A 0.322 B 0.488 C 0.588 D 0.741

5. (P) Treniramo binarni klasifikator za analizu predsjedničke izborne kampanje. Svrha klasifikatora jest predvidjeti hoće li kandidat ili kandidatkinja skupiti dovoljno potpisa za kandidaturu. Model koristi pet značajki: x_1 – politička orientacija (kategorička značajka s tri vrijednosti), x_2, x_3 – dob kandidata i politički staž (dvije numeričke značajke), x_4 – populist (binarna značajka) i x_5 – kandidat/kinja velike političke stranke (binarna značajka). Primijetite da u istom modelu kombiniramo diskretne i kontinuirane značajke, što je sasvim legitimno. Razmatramo tri modela različite složenosti:

\mathcal{H}_0 : Bayesov klasifikator bez ikakvih pretpostavki o uvjetnoj nezavisnosti

\mathcal{H}_1 : Polunaivan Bayesov klasifikator

\mathcal{H}_2 : Naivan Bayesov klasifikator

Polunaivan model \mathcal{H}_1 isti je kao i naivan model \mathcal{H}_2 , s tom razlikom da smo u jedan faktor združili značajke x_1 i x_4 , sluteći ipak da bi pokoji kandidat mogao dobro kapitalizirati populizam u kombinaciji s nekom etabliranom političkom orijentacijom. Kod naivnog Bayesovog klasifikatora naivnu pretpostavku uveli smo za sve varijable (i za diskretne i za kontinuirane). U sva tri modela za

značajke x_2 i x_3 koristimo dijeljenu kovarijacijsku matricu. Izračunajte broj parametara za svaki od ova tri modela. **Koliko parametara sveukupno imaju ova tri modela?**

- A 52 B 61 C 62 D 64

6. (N) Treniramo polunaivan Bayesov klasifikator sa $n = 3$ binarne varijable, x_1 , x_2 i x_3 . Zajednička vjerojatnost tih triju varijabli definirana je sljedećom tablicom:

		$x_3 = 0$		$x_3 = 1$	
		$x_2 = 0$	$x_2 = 1$	$x_2 = 0$	$x_2 = 1$
$x_1 = 0$	0.2	0.1	0.1	0.0	
	0.3	0.0	0.2	0.1	

Prije treniranja klasifikatora, koristimo uzajamnu informaciju kako bismo procijenili koje su varijable najviše statistički zavisne, jer se te varijable isplati združiti u zajednički faktor. Odlučili smo združiti onaj par varijabli koje imaju uzajamnu informaciju veću od 0.01. Ako to vrijedi za dva para varijabli, onda ćemo sve tri varijable združiti u jedan faktor. Izračunajte uzajamne informacije između svih parova varijabli te odredite koje varijable ćemo združiti u zajedničke faktore prema gornjem pravilu. **Kako glasi faktorizacija zajedničke vjerojatnosti tog polunaivnog Bayesovog klasifikatora?**

- A $P(y)P(x_1, x_2|y)P(x_3|y)$
 B $P(y)P(x_1, x_2, x_3|y)$
 C $P(y)P(x_1, x_3|y)P(x_2|y)$
 D $P(y)P(x_1|y)P(x_2|y)P(x_3|y)$

7. (N) Treniramo polunaivan Bayesov klasifikator sa tri binarne značajke, x_1 , x_2 i x_3 . Skup primjera za učenje \mathcal{D} sastoji se od sljedećih deset primjera:

x_1	x_2	x_3	y	x_1	x_2	x_3	y
1	1	0	1	1	1	0	0
0	1	0	1	1	0	1	1
1	1	0	0	1	0	0	0
0	1	1	1	0	1	1	1
0	1	0	1	0	0	1	0

Prije treniranja koristimo uzajamnu informaciju kako bismo procijenili koje su varijable najviše statistički zavisne, jer se te varijable isplati združiti u zajednički faktor. Izračun provodimo tako da za svaki par varijabli x_i i x_j procjenjujemo parametre zajedničke distribucije $P(x_i, x_j)$, a zatim iz zajedničke distribucije računamo marginalne vjerojatnosti i uzajamnu informaciju $I(x_i, x_j)$. Budući da je skup \mathcal{D} malen, za procjenu parametara distribucije $P(x_i, x_j)$ koristimo Laplaceov procjenitelj. **Koliko iznosi na taj način izračunata uzajamna informacija između varijabli x_1 i x_2 ?**

- A 0.0078 B 0.0112 C 0.0334 D 0.0423

V16 - Bayesov klasifikator II

1.1.

a)

$$\begin{aligned}
 P(y=1|x) &= \frac{P(x|y=1)P(y=1)}{P(x)} \\
 &= \frac{P(x|y=1)P(y=1)}{P(x|y=1)P(y=1) + P(x|y=0)P(y=0)} \\
 &= \frac{1}{1 + \frac{P(x|y=0)P(y=0)}{P(x|y=1)P(y=1)}} \\
 &= \frac{1}{1 + \exp\left(\ln \frac{P(x|y=0)P(y=0)}{P(x|y=1)P(y=1)}\right)} \\
 &= \frac{1}{1 + \exp(-\lambda)} = g(\lambda) = h_{\text{log}}(\vec{x}) \quad \begin{array}{l} \text{model logistickie} \\ \text{regresije uži} \\ \text{dani ujet da} \\ L = \vec{w}^T \vec{x} + w_0 \end{array}
 \end{aligned}$$

$$\begin{aligned}
 \lambda &= -\ln \frac{P(x|y=0)P(y=0)}{P(x|y=1)P(y=1)} = \underbrace{\ln [P(x|y=1)P(y=1)]}_{h_n(\vec{x})} - \underbrace{\ln [P(x|y=0)P(y=0)]}_{h_0(\vec{x})} \\
 &= \vec{w}^T \vec{x} + w_0
 \end{aligned}$$

$h_n(\vec{x})$ i $h_0(\vec{x})$ - Bayesov klasifikator

• gornji ujet svodi se na ujet da granica izmedju klasa mora biti LINEARNA \Leftrightarrow dijeljena kovarijacijska matrica

$$\begin{aligned}
 h_j(\vec{x}) &= \ln [P(\vec{x}|y=j)P(y=j)] \\
 &= \ln \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2} (\vec{x} - \vec{\mu}_j)^T \Sigma^{-1} (\vec{x} - \vec{\mu}_j)\right\} + \ln P(y=j) \\
 &= -\frac{n}{2} \ln(2\pi) - \frac{1}{2} \ln |\Sigma| - \frac{1}{2} (\vec{x} - \vec{\mu}_j)^T \Sigma^{-1} (\vec{x} - \vec{\mu}_j) + \ln P(y=j)
 \end{aligned}$$

$$\begin{aligned}
 h_n(\vec{x}) - h_0(\vec{x}) &= -\frac{n}{2} \ln(2\pi) - \frac{1}{2} \ln |\Sigma| - \frac{1}{2} (\vec{x} - \vec{\mu}_1)^T \Sigma^{-1} (\vec{x} - \vec{\mu}_1) + \ln P(y=1) \\
 &\quad + \cancel{\frac{n}{2} \ln(2\pi) + \frac{1}{2} \ln |\Sigma| + \frac{1}{2} (\vec{x} - \vec{\mu}_0)^T \Sigma^{-1} (\vec{x} - \vec{\mu}_0)} - \ln P(y=0)
 \end{aligned}$$

$$\begin{aligned}
 &= -\frac{1}{2} (\vec{x}^T \Sigma^{-1} \vec{x} - \vec{x}^T \Sigma^{-1} \vec{\mu}_1 - \vec{\mu}_1^T \Sigma^{-1} \vec{x} + \vec{\mu}_1^T \Sigma^{-1} \vec{\mu}_1) + \ln P(y=1) \\
 &\quad + \frac{1}{2} (\vec{x}^T \Sigma^{-1} \vec{x} - \underbrace{\vec{x}^T \Sigma^{-1} \vec{\mu}_0 - \vec{\mu}_0^T \Sigma^{-1} \vec{x}}_{n \times n : n \times 1} + \underbrace{\vec{\mu}_0^T \Sigma^{-1} \vec{\mu}_0}_{n \times 1}) + \ln P(y=0)
 \end{aligned}$$

$$\begin{aligned}
 h_1(\vec{x}) - h_0(\vec{x}) &= \\
 &= -\frac{1}{2}(-2\vec{x}^T \Sigma^{-1} \vec{\mu}_1 + \vec{\mu}_1^T \Sigma^{-1} \vec{\mu}_1) + \frac{1}{2}(-2\vec{x}^T \Sigma^{-1} \vec{\mu}_0 + \vec{\mu}_0^T \Sigma^{-1} \vec{\mu}_0) + \ln \frac{P(y=1)}{P(y=0)} \\
 &= \underbrace{\vec{w}^T \Sigma^{-1} (\vec{\mu}_1 - \vec{\mu}_0)}_{\vec{w}} - \underbrace{\frac{1}{2} \vec{\mu}_1^T \Sigma^{-1} \vec{\mu}_1 + \frac{1}{2} \vec{\mu}_0^T \Sigma^{-1} \vec{\mu}_0}_{w_0} + \ln \frac{P(y=1)}{P(y=0)} \\
 &= \vec{w}^T \vec{x} + w_0
 \end{aligned}$$

$$\boxed{z = \vec{w}^T \vec{x} + w_0}$$

→ už pretpostavku dijeljene kovarijacijske matrice Bayesov klasifikator jednak je modelu logističke regresije

b) $n = 100$

$$\# \text{param_LD} = n+1 = 101 \text{ parametar}$$

$$h_{LR}(\vec{x}) = \sigma(z) = \sigma(\vec{w}^T \vec{x} + w_0) = \frac{1}{1 + \exp(-\vec{w}^T \vec{x} + w_0)}$$

• odg. generativni model

⇒ Bayesov klasifikator ($K=2$, dij. Σ)

$$\# \text{param_BK} = \underbrace{\frac{n}{2}(n+1)}_{\Sigma} + \underbrace{n \cdot K}_{\vec{\mu}_j} + \underbrace{K-1}_{P(y=j)} = \frac{100}{2} \cdot 101 + 100 \cdot 2 + 1 = 5251$$

$$h_{OVO}(y=j)(\vec{x}) = -\frac{n}{2} \ln(2\pi) - \frac{1}{2} \ln |\Sigma| - \frac{1}{2} (\vec{x} - \vec{\mu}_j)^T \Sigma^{-1} (\vec{x} - \vec{\mu}_j) + \ln P(y=j)$$

c) $K=5$

$$\# \text{param_LD} = \binom{K}{2} (n+1) = 1010$$

↳ OVO schema

$$\# \text{param_BK} = \frac{n}{2}(n+1) + n \cdot K + K-1 = \frac{100}{2} \cdot 101 + 5 \cdot 100 + 4 = 5554$$

$$d) K = 10$$

$$\frac{K}{2}(n+1) \leq \frac{n^2}{2} + Kn + K - 1$$

$$45n + 45 \leq \frac{n^2}{2} + \frac{11n}{2} + 10n + 9 / \cdot 2$$

$$90n + 90 \leq n^2 + n + 20n + 18$$

$$n^2 - 69n - 72 \geq 0$$

$$n \leq -1.03 \quad ; \quad n \geq 70.03$$

$$n \geq 71$$

\rightarrow dimenzija prostora začajki mora biti ≥ 1 ili više da bi se logistička regresija ispolnila

1.2.

naivni Bayesov klasifikator

$$h_{\text{Naiv}}(\vec{x}) = \operatorname{argmax}_j P(\vec{x} | y=j) P(y=j)$$

$$= \operatorname{argmax}_j \prod_{e=1}^n P(x_e | y=j) \cdot P(y=j)$$

obično parametre procjenjujemo:

$$\bullet \text{MLE za } P(y=j) = \hat{p}_{yj} = \frac{1}{N} \sum_{i=1}^N \{y^i = j\} N = \frac{N_j}{N}$$

$$\bullet \text{MAP za } P(x_e | y=j) = \hat{p}_{x|yj} = \frac{\sum_{i=1}^N \mathbb{1}\{x_e^i = x_e \wedge y^i = j\} + 1}{\sum_{i=1}^N \mathbb{1}\{y^i = j\} + K_e}$$

$$= \frac{N_{ej} + 1}{N_j + K_e} \quad \begin{matrix} \text{broj vrednosti} \\ \text{začajke } e \text{ u klasi} \\ j \end{matrix} \quad \begin{matrix} \rightarrow \\ \text{broj klasa začajke } e \end{matrix}$$

a)

$$\begin{aligned} P(x_1 = \text{Istra} | y=\text{ne}) &= 0 \\ P(x_1 = \text{Dalmacija} | y=\text{ne}) &= 1/3 \\ P(x_1 = \text{Kvarner} | y=\text{ne}) &= 2/3 \end{aligned}$$

$$\begin{aligned} P(x_2 = \text{ne} | y=\text{da}) &= 1 \\ P(x_2 = \text{da} | y=\text{ne}) &= 0 \end{aligned}$$

$$\begin{aligned} P(x_3 = \text{privatni} | y=\text{ne}) &= 1/3 \\ P(x_3 = \text{kamp} | y=\text{ne}) &= 2/3 \\ P(x_3 = \text{otelj} | y=\text{ne}) &= 0 \end{aligned}$$

$$\begin{aligned} P(x_4 = \text{auto} | y=\text{ne}) &= 0 \\ P(x_4 = \text{bus} | y=\text{ne}) &= 2/3 \\ P(x_4 = \text{avion} | y=\text{ne}) &= 1/3 \end{aligned}$$

$$\begin{aligned} P(x_1 = \text{Istra} | y=\text{da}) &= 1/2 \\ P(x_1 = \text{Dalm.} | y=\text{da}) &= 1/2 \\ P(x_1 = \text{Kvarn.} | y=\text{da}) &= 0 \end{aligned}$$

MLE

\downarrow

$$\begin{aligned} P(x_2 = \text{ne} | y=\text{da}) &= 1/4 \\ P(x_2 = \text{da} | y=\text{da}) &= 3/4 \end{aligned}$$

svr.

par.

modela

$$\begin{aligned} P(x_3 = \text{privatni} | y=\text{da}) &= 1/2 \\ P(x_3 = \text{kamp} | y=\text{da}) &= 0 \\ P(x_3 = \text{otelj} | y=\text{da}) &= 1/2 \end{aligned}$$

$$\begin{aligned} P(x_4 = \text{auto} | y=\text{da}) &= 3/4 \\ P(x_4 = \text{bus} | y=\text{da}) &= 0 \\ P(x_4 = \text{avion} | y=\text{da}) &= 1/4 \end{aligned}$$

Predikcija

$$h(\text{Istra}, \text{ne}, \text{kamp}, \text{bus}) = \operatorname{argmax}_y P(y) \prod P(x_i | y)$$

$$\underline{y = \text{ne}}$$

$$h = \frac{3}{7} \cdot 0 \cdot 1 \cdot \frac{2}{3} \cdot \frac{2}{3} = 0$$

$$\underline{y = \text{da}}$$

$$h = \frac{4}{7} \cdot \frac{1}{2} \cdot \frac{1}{4} \cdot 0 \cdot 0 = 0$$

} ne možemo klasificirati.

$$h(\text{Dalm.}, \text{da}, \text{hotel}, \text{bus}) = \operatorname{argmax}_y P(y) \prod P(x_i | y)$$

$$\underline{y = \text{ne}}$$

$$h = \frac{2}{7} \cdot \frac{1}{3} \cdot 0 \cdot 0 \cdot \frac{2}{3} = 0$$

} ne možemo klasificirati

$$\underline{y = \text{da}}$$

$$h = \frac{4}{7} \cdot \frac{1}{2} \cdot \frac{2}{4} \cdot \frac{1}{2} \cdot 0 = 0$$

b) Laplaceove projene

$$k_1 = 3$$

$$N_{\text{ne}} = 3$$

$$k_2 = 2$$

$$N_{\text{da}} = 4$$

$$k_3 = 3$$

$$k_4 = 3$$

$$P(x_1 = \text{Istra} | y = \text{ne}) = \frac{1}{6}$$

$$P(x_1 = \text{Istra} | y = \text{da}) = \frac{3}{7}$$

$$P(x_1 = \text{Dalmacija} | y = \text{ne}) = \frac{2}{6}$$

$$P(x_1 = \text{Dalmacija} | y = \text{da}) = \frac{3}{7}$$

$$P(x_1 = \text{Kvarner} | y = \text{ne}) = \frac{3}{6}$$

$$P(x_1 = \text{Kvarner} | y = \text{da}) = \frac{1}{7}$$

$$P(x_2 = \text{ne} | y = \text{ne}) = \frac{4}{5}$$

$$P(x_2 = \text{ne} | y = \text{da}) = \frac{2}{6}$$

$$P(x_2 = \text{da} | y = \text{ne}) = \frac{1}{5}$$

$$P(x_2 = \text{da} | y = \text{da}) = \frac{4}{6}$$

$$P(x_3 = \text{priv.} | y = \text{ne}) = \frac{2}{6}$$

$$P(x_3 = \text{priv.} | y = \text{da}) = \frac{3}{7}$$

$$P(x_3 = \text{kamp} | y = \text{ne}) = \frac{3}{6}$$

$$P(x_3 = \text{kamp} | y = \text{da}) = \frac{1}{7}$$

$$P(x_3 = \text{hotel} | y = \text{ne}) = \frac{1}{6}$$

$$P(x_3 = \text{hotel} | y = \text{da}) = \frac{3}{7}$$

$$P(x_4 = \text{auto} | y = \text{ne}) = \frac{1}{6}$$

$$P(x_4 = \text{auto} | y = \text{da}) = \frac{4}{7}$$

$$P(x_4 = \text{bus} | y = \text{ne}) = \frac{3}{6}$$

$$P(x_4 = \text{bus} | y = \text{da}) = \frac{1}{7}$$

$$P(x_4 = \text{avion} | y = \text{ne}) = \frac{2}{6}$$

$$P(x_4 = \text{avion} | y = \text{da}) = \frac{2}{7}$$

$$x = (\text{Istra}, \text{ne}, \text{kamp}, \text{bus})$$

$$\begin{aligned} y = \text{ne} \quad h(x) &= \frac{4}{9} \cdot \frac{1}{6} \cdot \frac{2}{5} \cdot \frac{3}{6} \cdot \frac{3}{6} = \frac{2}{135} \\ y = \text{da} \quad h(x) &= \frac{5}{9} \cdot \frac{3}{7} \cdot \frac{2}{5} \cdot \frac{1}{7} \cdot \frac{1}{7} = \frac{5}{3087} \end{aligned} \quad \left. \begin{array}{l} h(x) = \text{ne} \\ h(x) = \text{da} \end{array} \right\}$$

$$x = (\text{Dalmacija}, \text{da}, \text{hotel}, \text{bus})$$

$$\begin{aligned} y = \text{ne} \quad h(x) &= \frac{4}{9} \cdot \frac{2}{5} \cdot \frac{1}{4} \cdot \frac{1}{6} \cdot \frac{3}{6} = \frac{1}{105} \\ y = \text{da} \quad h(x) &= \frac{5}{9} \cdot \frac{3}{7} \cdot \frac{4}{5} \cdot \frac{2}{7} \cdot \frac{1}{7} = \frac{10}{1029} \end{aligned} \quad \left. \begin{array}{l} h(x) = \text{ne} \\ h(x) = \text{da} \end{array} \right\}$$

1.3.

a)

• uvjetna nezavisnost slučajnih varijabli

= 2 varijable X i Y postaju rezavisne ako ravnje poznat ishod treće varijable Z

vrijedi:

$$P(X|Y, Z) = P(X|Z)$$

$$P(Y|X, Z) = P(Y|Z)$$

• preko zajedničke vjerojatnosti

$$P(X, Y|Z) = P(X|Z)P(Y|Z)$$

ISTOVJESENOST

$$\left. \begin{aligned} P(X, Y|Z) &= P(X|Z)P(Y|Z) \\ &= P(X|Z)P(Y|Z) \end{aligned} \right\}$$

ili

$$\left. \begin{aligned} P(X|Y, Z) &= P(X|Z) \\ P(Y|X, Z) &= P(Y|Z) \end{aligned} \right\},$$

$$\left. \begin{aligned} P(X, Y|Z) &= P(Y|Z)P(X|Y, Z) \\ &= P(Y|Z)P(X|Z) \end{aligned} \right\}$$

b)

I

P = danas je ponedjeljak

S = danas je subota

L = danas je četvrtak

Parovi

P ⊥ S ⊥ L

$$P(P|S, L) \neq P(P|L)$$

\Rightarrow nije uvjetno nezavisno

P ⊥ L | S

$$P(P|L, S) = P(P|S) \quad \checkmark$$

$$P(L|P, S) = P(L|S) \quad \checkmark$$

\rightarrow uvjetno nezavisno

S ⊥ L | P

$$P(S|L, P) = P(S|P) \quad \checkmark$$

$$P(L|S, P) = P(L|P) \quad \checkmark$$

III S = sunčano je

V = vruće je

K = ljudi se kupaju

S ⊥ V | K \rightarrow nije uvjetno nezavisno

S ⊥ K | V \rightarrow nije uvjetno nezavisno

III L = "čopta"

N = "rogomet"

S = "sport"

L ⊥ N | S \rightarrow nije uvj. nezavisno

L ⊥ S | N

N ⊥ S | L \rightarrow nije uvj. nezavisno

IV

K = poda kiša

C = pulka cjev

M = ulica je mokra

K ⊥ C | M \rightarrow nije uvjetno nezavisno

K ⊥ M | C

M ⊥ C | K \rightarrow nije uvjetno nezavisno

c) Vrijedi li

$$A \perp B \Rightarrow A \perp B | C ?$$

Općenito ne!

$$\rightarrow \text{Kontrapozit} \quad 1. Bb \quad IV \\ K \perp C \Rightarrow K \not\perp C | M$$

Vrijedi li obrat

$$A \perp B | C \Rightarrow A \perp B ?$$

Općenito ne!

$$\rightarrow \text{Kontrapozit. } 1. Bb \quad III$$

$$E \perp S | N, a \perp L \not\perp S$$

1.4.

a) $P(X, Y) = P(X)P(Y)$

Kullback - Leiblerova divergencija

- mjeri odstupanje jedne distribucije od druge

uzajamna informacija definirana je kao KL diverg. distribucija
 $P(X, Y) = P(X)P(Y)$

$$I(X, Y) = D_{KL}(P(X, Y) || P(X)P(Y)) = \sum_{x,y} P(x, y) \ln \frac{P(x, y)}{P(x)P(y)}$$

→ cilj izvoda

• veći I → više zavisne

IZVOD

Entropija

$$H(P) = - \sum_x P(x) \ln P(x)$$

Unakrsna entropija

$$H(P, Q) = - \sum_x P(x) \ln Q(x)$$

$$H(P, Q) - H(P) = - \sum_x P(x) \ln Q(x) + \sum_x P(x) \ln P(x) \quad //$$

$$= \sum_x P(x) \ln \frac{P(x)}{Q(x)}$$

$$= D_{KL}(P || Q) \quad //$$

b)

$$P(1,1) = 0.2$$

$$P(1,2) = 0.05$$

$$P(1,3) = 0.3$$

$$P(2,1) = 0.05$$

$$P(2,2) = 0.3$$

$$P(2,3) = 0.1$$

$$P(X=1) = \sum_y P(1,y) = 0.55$$

$$P(X=2) = \sum_y P(2,y) = 0.45$$

$$P(Y=1) = \sum_x P(x,1) = 0.25$$

$$P(Y=2) = \sum_x P(x,2) = 0.35$$

$$P(Y=3) = \sum_x P(x,3) = 0.4$$

$$I(x,y) = \sum P(x,y) \ln \frac{P(x,y)}{P(x)P(y)}$$

$$= \dots$$

$$= 0.194563$$

$\Rightarrow I(x,y)$ dosta velika $\Rightarrow X$ i Y nisu nezavisne varijable

Linearna zavisnost

$$\mathbb{E}[X] - \mu_x = 1.45$$

$$\sigma_x^2 = \frac{1}{2} \sum (x - \mu_x)^2 = 0.2525$$

$$\mathbb{E}[Y] = \mu_y = 2.15$$

$$\sigma_y^2 = 0.689167$$

$$\text{Cov}(X,Y) = \mathbb{E}[(X - \mu_x)(Y - \mu_y)]$$

$$= \mathbb{E}[XY] - \mu_x \mu_y = 3.1 - 1.45 \cdot 2.15$$

$$= -0.0175$$

$$\rho_{x,y} = \frac{\text{Cov}(X,Y)}{\sqrt{\sigma_x^2 \sigma_y^2}} = -0.0419 \Rightarrow \text{nisu linearno zavisni!}$$

c) * skip - van devira ishoda ucenja

1.5.

C_1 = visok rizik
 C_2 = umjeren rizik
 C_3 = nizak rizik

$$K = 3$$

$$n = 7$$

- 7 diskretiziranih ulaznih varijabli

- spol	- x_1	}	x ₃ , x ₄ , x ₅
- dob	- x_2		
- težina			
- visina			
- indeks tjelesne mase (BMI)			
- indikacija pušača	- x_6		

- indikacija sportske aktiv. - x_7

a)

Najviše Bayesov model ne bi ovdje bio dobar odobr
- velika zavisnost težine, visine i BMI

Polumarni model

$$h(\vec{x}) = \operatorname{argmax}_y P(y) P(\vec{x}|y)$$

$$= \operatorname{argmax}_y P(y) P(x_1|y) P(x_2|y) P(x_3|x_4, x_5|y) P(x_6|y) P(x_7|y)$$

CPT

b)

#param - naivni

$$= \underbrace{3(2-1)}_{x_1, x_2, x_3} + \underbrace{4 \cdot 2 \cdot K}_{x_4 = x_5} + \underbrace{K-1}_{\text{priori}} = 3 + 24 + 2 = 29$$

$$P(x_2|y) P(y)$$

$$P(x_3|x_4, x_5|y) P(y)$$

$$\hat{\mu}, \hat{\sigma}^2$$

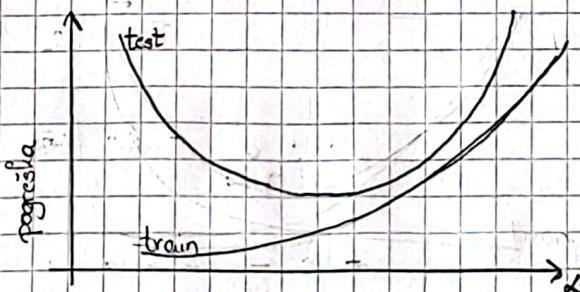
#param - polumarni

$$= \underbrace{K \cdot (2-1)}_{x_1, x_2, x_3} + \underbrace{2 \cdot 2 \cdot K}_{x_4, \{x_5, x_6, x_7\}} + \underbrace{K-1}_{\text{priori}} = 3 + 12 + 2 = 17$$

c)

H₂

$$I(x_i, x_j) \geq \lambda \Rightarrow \text{zduži } x_i, x_j$$



• veliki $\lambda \Rightarrow$
 \Rightarrow zdužene znacajke
 sa velikim I
 (zavisne zdužene)

• mali $\lambda \Rightarrow$
 \Rightarrow svi zduženi

2. Zadaci s ispitom

2.1.

- bitarna klas. : $K=2$
 $n=100$

Log reg

$$\# \text{param} = n+1 = 101$$

Bayes

$$h_j(\vec{x}) = -\frac{n}{2} \ln(2\pi) - \frac{1}{2} \ln(\Sigma) - \frac{1}{2} (\vec{x} - \vec{\mu}_j)^T \Sigma^{-1} (\vec{x} - \vec{\mu}_j) + c_j P(y=j)$$

$$\# \text{param} = \underbrace{\frac{n}{2}(n+1)}_{\text{dij. } 2} + \underbrace{K \cdot n}_{\vec{\mu}_j} + \underbrace{K-1}_{\text{priori}} = \frac{100}{2} \cdot 101 + 2 \cdot 100 + 1 = 5251$$

$$\Delta \text{param} = 5251 - 101 = 5150$$

(C)

2.2.

