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:

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

   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 ulaznome prostoru $\mathcal{X} = \{0,1\}^3$:

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

   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.*]

   (a) Pogreška hipoteze je očekivanje funkcije gubitka L. Nad kojom distribucijom je definirano to očekivanje? Koji je problem s takvom definicijom u praksi?

   (b) Definirajte *empirijsku* pogrešku preko funkcije gubitka L. Koja je pretpostavka implicitno ugrađena u tu definiciju?

   (c) 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.*]

   (a) Skicirajte krivulje pogreške učenje i ispitne pogreške u ovisnosti o složenosti modela. Naznačite područje prenaučenosti i podnaučenosti.

   (b) Objasnite zašto pogreška učenja s povećanjem složenosti modela teži k nuli.

   (c) Raspolažemo modelom $\mathcal{H}_\alpha$ 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?

   (d) Raspolažemo modelom $\mathcal{H}_\alpha$ 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 $\mathcal{H}_\alpha$ učenog algoritmom L_1, a α_2^* optimalnu vrijednost za $\mathcal{H}_\alpha$ učenog algoritmom L_2. Načinite skicu analognu onoj iz zadatka (a) i naznačite vrijednosti pogrešaka za modele $\mathcal{H}_{\alpha_1^*}$ i $\mathcal{H}_{\alpha_2^*}$.

   (e) 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 Zadatci s ispita

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

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

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:

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

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 dvodimenzijskome 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:

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

Razmatramo dva modela:

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

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 ulaznome prostoru $\mathcal{X} = \{0,1\}^2$. Razmatramo sljedeće modele:

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

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:

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

Optimizacijski postupak klasifikatora funkcionira 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 $h(x) \geqslant 0$ klasificiraju se pozitivno)

7. (P) Optimizacija parametara modela temelji se na funkciji gubitka $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_0^+$, 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 oznakama $\mathcal{Y} = \{0, 1\}$, 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 $\mathbb{R}^3$:

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

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 $0 \leqslant E(h|\mathcal{D}) \leqslant 1/4$

B $1/4 \leqslant E(h|\mathcal{D}) \leqslant 2/3$

C $\frac{1}{48} \leqslant E(h|\mathcal{D}) \leqslant 2/3$

D $1/12 \leqslant E(h|\mathcal{D}) \leqslant 3/4$

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

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

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 $E_1 = E_2 = E_3 > E_4$

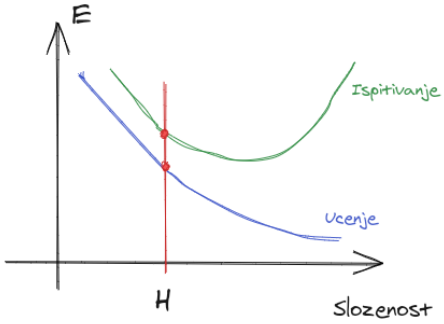
10. (P) Razmatramo klasifikacijski problem u ulaznome prostoru $\mathcal{X} = \mathbb{Z}^2$. Skup označenih primjera je $\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\} = \big\{((0,0),1), ((-1,-1),0), ((1,1),0)\big\}$. Razmatramo sljedeće modele $\mathcal{H}$ i funkcije preslikavanja $\boldsymbol{\phi}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

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

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 $\mathcal{H}_\alpha$, koji ima hiperparametar α kojim se može ugađati složenost modela. Isprobavamo dvije vrijednosti hiperparametra: α_1 i α_2. Treniramo modele $\mathcal{H}_{\alpha_1}$ i $\mathcal{H}_{\alpha_2}$ 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 $\mathcal{H}_{\alpha_2}$ je prenaučen

B Optimalan model je onaj s vrijednošću hiperparametra iz intervala $[\alpha_1, \alpha_2]$

C Model $\mathcal{H}_{\alpha_1}$ je podnaučen

D Model $\mathcal{H}_{\alpha_1}$ je manje složenosti od modela $\mathcal{H}_{\alpha_2}$

VL2 - Osnovni koncepti

I Zadaci za učenje

1.1.

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

Regresija
- na temelju ulaznog primjera model primjeru dodjeljuje numeričku vrijednost

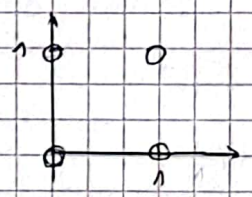
Pr

		OZNAKA
filtracija neželjene e-pošte	- klasifikacija	da/ne
predviđanje kretanja dionica	- regresija	vrijednost dionice
rangiranje rezultata tražilice	- klasifikacija	rang

1.2.

a) Hipoteza je fja koja preslikava skup primjera u skup oznaka, koje 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 najtočnije klasificira označene primjere, što procjenjujemo pomoću funkcije pogreške mjerene na skupu za učenje. Drugim riječima, učenje modela svodi se na optimizaciju parametara modela s empirijskom pogreškom kao kriterijskom fjom.

b) Binarna klasifikacija
$\mathcal{X} = \{0, 1\}^2$
- model primjere odvaja pravcem



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

$w_0 + w_1 x_1 + w_2 x_2 = 0$ ← pravac

MODEL

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

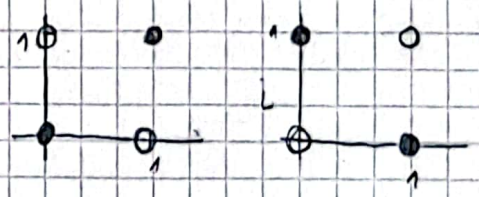
↳ klasificiramo: izlaz modela mora biti 0 ili 1

c) Dimenzija prostora parametara: 3 (w_0, w_1, w_2)
Broj različitih hipoteza: $16 - 2 = 14$

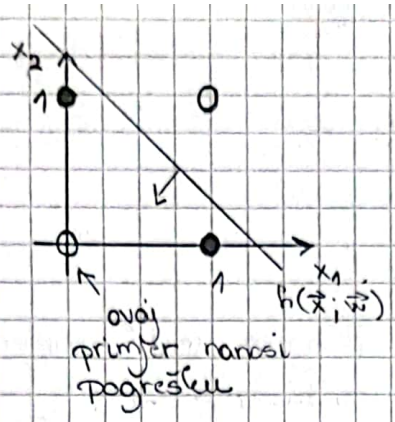
x_1	x_2	h
0	0	0 ili 1
0	1	0 ili 1
1	0	0 ili 1
1	1	0 ili 1

$2^4 = 16$

→ 2 skupa NISU linearno odvojiva XOR!!



d)



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

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

1.3

a) INDUKTIVNA PRISTRANOST

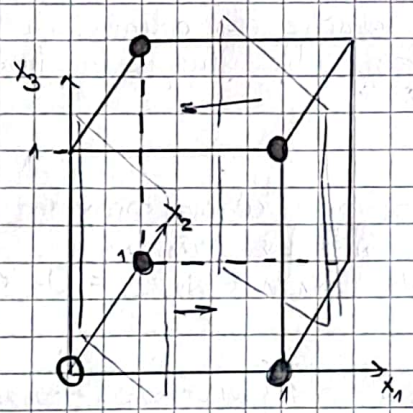
- neformalno: skup minimalnih pretpostavki pomoću kojih oznaka svih primjera slijedi dedukcijom
- formalno

inductive bias

$$\forall D\ \forall \mathcal{X}\ \forall \vec{x} \in \mathcal{X}$$
$$D \wedge \mathcal{X} \wedge \mathcal{B} \vdash h_L(\vec{x})$$

- 2 vrste induktivne pristranosti
  - pristranost jezikom (ograničenjem)
    - ograničavamo se na određeni skup hipoteza
  - pristranost pretraživanjem (preferencijom)
    - proizlazi iz same implementacije algoritma
    - iz više jednako dobrih hipoteza, algoritam pronalazi samo 1

b)



Pitanje: koja je klasifikacija neviđenih primjera?

Besmisleno pitanje
$\Rightarrow$ nema definiranog modela

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

$$h(\vec{x};\vec{w}) = 1\{w_0 + w_1x_1 + w_2x_2 + w_3x_3 \geq 0\}$$

$\hookrightarrow$ ovo je pristranost jezika

d) Pristranost iz modela nije dovoljna za jednoznačnu klasifikaciju primjera

1.4

a) Osnovne komponente dg. strojnog učenja

1. model
2. funkcija pogreške
3. optimizacijski postupak

b) MODEL - pristranost jezika / ograničenja
FJA POGREŠKE
OPT. POSTUPAK } pristranost pretraživ. / prefer.

1.5

a) pogreška hipoteze = očekivanje fje gubitka L
=> očekivanje definirano nad distribucijom podataka
=> problem je što u praksi ne znamo kojoj se distribuciji podaci pokoravaju

b) $\underbrace{\text{empirijska}}_{\text{nad podacima}}$ pogreška

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

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

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

c) 5 - false negative
2 - false positive

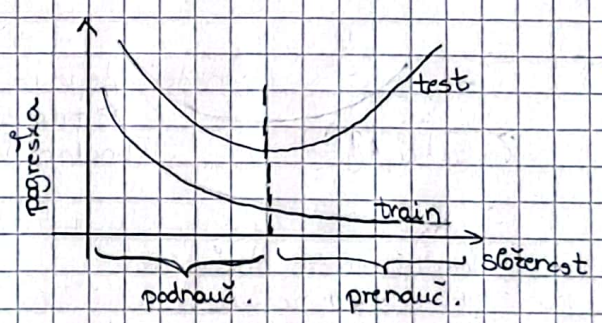
Matrica zabune

$$\text{stvarno}\begin{cases} T \\ N \end{cases} \underbrace{\begin{bmatrix} x & 5 \\ 2 & y \end{bmatrix}}_{\substack{T \quad N \\ \text{predviđanje}}} \Rightarrow E(\vec{w}|D) = \frac{5+2}{10} = \frac{7}{10}$$

$x + y = 3$

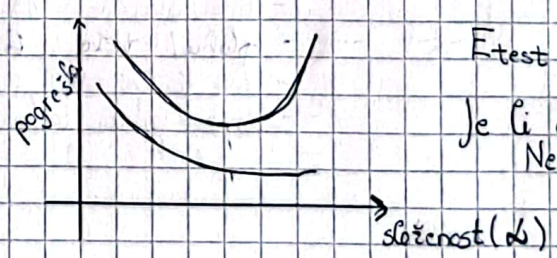
1.6

a)



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

c) H_α



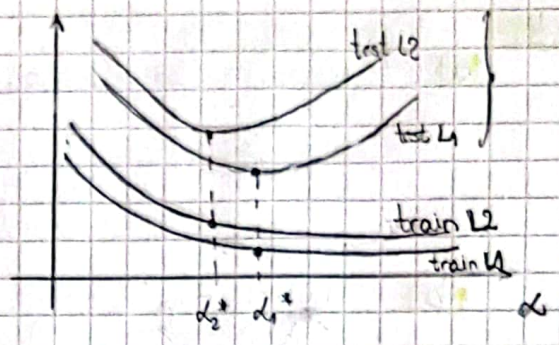
$E_{test} >> E_{train}$

Je li odabir parametra α subopt.?
Ne znamo, premalo informacija (znamo samo da je prosjep veći)

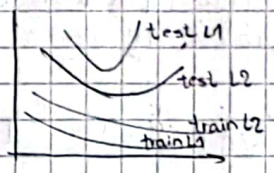
d) veći α - složeniji model

$E(\vec{\theta_2} \mid D) > E(\vec{\theta_1} \mid D)$

L_2 bolji model od L_1
α_2^* α_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 odvojivi
prema tome hipoteza ima manje od 256

B

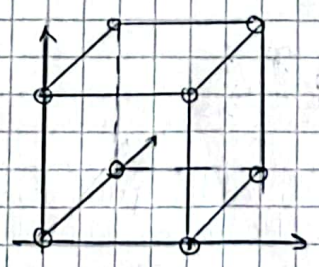
2.2.

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

$|\mathcal{H}| = ?$

$$\theta = \begin{bmatrix} \theta_{11} & \theta_{12} \\ \theta_{21} & \theta_{22} \\ \theta_{31} & \theta_{32} \end{bmatrix} \qquad \vec{x} \in \{0,1\}^3$$

- model definira kvadar
  $\Rightarrow$ primjer ima oznaku 1, ako je unutar kvadra, 0 inače



- kvadar može obuhvaćati
  - 1 točku : 8
  - brid / 2 točke : 12
  - plohu / 4 točke : 6
  - sve točke : 1
  - niti jednu točku : 1

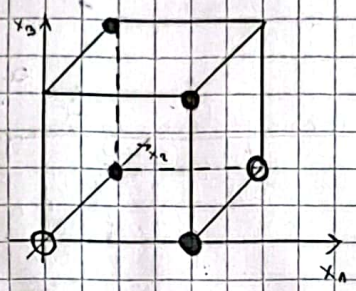
$|\mathcal{H}| = 28$

D

2.3. (B)

model nije definiran pa rasprava o empirijskoj pogrešci nema smisla

2.4.



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

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

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 dvije hiperravnine pa područje između tih ravnina može proglasiti kao područje klase 1

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

2.5

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

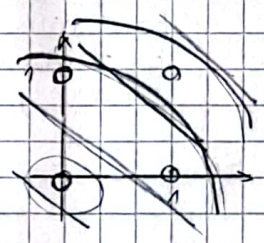
$\mathcal{H}_2: \; h_2(\vec{x};\vec{\theta}) = 1\{(x_1-\theta_1)^2 + (x_2-\theta_2)^2 \leq \theta_0^2\}$

$\mathcal{H}_3: \; h_3(\vec{x};\vec{\theta_1},\vec{\theta_2}) = h_1(\vec{x};\vec{\theta_1}) \wedge h_2(\vec{x},\vec{\theta_2})$

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

Klasifikacija !!

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



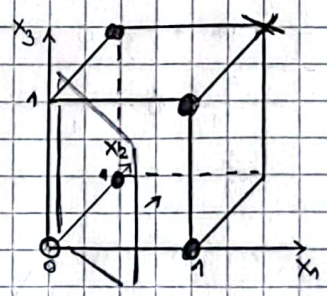
(B)

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

2.6

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

(C)



· fja gubitka 0-1

2.7.

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

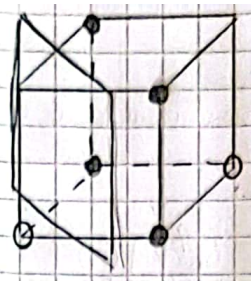
· bitnije je kažnjavati lažno negativne od lažno pozit.

$$\left.\begin{aligned} L(1,0) &> L(0,1) \\ L(1,1) &= L(0,0) = 0 \end{aligned}\right\} \text{(C)}$$

2.8.

FN - 1
FP - 1/2

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



minimalna pogreška

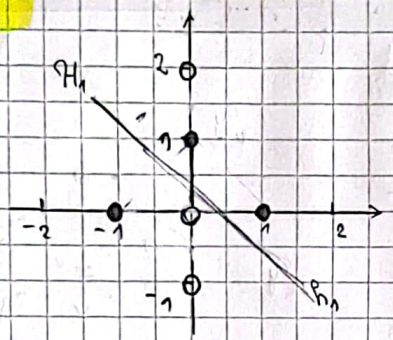
$$E(\vec{w}|D) = \frac{1}{6}\left(5\cdot 0 + \frac{1}{2}\right) = \frac{1}{12}$$

maksimalna greška

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

D

2.9.



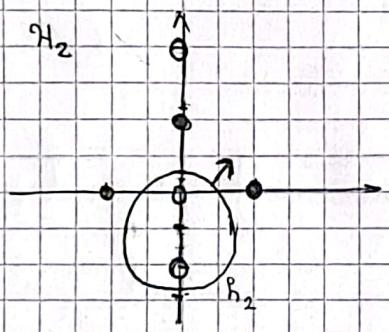
$\mathcal{H}_1$: $h_1(\vec{x}; \vec{\theta}) = 1\{\theta^T \vec{x} \geq 0\}$ (ravnina)

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

$\mathcal{H}_3$: $\mathcal{H}_1 \cup \mathcal{H}_2$ (ILI)

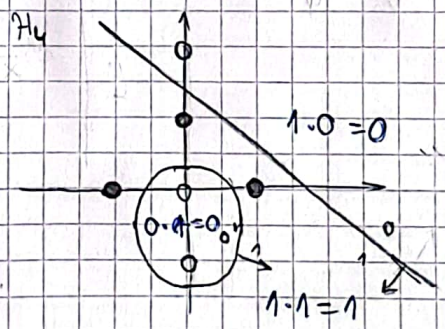
$\mathcal{H}_4$: $h_4(\vec{x}; \vec{\theta}) = h_1(\vec{x}) \cdot h_2(\vec{x})$ (I)

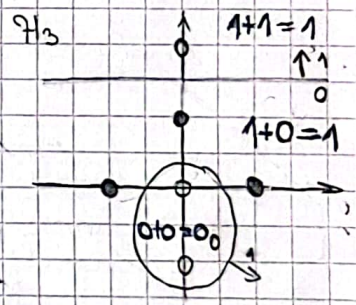
• linearno neodvojiv skup!



$\mathcal{H}_1$ minimalno 2 krivu
$\mathcal{H}_2$ minimalno 1 krivi } $E_1 > E_2$

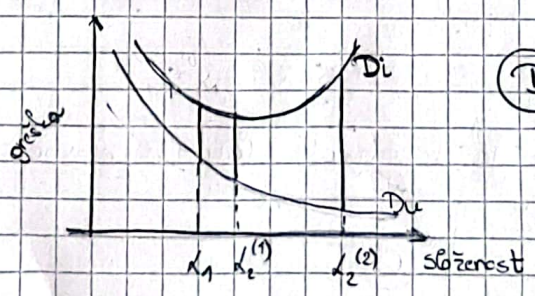
$\mathcal{H}_3$ minimalno 1 krivi
$\mathcal{H}_4$ 0 krivih





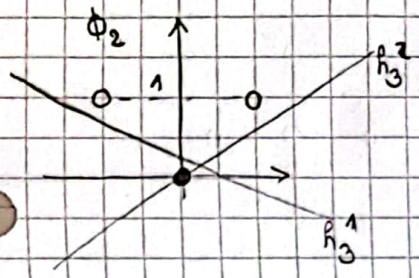
A $E_1 > E_2 = E_3 > E_4$

2.12.

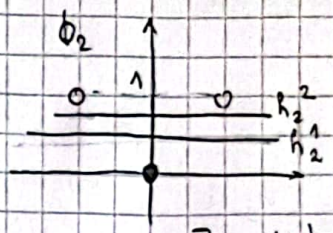


D $\mathcal{H}_{\lambda_1}$ manje složenosti od $\mathcal{H}_{\lambda_2}$

Napomena za 2.10.



Za ucrtane h_3^1 i h_3^2 radi se 2 različite hipoteze
$\Rightarrow$ ulazni prostor $\mathcal{X} = \mathbb{Z}^2$
- ne klasificiraju isto primjer $(-1, 0)$

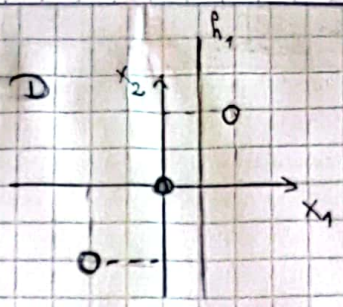


Za ucrtane h_2^1 i h_2^2 radi se o ISTOJ HIPOTEZI!
$\rightarrow$ ulazni prostor je $\mathbb{Z}^2$, a hipoteze sijeku y-os negdje na $\langle 0, 1 \rangle$ što isto klasificira sve primjere u $\mathbb{Z}$

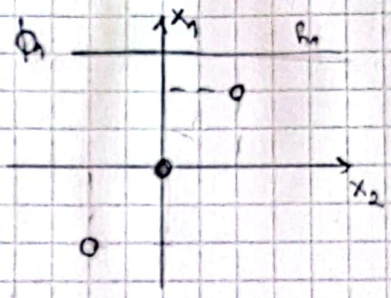
2.10.

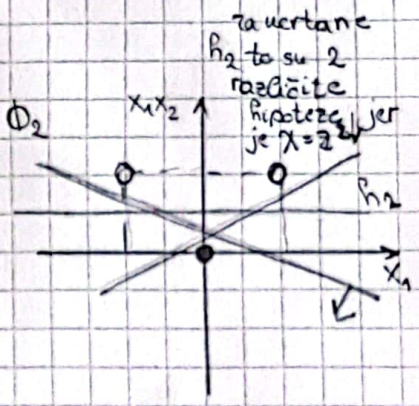
$\phi_1(\vec{x}) = (1, x_2, x_1)$
$\phi_2(\vec{x}) = (1, x_1, x_1 x_2)$

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



$h_1(\vec{x}) = 1\{\theta_1 x_1 + \theta_0 \geq 0\}$
$h_2(\vec{x}) = 1\{\theta_2 x_2 + \theta_0 \geq 0\}$
$h_3(\vec{x}) = 1\{\vec{\theta}^T \vec{x} \geq 0\}$
$h_4(\vec{x}) = 1\{\vec{x}_1^2 + \vec{x}_2^2 \geq \theta_0\}$





za ucrtane h_2 to su 2 različite hipoteze jer je $\mathcal{X} = \mathbb{Z}^2$

h_1 i h_2 - vertikalni / horizont. pravci u prostoru parametara

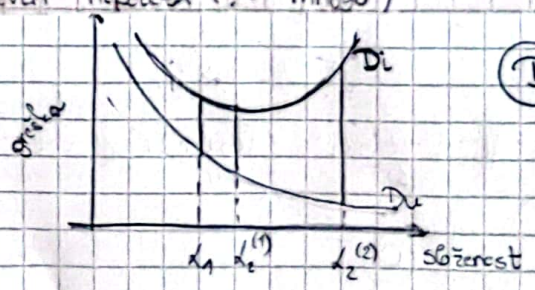
$h_4 + \phi_1$ - kružnica tj. van kruga
1 ishodište * centar kružnice
- ne može odvojiti dane primjere

$h_3 + \phi_2$ - primjeri će postati linearno odvojivi moguće ih razdvojiti pravcem, ali postoji više takvih hipoteza (∞ mnogo)

$h_2 + \phi_2$
- može odvojiti dani skup primjera
- postoji samo 1 takva hipoteza

(A)

2.12.



(D) $\mathcal{H}_{\alpha_1}$ manje složenosti od $\mathcal{H}_{\alpha_2}$

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 linearnog 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 $\mathbf{X}\mathbf{w} = \mathbf{y}$.

2. [***Svrha:** Razumjeti matrično rješenje za regulariziranu regresiju i izvježbati potrebnu matematku. 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:

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

   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 pretpostavke, funkcija kvadratne pogreške ima probabilističko tumačenje i opravdanje.*] Kod postupka najmanjih kvadrata empirijska je pogreška definirana kao:

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

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

2 Zadatci 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 sređivanja 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 $\mathbf{\Phi}$ koristimo funkciju $\boldsymbol{\phi}(\mathbf{x}) = (1, x_1, x_2, x_1x_2)$. Treniranjem modela na skupu $(\mathbf{\Phi}, \mathbf{y})$ dobili smo parametre $\mathbf{w} = (0.28, -0.58, 1.79, -0.75)^{\mathrm{T}}$. 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 $\mathbf{\Phi}$ s oznakama $\mathbf{y}$. **Koliko iznosi log-izglednost** $\ln \mathcal{L}(\mathbf{w}|\mathbf{\Phi}, \mathbf{y})$?

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

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

OPTIMIZACIJSKI POSTUPAK

$$\vec{w}^* = (X^T X)^{-1} X^T \vec{y} = X^+ \vec{y} = \underset{\vec{w}}{\operatorname{argmin}}\, E(\vec{w} \mid D)$$

b) Zašto koristimo kvadratni gubitak, a ne gubitak 0-1?
$\Rightarrow$ kvadratni gubitak derivabilan $\Rightarrow$ zatvorena forma
$\Rightarrow$ probabilistička interpretacija

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

X je matrica dimenzija $N \times (n+1)$ tj. nije kvadratna i nema inverz. Također u većini primjera tipično je $N >> n+1$

1.2.

a) Izvod (detaljnije vidi 4.2.)

$$E(\vec{w} \mid D) = \frac{1}{2} \sum_{i=1}^{N} (y^i - \vec{w}^T \vec{x}^i)^2 = \frac{1}{2} (X\vec{w} - \vec{y})^T (X\vec{w} - \vec{y})$$

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

$$= \frac{1}{2} (\vec{w}^T X^T X \vec{w} - \underbrace{\vec{w}^T X^T \vec{y}}_{1\times 1} - \underbrace{\vec{y}^T X \vec{w}}_{1\times 1} + \vec{y}^T \vec{y})$$

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

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

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

$$\Leftrightarrow \quad X^T X \vec{w} = X^T \vec{y}$$

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

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

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

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

e) Račun inverza $X^T X$ može biti računalno vrlo zahtjevno.
Problem je kada primjeri imaju puno značajki

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

Moguće izračunati rješenje nekom numeričkom metodom, npr. gradijentni spust

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

$\Rightarrow$ tada pseudoinverz možemo izračunati drugim metodama npr. SVD kompozicijom „

1.3.

$$P(\vec{y} \mid X, \vec{w}) \overset{\text{i.i.d.}}{=} \prod_{i=1}^{N} P(y^i \mid \vec{x}^i)$$

$$= \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(- \frac{(y^i - h(\vec{x}^i; \vec{w}))^2}{2\sigma^2} \right)$$

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

$$= -\left(-N \ln(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y^i - h(\vec{x}^i; \vec{w}))^2 \right)$$

$$= N \ln(\sigma\sqrt{2\pi}) + \frac{1}{2\sigma^2} \underbrace{\sum_{i=1}^{N} (y^i - h(\vec{x}^i; \vec{w}))^2}_{E(\vec{w} \mid D)} „$$

II Zadaci s ispita

2.1.

$$E(\vec{w}|D) = \frac{1}{2}\sum(\vec{w}^T\vec{x}^i - y^i)^2 = \frac{1}{2}(X\vec{w}-\vec{y})^T(X\vec{w}-\vec{y})$$

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

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

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

D

2.2.

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

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

$$E(\vec{w}|D) = \frac{1}{2}\sum(y^i + \vec{w}^T\vec{x}^i)^2 = \frac{1}{2}(X\vec{w}+\vec{y})^T(X\vec{w}+\vec{y})$$

$$= \dots = \frac{1}{2}(\vec{w}^TX^TX\vec{w} + 2\vec{y}^TX\vec{w} + \vec{y}^T\vec{y})$$

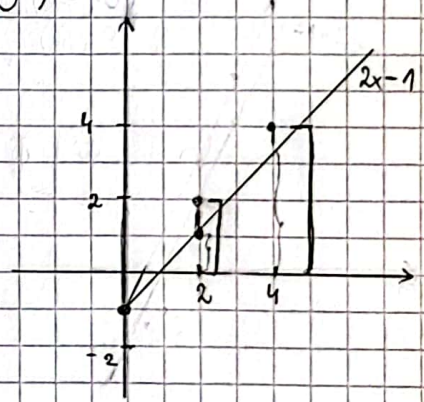
$$\nabla_w E = \frac{1}{2}(\vec{w}^T(X^TX + (XX^T)) + 2\vec{y}^TX)$$

$$= \vec{w}^T(X^TX) + \vec{y}^TX = 0$$

$$\vec{w}^T(X^TX) = -\vec{y}^TX \quad |^T$$

$$X^TX\vec{w} = -X^T\vec{y}$$

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

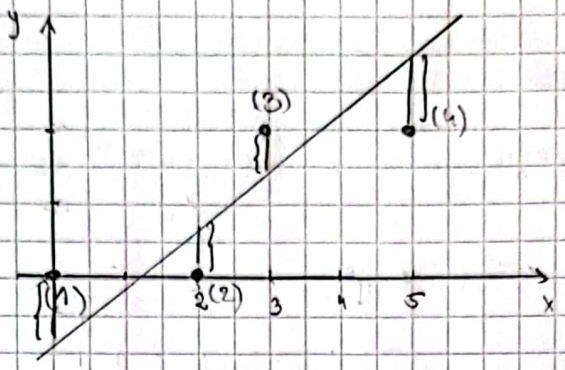


očekujemo $\begin{bmatrix}-1\\2\end{bmatrix}\cdot(-1) = \begin{bmatrix}1\\-2\end{bmatrix}$

A $\vec{w} = [1 \quad -2]$

točno rješenje s $(y-h(x))^2 = L$

2.3.



A - neizgledna opt.

B - $L1 = L4 = 1$, neizgledna opt.

C - baš nema smisla

ako je $L1 = L3 = 0$, onda bi model išao točno kroz (1) i (3), ali onda L2 i L4 nikako ne mogu biti manji od 1

(D) $L_1 = L_4 < L_2 = L_3$

<u>Napomena</u>: Zadatak moguće riješiti i računski

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

$$h(\vec{x}^i) = -\frac{2}{16} + \frac{6}{13} x$$

$$L(y^i, h(\vec{x}^i)) = (y^i - h(\vec{x}^i))^2 \Rightarrow L_1 = L_4 = \frac{4}{169} \quad L_2 = L_3 = \frac{100}{169}$$

2.4.

(B)

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

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

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

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

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

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

$$= \underbrace{N \cdot \ln\left(\frac{1}{\sqrt{2\pi}}\right)} - \frac{1}{2}\Big[(5-4.9)^2 + (-2+1.75)^2 + (1-0.84)^2 + (-3+3.3)^2 + (0-0.28)^2\Big]$$

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

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 linearnu funkciju odnosno (hiper)ravninu u prostoru značajki.***]

   (a) 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 $\boldsymbol{\phi}(x) = (1, x, x^2)$. Odredite vektor težina $\mathbf{w} = (w_0, w_1, w_2)$ tog modela.

   (b) 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.***]

   (a) Izvedite u matričnom obliku rješenje za vektor $\mathbf{w}$ za hrbatnu (L2-regulariziranu) regresiju.

   (b) Raspolažemo sljedećim skupom primjera za učenje:

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

   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.)

   (c) Komentirajte na koji način L2-regularizacija rješava problem numeričke nestabilnosti rješenja za $\mathbf{w}$.

   (d) 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 ulaznome prostoru.***] Linearnim modelom univarijatne regresije želimo aproksimirati jednu periodu funkcije $f(x) = \sin(\pi x)$. Raspolažemo sljedećim skupom primjera za učenje:

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

   (a) Izračunajte parametre linearnog modela regresije u ulaznom prostoru primjera, tj. s funkcijom preslikavanja u prostor značajki definiranom kao $\boldsymbol{\phi}(x) = (1, x)$. Skicirajte dobivenu regresijsku funkciju.

   (b) Izračunajte parametre modela polinomijalne regresije drugog stupnja, tj. modela koji koristi funkciju preslikavanja u prostor značajki definiranu kao $\boldsymbol{\phi}(x) = (1, x, x^2)$. Skicirajte dobivenu regresijsku funkciju.

(c) Izračunajte parametre modela polinomijalne regresije četvrtog stupnja, tj. modela koji koristi funkciju preslikavanja u prostor značajki definiranu kao $\boldsymbol{\phi}(x) = (1, x, x^2, x^3, x^4)$, 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 odabranih 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 ulaznome 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 pretpostavci 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, $h(\mathbf{x}; \mathbf{w}) = w_0 + w_1 x_1 + w_2 x_2$. 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 $\boldsymbol{\phi} : \mathbb{R}^n \to \mathbb{R}^{m+1}$, 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 irelevantne, odnosno zavisna varijabla u stvarnosti uopće ne ovisi o tim varijablama. Ako model nije regulariziran, koje su očekivane težine tih značajki?

2 Zadatci s ispita

1. (N) Raspolažemo sljedećim skupom primjera u dvodimenzijskome 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 polinomijalne 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

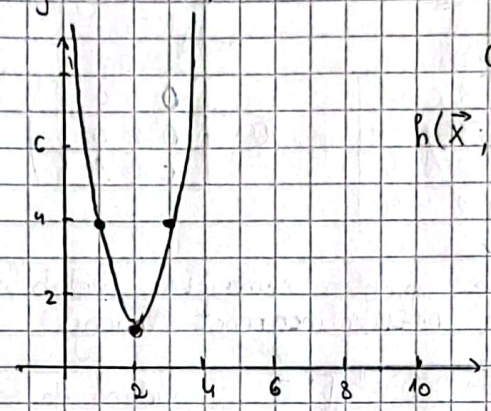
3. (P) Koristimo regresiju za predviđanje uspjeha na studiju. Kao značajke možemo koristiti ocjene u četiri razreda srednje škole (značajke x_1–x_4), prosjek ocjena sva četiri razreda (x_5) te uspjeh iz matematike (x_6) i fizike (x_7) na državnoj maturi (ukupno 7 značajki). Ne moramo iskoristiti sve 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 Zadaci za učenje

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



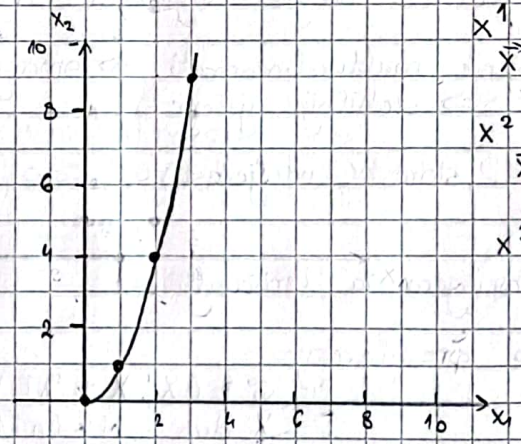
$$\phi(x) = (1, x, x^2)$$

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

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

$$\vec{w} = \begin{bmatrix} 13 & -12 & 3 \end{bmatrix}$$

b) $x_1 = x$
$x_2 = x^2$



$x^1 = 1 \quad h(x^1) = 4$
$\vec{x} = (1, 1)$

$x^2 = 2 \quad h(x^2) = 1$
$\vec{x} = (2, 4)$

$x^3 = 3 \quad h(x^3) = 4$ //
$\vec{x} = (3, 9)$

1.2. a)

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

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

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

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

$$\nabla_w E_a = \frac{1}{2} \Big(\vec{w}^T \big(\underbrace{(\Phi^T\Phi)^T + (\Phi^T\Phi)}_{2\Phi^T\Phi}\big) - 2\vec{y}^T \Phi\Big) + \frac{\lambda}{2} \vec{w}^T (I + I)$$

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

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

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

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

b)

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

$$h(\vec{x}) = w_0 + w_1 x + w_2 x^2$$
$$\lambda = 10$$

w_0 se ne regul.!

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

$$= \left(\begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 4 \\ 0 & 1 & 4 & 16 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 4 & 16 \end{bmatrix} + 10 \cdot \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 4 \\ 0 & 1 & 4 & 16 \end{bmatrix} \begin{bmatrix} 4 \\ 1 \\ 2 \\ 5 \end{bmatrix}$$

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

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

vidimo da se svakom dupcu Grammove matrice dodaje vrijednost λ na drugom mjestu.

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

d) $\vec{x}$ = (površ., starost, udaljenost)
y = cijena

minimalno primjera za izračunljivost

- bez preslikavanja

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

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 plitka – nestabilno rješenje
$N > (n+1)$
$\boxed{N \geq 4}$

- kvadratno preslikavanje s interakcijskim značajkama
$\Phi(\vec{x}) = (1, x_1, x_2, x_3, x_1x_2, x_1x_3, x_2x_3, x_1^2, x_2^2, x_3^2)$

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

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

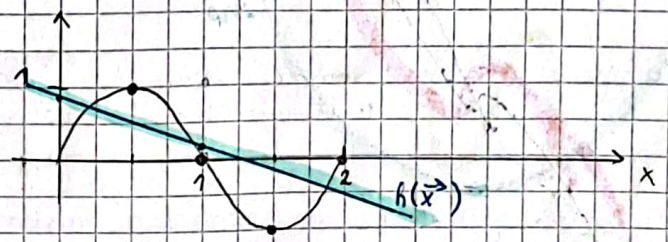
$D = \{(0.25, 0.707), (0.5, 1), (1, 0), (1.5, -1), (2, 0)\}$

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

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

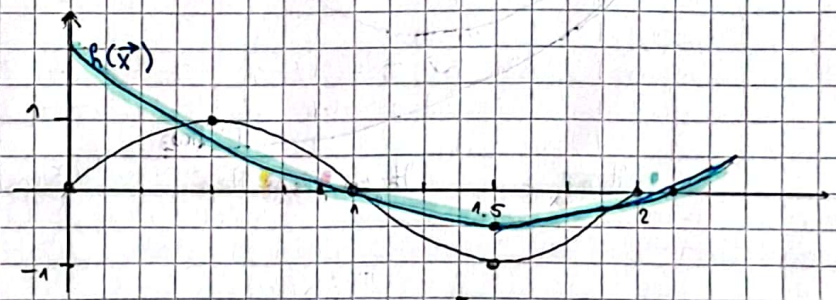
bez reg.

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



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

$$\Phi = \begin{bmatrix} 1 & 0.25 & 0.0625 \\ 1 & 0.5 & 0.25 \\ 1 & 1 & 1 \\ 1 & 1.5 & 2.25 \\ 1 & 2 & 4 \end{bmatrix} \qquad \vec{w} = (\Phi^T \Phi)^{-1} \Phi^T \vec{y} = \begin{bmatrix} 1.7538 \\ -2.9408 \\ 0.9755 \end{bmatrix}$$



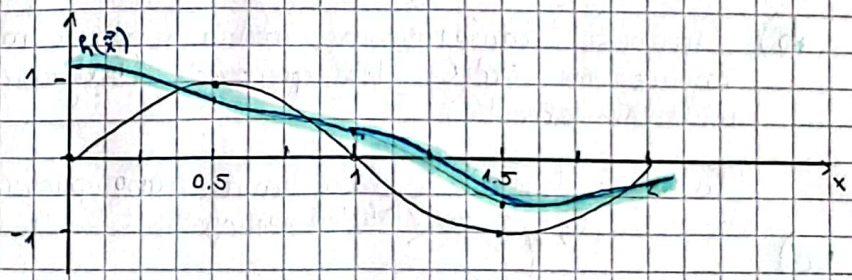
c) $\phi(x) = (1, x, x^2, x^3, x^4)$

$\lambda = 1$

$$\Phi = \begin{bmatrix} 1 & 0.25 & 0.0625 & 0.0156 & 3.906 \cdot 10^{-3} \\ 1 & 0.5 & 0.25 & 0.125 & 0.0625 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 1.5 & 2.25 & 3.375 & 5.0625 \\ 1 & 2 & 4 & 8 & 16 \end{bmatrix}$$

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

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



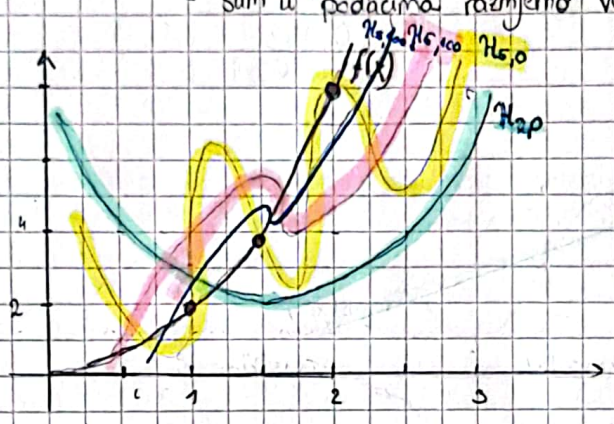
d) Najprikladniji je model pod c) jer ima najmanju kvadratnu pogrešku

1.4 $\mathcal{H}_{d,\lambda}$

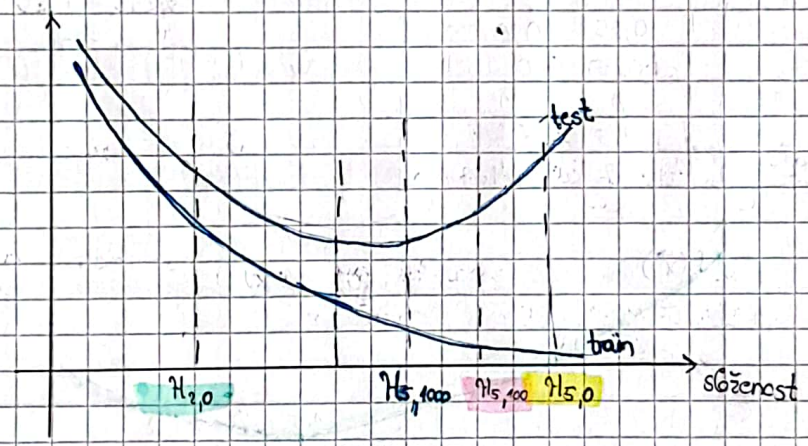
$\mathcal{X} = \mathbb{R}$

$\mathcal{H}_{2,0}$, $\mathcal{H}_{5,0}$, $\mathcal{H}_{5,100}$, $\mathcal{H}_{5,1000}$

a) regresijska fja
- podaci generir. polinomom 3. stupnja
- šum u podacima razmjerno velik



b)



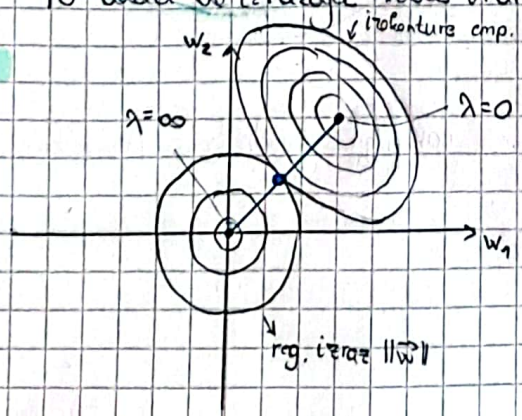
1.5 **a)** Regularizacija sprječava prenaučenost modela tako što ograničava rast vrijednosti parametara modela.

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

b) Prednost regulariziranog modela u odnosu na neregularizirani jest da je regularizirani model teže prenaučiti (+ daje stabilnija rješenja smanjivanjem multikolinearnosti)

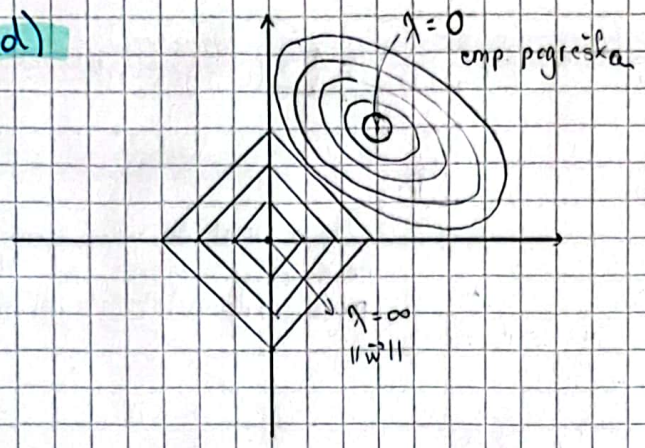
To dolazi do izražaja kada imamo malo primjera za učenje

c)



L2 regularizacijom parametri modela se priteže prema 0. Optimalni parametri w_1^* i w_2^* su oni u kojima se izokonture dodiruju

d)



- kod L1 regulariz. zbog regul. izraza čije su izokonture oštri rubovi lakše će se dogoditi da jedna od težina bude 0, tj. $\vec{w}^*$ će biti na koordinatnoj osi
- posljedica => rijetki modeli

1.6 $\phi : \mathbb{R}^n \to \mathbb{R}^{m+1}$
$m > n$
L2-reg.

a) Optimalan regularizacijski faktor λ odredili bismo unakrsnom provjerom (cross-validation)

b) Nakon treniranja modela uz L2-regularizaciju parametri modela koji imaju malu magnitudu (≈ 0) stoje uz NEBITNE ZNAČAJKE.

Izvorni model će biti PRESLOŽEN ako su parametri uz najsloženije značajke blizu 0!

c) L1 regularizirani model bi jače prigušio parametre modela ili čak rezultirao rijetkim modelom

d) Očekivane težine multikolineriranih značajki u neregulariziranom modelu imat će istu vrijednost za sve multikolinearne značajke.

II Zadaci s ispita

2.1. $\lambda = 1$

$\vec{w} = [2.12 \quad -0.94 \quad -0.08]$ $\qquad \|\vec{w}\| = \sum_{j=1}^{m} |w_j| = 0.94 + 0.08 = 1.02$ BEZ w_0

$$E(\vec{w}|D) = \frac{1}{2}\sum_{i=1}^{N}(y^i - h(\vec{x}^i))^2 + \frac{\lambda}{2}\|\vec{w}\|_1 \to 0.51$$

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

$$h(\vec{x}) = w_0 + w_1x_1 + w_2x_2 = 2.12 - 0.94x_1 - 0.08x_2$$

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

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

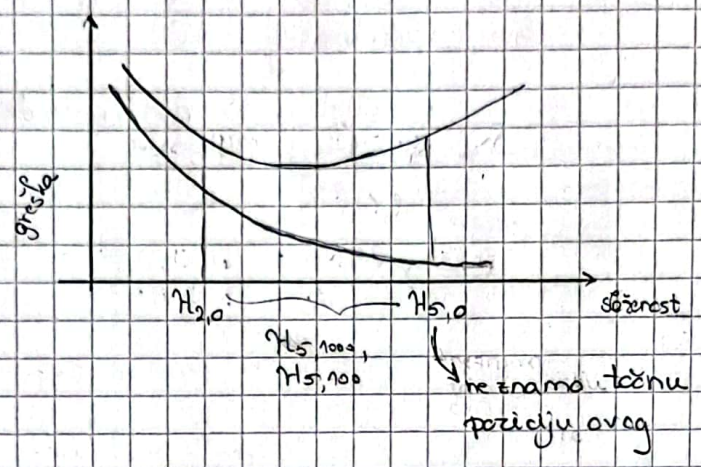
$$= 2.6946$$

2.2

$\mathcal{H}_{d,\lambda}$

$\mathcal{H}_{2,0}$ $\mathcal{H}_{5,100}$ $\mathcal{H}_{5,0}$ $\mathcal{H}_{5,1000}$

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



(B) $\mathcal{H}_{2,0}$ imat će veću pogrešku na skupu za učenje od $\mathcal{H}_{5,0}$, ali mogu podjednako lošе general.