- neural Bayes.

- progenc parametra: MAP - Laplace

$$\frac{N_j + 1}{N_j + K_e}$$

$$\vec{x} = (\text{Istra nc}, \text{kamp}, \text{bus})$$

$$P(y=1 | \vec{x}) = \frac{P(y=1) P(\vec{x} | y=1)}{P(\vec{x})} = \frac{P(y=1) P(\vec{x} | y=1)}{P(y=1) P(\vec{x} | y=1) + P(y=0) P(\vec{x} | y=0)}$$

$$\begin{aligned} P(y=1) &= \frac{3}{7} \\ P(y=0) &= \frac{4}{7} \end{aligned}$$

$$P(\vec{x} | y=1) = \frac{1}{6} \cdot \frac{1}{5} \cdot \frac{2}{6} \cdot \frac{2}{6} = \frac{1}{135}$$

$$\begin{aligned} P(\text{Istra} | 1) &= \frac{0+1}{3+3} = \frac{1}{6} \\ P(\text{nc} | 1) &= \frac{1+1}{3+3} = \frac{2}{6} \\ P(\text{kamp} | 1) &= \frac{1+1}{3+3} = \frac{2}{6} \\ P(\text{bus} | 1) &= \frac{1+1}{3+3} = \frac{2}{6} \end{aligned}$$

$$P(\vec{x} | y=0) = \frac{3}{7} \cdot \frac{1}{2} \cdot \frac{3}{7} \cdot \frac{2}{7} = \frac{9}{343}$$

$$\begin{aligned} P(\text{Istra} | 0) &= \frac{2+1}{4+3} = \frac{3}{7} \\ P(\text{nc} | 0) &= \frac{2+1}{4+3} = \frac{3}{7} \\ P(\text{kamp} | 0) &= \frac{2+1}{4+3} = \frac{3}{7} \\ P(\text{bus} | 0) &= \frac{1+1}{4+3} = \frac{2}{7} \end{aligned}$$

$$P(y=1 | \vec{x}) = \frac{\frac{3}{7} \cdot \frac{1}{135}}{\frac{3}{7} \cdot \frac{1}{135} + \frac{4}{7} \cdot \frac{9}{343}} = \frac{343}{1963} \approx 0.17473$$

(A)

2.3.

$$V = 2$$

- S = "staza"
- P = "pandemija"
- K = "koronavirus"
- G = "general"

• za koju jednost očekujemo da ne vrijedi

$$A \quad P(S|y=1) = P(S|P, y=1)$$

- kosa 1 - spominje se koronavirus, cvo očekuj. do vrijedi
(pandemija ne donosi dodatnu inf.)

$$(B) \quad P(G|y=0) = P(G|S, y=0)$$

- član vrgnog stazera

$$C \quad P(K|y=0) = P(K|G, y=0)$$

- vrijedi jer general ne donosi dodatnu inf.

$$D \quad P(P|y=1) = P(S|y=1)$$

- ne vrijedi, ali nema veze s Naivnim Bayesom

2.4.

• zavisnost Država i Stipendija
⇒ poluračun Bayes

$$\vec{x} = (\underbrace{\text{Italija}, \text{ne}}_{k_1=4}, \underbrace{\text{zimski}, \text{diplo. ne}}_{k_2=2}, \underbrace{\text{diplo.}, \text{ne}}_{k_3=2}, \underbrace{\text{ne}}_{k_4=2})$$

• MAP - Laplace

$$N = 8$$

$$P(y=1|\vec{x}) = \frac{P(\vec{x}|y=1)P(y=1)}{P(\vec{x}|y=1)P(y=1) + P(\vec{x}|y=0)P(y=0)}$$

$$P(y=1) = \frac{6}{8}$$

$$P(y=0) = \frac{2}{8}$$

$$P(\vec{x}|y=1) = \frac{3}{224}$$

$$P(\vec{x}|y=0) = \frac{9}{320}$$

$$P(\text{Italija, ne} | 1) = \frac{0+1}{6+8} = \frac{1}{14}$$

$$P(\text{Italija, ne} | 0) = \frac{0+1}{2+8} = \frac{1}{10}$$

$$P(\text{zimski} | 1) = \frac{0+1}{6+2} = \frac{1}{8}$$

$$P(\text{zimski} | 0) = \frac{2+1}{2+2} = \frac{3}{4}$$

$$P(\text{diplo.} | 1) = \frac{5+1}{6+2} = \frac{6}{8}$$

$$P(\text{diplo.} | 0) = \frac{1+1}{2+2} = \frac{2}{4}$$

$$P(\text{ne} | 1) = \frac{3+1}{6+2} = \frac{4}{8}$$

$$P(\text{ne} | 0) = \frac{2+1}{2+2} = \frac{3}{4}$$

$$P(y=1|\vec{x}) = \frac{\frac{6}{8} \cdot \frac{3}{224}}{\frac{6}{8} \cdot \frac{3}{224} + \frac{2}{8} \cdot \frac{9}{320}} = \frac{10}{17} = 0.5882 \quad C$$

2.5.

$K=2$

- x_1 - pol. čij. ($K_1 = 3$)
- x_2 = dobrovolník
- x_3 = polit. stav } numerická značky
- x_4 = populist (bunovna značka)
- x_5 = velká stránka (bin. značka)

H_0 = Bayes. bez uvj. nezávis.

H_1 = poluvoln. Bayes. (zdrožene x_1 i x_4)

H_2 = náhod. Bayes.

$$\cdot x_2 \text{ i } x_3 = \text{dijele } \Sigma$$

H_0

$$\# \text{param} = 3 + 4 + 22 + 1 = 30$$

$$\frac{x_2, x_3}{\Sigma : \frac{n(n+1)}{2} = \frac{2 \cdot 3}{2} = 3} \quad \mu : \frac{K \cdot n}{2 \cdot 2} = \frac{2 \cdot 2}{2} = 4$$

$$\frac{x_1, x_4, x_5}{y=0 \rightarrow K_1 \cdot K_4 \cdot K_5 - 1 = 3 \cdot 2 \cdot 2 - 1 = 11} \\ y=1 \rightarrow K_1 \cdot K_4 \cdot K_5 - 1 = 11$$

$$\text{prior} : K-1 = 1$$

H_1

$$\# \text{param} = 2 + 4 + 10 + 2 + 1 = 19$$

$$\frac{x_2, x_3}{\Sigma : \frac{n(n+1)}{2} \rightarrow \text{dijagonálna } (x_2 \text{ i } x_3 \text{ nezávisle}) \quad (\text{dg. navzájemne!})} \quad \mu : K \cdot n = 4$$

$$\frac{x_1, x_4}{K_1 \cdot (K_4 - 1)} \\ = 2 \cdot (3 - 2 - 1) \\ = 10$$

$$\frac{x_5}{K_5 \cdot (K_5 - 1)} = 2$$

$$\text{prior: } K-1 = 1$$

$$\text{Ukupno: } 30 + 19 + 15 \\ = 64$$

(D) //

H_2

$$\frac{x_2, x_3}{\Sigma : 2 \text{ (opt. diag. rbg. nez.)}} \quad \mu : K \cdot n = 4$$

$$\frac{x_1, x_4, x_5}{K_1 \cdot (K_4 - 1 + K_5 - 1)} \\ = 2 \cdot (2 + 1 + 1) = 2 \cdot 4 = 8$$

$$\text{prior: } K-1 = 1$$

$$\# \text{param} = 2 + 4 + 8 + 1 = 15$$

2.6

$$n=3$$

• x_1, x_2, x_3 = Binärne Variablen

$$\lambda = 0.01$$

$I(x_1, y) \geq \lambda \Rightarrow$ nur 1 par variabel

$$I(x_1, x_2) = ?$$

$$I(x_1, x_3) = ?$$

$$I(x_2, x_3) = ?$$

$$I(x_1, x_2) = \sum P(x_1, x_2) \ln \frac{P(x_1, x_2)}{P(x_1)P(x_2)} = 5.13 \cdot 10^{-3} < \lambda$$

$x_1 \backslash x_2$	0	1	
0	0.3	0.1	0.4
1	0.5	0.1	0.6
		0.8	0.2

$$I(x_1, x_3) = \sum P(x_1, x_3) \ln \frac{P(x_1, x_3)}{P(x_1)P(x_3)} = 0.0322 > \lambda$$

$x_1 \backslash x_3$	0	1	
0	0.3	0.1	0.4
1	0.3	0.3	0.6
		0.6	0.4

$\boxed{x_1, x_3}$

$$I(x_2, x_3) = \sum P(x_2, x_3) \ln \frac{P(x_2, x_3)}{P(x_2)P(x_3)} = 5.132 \cdot 10^{-3} < \lambda$$

$x_2 \backslash x_3$	0	1	
0	0.5	0.3	0.8
1	0.1	0.1	0.2
		0.6	0.4

$$P(y|\vec{x}) = P(y)P(x_1, x_2, x_3|y)$$

$$= P(y)P(x_1, x_3|y)P(x_2|y) \Rightarrow \textcircled{C}, //$$

2.7.

 $N = 10$

x_1	x_2	x_3	y
1	1	0	1
0	1	0	1
1	1	0	0
0	1	1	1
0	1	0	1
1	1	0	0
1	0	1	1
1	0	0	0
0	1	1	1
0	0	1	0

$$I(x_1, x_2) = \sum P(x_1, x_2) \ln \frac{P(x_1, x_2)}{P(x_1)P(x_2)}$$

Laplaceov průsledek

$$P(x_1 = 0, x_2 = 0) = \frac{1+1}{10+4} = \frac{2}{14}$$

$$P(x_1 = 0, x_2 = 1) = \frac{4+1}{10+4} = \frac{5}{14}$$

$$P(x_1 = 1, x_2 = 0) = \frac{2+1}{10+4} = \frac{3}{14}$$

$$P(x_1 = 1, x_2 = 1) = \frac{3+1}{10+4} = \frac{4}{14}$$

$x_1 \backslash x_2$	0	1	
0	$\frac{2}{14}$	$\frac{5}{14}$	$\frac{7}{14}$
1	$\frac{2}{14}$	$\frac{4}{14}$	$\frac{7}{14}$
		$\frac{5}{14}$	$\frac{9}{14}$

$$I(x_1, x_2) = 0.11168$$

↳ (b)

17. Probabilistički grafički modeli

Strojno učenje 1, UNIZG FER, ak. god. 2023./2024.

Jan Šnajder, vježbe, v2.3

1 Zadatci za učenje

1. [Svrha: Razumjeti što je to probabilistički grafički model. Shvatiti specifičnosti modela Bayesove mreže te kako taj model predstavlja zajedničku distribuciju. Shvatiti koje induktivne pristranosti ovakva reprezentacija koristi.]

- (a) Navedite tri osnovna aspekta svakog probabilističkog grafičkog modela (PGM).
- (b) Je li PGM parametarski ili neparametarski model? Je li generativni ili diskriminativni? Obrazložite odgovore.
- (c) Prepostavite zajedničku distribuciju četiriju varijabli $p(x, y, w, z)$. Faktorizirajte ovu distribuciju primjenom osnovnih pravila vjerojatnosti te skicirajte Bayesovu mrežu koja odgovara toj faktorizaciji. Topološki uređaj uzmite da je x, y, w, z .
- (d) Ponovite isto, ali ovaj put prepostavljajući $y \perp\!\!\! \perp w | x$ i $x \perp\!\!\! \perp z | y, w$. Kojoj vrsti induktivne pristranosti odgovaraju ove prepostavke o nezavisnosti? Obrazložite motivaciju za uvođenjem dodatnih prepostavki u model.
- (e) Formalno definirajte uređajno Markovljevo svojstvo i topološki uređaj čvorova mreže. Primjenom uređajnog Markovljevog svojstva izvedite uvjetne nezavisnosti kodirane Bayesovom mrežom koja odgovara faktorizaciji

$$P(x, y, w, z) = P(x)P(y|x, z)P(z)P(w|y).$$

- (f) Nacrtajte Bayesovu mrežu Skrivenog Markovljevog modela (HMM) i napišite pripadnu faktorizaciju zajedničke vjerojatnosti $p(\mathbf{x}, \mathbf{z})$. Koje je svrha latentnih varijabli \mathbf{z} i koje su uvjetne nezavisnosti kodirane ovom mrežom?
- 2. [Svrha: Izvježbati iščitavanje Bayesove mreže i uvjetnih nezavisnosti iz zadane faktorizacije zajedničke vjerojatnosti. Razumjeti kako uvjetne nezavisnosti, broj varijabli i njihovih vrijednosti određuju ukupan broj parametara Bayesove mreže.] Gradimo Bayesovu mrežu koja predviđa hoće li student/ica uspješno položiti SU. Mreža sadrži pet varijabli: pohađa li osoba konzultacije (x_1), je li osoba dobra u Pythonu (x_2), rješava li osoba samostalno domaće zadaće i laboratorijske vježbe (x_3), ocjenu iz predmeta UI (x_4) te varijablu koja govori je li osoba položila SU (y). Pritom vrijedi $x_1, x_2, x_3, y \in \{\top, \perp\}$ i $x_4 \in \{2, 3, 4, 5\}$.

- (a) Skicirajte Bayesovu mrežu ako je faktorizacija zajedničke distribucije sljedeća:

$$P(x_1, x_2, x_3, x_4, y) = P(x_1)P(x_2)P(x_3|x_1, x_2)P(x_4|x_2)P(y|x_3).$$

- (b) Koji je ukupan broj parametara ove mreže?
- (c) Koje su uvjetne nezavisnosti kodirane u strukturu ove mreže?

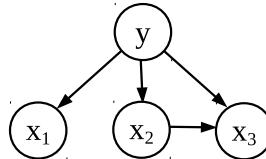
- 3. [Svrha: Razumjeti ideju d-odvajanja i kako se ona može provesti grafički. Shvatiti motivaciju iza ispitivanja uvjetne nezavisnosti parova varijabli.]

- (a) Zašto bismo htjeli znati koji parovi varijabli su uvjetno nezavisni? Nije li ta informacija već kodirana unutar strukture mreže? Objasnite.

- (b) Formalno definirajte d-odvajanje i objasnite koji uvjeti (i kada) moraju vrijediti da bi neke dvije varijable bile uvjetno nezavisne.
- (c) Na temelju Bayesove mreže iz zadatka 2, odredite pod kojim uvjetima su varijable prolaza SU (y) i ocjene iz predmeta UI (x_4) uvjetno nezavisne.
- (d) Svojim riječima objasnite efekt objašnjavanja (engl. *explaining away*) koristeći za primjer varijable x_1 , x_2 i x_3 .
4. [Svrha: Izvježbat iščitavanje uvjetnih nezavisnosti iz Bayesove mreže te određivanje (ne)zavistnosti proizvoljnog para varijabli primjenom pravila d-odvajanja.] Bayesovom mrežom modeliramo vjerojatnost oboljenja od kardiovaskularnih bolesti. Mreža sadrži četiri varijable: spol osobe (S), koliko često osoba tjedno odlazi u teretanu (T), je li osoba pušač (P) te varijablu koja govori o kakvom se riziku radi (R). Pritom vrijedi $s \in \{\text{muški}, \text{ženski}\}$, $p \in \{\perp, \top\}$, $t \in \{1, 3, 5\}$ i $r \in \{\text{nizak}, \text{umjeren}, \text{visok}\}$. Zajednička razdioba faktorizirana je kao:
- $$P(S, T, P, R) = P(S)P(T)P(P|S, T)P(R|P)$$
- (a) Skicirajte Bayesovu mrežu koja predstavlja izvedenu faktorizaciju. Primjenom uređajnoga Markovljevog svojstva izvedite pretpostavke uvjetne nezavisnosti varijabli koje su ugrađene u strukturu Bayesove mreže.
- (b) Koristeći pravila d-odvajanja, odredite pod kojim uvjetima su varijable x_1 i x_2 uvjetno nezavisne. Kako nam ta informacija može biti od koristi?

2 Zadaci s ispita

1. (P) Na slici ispod prikazana je Bayesova mreža koja odgovara polunaivnom Bayesovom klasifikatoru. Pretpostavite da su značajke x_1 , x_2 i x_3 binarne varijable te da je oznaka klase y također binarna varijabla. Označimo ovaj model sa \mathcal{H}_2 . Model \mathcal{H}_2 može se pojednostaviti ako se ukloni brid između varijabli x_2 i x_3 . Označimo takav model sa \mathcal{H}_1 . S druge strane, od modela \mathcal{H}_2 može se napraviti još složeniji model koji odgovara potpuno povezanom acikličkom grafu. Označimo takav model sa \mathcal{H}_3 .



Razmotrite koliko parametara imaju modeli \mathcal{H}_1 , \mathcal{H}_2 i \mathcal{H}_3 . **Koliko model \mathcal{H}_2 ima više parametara od modela \mathcal{H}_1 , a koliko manje parametara od modela \mathcal{H}_3 ?**

- A 2 više, 3 manje B 2 više, 6 manje C 4 više, 4 manje D 4 više, 8 manje
2. (P) Razmotrite Bayesovu mrežu koja zajedničku vjerojatnost faktorizira na sljedeći način:

$$P(w, x, y, z) = P(w)P(y)P(x|w, y)P(z|w)$$

Odredite topološki uređaj varijabli. Ako postoji više mogućih topoloških uređaja, izaberite onaj koji po leksičkom poretku dolazi prvi (npr. x, y, z dolazi prije x, z, y). Zatim primijenite uređajno Markovljevo svojstvo te izvedite sve uvjetne nezavisnosti koje su kodirane u ovoj Bayesovoj mreži. **Koje sve uvjetne nezavisnosti vrijede u ovoj Bayesovoj mreži?**

- A $w \perp y, z \perp \{x, y\} | w$ B $x \perp y | z, z \perp w | y$ C $w \perp y, z \perp x, x \perp w | \{z, y\}$ D $y \perp w, y \perp x | \{w, z\}$
3. (P) Bayesova mreža ima pet varijabli, od kojih su v , w i z binarne, a x i y ternarne varijable. Topološki uredaj varijabli neka je v, w, x, y, z . Uz takav uredaj, u mreži vrijede sljedeće marginalne i uvjetne nezavisnosti:

$$v \perp w \quad w \perp x | v \quad v \perp y | \{w, x\} \quad \{v, w\} \perp z | \{x, y\}$$

Izvedite faktorizaciju zajedničke distribucije koja odgovara ovoj Bayesovoj mreži. **Koliko parametara ima dotična Bayesova mreža?**

- A 10 B 22 C 25 D 27

4. (P) Bayesovom mrežom s pet binarnih varijabli modeliramo prometne prilike u gradu Zagrebu. U našoj mreži, jutarnje doba dana (J) i loše vrijeme (V) utječe na nastanak prometne gužve (G), u smislu da oba događaja povećavaju vjerojatnost nastanka prometne gužve. Loše vrijeme također utječe na nastupanje prometne nesreće (N), u smislu da povećava vjerojatnost prometne nesreće. Nadalje, nastupanje prometne nesreće utječe na nastanak prometne gužve, u smislu da povećava vjerojatnost nastanka prometne gužve. Loše vrijeme također utječe na zastoj tramvaja (T), u smislu da povećava vjerojatnost zastoja tramvaja. Međutim, nestanak struje (S) također uzorkuje zastoj tramvaja. Konačno, zastoj tramvaja uzrokuje masovno pješačenje putnika (P), što opet povećava vjerojatnost prometne nesreće. U ovom kauzalnom modelu može nastupiti efekt objašnjavanja. **Kako bi se efekt objašnjavanja konkretno manifestirao?**

- A $P(V = 1|P = 1, T = 1) < P(V = 1|T = 1)$
- B $P(G = 1|J = 1, V = 1) > P(G = 1|J = 1)$
- C $P(V = 1|P = 1, N = 1) < P(V = 1|N = 1)$
- D $P(T = 1|V = 1, P = 1) > P(T = 1|V = 1)$

5. (P) Bayesova mreža ima pet varijabli, s topološkim uređajem v, w, x, y, z . Uz takav uređaj, u mreži vrijede sljedeće uvjetne nezavisnosti:

$$\{v, w\} \perp y|x \quad \{v, x\} \perp z|\{w, y\}$$

Primjenom algoritma d-odvajanja ispitujemo zavisnosti između parova varijabli. **Koje od sljedećih tvrdnji o nezavisnosti vrijede u ovoj Bayesovoj mreži?**

- A $x \perp z|w$
- B $v \perp y|w$
- C $x \perp z|y$
- D $v \perp y|x$

6. (P) Razmotrite Bayesovu mrežu sa šest čvorova koja odgovara sljedećoj faktorizaciji:

$$P(u, v, w, x, y, z) = P(u|x)P(v|x)P(w|v, x)P(x)P(y|u, x, z)P(z|v, w)$$

Upotrijebite metodu d-odvajanja da biste ispitali uvjetnu nezavisnost varijabli x i z u ovisnosti o preostale četiri varijable. Ispitajte koje varijable od preostalih četiri varijabli trebaju biti opažene a koje neopažena, a da bi varijable x i z bile d-odvojene. **Za koju od preostale četiri varijable je svejedno je li opažena, ako su varijable x i z d-odvojene?**

- A u
- B v
- C w
- D y

V17 - Probabilistički grafički modeli

I Izdaci za učenje

1.1.

- a) 3 osnovna aspekta svakog probabilističkog grafičkog modela su:

 - reprezentacija
 - zaključivanje
 - učenje

- b) PGM-ovi učljučuju u glavnom generativne modele, ali i neke diskriminativne modele.

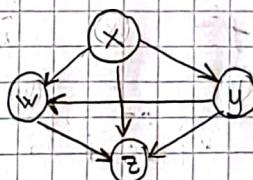
– npr. Bayesove mreže učimo generativno, a Markovljeve mreže diskriminativno.

Također PGM-ovi mogu biti parametarski (npr. Bayesove mreže) i neparametarski (npr. Markov. mreže)

- c) Topološki uređaj: x, y, w, z

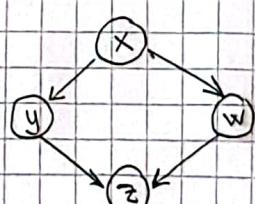
$$P(x, y, w, z) = P(x)P(y|x)P(w|x, y)P(z|x, y, w)$$

pravilo čarca



- d) Pretpostavke: $y \perp w \mid x$
 $x \perp z \perp y, w$ } pristranost ježka

$$P(x, y, w, z) = P(x)P(y|x)P(w|y, x)P(z|w, y, x)$$



Motivacija za uvođenjem dodatnih pretpostavki (uvjetne rezervnosti) zbog smanjivanja složenosti modela

e) TOPOLOŠKI UREĐAJ

= raspored (circularni) uređaj kada se svaki čvorovi roditelji dolaze
prije čvorova djece

UREĐAJNO MARKOVLIJEVO SVOJSTVO

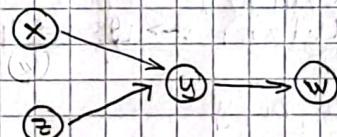
$$x_e \perp \text{pred}(x_e) \setminus \text{pa}(x_e) \mid \text{pa}(x_e)$$

$\rightarrow \text{pa}(x_e)$ = čvorovi roditelja čvora x_e

$\rightarrow \text{pred}(x_e)$ = čvorovi prethodnici čvora x_e

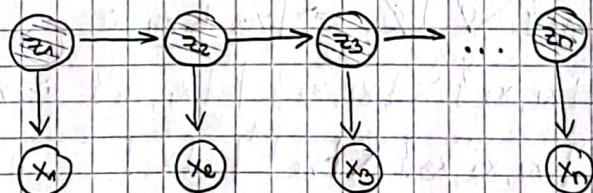
$$\begin{aligned} P(x, y, w, z) &= P(x) P(y|x, z) P(z) P(w|y) \\ &= P(x) P(z|x) P(y|x, z) P(w|y) \leftarrow x \perp z \\ &= P(x) P(z|x) P(y|x, z) P(w|x, y, z) \end{aligned}$$

$$\begin{matrix} w \perp x \mid y \\ w \perp z \mid y \\ x \perp z \end{matrix}$$



f)

Skriveni Markovljev model (HMM)



$$P(\vec{z}, \vec{x}) = P(z_1) P(x_1|z_1) \prod_{e=2}^n P(z_e|z_{e-1}) P(x_e|z_e)$$

- Latentne varijable \vec{z} = predstavljaju stanja kroz koja prolazi proces generiranja podataka

- uljetne nezavisnosti kodirane mrežom ??

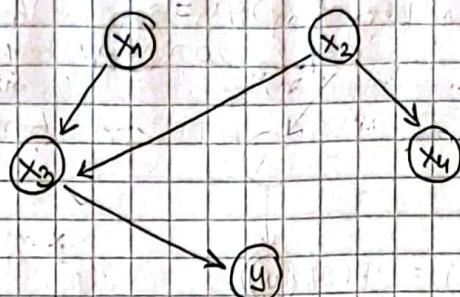
1.2.

$$x_1, x_2, x_3, x_4, y \in \{T, \perp\}$$

$x_4 \in \{2, 3, 4, 5\} \setminus \{y\}$

a)

$$P(x_1, x_2, x_3, x_4, y) = P(x_1)P(x_2)P(x_3 | x_1, x_2)P(x_4 | x_2)P(y | x_3)$$



b) Broj parametara = $1 + 1 + 4 + 6 + 2 = 14$

$$P(x_1) \rightarrow 1$$

$$P(x_2) \rightarrow 1$$

$$P(x_3 | x_1, x_2) \rightarrow 4 \cdot 1 = 4$$

$$P(x_4 | x_2) \rightarrow 3 \cdot 2 = 6$$

$$P(y | x_3) \rightarrow 2 = 1$$

c) Uvjetne nezavisnosti

TU: x_1, x_2, x_3, x_4, y

$$\therefore x_e \perp \text{pred}(x_e) \setminus \text{pa}(x_e) \perp \text{pa}(x_e)$$

$$x_1: x_1 \perp \emptyset | \emptyset$$

$$x_2: x_2 \perp \{x_1\} \setminus \emptyset | \emptyset \Rightarrow \boxed{x_2 \perp x_1}$$

$$x_3: x_3 \perp \{x_1, x_2\} \setminus \{x_1, x_2\} | \{x_1, x_2\} \Rightarrow x_3 \perp \emptyset$$

$$x_4: x_4 \perp \{x_1, x_2, x_3\} \setminus \{x_2\} | \{x_2\}$$

$x_4 \perp x_1$	x_2
$x_4 \perp x_3$	x_2

$$\therefore y \perp \{x_1, x_2, x_3, x_4\} \setminus \{x_3\} | \{x_3\}$$

$y \perp x_1$	x_3
$y \perp x_2$	x_3
$y \perp x_4$	x_3

1.3. a) Zašto bismo htjeli znati koji parovi varijabli su uvjetno nezavisni?

- bitno nam je znati koje su varijable uvjetno zavisne jer pri zaključivanju možemo ukloniti one varijable koje su nezavisne od upita i time smanjiti broj varijabli koje moramo analizirati

- također važno je znati koje su varijable uvjetno nezavisne jer na taj način možemo zaključivati o tome logičke statističke svojstva pretilaze iz određene kauzalne strukture.

Uvjetna nezavisnost nije prikazana Bayesovom mrežom. Prikazana je kauzalna struktura.

b) d-odvojanje

= analiza staza u grafu između čvorova x i y i zaključivanje o uvjetnoj nezavisnosti varijabla x i y na temelju te analize

• ako su svi stazi između čvorova x i y odvojene, onda su čvorovi uvjetno nezavisni

E = skup opaženih varijabli

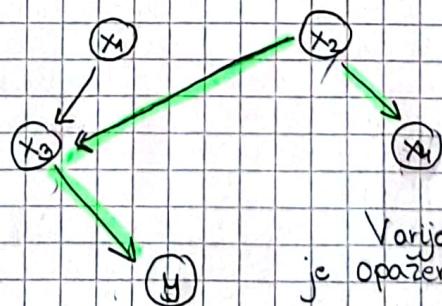
Ta stazu P od čvora x do čvora y kažemo da je d-odvojena
 \Leftrightarrow vrijedi barem jedno od sljedećeg

① P sadrži parac $x \rightarrow z \rightarrow y$ ili $x \leftarrow z \leftarrow y$ i
 $z \in E$

② P sadrži račvanje $x \leftarrow z \rightarrow y$ i
 $z \in E$

③ P sadrži sraz $x \rightarrow z \leftarrow y$ i
 $z \notin E$, ni jedan slijedbenik od $z \notin E$

c) Uvjetna nezavisnost y i $x_4 \Leftrightarrow$ sve staze d-odvojene



Jedina staza od y do x_4

$x_4 \leftarrow x_2 \rightarrow x_3 \rightarrow y$

račvanje

parac

Varijable y i x_4 su uvjetno nezavisne, aško je opažena varijabla x_2 ili x_3

d) Efekt objašnjavanja



$$P(x_1 | x_3) \neq P(x_1 | x_2, x_3) \Leftrightarrow x_1 \not\perp x_3 | x_2$$

• pojavljuje se kod sraza \rightarrow varijable x_1 i x_2 "natječu" se za objašnjavanje varijable x_3 \Rightarrow Ako su opažene varijable x_3 i x_2 smanjena je mjerljivost, i.e. npr. $x_1 \perp x_3 | x_2$

1.4.

$$S \in \{\text{muški, ženski}\}$$

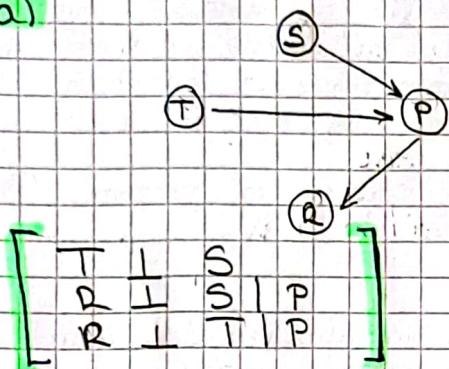
$$T \in \{1, 3, 5\}$$

$$P \in \{T, \perp\}$$

$$Q \in \{\text{nizak, umjeren, visok}\}$$

$$P(S, T, P, Q) = P(S) P(T) P(P|S, T) P(Q|P)$$

a)



Topološki uređaj: $S \rightarrow T \rightarrow P \rightarrow Q$
 $x_e \perp \text{pred}(x_e) \setminus \text{pa}(x_e) \mid \text{pa}(x_e)$

$$S: \quad S \perp \emptyset \mid \emptyset$$

$$T: \quad \begin{matrix} T \perp \{S\} \mid \alpha \mid \alpha \\ \rightarrow T \perp S \end{matrix}$$

$$P: \quad \begin{matrix} P \perp \{S, T\} \setminus \{S, T\} \mid \{S, T\} \\ P \perp \alpha \mid \alpha \end{matrix}$$

$$Q: \quad \begin{matrix} Q \perp \{S, T, P\} \setminus \{P\} \mid \{P\} \\ Q \perp S \mid P \\ Q \perp T \mid P \end{matrix}$$

b) Varijable x_1 i x_2 ujedno su nezavisne za dati E = skup opaženih varijabli \Leftrightarrow su sreć staze P d-odgovore za dati E

Staza P od čvora x_1 do čvora x_2 je d-odgovore \Leftrightarrow vrijedi barem jedno

$$x_2 \rightarrow z \rightarrow x_1 \text{ ili }$$

① P sadrži parac $x_1 \rightarrow z \rightarrow x_2$ i $z \in E$

② P sadrži račvanje $x_1 \leftarrow z \rightarrow x_2$ i $z \in E$

③ P sadrži sraz $x_1 \rightarrow z \leftarrow x_2$ i $z \notin E$ ni jedan slijedbenik od z nije u E

2. Zadatak s ispitom

T.1. Za Bayesovu mrežu poželjno da je generativni i parametarski model.
Rješto?

(D)

generativni = opisuje postupak kojim se mogu generirati podaci koji se pokoravaju određenoj riječi vjerijskoj distribuciji.

parametarski = svaki čvor Bayes. mreže definira uvjetnu vjerojatnost preko teorijske distribucije koja je definirana svojim parametrima.

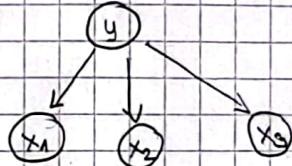
T.2. Koja je veza između uvje. nezavisnosti varijabli u Bayesovoj mreži i opasnosti od prenaučnosti?

(A) učešće pretpostavki o uvje. nezavisnosti pojednostavljuje strukturu Bayesove mreže i smanjuje broj parametara, čime se smanjuje i mogućnost prenaučnosti.

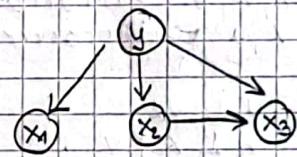
2.1

x_1, x_2, x_3 - binarne
 y - binarna

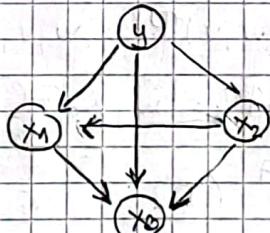
H₁



H₂



H₃



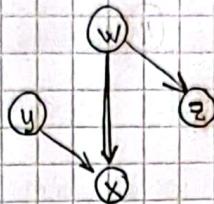
$$\begin{aligned}
 P_1(x_1, x_2, x_3, y) &= P(y) P(x_1|y) P(x_2|y) P(x_3|y) \\
 P_2(x_1, x_2, x_3, y) &= P(y) P(x_1|y) P(x_2|y) P(x_3|y, x_2) \\
 P_3(x_1, x_2, x_3, y) &= P(y) P(x_2|y) P(x_1|y, x_2) P(x_3|y, x_1, x_2)
 \end{aligned}$$

$$\begin{aligned}
 \#P_1 &= 1 + 2 + 2 + 2 = 7 \\
 \#P_2 &= 1 + 2 + 2 + 4 = 9 \\
 \#P_3 &= 1 + 2 + 4 + 8 = 15
 \end{aligned}$$

(B)

H₂ ima 2 više parametra od H₁, a 6 manje od H₃.

$$\underline{2.2} \quad P(w, x, y, z) = P(w)P(y)P(x|w,y)P(z|x)$$



$w, y, x, z \rightarrow$ leksički prvi
 w, y, z, x
 y, w, x, z
 y, w, z, x

TU: w, y, x, z

UMS: $x_e \perp \text{pred}\{x_e\} \setminus \text{pa}\{x_e\} \cup \text{ch}\{x_e\}$

$w \perp \emptyset | \emptyset$

$x \perp \{w, y\} \setminus \{w, y\} \cup \{w, y\}$

$y \perp \{w\} \setminus \emptyset | \emptyset$

$z \perp \{w, y, x\} \setminus \{w\} \cup \{w\}$

$\boxed{y \perp w}$

$\boxed{\begin{array}{c} z \perp y \\ z \perp x \end{array} \boxed{w}}$

(A) $w \perp y, z \perp \{x, y\} \mid w$

2.3.

$v, w, z \rightarrow \text{binarna}$
 $x, y \rightarrow \text{ternarna}$

TU: v, w, x, y, z

$v \perp w$

$w \perp x \mid v$

$v \perp y \mid \{w, x\}$

$z \perp \{v, w\} \mid \{x, y\}$

$$\begin{aligned} P(v, w, x, y, z) &= P(v)P(w|v)P(x|w,v)P(y|w,v,x)P(z|w,v,x,y) \\ &= P(v)P(w)P(x|v)P(y|w,x)P(z|x,y) \end{aligned}$$

Broj parametara = $1+1+4+12+9 = 27$

(D)

$P(v) \rightarrow 1$

$P(w) \rightarrow 1$

$P(x|v) \rightarrow 2 \cdot 2 = 4$

tern.

$P(y|x,w) \rightarrow 2 \cdot (3 \cdot 2) = 12$

$P(z|x,y) \rightarrow 3 \cdot 3 = 9$

Broj parametra čvora x_e
 $P(x_e|x_j)$

$(K_{x_e} - 1) \cdot K_j$

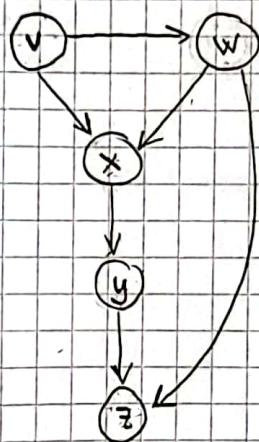
2.5

$$\cdot v, w, x, y, z$$

$$\{v, w\} \perp y | x$$

$$\{v, x\} \perp z | \{w, y\}$$

$$\begin{aligned} P(v, w, x, y, z) &= P(v) P(w|v) P(x|v, w) P(y|x, w, x) P(z|x, w, x, y) \\ &= P(v) P(w|v) P(x|v, w) P(y|x) P(z|y, w) \end{aligned}$$



Koje rezavisnosti vrijedti?

A $x \perp z | w$

Lanac: $x \rightarrow y \rightarrow z$ $y \in E$ (nije opažen)
Račvanje: $x \leftarrow w \rightarrow z$ $w \in E \checkmark$

B $v \perp y | w$

Lanac: $v \rightarrow x \rightarrow y \rightarrow z$ → nije odvojen
Lanac: $v \rightarrow w \rightarrow x \rightarrow y \checkmark$

C $x \perp z | y$

Lanac: $x \rightarrow y \rightarrow z \checkmark$
Račvanje: $x \leftarrow w \rightarrow z$ $x \in E$ (nije odvojeno)

D $v \perp y | x$

Lanac: $v \rightarrow x \rightarrow y \checkmark$

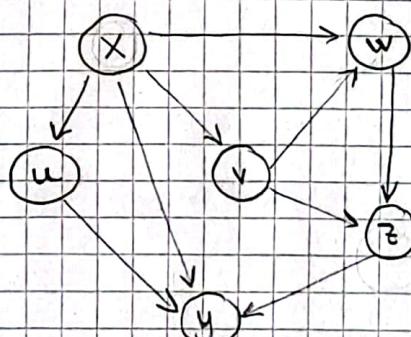
Lanac: $v \rightarrow w \rightarrow x \rightarrow y \checkmark$

ODNOJENO

(1) (v) → (z)

2.6

$$P(u, v, w, x, y, z) = P(u|x) P(v|x) P(w|v, x) P(x) P(y|u, x, z) \cdot P(z|v, w)$$



Promatramo x i z

① Staza

$$x \rightarrow w \rightarrow z$$

• Lanac da bi bila odvojena
 $w \in E$

② Staza (Lanac)

$$x \rightarrow v \rightarrow z$$

$\Rightarrow v \in E$

③ Staza (sraz)

$$x \rightarrow y \leftarrow z$$

$y \notin E$ i svi slijedbenici
od y u topološkom uređaju $\notin E$

Topološki uređaj

x, u, v, w, z, y

Da bi x bilo uvjetno rezavisno z mora vrijediti

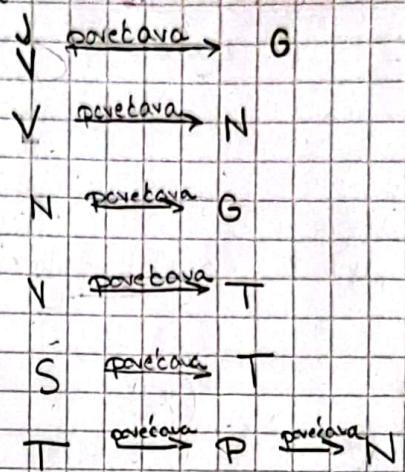
$$w, v \in E$$

$$y \notin E$$

→ ne ovisi o w A

2.4.

- 5 binarnih varijabli

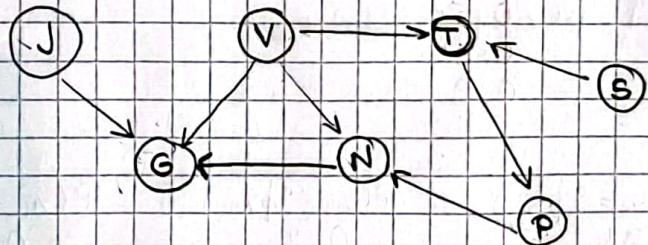


A $P(V=1 | P=1, T=1) < P(V=1 | T=1)$
 \Rightarrow ranac - nije efekt objašnjavanja $P \not\perp\!\!\! \perp V | T$

B $P(G=1 | J=1, V=1) > P(G=1 | J=1)$ \rightarrow nema središnjeg čvora
 $G \not\perp\!\!\! \perp V | J$

C $P(V=1 | P=1, N=1) < P(V=1 | N=1)$ $V \not\perp\!\!\! \perp P | N$

D $P(T=1 | N=1, P=1) > P(T=1 | V=1)$ $T \not\perp\!\!\! \perp P | V$
 $\rightarrow P$ nije srednji čvor



• efekt objašnjavanja \rightarrow u srazu!

$$\begin{aligned}
 P(x|z) &\neq P(x|y, z) \\
 \left. P(x=1 | z=1) \right\} &< \left. P(x=1 | y=0, z=1) \right\} \\
 \left. P(x=1 | z=1) \right\} &> \left. P(x=1 | y=1, z=1) \right\}
 \end{aligned}$$

• ako je z opažen, onda su x i y nezavisni (reprezentnost, točnije staza odvijena)

18. Probabilistički grafički modeli II

Strojno učenje 1, UNIZG FER, ak. god. 2023./2024.

Jan Šnajder, vježbe, v2.3

1 Zadatci za učenje

1. [Svrha: Razumjeti i izvježbati egzaktno zaključivanje kod Bayesovih mreža. Postati svjestan složenosti egzaktnog zaključivanja.] Skicirajte Bayesovu mrežu iz zadatka 2 iz cjeline 17. Parametri modela neka su sljedeći. Za čvorove x_1 i x_2 parametri su $P(x_1 = \top) = 0.2$ i $P(x_2 = \top) = 0.6$. Tablice uvjetnih vjerojatnosti za preostale čvorove su:

x_1	x_2	$P(x_3 = \top x_1, x_2)$	x_3	$P(y = \top x_3)$
\perp	\perp	0.3	\perp	0.2
\perp	\top	0.5	\top	0.9
\top	\perp	0.8		
\top	\top	0.9		

x_2	$P(x_4 = 2 x_2)$	$P(x_4 = 3 x_2)$	$P(x_4 = 4 x_2)$
\perp	0.4	0.2	0.3
\top	0.2	0.1	0.1

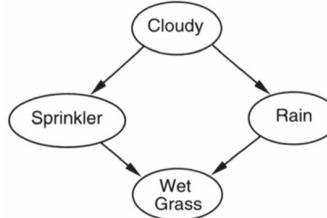
- (a) Postupkom egzaktnog zaključivanja izračunajte $P(y = \top | x_1 = \top, x_4 = 3)$.
- (b) Koja je razlika između posteriornog i MAP-upita? O kakvom tipu upita se radi u prošlom zadatku? Obrazložite.
- (c) Utječe li broj varijabli u mreži na učinkovitost zaključivanja? Zašto?
- (d) Objasnite ideju približnog zaključivanja uzorkovanjem. Koja je prednost tog postupka? U kratkim crtama objasnite kako biste uzorkovali $P(x_1, x_2, x_3, x_4, y)$ koristeći unaprijedno uzorkovanje (engl. *forward sampling*).
2. [Svrha: Razumjeti učenje Bayesovih mreža i njegovu povezanost s procjenom parametara. Znati kako pristupiti učenju modela ako su podaci nepotpuni.]
- (a) Što su parametri Bayesove mreže i na koji način ih učimo iz podataka?
- (b) Izvedite log-izglednost (proizvoljne) Bayesove mreže. Objasnite zašto je moguće procjenjivati parametre svakog čvora mreže zasebno.
- (c) Objasnite što to znači da neki model ima skrivene (latentne) varijable. Kako one utječu na postupak učenja modela?
3. [Svrha: Izvježbati procjenu parametara čvora Bayesove mreže na temelju zadanog skupa podataka. Izvježbati kako napisati izraz za egzaktno zaključivanje na temelju konkretne Bayesove mreže. Razumijeti prednosti i nedostatke egzaktnog zaključivanja naspram metoda uzorkovanja.] Skicirajte Bayesovu mrežu iz zadatka 4 iz cjeline 17. Parametre te mreže procjenjujemo na sljedećem skupu podataka:
- (a) Primjenom (Laplaceovog) MAP-procenitelja procijenite $P(P|S, T)$.

	S	P	T	R
<i>ženski</i>	T	1	<i>visok</i>	
<i>ženski</i>	T	5	<i>umjeren</i>	
<i>muški</i>	⊥	3	<i>nizak</i>	
<i>ženski</i>	⊥	1	<i>umjeren</i>	
<i>muški</i>	T	5	<i>nizak</i>	
<i>ženski</i>	⊥	1	<i>nizak</i>	

- (b) Korištenjem egzaktnog zaključivanja izvedite izraz za vjerojatnost visokog rizika oboljenja osobe koja je pušač i posjećuje teretanu pet puta tjedno. Za svaku od četiri varijable naznačite radi li se o varijabli upita, opaženoj varijabli ili varijabli smetnje.
- (c) Na ovoj mreži ilustrirajte prednosti i nedostatke metoda uzorkovanja nad metodom egzaktnog zaključivanja.
- (d) Na ovoj mreži ilustrirajte nedostatak unaprijednog uzorkovanja. Što su alternative unaprijednom uzorkovanju?

2 Zadatci s ispita

1. (N) Na slici ispod prikazana je Bayesova mreža za problem prskalice za travu, koji smo bili koristili na predavanjima. Varijable su: C (oblačno/*cloudy*), S (prskalica/*sprinkler*), R (kiša/*rain*) i W (mokra trava/*wet grass*). Dane su i tablice uvjetnih vjerojatnosti za svaki čvor.



		S		C		$P(S C)$		R		C		$P(R C)$		W	R	S	$P(W R, S)$
C	$P(C)$	0	0	0	0.5	0	0	0	0	0	1	0.8	0	0	0	1.0	
0	0.5	0	1	0	0.9	0	1	0	1	0	0.2	0	1	0	0.9		
1	0.5	1	0	0	0.5	1	0	1	0	0	0.2	0	1	1	0.1		
		1	1	0	0.1	1	1	1	1	0	0.8	1	0	0	0.01		
												1	0	0	0.0		
												1	0	1	0.1		
												1	1	0	0.9		
												1	1	1	0.99		

Izračunajte aposteriornu vjerojatnost da pada kiša ako je trava mokra i nije oblačno.

- A 0.112 B 0.491 C 0.709 D 0.825

2. (N) Bayesovom mrežom s četiri varijable modeliramo konstrukte pozitivne psihologije. Koristimo binarne varijable *Ljubav* (L), *Sreća* (S), *Tjeskoba* (T), s vrijednostima 0 (nema) i 1 (ima), te ternarnu varijablu *Novac* (N), s vrijednostima 0 (nema), 1 (ima malo) i 2 (ima puno). Strukturu Bayesove mreže definirali smo tako da ona modelira sljedeće pretpostavljene kauzalne odnose: L uzrokuje S , a N uzrokuje S i T . Tako definiranu Bayesovu mrežu zatim treniramo na sljedećem skupu od $N = 7$ primjera:

L	N	S	T
1	0	1	0
1	0	1	0
0	2	0	1
1	2	1	1
1	1	1	0
0	0	0	0
0	2	1	0

Parametre modela procjenjujemo MAP-procjeniteljem sa $\alpha = \beta = 2$ (za binarne varijable) odnosno $\alpha_k = 2$ (za ternarnu varijablu), što je istovjetno Laplaceovom zaglađivanju MLE procjene. Na kraju nas, naravno, zanima koja je vjerojatnost života uz ljubav, sreću i malo novaca. Napravite potrebne MAP-procjene parametara. **Koliko iznosi zajednička vjerojatnost $P(L = 1, S = 1, N = 1)$?**

- A 0.023 B 0.074 C 0.143 D 0.833

3. (P) Razmotrite jednostavnu Bayesovu mrežu koja odgovara faktorizaciji $P(x, y, z) = P(x)P(y)P(z|x, y)$. Sve varijable su binarne. Vrijedi $P(x = 1) = 0.2$ i $P(y = 1) = 0.3$. Tablica uvjetne vjerojatnosti za čvor z je sljedeća:

z	x	y	$p(z x, y)$	z	x	y	$p(z x, y)$
0	0	0	0.1	1	0	0	0.9
0	0	1	0.2	1	0	1	0.8
0	1	0	0.5	1	1	0	0.5
0	1	1	0.9	1	1	1	0.1

Postupkom uzorkovanja s odbijanjem uzorkujemo iz aposteriorne distribucije $P(y|x = 1, z = 0)$. Uzorkovanje smo ponovili ukupno $N = 1000$ puta. **Koja je očekivana veličina uzorka, odnosno koliko slučajnih vektora nećemo morati odbaciti?**

- A 54 B 124 C 200 D 739

4. (N) Bayesovu mrežu koristimo za medicinsku dijagnostiku te modeliramo sljedeće kauzalne odnose. Upala grla ($U = 1$) može biti uzrokovana virusom ($V = 1$) ili bakterijom ($B = 1$). Povišena temperatura ($T = 1$) može biti uzrokovana upalom grla ili sunčanicom ($S = 1$). Sve varijable su binarne. Na temelju podataka o pacijentima procijenili smo parametre mreže: $P(V = 1) = 0.3$, $P(B = 1) = 0.1$ i $P(S = 1) = 0.05$. Uvjetne vjerojatnosti za čvorove U i T su:

V	B	$P(U = 1 V, B)$	U	S	$P(T = 1 U, S)$
0	0	0.2	0	0	0
0	1	0.5	0	1	0.2
1	0	0.4	1	0	0.4
1	1	0.7	1	1	0.4

Zanima nas koje je najvjerojatnije objašnjenje izravnog uzroka povišene temperature u pacijenata kod kojih nije dokazano prisustvo virusa. U tu svrhu računamo MAP-upit za par varijabli upita U i S uz opažene varijable $V = 0$ i $T = 1$, tj. računamo $\text{argmax}_{U, S} P(U, S|V = 0, T = 1)$. Neka je p_1 vjerojatnost najvjerojatnijeg (MAP) objašnjenja za varijable U i S , a p_2 vjerojatnost drugog po redu najvjerojatnijeg obašnjenja za te varijable. **Koliko je puta najvjerojatnije objašnjenje vjerojatnije od drugog najvjerojatnijeg objašnjenja, tj. koliko iznosi p_1/p_2 ?**

- A 11.35 B 13.35 C 15.52 D 17.88

5. (N) Razmotrite Bayesovu mrežu koja odgovara faktorizaciji $P(w, x, y, z) = P(w)P(x)P(y|w, x)P(z|x)$. Sve varijable su binarne. Vrijedi $P(w = 1) = 0.1$, $P(x = 1) = 0.2$, $P(z = 1|x = 0) = 0.9$ i $P(z = 1|x = 1) = 0.7$. Tablica uvjetnih vjerojatnosti za čvor y je sljedeća:

w	x	$p(y = 1 w, x)$
0	0	0
0	1	0.4
1	0	0.2
1	1	0.7

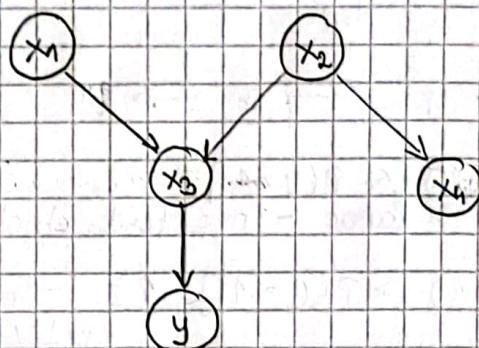
Postupkom uzorkovanja s odbijanjem želimo procijeniti parametar μ uvjetne distribucije $P(x = 0|y = 1, z = 0)$. Uzorkovanje smo ponovili ukupno $N = 100$ puta, od čega smo neke vektore morali odbaciti, pa je naš uzorak manji od N . Na temelju dobivenog uzorka parametar μ procjenjujemo MAP procjeniteljem uz $\alpha = \beta = 2$. **Koliko iznosi očekivana MAP procjena parametra μ ?**

- A 0.0490 B 0.0786 C 0.1274 D 0.1877

V18 - Probabilistički grafički modeli II

1. Zadatak za učenje

$$P(x_1, x_2, x_3, x_4, y) = P(x_1)P(x_2)P(x_3|x_1, x_2)P(x_4|x_2)P(y|x_3)$$



$$x_1, x_2, x_3, y \in \{1, 0\}$$

$$x_4 \in \{2, 3, 4, 5\}$$

$$\begin{aligned} P(x_1 = 1) &= 0.2 \\ P(x_1 = 0) &= 0.8 \end{aligned}$$

$$\begin{aligned} P(x_2 = 1) &= 0.6 \\ P(x_2 = 0) &= 0.4 \end{aligned}$$

Za x_3

x_1	x_2	$P(x_3 = 1 x_1, x_2)$
0	0	0.3
0	1	0.5
1	0	0.8
1	1	0.9

Za y

x_3	$P(y = 1 x_3)$
0	0.2
1	0.9

Za x_4

x_2	$P(x_4 = 2 x_2)$	$P(x_4 = 3 x_2)$	$P(x_4 = 4 x_2)$
0	0.4	0.2	0.3
1	0.2	0.1	0.1

a) Egzistencija zaključivanja:

$$P(\vec{x}_g | \vec{x}_o) = \frac{\sum_{\vec{x}_n} P(\vec{x}_g, \vec{x}_o, \vec{x}_n)}{\sum_{\vec{x}_n} \sum_{\vec{x}_2} P(\vec{x}_g, \vec{x}_o, \vec{x}_n)}$$

$$P(y = 1 | x_1 = 1, x_4 = 3) = \frac{P(y = 1 | x_1 = 1, x_4 = 3)}{P(x_1 = 1, x_4 = 3)}$$

$$\begin{aligned} \vec{x}_g &= y \\ \vec{x}_o &= \{x_1, x_4\} \\ \vec{x}_n &= \{x_2, x_3\} \end{aligned}$$

$$\begin{aligned} P(x_1 = 1, x_4 = 3) &= \sum_{\vec{x}_n} \sum_{\vec{x}_2} P(x_1, x_2, x_3, x_4, y) \\ &= \sum_{x_2} \sum_{x_3} \sum_y P(x_1 = 1) P(x_2) P(x_3 | x_1 = 1, x_2) P(x_4 = 3 | x_3) \end{aligned}$$

↳ 8 pribrojnika

x_2	x_3	y	$P(x_1=1)P(x_2)P(x_3 x_1=1, x_2)P(x_4=3 x_2)P(y x_2)$
0	0	0	$0.2 \cdot 0.4 \cdot 0.2 \cdot 0.2 \cdot 0.8 = 2.56 \cdot 10^{-3}$
0	0	1	$0.2 \cdot 0.4 \cdot 0.2 \cdot 0.2 \cdot 0.2 = 6.4 \cdot 10^{-4}$
0	1	0	$0.2 \cdot 0.4 \cdot 0.8 \cdot 0.2 \cdot 0.1 = 1.28 \cdot 10^{-3}$
0	1	1	$0.2 \cdot 0.4 \cdot 0.8 \cdot 0.2 \cdot 0.9 = 0.01152$
1	0	0	$0.2 \cdot 0.6 \cdot 0.1 \cdot 0.1 \cdot 0.8 = 9.6 \cdot 10^{-4}$
1	0	1	$0.2 \cdot 0.6 \cdot 0.1 \cdot 0.1 \cdot 0.2 = 2.4 \cdot 10^{-4}$
1	1	0	$0.2 \cdot 0.6 \cdot 0.9 \cdot 0.1 \cdot 0.1 = 1.08 \cdot 10^{-3}$
1	1	1	$0.2 \cdot 0.6 \cdot 0.9 \cdot 0.1 \cdot 0.9 = 9.72 \cdot 10^{-3}$

↑
zbroji

$$P(x_1=1, x_4=3) = 7/250 = 0.028$$

$$P(y=1, x_1=1, x_4=3) = \sum_{x_2} \sum_{x_3} P(x_1=1)P(x_2)P(x_3|x_1=1, x_2)P(x_4=3|x_2)P(y=1|x_2)$$

x_2 x_3 pribrojnik

0	0	$0.2 \cdot 0.4 \cdot 0.2 \cdot 0.2 \cdot 0.2 = 6.4 \cdot 10^{-4}$
0	1	$0.2 \cdot 0.4 \cdot 0.8 \cdot 0.2 \cdot 0.9 = 0.01152$
1	0	$0.2 \cdot 0.6 \cdot 0.1 \cdot 0.1 \cdot 0.2 = 2.4 \cdot 10^{-4}$
1	1	$0.2 \cdot 0.6 \cdot 0.9 \cdot 0.1 \cdot 0.9 = 9.72 \cdot 10^{-3}$

$$P(y=1, x_1=1, x_4=3) = 0.02212$$

$$\left[P(y=1 | x_1=1, x_4=3) = \frac{0.02212}{0.028} = 0.79 \right]$$

b)

Aposteriorni upiti odgovaraju na pitanje: Kako je aposteriorna distribucija skrivnih varijabli. MAP-upiti odgovaraju na pitanje: Koje su najvjerojatnije vrijednosti za sve varijable u upitu, uz dane opažene varijable.