2.3.

(C) $x_1 - x_4$ = ocjene srednje škole
x_5 = prosjek svih ocjena => kolinearnost (miičemo iz skupa)
x_6 = mat.
x_7 = fiz.

$\phi(x_1, x_2, x_3, x_4, x_6, x_7)$ =

1	-	1
x_i	-	6
x_i^2	-	6
$x_i x_j$	-	15
$x_i x_j x_z$	-	20

} 48 značajki

• rješenje stabilno i bez regularizacije

već uračunato kod preslik.

$N \geq m + 1$
$N \geq 48$

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 linearnog 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 linearna 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 ($K = 3$) klasa:

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

   (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 klasa. 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 modificirati primjer).

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

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

   (a) Objasnite zašto ne možemo izravno minimizirati $E_m(\mathbf{w}|\mathcal{D})$.

   (b) Je li pogreška perceptrona $E_p(\mathbf{w}|\mathcal{D})$ gornja ograda za pogrešku $E_m(\mathbf{w}|\mathcal{D})$? Objasnite.

   (c) Jedan nedostatak perceptrona 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 perceptrona jest da postupak ne konvergira ako primjeri nisu linearno odvojivi. 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 perceptrona 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 zašto 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 dvodimenzijskome ulaznome 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 klasa je između kategorija "politika" i "znanost". Očekivano, u kategoriji "politika" ima najviše članaka, dok ih u kategoriji "znanost" ima 5× 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 perceptrona. Koristimo algoritam perceptrona. 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× puta manje značajki od OVR, ali bi mogao raditi bolje na člancima iz kategorije "znanost"

   B OVR će imati 2× manje značajki od OVO, ali bi mogao raditi lošije na člancima iz kategorije "znanost"

   C OVO će imati 5× puta manje značajki od OVR, ali bi mogao raditi lošije na člancima iz kategorije "znanost"

   D OVR će imati 5× 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 dvodimenzijskome 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 šesterodimenzijski prostor značajki, koja je definirana na sljedeći način:

$$\boldsymbol{\phi}(\mathbf{x}) = (1, x_1, x_2, x_1x_2, x_1^2, x_2^2)$$

Početne težine perceptrona neka su sljedeće:

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

Koliko iznosi empirijska pogreška perceptrona na skupu za učenje prije početka treniranja (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 perceptrona neće konvergirati. Linearna neodvojivost podataka je konceptualni razlog zašto algoritam ne konvergira. **Koji je tehnički razlog zašto algoritam perceptrona 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

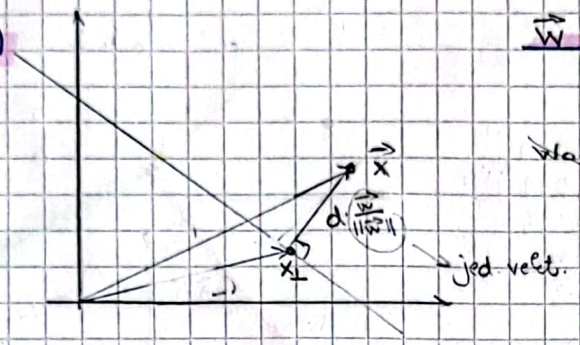
C U prostoru parametara ne postoji točka u kojoj je gradijent empirijske pogreške jednak nuli

D U prostoru parametara postoji više točaka za koje je empirijska pogreška jednaka nuli

V05 - Linearni diskriminativni modeli

I Zadaci za učenje

1.1. a)



$\vec{w}$ normala hiperravnine

$\vec{x}_1, \vec{x}_2$ na hiperravnini

$$h(\vec{x}_1) = h(\vec{x}_2)$$
$$\cancel{w_0} + \vec{w}^T \vec{x}_1 = \vec{w}^T \vec{x}_2 + \cancel{w_0}$$
$$\vec{w}^T \underbrace{(\vec{x}_1 - \vec{x}_2)}_{(1)} = 0$$

(1) vektor koji opis. leži na hiperravnini

$\Rightarrow$ skalarni produkt vektora 0

$\Rightarrow$ vektori okomiti

$\Downarrow$ (1)

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

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

$$\vec{w}^T \vec{x} = \vec{w}^T \vec{x}_\perp + d\|\vec{w}\| \quad / + w_0$$
$$\vec{w}^T \vec{x} + w_0 = \vec{w}^T \vec{x}_\perp + w_0 + d\|\vec{w}\|$$
$$h(\vec{x}) = \underbrace{h(\vec{x}_\perp)}_{0} + d\|\vec{w}\|$$

$$\left[d = \frac{h(\vec{x})}{\|\vec{w}\|} \right]$$

1.2.

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

a) OVR - 3 klasifikatora : $\vec{h}_k(\vec{x}) =$

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

$$\vec{y}_0 = [1\ 1\ 0\ 0\ 0\ 0]^T$$
$$\vec{y}_1 = [0\ 0\ 1\ 1\ 0\ 0]^T$$
$$\vec{y}_2 = [0\ 0\ 0\ 0\ 1\ 1]^T$$

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

$$\vec{w}_0 = [0.335 \quad -0.217 \quad -0.005]^T$$
$$\vec{w}_1 = [0.259 \quad 0.234 \quad 0.222]^T$$
$$\vec{w}_2 = [0.406 \quad -0.017 \quad -0.217]^T$$

$$h_0(\vec{x}) = 0.335 - 0.217x_1 - 0.005x_2$$
$$h_1(\vec{x}) = 0.259 + 0.234x_1 + 0.222x_2$$
$$h_2(\vec{x}) = 0.406 - 0.017x_1 - 0.217x_2$$

$$\left\{ h(\vec{x}) = \underset{k}{\operatorname{argmax}}\ h_k(\vec{x}) \right\}$$

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

→ diskriminacijske fje

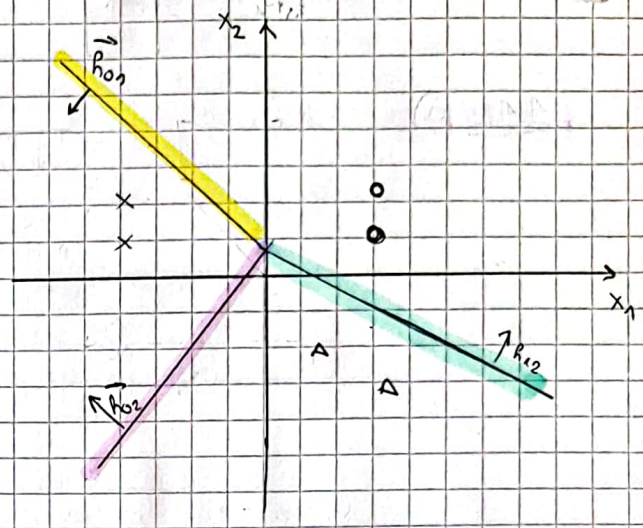
$h_{01}(\vec{x}) = w_{01,0} + w_{01,1} x_1 + w_{01,2} \cdot x_2$; $\vec{w}_{01} = \vec{w}_0 - \vec{w}_1$

↳ simetrala između h_0 i h_1

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

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

$$\vec{w}_{01} = [0.076 \quad -0.451 \quad -0.227]^T$$
$$\vec{w}_{02} = [-0.071 \quad -0.2 \quad 0.212]^T$$
$$\vec{w}_{03} = [-0.147 \quad 0.251 \quad 0.439]^T$$
$$\quad\quad w_0 \quad\quad w_1 \quad\quad w_2$$



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

$$\left.\begin{aligned} h_0(\vec{x}) &= 0.537 \\ h_1(\vec{x}) &= 0.691 \\ h_2(\vec{x}) &= 0.228 \end{aligned}\right\}$$ maksimum: 0.691 → klasa 1

(ili moguće iščitati iz gornje slike)

d) Ne možemo reći koja je vjerojatnost da primjer (-1,3) pripada klasi 1 jer linearni model koji smo koristili u shemi OVR nema probabilističku interpretaciju

e) <u>OVR vs OVO</u>

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
⇒ kao posljedica optimiz. postupak smanjuje $E(\vec{w}|D)$ naštrb primjera iz manje klase

f) Linearna regresija ne koristi se za klasifikaciju zbog
- nerobusnosti na stršeće vrijednosti
- nema probabilističke interpretacije

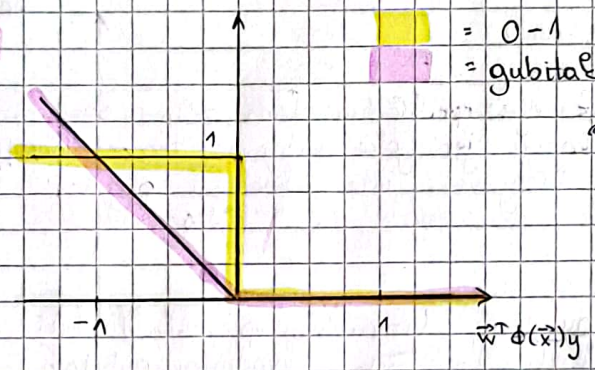
<u>Pr.</u> dodavanje primjera $((8,3), 1)$ u D

1.3.

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

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

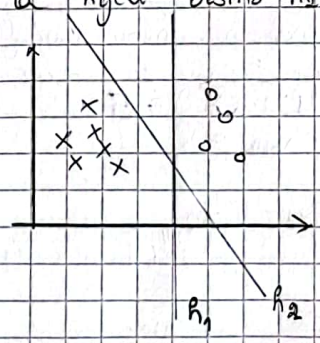
b)



= 0-1 gubitak
= gubitak perceptrona

Pogreška perceptrona 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 predočavanja primjera odabrati 1 hiperravninu od njih ∞. Algoritam perceptrona će se zaustaviti čim ispravno klasificira sve primjere za učenje, što ovisno o početnim uvjetima neće dati uvijek hiperravninu koja najbolje klasificira (dobit ćemo h_1, a htjeli bismo h_2)

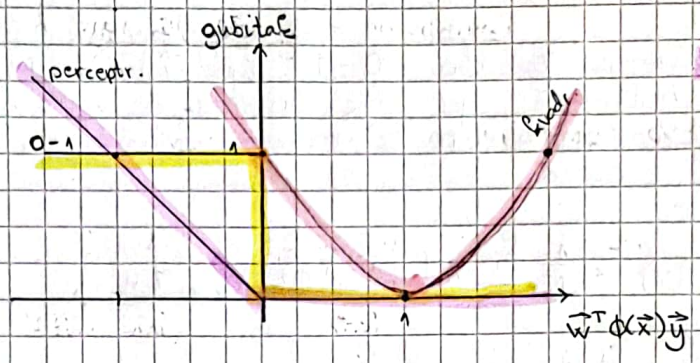


d) Budući da kod linearno neodvojivih primjera nije moguće postići savršenu klasifikaciju algoritam se ne zaustavlja
⇒ tj. ne može se postići $E_p(\vec{w}|D) = 0$
jer će se konstantno neki primjer kažnjavati sa odred. gubitkom

NEDOSTACI PERCEPTRONA
1. loša pristranost preferenc. ($\vec{w}^*$ ovisi o poč. uvj.)
2. algoritam ne konvergira za linearno neodvojive primjere
3. izlazi nemaju vjerojatn. interpretaciju

1.4.

a)



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 jako kažnjava točno klasificirane primjere, pa te primjere kažnjava isto kao i pogrešno klasificirane primjere

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

kvadr. gubitak > 0-1 loss

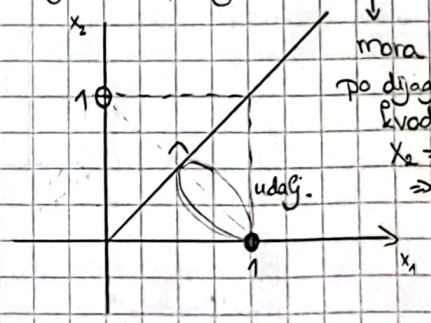
II Zadaci s ispita

2.1.

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

• maksim. minim. udaljenost primjera od hiperravnine
↓
mora biti pravac po dijag. I. kvadranta
$x_2 = x_1$
$\Rightarrow x_1 - x_2 = 0$

$y \cdot h(\vec{x}) = 5$
$w_2 = ?$



$h(\vec{x}) = \underbrace{w_0}_{0} + w_1x_1 + w_2x_2 = 5x_1 - 5x_2$

$y \cdot h(\vec{x}) = 5$
$1 \cdot (w_1 \cdot 1 + w_2 \cdot 0) = 5$
$w_1 = 5$

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

C

2.2. B 5 klasa

„sport"
„politika" 5x
„kriminal"
„zdravost" x
„lifestyle"

OVO, OVR

perceptron

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

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

↓

OVR će imati 2x manje značajki od OVO, ali mogao bi loše raditi za klasu „zdravost"

2.3.

$N = 1000$
$n = 555$

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

· nestabilnost zbog loše kondicije Φ

↳ Φ dimenzija $N \times (n+1)$ — dummy 1
stabilno ako $N \geq n+1$

OVO shema

	poz. prim.	neg. prim.	dim Φ	stabilno
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)

OVA shema → $\forall h_j$ dim Φ 1000×556 – stabilno

(C) 4

2.4.

$\vec{x}$	$\phi(\vec{x})$	y
(1,0)	(1,1,0,0,1,0)	1
(2,-3)	(1,2,-3,-6,4,9)	-1
(2,5)	(1,2,5,10,4,25)	-1

$\phi(\vec{x}) = (1, x_1, x_2, x_1x_2, x_1^2, x_2^2)$

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

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

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

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

(B)

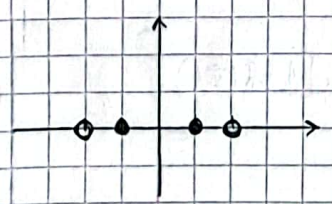
2.5.

$\vec{x}$	∇L	y
(-2,0)	( )	-1
(-1,0)		1
(1,0)		1
(2,0)		-1

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

Skup linearno neodvojiv
– perceptron ne konvergira

Tehnički razlog nekonvergencije?



Algor. perceptrona

$\vec{w}$

ponavljaj do konverg.
ako $f(\vec{w}^T \phi(\vec{x}^i)) \neq y^i$
$\vec{w} \leftarrow \vec{w} - \eta \nabla_w L$

$$\nabla_w L = -\phi(\vec{x}^i) y^i$$

(A) U svakoj točki prostora parametara postoji barem 1 primjer za koji je gradijent gubitka veći od nule.

VO6 - Logistička regresija

I Zadaci za učenje

1.1.

a) Poopćeni linearni model

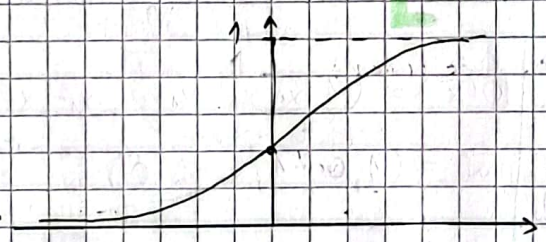
$$h(\vec{x}) = f(\vec{w}^T \phi(\vec{x}))$$

f = aktivacijska funkcija koja izlaz modela ograničava na interval $[0, 1]$ ili $[-1, 1]$

- ako je f nelinearna fja $\Rightarrow$ model je nelinearan u parametrima
- nelinearna granica $\Leftrightarrow$ nelinearnost 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 prikladan odabir za aktivac. fju jer

- izlaz modela će imati vjerj. interpre.
- fja gubitka slabo kažnjava TOČNO klasificirane primjere
- derivabilna je $\sigma(\alpha)(1-\sigma(\alpha))$ i slična je fji prijeđa

c) $E(\vec{w} | D) = -\ln P(\vec{y} | \vec{w})$

- izlaz modela - Bernoulijeva jdžba $P(y | \mu) = \mu^y (1-\mu)^{1-y}$

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

$$P(\vec{y} | X, \vec{w}) \overset{\text{i.i.d}}{=} \prod_{i=1}^{N} P(y^i | \vec{x}^i)$$

$$\hookrightarrow \ln P(\vec{y} | X, \vec{w}) = \ln \prod_{i=1}^{N} \left(h(\vec{x}^i)^{y^i} (1 - h(\vec{x}^i))^{1-y^i}\right)$$

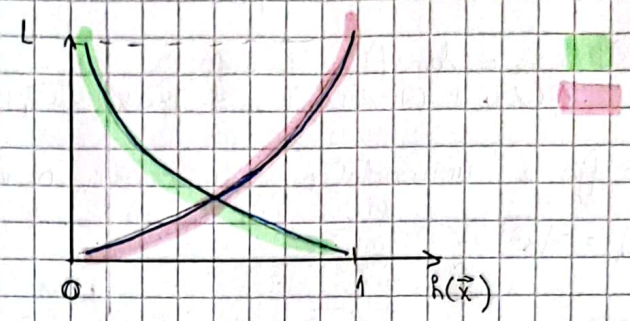
$$= \sum_{i=1}^{N} \ln \left[h(\vec{x})^{y^i} (1 - h(\vec{x}^i))^{1-y^i}\right]$$

$$= \sum_{i=1}^{N} y^i \ln[h(\vec{x}^i)] + (1-y^i) \ln[1 - h(\vec{x}^i)] \quad //$$

$$E(\vec{w} | D) \sim -\ln P(\vec{y} | X, \vec{w})$$

$$E(\vec{w} | D) = \frac{1}{N} \sum_{i=1}^{N} -y^i \ln[h(\vec{x}^i)] - (1-y^i) \ln[1 - h(\vec{x}^i)]$$

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 k ∞
Najmanji gubitak : ne postoji, teži k 0

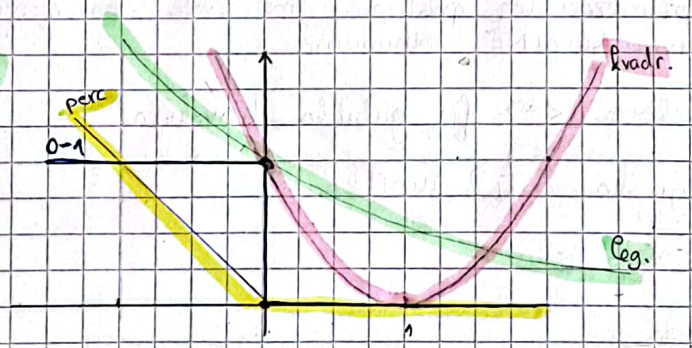
Uvijek postoji neki gubitak jer $h(\vec{x})$ kao izlaz nikad ne daje savršeno klasificiran primjer nego je izlaz na intervalu $\langle 0, 1 \rangle$

e) $y \in \{-1, 1\}$
i $L(y, 0) = 1$

$\Rightarrow$ vodi na reformulaciju:

$$\left[L(y, h(\vec{x})) = \frac{1}{\ln 2} \ln\left(1 + \exp(-\vec{w}^T \phi(\vec{x}) y)\right) \right]$$

f)



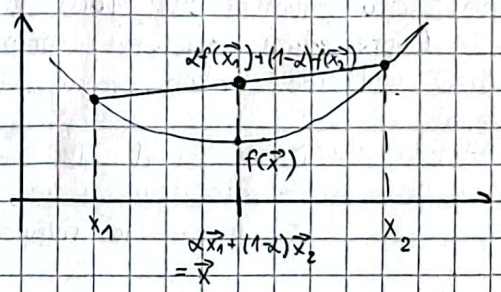
Zašto je logistički gubitak dobar za klasifikaciju?
Jako kažnjava netočno klasificirane primjere, a malo kažnjava točno klasi. primjere. Uvijek nanosi neki gubitak!

Je li logistički gubitak konveksni surogat 0-1 gubitka i što to znači?
Da jest! To znači da je log. gubitak konveksan i $\forall \vec{w}^T \phi(\vec{x}) y$ ima veću vrijednost od gubitka 0-1

1.2. a) $f: \mathbb{R}^n \to \mathbb{R}^n$ konveksna

$\Leftrightarrow$ ① domena f je konveksni skup
$\sum \alpha_i \vec{x}^i \in dom(f)$
ako $\sum \alpha_i = 1$ i $\vec{x}^i \in dom(f)$

② $\forall x_1, x_2 \in dom(f)$ i $\forall \alpha \in \langle 0, 1 \rangle$
$\Rightarrow f(\vec{x}) = f(\alpha \vec{x}_1 + (1-\alpha)\vec{x}_2)$
$\leq \alpha f(\vec{x}_1) + (1-\alpha) f(\vec{x}_2)$



b) f je kvazikonveksna (unimodalna)

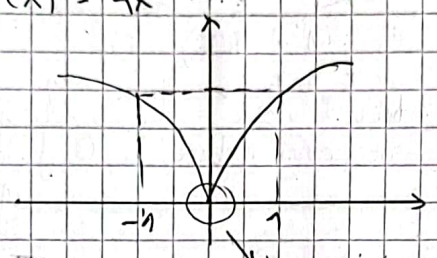
$\Leftrightarrow$ ① $dom(f)$ je konveksni skup

② $\forall x_1, x_2 \in dom(f) \quad \forall \alpha \in \langle 0,1 \rangle$

$$f(\alpha x_1 + (1-\alpha)x_2) \leq \max\{f(\vec{x}_1), f(\vec{x}_2)\}$$

Svaka konveksna fja je unimodalna, ali obrat ne vrijedi

Pr. $f(x) = \sqrt[3]{x^2}$



ima minimum, ali nije konveksna

$x_1 = 1 \quad x_2 = 2$

$\alpha = 0.5$

Konveksna: Ne!

$f(\alpha x_1 + (1-\alpha)x_2)$

$= f(1.5) = 1.31$

$\alpha f(x_1) + (1-\alpha) f(x_2)$

$= 1.29$

$1.31 > 1.29$

Kvazikonveks: DA!

$f(\alpha x_1 + (1-\alpha) f(x_2)) = 1.31$

$\max\{f(x_1), f(x_2)\} = 1.59$ ✓

c) U SU preferiramo konveksne fje pogreške jer nam to garantira da ćemo rješavanjem optimizacijskog postupka metodama 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 krće od činjenice da je u točki ekstrema $\nabla f(\vec{x}) = 0$
- u točkama nekstrema $\nabla f(\vec{x})$ pokazuje u smjeru najbržeg rasta fje f
- počeskom iz neke početne točke te pomicanjem u suprotnom smjeru od smjera gradijenta možemo stići u minimum fje

• Linijsko pretraživanje = odabire se η koji minimizira fju $g(\eta) = f(\vec{x} + \eta \Delta \vec{x})$ u smjeru spusta (suprotno od gradijenta)

$\rightarrow$ ubrzava konvergenciju

b) Grupni gradijentni spust radi ažuriranje težina tek nakon završene epohe (pregledani su svi primjeri), a stohastički grad. spust ažurira težine nakon svakog viđenog primjera.

Prednost SGD-a nad BGD-om jest

- što omogućuje online učenje
- manje je računalno zahtjevan

c) $\nabla_w E(w|D) = \sum_{i=1}^{N} \nabla_w L(y^i, h(\vec{x}^i)) = \sum_{i=1}^{N} (h(\vec{x}^i) - y^i)\,\phi(\vec{x}^i)$

BGD

1 $\vec{w} = \vec{0}$
2 ponavljaj do konvergencije
3 $\Delta\vec{w} = \vec{0}$
4 za $i = 1, \dots, N$
5 $h = \sigma(\vec{w}^T \phi(\vec{x}^i))$
6 $\Delta\vec{w} \mathrel{-}= (h - y^i)\,\phi(\vec{x}^i)$
7 η - linijsko pretraž.
8 $\vec{w} \mathrel{+}= \eta \cdot \Delta\vec{w}$

SGD

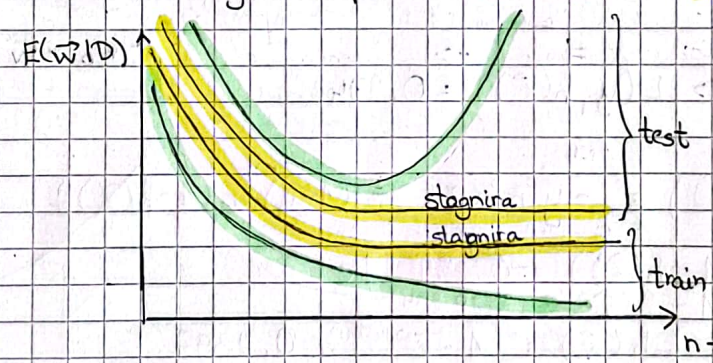
1 $\vec{w} = \vec{0}$
2 ponavljaj do konverg.
3 sluč. permutiraj D
4 za $i = 1, \dots, N$
5 $h = \sigma(\vec{w}^T \phi(\vec{x}^i))$
6 $\vec{w} \mathrel{-}= \eta (h - y^i)\,\phi(\vec{x}^i)$

1.4. $\lambda = 0$, $\lambda = 100$

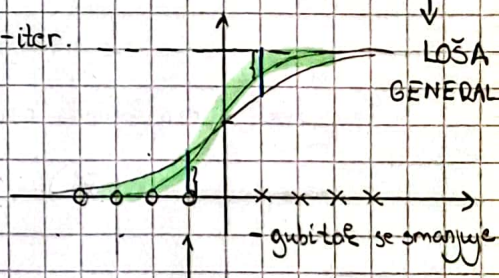
$$\left[\ \vec{w} = \vec{w}(1 - \eta\lambda) - \eta \sum_{i=1}^{N} (h(\vec{x}^i) - y^i)\,\phi(\vec{x}^i)\ \right]$$

[!! w_0 se ne regularizira $\lambda = 0$]

a) Linearno odvojiv skup $\lambda = 0$ i $\lambda = 100$

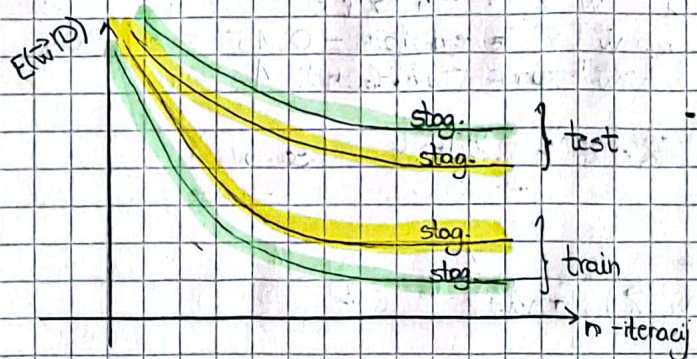


· za $\lambda = 0$
- nema regularizacije
- gradijent pogreške NIKAD neće biti 0
=> grad. spust neće konvergirati i težine rastu (otvrdnjavanje sigmoide)
↓
LOŠA GENERAL



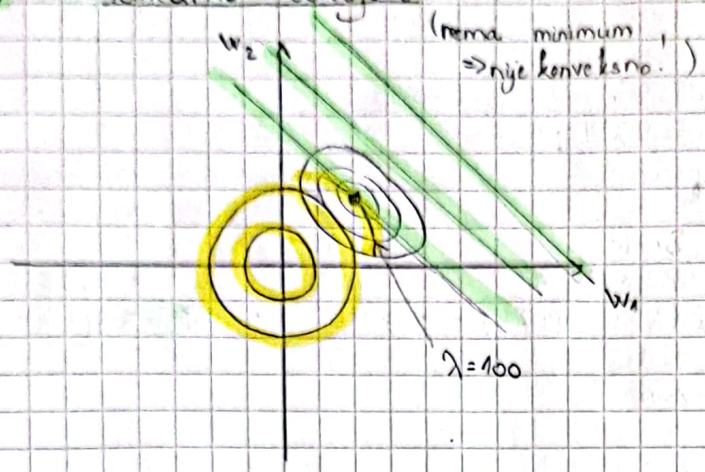
· za $\lambda = 100$
- regularizacija kažnjava složene modele (velike magnitude težina!)
- sprječava se efekt otvrdnjavanja sigmoide dolazi do konvergencije, stagnacije -> BOLJA GENERAL.

Linearno neodvojiv skup $\lambda = 0$, $\lambda = 100$

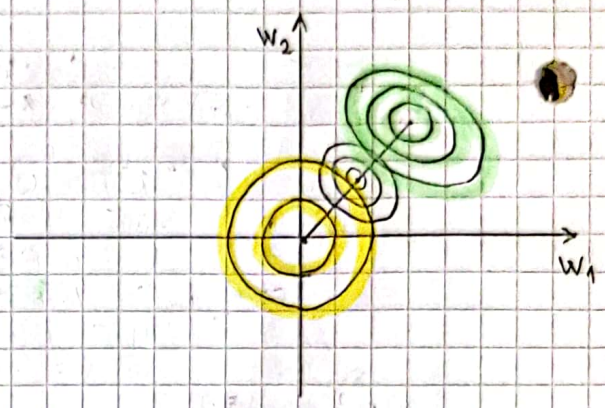


- model za linearno neodvojive primjere NIKADA SE NEĆE preraučiti (neki primjeri će uvijek biti krivo klasificirani)
=> dolazi do stag. i konvergencije i za $\lambda = 0$ i $\lambda = 100$
=> $\lambda = 100$ => bolja gener.

b) Linearno odvojivi



Linearno neodvojivi



$= E(\vec{w} \mid D)$

$=$ regularizacijski izraz $\|\vec{w}\|_2$

c) $= E_2(\vec{w} \mid D)$ za $\lambda = 100$

$\Rightarrow E(\vec{w} \mid D)$

$\Rightarrow$ reg. izraz

$= E_2(\vec{w} \mid D)$ za $\lambda = 100$

II Zadaci s ispita

2.1.

$w_0 = 0.15$

$\vec{w}$

$y = 0 \quad \Rightarrow L(y, h(\vec{x})) = 0.274$

$$L(y, h(\vec{x})) = -y \ln h(\vec{x}) - (1-y) \ln(1 - h(\vec{x}))$$

$y = 0$

$$\Rightarrow \ln(1 - h(\vec{x})) = -0.274 \quad | e$$

$$1 - h(\vec{x}) = e^{-0.274}$$

$$h(\vec{x}) = 1 - e^{-0.274} = 0.2397$$

[Ako promijenimo označeno: $L(y, h(\vec{x})) = -\ln h(\vec{x})$]

$$h(\vec{x}) = \sigma(-\vec{w}^T \vec{x}) = \frac{1}{1 + e^{-\vec{w}^T x}} = h(\vec{x})$$

$$h(\vec{x}) + h(\vec{x}) e^{-\vec{w}^T x} = 1$$

$$-\vec{w}^T x = \ln\left[\frac{1 - h(\vec{x})}{h(\vec{x})}\right] = 1.154$$

$$\vec{w}^T \vec{x} = -1.154$$

bez w_0 $\quad \vec{w}^T \vec{x} = -1.154 - 0.15 = -1.304$

bez w_0 $\quad \vec{w}^T \vec{x} = -2.604$

$$\Rightarrow -\vec{w}^T \vec{x} \cdot 2 = -2.454 \quad (\text{sa } w_0)$$

$y = 1$, značajke $\times 2$

$$h(\vec{x}) = \frac{1}{1 + \exp(-\vec{w}^T \vec{x})} = \frac{1}{1 + e^{2.45}} = 0.079$$

$$L(y, h(\vec{x})) = -\ln h(\vec{x}) = 2.538 \quad \text{B}$$

2.2. $L(y, h(\vec{x})) = 1.2 = -y \ln h(\vec{x}) - (1-y) \ln(1-h(\vec{x}))$

1° $y = 0$

$$L = -\ln(1-h(\vec{x}))$$
$$-1.2 = \ln(1-h(\vec{x})) \quad | e$$
$$1 - h(\vec{x}) = e^{-1.2}$$
$$h(\vec{x}) = 0.698$$

Promijeni oznaku $y = 1 \Rightarrow L = -\ln h(\vec{x}) = 0.359$

2° $y = 1$

$$L = -\ln(h(\vec{x}))$$
$$h(\vec{x}) = e^{-L} = 0.301$$

Promij. oznaku $y = 0 \Rightarrow L = -\ln(1-h(\vec{x})) = 0.3581$

$\Rightarrow$ svakako gubitak je 0.36 (C)

2.3. $\phi(\vec{x}) = (1, x_1, x_2, x_1 x_2)$

$\vec{w} = [0.2 \quad 0.5 \quad -1.1 \quad 2.7]$

L2-norma gradijenta gubitka za $(\vec{x}, y) = ((-0.5, 2), 1)$

$\phi(\vec{x}) = (1, -0.5, 2, -1)$ $\qquad \vec{w}^T \phi(\vec{x}) = -4.95$

$$\nabla_w L = (h(\vec{x}) - y)\phi(\vec{x})$$
$$h(\vec{x}) = \sigma(-\vec{w}^T \phi(\vec{x})) = 0.00703$$
$$\nabla_w L = -\underbrace{0.993}_{\approx 1} \phi(\vec{x}) = \phi(\vec{x}) = [1 \quad -0.5 \quad 2 \quad -1]$$

$\|\nabla_w L\|_2 \approx 2.5 \Rightarrow$ (B)

2.4.

$\vec{w}_0 = [1 \quad -4 \quad 4]$
$\vec{w}_1 = [1 \quad -4 \quad 6]$
$\vec{w}_2 = [1 \quad -1 \quad 7]$
$\vec{w}_3 = [1 \quad -7 \quad 1]$
$\vec{w}_4 = [1 \quad -7 \quad -3]$

Napomena:
- $w_0 - w^*$ mora biti normala kružnice $\frac{\lambda}{2}\|\vec{w}\|^2$
- $E_R(\vec{w}|D)$ je konveksna izgledom ista kao $E(\vec{w}|D)$, ali u koord. $w_1 \times w_2$ transl. za $\vec{w}_0 - \vec{w}^*$
- negativni gradijent $-\nabla_w E_R$ u w_4 na istom pravcu kao $w^* - w_4$ tj. w_4 na tang. kruž. u w^*

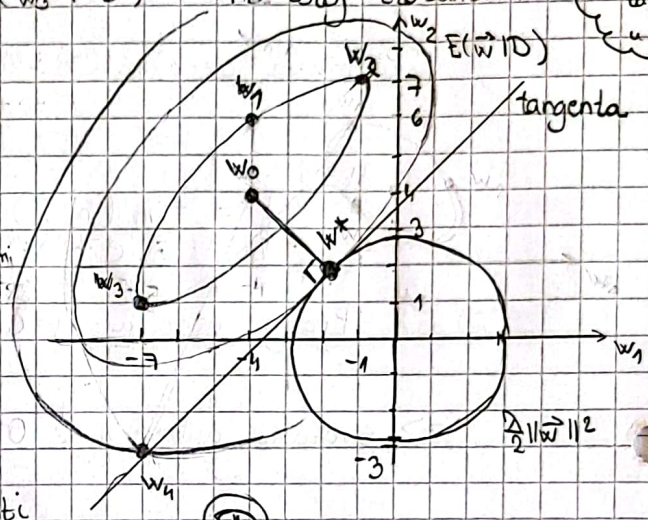
$E(\vec{w}_1|D) = E(\vec{w}_2|D) = E(\vec{w}_0|D)$ => na istoj izolinuri!

$\vec{w}_0$ min. $E(\vec{w}|D)$

$\lambda = 100$, $\frac{\lambda}{2}\|\vec{w}\|^2 = 400$
$\|\vec{w}\|^2 = 8$

$E_R(w|D) = E(w|D) + \frac{\lambda}{2}\|w\|^2$ ↓

kružnica $\sqrt{8} = 2\sqrt{2} \approx 2.8$ pravi

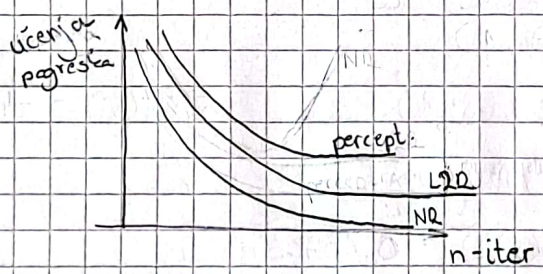


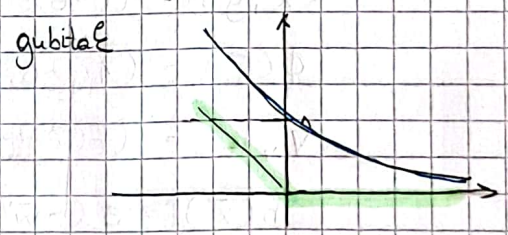
w^* – optimum $E_R(\vec{w}|D)$
=> najbolji inicijalni odabir je onaj koji najmanje krivuda, tj. onaj koji leži na tangenti kružnice $\|\vec{w}\|$ (ili je najbliži toj tangenti)
=> iz slike vidimo da je w_4 na toj tangenti
=> w_4 je najbolji inicijalni odabir

D

2.5.

- alg. perceptrona se zaustavio -> linearno odvojiv skup





- NR neće konvergirati – otvrdnjavanje sigmoide

D Pogreška učenja modela NR asimpt. teži k 0, dok pogreška učenja LR modela nakon određenog broja iteracija stagnira

2.6.

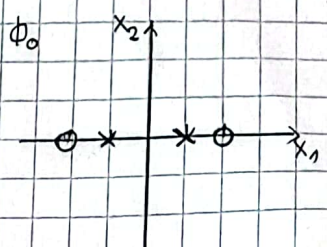
x_1	x_2	y	x_1^2	x_2^2	x_1x_2
-2	0	-1	4	0	0
-1	0	1	1	0	0
1	0	1	1	0	0
2	0	-1	4	0	0

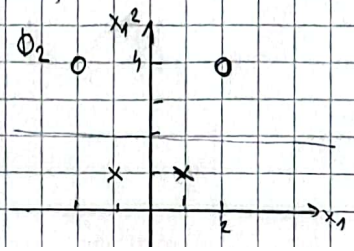
konverg. za lin. neodvojiv

konverg. za lin. odvojive

	LG	Percep
ϕ_0	Da	Ne
ϕ_1	Ne	Da
ϕ_2	Da	Ne

$\phi_0(\vec{x}) = (1, x_1, x_2)$ -> linearno neodvojivo
$\phi_1(\vec{x}) = (1, x_1, x_2, x_1^2, x_2^2)$ -> linearno odvojivo
$\phi_2(\vec{x}) = (1, x_1, x_2, x_1x_2)$ -> linearno neodvojivo





- minimizator emp pogreške

D LR + ϕ_2

2.7

SGD

$\lambda = 1000$

$\eta = 0.01$

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

$\vec{w} = [0.2 \quad 0.5 \quad -1.1 \quad 2.7]$

$\Delta w_1 = ?$

$(\vec{x}, y) = ((-1, 2), 1)$

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

$h(\vec{x}) = \sigma(-\vec{w}^T \phi(\vec{x})) = \sigma(7.9) = 3.71 \cdot 10^{-6}$

$\phi(\vec{x}) = (1, -1, 2, -2)$

$\nabla_w L = \underbrace{(3.71 \cdot 10^{-6} - 1)}_{\approx -1}\,\phi(\vec{x}) = [-1, 1, -2, 2]$

$w_1 = w_1(1 - \lambda\eta) - \eta \Delta_w L[1]$

$= 0.5(1 - 10) - 0.01 \cdot 1$

$= -4.51$

$\Delta w_1 = -4.51 - 0.5 = -5.01 \Rightarrow$ 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 $y\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x})$, 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 surogat* 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 $f : \mathbb{R}^n \to \mathbb{R}$ konveksna.

   (b*) Funkcija f je *kvazikonveksna* (ili *unimodalna*) akko je njezina domena $\mathbf{dom}\, f$ konveksna te ako za svaki $\mathbf{x}, \mathbf{y} \in \mathbf{dom}\, f$ i $0 \leqslant \alpha \leqslant 1$ vrijedi

   $$f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) \leqslant \max\{f(\mathbf{x}), f(\mathbf{y})\}\,.$$

   Kvazikonveksnost je poopćenje konveksnosti: svaka je konveksna funkcija unimodalna, ali obrat ne vrijedi. Pokažite primjerom da obrat ne vrijedi.

   (c) Zašto u strojnom 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. Demonstirati upoznatost s prednostima i nedostatcima 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 linearno neodvojive primjere.

   (c) Na grafikone iz prethodnoga zadatka docrtajte 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 x kada bismo njegove značajke pomnožili sa dva i promijenili 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 promijenili 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 dvodimenzijskog ulaznog prostora preslikali smo u prostor značajki funkcijom

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

   U jednoj iteraciji treniranja modela vektor parametara jednak je

$$\mathbf{w} = (0.2, 0.5, -1.1, 2.7)$$

   Koliko u toj iteraciji iznosi 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 dvodimenzijskoga 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 iteracije 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 asimptotski 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:

$$\begin{aligned}\boldsymbol{\phi_0}(\mathbf{x}) &= (1, x_1, x_2)\\ \boldsymbol{\phi_1}(\mathbf{x}) &= (1, x_1, x_2, x_1^2, x_2^2)\\ \boldsymbol{\phi_2}(\mathbf{x}) &= (1, x_1, x_2, x_1x_2)\end{aligned}$$

   Ukupno, dakle, isprobavamo šest kombinacija modela i funkcije preslikavanja. **Za koje će algoritme (model+preslikavanje) optimizacijski postupak pronaći minimizator empirijske pogreške?**

   A P+$\boldsymbol{\phi_0}$ B P+$\boldsymbol{\phi_2}$ C LR+$\boldsymbol{\phi_1}$ D LR+$\boldsymbol{\phi_2}$

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 $\boldsymbol{\phi}(\mathbf{x}) = (1, x_1, x_2, x_1x_2)$. U jednoj iteraciji treniranja modela vektor parametara jednak je $\mathbf{w} = (0.2, 0.5, -1.1, 2.7)$. **Koliko u toj iteraciji iznosi 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.***]

   (a) Definirajte funkciju *softmax*. Izračunajte $\text{softmax}(\boldsymbol{\alpha})$ za ulazni vektor $\boldsymbol{\alpha} = (2, 8, 1, 5)$. Koja su dva efekta funkcije softmax?

   (b) Definirajte model multinomijalne logističke regresije.

   (c) 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.***]

   (a) 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.

   (b) Objasnite prednost algortima LMS (odnosno stohastičkog gradijentnog spusta) nad grupnim (*batch*) gradijentnim spustom.

3. [***Svrha: Uočiti zajedničkosti poopćenih linearnih modela.***]

   (a) 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?

   (b) 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.***]

   (a) Objasnite što su bazne funkcije i koji je problem s fiksnim baznim funkcijama.

   (b) 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.