U zadatku pod a) radi se o aposteriornom upitu jer se traži vrijednost distribucije za određenu vrijednost varijable y : $P(y=1 | x_1=1, x_4=3)$.

c)

Utječi Ci broj varijabli u mreži na učinkovitost zaključivanja?

Da. Postoji problem kombinatorne eksploracije u postupku. Naime, čim je veći broj varijabli u mreži u izrazima za zajedničku gustoću (raste s povećanjem broja \vec{x}_n) i vjerovatnoću (raste s povećanjem broja \vec{x}_n i \vec{x}_g) treba marginalizirati po većem broju varijabli.

d)

Prednost približnog zaključivanja uzrokovanim → ne dovodi do kombinatorne eksploracije, a nadi uzrokovanim skupu se mogu relativno dobro procijeniti parametri tražene distribucije.

Uzrokovavanje unaprijednim uzrokovovanjem

$$P(x_1, x_2, x_3, x_4, y) = P(x_1)P(x_2)P(x_3|x_1, x_2)P(x_4|x_2)P(y|x_3)$$

- ① Po $P(x_1)$ generirati x_1
- ② Uz pomoć $P(x_2)$ generirati x_2
- ③ Uz pomoć $P(x_3|x_1, x_2)$ i ovisno o generiranim vrijednostima x_1 i x_2 generirati x_3
- ④ Uz pomoć $P(x_4|x_2)$ i ovisno o generiranom x_2 generirati x_4
- ⑤ Uz pomoć $P(y|x_3)$ i ovisno o generiranom x_3 generirati y
- ⑥ Postupak 1-5 ponavljamo dok ne dobimo uzorak željene veličine

1.2.

a) Parametri Bayesove mreže jesu parametri $\vec{\theta}$ koji određuju distribucije loje. opisuju podatke

$$\text{Bayes mreža } p(\vec{x}) = \prod_{e=1}^n p(x_e | pa(x_e))$$

svaka od svih n distribucija ima pripadne parametre $\vec{\theta}_e$

$$\vec{\theta} = [\vec{\theta}_1, \vec{\theta}_2, \dots, \vec{\theta}_e]$$

Parametre učimo iz podataka procjenama (MLE, MAP ili bayesovska procjena)

b) Cog-izglednost Bayesove mreže

$$\begin{aligned} \ln L(\vec{\theta} | D) &= \ln p(D|\vec{\theta}) \stackrel{\text{def.}}{=} \ln \prod_{i=1}^N p(\vec{x}^i | \vec{\theta}) = [\text{def. Bay. mreže}] \\ &= \ln \prod_{i=1}^N \prod_{e=1}^n p(x_e^i | pa(x_e^i; \vec{\theta})) \\ &= \ln \prod_{e=1}^n \prod_{i=1}^N p(x_e^i | pa(x_e^i; \vec{\theta})) \\ &= \sum_{e=1}^n \sum_{i=1}^N \ln p(x_e^i | pa(x_e^i; \vec{\theta})) \end{aligned}$$

Budući da cog-izglednost mreže dekomponira po čvorovima, parametri svakog čvora mreže mogu se procjenjivati zasobno.

c) Ako neli model ima skrivene (latentne) varijable, opaženi skup podataka nije potpun, odnosno on je nepotpun. Nod skupom nepotpunih podataka ne možemo procjenjivati parametre svakog čvora zasebno jer leg-izglednost mreže ne dekomponira po čvorovima.

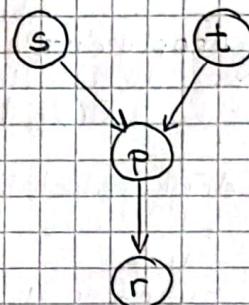
Posljedica toga je da ni za MLE, ni za MAP ni za bayesoveći procjenitelj ne postoji rješenje u zatvorenoj formi → EM-algoritam

1.3.

$$\begin{aligned} s &\in \{M, \bar{z}\} \\ p &\in \{0, 1\} \\ t &\in \{1, 3, 5\} \\ r &\in \{n, u, v\} \end{aligned}$$

s	p	t	r
\bar{z}	1	1	v
\bar{z}	1	5	u
M	0	3	n
\bar{z}	0	1	u
M	1	5	n
\bar{z}	0	1	n

$$\begin{aligned} P(s, t, p, r) \\ = P(s)P(t)P(p|s, t)P(r|p) \end{aligned}$$



a) MAP (Laplace) procjenitelj za $P(p|s, t)$ $L=2$

$$\begin{aligned} P(p=0 | s=M, t=1) &= 1/2 \\ P(p=0 | s=M, t=3) &= 2/3 \\ P(p=0 | s=M, t=5) &= 1/3 \end{aligned}$$

$$\begin{aligned} P(p=0 | s=\bar{z}, t=1) &= 3/5 \\ P(p=0 | s=\bar{z}, t=3) &= 1/2 \\ P(p=0 | s=\bar{z}, t=5) &= 1/3 \end{aligned}$$

$$\begin{aligned} P(p=1 | s=M, t=1) &= 1/2 \\ P(p=1 | s=M, t=3) &= 1/3 \\ P(p=1 | s=M, t=5) &= 2/3 \end{aligned}$$

$$\begin{aligned} P(p=1 | s=\bar{z}, t=1) &= 2/5 \\ P(p=1 | s=\bar{z}, t=3) &= 1/2 \\ P(p=1 | s=\bar{z}, t=5) &= 2/3 \end{aligned}$$

$$\text{MAP: } \frac{\text{Njet+1}}{\text{Njet+2}} \rightarrow \text{Ke}$$

$$b) P(r=v | p=1, t=5) = \frac{P(r=v, p=1, t=5)}{P(p=1, t=5)}$$

$$= \frac{\sum_s P(r=v, s, p=1, t=5)}{\sum_s \sum_r P(r, s, p=1, t=5)}$$

$$= \frac{\sum_s P(s)P(t=5)P(p=1 | s, t=5)P(r=v | p=1)}{\sum_s \sum_r P(s)P(t=5)P(p=1 | s, t=5)P(r | p=1)}$$

Varijable upita: $\vec{x}_q = [r]$

Varijable smatnje: $\vec{x}_n = [s]$

Opožene varijable: $\vec{x}_o = [p, t]$

c) Prednosti užročovanja nad egzakturnim zaključivanjem
 → rješava se problem kombinatorne eksplozije

U ovaj mreži treba procjeniti $1+1+2 \cdot 4 + 2 \cdot 3 =$

$$|S| = 2$$

$$|D| = 2$$

$$|T| = 3$$

$$|R| = 3$$

$$= 8 + 2 + 6 =$$

= 16 distribucija za crtežno zaključivanje

Nedostatci: zaključivanje je približno, a u nekim tehnikama za mreže vjerljivosti \vec{x}_0 (npr. ako je $\vec{x}_0 = [s, t]$ i vrijednosti koje su opažene $s = 2$ i $t = 3$) treba drugo vremena da se generira dovoljno velike uzorake za procjene

d) Nedostatak unaprijednog užročovanja je taj što generira vrijednosti iz zajedničke distribucije $P(\vec{x})$ (u ovom slučaju $P(s, p, t, r)$), a ne iz distribucije $P(\vec{x}_g | \vec{x}_0)$ (u ovom slučaju $P(p | s, t)$)

• vrijedi i drugi nedostatci pod c)

Alternative su

- užročovanje s odbacivanjem
- užročovanje po važnosti
- Gibbsovo užročovanje

2 Zadaci s ispitom

2.1. $P(r=1 | w=1, c=0) = \frac{P(r=1, w=1, c=0)}{P(w=1, c=0)} = \frac{0.0945}{0.1145} \approx 0.825$ D

$$\begin{aligned} P(r=1, w=1, c=0) &= \sum_s P(c=0) P(s|c=0) P(r=1|c=0) P(w=1|s, r=1) \\ &= \sum_s 0.5 \cdot P(s|c=0) \cdot 0.2 \cdot P(w=1|s, r=1) \\ &= 0.1 \cdot (P(s=0|c=0) P(w=1|s=0, r=1) \\ &\quad + P(s=1|c=0) P(w=1|s=1, r=1)) \end{aligned}$$

ova suma se može izvršiti za $r=1$
 ili za $r=0$ zasebno
 = $0.1 \cdot (0.5 \cdot 0.9 + 0.5 \cdot 0.89)$
 = 0.0945

$$P(w=1, c=0) = \sum_s \sum_r P(c=0) P(s|c=0) P(r|c=0) P(w=1|s, r) = 0.1145$$

$$s \quad r \quad P(c=0) \cdot P(s|c=0) \cdot P(r|c=0) \cdot P(w=1|s, r)$$

$$0 \quad 0 \quad 0.5 \cdot 0.5 \cdot 0.8 \cdot 0 = 0$$

$$0 \quad 1 \quad 0.5 \cdot 0.5 \cdot 0.2 \cdot 0.9 = 0.045$$

$$1 \quad 0 \quad 0.5 \cdot 0.5 \cdot 0.8 \cdot 0.1 = 0.02$$

$$1 \quad 1 \quad 0.5 \cdot 0.5 \cdot 0.2 \cdot 0.99 = 0.0495$$

2.2.

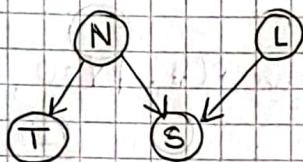
$$L, S, T \in \{0, 1\}$$

$$N \in \{0, 1, 2\}$$

$$L \rightarrow S$$

$$N \rightarrow S, T$$

- MAP progenity
($\lambda = \rho = 2$)



$$N=7$$

L	N	S	T
1	0	1	0
1	0	1	0
0	2	0	1
1	2	1	1
1	1	1	0
0	0	0	0
0	2	1	0

$$P(L=1, S=1, N=1) = \sum_T P(L=1) P(N=1) P(S=1 | L=1, N=1) P(T | N=1)$$

$$P(L=1) = \frac{4+1}{7+2} = \frac{5}{9} \quad P(N=1) = \frac{1+1}{7+3} = \frac{2}{10} = \frac{1}{5}$$

$$P(S=1 | L=1, N=1) = \frac{1+1}{1+2} = \frac{2}{3}$$

$$P(L=1, S=1, N=1) = \sum_T \frac{5}{9} \cdot \frac{1}{5} \cdot \frac{2}{3} \cdot P(T | N=1)$$

$$P(T=0 | N=1) = \frac{1+1}{1+2} = \frac{2}{3} = \frac{2}{27} (P(T=0 | N=1) + P(T=1 | N=1))$$

$$P(T=1 | N=1) = \frac{0+1}{1+2} = \frac{1}{3} = \frac{2}{27} \approx 0.0741 // \textcircled{B}$$

2.3. $P(x, y, z) = P(x) P(y) P(z | x, y)$

- sve varijable binarne

<u>z</u>	<u>x</u>	<u>y</u>	$P(z x, y)$
0	0	0	0.1
0	0	1	0.2
0	1	0	0.5
0	1	1	0.9
1	0	0	0.9
1	0	1	0.8
1	1	0	0.5
1	1	1	0.1

$$P(x=1) = 0.2$$

$$P(y=1) = 0.3$$

korrelacija s oddjeljnjem ($N=1000$)

$$P(y|x=1, z=0)$$

$$P(x=1) = 0.2$$

$$P(z=0 | x=1) = \frac{P(x=1, z=0)}{P(x=1)}$$

$$P(x=1, z=0) = \sum_y P(x=1) P(y) P(z=0 | x=1, y)$$

$$= 0.2 \cdot 0.7 \cdot 0.5 + 0.2 \cdot 0.3 \cdot 0.9$$

$$= 0.124$$

Vrijednost u=oreka

$$N \cdot P(x=1, z=0)$$

$$= 124 //$$

\textcircled{B}

$$P(x=1) = \sum_0^1 \sum_0^1 P(x=1) P(y) P(z | x=1, y) = 0.2$$

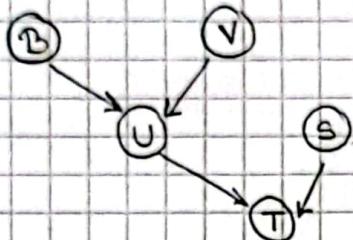
<u>y</u>	<u>z</u>	$P(x=1) \cdot P(y) \cdot P(z x=1, y)$
0	0	0.2 \cdot 0.7 \cdot 0.5 = 0.07
0	1	0.2 \cdot 0.7 \cdot 0.5 = 0.07
1	0	0.2 \cdot 0.3 \cdot 0.9 = 0.054
1	1	0.2 \cdot 0.3 \cdot 0.1 = 0.006

24.

$$\begin{array}{l} \mathcal{B}, V \rightarrow U \\ U, S \rightarrow T \end{array}$$

$$\begin{aligned} P(V=1) &= 0.3 \\ P(B=1) &= 0.1 \\ P(S=1) &= 0.05 \end{aligned}$$

• sve binarno



V	B	$P(U=1 V, B)$		U	S	$P(T=1 U, S)$
0	0	0.2		0	0	0
0	1	0.5		0	1	0.2
1	0	0.4		1	0	0.4
1	1	0.7		1	1	0.4

$$\text{MAP upit: } \underset{U, S}{\operatorname{argmax}} P(U, S | V=0, T=1)$$

$$\frac{P_1}{P_2} = ?$$

$$P(U, S, V, T) = P(B)P(V)P(S)P(U|B, V)P(T|U, S)$$

$$P(U, S | V=0, T=1) = \frac{P(U, S, V=0, T=1)}{P(U, S, V=0, T=1)} \leftarrow \text{možemo zanemariti jer se kratio u omjeru } \frac{P_1}{P_2}$$

$$\underline{P(U, S, V=0, T=1)} = \prod_B P(B)P(V=0)P(S)P(U|B, V=0)P(T=1|U, S)$$

B	U	S	$P(B) \cdot P(V=0)P(S)P(U B, V=0)P(T=1 U, S)$
0	0	0	$0.9 \cdot 0.7 \cdot 0.95 \cdot 0.8 \cdot 0 = 0$
	0	1	$0.9 \cdot 0.7 \cdot 0.05 \cdot 0.8 \cdot 0.2 = 5 \cdot 10^{-3}$
	1	0	$0.9 \cdot 0.7 \cdot 0.95 \cdot 0.2 \cdot 0.4 = 0.04788$
1	1	0	$0.9 \cdot 0.7 \cdot 0.05 \cdot 0.2 \cdot 0.4 = 2.52 \cdot 10^{-3}$
	0	0	$0.1 \cdot 0.7 \cdot 0.95 \cdot 0.5 \cdot 0 = 0$
	0	1	$0.1 \cdot 0.7 \cdot 0.05 \cdot 0.5 \cdot 0.2 = 3.5 \cdot 10^{-4}$
1	1	0	$0.1 \cdot 0.7 \cdot 0.95 \cdot 0.5 \cdot 0.4 = 0.0133$
	1	1	$0.1 \cdot 0.7 \cdot 0.05 \cdot 0.5 \cdot 0.4 = 7 \cdot 10^{-4}$

$$P(U=0, S=0, V=0, T=1) = 0$$

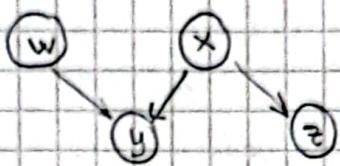
$$P(U=0, S=1, V=0, T=1) = 5 \cdot 10^{-3} = 0.00538 \leftarrow P_2$$

$$P(U=1, S=0, V=0, T=1) = 0.04788 \leftarrow P_1$$

$$P(U=1, S=1, V=0, T=1) = 2.52 \cdot 10^{-3} = 0.00322$$

$$\frac{P_1}{P_2} = \frac{874}{77} \approx 11.25 \Rightarrow \textcircled{A}$$

$$2.5 \quad P(w, x, y, z) = P(w) P(x) P(y|w, x) P(z|x) \quad (\text{sve binarno})$$



$$\begin{array}{ll} P(w=1) = 0.1 & P(w=0) = 0.9 \\ P(x=1) = 0.2 & P(x=0) = 0.8 \end{array}$$

$$\begin{array}{ll} P(z=1|x=0) = 0.9 & P(z=0|x=0) = 0.1 \\ P(z=1|x=1) = 0.7 & P(z=0|x=1) = 0.3 \end{array}$$

• učenje i prepoznavanje s oddijeljivanjem

$$P(x=0|y=1, z=0) = \mu$$

$$N = 100 \quad \text{uzorak}$$

$$\hat{\mu}_{MAP} = \frac{N_{x=0, y=1, z=0} + 1}{N_{y=1, z=0} + 2}$$

w	x	$P(y=1 w, x)$	$P(y=0 w, x)$
0	0	0	1
0	1	0.4	0.6
1	0	0.2	0.8
1	1	0.7	0.3

$$P(y=1, z=0) = \sum_w \sum_x P(w) P(x) P(y=1|w, x) P(z=0|x) = 0.0274$$

w	x	$P(w) P(x) P(y=1 w, x) P(z=0 x)$
0	0	0.9 \cdot 0.8 \cdot 0 \cdot 0.1 = 0
0	1	0.9 \cdot 0.2 \cdot 0.4 \cdot 0.3 = 0.0216
1	0	0.1 \cdot 0.8 \cdot 0.2 \cdot 0.1 = 0.0016 = 1.6 \cdot 10^{-3}
1	1	0.1 \cdot 0.2 \cdot 0.7 \cdot 0.3 = 0.0042 = 4.2 \cdot 10^{-3}

$$\begin{aligned} P(x=0, y=1, z=0) &= \sum_w P(w) P(x=0) P(y=1|w, x=0) P(z=0|x=0) \\ &= 0.8 \cdot 0.1 (P(w=0) P(y=1|w=0, x=0) + P(w=1) P(y=1|w=1, x=0)) \\ &= 0.08 (0.9 \cdot 0 + 0.1 \cdot 0.2) \\ &= 1.6 \cdot 10^{-3} = 0.0016 \end{aligned}$$

$$\hat{\mu}_{MAP} = \frac{N \cdot 0.0016 + 1}{N \cdot 0.0274 + 2} = \frac{58}{237} \approx 0.245$$

$$\begin{aligned} P(x=1, y=1, z=0) &= \sum_w P(w) P(x=1) P(y=1|w, x=1) P(z=0|x=1) \\ &= 0.2 \cdot 0.3 (P(w=0) P(y=1|w=0, x=1) + P(w=1) P(y=1|w=1, x=1)) \\ &= 0.06 (0.9 \cdot 0.4 + 0.1 \cdot 0.7) \\ &= 0.0158 \end{aligned}$$

19. Grupiranje

Strojno učenje 1, UNIZG FER, ak. god. 2023./2024.

Jan Šnajder, vježbe, v2.2

1 Zadatci za učenje

1. [Svrha: Razumjeti rad algoritma K-sredina u smislu minimizacije kriterija pogreške. Razumjeti kako rad algoritma ovisi o broju grupa K i odabiru početnih središta.]

Algoritam K-sredina minimizira kriterij pogreške $J(\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K | \mathcal{D})$. Vrijednost tog kriterija ovisi o broju grupa K , koji je unaprijed postavljen, te o položajima središta, koja se mijenjaju kroz iteracije.

- Nacrtajte skicu vrijednosti kriterija pogreške J kao funkcije broja grupa K . Koja je minimalna vrijednost funkcije J i zašto?
 - Izaberite na skici iz zadatka (a) tri vrijednosti za K i skicirajte na jednom grafikonu vrijednost kriterija pogreške J kao funkcije broja iteracija (tri krivulje).
 - Izaberite na skici iz zadatka (a) jednu vrijednost za K . Skicirajte na jednom grafikonu vrijednosti kriterija pogreške J kao funkcije broja iteracija, ali ovaj put uvezši u obzir stohastičnost uslijed slučajnog odabira početnih središta (nacrtajte nekoliko mogućih krivulja na istom grafikonu). Koje od tih krivulja su izglednije za algoritam K-means++?
2. [Svrha: Isprobati rad algoritma K-sredina i K-medoida na konkretnom primjeru. Shvatiti da je složenost ovog drugog puno nepovoljnija.] Raspolažemo skupom neoznačenih primjera:

$$\mathcal{D} = \{a = (5, 2), b = (7, 1), c = (1, 4), d = (6, 2), e = (2, 8), f = (3, 6), g = (0, 4)\}.$$

- Izvedite jedan korak algoritma K-sredina uz $K = 3$. Za početna središta odaberite $\boldsymbol{\mu}_1 = b$, $\boldsymbol{\mu}_2 = c$ i $\boldsymbol{\mu}_3 = e$.
 - Izvedite jedan korak algoritma K-medoida uz $K = 3$. Za početna središta odaberite primjere b , c i e .
 - Usporedite računalnu složenost algoritma K-sredina i K-medoida.
 - Što su prednosti, a što nedostatci algoritma K-medoida?
3. [Svrha: Isprobati izračun Randovog indeksa na konkretnom primjeru. Razumjeti primjenjivost Randovog indeksa.] Nedostatak svih algoritama grupiranja koje smo razmotrili jest što se broj grupa K mora zadati unaprijed. Osim u rijetkim slučajevima kada nam je taj broj unaprijed poznat, to predstavlja problem.

- Kada su primjeri ili podskup primjera označeni, kvaliteta grupiranja (uključivo i broj grupa K) može se procijeniti Randovim indeksom. Randov indeks zapravo izračunava točnost s kojom ćemo par jednakim označenih primjera smjestiti u istu grupu, odnosno par različito označenih primjera u različite grupe. Izračunajte Randov indeks za sljedeću particiju označenih primjera (podskupovi su grupe dobivene grupiranje, a brojke su označke klase primjera):

$$\{\{0, 0, 1, 2\}, \{1, 1\}, \{2, 2, 2, 1, 0\}\}.$$

- Skicirajte vrijednost Randovog indeksa kao funkcije broja grupa K .
- Randov indeks možemo koristiti samo ako su podaci označeni ili je podskup podataka označen. Međutim, čini se da to onda ujedno podrazumijeva da je unaprijed poznat broj grupa K . Imamo li koristi od Randovog indeksa čak i onda kada unaprijed znamo broj grupa? Možemo li ikako upotrijebiti Randov indeks, a da nam broj grupa nije unaprijed poznat?

2 Zadaci s ispita

1. (N) Raspolažemo sljedećim neoznačenim skupom primjera:

$$\mathcal{D} = \{\mathbf{x}^{(i)}\}_i = \{(1, 1), (1, 2), (2, 2), (2, 3), (3, 3)\}$$

Primjere grupiramo algoritmom K-sredina sa $K = 2$ grupe. Za početna središta odabrali smo primjere $\mathbf{x}^{(2)} = (1, 2)$ i $\mathbf{x}^{(5)} = (3, 3)$. Provedite prvu iteraciju algoritma K-sredina. **Koliko iznosi vrijednost kriterijske funkcije J nakon ažuriranja centroida?**

- A 2.962 B 1.833 C 1.667 D 2.414

2. (P) Skup neoznačenih primjera u dvodimenziskome ulaznom prostoru neka je sljedeći:

$$\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i=1}^5 = \{(0, 0), (0, 4), (2, 0), (2, 4), (4, 2)\}$$

Primjere grupiramo algoritmom K -sredina sa $K = 3$ grupe. Za početna središta grupa odaberemo nasumično primjere iz \mathcal{D} , pri čemu, naravno, pazimo da odaberemo različita središta. Ishod grupiranja i konačan iznos kriterijske funkcije J ovisit će o odabiru početnih središta. Neka je J^* vrijednost kriterijske funkcije u točki globalnog minimuma, dakle vrijednost koja odgovara najboljem grupiranju. Neka je J^+ vrijednost kriterijske funkcije u točki lokalnog minimuma, i to onoj točki lokalnog minimuma s najvećom vrijednošću funkcije J . **Koliko iznosi razlika $J^+ - J^*$?**

- A 4 B 6 C 8 D 12

3. (P) Skup neoznačenih primjera u dvodimenziskome ulaznom prostoru neka je sljedeći:

$$\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i=1}^4 = \{(0, 0), (0, 4), (2, 4), (4, 2)\}$$

Primjere grupiramo algoritmom K-sredina u $K = 2$ grupe. Grupiranje možemo shvatiti kao pretraživanje prostora stanja, gdje svako pojedino stanje odgovara jednom pridjeljivanju primjera grupama. Pritom dva stanja smatramo identičnima ako su grupiranja identična, neovisno o identitetu grupe (npr., grupiranje kod kojega prva grupa sadrži samo primjer $\mathbf{x}^{(1)}$ je identično kao i grupiranje kod kojega drugo grupa sadrži samo primjer $\mathbf{x}^{(1)}$). Neka je A_1 algoritam K-sredina s potpuno slučajno inicijaliziranim središtima, a A_2 algoritam K-sredina gdje su središta inicijalizirana algoritmom K-means++. Neka je $S(A_1)$ skup stanja koje pretražuje algoritam A_1 , a $S(A_2)$ skup stanja koje pretražuje algoritam A_2 . Izračunajte veličine ovih skupova, uvezvi u obzir mogućnost da pojedina grupa bude prazna te da dođe do izjednačenja udaljenosti primjera do centroida, što se razrješava slučajnim mehanizmom. **Koliko algoritam A_2 pretražuje manje stanja od algoritma A_1 , tj. koliko iznosi $|S(A_1)| - |S(A_2)|$?**

- A 4 B 6 C 12 D 10

4. (N) Algoritmom K-medoida (PAM) grupiramo $N = 5$ primjera. Za grupiranje koristimo mjeru različitosti, koja je za naših pet primjera definirana sljedećom matricom (matrica je simetrična, pa je donji trokut izostavljen):

$$\begin{array}{ccccc} & \mathbf{x}^{(1)} & \mathbf{x}^{(2)} & \mathbf{x}^{(3)} & \mathbf{x}^{(4)} & \mathbf{x}^{(5)} \\ \mathbf{x}^{(1)} & 0 & 0.2 & 0.9 & 0.7 & 0.5 \\ \mathbf{x}^{(2)} & & 0 & 0.9 & 0.1 & 0.6 \\ \mathbf{x}^{(3)} & & & 0 & 0.7 & 0.3 \\ \mathbf{x}^{(4)} & & & & 0 & 0.8 \\ \mathbf{x}^{(5)} & & & & & 0 \end{array}$$

Grupiramo u $K = 2$ grupe, s primjerima $\mathbf{x}^{(1)}$ i $\mathbf{x}^{(5)}$ kao početnim medoidima. Provedite prvu iteraciju algoritma K-medoida (PAM). **Koje medoide dobivamo nakon prve iteracije?**

- A $\mathbf{x}^{(1)}$ i $\mathbf{x}^{(3)}$ B $\mathbf{x}^{(3)}$ i $\mathbf{x}^{(4)}$ C $\mathbf{x}^{(1)}$ i $\mathbf{x}^{(2)}$ D $\mathbf{x}^{(2)}$ i $\mathbf{x}^{(5)}$

5. (N) Particijskim algoritmom grupiranja grupiramo $N = 1000$ primjera. Na temelju znanja o problemu zaključili smo da bi primjeri trebali formirati $K = 3$ grupe, pa smo s tim brojem grupa proveli grupiranje. Kako bismo evaluirali točnost grupiranja, slučajnim odabirom smo iz skupa

primjera uzorkovali 10 primjera, ručno smo označili primjere iz tog uzorka, i zatim na tom uzorku računamo Randov indeks. Označavanje smo proveli tako da smo svakom primjeru iz uzorka dodijelili oznaku točne grupe. Oznake grupe dobivene algoritmom grupiranja y_{pred} i oznake točnih grupa y_{true} za svih deset primjera u uzorku su sljedeće:

i	1	2	3	4	5	6	7	8	9	10
$y_{pred}^{(i)}$	0	1	2	2	1	0	0	2	1	2
$y_{true}^{(i)}$	1	1	0	2	0	0	1	1	1	2

Koliko iznosi Randov indeks grupiranja izračunat na ovom uzorku?

- A 0.27 B 0.56 C 0.64 D 0.70

6. (N) Želimo grupirati $N = 1000$ primjera, ali nemamo nikakvih saznanja o optimalnom broju grupa. Kako bismo odredili optimalan broj grupa, odlučili smo označiti uzorak primjera i na tom uzorku izračunati Randov indeks $RI(K)$ za grupiranja dobivena s različitim brojem grupa K . Naposlijetku ćemo onda kao optimalan broj grupa odabrati onaj K koji maksimizira Randov indeks, $K^* = \operatorname{argmax}_K RI(K)$. Budući da ne znamo koji je točan broj grupa, umjesto označavanja pojedinačnih primjera označavamo parove primjera. U tu svrhu smo iz skupa primjera uzorkovali 16 različitih primjera, uparili ih u 8 različitih parova primjera, te smo za svaki par primjera ručno označili trebaju li dotični primjeri pripadati istoj grupi ili ne. Rezultat označavanja je takav da tri para primjera trebaju pripadati istoj grupi (indeksi parova 1–3), a pet različitih grupama (indeksi parova 4–8). Nakon toga proveli smo grupiranje za $K \in \{3, 4, 5\}$ grupa. Za uzorak označenih primjera dobili smo ovakve grupe:

$$\begin{aligned} K = 3 : & \{1, 1, 2, 4, 8\} \{2, 3, 7\} \{4, 5, 3, 5, 6, 6, 7, 8\} \\ K = 4 : & \{1, 1, 2\} \{4, 8, 4\} \{2, 3, 7, 5, 7\} \{3, 5, 6, 6, 8\} \\ K = 5 : & \{1, 1\} \{3, 4, 8\} \{2, 2, 4\} \{7, 5, 7, 3, 5, 6\} \{6, 8\} \end{aligned}$$

Brojke označavaju indeks para primjera. Na primjer, u grupiranju sa $K = 3$ grupe par primjera s indeksom 1 našao se u istoj grupi, a par primjera s indeksom 2 u različitim grupama. Izračunajte Randov indeks $RI(K)$ te optimalan broj grupa K^* prema Randovom indeksu, za $K \in \{3, 4, 5\}$. **Koliko iznosi Randov indeks za optimalan broj grupa, $RI(K^*)$?**

- A 0.375 B 0.625 C 0.750 D 0.875

7. (N) Algoritmom K-sredina grupiramo $N = 1000$ primjera. U tom skupu nalazi se i uzorak od 11 primjera označenih oznakama $\mathcal{Y} = \{1, 2, 3, 4\}$. Međutim, nismo sigurni hoće li grupiranje u četiri grupe doista dati optimalne rezultate, pa isprobavamo grupiranje sa $K = 3$ i $K = 4$ grupe. Rezultati su sljedeći:

$$\begin{aligned} K = 3 : & \{\{1, 2, 4, 4\}, \{2, 3, 3\}, \{1, 1, 3, 4\}\} \\ K = 4 : & \{\{1, 2, 2, 4, 4\}, \{3, 3\}, \{1, 1\}, \{3, 4\}\} \end{aligned}$$

gdje podskupovi odgovaraju grupama, a brojke oznakama primjera. Izračunajte Randov indeks za oba ova grupiranja. **Koliko je Randov indeks za $K = 4$ veći od Randovog indeksa za $K = 3$?**

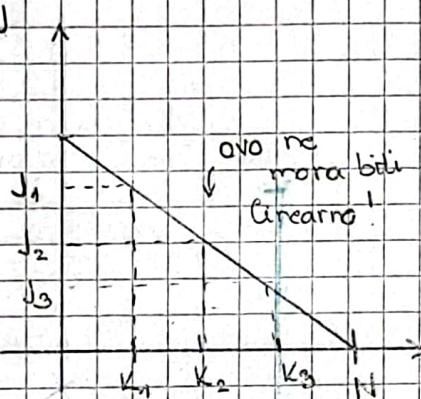
- A 0.0164 B 0.0945 C 0.0727 D 0.0682

V19 - Grupiranje

I Zadaci za učenje

1.1. $J(\vec{\mu}_1, \dots, \vec{\mu}_K | D) = \sum_{k=1}^K \sum_{i=1}^N b_k^i \|x^i - \vec{\mu}_k\|^2$

a)



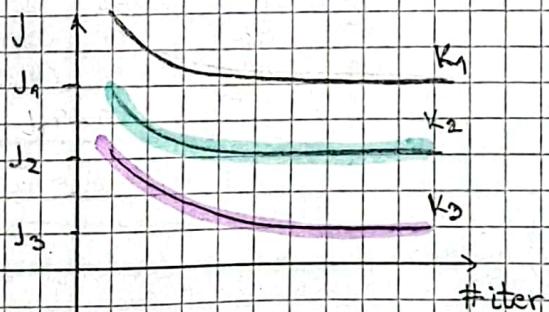
- minimalna vrijednost kriterijske funkcije je 0

- vrijednost $J = 0$ se postiže za $K = N$, tj. broj grupa jednak je broju primjera pa se efektivno svaki primjer nalazi u zadatnoj grupi

- također svaki primjer je onda i centroid svoje grupe jer je $\|x^i - \vec{\mu}_k\|^2 = 0 \Rightarrow J = 0$

b) $\Rightarrow K_1 < K_2 < K_3$
 $J_1 \geq J_2 \geq J_3$

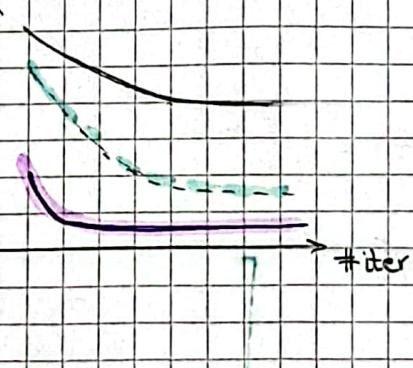
(može se desiti plato na tlim intervalima, ali za potrebe sljedeće recimo da nema plato)



- za veći broj grupa K greska postaje manja jer se radi o složenijem modelu

- vrijednost kriterijske funkcije J nakon određenog broja iteracija postaje konstantna jer algoritam K srednjih vrijednosti konvergira

c)



- fiksani K

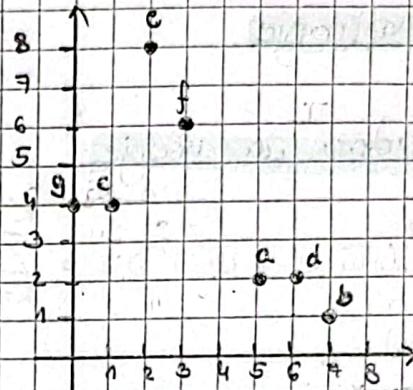
- različiti početni centroidi

- razlika je u brzini konvergencije i postignutoj vrijednosti kriterijske funkcije J

- za algoritam K-means++ najverovatnija je krivulja jer najbrže konvergira i jer ima najmanju početnu vrijednost

1.2.

$$\mathcal{D} = \{ \begin{array}{l} a = (5, 2) \\ b = (7, 1) \\ c = (1, 4) \\ d = (6, 2) \\ e = (2, 8) \\ f = (3, 6) \\ g = (0, 4) \end{array} \}$$



a) $k = 3$

pocádky
střediska

$$\begin{aligned} \vec{\mu}_1 &= b \\ \vec{\mu}_2 &= c \\ \vec{\mu}_3 &= e \end{aligned}$$

$$\vec{x}^i \quad \|\vec{x}^i - \vec{\mu}_1\|^2$$

$$(7, 1)$$

$$\|\vec{x}^i - \vec{\mu}_2\|^2$$

$$(1, 4)$$

$$\|\vec{x}^i - \vec{\mu}_3\|^2$$

$$\vec{b}_k^i = (b_1^i, b_2^i, b_3^i)$$

$a = (5, 2)$	5
$b = (7, 1)$	0
$c = (1, 4)$	45
$d = (6, 2)$	2
$e = (2, 8)$	74
$f = (3, 6)$	41
$g = (0, 4)$	58

20	45	(1, 0, 0)
45	74	(1, 0, 0)
0	17	(0, 1, 0)
29	52	(1, 0, 0)
17	0	(0, 0, 1)
8	5	(0, 0, 1)
1	20	(0, 1, 0)

NOVÍ CENTROIDY

$$\vec{\mu}_1 = \frac{\sum_{i=1}^6 b_1^i \vec{x}^i}{\sum b_1^i} = \frac{\vec{x}^1 + \vec{x}^2 + \vec{x}^4}{3} = \frac{1}{3}(a+b+d) = (6, \frac{5}{3})$$

$$\vec{\mu}_2 = \frac{c+g}{2} = (\frac{1}{2}, 4)$$

$$\vec{\mu}_3 = \frac{1}{2}(e+f) = (\frac{5}{2}, 7)$$

b) jedan krok K-medoida

$$k = 3$$

$$\vec{\mu}_1 = b = (7, 1)$$

$$\vec{\mu}_2 = c = (1, 4)$$

$$\vec{\mu}_3 = e = (2, 8)$$

$$b_k^i = \begin{cases} 1 & i = \arg \min V(\vec{x}^i, \vec{\mu}_j) \\ 0, \text{ ináče} & \end{cases}$$

\rightarrow koristimo euklidskou vzdálosť až su doložena správne.

$$\vec{\mu}_k = \arg \min_{\vec{\mu}_j \in \mathcal{D} \setminus \{ \vec{\mu}_k \}} \sum b_k^i V(\vec{x}^i, \vec{\mu}_j)$$

$$M = \{b, c, e\}$$

NOVI METODI

\vec{x}^i	$\vec{b}^i = (b_1^i, b_2^i, b_3^i)$
a = (5, 2)	(1, 0, 0)
b = (7, 1)	(1, 0, 0)
c = (1, 4)	(0, 1, 0)
d = (6, 2)	(1, 0, 0)
e = (2, 8)	(0, 0, 1)
f = (3, 6)	(0, 0, 1)
g = (0, 4)	(0, 1, 0)

$$\vec{\mu}_j = \underset{\vec{\mu}_j \in DM \cup \{\vec{\mu}_1\}}{\text{argmin}} \sum b_n^i \|\vec{x}^i - \vec{\mu}_j\|$$

$$\sum b_n^i \|\vec{x}^i - \vec{\mu}_j\| \Leftrightarrow \sum b_n^i \|\vec{x}^i - \vec{\mu}_j\|^2$$

$$\vec{\mu}_j \in \{a, b, d, f, g\}$$

$$b_n^i = 1 \rightarrow a \{a, b, d\}$$

$\sum b_n^i \cdot v$	$\vec{\mu}_j$	$\ a = (5, 2) - \vec{\mu}_j\ ^2$	$\ b = (7, 1) - \vec{\mu}_j\ ^2$	$\ d = (6, 2) - \vec{\mu}_j\ ^2$
6	a = (5, 2)	0	5	1
7	b = (7, 1)	5	0	8
3	d = (6, 2)	1	2	0
8	f = (3, 6)	20	41	25
12	g = (0, 4)	29	58	40

$$\boxed{\vec{\mu}_1 = d = (6, 2)}$$

$$\underline{\underline{\vec{\mu}_2}}$$

$$\vec{\mu}_j \in \{a, c, d, f, g\}$$

$$\|c = (1, 4) - \vec{\mu}_j\|^2$$

$$\|g = (0, 4) - \vec{\mu}_j\|^2$$

$$\sum b_n^i \cdot v$$

$\vec{\mu}_j$
a = (5, 2)
c = (1, 4)
d = (6, 2)
f = (3, 6)
g = (0, 4)

$$20$$

$$0$$

$$1$$

$$8$$

$$1$$

$$25$$

$$5$$

$$36$$

$$25$$

$$0$$

$$\sum b_n^i \cdot v$$

$$45$$

$$5$$

$$65$$

$$33$$

$$1$$

$$\boxed{\vec{\mu}_2 = g = (0, 4)}$$

$$\underline{\underline{\vec{\mu}_3}}$$

$$\vec{\mu}_j \in \{a, d, e, f, g\}$$

$$\|e = (2, 8) - \vec{\mu}_j\|^2$$

$$\|f = (3, 6) - \vec{\mu}_j\|^2$$

$$\sum b_n^i \cdot v$$

$\vec{\mu}_j$
a = (5, 2)
d = (6, 2)
e = (2, 8)
f = (3, 6)
g = (0, 4)

$$45$$

$$58$$

$$0$$

$$5$$

$$20$$

$$20$$

$$25$$

$$5$$

$$0$$

$$10$$

$$65$$

$$77$$

$$5$$

$$0$$

$$30$$

$$\boxed{\vec{\mu}_3 = e = (2, 8)}$$

övise o. ind.

c)

složenost algoritma K -sredina

$$O(TnNK) = T \cdot (O(nNK) + O(nN))$$

 $O(nNK)$

- = složenost pridjeljivanja koeficijenta b_i^i
- = složenost euklidiske udaljenosti $O(n)$, računa se $\forall \vec{x}^i \in \vec{\mu}_j$ jer je potrebno odrediti $\arg\min \|\vec{x}^i - \vec{\mu}_j\|$

 $O(nN)$

= složenost izračuna centroida

$$\frac{\sum b_i^i \vec{x}^i}{\sum b_i^i}$$

- Linearna složenost u svim parametrima

 $T = \text{broj iteracija}$

složenost

algoritma K -medoida

$$O(TK(N-K)^2) = T \cdot (O(K(N-K)) + O(K(N-K)^2))$$

 $O(K(N-K))$

- = svrstavanje u grupu, tj. određivanje koef. b_i^i
- = svaki kentralni se uspostavlja sa $\forall \vec{x}^i \in D \setminus M$
- čija je standardnost $N-K$

 $O(K(N-K)^2)$

= odabir novog medoida : $\arg\min_{\vec{\mu}_j \in D \setminus M} \sum_{\vec{x}^i \in D \setminus M} b_i^i v(\vec{x}^i, \vec{\mu}_j)$

- Kvadratna složenost u N

d)

Prednosti K -medoida

- primjeri ne moraju biti vektori
- sličnost se ne mora mjeriti euklid. udaljenšću

Nedostaci K -medoida

- velika vremenska složenost $O(TK(N-K)^2)$

1.3.

a) $\{\{0,0,1,2\}, \{1,1\}, \{2,2,2,1,0\}\}$

- brojke = stvarne označke klasa

- podeljivanje = grupe dobivene grupiranjem

$$Q = \frac{a+b}{\binom{N}{2}}$$

$$N = 10$$

$a = \text{broj jednako označenih parova u istim grupama}$

$b = \text{broj različito označenih parova u različitim grupama}$

parovi	0 u	I	$\binom{2}{2} = 1$	a = 5
parovi	1 u	II	$\binom{2}{2} = 1$	
parovi	2 u	III	$\binom{3}{2} = 3$	

I \leftrightarrow II

0 u I	s 1 u II	: $2 \cdot 2 = 4$
1 u II	s 1 u III	: 0
2 u I	s 1 u III	: $1 \cdot 2 = 2$

$$Q = \frac{a+b}{\binom{N}{2}} = \frac{28+5}{\binom{10}{2}}$$

I \leftrightarrow III

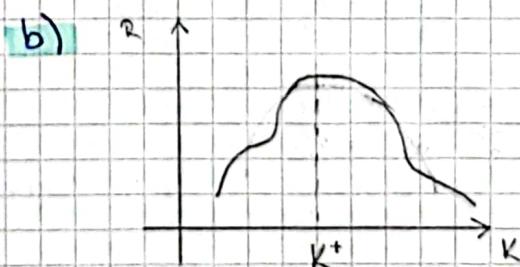
0 u II	s 2 u III	: $2 \cdot 3 = 6$
0 u II	s 1 u III	: $2 \cdot 1 = 2$
1 u II	s 2 u III	: $1 \cdot 3 = 3$
1 u I	s 0 u III	: $1 \cdot 1 = 1$
2 u I	s 1 u III	: $1 \cdot 1 = 1$
2 u I	s 0 u II	: $1 \cdot 1 = 1$

$$b = 28$$

$$Q = \frac{b}{5} = 0.6$$

II \leftrightarrow III

1 u II	s 2 u III	: $2 \cdot 3 = 6$
1 u II	s 0 u III	: $2 \cdot 1 = 2$



! nije kvadratna vrijednost, samo je fja unimodala

c)

Ako unaprijed znamo broj grupa imamo koristi od Randovog indeksa. Indeks nam u tom slučaju služi kao mjera uspješnog klasifikiranja podataka.

Ako nam broj grupa nije unaprijed poznat možemo koristit Randov indeks na manjem ručno označenom skupu podataka.

II zadaci s ispitom

2.1. $K = 2$ Kmeans

$$\mathcal{D} = \{(\vec{x}^i)\}_{i=1}^5 = \{(1,1), (1,2), (2,2), (2,3), (3,3)\}$$

$$\begin{aligned}\vec{\mu}_1 &= \vec{x}^2 = (1, 2) \\ \vec{\mu}_2 &= \vec{x}^5 = (3, 3)\end{aligned}$$

$$b_{jk}^i = \begin{cases} 1 & k = \operatorname{argmin} \|\vec{x}^i - \vec{\mu}_j\| \\ 0 & \text{inac}\end{cases}$$

$$\vec{\mu}_k = \frac{\sum b_{jk}^i \vec{x}^i}{\sum b_{jk}^i}$$

\vec{x}^i	$\ \vec{x}^i - \vec{\mu}_1\ ^2$
(1, 1)	0
(1, 2)	0
(2, 2)	1
(2, 3)	2
(3, 3)	5

\vec{x}^i	$\ \vec{x}^i - \vec{\mu}_2\ ^2$	$b^i = (b_1^i, b_2^i)$
(1, 1)	8	(1, 0)
(1, 2)	5	(1, 0)
(2, 2)	2	(1, 0)
(2, 3)	1	(0, 1)
(3, 3)	0	(0, 1)

$$\vec{\mu}_1 = \frac{\vec{x}^1 + \vec{x}^2 + \vec{x}^3}{3} = \left(\frac{4}{3}, \frac{5}{3}\right)$$

$$\vec{\mu}_2 = \frac{\vec{x}^4 + \vec{x}^5}{2} = \left(\frac{5}{2}, 3\right)$$

$$J = \sum_{k=1}^2 \sum_{i=1}^5 b_{ik}^i \|\vec{x}^i - \vec{\mu}_k\|^2 = \frac{11}{6} = 1.833 \quad (\text{B})$$

$$\begin{aligned}\|\vec{x}^1 - \vec{\mu}_1\| &= \sqrt{5}/3 \\ \|\vec{x}^2 - \vec{\mu}_1\| &= \sqrt{2}/3 \\ \|\vec{x}^3 - \vec{\mu}_1\| &= \sqrt{5}/3\end{aligned}$$

$$\begin{aligned}\|\vec{x}^4 - \vec{\mu}_2\| &= 1/2 \\ \|\vec{x}^5 - \vec{\mu}_2\| &= 1/2\end{aligned}$$

\vec{x}^i	$\ \vec{x}^i - \vec{\mu}_1\ $	$\ \vec{x}^i - \vec{\mu}_2\ $	$\frac{\vec{x}^i}{\ \vec{x}^i\ }$
(1, 1)	0.7453	2.5	(1, 0)
(1, 2)	0.4714	1.8027	(1, 0)
(2, 2)	0.7453	1.118	(1, 0)
(2, 3)	1.4907	0.5	(0, 1)
(3, 3)	2.1313	0.5	(0, 1)

2.2.

\vec{x}^i

- (0, 0)
- (0, 4)
- (2, 0)
- (2, 4)
- (4, 2)

$$N = 5$$

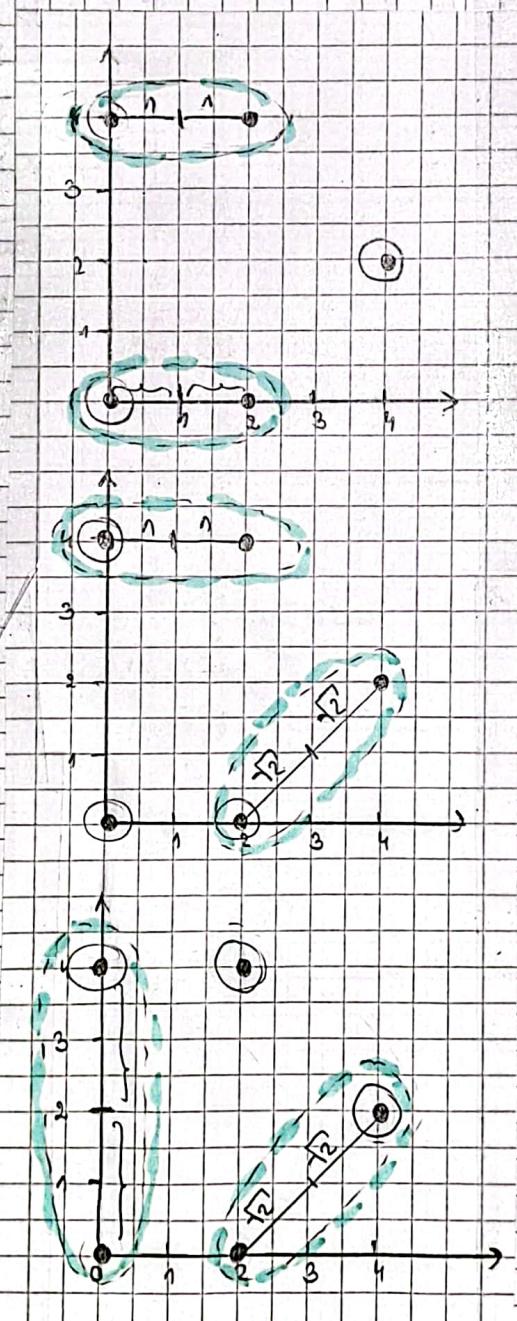
$$K = 3$$

početna središta rozmíčeno odabранa iz D

$$J = \sum_{k=1}^2 \sum_{i=1}^5 b_k^i \| \vec{x}^i - \vec{\mu}_j \|^2$$

$$J^+ - J^* = ?$$

J^+ = globální minimum
 J^* = globální minimum



$$J = 4 \cdot 1^2 = 4 = J^*$$

$$J = 2 \cdot 1^2 + 2 \cdot (\sqrt{2})^2 = 6$$

$$J = 2 \cdot 2^2 + 2 \cdot (\sqrt{2})^2 = 12 = J^+$$

$$J^+ - J^* = 12 - 4 = 8$$

$$\underline{2.3.} \quad D = \{\vec{x}^4\} = \{(0,0), (0,4), (2,4), (4,2)\}$$

$$K=2$$

K-means - pretražuje k^N particija
 $k^N = 2^M = 16$

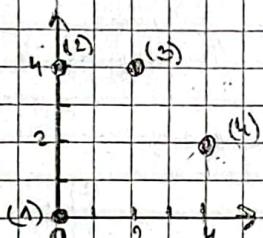
K - means

Primjeru
↑
 $O \rightarrow$ grupa

1	2	3	4
0	0	0	0
0	0	0	1
0	0	1	0
0	0	1	1
0	1	0	0
0	1	0	1
0	1	1	0
0	1	1	1
1	0	0	0

ostal sim. →

$$|S(A_1)| = 7$$



$$|S(A_1)| - |S(A_2)| = 3$$

24

K-medoids

Z = 5

$$V = 2$$

$$\vec{u} = \vec{x}$$

$$\mu_2 = x^b$$

	\vec{x}^1	\vec{x}^2	\vec{x}^3	\vec{x}^4	\vec{x}^5
\vec{x}^1	0	0.2	0.9	0.7	0.5
\vec{x}^2	0.2	0	0.9	0.1	0.6
\vec{x}^3	0.9	0.9	0	0.7	0.3
\vec{x}^4	0.7	0.1	0.7	0	0.8
\vec{x}^5	0.5	0.6	0.3	0.8	0

$$b_{\underline{z}}^i = \begin{cases} 1 & \text{ako } \underline{z} \in \operatorname{argmin}_{\underline{y}} \mathcal{J}(\vec{x}^i, \underline{w}_i) \\ 0 & \text{inče} \end{cases}$$

$$\mu_e = \operatorname{argmin}_{j \in DV \setminus \{\bar{\mu}_e\}} b_e^j \cup (\vec{x}^j, \bar{\mu}_j)$$

- | | | | |
|-----------|---------------------------------------------------------------------------------------------------|-------------------|----------|
| \bullet | $\left[\begin{array}{c} \vec{x} \\ \vec{x} \\ \vec{x} \\ \vec{x} \\ \vec{x} \end{array} \right]$ | $\xrightarrow{1}$ | $(1, 0)$ |
| | | $\xrightarrow{2}$ | $(1, 0)$ |
| | | $\xrightarrow{3}$ | $(0, 1)$ |
| | | $\xrightarrow{4}$ | $(1, 0)$ |
| \bullet | $\left[\begin{array}{c} \vec{x} \\ \vec{x} \end{array} \right]$ | $\xrightarrow{5}$ | $(0, 1)$ |

Kandidati za \vec{y}_1 e $\{\vec{x}^1, \vec{x}^2, \vec{x}^3, \vec{x}^4\}$

$v(\vec{x}^4, \vec{x}^1)$		$v(\vec{x}^4, \vec{x}^2)$		$v(\vec{x}^4, \vec{x}^3)$		$v(\vec{x}^4, \vec{x}^4)$	
\vec{x}^1	0	\vec{x}^2	0.2	\vec{x}^3	0.9	\vec{x}^4	0.7
\vec{x}^2	0.2	\vec{x}^1	0	\vec{x}^4	0.9	\vec{x}^3	0.1
\vec{x}^4	0.7				<th></th> <td>0</td>		0
0.9		0.3		2.7		0.8	0
		↓					
$\mu_1 = \vec{x}^2$							

Kandidati za $\vec{\mu}_2 \in \{\vec{x}^2, \vec{x}^3, \vec{x}^4, \vec{x}^5\}$

\vec{x}^i	$v(\vec{x}^i, \vec{x}^2)$	$v(\vec{x}^i, \vec{x}^3)$	$v(\vec{x}^i, \vec{x}^4)$	$v(\vec{x}^i, \vec{x}^5)$
\vec{x}^3	0.9	0	0.7	0.3
\vec{x}^5	0.6	0.3	0.8	0
\vec{x}^2	1.5	0.3	1.5	0.3

$$\boxed{\vec{\mu}_2 = \vec{x}^3 \text{ ili } \vec{x}^5}$$

- odluka ovisi o preferenciji pratrživog algoritma

(D) \vec{x}^2 i \vec{x}^5

2.5.

$$N = 1000$$

$$k=3 \quad \binom{?}{2} \quad \binom{?}{2} \quad \binom{?}{2}$$

$$\{ \{1, 1, 0\}, \{1, 0, 1\}, \{0, 1, 2\} \}^2$$

$$a = 1 + 1 + 1 = 3$$

$$\begin{aligned} b &= 2 \cdot 1 + 1 \cdot 2 + 2 \cdot 1 + 2 \cdot 2 + 1 \cdot 1 + 1 \cdot 2 + 2 \cdot 1 + 2 \cdot 2 + 1 \cdot 2 + 1 \cdot 1 \\ &= 2 + 2 + 2 + 4 + 1 + 2 + 2 + 4 + 2 + 1 \\ &= 10 + 1 + 10 + 1 \\ &= 22 \end{aligned}$$

$$R = \frac{a+b}{\binom{N}{2}} = \frac{22+3}{\binom{10}{2}} = \frac{5}{9} \approx 0.5556$$

(B)

2.6.

$$N = 1000$$

Parovi $1, 2, 3 \rightarrow$ parovi iz iste grupe

$4, 5, 6, 7, 8 \rightarrow$ parovi iz različitih grupa

$$\begin{array}{lll} k=3 & \{1, 1, 2, 4, 8\} & \{2, 3, 7\} \\ \{1, 4, 2\} & \{4, 8, 4\} & \{2, 3, 7, 5, 7\} \\ k=5 & \{1, 1\}, \{3, 4, 8\} & \{2, 2, 4\}, \{7, 5, 7, 3, 5, 6\} \\ & & \{6, 8\} \end{array}$$

Parovi	true ista grupa?	$k=3$ ista grupa?	$k=4$ ista grupa?	$k=5$ ista grupa?
1-1	1	0	1	1
2-2	1	0	0	1
3-3	1	0	0	0
4-4	0	0	1	0
5-5	0	0	0	1
6-6	0	1	1	0
7-7	0	0	1	1
8-8	1	0	0	0

$$RI(k) = \frac{TN + TP}{\binom{N}{2}}$$

$$RI(k=3) = \frac{81 \cdot 78}{100 \cdot 98} = 0.525 + 1 \text{ Optimalni } RI(k^* = 5) = 0.625$$

$$RI(k=4) = \frac{1+2}{8} = \frac{3}{8} = 0.375$$

$$RI(k=5) = \frac{2+3}{8} = \frac{5}{8} = 0.625$$

(B)

2.7.

$$N = 1000$$

$$y = \{1, 2, 3, 4\}$$

$$\begin{array}{l} K=3 : \\ \{ \{1, 2, 3, 4\} \} \\ \{ \{1, 2, 2, 4\} \} \\ \{ \{1, 2, 4, 4\} \} \end{array}$$

$$\begin{array}{l} \{ \{2, 3, 3\} \} \\ \{ \{3, 3\} \} \\ \{ \{1, 1, 3\} \} \\ \{ \{1, 1, 3\} \} \\ \{ \{3, 4\} \} \end{array}$$

zu $K=3$

$$a = 1 + 1 + 1 = 3$$

$$\begin{aligned} b = & 1 + 2 + 2 + 2 + 4 \\ & + 1 + 1 + 2 + 1 + 1 + 4 + 2 \\ & + 2 + 1 + 1 + 4 + 2 \end{aligned}$$

$$b = 11 + 12 + 10 = 33$$

$$Q = \frac{a+b}{\binom{N}{2}} = \frac{36}{\binom{11}{2}} = \frac{36}{55}$$

zu $K=4$

$$a = 1 + 1 + 1 + 1 = 4$$

$$\begin{aligned} b = & 2 + 4 + 4 \\ & + 4 + 4 \\ & + 1 + 4 + 2 + 2 + 2 \\ & + 4 + 2 \\ & + 2 + 2 \end{aligned}$$

$$b = 10 + 8 + 8 + 6 + 4 = 36$$

$$Q = \frac{a+b}{\binom{N}{2}} = \frac{40}{\binom{11}{2}} = \frac{8}{11}$$

$$\text{relative} = \frac{8}{11} - \frac{36}{55} = \frac{4}{55} \approx 0,07273 \Rightarrow \text{(C)}$$

20. Grupiranje II

Strojno učenje 1, UNIZG FER, ak. god. 2023./2024.