   (c) 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

1. (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 $\boldsymbol{\phi}$ 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:

$$h_k(\mathbf{x}) = \mathrm{softmax}_k\Big(\sum_{j=0}^{3} w_{j,k}\phi_j(\mathbf{x})\Big)$$

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 Zadaci za učenje

1.1.

a) $\text{softmax} : \mathbb{R}^n \to \mathbb{R}^n$

$$\text{softmax}_k(\vec{x}) = \frac{\exp(x_k)}{\sum_j \exp(x_j)}$$

2 efekta

1. normalizacija t.d. $\sum_k \text{softmax}_k(\vec{x}) = 1$
2. Vrijednosti se gnječe na $\langle 0, 1 \rangle$, manje vrijednosti još manje, a veće se pojačavaju

$$\alpha = [2 \;\; 8 \;\; 1 \;\; 5]$$

$$\left.\begin{aligned} \exp(2) &= e^2 \\ \exp(8) &= e^8 \\ \exp(1) &= e^1 \\ \exp(5) &= e^5 \end{aligned}\right\} \sum_j \exp(\alpha_j) = 3139.48$$

$$\left.\begin{aligned} \text{softmax}_2(\alpha) &= 0.0024 \\ \text{softmax}_8(\alpha) &= 0.9495 \\ \text{softmax}_1(\alpha) &= 0.00087 \\ \text{softmax}_5(\alpha) &= 0.0473 \end{aligned}\right\} \text{softmax}(\alpha) = [0.0024 \;\; 0.9495 \;\; 0.0009 \;\; 0.0473]$$

b) Model multinomijalne logist. regresije

$$h_k(\vec{x}; W) = \frac{\exp(\vec{w}_k^T \phi(\vec{x}))}{\sum_j \exp(\vec{w}_j^T \phi(\vec{x}))} = P(y = k \mid \vec{x}; W)$$

$$W = [\vec{w}_1 \;\; \vec{w}_2 \;\; \dots \;\; \vec{w}_K]$$

$$h(\vec{x}) = \underset{k}{\operatorname{argmax}}\, h_k(\vec{x}; W)$$

c) $$E(\vec{w} \mid D) = -\ln P(\vec{y} \mid \vec{x}; W)$$

$$\overset{\text{I.I.D.}}{=} -\ln \prod_{i=1}^{N} P(y^i \mid \vec{x}^i; W)$$

oznaka multinom. varijabla

$$P(y^i \mid \vec{x}^i; W) = \prod_{k=1}^{K} \mu_k^{y_k^i}$$

$$= -\ln \prod_{i=1}^{N} \prod_{k=1}^{K} h_k(\vec{x}^i; \vec{w}_k)^{y_k^i}$$

$$= -\sum_{i=1}^{N} \sum_{k=1}^{K} y_k^i \ln[h_k(\vec{x}^i; \vec{w}_k)]$$

1.2.

a) pravilo za ažur težina LMS poopćenih lin modela

$$\left[\vec{w} = \vec{w} - \eta H_E(\vec{w})^{-1} \nabla_{\vec{w}} E(\vec{w}|D) \right]$$

$H_E(\vec{w})$ = Hessova matrica

$$H_{E_{ij}} = \frac{\partial^2 E}{\partial w_i \partial w_j}$$

• za logističku regresiju: $H_E = \Phi^T \underbrace{S}_{\mathrm{diag}(h(\vec{x}^t)(1-h(\vec{x}^t)))} \Phi$

b) Prednost SGD (stoh.) nad BGD (batch)
- manje računalno intenzivnije
- pogodno za online učenje

1.3.

a)

	model	Distribucija zavisne var. y
Linearna reg.	$\vec{w}^T \phi(\vec{x})$	Normalna: $\mathcal{N}(h(\vec{x}), \sigma^2)$
Log. reg.	$\sigma(\vec{w}^T \phi(\vec{x}))$	Bernoulli: $\mathbb{P}(y\|\vec{x}, \vec{w}) = h(\vec{x})^y (1-h(\vec{x}))^{1-y}$
MNR	softmax $\Rightarrow \frac{\exp(\vec{w}_k^T \phi(\vec{x}))}{\sum_j \exp(\vec{w}_j^T \phi(\vec{x}))}$	Multinom.: $\mathbb{P}(y\|\vec{x}, W) = \prod_{i=1}^{K} h_k(\vec{x})^{y_k}$

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

distribucije eksponencijalne familije

b)

$$\underbrace{\vec{w}^*}_{\text{optim param}} = \underbrace{\underset{\vec{w}}{\mathrm{argmax}} \ln \mathbb{P}(D|\vec{w})}_{\text{maksimizacija log izglednosti}} = \underbrace{\underset{\vec{w}}{\mathrm{argmin}} E(\vec{w}|D)}_{\text{minimizacija emp. pogreške}}$$

1.4.

a) $\phi : \mathbb{R}^n \to \mathbb{R}^{m+1}$

$$\phi(\vec{x}) = (1, \phi_1(\vec{x}), \dots, \phi_m(\vec{x}))$$

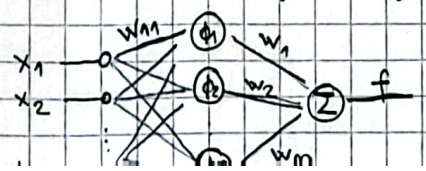
$\{\phi_1(\vec{x}), \dots, \phi_m(\vec{x})\}$ = m baznih fja

• problem fiksnih baznih funkcija
= njihov broj i oblik unaprijed određen
= ne znamo unaprijed koje su dobre bazne fje za naš problem

b)

$$h(\vec{x}; \vec{w}) = f(\vec{w}^T \phi(\vec{x})) = f\Big(\sum_{j=0}^{m} w_j^{(2)} \underbrace{f\Big(\sum_{i=0}^{n} w_{ji}^{(1)} x_i\Big)}_{\phi_j(\vec{x})}\Big)$$

$$= f(\vec{w}^T f(W^{(1)} \vec{x}))$$



c) Prednosti

i) u odnosu na poopćeni model bez baznih fja
- omogućivanje nelinearne granice

ii) u odnosu na poopćeni model s fiksnim baznim fjama
- bazne fje prilagodive, tj. ϕ se također uči iz podataka

Nedostatak
- složeniji postupak optimizacije → NEKONVEKSNOST FJE POGREŠ.
- veća mogućnost pretreniranosti modela

II Zadaci s ispita

2.1. $K = 3$

3d ulaz. prost. $n = 3$

MNR

$$W = \begin{bmatrix} 1 & 1 & 1 \\ 2 & 0 & -2 \\ 3 & -4 & 6 \\ -3 & 0 & 2 \end{bmatrix} \qquad \vec{x} = (3, 2, -1) \quad \vec{y} = (0, 1, 0) \qquad L(\vec{y}, h(\vec{x})) = ?$$

$$L(\vec{y}^i, h(\vec{x}^i)) = -\sum_{i=1}^{K} y^i \ln[h_k(\vec{x}^i)] + (1-y^i)\ln[1-h_k(\vec{x}^i)]$$

$$h_k(\vec{x}^i) = \frac{\exp(\vec{w}_k^T \phi(\vec{x}^i))}{\sum_j \exp(\vec{w}_j^T \phi(\vec{x}^i))} \quad \{e^{16} + e^{-7} + e^{5}\}$$

$$h(\vec{x}^i) = \text{softmax}(\vec{x}^T W) = \text{softmax}([16 \quad -7 \quad 5])$$

$$= [0.999\ldots \quad 1.03 \cdot 10^{-10} \quad 1.67 \cdot 10^{-5}]$$

$$\approx [1 \quad 0 \quad 0]$$

$$L(\vec{y}^i, h(\vec{x}^i)) = -\Big[\underbrace{0 \cdot \ln(1) + \overbrace{1 \cdot \ln(1-1)}^{0}}_{k=1} + \underbrace{1 \cdot \ln(1.03 \cdot 10^{-10}) + 0 \cdot \ln(1 - 1.03 \cdot 10^{-10})}_{k=2} + \underbrace{0 \cdot \ln(10^{-5}) + \underbrace{1 \cdot \ln(1 - 10^{-5})}_{0}}_{k=3}\Big]$$

$$= -\ln(1.03 \cdot 10^{-10}) = 22.996 \approx 23$$ // (C)

2.2.

$K = 3$

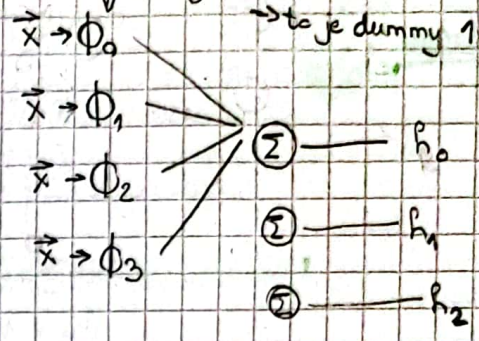
10 dim ulazni $\Rightarrow n = 10$

skalarno težina i prijelaz

$$h_k(\vec{x}) = \text{softmax}_k\left(\sum_{j=0}^{3} w_{jk}\, \phi_j(\vec{x})\right)$$

ovo nije adaptivno

$\rightarrow$ to je dummy 1



param.

$w_{jk} \Rightarrow 4 \cdot 3 = 12$

težine $\phi_j \Rightarrow 11 \cdot 3 = 33$

A 45 parametara

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.***]

   (a) Definirajte, korak po korak, problem maksimalne margine (tvrda margina).

   (b) Definirajte problem kvadratnog programiranja, pripadnu Lagrangeovu funkciju te dualnu Lagrangeovu funkciju i pripadne uvjete KKT. Obrazložite svaki uvjet KKT.

   (c) Definirajte, korak po korak, dualni problem maksimalne margine te pripadne uvjete KKT koji vrijede u točki rješenja.

   (d) Koje su prednosti formulacije problema kao dualnoga optimizacijskog problema?

   (e) Napišite primarnu i dualnu formulaciju modela SVM.

   (f) Objasnite što su to potporni vektori i kako znamo da oni sigurno leže na rubu margine.

   (g) 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:

$$\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_i = \\ \big\{((0,0),-1),((2,4),-1),((4,2),-1),((6,4),+1),((6,8),+1),((8,8),+1)\big\}$$

   (a) Skicirajte primjere u ulaznom prostoru $\mathbb{R}^2$ i granicu maksimalne margine. Napišite izraz za linearni model $h(\mathbf{x})$ koji odgovara toj granici.

   (b) Odredite širinu margine.

   (c) U ovom slučaju potporni vektori su $\mathbf{x}^{(3)} = (4,2)$ i $\mathbf{x}^{(4)} = (6,4)$. Odredite vektor Lagrangeovih koeficijenata $\boldsymbol{\alpha}$ temeljem izraza za ekspanziju težina $\mathbf{w}$ u potporne vektore.

   (d) Upoznajte se s formulom iz bilješke 20 iz skripte 8 te izračunajte pomak w_0.

   (e) 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 klasa sa znatno širom marginom nego na početnom skupu primjera. **Koliko je nova margina šira 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 $\boldsymbol{\alpha}$, odnosno Lagrangeove multiplikatore. Prisjetite 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 $\boldsymbol{\alpha}$. **Koliko iznosi najveća vrijednost parametra u vektoru dualnih parametara $\boldsymbol{\alpha}$?** (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 linearnom 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 x?**

A +907.43 B +541.53 C −373.22 D −739.13

4. (P) Skup za učenje čine sljedeći označeni primjeri:

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

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. $h(\mathbf{x}) = \mathbf{1}\{\mathbf{w}^{\mathrm{T}}\mathbf{x} \geqslant 0\}$. 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 Zadaci za učenje

1.1.

a) MODEL: $h(\vec{x}) = \vec{w}^T \phi(\vec{x}) + w_0$
oznake: $y \in \{-1, 1\}$

predikcija: $y = \text{sgn}(h(\vec{x}))$

Tvrda margina - primjeri linearno odvojivi

$$y^i = +1 \Rightarrow h(\vec{x}^i) \geq 0$$
$$y^i = -1 \Rightarrow h(\vec{x}^i) < 0$$

- predznačena udalj.: $d = \frac{h(\vec{x})}{\|\vec{w}\|}$
- nepredzn. udaljenost $d = \frac{y^i h(\vec{x}^i)}{\|\vec{w}\|}$

- MARGINA = udaljenost do najbližeg primjera

$$\xrightarrow{\text{bez } w_0} \frac{1}{\|\vec{w}\|} \min_i \{y^i(\vec{w}^T \phi(\vec{x}^i) + w_0)\}$$

$\Rightarrow$ možemo $\vec{w}, w_0$ skalirati tako da je $yh(\vec{x}) \geq 1$

$\hookrightarrow$ Problem: $\left[\underset{\vec{w}, w_0}{\text{argmax}} \frac{1}{\|\vec{w}\|} \quad \text{uz} \quad yh(\vec{x}) \geq 1 \right]$

$$\Leftrightarrow \left[\underset{\vec{w}, w_0}{\text{argmin}} \frac{1}{2}\|\vec{w}\|^2 \quad \text{uz} \quad yh(\vec{x}) \geq 1 \right]$$

b) Problem kvadratnog programiranja
- minimizacija konveksne fje uz ograničenja

$$\begin{cases} f(\vec{x}) \quad \text{minimizirati} \\ \text{ograničenja:} \quad g_i(\vec{x}) \leq 0 \quad i = 1, \dots, m \\ \qquad h_i(\vec{x}) = 0 \quad i = 1, \dots, p \end{cases}$$

Lagrangeova fja

$$L(\vec{x}, \vec{\alpha}, \vec{\beta}) = f(\vec{x}) + \sum_{i=1}^{m} \alpha_i g_i(\vec{x}) + \sum_{i=1}^{p} \beta_i h_i(\vec{x})$$

Dualna Lagrang. fja

$$\tilde{L}(\alpha, \beta) = \min_{\vec{x}} L(\vec{x}, \vec{\alpha}, \vec{\beta})$$

KKT uvjeti

$$\left.\begin{array}{l} g_i(x_i) \leq 0 \\ h_i(x_i) = 0 \end{array}\right\} \text{početni uvjeti}$$
$$\alpha_i \geq 0 \quad i = 1, \dots, m$$
$$\alpha_i g_i(\vec{x}) = 0 \quad i = 1, \dots, m$$

- komplementarna labavost
$\rightarrow$ ograničenje (ne)aktivno

c) <u>lagrang fja</u> (primarni problem)

$$L(\vec{w}, w_0, \vec{\alpha}) = \frac{1}{2}\|\vec{w}\|^2 - \sum_{i=1}^{N} \alpha_i \left[y^i (\vec{w}^T \vec{x}^i + w_0) - 1 \right]$$

→ prelazak u dual

$$\tilde{L}(\vec{\alpha}) = \min_{\vec{w}, w_0} L(\vec{w}, w_0, \vec{\alpha})$$

$$\frac{\partial L}{\partial \vec{w}} = 0 \quad \Rightarrow \quad \vec{w} = \sum_{i=1}^{N} \alpha_i y^i \vec{x}^i$$

$$\frac{\partial L}{\partial w_0} = 0 \quad \Rightarrow \quad 0 = \sum_{i=1}^{N} \alpha_i y^i$$

$$\tilde{L}(\alpha) = \frac{1}{2}\vec{w}^T \vec{w} - \sum_{i=1}^{N} \alpha_i y^i \vec{w}^T \vec{x}^i - \underbrace{\sum_{i=1}^{N} \alpha_i y^i}_{0} w_0 + \sum_{i=1}^{N} \alpha_i$$

$$= \frac{1}{2} \sum_{i=1}^{N} \alpha_i y^i (\vec{x}^i)^T \sum_{j=1}^{N} \alpha_j y^j \vec{x}^j - \sum_{i=1}^{N} \alpha_i y^i (\vec{x}^i)^T \sum_{j=1}^{N} \alpha_j y^j \vec{x}^j + \sum_{i=1}^{N} \alpha_i$$

$$\left[\tilde{L}(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y^i y^j (\vec{x}^i)^T \vec{x}^j \right]$$

ograničenja: $\alpha_i \geq 0 \quad i = 1, \dots, N$

$\sum_{i=1}^{N} \alpha_i y^i = 0$

• u točki rješenja vrijede uvjeti KKT:

$$\left.\begin{array}{ll} (1) & \alpha_i \geq 0 \\ (2) & y^i (\vec{w}^T \vec{x}^i + w_0) \geq 1 \\ (3) & \alpha_i (y^i h(\vec{x}^i) - 1) = 0 \end{array}\right] i = 1, \dots, N$$

→ potporni vektori $y^i h(\vec{x}^i) = 1$

d) Prednosti formulacije dualnog problema

- n+1 primarnih varijabli ⇒ N dualnih varijabli
  - u nekim slučajevima smanjena računalna složenost
- mogućnost korištenja jezgrenog trika
- korištenje optimizacijskog postupka SMO
  ⇒ $O(N^2)$ umjesto $O(n^3)$

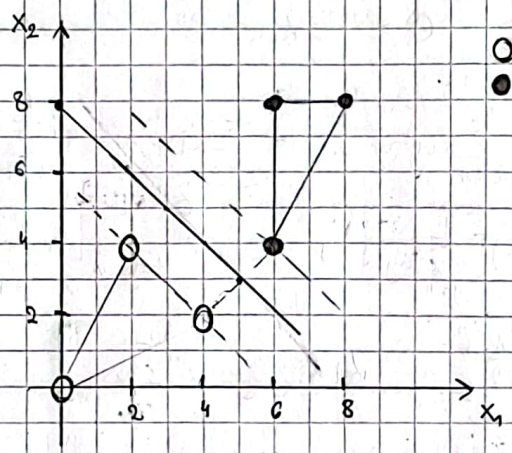
e) $$h(\vec{x}) = \underbrace{\vec{w}^T \vec{x} + w_0}_{\text{primarno}} = \underbrace{\sum_{i=1}^{N} \alpha_i y^i (\vec{x}^T (\vec{x}^i)) + w_0}_{\text{dualno}}$$

f) Potporni vektori su primjeri iz ulaznog 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 značajki u dualnoj formulaciji pojednostavljuje analizu problema i osigurava da za sve primjere vrijedi $y^i h(\vec{x}^i) \geq 1$ "

1.2.

a)



○ : $y = -1$
● : $y = 1$

$$h(\vec{x}) = x_1 + x_2 - 8$$

za potp. vektore
$$y^i h(\vec{x}^i) = 1$$
$(6, 4), 1$

$1 \cdot (6 + 4 - 8) = 2 \neq 1$

↳ gornje težine /: 2

$$\left[h(\vec{x}) = \frac{1}{2} x_1 + \frac{1}{2} x_2 - 4 \right]$$

b) Širina margine

$$\frac{1}{\|\vec{w}\|} = \frac{1}{\sqrt{\frac{1}{4} + \frac{1}{4}}} = \frac{1}{\sqrt{\frac{1}{2}}} = \sqrt{2}$$

bez w_0!!!

ili $$d = \frac{h(\vec{x}^3)}{\|\vec{w}\|} = -\sqrt{2}$$

opet bez w_0

c)
$$\vec{w} = \sum_{i=1}^{N} \alpha_i y^i \vec{x}^i$$

potporni vektori: $\vec{x}^3 = \begin{bmatrix} 4 & 2 \end{bmatrix}$ $y^3 = -1$
$\vec{x}^4 = \begin{bmatrix} 6 & 4 \end{bmatrix}$ $y^4 = 1$

$$\alpha_3 \cdot (-1) \cdot \begin{bmatrix} 4 \\ 2 \end{bmatrix} + \alpha_4 \cdot 1 \cdot \begin{bmatrix} 6 \\ 4 \end{bmatrix} = \begin{bmatrix} 1/2 \\ 1/2 \end{bmatrix}$$

$$\begin{cases} -4\alpha_3 + 6\alpha_4 = 1/2 \\ -2\alpha_3 + 4\alpha_4 = 1/2 \quad / \cdot (-2) \end{cases}$$

$$\begin{cases} -4\alpha_3 + 6\alpha_4 = 1/2 \\ 4\alpha_3 - 8\alpha_4 = -1 \end{cases}$$

$$-2\alpha_4 = -\frac{1}{2} \Rightarrow \begin{bmatrix} \alpha_3 = 0.25 \\ \alpha_4 = 0.25 \end{bmatrix}$$

d) $w_0 = ?$
$$y^i h(\vec{x}^i) = 1$$
$$y^i \left(\sum \alpha_j y^j \vec{x}^j (\vec{x}^i)^T + w_0 \right) = 1$$
$$w_0 = y^i - \sum_{N} \alpha_j y^j (\vec{x}^i)^T \vec{x}^j$$
$$\downarrow$$
$$w_0 = \frac{1}{|S|} \sum_{i \in S} y^i - \sum_{j \in S} \alpha_j y^j (\vec{x}^i)^T \vec{x}^j$$

$$w_0 = \frac{1}{2} \left(-1 - \left(\frac{1}{4}(-1) \begin{bmatrix} 4 \\ 2 \end{bmatrix} \begin{bmatrix} 4 & 2 \end{bmatrix} + \frac{1}{4} \cdot 1 \begin{bmatrix} 4 \\ 2 \end{bmatrix} \begin{bmatrix} 6 & 4 \end{bmatrix} \right) + 1 - \left(\frac{1}{4}(-1) \begin{bmatrix} 6 \\ 4 \end{bmatrix} \begin{bmatrix} 4 & 2 \end{bmatrix} + \frac{1}{4} \cdot 1 \begin{bmatrix} 6 \\ 4 \end{bmatrix} \begin{bmatrix} 6 & 4 \end{bmatrix} \right) \right)$$

$$= \frac{1}{2} \left(-1 - (-5 + 8) + 1 - (-8 + 13) \right)$$

$$= \frac{1}{2} (-1 - 3 + 1 - 5)$$

$$= -\frac{8}{2} = -4$$

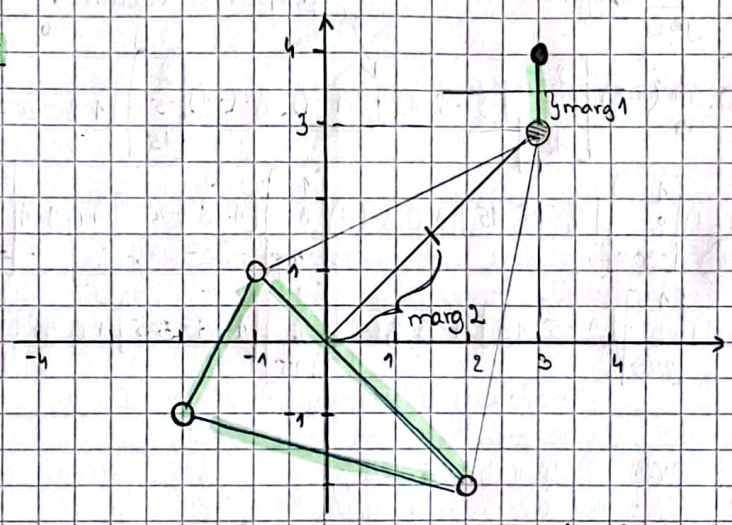
$$\boxed{w_0 = -4}$$

c) $\vec{x}^7 = (5,6)$

$$h(\vec{x}^7) = w_0 + \sum \alpha_i y^i (\vec{x})^T \vec{x}^i$$
$$= -4 + \left(\frac{1}{4}\cdot(-1)\begin{bmatrix}5\\6\end{bmatrix}\begin{bmatrix}4 & 2\end{bmatrix} + \frac{1}{4}\cdot 1\begin{bmatrix}5\\6\end{bmatrix}\begin{bmatrix}6 & 4\end{bmatrix}\right)$$
$$= -4 + (-8 + 13.5)$$
$$= 1.5 \quad \Rightarrow \quad \mathrm{sgn}(h(\vec{x})) = 1$$

II Zadaci s ispita

2.1.



$$\text{marg}2 = \frac{3-\sqrt{2}}{2}$$
$$\text{marg}1 = \frac{1}{2}$$
$$\frac{\text{marg}2}{\text{marg}1} = \left(\frac{\frac{3-\sqrt{2}}{2}}{\frac{1}{2}}\right) = 3-\sqrt{2}$$

↳ A

○ : $y = -1$
● : $y = 1$
⊖ : $y = \pm 1$

— konveksne ljuske s $((+3,3), 1)$

2.2.

$$X = \begin{bmatrix} 1 & 3 & 16 & -8 & 11 \\ 1 & -5 & 4 & -8 & -7 \\ 1 & 7 & -4 & 11 & 9 \\ 1 & 15 & -20 & 25 & 25 \end{bmatrix}$$

$$\vec{w} = [0.137 \quad -0.029 \quad +0.0194 \quad -0.0461 \quad -0.0388]$$
$$h(\vec{x}) = \vec{w}^T X$$
$$h(\vec{x}) = [1.156 \quad 1 \quad -1 \quad -2.81]$$

ovo će biti potporni vektori $\Rightarrow \alpha_2, \alpha_3 > 0$

bez w_0

$$\vec{w} = \sum \alpha_i y^i \vec{x}^i$$

$$\begin{bmatrix} -0.029 \\ 0.0194 \\ -0.0461 \\ -0.0388 \end{bmatrix} = \alpha_2 \cdot \begin{bmatrix} -5 \\ 4 \\ -8 \\ -7 \end{bmatrix} - \alpha_3 \begin{bmatrix} 7 \\ -4 \\ 11 \\ 9 \end{bmatrix}$$

$$4\alpha_2 + 4\alpha_3 = 0.0194 \quad | \cdot 5$$
$$-5\alpha_2 + 7\alpha_3 = -0.029 \quad | \cdot 4$$

$$\left.\begin{aligned} 20\alpha_2 + 20\alpha_3 &= 0.097 \\ -20\alpha_2 - 28\alpha_3 &= -0.116 \end{aligned}\right\}$$

$$\alpha_3 = 2.375 \cdot 10^{-3}$$
$$\alpha_2 = 2.475 \cdot 10^{-3} \qquad \alpha_2 > \alpha_3 \qquad \alpha_2 = 0.0025$$

B

2.3. $n = 3$

Potporni vektori

$((2, -5, 15), -1)$ $\quad \alpha_1 = 0.5$

$((1, 8, -305), -1)$ $\quad \alpha_2 = 0.8$

$((1, -6, 225), +1)$ $\quad \alpha_3 = 0.9$

$\vec{\mu} = [15 \quad -2 \quad 100]$

$\vec{\sigma} = [4 \quad 1 \quad 12]$ $\qquad \vec{x} = (1, 2, -30)$

$\Delta h(\vec{x}) = ?$

$$h(\vec{x}) = w_0 + \sum \alpha_i y^i (\vec{x})^T \vec{x}^i$$

$$w_0 = \frac{1}{|S|} \sum_{i \in S} y^i - \sum_{j \in S} \alpha_j y_j (\vec{x}^i)^T \vec{x}^j$$

potporne vektore! TREBA SKALIRATI!

$$= \frac{1}{3}\left[-1 - \left(0.5 \cdot (-1) \begin{bmatrix} 2 \\ -5 \\ 15 \end{bmatrix} [2 \; -5 \; 15] + 0.8 \cdot (-1) \begin{bmatrix} 2 \\ -5 \\ 15 \end{bmatrix} [1 \; 8 \; -305] + 0.9 \cdot 1 \begin{bmatrix} 2 \\ -5 \\ 15 \end{bmatrix} [1 \; -6 \; 225]\right)\right.$$

$$-1 - \left(0.5 \cdot (-1) \begin{bmatrix} 1 \\ 8 \\ -305 \end{bmatrix} [2 \; -5 \; 15] + 0.8 \cdot (-1) \begin{bmatrix} 1 \\ 8 \\ -305 \end{bmatrix} [1 \; 8 \; -305] + 0.9 \cdot 1 \begin{bmatrix} 1 \\ 8 \\ -305 \end{bmatrix} [1 \; -6 \; 225]\right)$$

$$\left. +1 - \left(0.5(-1) \begin{bmatrix} 1 \\ -6 \\ 225 \end{bmatrix} [2 \; -5 \; 15] + 0.8(-1) \begin{bmatrix} 1 \\ -6 \\ 225 \end{bmatrix} [1 \; 8 \; -305] + 0.9 \cdot 1 \begin{bmatrix} 1 \\ -6 \\ 225 \end{bmatrix} [1 \; -6 \; 225]\right)\right]$$

$$= \ldots = 418.826$$

$\vec{x}_{NESK.} = [1 \quad 2 \quad -30]$

$\vec{x}_{SKAL.} = [-3.5 \quad 4 \quad -10.\dot{8}\dot{3}]$

$h(\vec{x}_{NESK}) = w_0 + \sum \alpha_i y^i (\vec{x})^T \vec{x}^i = \ldots = 797.605$

$h(\vec{x}_{SKAL.}) = w_0 + \sum \alpha_i y^i (\vec{x})^T \vec{x}^i = \ldots = -58.019$

$\Delta h(\vec{x}) = h(\vec{x}_{SKAL}) - h(\vec{x}_{SKAL.}) = -739.13$

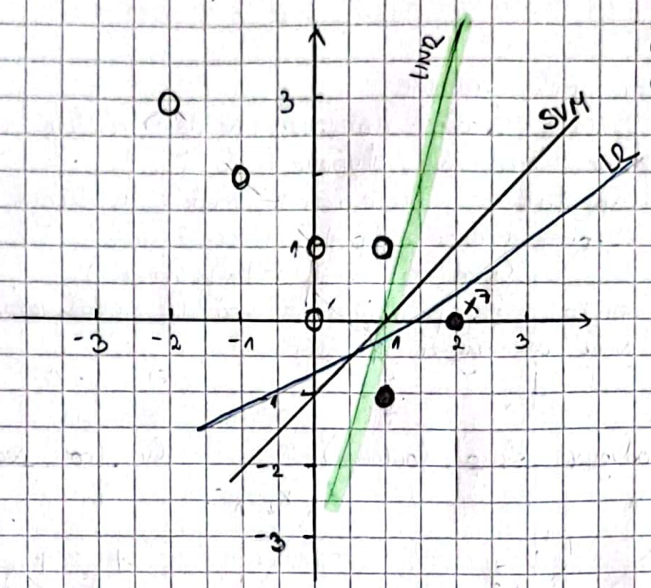
Nešt s brojevima nešto ima

D

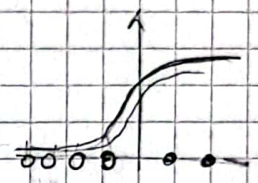
2.4.

LR - log. reg.
SVM
LINR - $h(\vec{x}) = 1\{\vec{w}^T \vec{x} \geq 0\}$

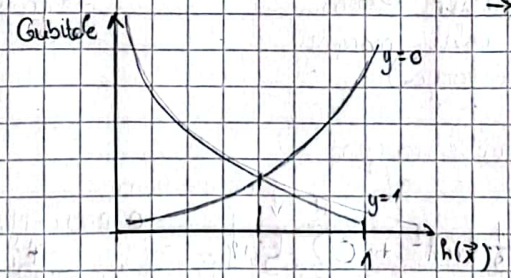
$\vec{x}^* = (2, 0)$



O : $y = 0$ ili $y = -1$
● : $y = 1$



- SVM - maksimizira minimalnu udaljenost
- LINR - kažnjava vrlo točno klasificirane primjere -> prikloniji primjerima klase $y = 0$
- LR - želi minimizirati gubitak
  (udaljava se od klase gdje je više primjera)
  -> naklonjeniji 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 robusnost klasifikacijske granice.***] Raspolažemo sljedećim primjerima za učenje:

$$\mathcal{D} = \left\{(\mathbf{x}^{(i)}, y^{(i)})\right\}_i = \left\{\big((1,1),0\big), \big((0,2),0\big), \big((2,3),0\big), \big((3,1),1\big), \big((4,3),1\big)\right\}.$$

   (a) Skicirajte funkcije gubitka L kao funkcije od $y\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x})$ za gubitak (1) linearne regresije, (2) perceptrona, (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 $\big((8,1),1\big)$. 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 dvodimenzijskome 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 ulaznome 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 trodimenzijskome 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 linearnom 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

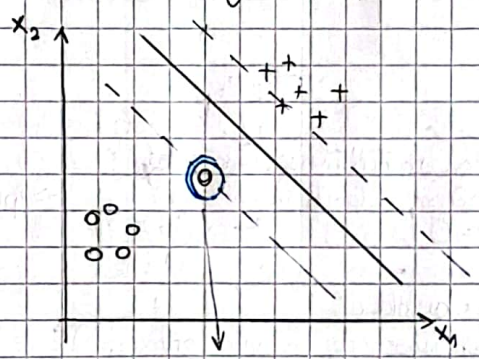
1.1.

a) Motivacija za uvođenje meke margine

- SVM s tvrdom marginom ne daje nikakvo rješenje za linearno neodvojive skupove
- sprječavamo prenaučenost modela u slučaju loda u laznim podacima postoji mnogo šuma (šira margina - bolja gener.)
- sprječavamo prenaučenost modela uzrokovanu preslikavanjem u veći dimenzijski prostor

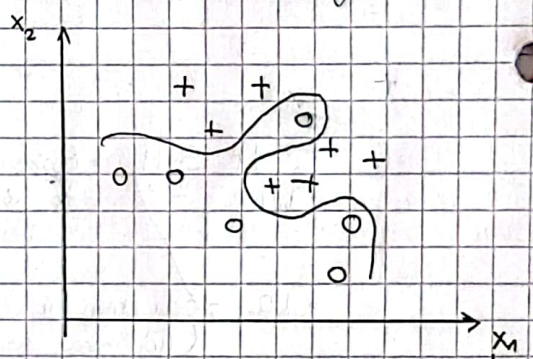
Prenaučenost.

linearno odvojiv skup (outliers)



outlier => šum
-> uzrokuje usku marginu
(loša generalizacija)

linearno neodvojiv skup



-> presložen model (prenaučen!)

b) Problem optimizacije meke margine

$$\left[\ \underset{\vec{w},\, w_0}{\arg\min} \left\{ \frac{1}{2}\|\vec{w}\|^2 + C\sum_{i=1}^{N} \xi_i \right\} \quad \text{OGRANIČENJA} \quad \begin{cases} \xi_i \geq 0 \\ y^i(\vec{w}^T\vec{x}^i + w_0) \geq 1 - \xi_i \end{cases} \Big\}_{i=1,\dots,N} \right]$$

(ξ_i - ulazak u marg.)

↳ $C = \frac{1}{\lambda} > 0$ => C određuje kompromis između veličine margine i greške
=> veći C = veća složenost modela i veća greška

c) Dualni kvadr. problem meke margine:

$$\left[\ \underset{\alpha}{\arg\max} \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i\alpha_j y^i y^j (\vec{x}^i)^T\vec{x}^j \ \right]$$

isto kao za tvrdu ALI RAZLIČ. OGRANIČ.

Ograničenja
① $0 \leq \alpha_i \leq C$
② $\sum_{i=1}^{N} \alpha_i y^i = 0$

d)

-> Potporni vektori
$0 < \alpha_i \leq C$
$\alpha_i = C$ unutar margine

d) BIL. 3, SVM II : Vrijedi $\underbrace{\alpha_i = C - \beta_i}_{\frac{\partial L}{\partial \xi_i} = 0}$ i $\underbrace{\beta_i \xi_i = 0}_{\text{komp. labav.}}$

$\alpha_i < C \Rightarrow \beta_i > 0 \Rightarrow \xi_i = 0$
$\alpha_i = C \Rightarrow \beta_i = 0 \Rightarrow \xi_i > 0$

12.

a)

- ispravna strana margine

$$y^i h(\vec{x}^i) \geq 1 \quad \leftarrow \text{ove primjere ne kažnjavamo}$$

- primjere unutar margine ili sa pogrešne strane margine kažnjavamo s $\xi_i = |y^i - h(\vec{x}^i)|$

$$= 1 - y^i h(\vec{x}^i)$$

$\Rightarrow$ funkcija gubitka prema tome je

$$L(y, h(\vec{x})) = \max(0, 1 - y h(\vec{x}))$$

b)

$$E(\vec{w} \mid D) = \sum_{i=1}^{N} \max(0, 1 - y^i h(\vec{x}^i)) + \frac{\lambda}{2} \|\vec{w}\|^2$$

$\rightarrow$ empirijska pogreška SVM-a je L2-regularizirano očekivanje gubitka zglobnice

- prim. opt. problem meke margine

$$\underset{\vec{w}, w_0, \vec{\xi}}{\operatorname{argmin}} \left\{ \underbrace{\frac{1}{2}\|\vec{w}\|^2}_{\text{odgovara regularizacijskom izrazu}} + \underbrace{C \sum_{i=1}^{N} \xi_i}_{\text{odgovara funkciji gubitka}} \right\}$$

$$\lambda = \frac{1}{C}$$

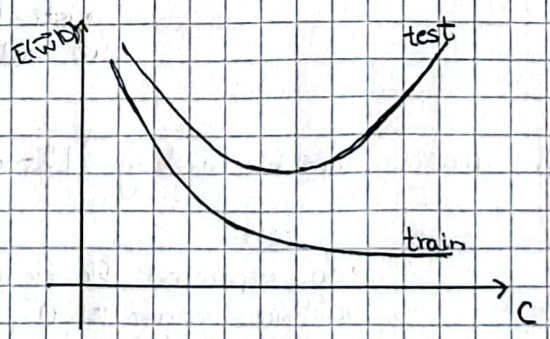
- manji λ daje manje regularizirani model, tj. složeniji model
- veći C daje složeniji model budući da više kažnjavamo netočne klasifikacije / ulaske u marginu

c)

$$\vec{x}^7 = [5 \quad 6] \qquad y = -1 \qquad (\text{V08, 2. zadatak})$$
$$h(\vec{x}^7) = 1.5$$

$$L(\vec{y}, h(\vec{x}^7)) = \max(0, 1 - \vec{y} h(\vec{x})) = \max(0, 1 - (-1) \cdot 1.5)$$
$$= 2.5$$

d)

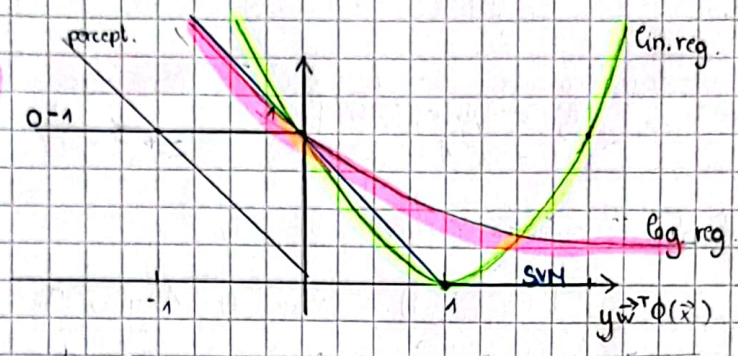


- veći $C \Rightarrow$ složeniji model

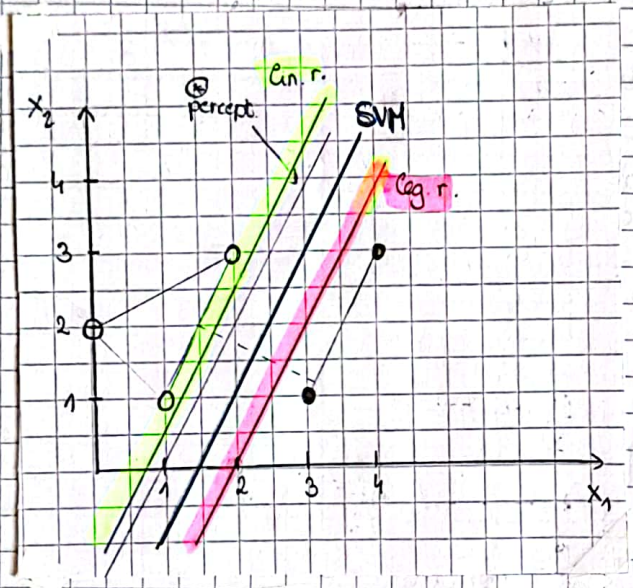
1.3. → pogledaj u ogledni DZ

$D = \{(\vec{x}^i, y^i)\} =$

a)



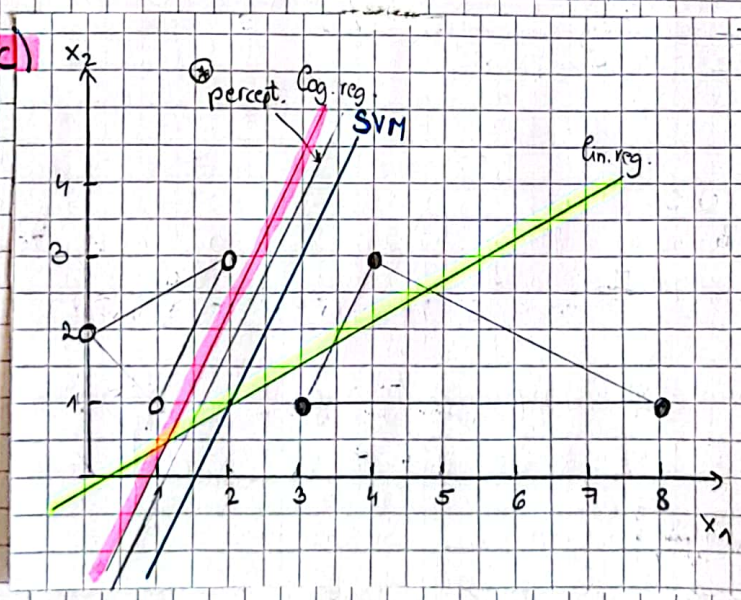
b)



○ : $y = 0$
● : $y = 1$

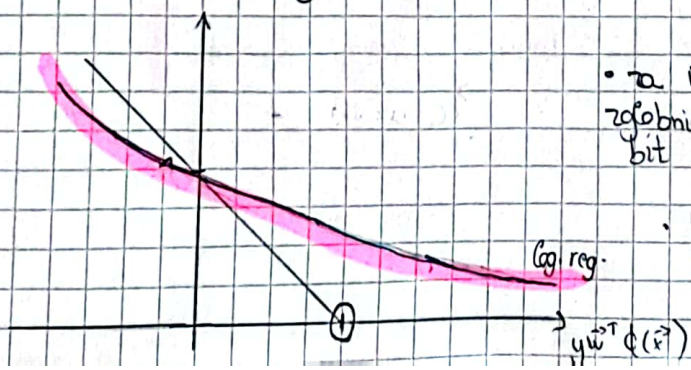
(*) perceptron
- ovisi o početnim težinama
- može biti bilo koja hipoteza koja savršeno klasificira D

c)



- SVM - ostaje isti (isti potp. vekt.)
- log. reg.
  - bit će pomaknuta u lijevo (podjednak je broj primjera dvije klase)
  - želi min. gubitak outliera ⇒ ako je on daleko od granice
- lin reg.
  - prikloniji granica klasi $y = 1$ zbog jake stršeće vrijednosti (8,1), 1)
- isti komentar za perceptron kao pod b)

d) SVM daje rijetke modele uvođenjem L2-regularizacije



- za ispravno klasificirane primjere fja gubitka jednaka je 0 ⇒ više težina bit će pritegnuto na 0
- kod logist. regresije i ispravno klasificirani primjeri donose grešku ⇒ neće rezultirati rijetkim modelom.

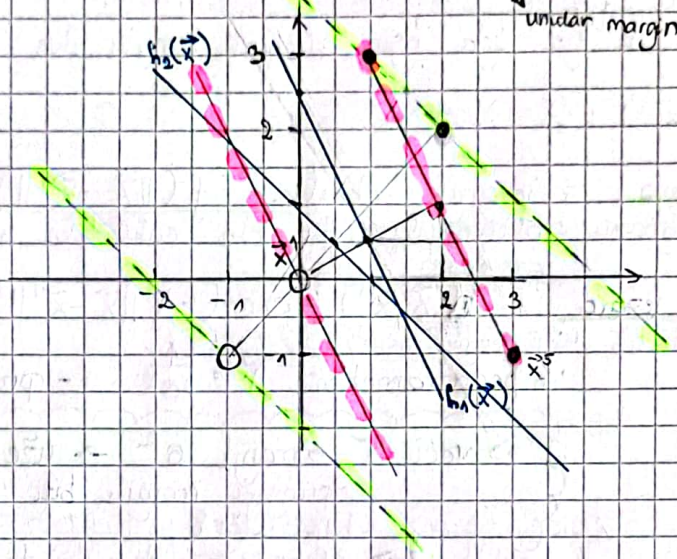
II Zadaci s ispita

2.1.

SVM_1 - tvrda margina $\Rightarrow h_1(\vec{x})$

SVM_2 $C = 1$ $\alpha_1 = \alpha_5 = 1$ $\alpha_2, \alpha_3, \alpha_4 > 0$ meka margina $\Rightarrow h_2(\vec{x})$

unutar margine



$$marg_{tvr} = \frac{1}{2}\sqrt{2^2 + 1^2} = \frac{\sqrt{5}}{2}$$

$$marg_{mek} = \frac{1}{2}\sqrt{3^2 + 3^2} = \frac{\sqrt{18}}{2} = \frac{3\sqrt{2}}{2}$$

$$\frac{marg_{mek}}{marg_{tvrd}} = \frac{3\sqrt{2}}{\sqrt{5}} = \frac{3}{5}\sqrt{10} \Rightarrow B$$

2.2.

$C = 0.01 \rightarrow \lambda = \frac{1}{C} = 100$

$\vec{\alpha} = [0 \quad 0.01 \quad 0.01]$

$w_0 = -0.8$

bez w_0

$$E(\vec{w} \mid D) = \sum_{i=1}^{N} \max(0, 1 - y^i h(\vec{x}^i)) + \frac{\lambda}{2}\|\vec{w}\|^2$$

$$\vec{w} = \sum \alpha_i y^i \vec{x}^i = 0.01 \cdot (-1)\begin{bmatrix} -4 \\ 4 \\ 4 \end{bmatrix} + 0.01 \cdot 1 \cdot \begin{bmatrix} -2 \\ 4 \\ 1 \end{bmatrix}$$

$$= \begin{bmatrix} 0.02 \\ 0 \\ -0.03 \end{bmatrix}$$

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

$h(\vec{x}) = [-1 \quad -1 \quad -0.87]$

$$E(\vec{w} \mid D) = \max(0, 1 - (-1)\cdot(-1)) + \max(0, 1 - (-1)(-1)) + \max(0, 1 - 1 \cdot (-0.87)) + 50 \cdot 0.0013$$

$$= 1.87 + 0.065 = 1.935 \Rightarrow A$$

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 jezgrenu funkciju, RBF-jezgru i Gaussovu 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 kojemu biste koristili tu jezgru umjesto Gaussove jezgre.

   (d) Definirajte jezgreni stroj i rijetki jezgreni (vektorski) stroj. Koji od njih je parametarski a koji neparametarski algoritam i što to znači?

2. [***Svrha:** Isprobati preslikavanje primjera u prostor značajki primjenom Gaussovih baznih funkcija. Razumjeti kako preslikavanje utječe na broj parametara i hiperparamera modela.*] Raspolažemo skupom primjera:

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

   (a) U ulaznome prostoru skicirajte diskriminacijsku granicu $h(\mathbf{x}) = 0$ koju biste dobili logističkom regresijom uz $\boldsymbol{\phi}(\mathbf{x}) = (1, \mathbf{x})$, tj. bez preslikavanja (izračun nije potreban).

   (b) Na isti skup primjera primijenite jezgreni stroj s baznim funkcijama:

$$\phi_j(\mathbf{x}) = \exp\Big(-\frac{\|\mathbf{x} - \boldsymbol{\mu}_j\|^2}{2\sigma_j^2}\Big).$$

   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 jezgreni 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 jezgreni trik kod SVM-a.*]

   (a) Za klasifikaciju primjera u ulaznome prostoru $X = \mathbb{R}^2$ koristimo polinomijalnu jezgrenu funkciju $\kappa(\mathbf{x}, \mathbf{z}) = (\mathbf{x}^{\mathrm{T}}\mathbf{z} + 1)^2$. Pokažite da je za $n = 2$ jezgra κ Mercerova jezgra. Zašto je to bitno?

   (b) Izvedite pripadno preslikavanje $\boldsymbol{\phi}(\mathbf{x})$ za $n = 2$. U koji će vektor u prostoru značajki efektivno biti preslikan primjer $\mathbf{x} = (2,3)$ primjenom jezgre $\kappa(\mathbf{x}, \mathbf{z}) = (\mathbf{x}^{\mathrm{T}}\mathbf{z} + 1)^2$?

   (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 jezgrenu funkciju $\kappa(\mathbf{x}, \mathbf{z}) = (\mathbf{x}^{\mathrm{T}}\mathbf{z})^2$?

4. [***Svrha: Izvježbati izračun predikcije pomoću jezgrenog trika.***] Veza između primarnih i dualnih parametara SVM-a jest $\mathbf{w} = \sum_{i=1}^{N} \alpha_i y^{(i)} \boldsymbol{\phi}(\mathbf{x}^{(i)})$. Na skupu za učenje trenirali smo SVM s polinomijalnom jezgrenom funkcijom, $\kappa(\mathbf{x}, \mathbf{z}) = (\mathbf{x}^{\mathrm{T}}\mathbf{z} + 1)^3$. 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 jezgreni trik te odredite klasifikaciju primjera $\mathbf{x}^{(4)} = (1, 2, 3)$.

5. [***Svrha: Razumjeti karakteristike Gaussove jezgre.***]

   (a) Primjenom operacija za izgradnju složenijih Mercerovih jezgri iz jednostavnijih Mercerovih jezgri, dokažite da je Gaussova jezgra Mercerova jezgra. (Pomoć: raspišite izraz $\|\mathbf{x} - \mathbf{x}'\|^2$.)

   (b) 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?

   (c) Skicirajte očekivana područja prenaučenosti i podnaučenosti modela SVM u prostoru hiperparametara $C \times \gamma$.

   (d) Koristimo Gaussovu jezgru uz $\gamma = 1$. Možemo li u ovom slučaju odrediti u koji vektor $\boldsymbol{\phi}(\mathbf{x})$ 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 beskonačnodimenzijski 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 jezgreni 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 dvodimenzijskome ulaznom prostoru:

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

   Na ovom skupu treniramo jezgreni 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 $\boldsymbol{\phi}(\mathbf{x}^{(3)}) = (1, 0.1, 0.2)$. 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:

$$\begin{aligned}\mathcal{D} =& \{(\mathbf{x}, y)\}_i \\ =& \big\{(\text{water}, 1), (\text{eau}, 0), (\text{dog}, 1), (\text{chien}, 0), (\text{paperclip}, 1), (\text{trombone}), 0), (\text{chance}, 1), (\text{hasard}, 0)\big\}\end{aligned}$$

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 $\boldsymbol{\phi}$ takva da $\kappa(\mathbf{x}, \mathbf{z}) = \boldsymbol{\phi}(\mathbf{x})^T \boldsymbol{\phi}(\mathbf{z})$. Konkretno, u slučaju dvodimenzijskoga ulaznog prostora ($n = 2$), ova jezgra odgovara preslikavanju u šesterodimenzijski prostor. Međutim, postoji odstupanje u konkretnim koeficijentima polinoma. Razmotrite primjer $\mathbf{x} = (1, 0)$ te izračunajte $\boldsymbol{\phi}_\kappa(\mathbf{x})$, koji dobivamo preslikavanjem definiranim implicitno preko jezgre, te $\boldsymbol{\phi}_p(\mathbf{x})$, koji dobivamo preslikavanjem definiranim kao polinom drugog stupnja. **Koliko iznosi euklidska udaljenost između $\boldsymbol{\phi}_\kappa(\mathbf{x})$ i $\boldsymbol{\phi}_p(\mathbf{x})$?**

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 zglobnice 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 $\boldsymbol{\alpha} = (0, 0.754, 0.754, 1, 1)$. **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 regularizacijski 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 - Jezgrene metode

I Zadaci za učenje

1.1. a) Jezgrena fja
= fja koja mjeri sličnost između dva primjera

$K(\vec{x}, \vec{x}') : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$

RBF jezgra = jezgra oblika $f(\|\vec{x} - \vec{x}'\|)$ koja ima omotanu aktivacijsku fju ob euklidske udalj. 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

=> veći γ -> manji σ^2 -> uže Gaussovo zvono
- primjeri moraju biti bliže da bi bili slični
  - ↳ Lako može dovesti do prenaučenosti kod SVM-a -> pretraživanje po rešetci (odabir parametara γ i λ)
  - ↳ $\gamma \to \infty$ => $K(\vec{x}, \vec{x}') \to 0$
    - svaki primjer u svojoj dimenziji + primjeri međusobno ortogonalni => SAVRŠENA LINEARNA ODVOJIVOST

b) RBF jezgre osjetljive su na razlike u skalama značajki jer koriste euklidsku udaljenost koja sve značajke tretira jednako važnima

c) Mahalanobisova udaljenost = poopćenje euklidske udaljenosti

RBF jezgra -> $\left[K(\vec{x}, \vec{x}') = \exp\left(-\frac{1}{2}(\vec{x} - \vec{x}')^T \Sigma^{-1} (\vec{x} - \vec{x}')\right) \right]$

Σ = kovarijacijska matrica između značajki

=> rješava problem osjetljivosti na skale :

Pr. Kada bi u ulaznom skupu imali značajku prihodi i dob
=> ako su sličnost i dob snažno korelirane u izračunu će ta razlika imati manju težinu nego razlika primjera u drugim značajkama

d) Jezgreni stroj

↳ poopćeni linearni model koji za preslikavanje koristi jezgrene funkcije

$$\phi(\vec{x}) = (1, K(\vec{x}, \vec{\mu}_1), K(\vec{x}, \vec{\mu}_2), \dots, K(\vec{x}, \vec{\mu}_m))$$

u smislu dimenzion.

$\vec{\mu}_i$ - unaprijed odabrani referentni primjeri iz prostora primjera

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

• rijetki jezgreni strojevi

⇒ kao referentne primjere $\vec{\mu}_j$ uzima se podskup ulaznih primjera

$$\phi(\vec{x}) = (1, K(\vec{x}, \vec{x}^1), \dots, K(\vec{x}, \vec{x}^N))$$

↳ prototipovi

lako se može prenaučiti ako je primjera previše $m = N$ → L1 regul.

⇒ rijetki jezgreni strojevi ovise o broju primjera u ulaznom skupu (N)

↳ rijetki jezgreni strojevi su neparametarski algoritam

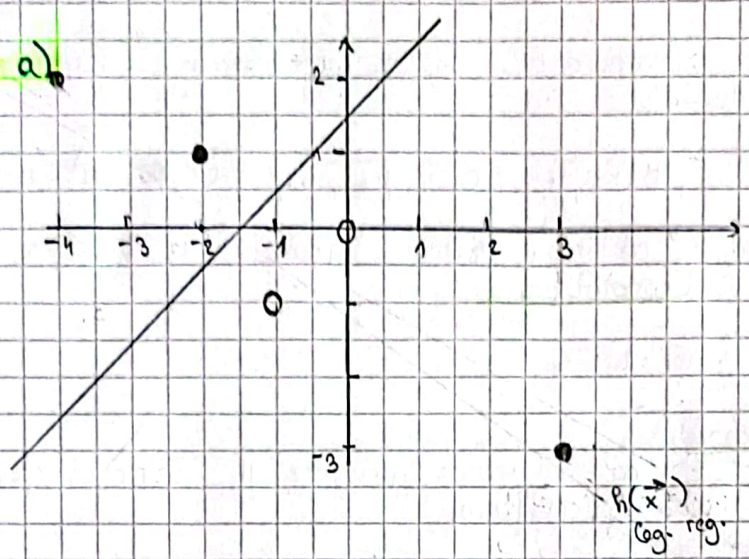
↳ jezgreni strojevi su paramet. algoritam

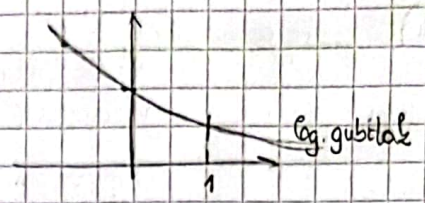
• neparametarski algoritam
- broj parametara ~ N
- ne pretpostavlja se model podataka
- lokal. aproksim. hipoteze ako pohranj. primjera

• parametarski algoritam
- broj parametra NE OVISI o N
- pretpostavlja distribucije podataka
- primjeri → globalno utječu na izgled hipoteze

1.2.

a)





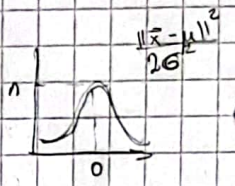
• skup je linearno neodvojiv

b)

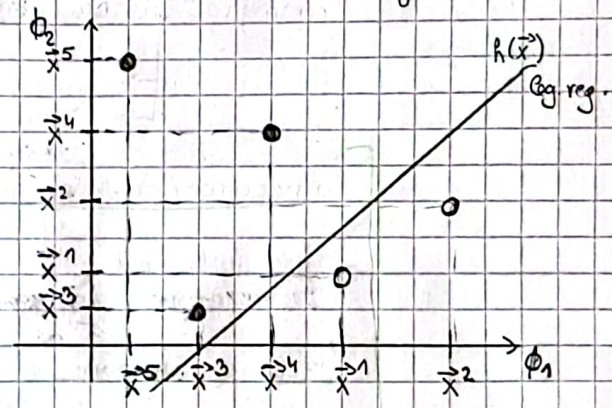
$\vec{\mu_1} = (0,0) \quad \sigma_1 = \sigma_2 = 1$

$\vec{\mu_2} = (-3,3)$

$$\phi_j = \exp\left(-\frac{1}{2\sigma_j^2}\|\vec{x}-\vec{\mu_j}\|^2\right) = \exp\left(-\frac{1}{2}\|\vec{x}-\vec{\mu_j}\|^2\right)$$



	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^{-36}
$\vec{x}^4$	1	-2	1	$e^{-5/2}$	$e^{-5/2}$
$\vec{x}^5$	1	-4	2	e^{-10}	e^{-1}



c)

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

• parametri : $\vec{w} = [w_0 \quad w_1 \quad w_2] \rightarrow 3$

• hiperparametri : $\underset{\sigma_1, \sigma_2}{\vec{\mu}} = [[\mu_1^1 \quad \mu_2^1] \quad [\mu_1^2 \quad \mu_2^2]]$ } 6 (ako su $\sigma_1 = \sigma_2$ uvijek onda 5)

\#param = 3
\#hiperparam. = 6

• u praksi vrijednosti hiperparametara određujemo unakrsnom provjerom

• u ovom slučaju odabir (hiperparametra)
- $\vec{\mu}$ ne određuje složenost modela
- $\vec{\sigma}$ određuje složenost
  → uže zvono ⇒ primjeri moraju biti bliže da budu sličniji → složeniji model

1.3.

a)

$$K(\vec{x},\vec{z}) = (\vec{x}^T\vec{z}+1)^2 \qquad n=2$$
$$= (x_1z_1 + x_2z_2 + 1)^2$$
$$= x_1^2z_1^2 + x_2^2z_2^2 + 1 + 2x_1x_2z_1z_2 + 2x_1z_1 + 2x_2z_2 +$$
$$= \begin{bmatrix} 1 \\ x_1^2 \\ x_2^2 \\ x_1x_2\sqrt{2} \\ x_1\sqrt{2} \\ x_2\sqrt{2} \end{bmatrix} \begin{bmatrix} 1 & z_1^2 & z_2^2 & z_1z_2\sqrt{2} & z_1\sqrt{2} & z_2\sqrt{2} \end{bmatrix}$$

dummy

$$= \phi(\vec{x})^T\phi(\vec{z})$$

uz defin. $\phi(\vec{a}) = (1, a_1^2, a_2^2, a_1a_2\sqrt{2}, a_1\sqrt{2}, a_2\sqrt{2})$

$\Rightarrow$ $K(\vec{x},\vec{z})$ odgovara skalarnom umnošku u 6-dim. prostoru značajki $\rightarrow$ K jest Mercerova jezgra

$\Rightarrow$ ovo svojstvo je bitno jer nam omogućuje da upotrijebimo jezgreni trik, tj. iskoristimo Mercerovu jezgru umjesto skalarnog umnoška $\phi(\vec{x})^T\phi(\vec{x})$ kod SVM-a

b) $\vec{x} = [2\ 3]$

$\hookrightarrow$ primjenom 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 dimenzijski prostor broj parametara neparametarske inačice ovog modela biti će veći od broja parametara parametarske inačice za $N > 5$

param. $\vec{w}^T\phi(\vec{x}) + w_0$

neparam. $\sum \alpha_i y^i K(\vec{x}, \vec{x}^i) + w_0$

d) XOR problem

$\vec{x}^i$	y	$\phi(\vec{x}^i) = (1, x_1^2, x_2^2, x_1x_2\sqrt{2}, x_1\sqrt{2}, x_2\sqrt{2})$
(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, $\sqrt{2}$, $\sqrt{2}$, $\sqrt{2}$, $\sqrt{2}$)
		ϕ_1 ϕ_2

$\rightarrow$ XOR je linearno odvojiv uporabom pripadnog preslikavanja

$\Rightarrow$ isto vrijedi i za jezgru $K(\vec{x},\vec{z}) = (\vec{x}^T\vec{z})^2$ jer je polinomijalna jezgra Mercerova jezgra

pripadno preslikavanje

$$\phi(x) = (x_1^2, x_2^2, x_1x_2\sqrt{2}),$$

1.4.

$$K(\vec{x}, \vec{z}) = (\vec{x}^T \vec{z} + 1)^3$$

$$\vec{w} = \sum_{i=1}^{N} \alpha_i y^i \phi(\vec{x}^i)$$

$\vec{x}^1 = (-2, 3, 5)$ $\quad \alpha_1 = 0.131$ $\quad y^1 = -1$

$\vec{x}^2 = (6, 4, 3)$ $\quad \alpha_2 = 0.048$ $\quad y^2 = 1$

$\vec{x}^3 = (8, 8, 2)$ $\quad \alpha_3 = 0.013$ $\quad y^3 = 1$

$w_0 = -0.51$

$h(\vec{x}^4 = [1\ 2\ 3]) = ?$

$$h(\vec{x}) = \sum \alpha_i y^i \underbrace{\phi(\vec{x})^T \phi(\vec{x}^i)}_{K(\vec{x}, \vec{x}^i)} + w_0$$

$$\begin{aligned} = \ & 0.131 \cdot (-1) \cdot \left([1\ 2\ 3]^T [-2\ 3\ 5] + 1 \right)^3 \\ & + 0.048 \cdot (+1) \cdot \left([1\ 2\ 3]^T [6\ 4\ 3] + 1 \right)^3 \\ & + 0.013 \cdot (+1) \cdot \left([1\ 2\ 3]^T [8\ 8\ 2] + 1 \right)^3 \\ & - 0.51 \end{aligned}$$

$$\begin{aligned} h(\vec{x}) &= -1048 + 663.552 + 387.283 - 0.51 \\ &= -2.325 \quad // \end{aligned}$$

1.5. a) Gaussova jezgra je Mercerova jezgra

Dokaz

$$\begin{aligned} K(\vec{x}, \vec{y}) &= \exp(-\gamma \|\vec{x} - \vec{y}\|^2) \\ &= \exp(-\gamma (\|\vec{x}\|^2 - 2\vec{x}^T\vec{y} + \|\vec{y}\|^2)) \\ &= \underbrace{\exp(-\gamma \|\vec{x}\|^2)}_{f(\vec{x})} \underbrace{\exp(2\gamma \vec{x}^T \vec{y})}_{\exp(K_1)} \underbrace{\exp(-\gamma \|\vec{y}\|^2)}_{f(\vec{y})} \end{aligned}$$

vidi str. 12 →

Treba vidjet je li $K_1(\vec{x}, \vec{y}) = 2\gamma \vec{x}^T \vec{y}$ Mercerova jezgra

$$K_1(\vec{x}, \vec{y}) = 2\gamma \vec{x}^T \vec{y} = \vec{x}^T \sqrt{2\gamma} \cdot \vec{y} \sqrt{2\gamma} = \phi(\vec{x})^T \phi(\vec{y}) \quad \blacksquare$$

b) $\gamma = \frac{1}{2\sigma^2}$ ⇒ veća preciznost (γ) uzrokuje manji σ^2, tj. uže Gaussovo zvono

⇒ primjeri će onda biti različitiji, tj. $K(\vec{x}, \vec{x}') \to 0$ što odgovara povećanju složenosti modela

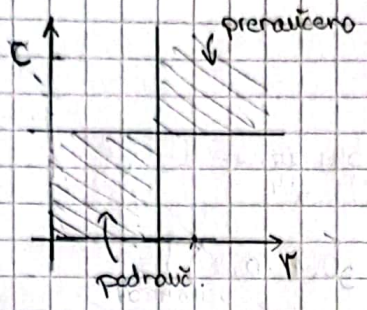
γ raste ⇒ raste složenost

$C = \frac{1}{\lambda}$ ⇒ porastom kazne (C) manje regulariziramo model, tj. složenost raste

C raste ⇒ raste složenost

- Ako odaberemo veliku preciznost moramo odabrati manji C kako bi smanjili složenost modela.

c)



d) $\gamma = 1$

$$K(\vec{x}, \vec{x}') = \exp(-\gamma \|\vec{x} - \vec{x}'\|^2)$$
$$= \exp(-\|\vec{x} - \vec{x}'\|^2)$$

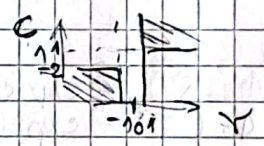
Ne možemo odrediti preslikavanje $\phi(\vec{x})$ jer je ono definirano implicitno.
Posljedično ne možemo odrediti ni težine jer su one dane s $\vec{w} = \sum \alpha_i y^i \phi(\vec{x}^i)$ „

e) Gaussova jezgra jamči visoku dimenzionalnost
- Grammova matrica K za Gauss. jezgru je punog ranga
$\Rightarrow \phi(\vec{x})$ su linearno nezavisni $\Rightarrow$ prostor značajki je beskonačno dim. (dim. kolko i primjera)

- jamči i savršenu linearnu odvojivost
  - $\gamma \to \infty \Rightarrow K(\vec{x}, \vec{x}') \to 0$ → primjeri ortogonalni i normirani
  - svaki primjer = 1 vrh (os) u prostoru značajki
- empirijska pogreška = 0 na skupu za učenje
  → zbog savrš. lin. odvojivosti

II Zadaci s ispita

2.1.



$N = 1000$
$n = 100$
nakon treniranja 28 prototipa

\# param. rač. mod = 28 prototipa, α, w_0 = $28 \cdot 100 + 28 \cdot \alpha + 1 = 2829$
\# hiperparam = ? = 1
\# kolik param. opt. = ? = $1000 \cdot \alpha + w_0 = 1001$ (tražimo potp. vektore)

$$h(\vec{x}) = w_0 + \sum_{i=1}^{N} \alpha_i y^i K(\vec{x}, \vec{x}^i)$$

→ $\exp(-\gamma \|\vec{x} - \vec{x}^i\|^2)$

→ sve Gauss. imaju isti γ / σ^2 ⇒ 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

→ Log. regresija

prototipovi: $\vec{x}^1$ $\vec{x}^4$

$$\left\{ \phi(\vec{x}^3) = (1 \;\; 0.1 \;\; 0.2) \right\}$$

$$\vec{w} = [0.2 \;\; 1 \;\; -1]$$

Točnost = ?

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

$$\phi(\vec{x}) = (1 \;\; \phi_1(\vec{x}) \;\; \phi_2(\vec{x}))$$

$$\phi_1(\vec{x}) = K(\vec{x}, \vec{x}^1) = \exp(-\gamma_1 \|\vec{x} - \vec{x}^1\|^2)$$

$$\phi_2(\vec{x}) = K(\vec{x}, \vec{x}^2) = \exp(-\gamma_2 \|\vec{x} - \vec{x}^4\|^2)$$

$$\phi_1(\vec{x}^3) = \exp\left(-\gamma_1 \left\|\begin{bmatrix}2\\3\end{bmatrix} - \begin{bmatrix}1\\1\end{bmatrix}\right\|^2\right) = 0.1 \quad / \ln$$

$$-\gamma_1 \left\|\begin{bmatrix}1\\2\end{bmatrix}\right\|^2 = \ln 0.1$$

$$-\gamma_1 \cdot 5 = \ln 0.1 \quad \Rightarrow \boxed{\gamma_1 = 0.46}$$

$$\phi_2(\vec{x}^3) = \exp\left(-\gamma_2 \left\|\begin{bmatrix}2\\3\end{bmatrix} - \begin{bmatrix}3\\4\end{bmatrix}\right\|^2\right) = 0.2 \quad / \ln$$

$$-\gamma_2 \cdot 2 = \ln 0.2 \quad \Rightarrow \boxed{\gamma_2 = 0.805}$$

$$\phi_1(\vec{x}) = \exp\left(-0.46 \left\|\vec{x} - \begin{bmatrix}1\\1\end{bmatrix}\right\|^2\right)$$

$$\phi_2(\vec{x}) = \exp\left(-0.805 \left\|\vec{x} - \begin{bmatrix}3\\4\end{bmatrix}\right\|^2\right)$$

$$h(\vec{x}^1) = \frac{1}{1 + \exp\left(-\begin{bmatrix}0.2\\-1\end{bmatrix}\begin{bmatrix}1 & \underset{\phi_1}{1} & \underset{\phi_2}{0}\end{bmatrix}\right)} = 0.7658 \geq 0.5 \Rightarrow 1 \checkmark$$

$$h(\vec{x}^2) = \ldots = 0.5789 \Rightarrow 1 \times$$

$$h(\vec{x}^3) = \ldots = 0.5249 \Rightarrow 1 \times$$

$$h(\vec{x}^4) = \ldots = 0.31 \Rightarrow 0$$

Točnost: $\frac{1}{2}$

C

2.3.

eng. $y=1$
franc. $y=0$

$$\vec{w} = [\underset{\uparrow\, w_0}{0.5}\ 0\ 0\ 0\ -3.5\ 0\ 1\ 0\ -1]$$

$$K(\vec{x}_1, \vec{x}_2) = \frac{|\vec{x}_1 \cap \vec{x}_2|}{|\vec{x}_1 \cup \vec{x}_2|}$$ → ne pon. slova

$\vec{x}^i$	y^i	$K(\vec{x}, \vec{x}^i)$	
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, e, r, c, l, i}
* TROMBONE	0	3/10	{t, r, o, m, b, n, e}
CHANCE	1	1/10	{c, h, a, n, e}
* HASARD	0	2/9	{h, a, s, r, d}

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

$$\Phi_j(\vec{x}) = K(\vec{x}, \vec{x}^j)$$

$$\Phi_0(\vec{x}) = 1$$

$$\Phi(\vec{x}) = [1\ \tfrac{1}{9}\ \tfrac{1}{7}\ \tfrac{1}{7}\ \tfrac{1}{9}\ \tfrac{1}{13}\ \tfrac{3}{10}\ \tfrac{1}{10}\ \tfrac{2}{9}]$$

• logistička regres. s kern.

$\vec{x}$ = rouncurs = {n, o, u, r, s}
$y = 0$

$$h(\vec{x}) = \sigma(-\vec{w}^T \Phi(\vec{x})) = \sigma\left(-\frac{13}{90}\right) = 0.547$$

$$L(y, h(\vec{x})) = -\ln(1 - h(\vec{x})) = 0.792$$

$\boxed{\text{UNIJA - sve (s ponavljanjem) - presjek}}$ → D

2.4.

D

$$K(\vec{x}, \vec{z}) = (\vec{x}^T \vec{z})^2 = (x_1 z_1 + x_2 z_2)^2 = x_1^2 z_1^2 + x_2^2 z_2^2 + 2 x_1 x_2 z_1 z_2$$

2dim → 6 dim

$$\Phi_K(\vec{x}) = (x_1^2, x_2^2, x_1 x_2 \sqrt{2})$$

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

pazi da se vračajke pokrivaju!

$$\vec{x} = (1, 0) \qquad \Phi_K(\vec{x}) = (1, 0, 0) \Rightarrow (\overset{1}{0}, \overset{x_1}{0}, \overset{x_2}{0}, \overset{x_1x_2}{0}, \overset{x_1^2}{1}, \overset{x_2^2}{0})$$

$$\Phi_P(\vec{x}) = (1, 1, 0, 0, 1, 0)$$

$$\|\Phi_K(\vec{x}) - \Phi_P(\vec{x})\| = \sqrt{1^2 + 1^2} = \sqrt{2}$$

2.5.

dim. ulaz. $n = 4$

$$K(\vec{x}, \vec{z}) = (\vec{x}^T \vec{z} + 2)^3$$

$\vec{x}^i$	y^i	α_i
[9 30 21]	-1	$2.214 \cdot 10^{-8}$
[-11 -20 -15]	-1	$3.803 \cdot 10^{-8}$
[-1 -7 -6]	1	$6.017 \cdot 10^{-8}$

$$\vec{x} = [3, 0\ -3]$$

$$h(\vec{x}) = w_0 + \sum_{i=1}^{N} \alpha_i y^i K(\vec{x}, \vec{x}^i) = \ldots = 0.9461$$

C

$$K([3\ 0\ -3], \vec{x}^i)$$

$(-36+2)^4 = 1\,336\,336$
$(12+2)^4 = 38\,416$
$(15+2)^4 = 83\,521$

$w_0 = ?$ $\quad h(\vec{x}^1) = \vec{w}^T \vec{x}^1 + w_0 = -1$

$\vec{w} = \sum \alpha_i y^i \vec{x}^i$ $\quad w_0 = -1 - \vec{w}^T \vec{x}^1$

$$\vec{w} = \begin{bmatrix} 1.589 \cdot 10^{-7} \\ -9.061 \cdot 10^{-8} \\ -2.555 \cdot 10^{-7} \end{bmatrix} \qquad \boxed{w_0 = -1}$$

2.6. $N = 5$

SVM s Gauss.

$$\vec{y} = [+1 \quad \underline{+1} \quad -1 \quad -1 \quad \underline{+1}]$$

$$K(\vec{x}, \vec{y}) = \exp(-\gamma d^2) = \exp(-d^2 \cdot 10^{-4})$$

Udaljenosti

$$D = \begin{array}{c|ccccc} & x^1 & x^2 & \cdots & & \\ \hline x^1 & 0 & 7.48 & 6.16 & 13.42 & 12.21 \\ x^2 & & 0 & 12.73 & 20.1 & 14.81 \\ \vdots & & & 0 & 10.49 & 9.95 \\ \vdots & & & & 0 & 20.02 \\ & & & & & 0 \end{array}$$

→ simetrična matrica

$C = 10$

$\gamma = 10^{-4}$

$$\vec{\alpha} = [10 \quad \underline{1.052} \quad 10 \quad 10 \quad \underline{8.948}]$$

$0 < \alpha_i \leq C$ → potporni vektori

$\alpha_i = C$ → unutar margine!

$$L(\hat{y}^1, h(\vec{x}^1)) = \max(0, 1 - y^1 h(\vec{x}^1)) = ? = 0.241$$

B

$$h(\vec{x}^1) = w_0 + \underbrace{\sum_{i=1}^{N} \alpha_i y^i K(\vec{x}, \vec{x}^i)}_{\text{svi primjeri unutar i na margini}} = \ldots = 0.759$$

$$y^i h(\vec{x}^i) = 1$$

$$\vec{w}^T \vec{x}^i + w_0 = y^i$$

$$w_0 = y^i - \vec{w}^T \vec{x}^i$$

$$w_0 = \frac{1}{|S|} \sum_{i \in S} y^i - \underbrace{\sum_{j=1}^{N} \alpha_j y^j K(\vec{x}^j, \vec{x}^i)}_{\text{tu svi primjeri}} = \ldots = 0.68$$

tu samo oni koji su na marg.

2.7. SVM, bin. klas.

$N = 5$

$\vec{y} = [+1, +1, -1, -1, +1]$

$$K = \begin{bmatrix} 1 & 0.97 & -0.949 & -0.555 & -0.986 \\ & 1 & -0.844 & -0.336 & -0.917 \\ & & 1 & 0.789 & 0.988 \\ & & & 1 & 0.684 \\ & & & & 1 \end{bmatrix}$$

↳ simetrično!

$C = 1 \Rightarrow \lambda = \frac{1}{C} = 1$

$$\vec{\alpha} = [\,0 \quad \underbrace{0.754 \quad 0.754}_{\text{na marginu}} \quad \underbrace{1 \quad 1}_{\text{unutar margine}}\,] \qquad 0 < \alpha_i \leq C$$

$$L(y^4, h(\vec{x}^4)) = \max(0, 1 - y^4 h(\vec{x}^4)) = ?$$
$$= \max(0, 1 - (-1)\cdot(-0.973)) = 0.027 \Rightarrow \text{C}$$

$$h(\vec{x}) = w_0 + \sum_{i=1}^{N} \alpha_i y^i K(\vec{x}, \vec{x}^i)$$

$y h(\vec{x}) = 1$

$y(\sum \alpha_i y^i \vec{x}^T \vec{x}^i + w_0) = 1$

$w_0 = y - \sum$

$w_0 = ?$

$$w_0 = \frac{1}{|S|} \sum_{i \in S} y^i - \sum_{j=1}^{N} \alpha_j y^j K(\vec{x}^i, \vec{x}^j)$$

$|S| = 2 \quad i \in \{2, 3\}$

$\underline{i = 2}$

$$y^2 - \sum_{j=1}^{N} \alpha_j y^j K(\vec{x}^2, \vec{x}^j) =$$

$$= 1 - (0 \cdot ___ + 0.754 \cdot 1 \cdot 1 + 0.754 \cdot (-1) \cdot (-0.844) + 1 \cdot (-1) \cdot (-0.336) + 1 \cdot 1 \cdot (-0.917))$$

$$= 1 - 0.8098 =$$

$$= 0.1906$$

$\underline{i = 3}$

$$y^3 - \sum_{j=1}^{N} \alpha_j y^j K(\vec{x}^3, \vec{x}^j) =$$

$$= -1 - (0 \cdot ___ + 0.754 \cdot 1 \cdot (-0.844) + 0.754 \cdot (-1) \cdot 1 + 1 \cdot (-1) \cdot 0.789 + 1 \cdot 1 \cdot 0.988)$$

$$= -1 + 1.1914 = 0.1914$$

$$\Rightarrow w_0 = \frac{1}{2}(0.1906 + 0.1914) = 0.191$$

$$h(\vec{x}^4) = w_0 + (-1.16425) = -0.973$$

$$h(\vec{x}^4) = w_0 + \sum \alpha_i y^i K(\vec{x}^4, \vec{x}^i) = w_0 + (0 \cdot ___ + 0.754 \cdot 1 \cdot (-0.336) + 0.754 \cdot (-1) \cdot 0.789 + 1 \cdot (-1) \cdot 1 + 1 \cdot 1 \cdot 0.684)$$

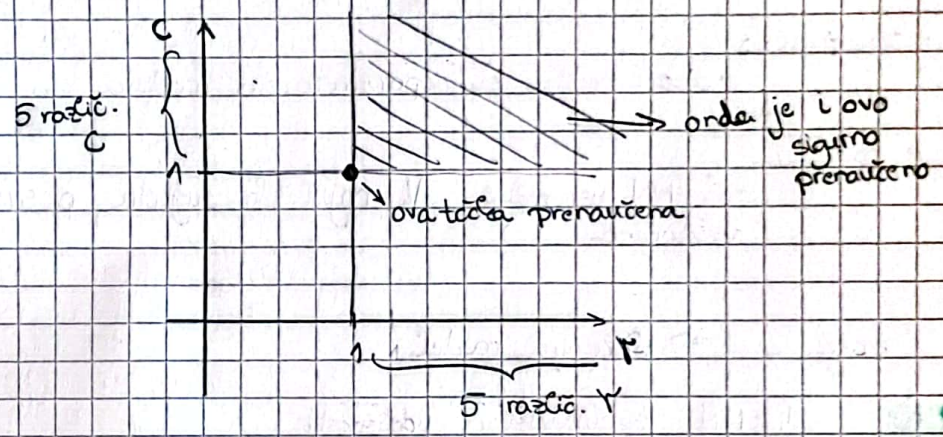
2.7. • isti kao 2.6. samo je zadana direktno jezgrena matrica, a ne euklidska udaljenost.

TIDEN. POSTUPAK RJEŠAVANJA

2.8.

$\left.\begin{matrix} C = 1 \\ \gamma = 1 \end{matrix}\right\}$ prenaučen

$\left.\begin{matrix} |C| = 11 \\ |\gamma| = 11 \end{matrix}\right\}$ 121 model



$\Rightarrow 6 \cdot 6 = 36$ modela prenaučeno $\quad 36 - 1 = 35$

(B) $\quad$ (1,1) $C = 1 = \gamma$

2.9.

$N = 1000$ primjera

$K = 4$ klase $\{400, 400, 100, 100\}$

• SVM u OVO ili OVR shemi

maksimalni omjer $\frac{OVO}{OVR} = ?$

OVO Grammova matrica jezgre

dim. $N = 800$ $\quad 800 \times 800$

$N = 500$ $\quad 500 \times 500$

$N = 200$ $\quad 200 \times 200$

OVR dim. 1000×1000

$$\frac{OVO}{OVR} = \frac{800 \times 800}{1000 \times 1000} = \frac{64}{100} = \frac{16}{25} = \frac{32}{50} \Rightarrow \text{(D)}$$

Napomena

Mogli smo računati s pola dim., ali bi se u omjeru 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 utječu na složenost modela.***]

   (a) Klasifikator 4-NN s euklidskom udaljenošću učen je na sljedećim primjerima iz $\mathbb{R}^3 \times \{0,1\}$:

   $$\mathcal{D} = \left\{\left((\mathbf{x}^{(i)}, y^{(i)}\right)\right\}_{i=1}^{6} = \left\{\left((4,4,0),1\right),\ \left((4,3,1),1\right),\ \left((6,0,2),1\right),\ \left((5,2,2),0\right),\ \left((5,1,1),0\right),\ \left((7,2,0),0\right)\right\}.$$

   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 naoko 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 pojmova, 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 riječi pretvori u drugu. Npr., $L(\textit{stroj}, \textit{straja}) = 2$. 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 $\mathbf{x} =$ *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 $\mathbf{x}$ = “zemlja” pomoću ova dva algoritma. U slučaju jednake sličnosti za dva primjera, kao susjed se uzima onaj koji je u skupu $\mathcal{D}$ naveden prvi. U slučaju izjednačenja glasova između klasa, prednost se daje klasi s numerički manjom oznakom y. **U koju će klasu biti klasificiran primjer 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 $y = 1$

V11 - Neparametarske metode

I Zadaci za učenje

1.1.

Dual. SVM: $h(\vec{x}) = \sum_{i=1}^{N} \alpha_i y^i \phi(\vec{x})^T \phi(\vec{x}^i) + w_0$

Alg. k-NN: $h(\vec{x}) = \underset{j \in \{0,\dots,k-1\}}{\operatorname{argmax}} \sum_{NN_k(\vec{x})} 1\{y^i = j\}$

- zajedničko
  - oba modela su neparametarska modela
- razlika
  - k-NN uspoređuje k najbližih susjeda, a SVM samo s potpornim vektorima

1.2.

manji k => složeniji model

a) 4-NN s euklidskom udaljenosti

	$\vec{x}^i$	y^i	$\lVert\vec{x}^i - \vec{x}^1\rVert$	$\lVert\vec{x}^i - \vec{x}^2\rVert$
1	[4 4 0]	1	$\sqrt{5} = 2.24$	$\sqrt{26} = 5.1$
2	[4 3 1]	1	1	$2\sqrt{5} \approx 4.47$
3	[6 0 2]	1	3	$\sqrt{46} \approx 6.78$
4	[5 2 2]	0	$\sqrt{2} = 1.41$	$3\sqrt{3} \approx 5.196$
5	[5 1 1]	0	$\sqrt{2} = 1.41$	$\sqrt{33} \approx 5.745$
6	[7 2 0]	0	$\sqrt{10} = 3.16$	$\sqrt{59} \approx 7.68$

$\vec{x}^1 = [4\ 2\ 1]$ => najbliži: $\vec{x}^1, \vec{x}^2, \vec{x}, \vec{x}$

$\vec{x}^2 = [0\ 3\ 3]$ => najbliži: $\vec{x}^1, \vec{x}^2, \vec{x}^4, \vec{x}^5$

$h(\vec{x}^1) = \underset{j \in \{0,1\}}{\operatorname{argmax}} (\underset{y=0}{2}, \underset{y=1}{2})$ => nema odluke

$h(\vec{x}^2) = \underset{j}{\operatorname{argmax}} (2, 2)$ -> nema odluke

b) Inverzna kvadr. jezgra $K(\vec{x}, \vec{x}^i) = \frac{1}{1 + \lVert\vec{x} - \vec{x}^i\rVert^2}$

	$\vec{x}^i$	y^i	$K(\vec{x}, \vec{x}^1)$	$K(\vec{x}, \vec{x}^2)$
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	$1/3 = 0.3\dot{3}$	$1/28 \approx 0.036$
5	(5,1,1)	0	$1/3 = 0.3\dot{3}$	$1/34 \approx 0.029$
6	(7,2,0)	0	$1/11 \approx 0.91$	$1/60 \approx 0.017$

težinski model: $h(\vec{x}) = \underset{j}{\operatorname{argmax}} \sum_{i=1}^{N} K(\vec{x}, \vec{x}^i) \cdot 1\{y^i = j\}$

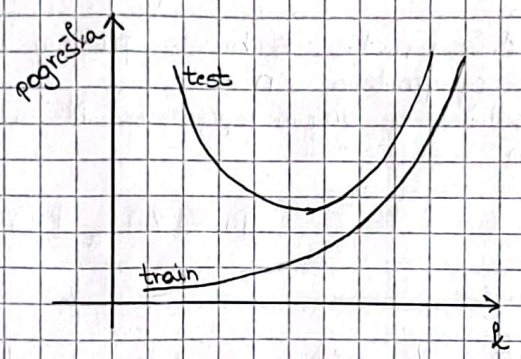
$$h(\vec{x}^{1}) = \underset{j}{\operatorname{argmax}} \left(\underbrace{\frac{1}{6} + \frac{1}{2}}_{y=1}, \underbrace{\frac{1}{3} + \frac{1}{3}}_{y=0} \right) = \operatorname{argmax} \left(\frac{2}{3}, \frac{2}{3} \right)$$

→ nema odluke

$$h(\vec{x}^{2}) = \underset{y}{\operatorname{argmax}} \left(\underbrace{\frac{1}{27} + \frac{1}{21}}_{y=1}, \underbrace{\frac{1}{28} + \frac{1}{34}}_{y=0} \right) = \operatorname{argmax} (0.085, 0.065) = 1$$

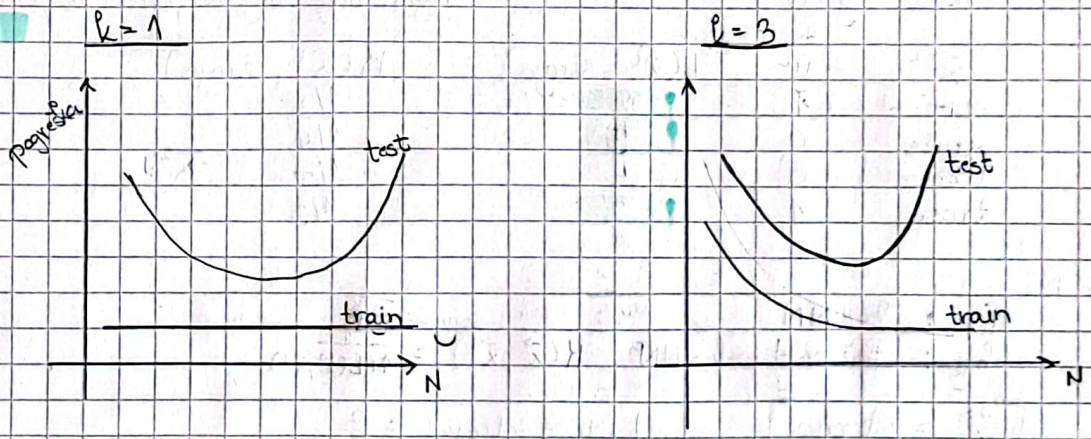
$h(\vec{x}^{2})$ bit će klasificiran kao $y = 1$ //

c)



[• veći k → jednostavniji model]

d)



mali N i mali k → složen model
veliki N i mali k → loša general.
mali N i veliki k → podnaučenost
veliki N i veliki k → najbolja komb. (pazi na prenaučenost!)

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.

slav. $y=1$
Cat. $y=0$

- fja udalj. = Levenshteinova udaljenost : $L(\vec{x}, \vec{x}')$
  → najm. broj zamjena, brisanja, umetanja

$\vec{x}^i$	y^i	$L(\vec{x}^i, \text{straja})$	$K(\vec{x}^i, \text{straja})$
stroj	1	2	1/3
strues	0	3	1/4
tracto	0	4	1/5
trasa	1	2	1/3

h_1 : 3-NN
h_2 : težinski k-NN $\quad K(\vec{x}, \vec{x}') = \frac{1}{1+L(\vec{x}, \vec{x}')}$

$[\vec{x} = \text{straja}]$

$L(\text{strues}, \text{straja}) = 3$
strues → straes → strajs → straja

$L(\text{tracto}, \text{straja}) = 4$

tracto → stracto → strajto → strajao → straja

$$h_1(\vec{x}) = \underset{j \in \{0,1\}}{\operatorname{argmax}} \sum_{\substack{i=1 \\ NN_3}}^{N} 1\{y^i = j\} = \operatorname{argmax}(\underbrace{2}_{y=1}, \underbrace{1}_{y=0}) = 1 \quad //$$

$$h_2(\vec{x}) = \operatorname{argmax} \sum K(\vec{x}^i, \vec{x}) 1\{y^i = j\} = \operatorname{argmax}\left(\underbrace{\tfrac{1}{3}+\tfrac{1}{3}}_{y=1}, \underbrace{\tfrac{1}{4}+\tfrac{1}{5}}_{y=0}\right)$$

$$= \operatorname{argmax}\left(\tfrac{2}{3}, \tfrac{9}{20}\right)$$

$$= 1 \quad //$$

→ B

2.2. k-NN

$\vec{x}^i$	y^i	$K(\vec{x}^i, \text{zemlja})$	
water	0	7/8	ae/zemljawtr
voda	1	1/8	a/zemljavod
zrak	1	2/7	za/zemljark
luft	2	0/9 = 0	ø/luftzemlja
feuer	2	1/8	e/feurzemlja

$$K(\vec{x}_1, \vec{x}_2) = \frac{|\vec{x}_1 \cap \vec{x}_2|}{|\vec{x}_1 \cup \vec{x}_2|}$$ → sličnost

najsličniji ⇒ najveća vrij.

h_1 : 3-NN višeklasni

h_2 : težinski k-NN $\qquad h_2 = \operatorname{argmax} \sum_{i=1}^{N} K(\vec{x}^i, \vec{x}) \cdot 1\{y^i = j\}$

$\vec{x}$ = zemlja

$$h_1(\vec{x}) = \underset{y \in \{0,1,2\}}{\operatorname{argmax}} [\overset{0}{1} \quad \overset{1}{2} \quad \overset{2}{0}] = 1$$

$$h_2(\vec{x}) = \operatorname{argmax} \left(\frac{2}{8}, \frac{1}{8} + \frac{2}{7}, \frac{1}{8} \right) = 1$$

D $\qquad y_1 = 1, \; y_2 = 1$ //

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 $\mathrm{Var}(X)$, kovarijancu $\mathrm{Cov}(X, Y)$, koeficijent korelacije $\rho_{X,Y}$ i kovarijacijsku matricu Σ.

   (c) Dokažite:

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

      ii. $\mathrm{Var}(aX) = a^2\mathrm{Var}(X)$

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

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 Zadaci za učenje

1.1.