Jan Šnajder, vježbe, v2.3

1 Zadatci za učenje

1. [Svrha: Razumjeti model miješane gustoće i razlog zašto maksimizacija log-izglednosti nije analitički rješiva. Razumjeti kako uvođenje latentnih varijabli rješava taj problem. Razumjeti, na općenitoj razini, E-korak i M-korak. Razumjeti rad algoritma kao maksimizacije log-izglednosti i razumjeti kako ishod ovisi o broju grupa i početnoj inicijalizaciji.] Algoritam maksimizacije očekivanja (EM-algoritam), kada se koristi za grupiranje, zapravo je poopćenje algoritma K-sredina.

- Što je prednost, a što nedostatak, algoritma maksimizacije očekivanja primjenjenog na GMM u odnosu na algoritam K-sredina?
- Napišite izraz za gustoću $p(\mathbf{x})$ za model miješane gustoće (bez latentnih varijabli) i izraz za pripadnu (nepotpunu) log-izglednost.
- Napišite izraz za mješavinu s latentnim varijablama i izvedite izraz za (potpunu) log-izglednost tog modela. Možemo li dalje raditi izravno s tom log-izglednošću? Zašto?
- Definirajte E-korak i M-korak algoritma maksimizacije očekivanja primjenjenog na Gaussovou mješavinu.
- Skicirajte vrijednost log-izglednosti $\ln \mathcal{L}(\boldsymbol{\theta}|\mathcal{D})$ modela Gaussove mješavine kao funkcije broja iteracija, i to za tri različite vrijednosti parametra K (broj grupa): $K = 1$, $K = 10$ i $K = 100$. Na istom grafikonu skicirajte krivulju za $K = 10$ kada se za inicijalizaciju središta koristi algoritam K-sredina.

2. [Svrha: Isprobati rad algoritma hijerarhijskog aglomerativnog grupiranja (HAC) na konkretnom primjeru, za slučaj kada primjeri nisu vektori. Uočiti razliku između udaljenosti i sličnosti te razliku između jednostrukih i potpunih povezanosti.] Jednako kao i algoritam K-medoida, algoritam hijerarhijskog aglomerativnog grupiranja može se primjeniti u slučajevima kada primjeri nisu prikazani kao vektori značajki te kada umjesto mjere udaljenosti između vektora raspoložemo općenitjom mjerom sličnosti (ili različitosti). Neka je *sličnost* primjera iz \mathcal{D} definirana sljedećom matricom sličnosti:

$$S = \begin{pmatrix} & a & b & c & d & e \\ a & 1.00 & 0.26 & 0.15 & 0.20 & 0.17 \\ b & 0.26 & 1.00 & 0.24 & 0.31 & 0.31 \\ c & 0.15 & 0.24 & 1.00 & 0.20 & 0.50 \\ d & 0.20 & 0.31 & 0.20 & 1.00 & 0.24 \\ e & 0.17 & 0.31 & 0.50 & 0.24 & 1.00 \end{pmatrix}$$

- Izgradite dendrogram uporabom jednostrukog povezivanja. Kada bi bilo potrebno napraviti particiju grupa, na kojoj biste razini presjekli taj dendrogram?
- Izgradite dendrogram uporabom potpunog povezivanja. Kada bi bilo potrebno napraviti particiju grupa, na kojoj biste razini presljekli taj dendrogram?
- [Svrha: Razumjeti kako se unutarnji kriterij algoritma grupiranja može (pokušati) upotrijebiti za provjeru grupiranja (odabir optimalnog broja grupa). Razumjeti da Akaikeov kriterij u stvari oponaša regulariziranu funkciju pogreške, koja pak aproksimira pogrešku generalizacije.]

- Skicirajte krivulju log-izglednosti kod EM-algoritma kao funkciju broja grupa K . Obrazložite izgled krivulje. Možete li temeljem ove krivulje odrediti optimalan broj grupa? Kako?

- (b) Optimizacija broja grupa K može se provesti nekim kriterijem koji kombinira funkciju pogreške (odnosno log-izglednost) i složenost modela. Takav kriterij odgovara strukturnome riziku modela, koji je minimalan za optimalan broj grupa. Jedan takav kriterij jest Akaikeov informacijski kriterij (AIC):

$$K^* = \operatorname{argmin}_K (-2 \ln \mathcal{L}(K) + 2q(K))$$

gdje je $-\ln \mathcal{L}(K)$ negativna log-izglednost podataka za K grupa, a $q(K)$ je broj parametara modela s K grupa.

Pretpostavite da podatci \mathcal{D} u stvarnosti dolaze iz $K = 5$ grupa. Podatke grupiramo dvjema varijantama EM-algoritma: standardni algoritam i preinačeni algoritam s dijeljenom kovarijacijskom matricom (zajednička kovarijacijska matrica procijenjena nad čitavim skupom primjera \mathcal{D} na početku izvođenja algoritma). Skicirajte za ta dva algoritma funkciju koju minimizira Akaikeov minimizacijski kriterij.

2 Zadaci s ispita

1. (P) Algoritam GMM koristimo za grupiranje $N = 10$ primjera u dvodimenzijskome ulaznom prostoru. Skup primjera koje grupiramo je sljedeći:

$$\mathcal{D} = \{(0,0), (1,1), (1,2), (2,2), (2,3), (5,0), (5,1), (6,0), (6,6), (7,7)\}$$

Razmatramo tri modela GMM:

- \mathcal{H}_1 : $K = 2$ grupa, puna kovarijacijska matrica
- \mathcal{H}_2 : $K = 2$ grupa, izotropna kovarijacijska matrica
- \mathcal{H}_3 : $K = 3$ grupe, izotropna kovarijacijska matrica

Za sva tri modela kovarijacijska matrica je nedijeljena, dakle svaka komponenta ima svoju kovarijacijsku matricu. Za početne centroide odabiremo nasumično dva odnosno tri primjera iz \mathcal{D} , ovisno o broju grupa K . Za svaki model grupiranje ponavljamo 100 puta te kao konačno grupiranje uzimamo ono s najvećom log-izglednošću na skupu \mathcal{D} . Zanima nas kojoj grupi najvjerojatnije pripada primjer $\mathbf{x}^{(5)} = (2, 3)$, to jest zanima nas k koji maksimizira odgovornost $h_k^{(5)} = P(y = k | \mathbf{x}^{(5)})$. Ta vrijednost će biti različita za ova tri modela. Označimo sa h_α maksimalnu odgovornost za primjer $\mathbf{x}^{(5)}$ u modelu \mathcal{H}_α , to jest vjerojatnost pripadanja tog primjera najvjerojatnijoj grupi dobivenoj grupiranjem pomoću modela \mathcal{H}_α . **Što možemo zaključiti o odgovornostima h_α za ova tri modela?**

- A $h_{\alpha_1} > h_{\alpha_2} > h_{\alpha_3}$ B $h_{\alpha_1} < h_{\alpha_2} < h_{\alpha_3}$ C $h_{\alpha_2} > h_{\alpha_1} > h_{\alpha_3}$ D $h_{\alpha_2} < h_{\alpha_1} < h_{\alpha_3}$

2. (P) Za grupiranje skupa primjera \mathcal{D} koristimo algoritam GMM. Koristimo nekoliko varijanti tog modela:

- \mathcal{H}_1 : Model sa $K = 50$ središta inicijaliziranim algoritmom K-sredina
- \mathcal{H}_2 : Model sa $K = 50$ središta inicijaliziranim algoritmom K-sredina i dijeljenom kov. matricom
- \mathcal{H}_3 : Model sa $K = 50$ slučajno inicijaliziranim središtima i dijeljenom kov. matricom
- \mathcal{H}_4 : Model sa $K = 10$ središta inicijaliziranim algoritmom K-sredina i dijeljenom kov. matricom

Sa svakim modelom grupiranje ponavljamo 1000 puta i zatim za svaki model crtamo graf funkcije log-izglednosti kroz iteracije EM-algoritma, uprosječen kroz svih 1000 ponavljanja. Neka je LL_α^0 prosječna log-izglednost za model \mathcal{H}_α na početku izvođenja EM-algoritma, a neka je LL_α^* prosječna log-izglednost za taj model na kraju izvođenja EM-algoritma. **Što možemo unaprijed zaključiti o ovim log-izglednostima?**

- A $LL_2^0 \geq LL_4^0, LL_1^* \geq LL_2^* \geq LL_3^*$
 B $LL_3^0 \geq LL_4^0, LL_1^* \geq LL_3^* \geq LL_4^*$
 C $LL_2^0 \geq LL_4^0 \geq LL_3^0, LL_1^* \geq LL_2^*$
 D $LL_2^0 \leq LL_4^0, LL_2^* \leq LL_1^* \geq LL_3^*$

3. (P) Skup neoznačenih primjera \mathcal{D} grupiramo modelom GMM treniranim EM-algoritmom. Koristimo nekoliko varijanti tog modela:

\mathcal{H}_1 : Model sa $K = 25$ središta inicijaliziranima algoritmom K-means++

\mathcal{H}_2 : Model sa $K = 50$ slučajno inicijaliziranim središtima i dijeljenom kovarijacijskom matricom

\mathcal{H}_3 : Model sa $K = 50$ središta inicijaliziranima algoritmom K-sredina i dijeljenom kovarijacijskom matricom

Sa svakim modelom grupiranje ponavljam 1000 puta i zatim za svaki model crtamo graf funkcije log-izglednosti kroz iteracije EM-algoritma, uprosječen kroz svih 1000 ponavljanja. Neka je LL_{α}^0 prosječna log-izglednost za model \mathcal{H}_{α} na početku izvođenja EM-algoritma, LL_{α}^* prosječna log-izglednost za taj model na kraju izvođenja EM-algoritma te neka je k_{α} broj iteracija EM-algoritma za taj model. **Što možemo zaključiti o očekivanim odnosima između ovih vrijednosti?**

- [A] $LL_1^0 \geq LL_2^0, LL_2^* \geq LL_3^*, k_1 \geq k_2$
- [B] $LL_1^0 \geq LL_3^0, LL_1^* \geq LL_3^*, k_2 \geq k_1$
- [C] $LL_2^0 \geq LL_3^0, LL_3^* \geq LL_2^*, k_3 \geq k_2$
- [D] $LL_3^0 \geq LL_2^0, LL_3^* \geq LL_2^*, k_2 \geq k_3$

4. (P) Algoritmom GMM grupiramo primjere u dvodimenzijskome ulaznom prostoru. Skup podataka u stvarnosti je uzorkovan iz zajedničke distribucije koja se može opisati sljedećim mješavinskim modelom:

$$p(\mathbf{x}) = \sum_{j=1}^3 \frac{1}{3} \mathcal{N}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \quad \text{gdje } \boldsymbol{\mu}_1 = (5, 5), \boldsymbol{\mu}_2 = (5, 10), \boldsymbol{\mu}_3 = (-10, -10), \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_3 = 2\mathbf{I}$$

Skup \mathcal{D} grupiramo u $K = 2$ grupe. Pritom isprobavamo tri modela, koji se međusobno razlikuju po pretpostavkama na kovarijacijsku matricu. Konkretno: dijeljena i puna kovarijacijska matrica (\mathcal{H}_1), nedijeljena i diagonalna kovarijacijska matrica (\mathcal{H}_2) i nedijeljena i izotropna kovarijacijska matrica (\mathcal{H}_3). Neka je \mathcal{L}_i izglednost parametara dobivena modelom \mathcal{H}_i nakon konvergencije algoritma. Za inicijalizaciju središta koristi se algoritam K-means++. **Što su očekivani odnosi između izglednosti za ova tri modela?**

- [A] $\mathcal{L}_1 = \mathcal{L}_2 > \mathcal{L}_3$
- [B] $\mathcal{L}_1 > \mathcal{L}_2 > \mathcal{L}_3$
- [C] $\mathcal{L}_1 > \mathcal{L}_3, \mathcal{L}_2 > \mathcal{L}_3$
- [D] $\mathcal{L}_2 > \mathcal{L}_1, \mathcal{L}_2 > \mathcal{L}_3$

5. (N) Algoritmom hijerarhijskog aglomerativnog grupiranja (HAC) grupiramo $N = 5$ primjera. Za grupiranje koristimo mjeru sličnosti, koja je za naših pet primjera definirana sljedećom matricom (matrica je simetrična, pa je donji trokut izostavljen):

$$\begin{array}{ccccc} & \mathbf{x}^{(1)} & \mathbf{x}^{(2)} & \mathbf{x}^{(3)} & \mathbf{x}^{(4)} & \mathbf{x}^{(5)} \\ \mathbf{x}^{(1)} & & 1 & 0.4 & 0.5 & 0.7 & 0.5 \\ \mathbf{x}^{(2)} & & & 1 & 0.9 & 0.3 & 0.6 \\ \mathbf{x}^{(3)} & & & & 1 & 0.7 & 0.1 \\ \mathbf{x}^{(4)} & & & & & 1 & 0.8 \\ \mathbf{x}^{(5)} & & & & & & 1 \end{array}$$

Provedite grupiranje algoritmom HAC s potpunim povezivanjem te nacrtajte pripadni dendrogram. Primijetite da dendrogram odgovara binarnom stablu, s pojedinim primjerima u listovima. **Kojem binarnom stablu odgovara dobiveni dendrogram?**

- [A] $((\mathbf{x}^{(2)}, \mathbf{x}^{(3)}), \mathbf{x}^{(4)}), (\mathbf{x}^{(5)}, \mathbf{x}^{(1)})$
- [B] $((\mathbf{x}^{(2)}, \mathbf{x}^{(3)}), \mathbf{x}^{(1)}), (\mathbf{x}^{(4)}, \mathbf{x}^{(5)})$
- [C] $((\mathbf{x}^{(2)}, \mathbf{x}^{(3)}), ((\mathbf{x}^{(4)}, \mathbf{x}^{(5)}), \mathbf{x}^{(1)}))$
- [D] $((\mathbf{x}^{(2)}, \mathbf{x}^{(3)}), ((\mathbf{x}^{(4)}, \mathbf{x}^{(1)}), \mathbf{x}^{(5)}))$

6. (N) Algoritmom hijerarhijskog aglomerativnog grupiranja (HAC) grupiramo $N = 5$ primjera. Za grupiranje koristimo mjeru sličnosti, definiranu sljedećom matricom:

$$\begin{pmatrix} 1.0 & 0.2 & 0.8 & 0.1 & 0.4 \\ 0.2 & 1.0 & 0.9 & 0.3 & 0.7 \\ 0.8 & 0.9 & 1.0 & 0.6 & 0.5 \\ 0.1 & 0.3 & 0.6 & 1.0 & 0.4 \\ 0.4 & 0.7 & 0.5 & 0.4 & 1.0 \end{pmatrix}$$

Provode grupiranje algoritmom HAC s potpunim povezivanjem. Pritom u svakoj iteraciji bilježite na kojoj razini sličnosti se odvija stapanje dviju grupa. **Koliko iznosi zbroj po svim razinama sličnosti na kojima se odvija stapanje grupa?**

- A 1.8 B 1.9 C 2.0 D 2.4

7. (N) Algoritmom HAC grupiramo riječi engleskog jezika. Neoznačeni skup podataka sastoji se od sljedećih riječi:

$$\mathcal{D} = \{\text{"water"}, \text{"watering"}, \text{"earth"}, \text{"air"}\}$$

Kao mjeru sličnosti između primjera koristimo jezgrentu funkciju nad znakovnim nizovima, definiranu kao $\kappa(\mathbf{x}_1, \mathbf{x}_2) = |\mathbf{x}_1 \cap \mathbf{x}_2| / |\mathbf{x}_1 \cup \mathbf{x}_2|$, gdje su operacije unije i presjeka definirane nad skupovima slova od kojih se riječi sastoje. Npr., $\kappa(\text{"water"}, \text{"watering"}) = 5/8 = 0.625$. Provode prve dvije iteracije grupiranja algoritmom HAC uz prosječno povezivanje. **Na kojoj se razini sličnosti spajaju grupe u drugoj iteraciji algoritma HAC?**

- A 0.292 B 0.478 C 0.535 D 0.583

V20 - Grupiranje II

I Zadaci za učenje

- 1.1. a) Prednost algoritma EM primjenjivog na GMM
- probabilistički izaz (mehko grupiranj)

Nedostatak EM

- složeniji od K-means
- nema rješenje u zatvorenoj formi

b)

model mješane gustoće: $P(\vec{x}) = \sum_{k=1}^K p(\vec{x}, y=k)$

$$= \sum_{k=1}^K P(y=k) p(\vec{x}|y=k)$$

$$= \sum_{k=1}^K \pi_k p(\vec{x}|\vec{\theta}_k)$$

nepotpuna log-izglednost:

$$\ln L(\vec{\theta} | D) = \ln P(D|\vec{\theta}) = \ln \prod_{i=1}^N p(\vec{x}^i)$$

$$= \ln \left[\prod_{i=1}^N \sum_{k=1}^K \pi_k p(\vec{x}^i | \vec{\theta}_k) \right]$$

$$= \sum_i \ln \sum_k \pi_k p(\vec{x}^i | \vec{\theta}_k)$$

- c) model mješane gustoće s latentnim varijablama

$$P(\vec{x}, \vec{z}) = P(\vec{z}) P(\vec{x} | \vec{z}, \vec{\theta})$$

$$P(\vec{z}^i = z_k) = \prod_{k=1}^K \pi_k^{z_k}$$

$$P(\vec{x}, \vec{z}) = \prod_{k=1}^K \pi_k^{z_k} \prod_{k=1}^K p(\vec{x} | \vec{\theta}_k)^{z_k} = \prod_{k=1}^K \pi_k^{z_k} p(\vec{x} | \vec{\theta}_k)^{z_k}$$

potpuna log-izglednost:

$$\ln L(\vec{\theta} | D, z) = \ln P(D, z | \vec{\theta}) = \ln \prod_{i=1}^N p(\vec{x}^i, \vec{z}^i)$$

$$= \ln \prod_i \prod_k \pi_k^{z_k^i} p(\vec{x}^i | \vec{\theta}_k)^{z_k^i}$$

$$= \sum_{i=1}^N \sum_{k=1}^K z_k^i (\ln \pi_k + \ln p(\vec{x}^i | \vec{\theta}_k))$$

↳ ne možemo da li roditi izrazno s tem izgledom jer ne znamo vrijednosti latentnih varijabli

d) E-korak (program očuvanja potpuno log-izg.)

$$\begin{aligned} Q(\Theta | \Theta^t) &= \mathbb{E}_{z|D, \Theta^t} [C_L(\vec{\theta}|D, z)] \\ &= \mathbb{E} \left[\sum_{i=1}^N \sum_{k=1}^K z_e^i (\ln \pi_k + \ln p(\vec{x}^i | \vec{\theta}_k)) \right] \\ &= \sum_{i=1}^N \sum_{k=1}^K \mathbb{E}[z_e^i | \Theta^t, D] (\ln \pi_k + \ln p(\vec{x}^i | \vec{\theta}_k)) \end{aligned}$$

→ očuvanje potpune izgled: uz fiks. parametre $\vec{\theta}^t$

$$\left[\mathbb{E}[z_e^i | D, \vec{\theta}^t] = \mathbb{P}(z_e^i = 1 | \vec{x}^i, \Theta^t) = \frac{p(\vec{x}^i | \vec{\theta}_e^t)}{\sum_{j=1}^K p(\vec{x}^i | \vec{\theta}_j)} \pi_j^t = \frac{p_i}{h_e} \right]$$

↓
odgovornost!

M-korak (maksimiz. potpune izglednosti po Θ_e i $\vec{\theta}_e$)

$$\Theta^{t+1} = \underset{\Theta^t}{\operatorname{argmax}} Q(\Theta | \Theta^t)$$

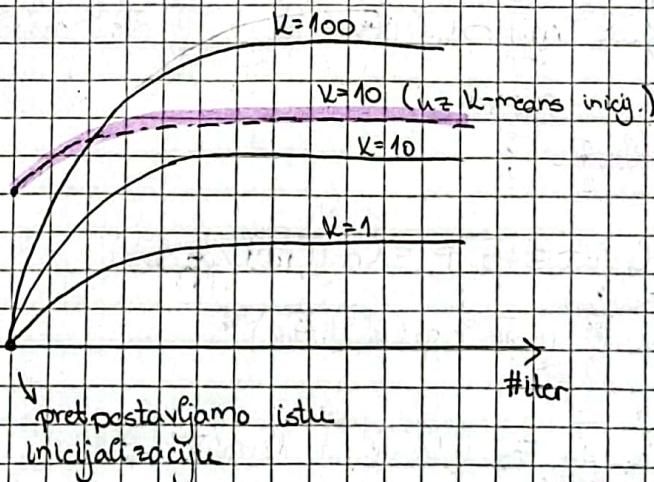
$$\nabla_{\Theta} Q(\Theta | \Theta^t) = 0$$

$$\Rightarrow \pi_e^{t+1} = \frac{1}{N} \sum_{i=1}^N p_e^i$$

$$\vec{\mu}_e^{t+1} = \frac{\sum p_e^i \vec{x}^i}{\sum p_e^i}$$

$$\vec{\Sigma}_e^{t+1} = \frac{\sum p_e^i (\vec{x}^i - \vec{\mu}_e^{t+1})(\vec{x}^i - \vec{\mu}_e^{t+1})^T}{\sum p_e^i}$$

e) $C_L(\vec{\theta}|D)$



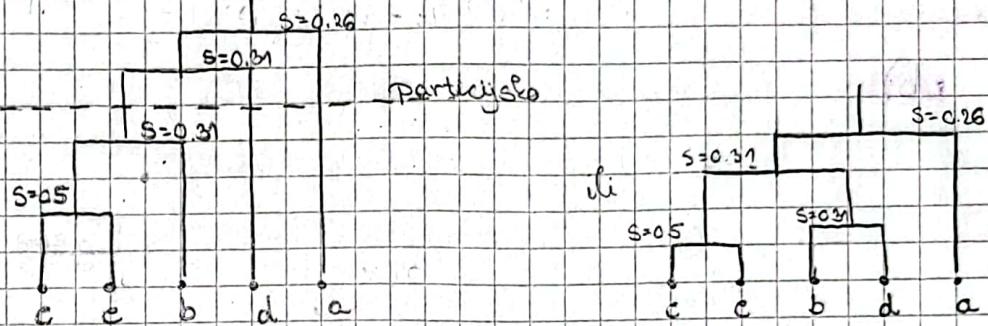
1.2.

HAC

	a	b	c	d	e
a	1	0.26	0.15	0.2	0.17
b	0.26	1	0.24	0.21	0.31
c	0.15	0.24	1	0.2	0.5
d	0.2	0.31	0.2	1	0.24
e	0.17	0.31	0.5	0.24	1

a) Dendrogram dobiven jednostrukim povezivanjem

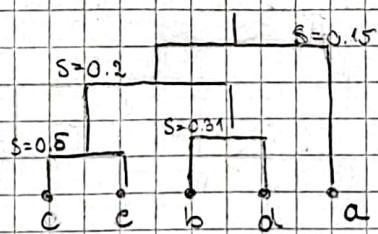
$$d_{\min}(G_i, G_j) = \min_{\vec{x} \in G_i, \vec{y} \in G_j} d(\vec{x}, \vec{y}) = \max S(\vec{x}, \vec{y})$$



b) Dendrogram dobiven polplunim povezivanjem

$$d_{\max}(G_i, G_j) = \max_{\vec{x} \in G_i, \vec{y} \in G_j} d(\vec{x}, \vec{y}) = \min S(\vec{x}, \vec{y})$$

↳ za grupe istovremeno spojaju se one koje su najblize!