$P(1,1) = 0.2$ $\quad$ $P(2,1) = 0.05$
$P(1,2) = 0.05$ $\quad$ $P(2,2) = 0.3$
$P(1,3) = 0.3$ $\quad$ $P(2,3) = 0.1$

(1) $P(X|Y) = \frac{P(X,Y)}{P(Y)}$

$X \in \{1,2\}$ $\quad$ $Y \in \{1,2,3\}$

a) $P(X)$, $P(Y) = ?$
$P(X|Y)$, $P(Y|X) = ?$

Bayesov teorem

$$P(X|Y) = \frac{P(Y|X) \cdot P(X)}{P(Y)}$$

$P(X=1) = 0.2 + 0.05 + 0.3$
$= 0.55$

$P(X=2) = 0.05 + 0.3 + 0.1$
$= 0.45$

$P(Y=1) = 0.2 + 0.05 = 0.25$
$P(Y=2) = 0.05 + 0.3 = 0.35$
$P(Y=3) = 0.1 + 0.3 = 0.4$

po 1

$P(X|Y)$

$$\left.\begin{aligned} P(X=1|Y=1) &= \frac{P(1,1)}{P(Y=1)} = \frac{0.2}{0.25} = \frac{4}{5} = 0.8 \\ P(X=2|Y=1) &= \frac{P(2,1)}{P(Y=1)} = \frac{0.05}{0.25} = \frac{1}{5} = 0.2 \end{aligned}\right\}$$

$$\left\{\begin{aligned} P(X=1|Y=2) &= \frac{0.05}{0.35} = \frac{1}{7} \\ P(X=2|Y=2) &= \frac{0.3}{0.35} = \frac{6}{7} \end{aligned}\right\} \quad \left\{\begin{aligned} P(X=1|Y=3) &= \frac{0.3}{0.4} = \frac{3}{4} \\ P(X=2|Y=3) &= \frac{0.1}{0.4} = \frac{1}{4} \end{aligned}\right\}$$

po 1

$P(Y|X)$

$$\left\{\begin{aligned} P(Y=1|X=1) &= \frac{P(1,1)}{P(X=1)} = \frac{0.2}{0.55} = \frac{4}{11} \\ P(Y=2|X=1) &= \frac{P(1,2)}{P(X=1)} = \frac{0.05}{0.55} = \frac{1}{11} \\ P(Y=3|X=1) &= \frac{P(1,3)}{P(X=1)} = \frac{0.3}{0.55} = \frac{6}{11} \end{aligned}\right\}$$

$$\left\{\begin{aligned} P(Y=1|X=2) &= \frac{0.05}{0.45} = \frac{1}{9} \\ P(Y=2|X=2) &= \frac{0.3}{0.45} = \frac{2}{3} = \frac{6}{9} \\ P(Y=3|X=2) &= \frac{0.1}{0.45} = \frac{2}{9} \end{aligned}\right\}$$

Provjera Bayesovog TEOREMA

$$P(X|Y) = \frac{P(Y|X)\cdot P(X)}{P(Y)} \iff P(X|Y)P(Y) = P(Y|X)P(X)$$

• samo za neke primjere

1 $\quad P(X=1|Y=1)\cdot P(Y=1) \overset{?}{=} P(Y=1|X=1)\,P(X=1)$

$$\frac{4}{5}\cdot\frac{1}{4} = \frac{4}{11}\cdot\frac{11}{20}$$

$$\frac{1}{5} = \frac{1}{5} \quad \checkmark$$

2 $\quad P(X=2|Y=3)\,P(Y=3) \overset{?}{=} P(Y=3|X=2)\cdot P(X=2)$

$$\frac{1}{4}\cdot\frac{4}{10} = \frac{2}{9}\cdot\frac{9}{20}$$

$$\frac{1}{10} = \frac{1}{10} \quad \checkmark$$

Jednakost slijedi direktno iz činjenice
$P(A,B) = P(A)\cdot P(B|A)$
$= P(B)\cdot P(A|B)$

b)

$P(X=1) = 0.55$
$P(X=2) = 0.45$

$P(Y=1) = 0.25$
$P(Y=2) = 0.35$
$P(Y=3) = 0.4$

$$\mathbb{E}[X] = \sum x\cdot P(X=x) = 1\cdot 0.55 + 2\cdot 0.45 = 1.45 = \mu_X$$
$$\mathbb{E}[Y] = \sum y\,P(y) = 1\cdot 0.25 + 2\cdot 0.35 + 3\cdot 0.4 = 2.15 = \mu_Y$$

$$Var(X) = \sigma^2_X = \mathbb{E}[(X-\mathbb{E}[X])^2] = \sum (x-\mu_X)^2 p(x)$$
$$= \sum x^2 p(x) - 2\mu_X \underbrace{\sum x\,p(x)}_{\mu_X} + \mu_X^2 \underbrace{\sum p(x)}_{1}$$
$$= \sum x^2 p(x) - \mu_X^2 = 1\cdot 0.55 + 4\cdot 0.45 - 1.45^2$$
$$= 0.2475$$

$$Var(Y) = \sigma^2_Y = \mathbb{E}[(Y-\mathbb{E}[Y])^2] = \ldots = \mathbb{E}[Y^2] - \mathbb{E}[Y]^2$$
$$= \sum y^2 P(Y) - \mu_Y^2$$
$$= 1\cdot 0.25 + 2^2\cdot 0.35 + 3^2\cdot 0.4 - 2.15^2$$
$$= 0.6275$$

$$Cov(X,Y) = \mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])]$$
$$= \sum_x\sum_y (x-\mu_X)(y-\mu_Y)P(x,y)$$
$$= P(1,1)(1-1.45)(1-2.15) + P(1,2)(1-1.45)(2-2.15) + P(1,3)\cdot(1-1.45)\cdot(3-2.15)$$
$$+ P(2,1)(2-1.45)(1-2.15) + P(2,2)(2-1.45)(2-2.15) + P(2,3)(2-1.45)\cdot(3-2.15)$$

$$= \ldots = -0.45\cdot 0.0175 + 0.55\cdot(-0.0175) = -0.0175$$

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} \approx -0.04441$$

$$\Sigma = \begin{bmatrix} \sigma^2_X & \sigma_{X,Y} \\ \sigma_{Y,X} & \sigma^2_Y \end{bmatrix} = \begin{bmatrix} 0.2475 & -0.0175 \\ -0.0175 & 0.6275 \end{bmatrix}$$

c) i) $Var(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$

$$Var(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \left[\text{označimo radi jednost. } \mu_X = \mathbb{E}[X]\right]$$

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

$$= \sum_x (x-\mu_X)^2 P(X=x) = \sum (x^2 - 2x\mu_X + \mu_X^2) P(x)$$

$$= \underbrace{\sum_x x^2 P(x)}_{\mathbb{E}[X^2]} - 2\mu_X \underbrace{\sum_x xP(x)}_{\mu_X} + \mu_X^2 \underbrace{\sum_x P(x)}_{1}$$

$$= \mathbb{E}[X^2] - 2\mu_X^2 + \mu_X^2$$

$$= \mathbb{E}[X^2] - \mathbb{E}[X]^2 \blacksquare$$

ii)

$$Var(aX) = \mathbb{E}[(aX - \mathbb{E}[aX])^2] \overset{(1)}{=}$$

$$= \mathbb{E}[(aX - a\mathbb{E}[X])^2] =$$

$$= \mathbb{E}[a^2(X - \mathbb{E}[X])^2] \overset{(1)}{=}$$

$$= a^2\,\mathbb{E}[(X - \mathbb{E}[X])^2] =$$

$$= a^2 Var(X)$$

$$\overset{(1)}{\mathbb{E}[aX]} = \sum_x a \cdot xP(x) = a\sum xP(x) = a\mathbb{E}[X]$$

iii)

$$Cov(X,Y) = \mathbb{E}[(X - \mathbb{E}[X]) \cdot (Y - \mathbb{E}[Y])]$$

$$= \mathbb{E}[XY - X\cdot\mathbb{E}[Y] - Y\cdot\mathbb{E}[X] + \mathbb{E}[X]\mathbb{E}[Y]]$$

$$= \mathbb{E}[XY] - \mathbb{E}[X\cdot\mathbb{E}[Y]] - \mathbb{E}[Y\mathbb{E}[X]] + \mathbb{E}[\mathbb{E}[X]\mathbb{E}[Y]]$$

$$= \begin{cases} \mathbb{E}[X\cdot\mathbb{E}[Y]] = \sum x\mu_Y P(x) = \mu_X\mu_\mu = \mathbb{E}[X]\mathbb{E}[Y] \\ \mathbb{E}[\mathbb{E}[X]\mathbb{E}[Y]] = \mathbb{E}[X]\mathbb{E}[Y] \end{cases}$$

$$= \mathbb{E}[XY] - 2\mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[Y]$$

$$= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \,//$$

1.2

a) Nezavisnost slučajnih varijabli X i Y

$\Leftrightarrow$ preko zajedničke vjerojatnosti : $\boxed{P(X,Y) = P(X)P(Y)}$

$\Leftrightarrow$ preko uvjetne vjerojatnosti : $\boxed{\begin{array}{l} P(Y|X) = P(Y) \\ P(X|Y) = P(X) \end{array}}$

$$\begin{aligned} P(X,Y) &= P(X|Y)P(Y) \\ &= P(Y|X)\,P(X) \\ &= P(X)P(Y) \end{aligned}$$

b) Sudeći po $\rho_{X,Y}$ iz 1. zadatka jesu li X i Y linearno zavisne?

$\rho_{X,Y} = -0.04441 \approx 0$

$\Rightarrow$ X i Y nisu linearno zavisne

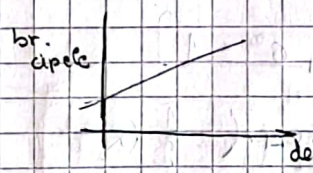
Jesu li nezavisne?

X i Y su linearno nezavisne, međutim ne možemo reći da su općenito X i Y nezavisne. Razlog tome je što Pearsonov koeficijent korelacije mjeri isključivo LINEARNU ZAVISNOST. Drugim riječima moguće je da su X i Y nelinearno zavisne (npr. kvadratno), a da je $\rho_{X,Y} \approx 0$

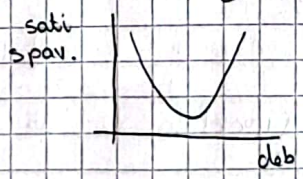
c)

I DOB i VELIČINA CIPELA = linearno zavisno

└ možda i samo zavisno (ovisno o domeni dobi)

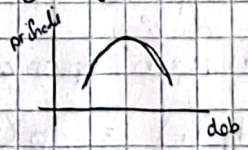


II DOB i SATI SPAVANJA = zavisno



III RAZINA BUKE i UDALJ. OD IZVORA
- linearno zavisno

IV DOB i PRIHODI
- zavisno



d) X i Y nezavisne varijable

$\Rightarrow P(X,Y) = P(X)P(Y)$

vidi 1.1c

$$\begin{aligned} \Rightarrow Cov(X,Y) &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = \sum\sum xy\,P(x,y) - \mu_X\mu_Y \\ &= \sum\sum xy\,P(X)P(Y) - \mu_X\mu_Y \\ &= \sum x\,P(x) \sum y\,P(y) - \mu_X\mu_Y \\ &= \mu_X\mu_Y - \mu_X\mu_Y \\ &= 0 \end{aligned}$$

$\Rightarrow \rho_{X,Y} = \dfrac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = 0 \Rightarrow$ 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 pretpostavci o skupu $\mathcal{D}$ temelji ta definicija?

   (b) Raspolažemo skupom (neoznačenih) primjera $\mathcal{D} = \{x^{(i)}\}_i = \{-2, -1, 1, 3, 5, 7\}$. Pretpostavljamo 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-procjenitelja dviju osnovnih univarijatnih razdioba.***]

   (a) Definirajte MLE-procjenitelj $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$.

   (b) Izvedite MLE-procjenitelj $\hat{\mu}_{\mathrm{ML}}$ za parametar μ Bernoullijeve razdiobe $P(x|\mu)$.

   (c*) Izvedite MLE-procjenitelj $\hat{\mu}_{k,\mathrm{MLE}}$ 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-procjenitelje $\hat{\mu}_{\mathrm{ML}}$ 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-procjenitelj može biti pristran, tj. da najveća izglednost ne jamči nepristranost.***]

   (a) Dokažite da je $\hat{\mu}_{\mathrm{ML}}$ nepristran, a $\hat{\sigma}^2_{\mathrm{ML}}$ 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 pretpostavljamo 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-procjenitelj i način njegovog izračuna za Bernoullijevu distribuciju (beta-Bernoullijev model). Uočiti kako svojstvo konjugatnosti olakšava izračun aposteriorne distribucije.***]

   (a) Definirajte MAP-procjenitelj $\hat{\boldsymbol{\theta}}_{\mathrm{MAP}}$ i objasnite zašto je on bolji od MLE-procjenitelja $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$.

   (b) Objasnite što je to (1) konjugatna distribucija i (2) konjugatna apriorna distribucija. Zašto nam je svojstvo konjugatnosti bitno?

   (c) Apriornu distribuciju parametra μ Bernullijeve 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) $\alpha = 4$, $\beta = 2$.

   (d) Izvedite izraz za aposteriornu distribuciju parametra, $p(\mu|N, m, \alpha, \beta)$.

   (e) 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 $\frac{\alpha-1}{\alpha+\beta-2}$.

   (f) Izračunajte $\hat{\mu}_{\mathrm{MAP}}$ i $\hat{\mu}_{\mathrm{ML}}$ te komentirajte razliku. Kako bi porast broja primjera N utjecao na ovu razliku?

   (g) Pokažite da se MAP-procjenitelj za parametar μ Bernoullijeve distribucije svodi na Laplaceov procjenitelj, ako se apriorna distribucija parametra modelira beta-distribucijom te ako se odaberu odgovarajući (koji?) parametri α i β.

6. [***Svrha: Razumjeti MAP-procjenitelj i način njegovog izračuna za kategorijsku (multinulijevu) varijablu (Dirichlet-kategorijski model).***]

   (a) Definirajte Dirichletovu distribuciju.

   (b) Definirajte Dirichlet-kategorijski model i izvedite MAP procjenitelj za $\alpha_k = 2$.

7. [***Svrha: Razumjeti vezu između probabilističkih modela i poopćenih linearnih modela preko veze između MLE-procjenitelja i minimizacije empirijske pogreške. Razumjeti vezu između MAP-procjenitelja i minimizacije L2-regularizirane empirijske pogreške.***]

   (a) Pokažite da je MLE-procjena za parametre $\mathbf{w}$ kod linearne regresije (uz pretpostavku normalno distribuiranog šuma) ekvivalentna postupku najmanjih kvadrata.

   (b) Pokažite da je MLE-procjena za parametre $\mathbf{w}$ kod logističke regresije (uz pretpostavku Bernoullijeve distribucije oznaka) ekvivalentna minimizacije pogreške unakrsne entropije.

   (c*) Gornja dva zadataka demonstriraju vezu između MLE-procjenitelja i minimizacije empirijske pogreške. Postoji analogna veza između MAP-procjenitelja i minimizacije L2-regularizirane empirijske pogreške. Razmotrimo konkretno linearnu regresiju. Ako se apriorna gustoća vjerojatnosti težina $\mathbf{w}$ definira kao:

   $$p(\mathbf{w}) = \mathcal{N}(0, \alpha^{-1}\mathbf{I})$$

   tj. kao multivarijatna normalna razdioba sa središtem u ishodištu prostora parametara i s izotropnom kovarijacijskom matricom pomnoženom nekim hiperparametrom α^{-1}, onda je MAP-procjenitelj ekvivalentan L2-regulariziranoj kvadratnoj pogrešci. Dokažite to. (Pomoć: slajdovi 30–31 ovdje i poglavlje 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 Zadatci 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 klasa $p(x|y)$. Te su izglednosti definirane Gaussovom gustoćom vjerojatnosti:

$$p(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\Big\{ -\frac{(x-\mu)^2}{2\sigma^2} \Big\}$$

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:

$$p(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\Big\{ -\frac{(x-\mu)^2}{2\sigma^2} \Big\}$$

Koristimo MLE kako bismo na uzorku $\mathcal{D}$ procijenili parametre $\hat{\mu}_{\text{MLE}}$ i $\hat{\sigma}^2_{\text{MLE}}$. Usporedbe radi, parametar σ^2 dodatno procjenjujemo nepristranim procjeniteljem $\hat{\sigma}^2_{\text{UB}}$. Izračunajte log-izglednost parametara $(\hat{\mu}_{\text{MLE}}, \hat{\sigma}^2_{\text{MLE}})$ te log-izglednost parametara $(\hat{\mu}_{\text{MLE}}, \hat{\sigma}^2_{\text{UB}})$. **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 $(\mu_{\text{MLE}}, \sigma^2_{\text{MLE}})$ parametri distribucije procijenjeni MLE-om nad uzorkom $\mathcal{D}$, te neka je σ^2_{UB} nepristrana procjena varijance, izračunata kao $\sigma^2_{\text{UB}} = \frac{N}{N-1}\sigma^2_{\text{MLE}}$. 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 $\mathcal{L}(\mu, \sigma^2|\mathcal{D})$:

$$\begin{aligned}
\mathcal{L}_0 &= \mathcal{L}(\mu_{\text{MLE}}, \sigma^2_{\text{MLE}}|\mathcal{D}) \\
\mathcal{L}_1 &= \mathcal{L}(0, 1|\mathcal{D}) \\
\mathcal{L}_2 &= \mathcal{L}(\mu_{\text{MLE}}, \sigma^2_{\text{UB}}|\mathcal{D}) \\
\mathcal{L}_3 &= \mathcal{L}(\mu_{\text{MLE}}, \sigma^2_{\text{UB}}|\mathcal{D}')
\end{aligned}$$

Što možemo zaključiti o odnosima između ovih vrijednosti funkcije log-izglednosti?

A $\mathcal{L}_0 > \mathcal{L}_1$, $\mathcal{L}_1 \neq \mathcal{L}_2$, $\mathcal{L}_2 \geqslant \mathcal{L}_3$

B $\mathcal{L}_1 \geqslant \mathcal{L}_0$, $\mathcal{L}_2 \geqslant \mathcal{L}_3$

C $\mathcal{L}_0 < \mathcal{L}_3$, $\mathcal{L}_0 \leqslant \mathcal{L}_2$

D $\mathcal{L}_0 \geqslant \mathcal{L}_1$, $\mathcal{L}_0 > \mathcal{L}_2$, $\mathcal{L}_2 \neq \mathcal{L}_3$

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 aposteriornu 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-procjenitelj kako bismo procijenili parametre distribucije kategoričke (multinulijeve) varijable X. Varijabla može poprimiti tri vrijednosti, x_1, x_2 i x_3, pa dakle trebamo procijeniti 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 definirali 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 definirali 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 igraće kocke modeliramo kategoričkom varijablom $\mathbf{x}$, gdje indikatorske varijable $x_1, \ldots, x_6$ odgovaraju vrijednosti koju dobivamo bacanjem kocke. Za procjenu parametara $\boldsymbol{\mu}$ kategoričke distribucije koristimo MAP-procjenitelj 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-procjeniteljem 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 proprimiti 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, $\alpha \geqslant 1$. 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 procijenjenoj 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:

$$p(\mu|\alpha,\beta) = \frac{1}{B(\alpha,\beta)}\mu^{\alpha-1}(1-\mu)^{\beta-1} \qquad \mu^* = \frac{\alpha-1}{\alpha+\beta-2}$$

Na temelju beta-Bernoullijevog modela na skupu $\mathcal{D}$ računamo MAP procjenu parametra μ Bernoulijeve 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 učenje

1.1.

a) fja izglednosti $\mathcal{L}(\vec{\theta}|D) : \vec{\theta} \to p(D|\vec{\theta})$

$$\mathcal{L}(\vec{\theta}|D) = p(D|\vec{\theta}) = p(\vec{x}^1, \vec{x}^2, \dots, \vec{x}^N | \vec{\theta})$$
$$= \prod_{i=1}^{N} p(\vec{x}^i | \vec{\theta})$$

- pretpostavka na skupu $D = \{\vec{x}^i\}_{i=1}^N$
  - podaci nezavisno i identično distribuirani (I.I.D.)

b) $D = \{x^i\} = \{-2, -1, 1, 3, 5, 7\} \sim \mathcal{N}(\mu, \sigma^2)$

$$\mathcal{L}(\mu, \sigma | D) = p(D|\mu, \sigma)$$
$$= \prod_{i=1}^{6} p(x^i | \mu, \sigma)$$
$$= \prod_{i=1}^{6} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(x^i - \mu)^2\right)$$
$$= [\mu = 0, \sigma^2 = 1]$$
$$= \prod_{i=1}^{6} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x^i)^2\right)$$
$$= \left(\frac{1}{\sqrt{2\pi}}\right)^6 \cdot \prod_{i=1}^{6} e^{-\frac{1}{2}(x^i)^2}$$
$$\approx 1 \cdot 10^{-22}$$

→ izglednost i vjerojatnost uzorka D uz parametre $\mu = 0$ i $\sigma^2 = 1$ iznosi 10^{-22}.

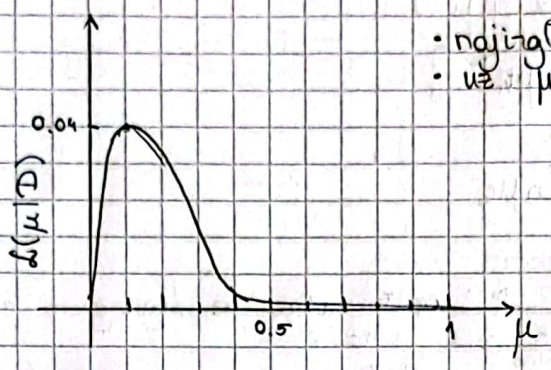
c) Trovčić N puta
↳ m glava
↳ $N-m$ pisma

Bernoulli

$$\mathcal{L}(\mu | N, m) = p(D | N, m, \mu) = \prod_{i=1}^{N} p(x^i | N, m, \mu)$$
$$= \prod_{i=1}^{N} \mu^{x^i} (1-\mu)^{1-x^i}$$
$$= \mu^m (1-\mu)^{N-m}$$

d) $\left.\begin{matrix} N = 10 \\ m = 1 \end{matrix}\right\}$ $\mathcal{L}(\mu|D) = \mu^m (1-\mu)^{N-m}$

$= \mu (1-\mu)^9$



- najizglednija vrijednost parametra $\mu = 0.1$
- uz $\mu = 0.1$ skup D je najvjerojatniji

$\mathcal{L}(\mu|D) = P(D|\mu)$

1.2. progenitelj MLE

a) definicija $\vec{\theta}_{ML} = \underset{\vec{\theta}}{\operatorname{argmax}}\, \mathcal{L}(\vec{\theta}|D)$

↳ također zbog matematičke jednostavnosti i monotonosti logaritamske fje

log-izglednost

$\vec{\theta}_{ML} = \underset{\vec{\theta}}{\operatorname{argmax}}\left(\ln[\mathcal{L}(\vec{\theta}|D)]\right)$

b) $\hat{\mu}_{ML} = ?$ (Bernoullijeva razdioba)

Log-izglednost

$\ln[\mathcal{L}(\hat{\mu}_{ML}|D)] = \ln[P(D|\hat{\mu}_{ML})] = \ln \prod_{i=1}^{N} P(x^i|\hat{\mu}_{ML}) = \ln \prod \mu^{x^i}(1-\mu)^{1-x^i}$

$= \sum x^i \ln(\mu) + (1-x^i)\ln(1-\mu)$

$= m\ln(\mu) + (N-m)\ln(1-\mu)$

- maksimizacija $\mathcal{L}$

$$\frac{d\mathcal{L}}{d\mu} = \frac{m}{\mu} + \frac{N-m}{1-\mu}\cdot(-1) = 0$$

$$m(1-\mu) - (N-m)\mu = 0$$

$$m - \cancel{m\mu} - N\mu + \cancel{m\mu} = 0$$

$$\left[\hat{\mu}_{ML} = \frac{m}{N} = \frac{1}{N}\sum_{i=1}^{N} x^i\right]$$

c*) $\hat{\mu}_{k,ML} = ?$ (Kategorička „multinulli" razdioba)

$$\ln \mathcal{L}(\vec{\mu} \mid D) = \ln\left[P(D \mid \vec{\mu})\right]$$
$$= \ln\left[\prod_{i=1}^{N} P(\vec{x}^i \mid \vec{\mu})\right]$$
$$= \ln\left[\prod_{i=1}^{N}\prod_{k=1}^{K} \mu_k^{x_k^i}\right]$$
$$= \sum_{i=1}^{N}\sum_{k=1}^{K} x_k^i \ln \mu_k$$

• ograničenje: $\sum_{k=1}^{K} \mu_k = 1$ → optimizacija uz Lagrangeove multipl.

Lagrang. fja: $f(\vec{\mu}, \lambda) = \sum_{i=1}^{N}\sum_{k=1}^{K} x_k^i \ln \mu_k + \lambda\left(\sum_{k=1}^{K} \mu_k - 1\right)$

$$\frac{\partial f}{\partial \mu_k} = 0 \Rightarrow \mu_k = -\frac{1}{\lambda}\sum_{i=1}^{N} x_k^i$$

<u>ograničenje</u>

$$\sum_{k=1}^{K} \mu_k = -\frac{1}{\lambda}\underbrace{\sum_{k=1}^{K}\sum_{i=1}^{N} x_k^i}_{N} = 1$$

$$\Rightarrow \lambda = -N$$

$$\hookrightarrow \boxed{\hat{\mu}_{k,ML} = \frac{1}{N}\sum_{i=1}^{N} x_k^i = \frac{N_k}{N}}$$

d) $\hat{\mu}_{ML}, \hat{\sigma}^2 = ?$ (Univarijantna Gaussova razdioba)

$$f(\mu, \sigma^2) = \ln\left[\mathcal{L}(\mu, \sigma^2 \mid D)\right] = \ln\left[p(D \mid \mu, \sigma^2)\right] = \ln\left[\prod_{i=1}^{N} p(x^i \mid \mu, \sigma^2)\right]$$
$$= \ln\left[\prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(x^i - \mu)^2\right)\right]$$
$$= \sum_{i=1}^{N}\left[\ln(\sigma\sqrt{2\pi})^{-1} + \left(-\frac{1}{2\sigma^2}\right)(x^i - \mu)^2\right]$$
$$= \sum_{i=1}^{N} -\ln\sigma - \frac{1}{2}\ln(2\pi) - \frac{1}{2\sigma^2}(x^i - \mu)^2$$
$$= -N\ln\sigma \underbrace{- \frac{N}{2}\ln(2\pi)}_{\text{konstanta}} - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x^i - \mu)^2 \Rightarrow \nabla \ln[\mathcal{L}] = 0$$

$$\frac{\partial f}{\partial \mu} = -\frac{1}{2\sigma^2}\sum_{i=1}^{N} -2x^i + 2\mu = 0$$
$$-2\sum x^i - \mu = 0$$
$$-N\mu + \sum x^i = 0 \Rightarrow \boxed{\mu_{ML} = \frac{1}{N}\sum_{i=1}^{N} x^i}$$

$$\frac{\partial f}{\partial \sigma} = 0$$
$$\hookrightarrow \boxed{\sigma^2_{ML} = \frac{1}{N}\sum_{i=1}^{N}(x^i - \hat{\mu}_{ML})^2}$$

↳ pristran!

1.3.

a)

$$\hat{\mu}_{ML} = \frac{1}{N}\sum_{i=1}^{N} x^i \qquad \sigma^2_{ML} = \frac{1}{N}\sum_{i=1}^{N}(x^i - \hat{\mu})^2$$

$$\mathbb{E}[\hat{\mu}] = \mathbb{E}\left[\frac{1}{N}\sum x^i\right] = \frac{1}{N}\mathbb{E}\left[\sum x^i\right] = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[x^i]$$

$$= \frac{1}{N}\sum \mu = \frac{1}{N}\cdot N\cdot\mu = \mu$$

$\Rightarrow$ $\hat{\mu}_{ML}$ je nepristran procjenitelj

$$\mathbb{E}[\sigma^2] = \frac{1}{N}\mathbb{E}\left[\sum x^{i2} - 2\hat{\mu}x^i + \hat{\mu}^2\right]$$

$$= \frac{1}{N}\mathbb{E}\left[\sum x^{i2} - 2\hat{\mu}\sum x^i + \sum \hat{\mu}^2\right]$$

$$= \frac{1}{N}\mathbb{E}\left[\sum (x^i)^2 - 2\cdot N\hat{\mu}^2 + N\hat{\mu}\right]$$

$$= \frac{1}{N}\mathbb{E}\left[\sum (x^i)^2 - N\hat{\mu}^2\right]$$

$$= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[x^{i2}] - \frac{1}{N}\cancel{N}\,\mathbb{E}[\hat{\mu}^2]$$

$$= \mathbb{E}[(x^i)^2] - \mathbb{E}[\hat{\mu}^2]$$

$$\mathbb{E}(\hat{\mu}^2) = \frac{\sigma^2}{n} + \mu^2$$
$$\mathbb{E}[(x^i)^2] = \sigma^2 + \mu^2$$

$$= \ldots$$

$$= \sigma^2 + \mu^2 - \left(\frac{\sigma^2}{n} + \mu^2\right)$$

$$= \frac{n-1}{n}\sigma^2 \neq \sigma^2$$

$\Rightarrow$ $\hat{\sigma}^2_{ML}$ je pristran procjenitelj

$$b(\hat{\sigma}^2_{ML}) = \frac{N-1}{N}\sigma^2 - \sigma^2 = -\frac{\sigma^2}{N}$$

b) Pristranost u praksi nije problematična
- za male uzorke $\Rightarrow$ korekcija $\hat{\sigma}^2_{ML}$ s faktorom $\frac{N}{N-1}$
- za veliki uzorak $(N \to \infty)$ $\Rightarrow$ $b(\hat{\sigma}^2) \to 0$.

1.4

$$D = \{\vec{x}^i\}_{i=1}^{6}$$

$$\vec{x}^1 = (9.5, -0.7, -2.8)$$
$$\vec{x}^2 = (8.8, -0.8, +3.2)$$
$$\vec{x}^3 = (6.5, -0.2, -0.8)$$
$$\vec{x}^4 = (2.3, 0.3, 1.2)$$
$$\vec{x}^5 = (2.2, 0, 0)$$
$$\vec{x}^6 = (3.6, 0.3, 1.2)$$

a) $$\left[\hat{\vec{\mu}}_{MLE} = \frac{1}{N}\sum_{i=1}^{N}\vec{x}^i\right] = \left(\frac{329}{60}, \frac{-11}{60}, \frac{-11}{15}\right)$$

Multivarijatna Gaussova razdioba

$$\hat{\Sigma}_{MLE} = ?$$

$$\left[\hat{\Sigma}_{MLE} = \frac{1}{N}\sum_{i=1}^{N}\underbrace{(\vec{x}^i - \hat{\vec{\mu}}_{ML})}_{3\times 1}\underbrace{(\vec{x}^i - \vec{\mu}_{ML})^T}_{1\times 3}\right]$$

$$= \dots$$

$$= \begin{bmatrix} 8.771 & -1.198 & -4.792 \\ -1.198 & 0.191 & 0.766 \\ -4.792 & 0.766 & 3.06 \end{bmatrix}$$

b) $$\vec{x} = (-2, 1, 0)$$

$$p(\vec{x} \mid \hat{\vec{\mu}}, \hat{\Sigma}) = \frac{1}{(2\pi)^{\frac{3}{2}} |\hat{\Sigma}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\vec{x} - \hat{\vec{\mu}})^T \hat{\Sigma}^{-1} (\vec{x} - \hat{\vec{\mu}})\right\}$$

$$|\hat{\Sigma}| = \dots = 0$$

→ Gustoća nije dobro definirana jer je determinanta kovarijacijske matrice 0

→ dijeljenje s 0 + ne postoji inverz

↳ naznaka multikolinearnosti i/ili konstantnih značajki

c) $\rho_{X,Y} = \dfrac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y}$

$$\hat{\Sigma}_{MLE} = \begin{array}{c} \\ \\ \\ \end{array}\begin{array}{cc} \begin{array}{ccc} X1 & X2 & X3 \end{array} & \\ \begin{bmatrix} 8.771 & -1.198 & -4.792 \\ -1.198 & 0.191 & 0.766 \\ -4.792 & 0.766 & 3.06 \end{bmatrix} & \begin{array}{c} X1 \\ X2 \\ X3 \end{array} \end{array}$$

- na dijagonali – varijanca
- van dijagonale – kovarijanca

$\sigma_{X_1} = \sqrt{8.771}$ $\quad Cov(X1, X2) = -1.198$

$\sigma_{X_2} = \sqrt{0.191}$ $\quad Cov(X1, X3) = -4.792$

$\sigma_{X_3} = \sqrt{3.06}$ $\quad Cov(X2, X3) = 0.766$

$$\rho_{X_1,X_2} = \frac{Cov(X_1, X_2)}{\sigma_{X_1} \cdot \sigma_{X_2}} \approx -0.9256$$

$\rho_{X_1,X_3} \approx -0.92498$

$\rho_{X_2,X_3} \approx 1.002$

→ iz ulaznog prostora izbacujemo značajku X_2 jer ima najveću korelaciju s ostalim značajkama

→ nakon izbacivanja značajke

$\hat{\mu}_{ML} = (5.48, -0.73)$

$$\hat{\Sigma}_{ML} = \frac{1}{N} \sum_{i=1}^{N} (\vec{x}^i - \hat{\mu}_{ML})(\vec{x}^i - \hat{\mu}_{ML})^T = \begin{bmatrix} 8.771 & -4.792 \\ -4.792 & 3.06 \end{bmatrix}$$

$|\hat{\Sigma}_{ML}| = 3.8945$

$$\hat{\Sigma}^{-1}_{ML} = \begin{bmatrix} 0.7863 & 1.2304 \\ 1.2304 & 2.2522 \end{bmatrix}$$

$\vec{x} = (-2, \underset{\text{ignoriramo}}{1}, 0)$

$$p(\vec{x} \mid \hat{\mu}_{ML}, \hat{\Sigma}_{ML}) = \frac{1}{(2\pi)^{\frac{n}{2}} |\hat{\Sigma}_{ML}|^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2} (\vec{x} - \hat{\vec{\mu}})^T \hat{\Sigma}^{-1} (\vec{x} - \hat{\vec{\mu}}) \right\}$$

$\cdot [n = 2]$

$= \ldots$

$= 1.0346 \cdot 10^{-8}$

1.5

a) MAP procjenitelj

$$\vec{\theta}_{MAP} = \underset{\vec{\theta}}{\operatorname{argmax}}\ \mathcal{L}(\vec{\theta}\,|\,D)\,p(\vec{\theta})$$

$$= \underset{\vec{\theta}}{\operatorname{argmax}}\ p(\vec{\theta}\,|\,D)$$

Procjenitelj MAP procjenjuje maksimalnu aposteriornu distribuciju parametara. Kombinira izglednost parametra $\vec{\theta}$ (informacije iz podataka) s apriornom distribucijom parametara (prijašnje znanje)

U odnosu na MLE procjenitelj MAP
- omogućava ugrađivanje 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 $p(\vec{\theta})$ odabrane 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.: $p(\mu\,|\,\alpha,\beta) = \frac{1}{B(\alpha,\beta)}\,\mu^{\alpha-1}(1-\mu)^{\beta-1}$

• slika za

$\mu = \frac{1}{2}$, $\alpha = \beta = 1$

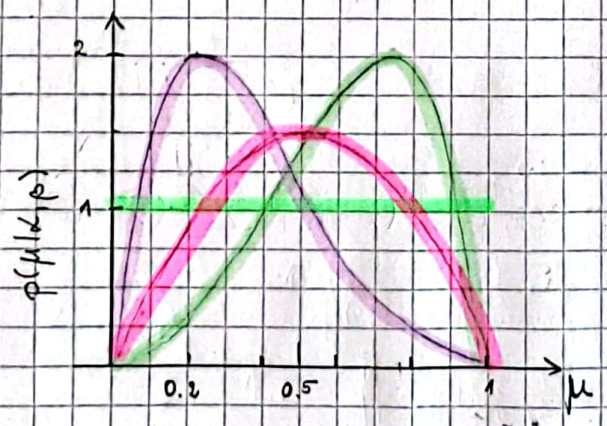
$\mu = \frac{1}{4}$, $\alpha = \beta = 2$

$\mu = \frac{3}{4}$, $\alpha = 2$, $\beta = 4$

$\alpha = 4$, $\beta = 2$

• mod beta-distribucije:

$$\left[\mu = \frac{\alpha - 1}{\alpha + \beta - 2}\right]$$



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)

$$p(\mu) = \frac{1}{B(\alpha,\beta)}\,\mu^{\alpha-1}(1-\mu)^{\beta-1}$$

$$p(D|\mu) = \mu^{m}(1-\mu)^{N-m}$$

$$p(\mu|N,m,\alpha,\beta) = \frac{1}{p(D)}\cdot\frac{1}{B(\alpha,\beta)}\,\mu^{\alpha-1}(1-\mu)^{\beta-1}\cdot\mu^{m}(1-\mu)^{N-m}$$

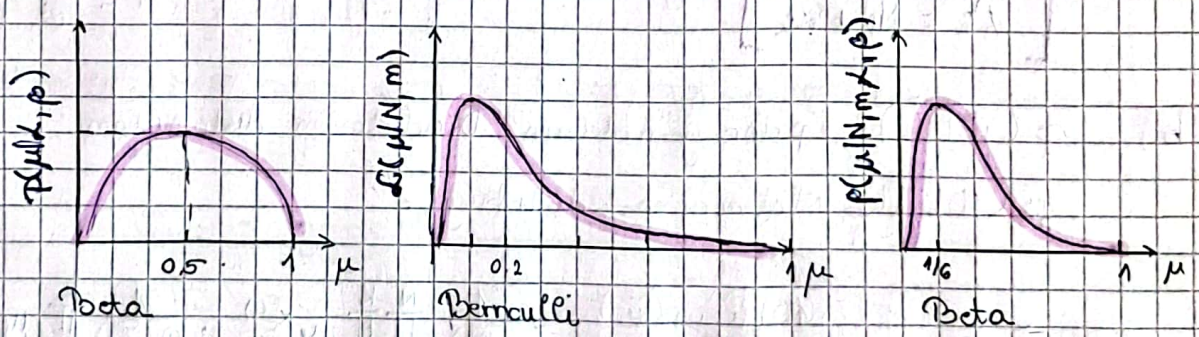
$$= \frac{1}{p(D)B(\alpha,\beta)}\,\mu^{\alpha+m-1}(1-\mu)^{N-m+\beta-1}$$

$$= \frac{1}{B(\alpha',\beta')}\,\mu^{\alpha'-1}(1-\mu)^{\beta'-1}$$

$$\left[\begin{array}{l}\alpha' = \alpha + m\\ \beta' = N - m + \beta\end{array}\right]$$

c) apriorna d. $p(\mu|\alpha=2,\beta=2) \rightarrow \text{mod} = \frac{\alpha-1}{\alpha+\beta-2} = \frac{1}{2}$
$N=10$, $m=1$

$\mathcal{L}(\mu|N,m) = \mu(1-\mu)^{9} \rightarrow \text{mod} = \frac{m}{N} = 0.1$



· 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)

$$\hat{\mu}_{MAP} = \frac{\alpha' - 1}{\alpha' + \beta' - 2} = \frac{m + \alpha - 1}{m + \alpha + N - m + \beta - 2} = \frac{m + \alpha - 1}{\alpha + \beta + N - 2} = \frac{1 + 2 - 1}{2 + 2 + 10 - 2}$$

$$= \frac{2}{12} = \frac{1}{6} \approx 0.1667$$

$$\hat{\mu}_{MLE} = \frac{m}{N} = \frac{1}{10} = 0.1$$

Porastom broja primjera N razlika između $\hat{\mu}_{MAP}$ i $\hat{\mu}_{MLE}$ bi se smanjila jer veći utjecaj na $\hat{\mu}_{MAP}$ ima N u razvnicku
- više se vjeruje podacima, a ne apriornom znanju

g) $\alpha = \beta = 2 \rightarrow \hat{\mu}_{MAP}$ se svodi na Laplaceov procjenitelj

$$\hat{\mu}_{MAP} = \frac{m + \alpha - 1}{\alpha + \beta + N - 2} = \frac{m + 1}{N + 2} \quad //$$

1.6. Dirichlet - kategorički model

a) Dirichletova distribucija

$$P(\vec{\mu} \mid \vec{\alpha}) = \frac{1}{B(\vec{\alpha})} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}, \quad \sum_{k=1}^{K} \mu_k = 1 \quad i \quad \mu_k \geq 0$$

$$\left[\hat{\mu}_{k,MAP} = \frac{\alpha'_k - 1}{\sum_{k=1}^{K} \alpha'_k - K} \right]$$

$$\left[\alpha'_k = N_k + \alpha_k \right]$$

b) $p(\vec{\mu} \mid \vec{\alpha})$ - prior je modeliran Dirichletovom distribucijom

$p(D \mid \vec{\mu})$ - kategorička distribucija.

$$p(\vec{\mu} \mid \vec{\alpha}, D) = \frac{p(D \mid \vec{\mu})\, p(\vec{\mu})}{p(D)} = \frac{1}{p(D)} \prod_{i=1}^{N} \prod_{k=1}^{K} \mu_k^{x_k^{(i)}} \cdot \frac{1}{B(\vec{\alpha})} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}$$

$$= \frac{1}{B(\vec{\alpha})\, p(D)} \prod_{k=1}^{K} \mu_k^{N_k} \cdot \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}$$

$$= \frac{1}{B(\vec{\alpha})\, p(D)} \prod_{k=1}^{K} \mu_k^{\alpha_k + N_k - 1} = \frac{1}{B(\vec{\alpha}')} \prod_{k=1}^{K} \mu_k^{\alpha_k' - 1}$$

$$\left[\alpha_k' = N_k + \alpha_k \right] \quad \underline{\text{mod}} \quad \left[\hat{\mu}_{k,MAP} = \frac{\alpha_k' - 1}{\sum_{k=1}^{K} \alpha_k' - K} = \frac{\alpha_k + N_k - 1}{N + \sum_{k=1}^{K} \alpha_k - K} \right]$$

$\alpha_k = 2$

$$\hat{\mu}_{k,MAP} = \frac{\alpha_k + N_k - 1}{\sum \alpha_k + N - K} = \frac{N_k + 1}{2K + N - K} = \frac{N_k + 1}{N + K}$$

1.7.

a) MLE procjena za $\vec{w}$ (kod linearne regresije)

$y \sim \mathcal{N}(\mu, \sigma^2)$

$$\hat{\vec{w}}_{MLE} = \underset{\vec{w}}{\operatorname{argmax}} \left[\ln \mathcal{L}(\vec{w} | D) \right]$$

$$\ln\left[\mathcal{L}(\vec{w}|D)\right] = \ln\left[p(D|\vec{w})\right] = \ln\left[\prod_{i=1}^{N} p(y^i | \vec{x}^i)\right]$$

$$= \ln\left[\prod_{i=1}^{N} \mathcal{N}(h(\vec{x}^i; \vec{w}), \sigma^2)\right]$$

$$= \ln\left[\prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(y^i - h(\vec{x}^i))^2\right)\right]$$

$$= \sum_{i=1}^{N} -\ln \sigma\sqrt{2\pi} + -\frac{1}{2\sigma^2}(y^i - h(\vec{x}^i))^2$$

$$= \underbrace{-N\ln\sigma - \frac{N}{2}\ln 2\pi}_{\text{konstanta}} - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y^i - h(\vec{x}^i; \vec{w}))^2$$

$$\propto \underbrace{-\frac{1}{2}\sum_{i=1}^{N}(y^i - h(\vec{x}^i; \vec{w}))^2}_{E(\vec{w}|D)}$$

- maksimizacija log-izglednosti $\Leftrightarrow$ minimizacija kvadratne pogreške

b) MLE procjena za $\vec{w}$ (kod logističke regresije)

$y \sim \mathcal{B}(\mu)$

$$\hat{\vec{w}}_{MLE} = \underset{\vec{w}}{\operatorname{argmax}} \left[\ln \mathcal{L}(\vec{w}|D)\right]$$

$$\ln \mathcal{L}(\vec{w}|D) = \ln\left[p(\vec{y}|\vec{w})\right] = \ln\left[\prod_{i=1}^{N} p(y^i|\vec{w})\right]$$

$$= \ln\left[\prod_{i=1}^{N} \mu^{y^i}(1-\mu)^{1-y^i}\right]$$

$$= \sum_{i=1}^{N} y^i \ln(\mu) + (1-y^i)\ln(1-\mu)$$

- maksim. log izglednosti $\Leftrightarrow$ minimizacija unakrsne pogreške

c) i d) – Skip.

II Zadaci s ispita

2.1.

(x^i, y^i)
$(-2, 1)$
$(-2, 1)$
$(-1, 0)$
$(0, 0)$
$(1, 1)$
$(3, 1)$

$$p(x|y) \sim \mathcal{N}(\mu, \sigma^2)$$

$$\mathcal{L}(\mu_1, \sigma_1^2 | D_y = 1) = p(D_y = 1 | \mu_1, \sigma_1^2)$$
$$= \prod_{i=1}^{4} p(x^i | \mu_1, \sigma_1^2)$$
$$= \prod_{i=1}^{4} \frac{1}{\sigma_1\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma_1^2}(x^i - \mu_1)^2\right\}$$
$$= \prod_{i=1}^{4} \sqrt{\frac{2}{9\cdot 2\pi}} \exp\left\{-\frac{1}{2}\cdot\frac{2}{9}(x^i)^2\right\}$$
$$= \prod_{i=1}^{4} \frac{1}{3\sqrt{\pi}} \exp \frac{-(x^i)^2}{9}$$

podskup $D_y = 1$, $N = 4$

$$\begin{cases} \hat{\mu}_1 = \frac{1}{N}\sum_{i=1}^{N} x^i = \frac{1}{4}(-2-2+1+3) = 0 \\ \hat{\sigma}_1^2 = \frac{1}{N}\sum_{i=1}^{N}(x^i - \hat{\mu}_1)^2 = \frac{9}{2} \end{cases}$$

Log-izglednost

$$\ln \mathcal{L}(\mu_1, \sigma_1^2 | D) = \ln \prod_{i=1}^{4} \frac{1}{3\sqrt{\pi}} \exp \frac{-1}{9} x^{i2}$$
$$= \sum_{i=1}^{4} \ln\frac{1}{3\sqrt{\pi}} + -\frac{1}{9}x^{i2}$$
$$= -4\ln\frac{1}{3\sqrt{\pi}} - \frac{1}{9}\sum x^{i2} = -8.684 \Rightarrow \text{B}$$

//

2.2.