1. grupa $\{c\}$ najbliza $\{e\}$

$$2. \quad \{c, e\} - b = 0.24$$

$$\{c, e\} - d = 0.2$$

$$\{c, e\} - a = 0.15$$

$$b - d = 0.31 \rightarrow \{b\} i \{d\} \text{ najblizi}$$

$$b - a = 0.26$$

$$d - a = 0.2$$

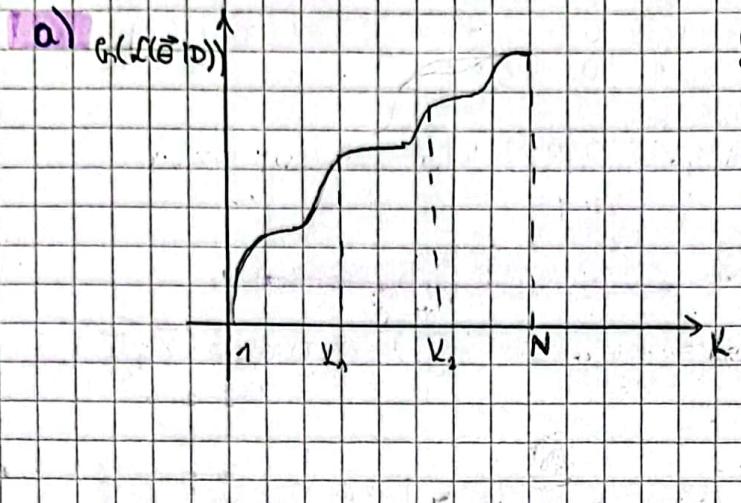
3. $\{c, e\} - \{b, d\} = 0.2 \rightarrow \{c, e\} i \{b, d\} \text{ najblizi}$

$$\{c, e\} - a = 0.15$$

$$\{b, d\} - a = 0.2$$

1.3.

$$| \text{Ch}(L(\vec{\theta} | D) \propto -J |$$



log-izglednost raste tako broj grupa K raste jer vrijedi veza $\text{Ch}(L(\vec{\theta} | D)) \propto -J$. Također raste složenost modela pa ima smisla da izglednost raste.

Temeđem ove privrede može se pokušati odrediti optimalan K metodom Cekta

→ kandidati sa slice K_1 i K_2
↳ ako vidimo stagnaciju dodatno dijelimo prirodne grupe

b) Akaikeov informacijski kriterij (AIC)

$$K^* = \underset{K}{\operatorname{argmin}} (-2\text{Ch}(L(K)) + 2g(K))$$

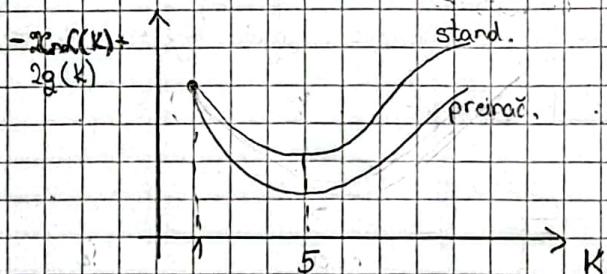
$g(K)$ = broj parametara modela s K grupa

$$K=5$$

standardni aig.
pretraženi - dijeljeno Σ

$$\begin{aligned} g(K) &= K \cdot \frac{n}{2}(n+1) + K-1 + nK = 5n + 4 + 5 \frac{n}{2}(n+1) \\ g(K) &= \frac{n}{2}(n+1) + K-1 + nK \\ &= \frac{n}{2}(n+1) + 5n + 4 \end{aligned}$$

ovaj se više kažnjava



II Zadaci s ispita

2.1.

$$N = 10$$

$$\mathcal{D} = \{(0,0), (1,1), (1,2), (2,2), (2,3), (5,0), (5,1), (6,0), (6,6), (7,7)\}$$

$$\begin{aligned} H_1: & K=2, \text{ puna } \Sigma \\ H_2: & K=2, \text{ izotropna } \Sigma \\ H_3: & K=3, \text{ izotropna } \Sigma \end{aligned} \quad \left. \begin{array}{l} \\ \\ \end{array} \right\} \text{ nedjelj.}$$

→ očekujemo da je odgovornost veća za manji broj klasa.

→ usporedjivi modeli

$$\cdot H_1 - H_2$$

→ H_1 bolje opisuje dati skup podataka $\Rightarrow h_{11} \geq h_{12}$
⇒ zbog transitivnosti $\Rightarrow h_{11} \geq h_{12} \geq h_{22}$

$$h_{11} > h_{12} > h_{22}$$

$$H_2 - H_3$$

⇒ zbog broja klasa $h_{12} \geq h_{13}$

$$h_{12} \geq h_{13}$$

(A)

$$H_1: K=50, \text{ inicij. K-means}$$

$$H_2: K=50, \text{ inicij. K-means i dij. } \Sigma$$

$$H_3: K=50, \text{ inicij. } \Sigma \text{ i dij. } \Sigma$$

$$H_4: K=10, \text{ inicij. K-means i dij. } \Sigma$$

K-means ubrzava konvergenciju

- usporedjivi modeli:

$$\cdot H_1 - H_2: LL_1^0 \geq LL_2^0$$

- očekujemo da će nedjeljena Σ biti pregeniti * parametre grupe
 $LL_1^0 \geq LL_2^0$

$$\cdot H_2 - H_3: LL_2^0 \geq LL_3^0$$

- očekujemo da će inicijalizacija K-meansom dati bolje rezultate
 $LL_2^0 \geq LL_3^0 \Rightarrow$ ovo bi moglo konv. u isto

$$\cdot H_2 - H_4: LL_2^0 \geq LL_4^0$$

$$LL_2^* \geq LL_4^*$$

veća klasa bolja izglednost

Zaključci

$$\begin{aligned} LL_2^0 &\geq LL_4^0 \\ LL_1^* &\geq LL_2^* \geq LL_3^* \end{aligned}$$

(A)

2.3.

H_1 : $K=25$, inicij. K-means++

H_2 : $K=50$, sruč inicij. i dij. Σ

H_3 : $K=50$, sred inicij. K-means, dij. Σ

k_x = broj iteracija (ekviv. koji broj konvergira)

- usporedivi modeli

$H_2 = H_3$

- očekujemo da K-means daje bolje rezultate i to u manjem broju iteracija

$$\left. \begin{array}{l} \mu_2 = \mu_3 \\ LL_3^o \geq LL_2^o \\ LL_3^* \geq LL_2^* \end{array} \right\} D$$

- model H_1 hipotetski je takođe možemo usporediti sa H_2 i H_3

2.4.

$$P(\vec{x}) = \sum_{j=1}^3 \frac{1}{3} p(\vec{x}_j | \mu_j, \Sigma_j)$$

$\Sigma_1 = \Sigma_2 = \Sigma_3 = 2I$ → matrice u stvarnosti isotropne i dijeljenje izokonture su kružnice

$$\vec{\mu}_1 = (5, 5)$$

$$\vec{\mu}_2 = (5, 10)$$

$$\vec{\mu}_3 = (-10, -10)$$

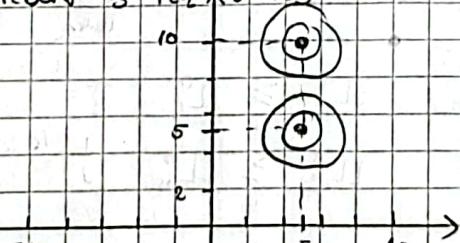
S obzirom na dane modele i dati skup možemo zaključiti:

$$\left| \begin{array}{l} L_2 > L_3 \\ L_2 > L_1 \end{array} \right\} D$$

$K=2$

H_1 - dij. i puna Σ
 H_2 - sruč. Σ
od H_3 } H_3 - nedij. i dija. Σ

→ u kontekstu složenosti H_1 ne
usporedivi se H_2 i H_3



• inicijalizacija K-means++

L_i = izgled. modela H_i

$$L(\vec{\theta}|D) = P(D|\vec{\theta})$$

$$= \prod_{i=1}^N p(\vec{x}_i | \vec{\theta})$$

$$= \prod_{i=1}^N \sum_{k=1}^K p(\vec{x}_i | \vec{\theta}_k)$$

H_1 : puna $\Sigma_1 = \Sigma_2$

→ omogućava zaklošene elipse u izokonturama

→ no pomaze baš s obzirom da izokonture moraju biti iste za obje grupe

H_2 : izokonture elipse i to različite za iste grupe

H_3 : izokonture kružnice, ali iste za obje grupe

2.5.

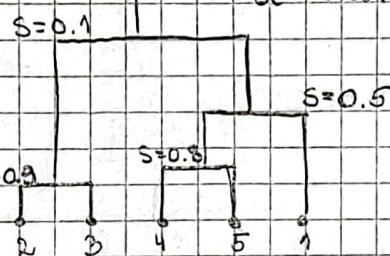
$N=5$

• sličnost

\vec{x}^1	\vec{x}^2	\vec{x}^3	\vec{x}^4	\vec{x}^5
1	0.4	0.5	0.7	0.5
0.4	1	0.9	0.3	0.6
0.5	0.9	1	0.7	0.1
0.7	0.3	0.7	1	0.8
0.5	0.6	0.1	0.8	1

• potpuno povezivanje

$$d = \max d(\vec{x}, \vec{y}) = \min S(\vec{x}, \vec{y})$$



(I) nejednacina $\vec{x}^2 \text{ i } \vec{x}^5$

$$\begin{aligned} \text{(II)} \quad & \{2, 3\} - 1 = 0.4 \\ & \{2, 3\} - 4 = 0.3 \\ & \{2, 3\} - 5 = 0.1 \\ & 1 - 4 = 0.7 \\ & 1 - 5 = 0.5 \\ & 4 - 5 = 0.8 \end{aligned}$$

(II)

$$\begin{aligned} \{2, 3\} - 1 &= 0.4 \\ \{2, 3\} - \{4, 5\} &= 0.1 \\ \{4, 5\} - 1 &= 0.5 \end{aligned}$$

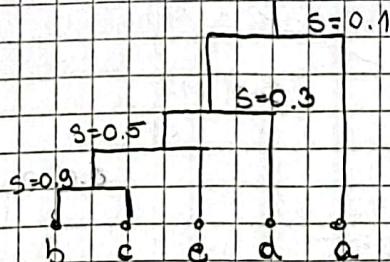
(C) $((\vec{x}^2, \vec{x}^3), ((\vec{x}^4, \vec{x}^5), \vec{x}^1))$

2.6.

$N=5$

	a	b	c	d	e
a	1	0.2	0.8	0.1	0.4
b	0.2	1	0.9	0.3	0.7
c	0.8	0.9	1	0.6	0.5
d	0.1	0.3	0.6	1	0.4
e	0.4	0.7	0.5	0.4	1

• potpuno poveziv.
 $d = \min S$



$$= 0.9 + 0.5 + 0.3 + 0.1$$

$$= 1.4 + 0.4 = 1.8$$

(A)

(I) $\{b\} \text{ i } \{c\}$

$$\begin{aligned} \{b, c\} - a &= 0.2 \\ -d &= 0.3 \\ (-e) &= 0.5 \end{aligned}$$

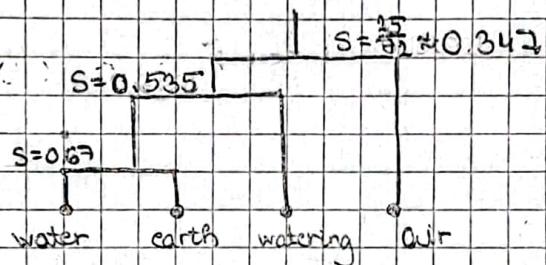
$$\begin{aligned} a - d &= 0.1 \\ -e &= 0.4 \\ d - e &= 0.4 \end{aligned}$$

$$\begin{aligned} \{b, c, e\} - a &= 0.2 \\ -d &= 0.3 \\ a - d &= 0.1 \end{aligned}$$

$$2.7. \quad K(\vec{x}_1, \vec{x}_2) = \frac{|\vec{x}_1 \cap \vec{x}_2|}{|\vec{x}_1 \cup \vec{x}_2|}$$

	water	watering	earth	air
water	1	$\frac{5}{8} = 0.625$	$\frac{4}{6} = 0.67$	$\frac{2}{6} = 0.33$
watering	$\frac{1}{8} = 0.125$	1	$\frac{4}{9} = 0.44$	$\frac{3}{8} = 0.375$
earth	$\frac{1}{8} = 0.125$	$\frac{1}{9} = 0.111$	1	$\frac{2}{6} = 0.33$
air	$\frac{1}{8} = 0.125$	$\frac{1}{8} = 0.125$	$\frac{1}{6} = 0.167$	1

• prosječno povezivanje $d(G_i, G_j) = \frac{1}{N_i \cdot N_j} \sum \sum d(\vec{x}^i, \vec{x}^j)$



(C) 0.535

(I) water, earth

(II)

$$\{w, e\} - \text{wing} = \frac{77}{144} \approx 0.535$$

$$\{w, e\} - a = \frac{3}{8} \approx 0.375$$

21. Vrednovanje modela

Strojno učenje 1, UNIZG FER, ak. god. 2023./2024.

Jan Šnajder, vježbe, v1.5

1 Zadatci za učenje

1. [Svrha: Izvježbati izračun mjera uspješnosti modela na konkretnom primjeru.]

Raspolažemo skupom od 11 ispitnih primjera koje želimo klasificirati u tri klase. Oznaka $y^{(i)}$ i izlaz modela $h(\mathbf{x}^{(i)})$ za svaki od 11 primjera su sljedeći:

$$\{(y^{(i)}, h(\mathbf{x}^{(i)}))\}_{i=1}^{11} = \{(1, 1), (0, 2), (2, 2), (1, 2), (1, 1), (0, 0), (1, 1), (2, 1), (0, 1), (2, 0), (2, 1)\}.$$

- (a) Izračunajte točnost klasifikatora.
(b) Izračunajte preciznost, odziv i mjeru F_1 , i to *mikro* i *makro* varijante.
2. [Svrha: Izvježbati izračun mjere F_1 na temelju parcijalno zadane matrice zabune.] Od $N = 1000$ primjera, klasifikator je za prvu, drugu i treću klasu ispravno pozitivno klasificirao njih 620, 146 odnosno 134. Od preostalih 100 neispravno klasificiranih primjera, 50 ih je klasificirano u drugu klasu umjesto u prvu, 20 u drugu umjesto u treću, a 30 u treću umjesto u drugu klasu. Izračunajte makro- F_1 .
3. [Svrha: Znati kako na temelju probabilističkog izlaza klasifikatora skicirati krivulju ROC. Znati da mjerom AUC možemo usporediti klasifikator s nasumičnim klasifikatorom. Znati kako pomoću krivulje ROC uspoređivati klasifikatore međusobno.] Na ispitnome skupu od $N = 10$ primjera evaluiramo tri binarna klasifikatora: logističku regresiju (h_{LR}), naivan Bayesov klasifikator (h_{NB}) i stroj potpornih vektora s probabilističkim izlazom dobivenim metodom Plattove kalibracije (h_{SVM}). Stvarne označke primjera $y^{(i)}$ i vjerojatnosne predikcije triju klasifikatora $h(\mathbf{x}^{(i)}) = p(y = 1 | \mathbf{x}^{(i)})$ na tom skupu su sljedeće:

i	1	2	3	4	5	6	7	8	9	10
$y^{(i)}$	1	1	0	0	1	1	1	0	0	1
$h_{LR}(\mathbf{x})$	0.8	0.6	0.8	0.6	0.8	0.8	0.8	0.2	0.2	0.2
$h_{NB}(\mathbf{x})$	0.3	0.8	0.3	0.5	0.8	0.3	0.8	0.5	0.3	0.5
$h_{SVM}(\mathbf{x})$	0.6	0.1	0.7	0.6	0.1	0.7	0.7	0.6	0.1	0.7

Na temelju ovog uzorka želimo procijeniti krivulju ROC te mjeru AUC (površinu ispod krivulje ROC). Prisjetite se da krivulja ROC opisuje TPR (odziv) kao funkciju od FPR (stopa lažnog alarmu).

- (a) Skicirajte krivulje ROC za ova tri klasifikatora, linearno interpolirajući između točaka dobivenih na temelju gornjeg uzorka.
(b) Izračunajte mjeru AUC za sva tri klasifikatora.
(c) Kako izgleda krivulja ROC za nasumični klasifikator. Zašto?
(d) Koji je od navedenih klasifikatora lošiji od nasumičnog klasifikatora, a koji biste klasifikator odabrali kao najbolji?

4. [Svrha: Razumjeti na koji se način provodi ugniježđena unakrsna provjera, kako se razdjeljuju primjeri kroz iteracije petlji te kako ugraditi dodatne predobradbe značajki, a pritom ne kompromitirati podjelu na skup za učenje i skup za ispitivanje.] Raspolažemo sa 1000 označenih primjera. Za vrednovanje SVM-a s hiperparametrima C i γ koristimo ugniježđenu unakrsnu provjeru sa po 5 ponavljanja u obje petlje. Hiperparametre optimiramo rešetkastim pretraživanjem u rasponima $C \in \{2^{-5}, 2^{-4}, \dots, 2^{15}\}$ i $\gamma \in \{2^{-15}, 2^{-14}, \dots, 2^3\}$.

- (a) Koliko ćemo ukupno puta trenirati model?
- (b) Koliko ćemo primjera u svakoj od iteracija koristiti za treniranje, koliko za provjeru, a koliko za ispitivanje?
- (c) Kako glase odgovori na prethodna dva pitanja, ako bismo u vanjskoj petljici umjesto petrostrukih unakrsnih provjera koristili unakrsnu provjeru *izdvoji jednoga* (engl. *leave one out, LOOCV*)?
- (d) Klasifikator SVM posebno je osjetljiv na razlike u rasponima između značajki (zašto?), pa se preporuča standardizirati značajke. Što to točno znači i kako biste standardizaciju značajki ugradili u ugniježđenu unakrsnu provjeru?
- (e) Gdje biste u ugniježđenu unakrsnu provjeru ugradili odabir značajki modela i optimizaciju praga po mjeri AUC?

2 Zadatci s ispita

1. (N) Na ispitnome skupu evaluiramo klasifikator sa $K = 3$ klase. Dobili smo sljedeću matricu zabune (stupci su stvarne oznake, a retci oznake koje daje klasifikator):

$$\begin{array}{ccc} & y = 1 & y = 2 & y = 3 \\ y = 1 & 15 & 3 & 1 \\ y = 2 & 6 & 5 & 4 \\ y = 3 & 4 & 2 & 23 \end{array}$$

Izračunajte mikro-F1 (F_1^μ) i makro-F1 (F_1^M) mjere na ovoj matrici zabune. **Koliko iznosi razlika između vrijednosti mikro-F1 i makro-F1 mjere, $F_1^\mu - F_1^M$?**

- A 0.01 B 0.05 C 0.09 D 0.13

2. (N) Na ispitnome skupu evaluiramo model multinomijalne logističke regresije (MLR) za klasifikaciju u $K = 3$ klase. Dobili smo sljedeću matricu zabune (stupci su stvarne oznake, a retci oznake koje daje klasifikator):

$$\begin{array}{ccc} & y = 1 & y = 2 & y = 3 \\ y = 1 & 30 & 18 & 3 \\ y = 2 & 11 & 25 & 2 \\ y = 3 & 4 & 2 & 5 \end{array}$$

Klasifikator MLR uspoređujemo s klasifikatorom RAND koji primjere klasificira nasumično, i to tako da oznaku $y = j$ dodjeljuje s vjerojatnošću proporcionalnoj udjelu klase j u ispitnome skupu. Izračunajte mikro- F_1 za klasifikator MLR i očekivani mikro- F_1 za klasifikator RAND. **Koliko iznosi očekivana razlika u vrijednostima mikro- F_1 klasifikatora MLR i RAND?**

- A 0.085 B 0.155 C 0.185 D 0.205

3. (N) Logističku regresiju vrednjujemo na ispitnome skupu od $N = 10$ primjera. Stvarne oznake primjera $y^{(i)}$ i vjerojatnosne predikcije klasifikatora $h(\mathbf{x}^{(i)}) = p(y = 1 | \mathbf{x}^{(i)})$ na tom skupu su sljedeće:

$$\{(y^{(i)}, h(\mathbf{x}^{(i)}))\}_{i=1}^{10} = \{(1, 0.8), (0, 0.2), (0, 0.6), (0, 0.6), (1, 0.8), (0, 0.8), (1, 0.6), (1, 0.2), (0, 0.6), (1, 0.8)\}$$

Na temelju ovog uzorka želimo procijeniti mjeru AUC (površinu ispod krivulje ROC). Prisjetite se da krivulja ROC opisuje TPR (odziv) kao funkciju od FPR (stopa lažnog alarma). Skicirajte krivulju ROC, linearno interpolirajući između točaka dobivenih na temelju gornjeg uzorka. **Koliko je ovaj klasifikator prema mjeri AUC bolji od nasumičnog klasifikatora?**

- A 0 B 0.16 C 0.24 D 0.35

4. (P) Na istom skupu označenih primjera vrednujemo četiri binarna klasifikatora s vjerojatnosnim izlazima. Za vrednovanje koristimo krivulju ROC. Svaki smo klasifikator ispitali s tri vrijednosti klasifikacijskog praga te smo za te vrijednosti izračunali FPR (stopa lažnog alarma) i TPR (odziv). Dobiveni parovi vrijednosti (FPR, TPR) za sva četiri klasifikatora su sljedeći:

$$\begin{aligned} h_1 &: (0.4, 0.2), (0.6, 0.2), (0.9, 0.4) \\ h_2 &: (0.3, 0.1), (0.5, 0.4), (0.8, 0.6) \end{aligned}$$

$$\begin{aligned} h_3 &: (0.1, 0.5), (0.6, 0.6), (0.7, 0.8) \\ h_4 &: (0.3, 0.8), (0.4, 0.9), (0.6, 1.0) \end{aligned}$$

Skicirajte odgovarajuće krivulje ROC, linearno interpolirajući između izmjerjenih točaka, a dodajte i točke koje odgovaraju krajnjim vrijednostima klasifikacijskog praga (0 i 1). Pritom, ako je neki klasifikator lošiji od nasumičnog klasifikatora, umjesto tog klasifikatora razmatrajte njegovu pravljenu varijantu koju ćete dobiti invertiranjem izlaza ($1 - h(\mathbf{x})$ umjesto $h(\mathbf{x})$). **Koji su od ispitanih klasifikatora (eventualno nakon popravka) najbolji prema krivulji ROC?**

- A h_3 i h_4 B h_1 i h_4 C h_2 i h_3 D h_1 i h_3

5. (P) Krivuljom ROC vrednujemo binarne klasifikatore. Neka $h_1 > h_2$ označava da naučeni model h_1 striktno dominira nad naučenim modelom h_2 prema krivulji ROC, tj. da vrijedi

$$\forall \theta. \left((\text{FPR}_\theta(h_1) = \text{FPR}_\theta(h_2)) \Rightarrow (\text{TPR}_\theta(h_1) > \text{TPR}_\theta(h_2)) \right)$$

gdje su $\text{FPR}_\theta(h)$ i $\text{TPR}_\theta(h)$ vrijednosti stope lažnih pozitiva odnosno stvarnih pozitiva hipoteze h s pragom θ . Nadalje, neka $h_1 > h_2$ označava da vrijedi $\text{AUC}(h_1) > \text{AUC}(h_2)$. **Što od sljedećeg vrijedi?**

- A $h_1 > h_2 \Rightarrow h_1 > h_2$
 B $h_1 > h_2 \Rightarrow h_2 > h_1$
 C $h_1 > h_2 \Rightarrow h_1 > h_2$
 D $h_1 > h_2 \Rightarrow h_2 > h_1$

6. (P) Raspolažemo sa 1000 označenih primjera. Na tom skupu treniramo i evaluiramo algoritam SVM. Pritom razmatramo tri hiperparametra: jezgra (linearna ili RBF), regularizacijski faktor C i preciznost RBF jezgre γ . Posljednja dva hiperparametra optimiramo rešetkastim pretraživanjem u rasponima $C \in \{2^{-15}, 2^{-14}, \dots, 2^{15}\}$ i $\gamma \in \{2^{-15}, 2^{-14}, \dots, 2^{15}\}$. Naravno, ako ne koristimo RBF-jezgru, onda hiperparametar γ ne optimiramo. Za treniranje i evaluaciju modela koristimo ugniježđenu unakrsnu provjeru s 10 ponavljanja u vanjskoj petlji i 5 ponavljanja u unutarnjoj petlji. **Koliko će puta svaki primjer biti iskorišten za treniranje modela?**

- A 35721 B 44640 C 49600 D 69201

7. (P) Raspolažemo sa 1000 označenih primjera. Na tom skupu treniramo i vrednujemo algoritam SVM, optimizirajući hiperparametre C i γ . Za vrednovanje koristimo ugniježđenu unakrsnu provjeru sa 5 preklopa u vanjskoj petlji i 5 preklopa u unutarnjoj petlji. To znači da za svaku kombinaciju vrijednosti hiperparametara C i γ treniramo pet modela. Ti modeli trenirani su na skupovima za učenje koji nisu disjunktni: svaki par treniranih modela dijele određeni broj primjera za učenje. Izračunajte koliko primjera za učenje dijele svaki par modela koje treniramo u unutarnjoj petlji ugniježđene unakrsne provjere. **Za koliko bi taj broj narastao kada bismo broj preklopa u unutarnjoj petlji povećali na 10?**

- A 80 B 100 C 160 D 192

8. (P) Evaluiramo model L_2 -regularizirane logističke regresije. Za evaluaciju koristimo ugniježđenu unakrsnu provjeru u kojoj optimiramo regularizacijski faktor λ . Neka je λ_1 prosjek optimalnih vrijednosti regularizacijskog faktora, i neka je F_1^1 prosječna F_1 -mjera na ispitnom skupu vanjske petlje. Međutim, naknadno smo ustanovili da nam se potkrala pogreška i da smo u unutarnjoj petlji model uvijek ispitivali na prvom preklopu. Kada to ispravimo, dobivamo λ_2 kao prosjek optimalnih vrijednosti regularizacijskog faktora i F_1^2 kao prosjek F_1 -mjere na ispitnom skupu vanjske petlje. Nažalost, kasnije smo ustanovili da nam se potkrala još jedna pogreška: umjesto da u vanjskoj petlji optimalan model treniramo na cijelom skupu za treniranje, mi smo ga trenirali samo na skupu za

treniranje zadnje iteracije unutarnje petlje. Kada i tu pogrešku ispravimo, dobivamo λ_3 odnosno F_1^3 . **Što možemo očekivati o odnosima između procjena za optimalni λ i za F_1 -mjeru na ispitnom skupu?**

- [A] $\lambda_1 > \lambda_2 > \lambda_3$, $F_1^1 < F_1^2$, $F_1^3 < F_1^2$
- [B] $\lambda_1 < \lambda_3$, $F_1^1 < F_1^2 < F_1^3$
- [C] $\lambda_1 < \lambda_2 = \lambda_3$, $F_1^1 > F_1^2$, $F_1^3 > F_1^2$
- [D] $\lambda_1 = \lambda_3 < \lambda_2$, $F_1^2 < F_1^1$, $F_1^3 < F_1^2$

V21 - Vrednovanje modela

I Podaci za učenje

1.1. $N = 11$
 $K = 3$

y^i	1	0	2	1	1	0	1	2	0	2	2
$f(\vec{x}^i)$	1	2	2	2	1	0	1	1	1	0	1

a) $\text{Acc} = ?$

$$\text{Acc} = \frac{1}{N} \sum_{i=1}^N \mathbb{1}[y^i = f(\vec{x}^i)] = \frac{1}{11} \cdot 5 = \frac{5}{11} \approx 0.455$$

b)

pred	true		
	0	1	2
0	1	1	1
1	1	3	2
2	1	1	1

pred	true	
	$y = 0$	$y \neq 0$
$y = 0$	1	1
$y \neq 0$	2	7

pred	true	
	$y = 1$	$y \neq 1$
$y = 1$	1	1
$y \neq 1$	1	4

pred	true	
	$y = 2$	$y \neq 2$
$y = 2$	1	2
$y \neq 2$	3	5

Mikro procene

$$\sum TP = 5$$

$$\sum FP = 6$$

$$\sum FN = 6$$

$$\sum TN = 16$$

$$P^\mu = \frac{TP}{TP+FP} = \frac{5}{11}$$

$$Q^\mu = \frac{TP}{TP+FN} = \frac{5}{11}$$

$$F^\mu = \frac{2PQ}{P+Q} = \frac{5}{11}$$

Makro procene

$$P_1 = \frac{1}{2}, \quad Q_1 = \frac{1}{3}, \quad F_{1,1} = \frac{2}{5}$$

$$P_2 = \frac{1}{2}, \quad Q_2 = \frac{2}{4}, \quad F_{1,2} = \frac{2}{5}$$

$$P_3 = \frac{1}{3}, \quad Q_3 = \frac{1}{4}, \quad F_{1,3} = \frac{2}{7}$$

$$P^M = \frac{1}{3} \sum P_i = \frac{1}{3} = 0.44$$

$$Q^M = \frac{1}{3} \sum Q_i = \frac{1}{3} = 0.44$$

$$F^M = \frac{1}{3} \sum F_{1,i} = \frac{2}{7} \approx 0.429$$

1.2.