$D = \{-3, -2, 1, 4\}$

kontinuirana slučajna varijabla

$\rightarrow \mathcal{N}(\mu, \sigma^2)$

$$\ln[\mathcal{L}(\mu, \sigma^2 | D)] = \ln\left[\prod_{i=1}^{N} p(x^i | \mu, \sigma^2)\right]$$
$$= \ln\left[\prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(x^i - \mu)^2\right)\right]$$
$$= \sum_{i=1}^{N} -\ln\sigma - \frac{1}{2}\ln 2\pi - \frac{1}{2\sigma^2}(x^i - \mu)^2$$
$$= -N\ln\sigma \underbrace{- \frac{N}{2}\ln 2\pi - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i - \mu)^2}$$

$$\begin{cases} \hat{\mu}_{MLE} = \frac{1}{N}\sum x^i = 0 \\ \hat{\sigma}^2_{MLE} = \frac{1}{N}\sum(x^i - \hat{\mu}_{MLE})^2 = \frac{15}{2} \\ \hat{\sigma}^2_{UB} = \frac{1}{N-1}\sum(x^i - \hat{\mu}_{MLE})^2 = 10 \end{cases}$$

$\ln\mathcal{L}_1 = -9.7055$

$\ln\mathcal{L}_2 = -9.7809$

$$\Delta\mathcal{L} = \ln\mathcal{L}_1 - \ln\mathcal{L}_2 = 0.075 \rightarrow \text{C}$$

2.3.

$$\mathcal{L}(\mu, \sigma^2 | D) = \ln \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x^i - \mu)^2\right)$$

$$= \dots$$

$$= -N \ln \sigma - \frac{N}{2} \ln 2\pi - \frac{1}{2} \sum_{i=1}^{N} (x^i - \mu)^2$$

$$\begin{cases} \hat{\mu}_{MLE} = \frac{1}{N} \sum x_i \\ \hat{\sigma}^2_{MLE} = \frac{1}{N} \sum (x_i - \hat{\mu}_{MLE})^2 \\ \hat{\sigma}^2_{UB} = \frac{N}{N-1} \hat{\sigma}^2_{MLE} \end{cases} \qquad D' \subset D$$

$$\mathcal{L}_0 = \mathcal{L}(\hat{\mu}_{MLE}, \hat{\sigma}^2_{MLE} | D)$$
$$\mathcal{L}_1 = \mathcal{L}(0, 1 | D)$$
$$\mathcal{L}_2 = \mathcal{L}(\hat{\mu}_{MLE}, \hat{\sigma}^2_{UB} | D)$$
$$\mathcal{L}_3 = \mathcal{L}(\hat{\mu}_{MLE}, \hat{\sigma}^2_{UB} | D')$$

$\hat{\sigma}^2_{MLE}$ podcjenjuje σ^2 $\left(b = -\frac{\sigma^2}{N}\right)$

- MLE - most likelihood estimation (najveći $\mathcal{L}$ za MLE procjenitelje)

$\rightarrow$ $\mathcal{L}_0 \geq \mathcal{L}_1$, $\mathcal{L}_0 > \mathcal{L}_2$ $(\hat{\sigma}^2_{UB} \neq \hat{\sigma}^2_{MLE})$

$\rightarrow$ otpada B $(\mathcal{L}_1 \geq \mathcal{L}_0)$ i C $(\mathcal{L}_0 < \mathcal{L}_3, \mathcal{L}_0 \leq \mathcal{L}_2)$

A odgovor netočan

$\mathcal{L}_1 \neq \mathcal{L}_2 \Rightarrow$ može se desiti da je skup D baš takav da $\hat{\mu}_{MLE} = 0$ i $\hat{\sigma}^2_{UB} = 1$

- u principu o $\mathcal{L}_1$ ne možemo ništa zaključiti osim da je $\mathcal{L}_0 \geq \mathcal{L}_1$

D

$\left.\begin{array}{l} \mathcal{L}_0 \geq \mathcal{L}_1 \\ \mathcal{L}_0 > \mathcal{L}_2 \end{array}\right\}$ zbog prirode MLE

$\mathcal{L}_2 \neq \mathcal{L}_3$ - ovisi kakav D' uzrokujemo, može se dogoditi $\mathcal{L}_2 > \mathcal{L}_3$ i $\mathcal{L}_2 < \mathcal{L}_3$, ali ne i $\mathcal{L}_2 = \mathcal{L}_3$ jer je D' podskup od D

2.4. $\alpha = \beta = 2$

$\hat{\mu}_{MAP} = ?$

① D_1 ② D_2

$N_1 = 50$ $N_2 = 15$

$m_1 = 42$ $m_2 = 3$

– „online" učenje

• <u>učenje na skupu D1</u>

$$\alpha = \beta = 2$$

$$\alpha' = \alpha + m_1 = 2 + 42 = 44$$

$$\beta' = \beta + N_1 - m_1 = 2 + 50 - 42 = 2 + 8 = 10$$

$$\hat{\mu}_{MAP,1} = \frac{\alpha' - 1}{\alpha' + \beta' - 2} = \frac{43}{54 - 2} = \frac{43}{52}$$

• <u>učenje na skupu D2</u>

$$\alpha = 44$$

$$\beta = 10$$

$$\alpha' = \alpha + m_2 = 44 + 3 = 47$$

$$\beta' = \beta + N_2 - m_2 = 10 + 15 - 3 = 22$$

$$\hat{\mu}_{MAP,2} = \frac{\alpha' - 1}{\alpha' + \beta' - 2} = \frac{46}{67}$$

$$\Delta\mu = \hat{\mu}_{MAP,1} - \hat{\mu}_{MAP,2} = 0.1404 \quad \Rightarrow \text{(C)} \quad \Delta\mu = \hat{\mu}_2 - \hat{\mu}_1 = -0.14$$

2.5. $\hat{\mu}_{k,MAP}$

$K = 3$

• Dirichlet – kateg. model

$(\alpha_1, \alpha_2, \alpha_3) = (2, 2, 1)$ ← apriorno

• realizacija $(0, \frac{1}{2}N, \frac{1}{2}N)$

$\vec{\mu}_{MAP} = ?$

$$\alpha_k' = \alpha_k + N_k$$

$$\hat{\mu}_{k,MAP} = \frac{\alpha_k' - 1}{\sum \alpha_k' - K}$$

$$\left[\begin{aligned} \alpha_1' &= 2 + 0 = 2 \\ \alpha_2' &= 2 + \tfrac{1}{2}N = \frac{N+4}{2} \\ \alpha_3' &= 1 + \tfrac{1}{2}N = \frac{N+2}{2} \end{aligned}\right.$$

$$\left[\ \sum \alpha_k' - K = \sum_{k=1}^{3} \alpha_k' - 3 = \frac{2N+6}{2} + 2 - 3 = N + 3 + 2 - 3 = N + 2\right.$$

$$\hat{\mu}_1 = \frac{2-1}{N+2} = \frac{1}{N+2}$$

$$\hat{\mu}_2 = \frac{\frac{N+4}{2} - 1}{N+2} = \frac{\frac{N+2}{2}}{N+2} = \frac{1}{2}$$

$$\hat{\mu}_3 = \frac{\frac{N+2}{2} - 1}{N+2} = \frac{N}{2N+4}$$

A ne može $\hat{\mu}_1 \neq 0$

B ne može $\hat{\mu}_2 = 1/2$

C ne može (nije nužno da $\hat{\mu}_1 < \hat{\mu}_3$)

D $0 < \hat{\mu}_1 < \frac{1}{3}$

$\frac{1}{3} < \hat{\mu}_2 < 1$

$0 < \hat{\mu}_3, \hat{\mu}_2 < 1$

$\underline{N = 2}$

$\hat{\mu}_1 = \frac{1}{4} = 0.25$

$\hat{\mu}_2 = \frac{1}{2} = 0.5$

$\hat{\mu}_3 = \frac{2}{8} = \frac{1}{4} = 0.25$

2.7

$\vec{\mu}_{MAP}$ → Dirichlet - kateg. model

$K = 3$

prior: $P(\vec{\mu} | \vec{\alpha})$

$\vec{\alpha} = \alpha(1, 4, 2)$, $\alpha \geq 1$

$\sum \alpha_k' = 7\alpha$

$$\left.\begin{array}{l} N_1 = 50 \\ N_2 = 29 \\ N_3 = 40 \end{array}\right\} N = 119$$

$$P(\vec{\mu} | \vec{\alpha}) = \prod_{k=1}^{K} \mu_k^{\alpha_k - 1} \cdot \frac{1}{B(\vec{\alpha})}$$

<u>MAP proc.</u>

$$\hat{\mu}_{MAP} = \underset{\vec{\mu}}{\operatorname{argmax}}\, P(D|\vec{\mu}) P(\vec{\mu}|\alpha) = \underset{\vec{\mu}}{\operatorname{argmax}}\, P(\vec{\mu}|D)$$

$$= \underset{\vec{\mu}}{\operatorname{argmax}}\, P(\vec{x}|\vec{\mu}) P(\vec{\mu}|\alpha)$$

$$= \underset{\vec{\mu}}{\operatorname{argmax}}\, \mathcal{L}(\vec{\mu}|\vec{x}) P(\vec{\mu})$$

$$\begin{bmatrix} N = 119 \\ K = 3 \end{bmatrix}$$

$$\hat{\mu}_{k,MAP} = \frac{\alpha_k' - 1}{\sum \alpha_k' - K} = \frac{N_k + \alpha_k - 1}{\sum \alpha_k + N - K} = \frac{\alpha_k + N_k - 1}{7\alpha + 116}$$

$$\vec{\mu}_{MAP} = \left(\frac{\alpha + 49}{7\alpha + 116}, \frac{4\alpha + 28}{7\alpha + 116}, \frac{2\alpha + 38}{7\alpha + 116} \right)$$

$4\alpha + 28 > \alpha + 49$ i $4\alpha + 28 > 2\alpha + 38$

$3\alpha > 21$ $\quad$ $2\alpha > 11$

$\alpha > 7$ $\quad$ $\alpha > 5.5$

$\Rightarrow \boxed{\alpha > 7}$ C

2.6.

$\vec{x} = (x_1, \dots, x_6)$
$k = 6$

$\vec{\mu}_{MAP}$

- apriorna PDF → Dirichlet

- realizacija → češće 6-ica (u stvarnosti)
- opažanje → najčešće 5-ica, ali mali broj opažanja

$\vec{\alpha}$ određuje vjerojatnost $\vec{\mu}$, uz $\sum_{k=1}^{K} \mu_k = 1$

Uz koji $\vec{\alpha}$ će naša procjena za $\vec{\mu}$ biti najbliža stvarnoj vrijednosti tih parametara?

a) $\vec{\alpha} = (1, 1, 1, 1, 1, 1)$
→ svaka komb. $\vec{\mu}$ jednako vjerojatna (ne unosimo apriorno znanje, vjerujemo podacima)

b) $\vec{\alpha} = (1/6, 1/6, 1/6, 1/6, 1/6, 1/6)$
→ najvjerojatniji $\vec{\mu}$ na rubovima simpleksa

c) $\vec{\alpha} = (2, 2, 2, 2, 2, 2)$
→ $\vec{\mu}$ najvjerojatnije svi međusobno jednaki ($\mu_i = \mu_j$, jer je $\alpha_i > 1$)

d) $\vec{\alpha} = (1, 1, 1, 1, 3, 1)$
→ asimetrične kombinacije davat će veću vjerojatnost kombinacijama kod kojih μ_k nisu jednaki (najčešći je 5. ishod)

2.8.

Beta-Bernoulli model

$\mu_{MAP} = ?$

$\mu_{MLE} = 0.2 = \frac{m}{N} \rightarrow m = 0.2N$

$$\mu_{MAP} = \frac{\alpha' - 1}{\alpha' + \beta' - 2} = \frac{\alpha + m - 1}{\alpha + \beta + N - 2} = 0.2$$

$$\alpha + m - 1 = 0.2\alpha + 0.2\beta + 0.2N - 0.4$$

$$0.8\alpha - 0.2\beta = 0.6$$
$$8\alpha - 2\beta = 6$$
$$4\alpha - \beta = 3$$

$\hookrightarrow \beta = 4\alpha - 3$

$\alpha = 2$
$\beta = 5$ → 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 klasifkatora 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 klasifikaciji 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 vjerojatnost 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 dijagonalnom kovarijacijskom matricom.

   (b) Napišite broj parametara ovog modela.

   (c) Objasnite zašto je izglednost faktorizirana u produkt univarijatnih 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 dijagonalnom (i dijeljenom) kovarijacijskom matricom i modelom $\mathcal{H}_3$ s izotropnom kovarijacijskom matricom.

(a) Koliko svaki od ova tri modela ima parametara?

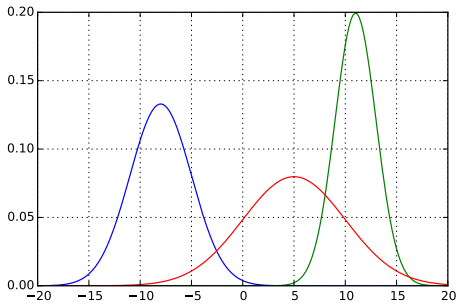
(b) Za koji od ova tri modela očekujete da će najbolje generalizirati u ovom konkretnom slučaju (uzmite u obzir prirodu problema i očekivane odnose između značajki)? Zašto?

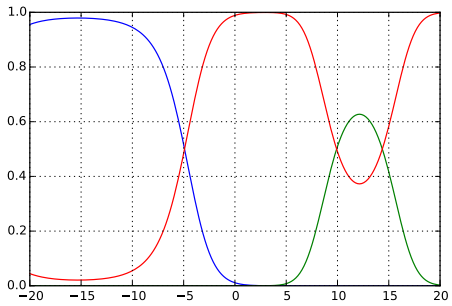
(c) Nacrtajte skicu funkcije empirijske pogreške i pogreške generalizacije i naznačite na njoj točke koje označavaju navedenim trima modelima.

(d) Kako biste u praksi odredili koji ćete model upotrijebiti?

2 Zadatci 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 multivarijantnom Gaussovom gustoćom vjerojatnosti:

$$p(\mathbf{x}|y = j) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}_j|^{1/2}} \exp\Big\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_j)^{\mathrm{T}}\boldsymbol{\Sigma}_j^{-1}(\mathbf{x} - \boldsymbol{\mu}_j)\Big\}$$

gdje je Σ_j matrica kovarijacije za klasu j. Treniranjem modela dobili smo sljedeće procjene za parametre:

$$\hat{\mu}_1 = 0.2 \qquad \hat{\boldsymbol{\mu}}_1 = (1, 0, -2) \qquad \hat{\Sigma}_1 = \begin{pmatrix} 5 & 2 & 4 \\ 2 & 5 & 3 \\ 4 & 3 & 6 \end{pmatrix}$$

$$\hat{\mu}_2 = 0.8 \qquad \hat{\boldsymbol{\mu}}_2 = (2, -1, 5) \qquad \hat{\Sigma}_2 = \begin{pmatrix} 6.25 & -0.5 & -1 \\ -0.5 & 1.25 & -0.75 \\ -1 & -0.75 & 3.5 \end{pmatrix}$$

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 klasfikaciju jednodimenzijskih podataka u tri klase. Procijenjene izglednosti klasa 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 klasa 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 izmijeni, i to tako da se varijanca izglednosti klase $y = 1$ postavi na 5 i da se apriorne vjerojatnosti klasa 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 $-10 \leqslant x \leqslant 10$. **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 dvodimenzijskome ulaznom prostoru ($\mathcal{X} = \mathbb{R}^2$). Apriorne vjerojatnosti klasa su jednake, dok su izglednosti klasa 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 ulaznome 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:

$$p(\mathbf{x}|y = j) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}_j|^{1/2}} \exp\Big\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_j)^{\mathrm{T}}\boldsymbol{\Sigma}_j^{-1}(\mathbf{x} - \boldsymbol{\mu}_j)\Big\}$$

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.

$$\left[h_j(\vec{x}; \vec{\theta}) = P(y=j|\vec{x}) = \frac{p(\vec{x}|y)P(y)}{\sum_{y'} p(\vec{x}|y')P(y')} = \frac{p(\vec{x}|y)P(y)}{p(\vec{x})} \right]$$

$p(\vec{x}|y)$ = izglednost klase
- gustoća vjerojatnosti primjera uz zadanu klasu
- „vjerojatnost primjera za zadanu klasu"

$P(y)$ = apriorna vjerojatnost klase

$p(\vec{x})$ = gustoća vjerojatnosti primjera neovisno o klasi

- brojnik faktoriziramo kako bismo lakše modelirali složenu zajedničku distribuciju $p(\vec{x}, y)$
- $p(\vec{x})$ dolazi iz Bayesovog pravila i modelira gustoću vjerojat. primjera
  - slično kao i PDF $p(\vec{x}, y)$ može biti složena distribucija pa ju faktoriziramo na $\sum p(x|y')p(y')$
- ako nas ne zanima vjerojatnost klasifikacije primjera u klasu j nazivnik možemo zanemariti pa model postaje
  $h(\vec{x}; \vec{\theta}) = \underset{y}{\operatorname{argmax}}\, p(\vec{x}|y)P(y)$
  → MAP HIPOTEZA

b) Bayesov klasifikator jest parametarski model
↳ model pretpostavlja da se primjeri $\vec{x}$ i oznake y pokoravaju nekoj teoretskoj distribuciji

c) Bayesov klasifikator nazivamo generativnim jer modelira zajedničku vjerojatnost $p(\vec{x}, y)$
→ modelira postupak generiranja podataka

GENERATIVNA PRIČA BAY. KLAS.
Primjeri $D = \{(\vec{x}^i, y^i)\}_i$ nastaju u 2 koraka

1. odabrana je oznaka y po distribuciji $P(y)$
2. Za odabranu oznaku y odabran je primjer $\vec{x}$ po distribuciji $p(\vec{x}|y)$

d)

Generativni modeli

- modeliraju zajed. distribuciju $P(\vec{x}, y)$

Diskriminativni modeli

- izravno modeliraju aposteriornu vjerojatnost $P(y|\vec{x})$
- dobiva se decizijska granica između klasa

Prednosti gener. modela

1. Lako ugraditi apriorno znanje o problemu
   → apriorna distribucija
2. intuitivna interpretabilnost rezultata i mogućnost različitih analiza

- ocjena pouzdanosti klasifikacije
  → izlaz modela moguće tumačiti kao pouzdanost
- odbijanje klasifikacije
  ↳ pouzdanost manja od praga
- Lako detektiranje stršećih vrijednosti

Nedostaci generat. modela

1. velik broj parametara
2. nepotrebna složenost modeliranja

1.2.

apriorne vjeroj.

$P(y=1) = 0.2$
$P(y=2) = 0.05$
$P(y=3) = 0.75$

$P(y|\vec{x}) = ? = h_j(\vec{x})$

$$h_j(\vec{x}) = \frac{P(\vec{x}|y)P(y)}{P(\vec{x})}$$

izgl

$P(\vec{x}|y=1) = 0.8$
$P(\vec{x}|y=2) = 0.5$
$P(\vec{x}|y=3) = 0.5$

$$P(\vec{x}) = \sum_{y'} P(\vec{x}|y')P(y')$$
$$= 0.8 \cdot 0.2 + 0.5 \cdot 0.05 + 0.5 \cdot 0.75$$
$$= 0.56$$

$$h_1(\vec{x}) = P(y=1|\vec{x}) = \frac{P(\vec{x}|y=1)P(y=1)}{P(\vec{x})} = \frac{2}{7} \approx 0.286$$

$$h_2(\vec{x}) = P(y=2|\vec{x}) = \frac{5}{112} \approx 0.0446$$

$$h_3(\vec{x}) = P(y=3|\vec{x}) = \frac{75}{112} \approx 0.67$$

MAP hipoteza: $h(\vec{x}) = \underset{j}{\operatorname{argmax}}\, P(\vec{x}|y=j)P(y=j) = \underset{j}{\operatorname{argmax}} \{0.16, 0.025, 0.375\}$
$= 3$ (klasa „normal")

1.3.

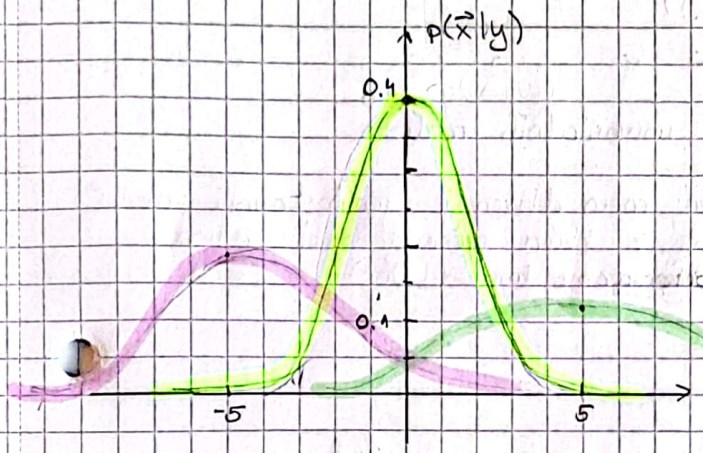
$K=3$ $\qquad X = \mathbb{R}$ $\qquad p(\vec{x}|y)$

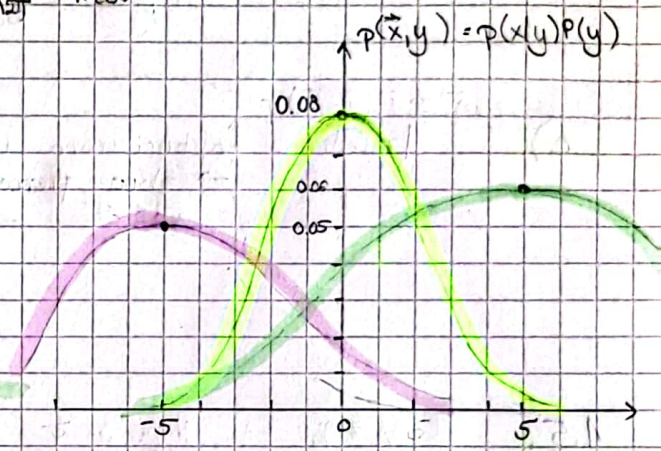
0.0534	$P(y=1) = 0.3$	$\mu_1 = -5$	$\sigma_1^2 = 5$	$\mathcal{N}(-5, 5)$	0.178
0.08	$P(y=2) = 0.2$	$\mu_2 = 0$	$\sigma_2^2 = 1$	$\mathcal{N}(0, 1)$	0.4
0.063	$P(y=3) = 0.5$	$\mu_3 = 5$	$\sigma_3^2 = 10$	$\mathcal{N}(5, 10)$	0.126

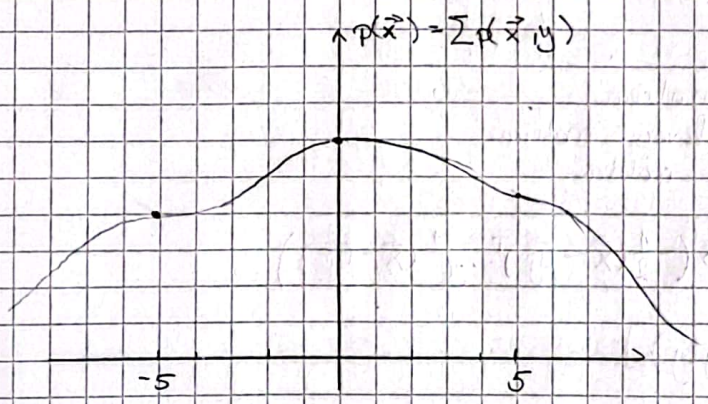
$$p(x|\mu,\sigma,y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

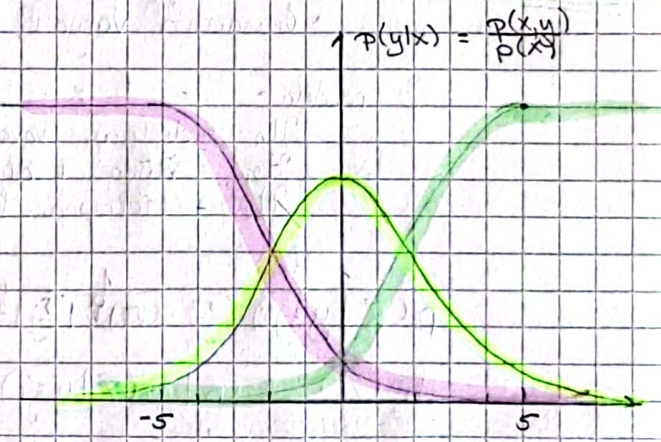
→ za mod Gaussa za x uvrstiti $x = \mu$

→ $\frac{1}{\sigma\sqrt{2\pi}}$ mod









1.4.

a) model višedim. Bayesov. klasifikatora

- kontinuirani ulaz
- dijeljena dijagonalna kovarijacijska matrica

$\Sigma^{-1} = \mathrm{diag}(1/\sigma_i^2)$

$|\Sigma| = \prod_{i=1}^{n} \sigma_i^2$

$$h_j(\vec{x}) = \ln\left[p(\vec{x}|y=j)P(y=j)\right] = \ln p(\vec{x}|y) + \ln P(y=j)$$

$$\ln\left[p(\vec{x}|y=j)\right] = \ln\left[\frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu}_j)^T \Sigma^{-1} (\vec{x}-\vec{\mu}_j)\right)\right]$$

$$= \ln \frac{1}{(2\pi)^{\frac{n}{2}} \prod_{i=1}^{N} \sigma_i} \exp\left(-\frac{1}{2}\sum_{i=1}^{N} \frac{1}{\sigma_i^2}(x^i - \mu_j^i)^2\right)$$

$$= \ln \prod_{i=1}^{N} \frac{1}{\sigma_i\sqrt{2\pi}} \prod_{i=1}^{N} \exp\left(-\frac{1}{2\sigma_i^2}(x^i - \mu_j^i)^2\right)$$

$$= \ln \prod_{i=1}^{N} \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x^i - \mu_j^i}{\sigma_i}\right)^2\right) = \ln \prod_{i=1}^{N} \mathcal{N}(\mu_j^i, \sigma_i^2)$$

$$\left[h_j(\vec{x}) = \ln p(\vec{x} \mid y=j) + \ln P(y=j) \right.$$

$$= \dots$$

$$\left. = -\frac{1}{2} \sum_{i=1}^{n} \left(\frac{x_i - \mu_j^i}{\sigma_i} \right)^2 + \ln P(y=j) \right]$$

b) $$\#\text{param} = \underbrace{K \cdot n}_{\vec{\mu}_j} + \underbrace{n}_{\sigma_i^2} + \underbrace{K-1}_{P(y=j)}$$

c) Izglednost faktorizirana u produkt univarijatnih razdioba
→ zbog preraučenosti
⇒ na ovaj način dobiven je jednostavniji model koji je manje sklon preraučenosti i ima manji broj parametara ($O(n)$)

15. $K = 2$
- 6 ulaznih varijabli ($n = 6$)

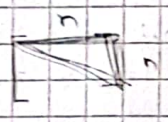
3 modela
$\mathcal{H}_1$ = dijeljena kovarij. matrica
$\mathcal{H}_2$ = dijag. i dijelj. kovarij. matrica
$\mathcal{H}_3$ = izotropna kovarij. matrica

$$p(\vec{x} \mid y=j) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma_j|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (\vec{x} - \vec{\mu}_j)^T \Sigma_j^{-1} (\vec{x} - \vec{\mu}_j) \right)$$

$$h_j(\vec{x}) = \ln \left[p(\vec{x} \mid y=j) P(y=j) \right]$$

$$= \ln p(\vec{x} \mid y) + \ln P(y=j)$$

$$= \underbrace{-\frac{n}{2} \ln 2\pi}_{\text{konst.}} - \frac{1}{2} \ln |\Sigma_j| - \frac{1}{2} (\vec{x} - \vec{\mu}_j)^T \Sigma_j^{-1} (\vec{x} - \vec{\mu}_j) + \ln P(y=j)$$



a) $\mathcal{H}_1$

$$\#\text{param} \quad \underbrace{\frac{n}{2}(n+1)}_{\text{gornji } \Delta\, \Sigma_j} + \underbrace{K \cdot n}_{\vec{\mu}_j \text{ (centroidi)}} + \underbrace{K-1}_{\text{priori}} \qquad O(n^2)$$

$\mathcal{H}_2$

$$\#\text{param} \quad n + K \cdot n + K - 1$$

$\mathcal{H}_3$

$$\#\text{param:} \quad 1 + K \cdot n + K - 1$$

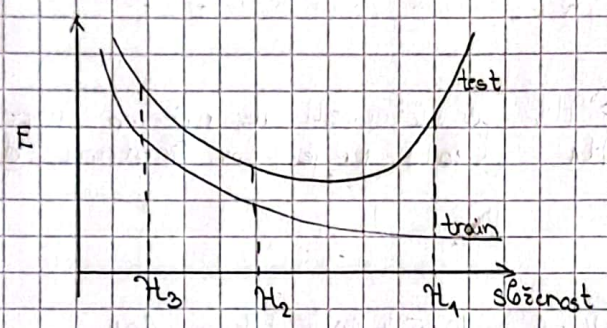
$K = 2$
$n = 6$

	# param
$\mathcal{H}_1$	34
$\mathcal{H}_2$	19
$\mathcal{H}_3$	14

b) Očekujemo da će najbolje generalizirati model $\mathcal{H}_2$

- model $\mathcal{H}_1$
  - sklon prenaučenosti - modeliraju se korelacije svih značajki - u ovoj primjeni loše
- model $\mathcal{H}_3$
  - podnaučen model
  - razlike u skalama značajki
- model $\mathcal{H}_2$
  - nosjet. 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!

$$p(y|x) = \frac{p(x|y)\,p(y=j)}{p(x)}$$

$p(x)$ → za sve isti

→ sjecišta aposteriornih distribucija <=> sjecišta zajedničkih vjerojatnosti

Plava ($y=1$) i crvena ($y=2$) klasa

- aposteriorno sjecište ≈ -5
- sjecište izglednosti ≈ -3

↑ pomak ulijevo => crvena krivulja širila više od plave

„lijevi rep"

=> $P(y=2) > P(y=1)$

Crvena ($y=2$) i zelena ($y=3$) klasa

① apost. sjecište ≈ 10
sjecište izgled. ≈ 7 ↑ pomak desno
„lijevi rep"

② apost. sjecište ≈ 14
sjecište izgled. ≈ 16 ↑ pomak lijevo
„desni rep"

crvena krivulja širila se više od zelene

=> $P(y=2) > P(y=3)$

=> (A)

2.2. $K = 10$
$n = 5$

$\mathcal{H}_1$ = nedijeljena + dijagonalna Σ_j
$\mathcal{H}_2$ = nedijeljena + izotropna Σ_j
$\mathcal{H}_3$ = puna, dijeljena Σ_j

$$p(x|y) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma_j|^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2} (\vec{x} - \vec{\mu}_j)^T \Sigma^{-1} (\vec{x} - \vec{\mu}_j) \right\}$$

$$h_j(\vec{x}) = -\frac{n}{2} \ln(2\pi) - \frac{1}{2} \ln|\Sigma_j| - \frac{1}{2} (\vec{x} - \vec{\mu}_j)^T \Sigma^{-1} (\vec{x} - \vec{\mu}_j) + \ln P(y=j)$$

<u>Složenost</u>: $\mathcal{H}_1 \supset \mathcal{H}_2$
$\Rightarrow \mathcal{H}_3$ po složenosti ne možemo usporediti s $\mathcal{H}_1$ (zato jer je kov. matrica dijeljena)

	#param
$\mathcal{H}_1$	$K \cdot n + K \cdot n + K - 1 = 10 \cdot 5 + 10 \cdot 5 + 4 = 104$
$\mathcal{H}_2$	$K + K \cdot n + K - 1 = 10 + 50 + 4 = 64$
$\mathcal{H}_3$	$\frac{n}{2}(n+1) + K \cdot n + K - 1 = \frac{5 \cdot 6}{2} + 10 \cdot 5 + 4 = 69$

$$\left.\begin{array}{l} \mathcal{H}_1 > \mathcal{H}_3 > \mathcal{H}_2 \\ \mathcal{H}_1 \supset \mathcal{H}_2 \end{array}\right\} \text{A}$$

2.3. $n = 3$
$K = 2$
• dijeljena kovarijacijska matrica

$P(y=1) = \hat{\mu}_1 = 0.2$ $\quad \hat{\vec{\mu}}_1 = (1, 0, -2)$
$\hat{\mu}_2 = 0.8$ $\quad \hat{\vec{\mu}}_2 = (2, -1, 5)$

$$\hat{\Sigma}_1 = \begin{bmatrix} 5 & 2 & 4 \\ 2 & 5 & 3 \\ 4 & 3 & 6 \end{bmatrix} \qquad \hat{\Sigma}_2 = \begin{bmatrix} 6.25 & -0.5 & -1 \\ -0.5 & 1.25 & -0.75 \\ -1 & -0.75 & 3.5 \end{bmatrix}$$

$$\hat{\Sigma} = \sum_j \hat{\mu}_j \hat{\Sigma}_j = \begin{bmatrix} 6 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 4 \end{bmatrix} \qquad |\hat{\Sigma}| = 6 \cdot 2 \cdot 4 = 48 \qquad \hat{\Sigma}^{-1} = \begin{bmatrix} 1/6 & 0 & 0 \\ 0 & 1/2 & 0 \\ 0 & 0 & 1/4 \end{bmatrix}$$

$\vec{x} = (0, 0, 0)$

$$h_1(\vec{x}) = -\frac{1}{2} \ln|\hat{\Sigma}| - \frac{1}{2} (\vec{x} - \hat{\vec{\mu}}_1)^T \hat{\Sigma}^{-1} (\vec{x} - \vec{\mu}_1) + \ln P(y=j) - \frac{n}{2} \ln(2\pi)$$

$$= -\frac{1}{2} \ln|\hat{\Sigma}| - \frac{1}{2} \cdot \frac{7}{6} + \ln 0.2 - \frac{3}{2} \ln(2\pi)$$

$$= -\frac{1}{2} \ln 48 - \frac{7}{12} + \ln 0.2 - \frac{3}{2} \ln(2\pi)$$

$$= -6.885 \qquad \text{A}$$

2.4.

$K = 3$
$n = 1$

Ispravci
$\sigma^2 = 5$

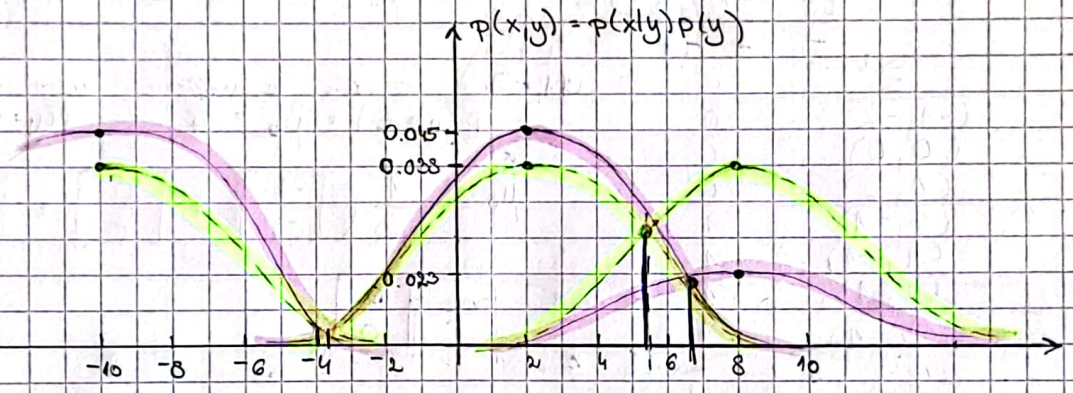
prije

$p(x|y=1) = \mathcal{N}(-10, 2)$
$p(x|y=2) = \mathcal{N}(2, 2)$
$p(x|y=3) = \mathcal{N}(8, 2)$

$P(y=1) = \frac{2}{5}$
$P(y=2) = 2/5$
$P(y=3) = 1/5$

$P(y=1) = 1/3$
$P(y=2) = 1/3$
$P(y=3) = 1/3$

$$p(x|y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$



— = prije ispravka
- - - = nakon ispravka

$[-4, -4+a]$ ← razlika h_1
$[5, 5+b]$ ← razlika h_2

$h_1(\vec{x}) = p(x|y=1)\,P(y=1)$
$h_2(\vec{x}) = p(x|y=2)\,P(y=2)$

D

2.5.

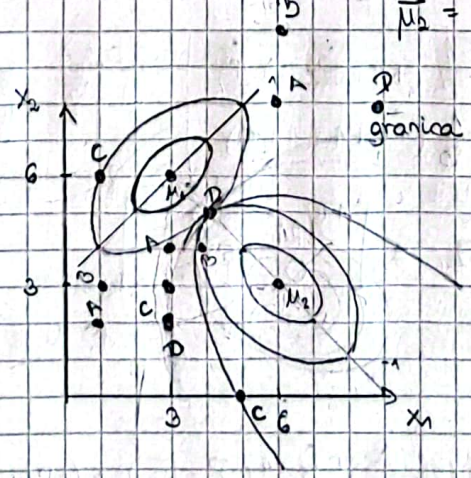
$K = 2$
$n = 2$

$P(y=1) = P(y=2) = 0.5$

Parametri izgledanosti

$\vec{\mu}_1 = (3, 6)$ $\quad \Sigma_1 = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$ ← pozitiv. korel.

$\vec{\mu}_2 = (6, 3)$ $\quad \Sigma_2 = \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix}$ ← neg. korel.



$$\vec{x}^T(\Sigma_1^{-1} - \Sigma_2^{-1})\vec{x} = \vec{x}^T \begin{bmatrix} 1 & -2 \\ -2 & -1 \end{bmatrix} \vec{x}$$

$\boxed{D}$ $(3,2), (4,5), (9,7)$

2.6.

$n = 2$
$k = 2$

• dijeljena i dijagonalna Σ

$h_j(\vec{x}) = \ln p(\vec{x}, y=j)$

$h_0(\vec{x}) = ?$
$\vec{x} = (0,0)$

$$h_0(\vec{x}) = \ln p(\vec{x}|y=0) + \ln P(y=0)$$

$$= -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln|\hat{\Sigma}| - \frac{1}{2}(\vec{x}-\hat{\vec{\mu}}_0)^T \hat{\Sigma}^{-1}(\vec{x}-\hat{\vec{\mu}}_0) + \ln P(y=0)$$

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

$N = 7$

$P(y=0) = \hat{\mu}_0 = \frac{3}{7}$ $\quad P(y=1) = \hat{\mu}_1 = \frac{4}{7}$

$$\hat{\vec{\mu}}_0 = \frac{1}{3}\left(\begin{bmatrix}-1\\-2\end{bmatrix} + \begin{bmatrix}0\\0\end{bmatrix} + \begin{bmatrix}1\\2\end{bmatrix}\right) = \begin{bmatrix}0\\0\end{bmatrix}$$

$$\hat{\Sigma}_0 = \frac{1}{N_0-1}\left(\underbrace{\begin{bmatrix}-1\\-2\end{bmatrix}}_{2\times 1}\underbrace{\begin{bmatrix}-1 & -2\end{bmatrix}}_{1\times 2} + \begin{bmatrix}0\\0\end{bmatrix}\begin{bmatrix}0 & 0\end{bmatrix} + \begin{bmatrix}1\\2\end{bmatrix}\begin{bmatrix}1 & 2\end{bmatrix}\right)$$

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

Napomena
Svejedno koristiš li
$\hat{\Sigma}_j = \frac{1}{N}\sum(x-\mu)(x-\mu)^T$
ili procjenjuješ Cov i Var za svaku četiju zasebno!
Samo pazi na faktor pristranosti

$$\hat{\vec{\mu}}_1 = \frac{1}{4}\cdot\begin{bmatrix}3+4+4+5\\-1-1+1+1\end{bmatrix} = \begin{bmatrix}4\\0\end{bmatrix}$$

$$\hat{\Sigma}_1 = \begin{bmatrix}Var(X_1) & Cov(X_1,X_2)\\ & Var(X_2)\end{bmatrix} = \begin{bmatrix}\frac{1}{N_1-1}\sum(x_{i1}-\mu_{11})^2 & \frac{1}{N_1-1}\sum(x_{i1}-\mu_{11})(x_{i2}-\mu_{12})\\ & \frac{1}{N_1-1}\sum(x_{i2}-\mu_{12})^2\end{bmatrix}$$

$$= \begin{bmatrix}\frac{1}{3}(1+0+0+1) & \frac{1}{3}(+1+0+0+1)\\ & \frac{1}{3}(+1+1+1+1)\end{bmatrix} = \begin{bmatrix}\frac{2}{3} & \frac{2}{3}\\ \frac{2}{3} & \frac{5}{3}\end{bmatrix}$$

$$\hat{\Sigma} = \sum \mu_j \hat{\Sigma}_j = \frac{3}{7}\hat{\Sigma}_0 + \frac{4}{7}\hat{\Sigma}_1 = \begin{bmatrix}17/21 & 0\\ 0 & 52/21\end{bmatrix}$$

piše u zad. da je dijag.!

$$|\hat{\Sigma}| = \frac{17}{21}\cdot\frac{52}{21}$$

$$h_0(\vec{x}) = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln|\hat{\Sigma}| - \underbrace{\frac{1}{2}(\vec{x}-\hat{\vec{\mu}}_0)^T\hat{\Sigma}^{-1}(\vec{x}-\hat{\vec{\mu}}_0)}_{0} + \ln P(y=0)$$

$$= -\ln(2\pi) - \frac{1}{2}\ln\frac{17\cdot 52}{21^2} + \ln\frac{3}{7} = -3.03 \Rightarrow \text{C}$$

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 probabilističku interpretaciju logističke regresije. Razumjeti razliku u broju parametara između diskriminativnog i generativnog modela te utjecaj broja klasa i broja primjera na taj odnos.***]

   (a) 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.

   (b) Model logističke regresije koristimo za binarnu klasifikaciju primjera s $n = 100$ značajki. Odredite broj parametara modela logističke regresije te njemu odgovarajućeg generativnog modela.

   (c) Izračunajte broj parametara za isti slučaj, ali sa $K = 5$ klasa.

   (d) Pretpostavite 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

   (a) Izračunajte MLE procjene svih parametara modela te klasificirajte primjere $(\text{Istra}, \text{ne}, \text{kamp}, \text{bus})$ i $(\text{Dalmacija}, \text{da}, \text{hotel}, \text{bus})$.

   (b) Izračunajte Laplaceove (zaglađene) procjene za sve parametre modela te klasificajte nanovo iste primjere.

3. [***Svrha: Razviti intuiciju o uvjetnoj nezavisnosti i odnosu između nezavisnosti i uvjetne nezavisnosti.***]

   (a) Definirajte uvjetnu nezavisnost slučajnih varijabli. Pokažite da je definicija pomoću zajedničke vjerojatnosti istovjetna definiciji pomoću uvjetne vjerojatnosti.

   (b) Za sljedeće primjere razmotrite sve parove varijabli i odredite za koje parove možemo pretpostaviti nezavisnost odnosno uvjetnu nezavisnost:

      i. $P \equiv$ danas je ponedjeljak, $S =$ danas je subota, $L \equiv$ danas je listopad.

      ii. $S \equiv$ 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 lienarne 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 $I(X, Y) \geqslant 0$.