$$N = 1000$$

$$TP_1 = 620$$

$$TP_2 = 146$$

$$TP_3 = 134$$

pred, y_i

$$FN_{2,1} = 50$$

$$FN_{2,3} = 20$$

$$FN_{3,2} = 30$$

		true		
		1	2	3
pred	1	620	0	0
	2	50	146	20
	3	0	30	134

• dekompozicija

$$\begin{array}{l} y=1 \quad y \neq 1 \\ \text{pred} \left\{ \begin{array}{l} y=1 \quad \begin{bmatrix} 620 & 0 \\ 50 & 330 \end{bmatrix} \\ y \neq 1 \quad \begin{bmatrix} 0 & 146 \\ 30 & 754 \end{bmatrix} \end{array} \right. \end{array}$$

$$\begin{array}{l} y \neq 2 \quad y \in 2 \\ \text{pred} \quad \begin{bmatrix} 146 & 70 \\ 30 & 754 \end{bmatrix} \end{array}$$

$$\begin{array}{l} y_{\text{true}} = 3 \quad y_{\text{true}} \neq 3 \\ \text{pred} \quad \begin{bmatrix} 134 & 30 \\ 20 & 816 \end{bmatrix} \end{array}$$

$$P_1 = \frac{TP}{TP+FP} = 1$$

$$P_2 = \frac{146}{216}$$

$$P_3 = \frac{134}{164}$$

$$R_1 = \frac{TP}{TP+FN} = \frac{620}{670}$$

$$R_2 = \frac{146}{176}$$

$$R_3 = \frac{134}{154}$$

$$F_{1,1} = \frac{2P_1R_1}{P_1+R_1} = \frac{124}{129}$$

$$F_{1,2} = \frac{73}{98}$$

$$F_{1,3} = \frac{134}{159}$$

$$F_1^M = \frac{1}{3} \sum_{i=1}^3 F_{1,i} = 0.8496$$

1.3.

$$N = 10$$

- 3 binarna klasifikatora

 P_{LR} P_{NB} P_{SVM}

y_i	1	2	3	4	5	6	7	8	9	10
$P_{\text{LR}}(\vec{x}_i)$	0.8	0.6	0.8	0.6	0.8	0.8	0.8	0.2	0.2	0.2
$P_{\text{NB}}(\vec{x}_i)$	0.3	0.8	0.3	0.5	0.8	0.3	0.8	0.5	0.3	0.5
$P_{\text{SVM}}(\vec{x}_i)$	0.6	0.1	0.7	0.6	0.1	0.7	0.7	0.6	0.1	0.7

a)

$$TPD_0 = \frac{TP}{TP+FN}$$

$$FPD_0 = \frac{FP}{FP+TN}$$

• za različite programe treba provesti klasifikaciju i izračunati
prirodne TPD_0 i FPR

=> rodimo programski - python

- izračun čemo provesti za progove: 1, 0.8, 0.6, 0.4, 0.2, 0

$$\text{pred} \begin{bmatrix} 1 & \frac{1}{TP} & \frac{FP}{TP+FN} \\ 0 & \frac{FN}{TP+FN} & \frac{TN}{TP+TN} \end{bmatrix}$$

Logistička regresija

i.	1	2	3	4	5	6	7	8	9	10
$\text{Rho}(\vec{x}^i)$	0.8	0.6	0.8	0.6	0.8	0.8	0.8	0.2	0.2	0.2
y_i^i	1	1	0	0	1	1	1	0	0	1
1	0	0	0	0	0	0	0	0	0	0
0.8	1	0	1	0	1	1	1	0	0	0
0.6	1	1	1	1	1	1	1	0	0	0
0.4	1	1	1	1	1	1	1	0	0	0
0.2	1	1	1	1	1	1	1	1	1	1
0	1	1	1	1	1	1	1	1	1	1

Progovi 0.6 i 0.4 te progovi 0.2 i 0 rezultiraju istim klasifikacijama pa te vrijednosti TPR i FPR biti jednakie za korespondentne progove.

$$FPR = \frac{FP}{FP+TN}$$

$$TPR = \frac{TP}{TP+FN}$$

prog

$$\begin{aligned} 1 & \quad \frac{0}{0+4} = 0 \\ 0.8 & \quad \frac{1}{1+3} = \frac{1}{4} \\ 0.6 & \quad \frac{2}{2+2} = \frac{1}{2} \\ 0.2 & \quad \frac{1}{4} = 0.25 \end{aligned}$$

$$\begin{aligned} 1 & \quad \frac{0}{0+6} = 0 \\ 0.8 & \quad \frac{4}{4+2} = \frac{4}{6} = \frac{2}{3} \\ 0.6 & \quad \frac{5}{5+1} = \frac{5}{6} \\ 0.2 & \quad \frac{6}{6+0} = 1 \end{aligned}$$

Nainan Bayesov klasifikator

i.	1	2	3	4	5	6	7	8	9	10
$\text{Rho}(\vec{x}^i)$	0.3	0.8	0.3	0.5	0.8	0.3	0.8	0.5	0.3	0.5
y_i^i	1	1	0	0	1	1	1	0	0	1
1	0	0	0	0	0	0	0	0	0	0
0.8	0	1	0	0	1	0	1	0	0	0
0.6	0	1	0	0	1	0	1	0	0	0
0.4	0	1	0	1	1	0	1	1	0	1
0.2	1	1	1	1	1	1	1	1	1	1
0	1	1	1	1	1	1	1	1	1	1

$$FPR = \frac{FP}{FP+TN}$$

$$TPR = \frac{TP}{TP+FN}$$

prog

$$\begin{aligned} 1 & \quad \frac{0}{0+4} = 0 \\ 0.8 & \quad \frac{0}{0+4} = 0 \\ 0.6 & \quad \frac{2}{2+2} = \frac{1}{2} \\ 0.4 & \quad \frac{4}{4+0} = 1 \end{aligned}$$

$$\begin{aligned} 1 & \quad \frac{0}{0+6} = 0 \\ 0.8 & \quad \frac{3}{3+3} = 0.5 \\ 0.6 & \quad \frac{4}{4+2} = \frac{2}{3} \\ 0.4 & \quad \frac{6}{6+0} = 1 \end{aligned}$$

SVM

i	1	2	3	4	5	6	7	8	9	10
$h_{SVM}(\vec{x}_i)$	0.6	0.1	0.7	0.6	0.1	0.7	0.7	0.0	0.1	0.7
y_i	1	1	0	0	1	1	1	0	0	1
lista ccf	0	0	0	0	0	0	0	0	0	0
0.8	0	0	0	0	0	0	0	0	0	0
lista ccf	0.6	1	0	1	1	0	1	1	0	1
0.4	1	0	1	1	0	1	1	1	0	1
0.2	1	0	1	1	0	1	1	1	0	1
0	1	1	1	1	1	1	1	1	1	1

prag

$$FPR = \frac{FP}{FP + TN}$$

$$\frac{0}{0+4} = 0$$

$$\frac{3}{3+1} = \frac{3}{4}$$

$$\frac{4}{4+0} = 1$$

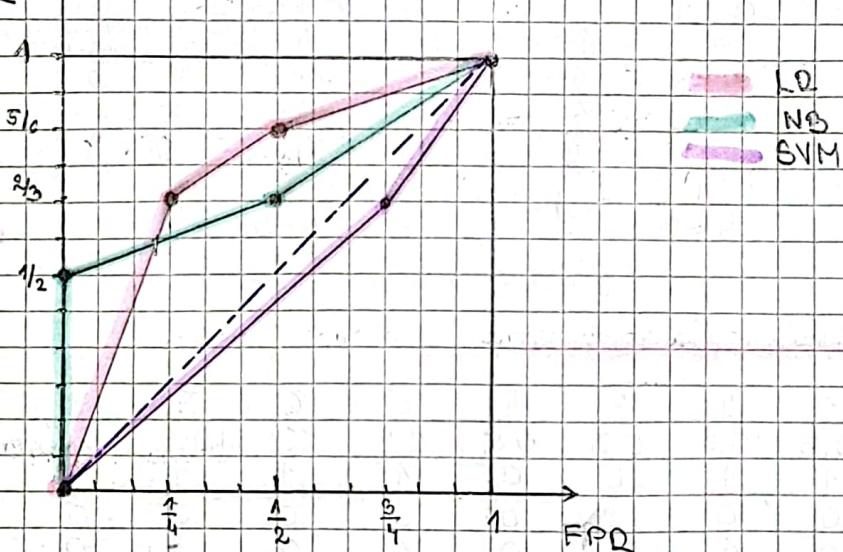
$$TPR = \frac{TP}{TP + FN}$$

$$\frac{0}{0+6} = 0$$

$$\frac{4}{4+2} = \frac{2}{3}$$

$$\frac{6}{6+0} = 1$$

TPR



b) $AUC = \text{računamo } \text{broj površinu ispod ROC krivulje} \Rightarrow \text{zbroj površina poligona}$

$$ROC_{LDA} = \frac{2}{3} \cdot \frac{1}{4} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{\frac{2}{3} + \frac{5}{6}}{2} + \frac{\frac{5}{6} + 1}{2} \cdot \frac{1}{2} = \frac{35}{48} \approx 0.729$$

$$ROC_{NB} = \frac{1}{2} \cdot \frac{\frac{1}{2} + \frac{2}{3}}{2} + \frac{1}{2} \cdot \frac{\frac{2}{3} + 1}{2} = \frac{17}{24} \approx 0.708$$

$$ROC_{SVM} = \frac{3}{4} \cdot \frac{2}{3} \cdot \frac{1}{2} + \frac{1}{4} \cdot \frac{\frac{2}{3} + 1}{2} = \frac{11}{24} \approx 0.458$$

c) ROC krivulja točke $(0,0)$ do $(1,1)$. Nasumični klasifikator slučajno odabire 2% primjera i proglašava ih pozitivnim tj. $TPR = \frac{TP}{TP+FN} = 2\%$. od svih α

Također nasumični klasifikator slučajno odabire 2% primjera od ukupne brojevi negativnih primjera i proglašava ih negativnim, tj. $FPR = \frac{FP}{FP+TN} = 2\%$.

d) S obzirom na dobivene ROC krivulje pošji klasifikator od nasumičnog jest SVM.

Odabir najboljeg klasifikatora ovisi o razini FPR. Ako želimo FPR u intervalu od 0 do a gdje je a nešto manji od $1/4$ najbolji klasifikator je ravnans Bayesov klasifikator, a inače Logistička regresija.

1.4.

$N = 1000$ primjera
ugrijevštena unutarnja projekcija 5×5

$$C \in \{2^{-5}, 2^{-4}, \dots, 2^{15}\}$$

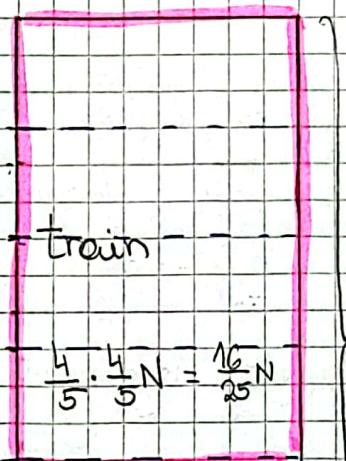
$$|C| = 15 + 5 + 1 = 21$$

$$\gamma \in \{2^{-15}, 2^{-14}, \dots, 2^5\}$$

$$|\gamma| = 15 + 3 + 1 = 19$$

Ukupno $21 \cdot 19 = 399$ kombinacija hiperparametara

a) Koliko puta ćemo ukupno trenirati model?



k = vanjski preklop
 e = unutarnji preklop

izmjena vanjskih preklopa

$$k \cdot (\underbrace{e \cdot |C| \cdot |\gamma|}_{\text{optimizacija hiperparam.}} + 1)$$

model za optim. hiperparametre

Ukupni broj trniranja

$$k \cdot (e \cdot |C| \cdot |\gamma| + 1)$$

$$= 5 \cdot (5 \cdot 21 \cdot 19 + 1)$$

$$= 9980$$

test $\cdot \frac{1}{5}N$

stvaraju za $k=1$

b) Koliko ćemo primjera koristiti u svakoj iteraciji, koristiti za treniranje, soliter za provjeru, a koliko za ispitivanje?

$$\begin{aligned}
 N &= 1000 \\
 \frac{4}{5}N &= 800 \\
 \text{train: } \frac{1}{5}N &= \frac{1}{5}1000 = 200 \\
 \text{val: } \frac{1}{5}\frac{4}{5}N &= \frac{1}{5}800 = 160 \\
 \text{train: } \frac{4}{5}\frac{4}{5}N &= 640
 \end{aligned}$$

Za treniranje koristimo 640 primjera, za validaciju 160 primjera, a za ispitivanje 200 primjera.

c) U vanjskoj petlji koristimo 1000, tj. $k = N$

$$\begin{aligned}
 \text{Ukupan broj treniranja: } & \frac{1}{5}(C \cdot |C| \cdot |V| + 1) \\
 &= 1000 \cdot 5 \cdot 21 \cdot 19 + 1000 \\
 &= 1.930.000
 \end{aligned}$$

U svakoj iteraciji

$$\begin{aligned}
 N(\text{test}) &= 1 \\
 N(\text{val}) &= \frac{1}{5} \cdot 930 = 200 \\
 N(\text{train}) &= \frac{4}{5} \cdot 930 = 730
 \end{aligned}$$

d) SVM

- standarnizacija značajki podrazumijeva skidanje značajki na interval $[0, 1]$ na slijedeći način $\vec{x} = \frac{\vec{x} - \mu_x}{\sigma_x}$

- SVM je osjetljiv na skale jer promatra euklidsku udaljenost između primjera pa značajke većin skala više utječu na udaljenost

- u postupku ugniježdene unakrsne provjere standardizacija bi se mogla ugraditi neposredno prije treniranja / validacije / testiranja modela (bez cjevovoda)

e)

U postupku ugniježdene unakrsne provjere bi postupak

- odabira značajku ugradio prije vanjsku petlju, tj. prije podjele na vanjsku preklopne
- optimizacije praga po mjeri AUC ugradio u unutarnju petlju, tj. u optimizaciju hiperparametara

II Tabaci s ispitom

2.1.

$K=3$

$$\text{true} \quad \begin{matrix} 1 & 2 & 3 \end{matrix}$$

pred	1	15	3	1
	2	6	5	4
	3	4	2	23

$$\begin{matrix} \text{true} \\ \downarrow \\ \begin{cases} i \\ \neq i \end{cases} \end{matrix} \quad \begin{bmatrix} \text{TP} & \text{FP} \\ \text{FN} & \text{TN} \end{bmatrix}$$

$$F_1^M - F_1^U = ?$$

$$F_1^M = \text{Acc} = \frac{1}{N} \sum_{i=1}^N \mathbb{1}\{y_{\text{pred}}^i = y_{\text{true}}^i\}$$

$$= \frac{1}{63} (15+5+23) = \frac{43}{63}$$

$$F_1 = \frac{2PR}{P+R}$$

Dekompozicija po klasama

$y=1$

15	4
10	29

$$P_1 = \frac{15}{19}, R_1 = \frac{15}{25}$$

$$F_1^1 = \frac{15}{22}$$

$y=2$

5	10
5	10

$$P_2 = \frac{5}{15}, R_2 = \frac{5}{10}$$

$$F_1^2 = \frac{2}{5}$$

$y=3$

23	6
5	28

$$P_3 = \frac{23}{28}, R_3 = \frac{23}{28}$$

$$F_1^3 = \frac{46}{57}$$

$$P = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$R = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$F_1^M = \frac{1}{K} \sum_{k=1}^K F_1^k \approx 0.6286$$

$$F_1^M - F_1^U = 0.0529 \Rightarrow \textcircled{B}$$

2.2.

MLQ

$K=3$

true

pred	1	2	3
	30	18	3
	11	25	2

	\downarrow	\downarrow	\downarrow
	45	45	10
\rightarrow	0.45	0.45	0.1

$$F_1^M (\text{MLQ}) = \text{Acc} = \frac{1}{N} \sum_{i=1}^N \mathbb{1}\{y_{\text{pred}}^i = y_{\text{true}}^i\} = \frac{3}{5}$$

RAND

	1	2	3
1	$0.45 \cdot 45$	$0.45 \cdot 45$	$0.45 \cdot 10$
2	$0.45 \cdot 45$	$0.45 \cdot 45$	$0.45 \cdot 10$
3	$0.1 \cdot 45$	$0.1 \cdot 45$	$0.1 \cdot 10$

$$= \begin{bmatrix} 20.25 & 20.25 & 4.5 \\ 20.25 & 20.25 & 4.5 \\ 4.5 & 4.5 & 1 \end{bmatrix}$$

$$F_1^M (\text{RAND}) = \frac{1}{100} (20.25 + 20.25 + 4) = \frac{83}{200}$$

$$F_1^M (\text{MLQ}) - F_1^M (\text{RAND}) = \frac{37}{200} = 0.185 \Rightarrow \textcircled{C}$$

2.3

i	1	2	3	4	5	6	7	8	9	10
$h(x^i)$	0.8	0.2	0.6	0.6	0.8	0.8	0.6	0.2	0.6	0.8
y_i	1	0	0	0	1	0	1	1	0	1
0.8	1	0	0	0	1	1	0	0	0	1
0.6	1	0	1	1	1	1	1	0	1	1
0.2	1	1	1	1	1	1	1	1	1	1

$$\begin{bmatrix} TP & FP \\ FN & TN \end{bmatrix}$$

$$TPR = \frac{TP}{TP+FN}$$

$$FPR = \frac{FP}{FP+TN}$$

$$P=0.8 \quad \begin{bmatrix} 3 & 1 \\ 2 & 4 \end{bmatrix}$$

$$P=0.6 \quad \begin{bmatrix} 4 & 4 \\ 1 & 1 \end{bmatrix}$$

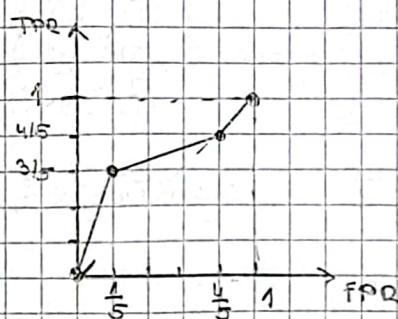
$$P=0.2 \quad \begin{bmatrix} 5 & 5 \\ 0 & 0 \end{bmatrix}$$

$$\text{prog} \quad PPR \quad TPR$$

$$0.8 \quad \frac{1}{5} \quad \frac{3}{5}$$

$$0.6 \quad \frac{4}{5} \quad \frac{4}{5}$$

$$0.2 \quad -1 \quad 1$$



$$AUC(\text{RAND}) = 0.5$$

$$AUC(LP) = \frac{1}{2} \cdot \frac{1}{5} \cdot \frac{3}{5} + \frac{2}{5} \cdot \frac{\frac{5}{5} + \frac{4}{5}}{2} + \frac{1}{5} \cdot \frac{1 + \frac{4}{5}}{2} = \frac{32}{50} = 0.66$$

$$AUC(LP) - AUC(\text{RAND}) = 0.16$$

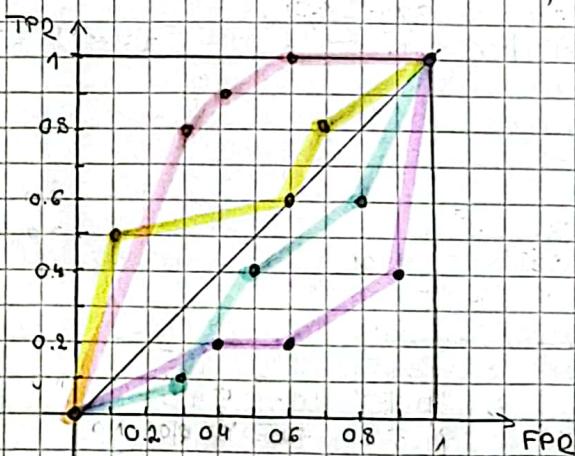
(15)

(FPR, TPR)

2.4

- $h_1: (0.4, 0.2), (0.6, 0.2), (0.9, 0.4)$
- $h_2: (0.3, 0.1), (0.5, 0.4), (0.8, 0.6)$
- $h_3: (0.1, 0.5), (0.6, 0.6), (0.7, 0.8)$
- $h_4: (0.3, 0.8), (0.4, 0.8), (0.6, 1)$

• binarni klasiifikatori $K=2$



• za h_1 i h_2 trebamo razmatriti klasi.

$$1-h_1 \cup 1-h_2$$

• komplementarni klasiifikatori

$$TP_R \rightarrow \bar{FN}_{1-R}$$

$$FN_R \rightarrow \bar{TP}_{1-R}$$

$$TN_R \rightarrow \bar{FP}_{1-R}$$

$$FP_R \rightarrow \bar{TN}_{1-R}$$

$$TP_{R,h} = \frac{TP_h}{TP_h + FN_h} = \frac{FN_{1-h}}{FN_{1-h} + TP_{1-h}} = 1 - TP_{1-h}$$

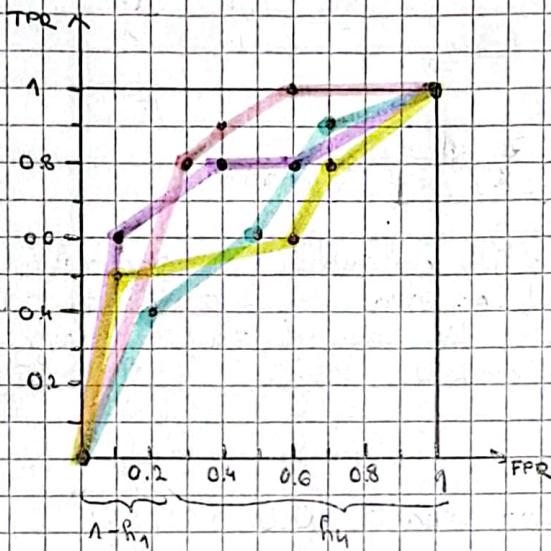
$$FP_{R,h} = \frac{FP_h}{TN_h + FP_h} = \frac{TN_{1-h}}{TN_{1-h} + FP_{1-h}} = 1 - FP_{1-h}$$

$$FPR_{1-h} = 1 - FP_{R,h}$$

$$TPR_{1-h} = 1 - TPR_{R,h}$$

(FPR, TPR)

- $1-h_1 : (0.6, 0.8), (0.4, 0.8), (0, 1, 0.6)$
- $1-h_2 : (0.7, 0.8), (0.5, 0.6), (0.2, 0.4)$
- $h_2 : (0.1, 0.5), (0.6, 0.6), (0.7, 0.8)$
- $h_1 : (0.3, 0.8), (0.4, 0.9), (0.6, 1)$



B $h_1 \succ h_2$

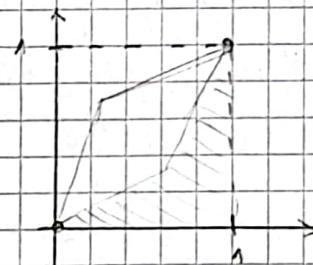
2.5

$h_1 \succ h_2$

$$\Leftrightarrow \forall \theta ((FPR_{\theta}(h_1) = FPR_{\theta}(h_2)) \Leftrightarrow (TPR_{\theta}(h_1) > TPR_{\theta}(h_2)))$$

$h_1 \succ h_2$

$$\Leftrightarrow AUC(h_1) > AUC(h_2)$$



Ovo što sigurno vrijedi jest
 $h_1 \succ h_2 \Leftrightarrow h_1 > h_2$

oko h_1 dominira nad h_2 , onda je
sigurno $AUC(h_1) > AUC(h_2)$

C

Relacija $h_1 > h_2 \Rightarrow h_1 \succ h_2$ ne mora
nužno vrijediti. Npr h_1 dominira nad
 h_2 samo na nekom intervalu FPR .

2.6.

$$N = 1000$$

3 hiperparametra

Broj paramet.
(komb.)

$$31^2 + 31 = 992$$

jezgra - linearna ili RBF

$$\therefore C - |C| = 15 + 15 + 1 = 31$$

$$\therefore \gamma - |\gamma| = 15 + 15 + 1 = 31$$

Koliko će puta
svaki primjer biti
iskorišten za trening?

$$k=10$$

$$l=1$$

$$c=5$$

train

$$\frac{4}{5} \cdot \frac{9}{10} N = \frac{18}{25} N$$

$$val \quad \frac{1}{5} \cdot \frac{9}{10} N$$

$$test \quad \frac{1}{10} N$$

$$\begin{aligned} & (k-1) \cdot ((c-1) \cdot 992 + 1) \\ & = 9 \cdot (4 \cdot 992 + 1) \\ & = 35721 \end{aligned}$$

A

2.7.

$$N = 1000$$

SVM

$$\begin{aligned} k &= 5 \\ c &= 5 \end{aligned}$$

	train
1000	$\frac{k-1}{k} \cdot \frac{c-1}{c} \cdot N$
val	$\frac{k-1}{k} \cdot \frac{1}{c} \cdot N$
test	$\frac{1}{k} \cdot N$

Svaku par modela nad pojma se optimiraju hiperparametri dijeli sečinu primjera. Razlikuju se samo u 2 prečekpa - validacijski prečekop je u drugom modelu dio skupca za treniranje. Drugim riječima svaku par modela dijeli $\frac{k-1}{k} \cdot \frac{c-1}{c} \cdot N$ primjera.

Ako je $k=5, c=5 \rightarrow$ par modela dijeli

$$\frac{4}{5} \cdot \frac{3}{5} \cdot N = 480 \text{ primjera}$$

Ako je $k=5, c=10 \rightarrow$ par modela dijeli

$$\frac{4}{5} \cdot \frac{8}{10} \cdot N = 640 \text{ primjera}$$

Razlika u broju primjera loguje se dijeli je onda $640 - 480 = 160$ primjera.

(C)

2.8.

$k \times c$ unakrsna projekcija

- optimiramo reg. faktor λ

λ_1 = prosjek optimalnih vrijed. reg. faktora
 F_1^1 = prosječna F_1 mjeru na ispitnom skupu

unitarnja petlja
 → validacija uvijek na 1. prečekopu

λ_2 = prosjek. optim. vrij. reg. faktora
 F_1^2 = prosječne F_1 mjeru na ispitnom skupu] optim. model treniran samo na F_{train}

λ_3 = prosjek. optim. vrij. reg. faktora
 F_1^3 = prosječne F_1 mjeru na ispitnom skupu]

$\lambda_2 = \lambda_3 > \lambda_1$
 $F_1^3 > F_1^2 > F_1^1$

(b)

U 1. varijaciji prenaučili smo hiperparametar λ_1 u 2. i 3. varijaciji dobivamo isti vrijednosti za hiperparametar λ , tj.
 $\lambda_2 = \lambda_3$ i očekujemo da je $\lambda_2 = \lambda_3 > \lambda_1$

(→ prema rezultatima za λ zaključujemo $F_1^3 > F_1^2 > F_1^1$
 $F_1^3 > F_1^2$ zbog pogreške u tren. optimalnog modela