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 $\mathcal{C}_1 = \mathit{VisokRizik}$, $\mathcal{C}_2 = \mathit{UmjerenRizik}$, $\mathcal{C}_3 = \mathit{NizakRizik}$. 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 $\mathcal{H}_\alpha$ kod kojeg se združivanje varijabli provodi za sve parove varijabli (x_i, x_j) za koje $I(x_i, x_j) \geqslant \alpha$. Skicirajte pogreške učenja i ispitivanja modela $\mathcal{H}_\alpha$ 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 ulaznome 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 ulaznome 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-procjeniteljem. Zanima nas klasifikacija sljedećeg primjera:

$$\mathbf{x} = (\text{Istra}, \text{ne}, \text{kamp}, \text{bus})$$

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 $P(\text{stožer}|y = 1) = P(\text{stožer}|\text{pandemija}, y = 1)$

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	GovoriJezik	$y^{(i)}$
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-procjeniteljem. Zanima nas klasifikacija za $\mathbf{x} = (\text{Italija}, \text{ne}, \text{zimski}, \text{dipl}, \text{ne})$. **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 orijentacija (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
$x_1 = 1$	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)

$$P(y=1|x) = \frac{p(x|y=1)P(y=1)}{p(x)}$$

$$= \frac{p(x|y=1)P(y=1)}{p(x|y=1)P(y=1)+p(x|y=0)P(y=0)}$$

$$= \frac{1}{1+\frac{p(x|y=0)P(y=0)}{p(x|y=1)P(y=1)}}$$

$$= \frac{1}{1+\exp\Big(\underbrace{\ln\frac{p(x|y=0)P(y=0)}{p(x|y=1)P(y=1)}}_{-\alpha}\Big)}$$

$$= \frac{1}{1+\exp(-\alpha)} = \sigma(\alpha) = h_{LR}(\vec{x})$$

model logističke regresije uz donji uvjet da $\alpha = \vec{w}^T x + w_0$

$$\alpha = -\ln\frac{p(x|y=0)P(y=0)}{p(x|y=1)P(y=1)} = \underbrace{\ln[p(x|y=1)P(y=1)]}_{h_1(\vec{x})} - \underbrace{\ln[p(x|y=0)P(y=0)]}_{h_0(\vec{x})}$$

$$= \vec{w}^T x + w_0$$

$h_1(\vec{x})$ i $h_0(\vec{x})$ - Bayesov klasifikator

- gornji uvjet svodi se na uvjet da granica između klasa mora biti <u>LINEARNA</u> $\Leftrightarrow$ dijeljena kovarijacijska matrica

$$h_j(\vec{x}) = \ln[p(\vec{x}|y=j)P(y=j)]$$

$$= \ln\frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(\vec{x}-\vec{\mu}_j)^T\Sigma^{-1}(\vec{x}-\vec{\mu}_j)\right\} + \ln P(y=j)$$

$$= -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma| - \frac{1}{2}(\vec{x}-\vec{\mu}_j)^T\Sigma^{-1}(\vec{x}-\vec{\mu}_j) + \ln P(y=j)$$

$$h_1(\vec{x}) - h_0(\vec{x}) = \cancel{-\frac{n}{2}\ln(2\pi)} \cancel{-\frac{1}{2}\ln|\Sigma|} - \frac{1}{2}(\vec{x}-\vec{\mu}_1)^T\Sigma^{-1}(\vec{x}-\vec{\mu}_1) + \ln P(y=1)$$
$$\cancel{+\frac{n}{2}\ln(2\pi)} \cancel{+\frac{1}{2}\ln|\Sigma|} + \frac{1}{2}(\vec{x}-\vec{\mu}_0)^T\Sigma^{-1}(\vec{x}-\vec{\mu}_0) - \ln P(y=0)$$

$$= -\frac{1}{2}(\cancel{\vec{x}^T\Sigma^{-1}\vec{x}} - \vec{x}^T\Sigma^{-1}\vec{\mu}_1 - \vec{\mu}_1^T\Sigma^{-1}\vec{x} + \vec{\mu}_1^T\Sigma^{-1}\vec{\mu}_1) + \ln P(y=1)$$
$$+\frac{1}{2}(\cancel{\vec{x}^T\Sigma^{-1}\vec{x}} - \underbrace{\vec{x}^T\Sigma^{-1}\vec{\mu}_0}_{1\times n\cdot n\times n\cdot n\times 1 \,=\, 1\times 1} - \underbrace{\vec{\mu}_0^T\Sigma^{-1}\vec{x}}_{1\times 1} + \underbrace{\vec{\mu}_0^T\Sigma^{-1}\vec{\mu}_0}_{1\times 1}) + \ln P(y=0)$$

$h_1(\vec{x}) - h_0(\vec{x}) =$

$$= -\frac{1}{2}\left(-2\vec{x}^T\Sigma^{-1}\vec{\mu_1} + \vec{\mu_1}^T\Sigma^{-1}\vec{\mu_1}\right) + \frac{1}{2}\left(-2\vec{x}^T\Sigma^{-1}\vec{\mu_0} + \vec{\mu_0}^T\Sigma^{-1}\vec{\mu_0}\right) + \ln\frac{P(y=1)}{P(y=0)}$$

$$= \vec{x}^T\underbrace{\Sigma^{-1}(\vec{\mu_1}-\vec{\mu_0})}_{\vec{w}} \underbrace{-\frac{1}{2}\vec{\mu_1}^T\Sigma^{-1}\vec{\mu_1} + \frac{1}{2}\vec{\mu_0}^T\Sigma^{-1}\vec{\mu_0} + \ln\frac{P(y=1)}{P(y=0)}}_{w_0}$$

$$= \vec{w}^T\vec{x} + w_0 \quad //$$

$$\boxed{\alpha = \vec{w}^T\vec{x} + w_0}$$

→ uz pretpostavku dijelјene kovarijacijske matrice Bayesov klasifikator jednak je modelu logističke regresije

b) $n = 100$

$$\#param_LR = \underset{\vec{w}}{n} + \underset{w_0}{1} = 101 \text{ parametar}$$

$$h_{LR}(\vec{x}) = \sigma(\alpha) = \sigma(\vec{w}^T\vec{x} + w_0) = \frac{1}{1+\exp(-\vec{w}^T x + w_0)}$$

• odg. generativni model

⇒ Bayesov klasifikator ($K=2$, dij. Σ)

$$\#param_BK = \underbrace{\frac{n}{2}(n+1)}_{\Sigma} + \underbrace{n\cdot K}_{\vec{\mu_j}} + \underbrace{K-1}_{P(y=j)} = \frac{100}{2}\cdot 101 + 100\cdot 2 + 1 = 5251$$

$$h_{BK,j}(\vec{x}) = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma| - \frac{1}{2}(\vec{x}-\vec{\mu_j})^T\Sigma^{-1}(\vec{x}-\vec{\mu_j}) + \ln P(y=j)$$

c) $K = 5$

$$\#param_LR = \binom{K}{2}(n+1) = 1010$$

↳ OVO shema

$$\#param_BK = \frac{n}{2}(n+1) + n\cdot K + K - 1 = \frac{100}{2}\cdot 101 + 5\cdot 100 + 4 = 5554$$

d) $K = 10$

$$\binom{K}{2}(n+1) \leq \frac{n}{2}(n+1) + Kn + K - 1$$

$$45n + 45 \leq \frac{n^2}{2} + \frac{n}{2} + 10n + 9 \quad / \cdot 2$$

$$90n + 90 \leq n^2 + n + 20n + 18$$

$$n^2 - 69n - 72 \geq 0$$

$$n \leq -1.03 \quad i \quad n \geq 70.03$$

$$\boxed{n \geq 71}$$

→ dimenzija prostora značajki mora biti 71 ili više da bi se logistička regresija isplatila

1.2.

• naivan Bayesov klasifikator

$$h_{NB}(\vec{x}) = \underset{j}{\operatorname{argmax}}\, p(\vec{x} \mid y=j)\, P(y=j)$$

$$= \underset{j}{\operatorname{argmax}} \prod_{k=1}^{n} P(x_k \mid y=j) \cdot P(y=j)$$

obično parametre procjenjujemo:

• MLE za $P(y=j) = \hat{\mu}_j = \frac{1}{N} \sum_{i=1}^{N} \{y^i = j\} N = \frac{N_j}{N}$

• MAP za $P(x_k \mid y=j) = \hat{\mu}_{kj} = \frac{\sum_{i=1}^{N} 1\{x_k^i = x_k \wedge y^i = j\} + 1}{\sum_{i=1}^{N} 1\{y^i = j\} + K_k}$

zaglađivanje ⇒

$$= \frac{N_{kj} + 1}{N_j + K_k}$$

broj vrijednosti značajke k u klasi j

→ broj klasa značajke k

a)

$P(x_1 = \text{Istra} \mid y = ne) = 0$
$P(x_1 = \text{Dalmacija} \mid y = ne) = 1/3$
$P(x_1 = \text{Kvarnerk} \mid y = ne) = 2/3$

$P(x_2 = ne \mid y = ne) = 1$
$P(x_2 = da \mid y = ne) = 0$

$P(x_3 = \text{privatni} \mid y = ne) = 1/3$
$P(x_3 = \text{kamp} \mid y = ne) = 2/3$
$P(x_3 = \text{hotel} \mid y = ne) = 0$

$P(x_4 = \text{auto} \mid y = ne) = 0$
$P(x_4 = \text{bus} \mid y = ne) = 2/3$
$P(x_4 = \text{avion} \mid y = ne) = 1/3$

$P(x_1 = \text{Istra} \mid y = da) = 1/2$
$P(x_1 = \text{Dalm} \mid y = da) = 1/2$
$P(x_1 = \text{Kvarn.} \mid y = da) = 0$

$P(x_2 = ne \mid y = da) = 1/4$
$P(x_2 = da \mid y = da) = 3/4$

$P(x_3 = \text{privatni} \mid y = da) = 1/2$
$P(x_3 = \text{kamp} \mid y = da) = 0$
$P(x_3 = \text{hotel} \mid y = da) = 1/2$

$P(x_4 = \text{auto} \mid y = da) = 3/4$
$P(x_4 = \text{bus} \mid y = da) = 0$
$P(x_4 = \text{avion} \mid y = da) = 1/4$

MLE
↓
za sve par. modela

Predikcija

$$h(\text{Istra}, \text{ne}, \text{kamp}, \text{bus}) = \underset{y}{\operatorname{argmax}}\, P(y) \prod P(x_i|y)$$

$y = ne$ $\quad h = \frac{3}{7} \cdot 0 \cdot 1 \cdot \frac{2}{3} \cdot \frac{2}{3} = 0$

$y = da$ $\quad h = \frac{4}{7} \cdot \frac{1}{2} \cdot \frac{1}{4} \cdot 0 \cdot 0 = 0$

ne možemo klasific.

$$h(\text{Dalm.}, \text{da}, \text{hotel}, \text{bus}) = \underset{y}{\operatorname{argmax}}\, P(y) \prod P(x_i|y)$$

$y = ne$ $\quad h = \frac{3}{7} \cdot \frac{1}{3} \cdot 0 \cdot 0 \cdot \frac{2}{3} = 0$

$y = da$ $\quad h = \frac{4}{7} \cdot \frac{1}{2} \cdot \frac{2}{4} \cdot \frac{1}{2} \cdot 0 = 0$

ne možemo klasificirati

b) Laplaceove progjene

$K_1 = 3$ $\quad N_{ne} = 3$

$K_2 = 2$ $\quad N_{da} = 4$

$K_3 = 3$

$K_4 = 3$

$P(x_1 = \text{Istra} | y = ne) = 1/6$ $\quad P(x_1 = \text{Istra} | y = da) = 3/7$

$P(x_1 = \text{Dalmacija} | y = ne) = 2/6$ $\quad P(x_1 = \text{Dalm.} | y = da) = 3/7$

$P(x_1 = \text{Kvarner} | y = ne) = 3/6$ $\quad P(x_1 = \text{Kvarner} | y = da) = 1/7$

$P(x_2 = ne | y = ne) = 4/5$ $\quad P(x_2 = ne | y = da) = 2/6$

$P(x_2 = da | y = ne) = 1/5$ $\quad P(x_2 = da | y = da) = 4/6$

$P(x_3 = \text{priv.} | y = ne) = 2/6$ $\quad P(x_3 = \text{priv.} | y = da) = 3/7$

$P(x_3 = \text{kamp} | y = ne) = 3/6$ $\quad P(x_3 = \text{kamp} | y = da) = 1/7$

$P(x_3 = \text{hotel} | y = ne) = 1/6$ $\quad P(x_3 = \text{hotel} | y = da) = 3/7$

$P(x_4 = \text{auto} | y = ne) = 1/6$ $\quad P(x_4 = \text{auto} | y = da) = 4/7$

$P(x_4 = \text{bus} | y = ne) = 3/6$ $\quad P(x_4 = \text{bus} | y = da) = 1/7$

$P(x_4 = \text{avion} | y = ne) = 2/6$ $\quad P(x_4 = \text{avion} | y = da) = 2/7$

$x = (\text{Istra}, \text{ne}, \text{kamp}, \text{bus})$

$y = ne$ $\quad h(x) = 4/9 \cdot 1/6 \cdot 4/5 \cdot 3/6 \cdot 3/6 = 2/135$

$y = da$ $\quad h(x) = 5/9 \cdot 3/7 \cdot 2/6 \cdot 1/7 \cdot 1/7 = 5/3087$

$h(x) = ne$

$x = (\text{Dalmacija}, \text{da}, \text{hotel}, \text{bus})$

$y = ne$ $\quad h(x) = 4/9 \cdot 2/6 \cdot 1/5 \cdot 1/6 \cdot 3/6 = 1/405$

$y = da$ $\quad h(x) = 5/9 \cdot 3/7 \cdot 4/6 \cdot 3/7 \cdot 1/7 = 10/1029$

$h(x) = da$

1.3.

a) • uvjetna nezavisnost slučajnih varijabli

= 2 varijable X i Y postaju nezavisne ako nam je poznat ishod treće varijable Z

vrijedi:

$$P(X|Y,Z) = P(X|Z)$$
$$P(Y|X,Z) = P(Y|Z)$$

• preko zajedničke vjerojatnosti

$$P(X,Y|Z) = P(X|Z)P(Y|Z)$$

ISTOVJETNOST

$$P(X,Y|Z) = P(X|Z)P(Y|X,Z) = P(X|Z)P(Y|Z)$$

ili

$$P(X,Y|Z) = P(Y|Z)P(X|Y,Z) = P(Y|Z)P(X|Z)$$

$$P(X|Y,Z) = P(X,Z)$$
$$P(Y|X,Z) = P(Y,Z)$$ "

b) I

P = danas je ponedjeljak
S = danas je subota
L = danas je listopad

Parovi $P \perp S \mid L$

$$P(P|S,L) \neq P(P|L)$$

=> nije uvjetno nezavisno

$\boxed{P \perp L \mid S}$

$$P(P|L,S) = P(P|S) \checkmark$$
$$P(L|P,S) = P(L|S) \checkmark$$

→ uvjetno nezavisno

$\boxed{S \perp L \mid P}$

$$P(S|L,P) = P(S|P) \checkmark$$
$$P(L|S,P) = P(L|P) \checkmark$$

II

S = sunčano je
V = vruće je
K = ljudi se kupaju

$S \perp V \mid K$ → nije uvjetno nezavisno

$\boxed{S \perp K \mid V}$

$V \perp K \mid S$ → nije uvjetno nezavisno

III

L = „lepta"
N = „nogomet"
S = „sport"

$L \perp N \mid S$ → nije uvj. nezavisno

$\boxed{L \perp S \mid N}$

$N \perp S \mid L$ → nije uvj. nezavisno

IV

K = pada kiša
C = puka cjev
M = „ulica" je mokra

$K \perp C \mid M$ → nije uvjetno nezavisno

$\boxed{K \perp M \mid C}$

$M \perp C \mid K$ → nije uvjetno nezavisno

c) Vrijedi li

$$A \perp B \Rightarrow A \perp B \mid C \ ?$$

Općenito **ne!**

→ kontraprimjer 1.3b IV

$$K \perp C \Rightarrow K \not\perp C \mid M$$

Vrijedi li obrat

$$A \perp B \mid C \Rightarrow A \perp B \ ?$$

Općenito **ne!**

→ kontraprimjer 1.3b III

$$L \perp S \mid N, \text{ a } L \not\perp S$$

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 $P(X,Y)$ i $P(X)P(Y)$

$$\left[I(X,Y) = D_{KL}(P(X,Y) \,||\, P(X)P(Y)) = \sum_{x,y} P(x,y) \ln \frac{P(x,y)}{P(x)P(y)} \right]$$

→ cilj izvoda

- veći I → više zavisne

IZVOD

Entropija

$$H(P) = -\sum_x P(x) \ln P(x)$$

Unakrsna entropija

$$H(P,Q) = -\sum_x P(x) \ln Q(x)$$

$$H(P,Q) - H(P) = -\sum_x P(x) \ln Q(x) + \sum_x P(x) \ln P(x)$$

$$= \sum_x P(x) \ln \frac{P(x)}{Q(x)}$$

$$= D_{KL}(P \,||\, Q)$$

b)

$P(1,1) = 0.2$ $\quad$ $P(2,1) = 0.05$

$P(1,2) = 0.05$ $\quad$ $P(2,2) = 0.3$

$P(1,3) = 0.3$ $\quad$ $P(2,3) = 0.1$

$$P(X=1) = \sum_y P(1,y) = 0.55$$
$$P(X=2) = \sum_y P(2,y) = 0.45$$

$$P(Y=1) = \sum_x P(x,1) = 0.25$$
$$P(Y=2) = \sum_x P(x,2) = 0.35$$
$$P(Y=3) = \sum_x P(x,3) = 0.4$$

$$I(x,y) = \sum P(x,y) \ln \frac{P(x,y)}{P(x)P(y)}$$
$$= \dots$$
$$= 0.194563$$

$\Rightarrow I(x,y)$ dosta velika $\Rightarrow$ X i Y nisu nezavisne varijable

<u>Linearna zavisnost</u>

$\mathbb{E}[X] = \mu_X = 1.45$ $\quad$ $\mathbb{E}[Y] = \mu_Y = 2.15$

$\sigma_X^2 = \frac{1}{2}\sum(x-\mu_X)^2 = 0.2525$ $\quad$ $\sigma_Y^2 = 0.689167$

$$Cov(X,Y) = \mathbb{E}[(x-\mu_X)(y-\mu_Y)]$$
$$= \mathbb{E}[XY] - \mu_X\mu_Y = 3.1 - 1.45 \cdot 2.15$$
$$= -0.0175$$

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sqrt{\sigma_X^2 \, \sigma_Y^2}} = -0.0419$$ $\Rightarrow$ nisu linearno zavisni!

c) * skip - van okvira istoda učenja

1.5.

C_1 = visok rizik
C_2 = umjeren rizik
C_3 = nizak rizik

$K = 3$
$n = 7$

- 7 diskretiziranih ulaznih varijabli

kontinuiranih:
- spol – x_1
- dob – x_2
- težina
- visina
- indeks tjelesne mase (BMI)

} x_3, x_4, x_5

- indikacija pušača – x_6
- indik. sportske aktiv. – x_7

a) Naivan Bayesov model ne bi ovdje bio dobar odabir
– velika zavisnost težine, visine i BMI

Polunaivni model

$$h(\vec{x}) = \underset{y}{\operatorname{argmax}}\, P(y)\, P(\vec{x}|y)$$

$$= \underset{y}{\operatorname{argmax}}\, P(y)\, P(x_1|y)\, P(x_2|y)\, P(x_3, x_4, x_5|y)\, P(x_6|y)\, P(x_7|y)$$

CPT

b) #param_naivan

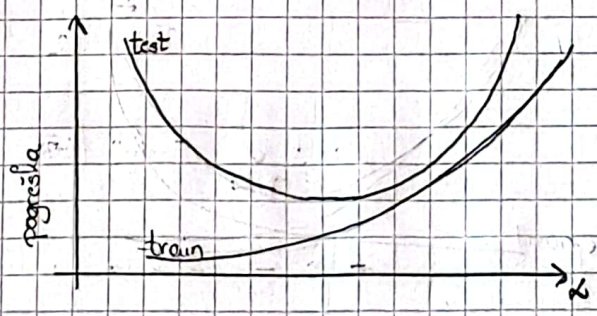
$$= \underbrace{3(2-1)}_{x_1, x_6, x_7} + \underbrace{\overbrace{4 \cdot 2 \cdot K}^{P(x_2|y)\,P(y)}}_{\substack{x_2 - x_5 \\ \hat{\mu},\ \hat{\sigma}^2}} + \underbrace{K-1}_{\text{priori}} = 3 + 24 + 2 = 29$$

#param_polunaivan

$$= \underbrace{K \cdot (2-1)}_{x_1, x_6, x_7} + \underbrace{2 \cdot 2 \cdot K}_{\substack{x_2,\ \{x_3, x_4, x_5\} \\ \hat{\mu},\ \hat{\sigma}^2}} + \underbrace{K-1}_{\text{priori}} = 3 + 12 + 2 = 17$$

c) H_α

$I(x_i, x_j) \geq \alpha \Rightarrow$ združi x_i, x_j



- veliki α $\Rightarrow$ združene značajke sa velikim I (zavisne združene)
- mali α $\Rightarrow$ sve združeno

2 Zadaci s ispita

2.1.

- binarna klas. $K=2$
  $n=100$

Log reg

$$\#param = n+1 = 101$$

Bayes

$$h_j(\vec{x}) = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma| - \frac{1}{2}(\vec{x}-\vec{\mu}_j)^T\Sigma^{-1}(\vec{x}-\vec{\mu}_j) + \ln P(y=j)$$

$$\#param = \underbrace{\frac{n}{2}(n+1)}_{\text{dij. } \Sigma} + \underbrace{K\cdot n}_{\vec{\mu}_j} + \underbrace{K-1}_{\text{priori}} = \frac{100}{2}\cdot 101 + 2\cdot 100 + 1 = 5251$$

$$\Delta param = 5251 - 101 = 5150 \quad \text{C}$$

2.2.

- naivan Bayes.
- procjene parametra: MAP - Laplace $\quad \frac{N_{ij}+1}{N_j+K_i}$

$\vec{x} = (\text{lstra}, \text{nc}, \text{kamp}, \text{bus})$

$$P(y=1|\vec{x}) = \frac{P(y=1)P(\vec{x}|y=1)}{P(\vec{x})} = \frac{P(y=1)P(\vec{x}|y=1)}{P(y=1)P(\vec{x}|y=1)+P(y=0)P(\vec{x}|y=0)}$$

$P(y=1) = \frac{3}{7}$

$P(y=0) = \frac{4}{7}$

$$P(\vec{x}|y=1) = \frac{1}{6}\cdot\frac{2}{5}\cdot\frac{2}{6}\cdot\frac{2}{6} = \frac{1}{135}$$

$P(\text{lstra}|1) = \frac{0+1}{3+3} = \frac{1}{6}$

$P(\text{nc}|1) = \frac{1+1}{3+2} = \frac{2}{5}$

$P(\text{kamp}|1) = \frac{1+1}{3+3} = \frac{2}{6}$

$P(\text{bus}|1) = \frac{1+1}{3+3} = \frac{2}{6}$

$$P(\vec{x}|y=0) = \frac{3}{7}\cdot\frac{1}{2}\cdot\frac{3}{7}\cdot\frac{2}{7} = \frac{9}{343}$$

$P(\text{lstra}|0) = \frac{2+1}{4+3} = \frac{3}{7}$

$P(\text{nc}|0) = \frac{2+1}{4+2} = \frac{3}{6}$

$P(\text{kamp}|0) = \frac{2+1}{4+3} = \frac{3}{7}$

$P(\text{bus}|0) = \frac{1+1}{4+3} = \frac{2}{7}$

$$P(y=1|\vec{x}) = \frac{\frac{3}{7}\cdot\frac{1}{135}}{\frac{3}{7}\cdot\frac{1}{135}+\frac{4}{7}\cdot\frac{9}{343}} = \frac{343}{1963} \approx 0.17473 \quad \text{A}$$

2.3.

$K = 2$

S = „stožer"
P = „pandemija"
K = „koronavirus"
G = „general"

• za koju jednakost očekujemo da ne vrijedi

A $\quad P(S|y=1) = P(S|P, y=1)$
– klasa 1 – spominje se koronavirus, ovo očekuj. da vrijedi (pandemija ne donosi dodatnu inf.)

B $\quad P(G|y=0) = P(G|S, y=0)$
– član vojnog stožera

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
→ poluraivan Bayes

$$\vec{x} = (\underbrace{\text{Italija}, \text{ne}}_{K_1 = 4\cdot 2 = 8}, \underbrace{\text{zimski}}_{K_2=2}, \underbrace{\text{dipl}}_{K_3=2}, \underbrace{\text{ne}}_{K_4=2})$$

• MAP – Laplace

$N = 8$

$$P(y=1|\vec{x}) = \frac{P(\vec{x}|y=1)P(y=1)}{P(\vec{x}|y=1)P(y=1) + P(\vec{x}|y=0)P(y=0)}$$

$P(y=1) = \frac{6}{8} \qquad P(y=0) = \frac{2}{8}$

$\underline{P(\vec{x}|y=1)} = \frac{3}{224} \qquad \underline{P(\vec{x}|y=0)} = \frac{9}{320}$

$P(\text{Italija}, \text{ne}|1) = \frac{0+1}{6+8} = \frac{1}{14} \qquad P(\text{Italija}, \text{ne}|0) = \frac{0+1}{2+8} = \frac{1}{10}$

$P(\text{zimski}|1) = \frac{3+1}{6+2} = \frac{4}{8} \qquad P(\text{zimski}|0) = \frac{2+1}{2+2} = \frac{3}{4}$

$P(\text{dipl.}|1) = \frac{5+1}{6+2} = \frac{6}{8} \qquad P(\text{dipl}|0) = \frac{1+1}{2+2} = \frac{2}{4}$

$P(\text{ne}|1) = \frac{3+1}{6+2} = \frac{4}{8} \qquad P(\text{ne}|0) = \frac{2+1}{2+2} = \frac{3}{4}$

$$P(y=1|\vec{x}) = \frac{\frac{6}{8}\cdot\frac{3}{224}}{\frac{6}{8}\cdot\frac{3}{224} + \frac{2}{8}\cdot\frac{9}{320}} = \frac{10}{17} = 0.5882 \quad \text{C}$$

2.5.

$K = 2$

x_1 – pol. orj. $(K_1 = 3)$
x_2 = dob kandid
x_3 = polit. staž } numeričke značajke
x_4 = populist (binarna zroč.)
x_5 = velika stranka (bin. zroč.)

$\mathcal{H}_0$ = Bayes. bez uvj. nezavis.

$\mathcal{H}_1$ = polunaiv. Bayes (združene x_1 i x_4)

$\mathcal{H}_2$ = naivan Bayes.

• x_2 i x_3 = dijele Σ

$\mathcal{H}_0$

$\#param = 3 + 4 + 22 + 1 = 30$

x_2, x_3

Σ : $\frac{n(n+1)}{2} = \frac{2 \cdot 3}{2} = 3$

μ $K \cdot n = 2 \cdot 2 = 4$

x_1, x_4, x_5

$y = 0 \rightarrow K_1 \cdot K_4 \cdot K_5 - 1 = 3 \cdot 2 \cdot 2 - 1 = 11$

$y = 1 \rightarrow K_1 \cdot K_4 \cdot K_5 - 1 = 11$

prior : $K - 1 = 1$

$\mathcal{H}_1$

$\#param = 2 + 4 + 10 + 2 + 1 = 19$

x_2, x_3

! Σ $\rightarrow$ dijagonalna (x_2 i x_3 nezavisne) (dg. varijanci.)
2

μ $K \cdot n = 4$

x_1, x_4

$K \cdot (K_1 \cdot K_4 - 1)$
$= 2 \cdot (3 \cdot 2 - 1)$
$= 10$

x_5 $K \cdot (K_5 - 1) = 2$

prior: $K - 1 = 1$

Ukupno : $30 + 19 + 15$
$= 64$

D //

$\mathcal{H}_2$

x_2, x_3 Σ : 2 (opet dijag. zbog nez.)
μ : $K \cdot n = 4$

x_1, x_4, x_5

$K_y (K_1 - 1 + K_4 - 1 + K_5 - 1)$
$= 2 \cdot (2 + 1 + 1) = 2 \cdot 4 = 8$

prior $K - 1 = 1$

$\#param = 2 + 4 + 8 + 1 = 15$.

2.6

$n = 3$

- x_1, x_2, x_3 = Binarne varijable

$\alpha = 0.01$

$I(x, y) \geq \alpha \Rightarrow$ združi par varijabli

$I(x_1, x_2) = ?$

$I(x_1, x_3) = ?$

$I(x_2, x_3) = ?$

$$I(x_1, x_2) = \sum P(x_1, x_2) \ln \frac{P(x_1, x_2)}{P(x_1)P(x_2)} = 5.13 \cdot 10^{-3} < \alpha$$

$x_1 \backslash x_2$	0	1	
0	0.3	0.1	0.4
1	0.5	0.1	0.6
	0.8	0.2	

$$I(x_1, x_3) = \sum P(x_1, x_3) \ln \frac{P(x_1, x_3)}{P(x_1)P(x_3)} = 0.0322 > \alpha$$

$x_1 \backslash x_3$	0	1	
0	0.3	0.1	0.4
1	0.3	0.3	0.6
	0.6	0.4	

$\boxed{x_1, x_3}$

$$I(x_2, x_3) = \sum P(x_2, x_3) \ln \frac{P(x_2, x_3)}{P(x_2)P(x_3)} = 5.132 \cdot 10^{-3} < \alpha$$

$x_2 \backslash x_3$	0	1	
0	0.5	0.3	0.8
1	0.1	0.1	0.2
	0.6	0.4	

$$P(y|\vec{x}) = P(y)P(x_1, x_2, x_3|y)$$

$$= P(y)P(x_1, x_3|y)P(x_2|y) \Rightarrow \text{C}$$

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

$$I(x_1, x_2) = \sum P(x_1, x_2) \ln \frac{P(x_1, x_2)}{P(x_1)P(x_2)}$$

<u>Laplaceov progenitelj</u>

$$P(x_1 = 0, x_2 = 0) = \frac{1+1}{10+4} = \frac{2}{14}$$

$$P(x_1 = 0, x_2 = 1) = \frac{4+1}{10+4} = \frac{5}{14}$$

$$P(x_1 = 1, x_2 = 0) = \frac{2+1}{10+4} = \frac{3}{14}$$

$$P(x_1 = 1, x_2 = 1) = \frac{3+1}{10+4} = \frac{4}{14}$$

$x_1 \backslash x_2$	0	1	
0	$\frac{2}{14}$	$\frac{5}{14}$	$\frac{7}{14}$
1	$\frac{3}{14}$	$\frac{4}{14}$	$\frac{7}{14}$
	$\frac{5}{14}$	$\frac{9}{14}$	

$I(x_1, x_2) = 0.11168$

$\hookrightarrow$ 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) Pretpostavite 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 pretpostavljajući $y \perp w | x$ i $x \perp z | y, w$. Kojoj vrsti induktivne pristranosti odgovaraju ove pretpostavke o nezavisnosti? Obrazložite motivaciju za uvođenjem dodatnih pretpostavki 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 vrijednsti 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 $x_1, x_2, x_3, y \in \{\top, \bot\}$ 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: Izvježbati iščitavanje uvjetnih nezavisnosti iz Bayesove mreže te određivanje (ne)zavinosti 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 $s \in \{\textit{muški}, \textit{ženski}\}$, $p \in \{\bot, \top\}$, $t \in \{1, 3, 5\}$ i $r \in \{\textit{nizak}, \textit{umjeren}, \textit{visok}\}$. Zajednička razdioba faktorizirana je kao:

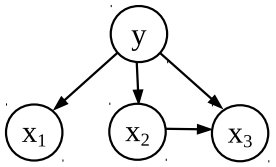
$$P(S, T, P, R) = P(S)P(T)P(P|S, T)P(R|P)$$

(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 Zadatci 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 $w \bot y$, $z \bot \{x, y\}|w$ B $x \bot y|z$, $z \bot w|y$ C $w \bot y$, $z \bot x$, $x \bot w|\{z, y\}$ D $y \bot w$, $y \bot x|\{w, z\}$

3. (P) Bayesova mreža ima pet varijabli, od kojih su v, w i z binarne, a x i y ternarne varijable. Topološki uređaj varijabli neka je v, w, x, y, z. Uz takav uređaj, u mreži vrijede sljedeće marginalne i uvjetne nezavisnosti:

$$v \bot w \qquad w \bot x|v \qquad v \bot y|\{w, x\} \qquad \{v, w\} \bot z|\{x, y\}$$

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) utječu 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:

$$\{v, w\} \perp y|x \qquad \{v, x\} \perp z|\{w, y\}$$

   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 Zadaci za učenje

1.1.

a) 3 osnovna aspekta svakog probabilističkog grafičkog modela su
- reprezentacija
- zaključivanje
- učenje

b) PGM-ovi uključuju uglavnom generativne modele, ali i neke diskriminativne modele.

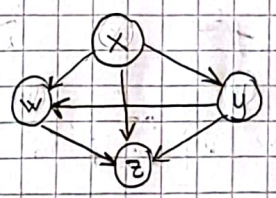
↳ npr. Bayesove mreže učimo generativno, a Markovljeve mreže diskriminativno

Također PGM-ovi mogu biti parametarski (npr. Bayesove mreže) i neparametarski (npr. Markov. mreže)

c) Topološki uređaj: x, y, w, z

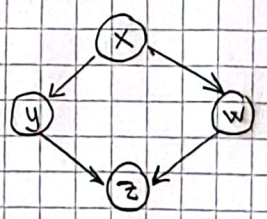
$$p(x,y,w,z) = p(x)p(y|x)p(w|x,y)p(z|x,y,w)$$

$\Rightarrow$ pravilo lanca



d) Pretpostavke: $\left.\begin{array}{l} y \perp w \mid x \\ x \perp z \perp y, w \end{array}\right\}$ pristranost jezika

$$p(x,y,w,z) = p(x)p(y|x)p(w|x,y)p(z|x,y,w)$$
$$= p(x)p(y|x)p(w|x)p(z|y,w)$$



Motivacija za uvođenjem dodatnih pretpostavki (uvjetne nezavisnosti) zbog smanjivanja složenosti modela

e) **TOPOLOŠKI UREĐAJ**

= potpuni (linearni) uređaj kod kojeg čvorovi roditelji dolaze prije čvorova djece

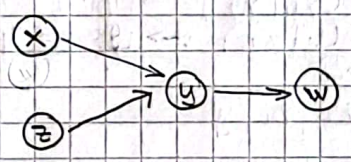
UREĐAJNO MARKOVLJEVO SVOJSTVO

$$x_k \perp pred(x_k) \setminus pa(x_k) \mid pa(x_k)$$

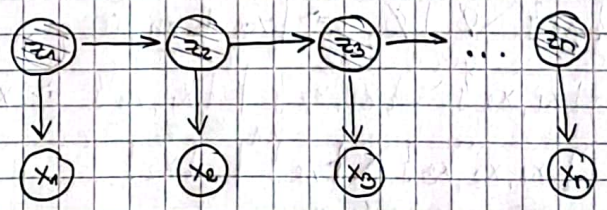
→ $pa(x_k)$ = čvorovi roditelja čvora x_k
→ $pred(x_k)$ = čvorovi prethodnici čvora x_k

$$\begin{aligned} P(x,y,w,z) &= P(x)P(y|x,z)P(z)P(w|y) \\ &= P(x)P(z|x)P(y|x,z)P(w|y) \quad \Leftarrow \quad x \perp z \\ &= P(x)P(z|x)P(y|x,z)P(w|x,y,z) \end{aligned}$$

$w \perp x \mid y$
$w \perp z \mid y$
$x \perp z$



f) **Sakriveni Markovljev model (HMM)**



$$p(\vec{x},\vec{z}) = p(z_1)p(x_1|z_1)\prod_{k=2}^{n} p(z_k|z_{k-1})p(x_k|z_k)$$

- Latentne varijable $\vec{z}$ = predstavljaju stanja kroz koja prolazi proces generiranja podataka

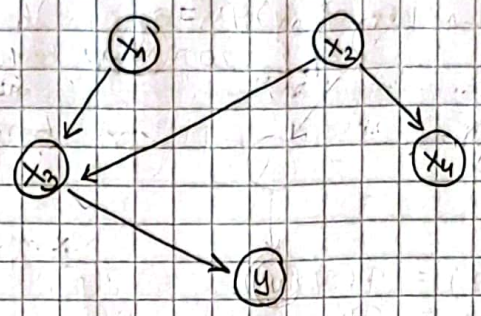
- uvjetne nezavisnosti kodirane mrežom ??

1.2.

$x_1, x_2, x_3, y \in \{T, \bot\}$
$x_4 \in \{2, 3, 4, 5\}$

a)

$$P(x_1, x_2, x_3, x_4, y) = P(x_1)P(x_2)P(x_3 \mid x_1, x_2)P(x_4 \mid x_2)P(y \mid x_3)$$



b) Broj parametra $= 1 + 1 + 4 + 6 + 2 = 14$

$P(x_1) \to 1$
$P(x_2) \to 1$
$P(x_3 \mid x_1, x_2) \to 4 \cdot 1 = 4$
$P(x_4 \mid x_2) \to 3 \cdot 2 = 6$
$P(y \mid x_3) \to 2 \cdot 1 = 2$

c) Uvjetne nezavisnosti

TU: x_1, x_2, x_3, x_4, y

$x_e \perp pred(x_e) \setminus pa(x_e) \mid pa(x_e)$

x_1: $x_1 \perp \emptyset \mid \emptyset$

x_2: $x_2 \perp \{x_1\} \setminus \emptyset \mid \emptyset \Rightarrow \boxed{x_2 \perp x_1}$

x_3: $x_3 \perp \{x_1, x_2\} \setminus \{x_1, x_2\} \mid \{x_1, x_2\} \Rightarrow x_3 \perp \emptyset$

x_4: $x_4 \perp \{x_1, x_2, x_3\} \setminus \{x_2\} \mid \{x_2\}$

$$\boxed{\begin{array}{l} x_4 \perp x_1 \mid x_2 \\ x_4 \perp x_3 \mid x_2 \end{array}}$$

y: $y \perp \{x_1, x_2, x_3, x_4\} \setminus \{x_3\} \mid \{x_3\}$

$$\boxed{\begin{array}{l} y \perp x_1 \mid x_3 \\ y \perp x_2 \mid x_3 \\ y \perp x_4 \mid x_3 \end{array}}$$

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 koja statistička svojstva proizlaze iz određene kauzalne strukture.

Uvjetna nezavisnost nije prikazana Bayesovom mrežom. Prikazana je kauzalna struktura.

b) <u>d-odvajanje</u>

= analiza staza u grafu između čvorova x i y i zaključivanje o uvjetnoj nezavisnosti varijabla x i y na temelju te analize

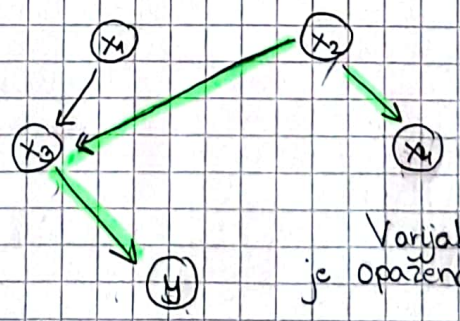
- ako su <u>sve staze</u> između čvorova x i y <u>odvojene</u>, onda su <u>čvorovi uvjetno nezavisni</u>

E = skup opaženih varijabli

Za stazu P od čvora x do čvora y kažemo da je d-odvojena $\Leftrightarrow$ vrijedi barem jedno od sljedećeg

1. P sadrži lanac $x \to z \to y$ ili $x \leftarrow z \leftarrow y$ i $z \in E$
2. P sadrži račvanje $x \leftarrow z \to y$ i $z \in E$
3. P sadrži sraz $x \to z \leftarrow y$ i $z \notin E$, ni jedan sljedbenik od $z \notin E$

c) Uvjetna nezavisnost y i x_4 $\Leftrightarrow$ sve staze d-odvojene



Jedina staza od y do x_4

$$\underbrace{x_4 \leftarrow \underbrace{x_2 \to x_3}_{\text{račvanje}} \to y}_{\text{lanac}}$$

Varijable y i x_4 su uvjetno nezavisne, ako je opažena varijabla x_2 ili x_3

d) <u>Efekt objašnjavanja</u>

$$P(x_1 \mid x_3) \neq P(x_1 \mid x_2, x_3) \iff x_1 \not\perp x_2 \mid x_3$$

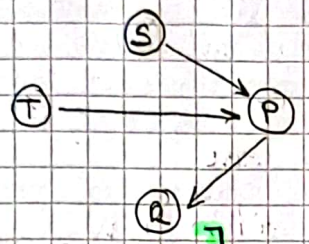
- pojavljuje se kod sraza $\to$ varijable x_1 i x_2 „natječu" se za objašnjavanje varijable x_3 $\Rightarrow$ Ako su opažene varijable x_3 i x_2 smanjena je vjerojatnost da je opažena x_1 u odnosu kada je opažena samo x_3

1.4.

$S \in \{\text{muški}, \text{ženski}\}$
$T \in \{1, 3, 5\}$
$P \in \{T, \bot\}$
$R \in \{\text{nizak}, \text{umjeren}, \text{visok}\}$

$$P(S,T,P,R) = P(S)P(T)P(P|S,T)P(R|P)$$

a)



$$\begin{bmatrix} T \perp S \\ R \perp S \mid P \\ R \perp T \mid P \end{bmatrix}$$

Topološki uređaj: $S\ T\ P\ R$

$x_e \perp \text{pred}(x_e) \setminus \text{pa}(x_e) \mid \text{pa}(x_e)$

$\underline{S:}$ $S \perp \emptyset \mid \emptyset$

$\underline{T:}$ $T \perp \{S\} \setminus \emptyset \mid \emptyset$
$\rightarrow T \perp S$

$\underline{P:}$ $P \perp \{S, T\} \setminus \{S, T\} \mid \{S, T\}$
$P \perp \emptyset \mid \emptyset$

$\underline{R}$ $R \perp \{S, T, P\} \setminus \{P\} \mid \{P\}$
$\rightarrow R \perp S \mid P$
$R \perp T \mid P$

b) Varijable x_1 i x_2 uvjetno su nezavisne za dani E = skup opaženih varijabli $\Leftrightarrow$ su sve staze P d-odvojene za dani E

Staza P od čvora x_1 do čvora x_2 je d-odvojena $\Leftrightarrow$ vrijedi barem jedno

1. P sadrži lanac $x_2 \rightarrow z \rightarrow x_1$ ili $x_1 \rightarrow z \rightarrow x_2$ 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 sljedbenik od z nije u E

2 Zadaci s ispita

T.1. Za Bayesovu mrežu kažemo da je generativni i parametarski model. Zašto?

(D)

generativni = opisuje postupak kojim se mogu generirati podaci koji se pokoravaju određenoj zajed. vjerj. distribuciji

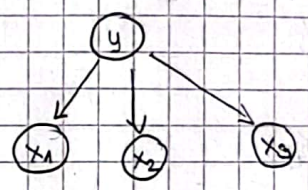
parametarski = svaki čvor Bayes. mreže definira uvjetnu vjerojatnost preko teorijske distribucije koja je definirana svojim parametrima

T.2. Koja je veza između uvje. nezavisnosti varijabli u Bayesovoj mreži i opasnosti od preraučenosti?

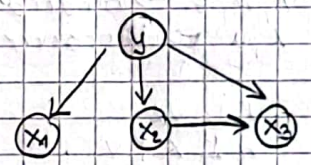
(A) uvođenje pretpostavki o uvj nezavisnosti pojednostavljuje strukturu Bayesove mreže i smanjuje broj parametara, čime se smanjuje i mogućn. preraučenosti

2.1. x_1, x_2, x_3 – binarne
y – binarna

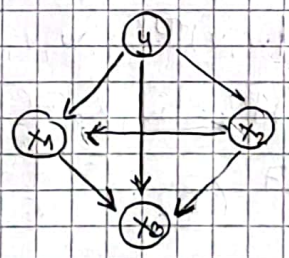
H_1



H_2



H_3



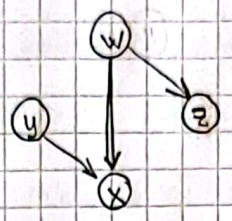
$P_1(x_1, x_2, x_3, y) = P(y)P(x_1|y)P(x_2|y)P(x_3|y)$
$P_2(x_1, x_2, x_3, y) = P(y)P(x_1|y)P(x_2|y)P(x_3|y, x_2)$
$P_3(x_1, x_2, x_3, y)$
$= P(y)P(x_2|y)P(x_1|y, x_2)P(x_3|y, x_1, x_2)$

$\#p1 = 1 + 2 + 2 + 2 = 7$
$\#p2 = 1 + 2 + 2 + 4 = 9$
$\#p3 = 1 + 2 + 4 + 8 = 15$

(B) H_2 ima 2 više parametra od H_1, a 6 manje od H_3

2.2

$P(w,x,y,z) = P(w)P(y)P(x|w,y)P(z|w)$



$w, y, x, z \rightarrow$ čeksiđei prvi
w, y, z, x
y, w, x, z
y, w, z, x

TU: w, y, x, z

UMS: $x_i \perp pred\{x_i\} \setminus pa\{x_i\} \mid pa\{x_i\}$

w: $w \perp \emptyset \mid \emptyset$

y: $y \perp \{w\} \setminus \emptyset \mid \emptyset$

$\boxed{y \perp w}$

x: $x \perp \{w,y\} \setminus \{w,y\} \mid \{w,y\}$
$x \perp \emptyset \mid \{w,y\}$

z: $z \perp \{w,y,x\} \setminus \{w\} \mid \{w\}$

$\boxed{z \perp y \mid w}$
$\boxed{z \perp x \mid w}$

(A) $w \perp y$, $z \perp \{x,y\} \mid w$

2.3

$v, w, z \rightarrow$ binarno
$x, y \rightarrow$ ternarno

TU: v, w, x, y, z

$v \perp w$
$w \perp x \mid v$
$v \perp y \mid \{w,x\}$
$z \perp \{v,w\} \mid \{x,y\}$

$P(v,w,x,y,z) = P(v)P(w|v)P(x|w,v)P(y|w,v,x)P(z|w,v,x,y)$
$= P(v)P(w)P(x|v)P(y|w,x)P(z|x,y)$

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_i
$P(x_i|x_j)$
$(k_i - 1) \cdot k_j$

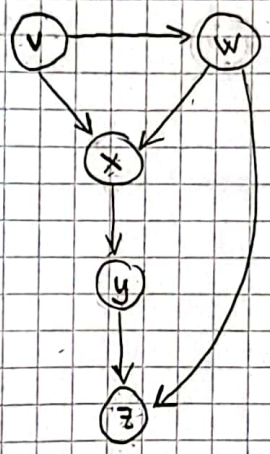
2.5

• v, w, x, y, z

$\{v, w\} \perp y \mid x$

$\{v, x\} \perp z \mid \{w, y\}$

$$P(v, w, x, y, z) = P(v)P(w|v)P(x|v,w)P(y|v,w,x)P(z|v,w,x,y)$$
$$= P(v)P(w|v)P(x|v,w)P(y|x)P(z|y,w)$$



Koje nezavisnosti vrijedi?

A $x \perp z \mid w$

Lanac: $x \to y \to z$ $y \in E$ (nije opažen) :

Račvanje: $x \leftarrow w \to z$ $w \in E$ ✓

B $v \perp y \mid w$

Lanac: $v \to x \to y$ → nije odvojeno

Lanac: $v \to w \to x \to y$ ✓

C $x \perp z \mid y$

Lanac: $x \to y \to z$ ✓

Račvanje: $x \leftarrow w \to z$ ✗ (nije odvojeno)

D $v \perp y \mid x$

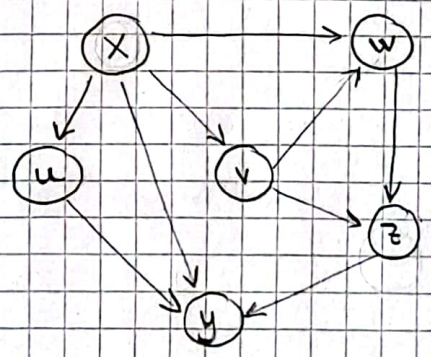
Lanac: $v \to x \to y$ ✓

Lanac: $v \to w \to x \to y$ ✓

ODVOJENO

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
   $x \to w \to z$
   • lanac da bi bila odvojena
   $w \in E$

2. Staza (lanac)
   $x \to v \to z$
   $\Rightarrow v \in E$

3. Staza (sraz)
   $x \to y \leftarrow z$
   $y \notin E$ i svaki sljedbenik od y u topološkom uređaju $\notin E$

Topološki uređaj

x, u, v, w, z, y

Da bi x bilo uvjetno nezavisno z mora vrijediti

$w, v \in E$

$y \notin E$

→ ne ovisi o u (A)

2.4.

• 5 binarnih varijabli

$J \xrightarrow{\text{povećava}} G$

$V \xrightarrow{\text{povećava}} N$

$N \xrightarrow{\text{povećava}} G$

$V \xrightarrow{\text{povećava}} T$

$S \xrightarrow{\text{povećava}} T$

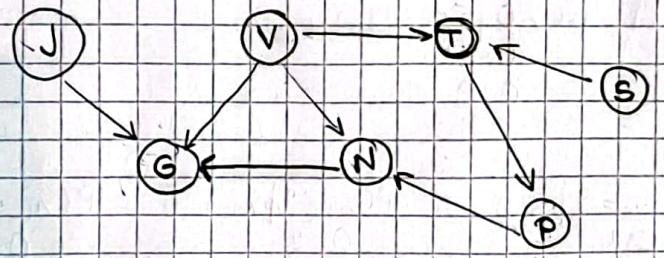
$T \xrightarrow{\text{povećava}} P \xrightarrow{\text{povećava}} N$

A $P(V=1 \mid P=1, T=1) < P(V=1 \mid T=1)$
=> lanac – nije efekt objašnjavanja $P \not\perp V \mid T$

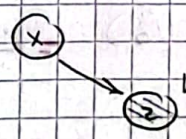
B $P(G=1 \mid J=1, V=1) > P(G=1 \mid J=1)$ -> nema središnjeg čvora
$G \not\perp V \mid J$

C $P(V=1 \mid P=1, N=1) < P(V=1 \mid N=1)$
$V \not\perp P \mid N$

D $P(T=1 \mid V=1, P=1) > P(T=1 \mid V=1)$
$T \not\perp P \mid V$
-> P nije srednji čvor



• efekt objašnjavanja -> u srazu!



$P(x \mid z) \neq P(x \mid y, z)$
$P(x=1 \mid z=1) < P(x=1 \mid y=0, z=1)$
$P(x=1 \mid z=1) > P(x=1 \mid y=1, z=1)$

• ako je z opažen, onda su x i y nezavisni (nepreciznost, točnije staza odvojena)

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 modele 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	$P(x_3 = \top \mid x_1, x_2)$
$\bot$	$\bot$	0.3
$\bot$	$\top$	0.5
$\top$	$\bot$	0.8
$\top$	$\top$	0.9

x_3	$P(y = \top \mid x_3)$
$\bot$	0.2
$\top$	0.9

x_2	$P(x_4 = 2 \mid x_2)$	$P(x_4 = 3 \mid x_2)$	$P(x_4 = 4 \mid x_2)$
$\bot$	0.4	0.2	0.3
$\top$	0.2	0.1	0.1

   (a) Postupkom egzaktnog zaključivanja izračunajte $P(y = \top \mid x_1 = \top, x_4 = 3)$.

   (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 podatci 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-procjenitelja procijenite $P(P \mid S, T)$.

S	P	T	R
ženski	$\top$	1	*visok*
ženski	$\top$	5	*umjeren*
muški	$\bot$	3	*nizak*
ženski	$\bot$	1	*umjeren*
muški	$\top$	5	*nizak*
ženski	$\bot$	1	*nizak*

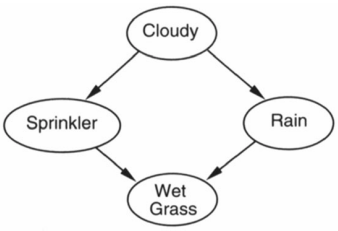
(b) Korištenjem egzaktnog zaključivanja izvedite izraz za vjerojatnost visokog rizika oboljenja osobe koja je pušač i posjećuje teretanu pet puta tjedno. Za svaku od četiri varijable naznačite radi li se o varijabli upita, opaženoj varijabli ili varijabli smetnje.

(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.



C	$P(C)$
0	0.5
1	0.5

S	C	$P(S\|C)$
0	0	0.5
0	1	0.9
1	0	0.5
1	1	0.1

R	C	$P(R\|C)$
0	0	0.8
0	1	0.2
1	0	0.2
1	1	0.8

W	R	S	$P(W\|R,S)$
0	0	0	1.0
0	0	1	0.9
0	1	0	0.1
0	1	1	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

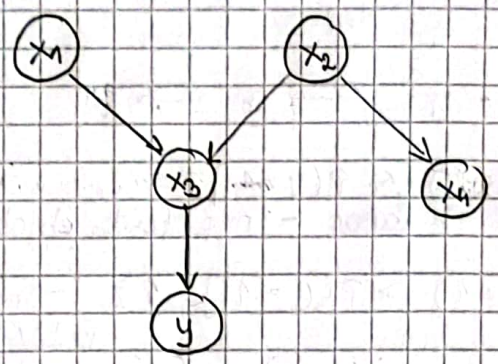
Postupkom uzorkovanja s odbijanjem želimo procijeniti parametar μ uvjetne distribucije $P(x = 0|y = 1, z = 0)$. Uzorkovanje smo ponovili ukupno $N = 100$ puta, od čega smo neke vektore morali odbaciti, pa je naš uzorak manji od N. Na temelju dobivenog uzorka parametar μ procjenjujemo MAP procjeniteljem 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 Zadaci za učenje

1.1. $P(x_1, x_2, x_3, x_4, y) = P(x_1)P(x_2)P(x_3|x_1,x_2)P(x_4|x_2)P(y|x_3)$



$x_1, x_2, x_3, y \in \{1, 0\}$

$x_4 \in \{2, 3, 4, 5\}$

$P(x_1 = 1) = 0.2$ $\quad$ $P(x_2 = 1) = 0.6$

$P(x_1 = 0) = 0.8$ $\quad$ $P(x_2 = 0) = 0.4$

Za x_3

x_1	x_2	$P(x_3 = 1 \mid x_1, x_2)$
0	0	0.3
0	1	0.5
1	0	0.8
1	1	0.9

Za y

x_3	$P(y=1 \mid x_3)$
0	0.2
1	0.9

Za x_4

x_2	$P(x_4 = 2 \mid x_2)$	$P(x_4 = 3 \mid x_2)$	$P(x_4 = 4 \mid x_2)$
0	0.4	0.2	0.3
1	0.2	0.1	0.1

a) Egzaktno zaključivanje: $\left[P(\vec{x}_q \mid \vec{x}_o) = \frac{\sum_{\vec{x}_n} P(\vec{x}_q, \vec{x}_o, \vec{x}_n)}{\sum_{\vec{x}_n} \sum_{\vec{x}_q} P(\vec{x}_q, \vec{x}_o, \vec{x}_n)} \right]$

$$P(y=1 \mid x_1 = 1, x_4 = 3) = \frac{P(y=1, x_1 = 1, x_4 = 3)}{P(x_1 = 1, x_4 = 3)}$$

$\vec{x}_q = y$

$\vec{x}_o = \{x_1, x_4\}$

$\vec{x}_n = \{x_2, x_3\}$

$$P(x_1 = 1, x_4 = 3) = \sum_{x_3} \sum_{x_2} P(x_1, x_2, x_3, x_4, y)$$

$$= \sum_{x_2} \sum_{x_3} \sum_{y} P(x_1 = 1) P(x_2) P(x_3 \mid x_1 = 1, x_2) P(x_4 = 3 \mid x_2) \cdot P(y \mid x_3)$$

↳ 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	$0.2 \cdot 0.6 \cdot 0.9 \cdot 0.1 \cdot 0.9 = 9.72 \cdot 10^{-3}$

$\uparrow$ 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	$0.2 \cdot 0.4 \cdot 0.2 \cdot 0.2 \cdot 0.2 = 6.4 \cdot 10^{-4}$	0.02212
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	$0.2 \cdot 0.6 \cdot 0.9 \cdot 0.1 \cdot 0.9 = 9.72 \cdot 10^{-3}$	

$$P(y=1, x_1=1, x_4=3) = 0.02212$$

$$\left[P(y=1 \mid x_1=1, x_4=3) = \frac{0.02212}{0.028} = 0.79 \right]$$

b) <u>Aposteriorni upiti</u> odgovaraju na pitanje kakva je aposteriorna distribucija skrivenih varijabli <u>MAP-upiti</u> odgovaraju na pitanje koje su najvjerojatnije vrijednosti za sve varijable upita, uz dane opažene varijable

U zadatku pod a) radi se o aposteriornom upitu jer se traži vrijednost distribucije za određenu vrijednost varijable y $P(y=1 \mid x_1=1, x_4=3)$

c) Utječe li broj varijabli u mreži na učinkovitost zaključivanja?
Da. Postoji problem <u>kombinatorne eksplozije</u> u postupku. Naime čim je veći broj varijabli u mreži u izrazima za <u>zajedničku gustoću</u> (raste s povećanjem broja $\vec{x}_n$) i <u>vjerojatnost dokaza</u> (raste s povećanjem broja $\vec{x}_n$ i $\vec{x}_q$) treba **marginalizirati** po većem broju varijabli

d) Prednost približnog zaključivanja uzrokovanjem
- ne dovodi do kombinatorne eksplozije, a nad uzrokovanom skupu se mogu relativno dobro procijeniti parametri tražene distribucije.

Uzrokovanje unaprijednim uzrokovanjem

$$P(x_1, x_2, x_3, x_4, y) = P(x_1)P(x_2)P(x_3|x_1, x_2)P(x_4|x_2)P(y|x_3)$$

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 koje opisuju podatke

Bayes. mreža $p(\vec{x}) = \prod_{k=1}^{n} p(x_k | pa(x_k))$

svaka od ovih n distribucija ima pripadne parametre $\vec{\theta}_k$

$$\vec{\theta} = [\vec{\theta}_1 \ \vec{\theta}_2 \ \dots \ \vec{\theta}_k]$$

Parametre učimo iz podataka procjenama (MLE, MAP ili bayesovska procjena)

b) Log-izglednost Bayesove mreže

$$\ln \mathcal{L}(\vec{\theta}|D) = \ln p(D|\vec{\theta}) \overset{i.i.d}{=} \ln \prod_{i=1}^{N} p(\vec{x}^i|\vec{\theta}) = [\text{def. Bay. mreže}]$$

$$= \ln \prod_{i=1}^{N} \prod_{k=1}^{n} p(x_k^i | pa(x_k^i; \vec{\theta}))$$

$$= \ln \prod_{k=1}^{n} \prod_{i=1}^{N} p(x_k^i | pa(x_k^i; \vec{\theta}))$$

$$= \sum_{k=1}^{n} \sum_{i=1}^{N} \ln p(x_k^i | pa(x_k^i; \vec{\theta}))$$

Budući da log-izglednost mreže dekomponira po čvorovima, parametri svakog čvora mreže mogu se procjenjivati zasebno

c) Ako neki model ima skrivene (latentne) varijable, opaženi skup podataka nije potpun, odnosno on je nepotpun. Nad skupom nepotpunih podataka ne možemo procjenjivati parametre svakog čvora zasebno jer log-izgledrost mreže ne dekomponira po čvorovima.

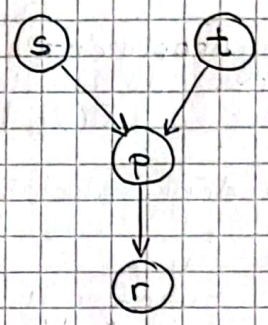
Posljedica toga je da ni za MLE, ni za MAP ni za bayesovski procjenitelj ne postoji rješenje u zatvorenoj formi $\rightarrow$ EM-algoritam

1.3.

$s \in \{M, Ž\}$
$p \in \{0, 1\}$
$t \in \{1, 3, 5\}$
$r \in \{n, u, v\}$

$$P(s,t,p,r) = P(s)P(t)P(p|s,t)P(r|p)$$

s	p	t	r
Ž	1	1	v
Ž	1	5	u
M	0	3	n
Ž	0	1	u
M	1	5	n
Ž	0	1	n



a) MAP (Laplace) procjenitelj za $P(p|s,t)$ $\quad k = 2$

$P(p=0|s=M,t=1) = 1/2$
$P(p=0|s=M,t=3) = 2/3$
$P(p=0|s=M,t=5) = 1/3$

$P(p=0|s=Ž,t=1) = 3/5$
$P(p=0|s=Ž,t=3) = 1/2$
$P(p=0|s=Ž,t=5) = 1/3$

$P(p=1|s=M,t=1) = 1/2$
$P(p=1|s=M,t=3) = 1/3$
$P(p=1|s=M,t=5) = 2/3$

$P(p=1|s=Ž,t=1) = 2/5$
$P(p=1|s=Ž,t=3) = 1/2$
$P(p=1|s=Ž,t=5) = 2/3$

MAP: $\dfrac{N_{ej}+1}{N_{ej}+2}$ $\rightarrow k_e$

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 smetnje: $\vec{x_n} = [s]$
opažene varijable: $\vec{x_o} = [p,t]$

c) **Prednosti** uzorkovanja nad egzaktnim zaključivanjem
→ rješava se problem kombinatorne eksplozije

U ovoj mreži treba procijeniti $1+1+2\cdot 4+2\cdot 3 =$
$= 8+2+6 =$
$= 16$ distribucija za egzaktno zaključivanje

$|s| = 2$
$|p| = 2$
$|t| = 3$
$|r| = 3$

Nedostatak: zaključivanje je približno, a u nekim tehnikama za male vjerojatnoće $\vec{x}_0$ (npr. ako je $\vec{x}_0 = [s, t]$ i vrijednosti koje su opažene $s =$ ženski, a $t = 3$) treba dugo vremena da se generira dovoljno velike uzorke za procjenu

d) Nedostatak unaprijednog uzorkovanja je taj što generira vrijednosti iz zajedničke distribucije $P(\vec{X})$ (u ovom slučaju $P(s,p,t,r)$), a ne iz distribucije $P(\vec{x}_g \mid \vec{x}_0)$ (u ovom slučaju $P(p \mid s,t)$)

- vrijedi i drugi nedostatak pod c)

Alternative su
- uzorkovanje s odbacivanjem
- uzorkovanje po važnosti
- Gibbsovo uzorkovanje

2 Zadaci s ispita

2.1. D

$$P(r=1 \mid w=1, c=0) = \frac{P(r=1, w=1, c=0)}{P(w=1, c=0)} = \frac{0.0945}{0.1145} \approx 0.825$$

$$P(r=1, w=1, c=0) = \sum_s P(c=0)P(s \mid c=0)P(r=1 \mid c=0)P(w=1 \mid s, r=1)$$
$$= \sum_s 0.5 \cdot P(s \mid c=0) \cdot 0.2 \cdot P(w=1 \mid s, r=1)$$
$$= 0.1 \cdot \left(P(s=0 \mid c=0)P(w=1 \mid s=0, r=1) + P(s=1 \mid c=0)P(w=1 \mid s=1, r=1) \right)$$
$$= 0.1 \cdot (0.5 \cdot 0.9 + 0.5 \cdot 0.99)$$
$$= 0.0945$$

ova suma se može okrenut za $r=1$ zasebno pa raširit i $r=0$

$$P(w=1, c=0) = \sum_s \sum_r P(c=0)P(s \mid c=0)P(r \mid c=0)P(w=1 \mid s, r) = 0.1145$$

s	r	$P(c=0)$	$\cdot$	$P(s \mid c=0)$	$\cdot$	$P(r \mid c=0)$	$\cdot$	$P(w=1 \mid s, r)$	
0	0	0.5	$\cdot$	0.5	$\cdot$	0.8	$\cdot$	0	$= 0$
0	1	0.5	$\cdot$	0.5	$\cdot$	0.2	$\cdot$	0.9	$= 0.045$
1	0	0.5	$\cdot$	0.5	$\cdot$	0.8	$\cdot$	0.1	$= 0.02$
1	1	0.5	$\cdot$	0.5	$\cdot$	0.2	$\cdot$	0.99	$= 0.0495$

2.2.

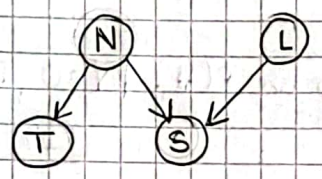
$L, S, T \in \{0, 1\}$
$N \in \{0, 1, 2\}$

$L \to S$
$N \to S, T$

$N = 7$

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

• MAP procjenitelj
$(\alpha = \beta = 2)$



$$P(L=1, S=1, N=1) = \sum_T P(L=1)P(N=1)P(S=1 \mid L=1, N=1)P(T \mid N=1)$$

$$P(L=1) = \frac{4+1}{7+2} = \frac{5}{9} \qquad P(N=1) = \frac{1+1}{7+3} = \frac{2}{10} = \frac{1}{5}$$

$$P(S=1 \mid L=1, N=1) = \frac{1+1}{1+2} = \frac{2}{3}$$

$$P(L=1, S=1, N=1) = \sum_T \frac{5}{9} \cdot \frac{1}{5} \cdot \frac{2}{3} \cdot P(T \mid N=1)$$

$$= \frac{2}{27}\left(P(T=0 \mid N=1) + P(T=1 \mid N=1)\right)$$

ovo svakako 1

$$P(T=0 \mid N=1) = \frac{1+1}{1+2} = \frac{2}{3}$$

$$= \frac{2}{27} \approx 0.0741 \quad // \quad \text{B}$$

$$P(T=1 \mid N=1) = \frac{0+1}{1+2} = \frac{1}{3}$$

2.3.

$$P(x, y, z) = P(x)P(y)P(z \mid x, y)$$

• sve varijable binarne

$P(x=1) = 0.2$
$P(y=1) = 0.3$

z	x	y	$p(z \mid x, y)$
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

• uzorkovanje s odbijanjem $(N = 1000)$
$P(y \mid x=1, z=0)$

$P(x=1) = 0.2$

$$P(z=0 \mid x=1) = \frac{P(x=1, z=0)}{P(x=1)}$$

$$P(x=1, z=0) = \sum_y P(x=1)P(y)P(z=0 \mid x=1, y)$$
$$= 0.2 \cdot 0.7 \cdot 0.5 + 0.2 \cdot 0.3 \cdot 0.9$$
$$= 0.124$$

Veličina uzorka
$N \cdot P(x=1, z=0)$
$= 124$ //

B

$$P(x=1) = \sum_y \sum_z P(x=1)P(y)P(z \mid x=1, y) = 0.2$$

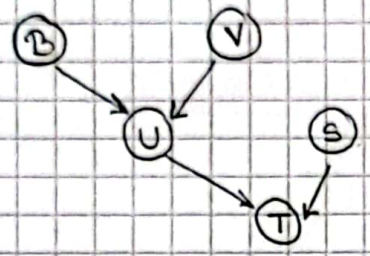
y	z	$P(x=1)$	$P(y)$	$P(z \mid x=1, y)$	
0	0	0.2	0.7	0.5	= 0.07
0	1	0.2	0.7	0.5	= 0.07
1	0	0.2	0.3	0.9	= 0.054
1	1	0.2	0.3	0.1	= 0.006

2.4.

$B, V \to U$
$U, S \to T$

$P(V=1) = 0.3$
$P(B=1) = 0.1$
$P(S=1) = 0.05$

• sve binarno



V	B	$P(U=1\|V,B)$
0	0	0.2
0	1	0.5
1	0	0.4
1	1	0.7

U	S	$P(T=1\|U,S)$
0	0	0
0	1	0.2
1	0	0.4
1	1	0.4

MAP upit: $\underset{U,S}{\operatorname{argmax}}\, P(U,S \mid V=0, T=1)$

$$\frac{p_1}{p_2} = ?$$

$$P(U,S,V,T) = P(B)P(V)P(S)P(U|B,V)P(T|U,S)$$

$$P(U,S|V=0,T=1) = \frac{P(U,S,V=0,T=1)}{P(U,S,V=0,T=1)}$$ ← možemo zanemarit jer se krati u omjeru $\frac{p_1}{p_2}$

$$\underline{P(U,S,V=0,T=1)} = \sum_B P(B)P(V=0)P(S)P(U|B,V=0)P(T=1|U,S)$$

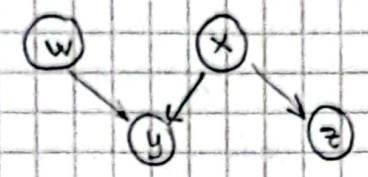
B	U	S	$P(B)$	$P(V=0)$	$P(S)$	$P(U\|B,V=0)$	$P(T=1\|U,S)$	
0	0	0	0.9	0.7	0.95	0.8	0	$= 0$
	0	1	0.9	0.7	0.05	0.8	0.2	$= 5.04 \cdot 10^{-3}$
	1	0	0.9	0.7	0.95	0.2	0.4	$= 0.04788$
	1	1	0.9	0.7	0.05	0.2	0.4	$= 2.52 \cdot 10^{-3}$
1	0	0	0.1	0.7	0.95	0.5	0	$= 0$
	0	1	0.1	0.7	0.05	0.5	0.2	$= 3.5 \cdot 10^{-4}$
	1	0	0.1	0.7	0.95	0.5	0.4	$= 0.0133$
	1	1	0.1	0.7	0.05	0.5	0.4	$= 7 \cdot 10^{-4}$

$P(U=0, S=0, V=0, T=1) = 0$
$P(U=0, S=1, V=0, T=1) = 5.39 \cdot 10^{-3} = 0.00539$ ← p_2
$P(U=1, S=0, V=0, T=1) = 0.06118$ ← p_1
$P(U=1, S=1, V=0, T=1) = 3.22 \cdot 10^{-3} = 0.00322$

$$\frac{p_1}{p_2} = \frac{874}{77} \approx 11.35 \Rightarrow \text{A}$$

2.5

$P(w,x,y,z) = P(w)P(x)P(y|w,x)P(z|x)$ (sve binarno)



$P(w=1) = 0.1$ $\quad P(w=0) = 0.9$

$P(x=1) = 0.2$ $\quad P(x=0) = 0.8$

$P(z=1|x=0) = 0.9$ $\quad P(z=0|x=0) = 0.1$

$P(z=1|x=1) = 0.7$ $\quad P(z=0|x=1) = 0.3$

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

- uzorkovanje s odbijanjem

$P(x=0|y=1, z=0) = \mu$

$N = 100$ uzoraka

$$\hat{\mu}_{MAP} = \frac{N_{x=0,y=1,z=0} + 1}{N_{y=1,z=0} + 2}$$

$\to k_\varepsilon$

$$P(y=1, z=0) = \sum_w \sum_x P(w)P(x)P(y=1|w,x)P(z=0|x) = 0.0274$$

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	$0.1 \cdot 0.2 \cdot 0.7 \cdot 0.3 = 0.0042 = 4.2 \cdot 10^{-3}$

$$P(x=0, y=1, z=0) = \sum_w P(w)P(x=0)P(y=1|w, x=0)P(z=0|x=0)$$

$$= 0.8 \cdot 0.1 \left(P(w=0)P(y=1|w=0, x=0) + P(w=1)P(y=1|w=1, x=0)\right)$$

$$= 0.08(0.9 \cdot 0 + 0.1 \cdot 0.2)$$

$$= 1.6 \cdot 10^{-3} = 0.0016$$

$$\hat{\mu}_{MAP} = \frac{N \cdot 0.0016 + 1}{N \cdot 0.0274 + 2} = \frac{58}{237} \approx 0.245$$

$$P(x=1, y=1, z=0) = \sum_w P(w)P(x=1)P(y=1|w, x=1)P(z=0|x=1)$$

$$= 0.2 \cdot 0.3 \left(P(w=0)P(y=1|w=0, x=1) + P(w=1)P(y=1|w=1, x=1)\right)$$

$$= 0.06(0.9 \cdot 0.4 + 0.1 \cdot 0.7)$$

$$= 0.0258$$

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 minimizicije kriterija pogreške. Razumjeti kako rad algoritma ovisi o broju grupa K i odabiru početnih središta.***]

   Algoritam K-sredina minimizira kriterij pogreške $J(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K|\mathcal{D})$. Vrijednost tog kriterija ovisi o broju grupa K, koji je unaprijed postavljen, te o položajima središta, koja se mijenjaju kroz iteracije.

   (a) Nacrtajte skicu vrijednosti kriterija pogreške J kao funkcije broja grupa K. Koja je minimalna vrijednost funkcije J i zašto?

   (b) 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).

   (c) 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 uzevši 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:

$$\mathcal{D} = \left\{a = (5,2), b = (7,1), c = (1,4), d = (6,2), e = (2,8), f = (3,6), g = (0,4)\right\}.$$

   (a) Izvedite jedan korak algoritma K-sredina uz $K = 3$. Za početna središta odaberite $\boldsymbol{\mu}_1 = b$, $\boldsymbol{\mu}_2 = c$ i $\boldsymbol{\mu}_3 = e$.

   (b) Izvedite jedan korak algoritma K-medoida uz $K = 3$. Za početna središta odaberite primjere b, c i e.

   (c) Usporedite računalnu složenost algoritma K-sredina i K-medoida.

   (d) 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.

   (a) 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 jednako 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 oznake klasa primjera):

$$\left\{\{0,0,1,2\}, \{1,1\}, \{2,2,2,1,0\}\right\}.$$

   (b) Skicirajte vrijednost Randovog indeksa kao funkcije broja grupa K.

   (c) Randov indeks možemo koristiti samo ako su podatci 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 Zadatci 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 dvodimenzijskome 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 dvodimenzijskome 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 grupa (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, uzevši 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):

$$\begin{array}{c} \\ \mathbf{x}^{(1)} \\ \mathbf{x}^{(2)} \\ \mathbf{x}^{(3)} \\ \mathbf{x}^{(4)} \\ \mathbf{x}^{(5)} \end{array} \begin{array}{c} \begin{array}{ccccc} \mathbf{x}^{(1)} & \mathbf{x}^{(2)} & \mathbf{x}^{(3)} & \mathbf{x}^{(4)} & \mathbf{x}^{(5)} \end{array} \\ \begin{pmatrix} 0 & 0.2 & 0.9 & 0.7 & 0.5 \\ & 0 & 0.9 & 0.1 & 0.6 \\ & & 0 & 0.7 & 0.3 \\ & & & 0 & 0.8 \\ & & & & 0 \end{pmatrix} \end{array}$$

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 $\mathrm{RI}(K)$ za grupiranja dobivena s različitim brojem grupa K. Naposlijetku ćemo onda kao optimalan broj grupa odabrati onaj K koji maksimizira Randov indeks, $K^* = \mathrm{argmax}_K \mathrm{RI}(K)$. 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 različitim 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:

$$K = 3 : \{1, 1, 2, 4, 8\}\ \{2, 3, 7\}\ \{4, 5, 3, 5, 6, 6, 7, 8\}$$
$$K = 4 : \{1, 1, 2\}\ \{4, 8, 4\}\ \{2, 3, 7, 5, 7\}\ \{3, 5, 6, 6, 8\}$$
$$K = 5 : \{1, 1\}\ \{3, 4, 8\}\ \{2, 2, 4\}\ \{7, 5, 7, 3, 5, 6\}\ \{6, 8\}$$

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 $\mathrm{RI}(K)$ te optimalan broj grupa K^* prema Randovom indeksu, za $K \in \{3, 4, 5\}$. **Koliko iznosi Randov indeks za optimalan broj grupa,** $\mathrm{RI}(K^*)$**?**

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:

$$K = 3 : \quad \big\{\{1, 2, 4, 4\}, \{2, 3, 3\}, \{1, 1, 3, 4\}\big\}$$
$$K = 4 : \quad \big\{\{1, 2, 2, 4, 4\}, \{3, 3\}, \{1, 1\}, \{3, 4\}\big\}$$

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** $K = 3$**?**

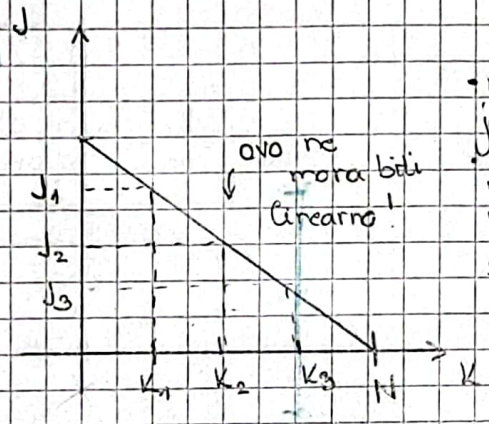
A 0.0164 B 0.0945 C 0.0727 D 0.0682

V19 – Grupiranje

I Zadaci za učenje

1.1. $J(\vec{\mu}_1, \dots, \vec{\mu}_K | D) = \sum_{k=1}^{K} \sum_{i=1}^{N} b_k^i \|\vec{x}^i - \vec{\mu}_k\|^2$

a)

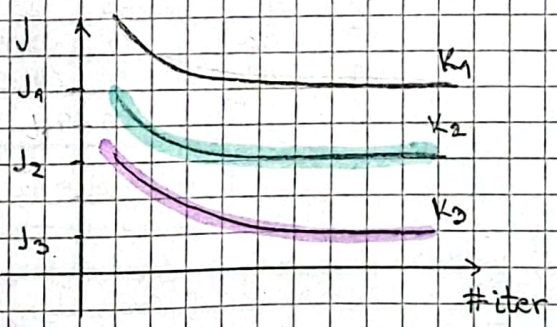


- minimalna vrijednost kriterijske fje je 0
- vrijednost $J = 0$ se postiže za $K = N$, tj. broj grupa jednak je broju primjera pa se efektivno svaki primjer nalazi u zasebnoj grupi
- također svaki primjer je onda i centroid svoje grupe pa je $\|\vec{x}^i - \vec{\mu}_k\| = 0 \Rightarrow J = 0$ //

b)

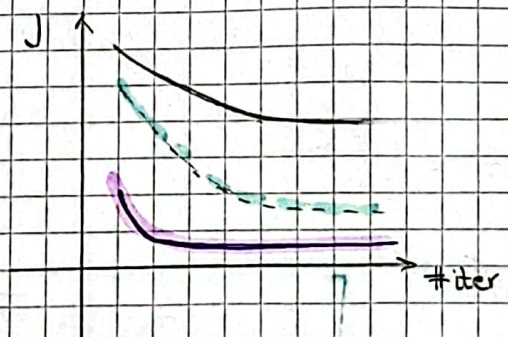
$K_1 < K_2 < K_3$

$\Rightarrow J_1 \geq J_2 \geq J_3$ (može se desiti plato na tim intervalima ali za potrebe skice recimo da nema platea)



- za veći broj grupa K greška postaje manja jer se radi o složenijem modelu
- vrijednost kriterijske fje J nakon određenog broja iteracija postaje konstantna jer algoritam K srednjih vrijednosti konvergira

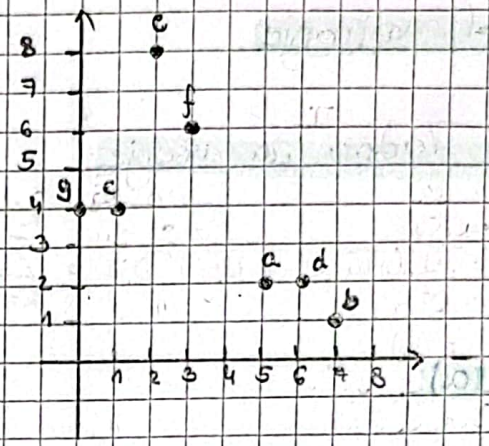
c)



- fiksan K
  - različiti početni centroidi
- razlika je u brzini konvergencije i postignutoj vrijednost kriterijske fje J
- za algoritam K-means++ najvjerojatnija je krivulja — jer najbrže konvergira i jer ima najmanju početnu vrijednost

1.2.

$$D = \{ a=(5,2),\ b=(7,1),\ c=(1,4),\ d=(6,2),\ e=(2,8),\ f=(3,6),\ g=(0,4) \}$$



a) $K = 3$

početna središta $\begin{cases} \vec{\mu_1} = b \\ \vec{\mu_2} = c \\ \vec{\mu_3} = e \end{cases}$

jedan korak algoritma K-means

$$\vec{\mu_k} = \frac{\sum b_k^i \vec{x}^i}{\sum b_k}$$

$$b_k^i = \begin{cases} 1 & \text{ako } k = \operatorname{argmin}_j \|\vec{x}^i - \vec{\mu_j}\| \\ 0 & \text{inače} \end{cases}$$

$\vec{x}^i$	$\|\vec{x}^i - \vec{\mu_1}\|^2$ ($\vec{\mu_1}=(7,1)$)	$\|\vec{x}^i - \vec{\mu_2}\|^2$ ($\vec{\mu_2}=(1,4)$)	$\|\vec{x}^i - \vec{\mu_3}\|^2$ ($\vec{\mu_3}=(2,8)$)	$\vec{b}^i = (b_1^i, b_2^i, b_3^i)$
$a=(5,2)$	5	20	45	(1,0,0)
$b=(7,1)$	0	45	74	(1,0,0)
$c=(1,4)$	45	0	17	(0,1,0)
$d=(6,2)$	2	29	52	(1,0,0)
$e=(2,8)$	74	17	0	(0,0,1)
$f=(3,6)$	41	8	5	(0,0,1)
$g=(0,4)$	58	1	20	(0,1,0)

NOVI CENTROIDI

$$\vec{\mu_1} = \frac{\sum_{i=1}^{6} b_1^i \vec{x}^i}{\sum b_1^i} = \frac{\vec{x}^1 + \vec{x}^2 + \vec{x}^4}{3} = \frac{1}{3}(a+b+d) = \left(6, \frac{5}{3}\right)$$

$$\vec{\mu_2} = \frac{c+g}{2} = \left(\frac{1}{2}, 4\right)$$

$$\vec{\mu_3} = \frac{1}{2}(e+f) = \left(\frac{5}{2}, 7\right)$$

b) jedan korak K-medoida

$K = 3$

$\vec{\mu_1} = b = (7,1)$

$\vec{\mu_2} = c = (1,4)$

$\vec{\mu_3} = e = (2,8)$

$$\vec{\mu_k} = \operatorname{argmin}_{\vec{\mu_j} \in D \setminus M \cup \{\vec{\mu_k}\}} \sum b_k^i \, d(\vec{x}^i, \vec{\mu_j})$$

$$b_k^i = \begin{cases} 1 & k = \operatorname{argmin}_j d(\vec{x}^i, \vec{\mu_j}) \\ 0, & \text{inače} \end{cases}$$

↳ za d koristimo euklidsku udaljenost pa su dobivena područ.

$M = \{b, c, e\}$

NOVI MEDOIDI

$\vec{x}^i$	$\vec{b}^i = (b_1^i, b_2^i, b_3^i)$
$a = (5,2)$	$(1,0,0)$
• $b = (7,1)$	$(1,0,0)$
• $c = (1,4)$	$(0,1,0)$
$d = (6,2)$	$(1,0,0)$
• $e = (2,8)$	$(0,0,1)$
$f = (3,6)$	$(0,0,1)$
$g = (0,4)$	$(0,1,0)$

$$\vec{\mu_1} = \underset{\vec{\mu_j} \in D \setminus M \cup \{\vec{\mu_1}\}}{\operatorname{argmin}} \sum b_1^i \|\vec{x}^i - \vec{\mu_j}\|$$

$$\sum b_1^i \|\vec{x}^i - \vec{\mu_j}\| \iff \sum b_1^i \|\vec{x}^i - \vec{\mu_j}\|^2$$

$$\vec{\mu_j} \in \{a, b, d, f, g\}$$

$$b_1^i = 1 \quad \text{za } \{a, b, d\}$$

$\sum b_1^i v$	$\vec{\mu_j}$	$\|a = (5,2) - \vec{\mu_j}\|^2$	$\|b = (7,1) - \vec{\mu_j}\|^2$	$\|d = (6,2) - \vec{\mu_j}\|^2$
6	$a = (5,2)$	0	5	1
7	$b = (7,1)$	5	0	2
3	$d = (6,2)$	1	2	0
86	$f = (3,6)$	20	41	25
127	$g = (0,4)$	29	58	40

$$\boxed{\vec{\mu_1} = d = (6,2)}$$

$\vec{\mu_2}$

$$\vec{\mu_j} \in \{a, c, d, f, g\}$$

$$b_2^i = 1 \quad \text{za } \{c, g\}$$

$\vec{\mu_j}$	$\|c = (1,4) - \vec{\mu_j}\|^2$	$\|g = (0,4) - \vec{\mu_j}\|^2$	$\sum b_2^i v$
$a = (5,2)$	20	25	45
$c = (1,4)$	0	5	5
$d = (6,2)$	29	36	65
$f = (3,6)$	8	25	33
$g = (0,4)$	1	0	1

$$\boxed{\vec{\mu_2} = g = (0,4)}$$

$\vec{\mu_3}$

$$\vec{\mu_j} \in D / M \cup \{\vec{\mu_3}\} \{a, d, e, f, g\}$$

$$b_3^i = 1 \quad \text{za } \{e, f\}$$

$\vec{\mu_j}$	$\|e = (2,8) - \vec{\mu_j}\|^2$	$\|f = (3,6) - \vec{\mu_j}\|^2$	$\sum b_3^i v$
$a = (5,2)$	45	20	65
$d = (6,2)$	52	25	77
$e = (2,8)$	0	5	5
$f = (3,6)$	5	0	5
$g = (0,4)$	20	10	30

ovisi o ind.

$$\boxed{\vec{\mu_3} = e = (2,8)}$$

c) složenost algoritma K-sredina

$$O(TnNK) = T \cdot (O(nNK) + O(nN))$$

$O(nNK)$

- složenost pridjeljivanja koeficijenta b_k^i
- složenost euklidske udaljenosti $O(n)$ računa se $\forall \vec{x}^i$ i $\vec{\mu}_j$ jer je potrebno odrediti $\underset{j}{\operatorname{argmin}} \| \vec{x}^i - \vec{\mu}_j \|$

$O(nN)$

- složenost izračuna centroida $\frac{\sum b_k^i \vec{x}^i}{\sum b_k^i}$

• linearna složenost u svim parametrima
T = 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_k^i
- svaki centroid se uspoređuje sa $\vec{x}^i \in D \setminus M$ čija je kardinalnost $N-K$

$O(K(N-K)^2)$

- odabir novog medoida: $\underset{\vec{\mu}_j \in D \setminus M \cup \{\vec{\mu}_k\}}{\operatorname{argmin}} \sum b_k^i \, v(\vec{x}^i, \vec{\mu}_j)$

• kvadratna složenost u N

d) Prednosti K-medoida

- primjeri ne moraju biti vektori
- sličnost se ne mora mjeriti euklid. udaljenošću

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 oznake klasa
- podskupovi = grupe dobivene grupiranjem

$$R = \frac{a+b}{\binom{N}{2}}$$

a = broj jednako označenih parova u istim grupama

b = broj različito označenih parova u različitim grupama

$N = 10$

parovi 0 u I $\binom{2}{2} = 1$
parovi 1 u II $\binom{2}{2} = 1$
parovi 2 u III $\binom{3}{2} = 3$

$a = 5$

I s II:
- 0 iz I s 1 iz II : $2 \cdot 2 = 4$
- 1 iz I s 1 iz II : 0
- 2 iz I s 1 iz I : $1 \cdot 2 = 2$

I s III:
- 0 iz I s 2 iz III : $2 \cdot 3 = 6$
- 0 iz I s 1 iz III : $2 \cdot 1 = 2$
- 1 iz I s 2 iz III : $1 \cdot 3 = 3$
- 1 iz I s 0 iz III : $1 \cdot 1 = 1$
- 2 iz I s 1 iz III : $1 \cdot 1 = 1$
- 2 iz I s 0 iz III : $1 \cdot 1 = 1$

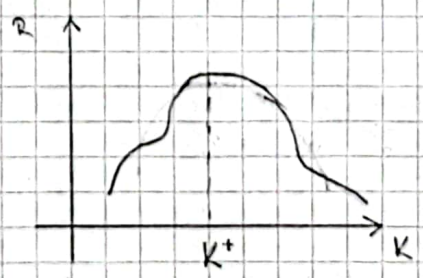
II s III:
- 1 iz II s 2 iz III : $2 \cdot 3 = 6$
- 1 iz II s 0 iz III = $2 \cdot 1 = 2$

$b = 28$

$$R = \frac{a+b}{\binom{N}{2}} = \frac{28+5}{\binom{11}{2}}$$

$$R = \frac{3}{5} = 0.6$$

b)



! nije kvadratna ovisnost, samo je fja unimodalna

c) Ako unaprijed znamo broj grupe imamo koristi od Randovog indeksa. Indeks nam u tom slučaju služi kao mjera uspješnog klasificiranja podataka.

Ako nam broj grupa nije unaprijed poznat možemo koristit Randov indeks na manjem ručno označenom skupu podataka.

II Zadaci s ispita

2.1.

$K = 2$ Kmeans

$$D = \{(\vec{x}^i)\}_i = \{(1,1), (1,2), (2,2), (2,3), (3,3)\}$$

$$\vec{\mu_1} = \vec{x}^2 = (1,2)$$
$$\vec{\mu_2} = \vec{x}^5 = (3,3)$$

$$b_k^i = \begin{cases} 1 & k = \operatorname{argmin} \|\vec{x}^i - \vec{\mu_j}\| \\ 0 & \text{inače} \end{cases} \qquad \vec{\mu_k} = \frac{\sum b_k^i \vec{x}^i}{\sum b_k^i}$$

$\vec{x}^i$	$\|\vec{x}^i - \vec{\mu_1} = (1,2)\|^2$	$\|\vec{x}^i - \vec{\mu_2} = (3,3)\|^2$	$\vec{b}^i = (b_1^i, b_2^i)$
(1,1)	1	8	(1,0)
(1,2)	0	5	(1,0)
(2,2)	1	2	(1,0)
(2,3)	2	1	(0,1)
(3,3)	5	0	(0,1)

$$\vec{\mu_1} = \frac{\vec{x}^1 + \vec{x}^2 + \vec{x}^3}{3} = \left(\frac{4}{3}, \frac{5}{3}\right)$$

$$\vec{\mu_2} = \frac{\vec{x}^4 + \vec{x}^5}{2} = \left(\frac{5}{2}, 3\right)$$

$$J = \sum_{k=1}^{K} \sum_{i=1}^{N} b_k^i \|\vec{x}^i - \vec{\mu_k}\|^2 = \frac{11}{6} = 1.83\dot{3} \quad \text{B}$$

$$\|\vec{x}^1 - \vec{\mu_1}\| = \sqrt{5}/3 \qquad \|\vec{x}^4 - \vec{\mu_2}\| = 1/2$$
$$\|\vec{x}^2 - \vec{\mu_1}\| = \sqrt{2}/3 \qquad \|\vec{x}^5 - \vec{\mu_2}\| = 1/2$$
$$\|\vec{x}^3 - \vec{\mu_1}\| = \sqrt{5}/3$$

$\vec{x}^i$	$\|\vec{x}^i - \vec{\mu_1}\|$	$\|\vec{x}^i - \vec{\mu_2}\|$	$\vec{b}^i$
(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)	2.1343	0.5	(0,1)

2.2.

$\vec{x}^i$
(0,0)
(0,4)
(2,0)
(2,4)
(4,2)

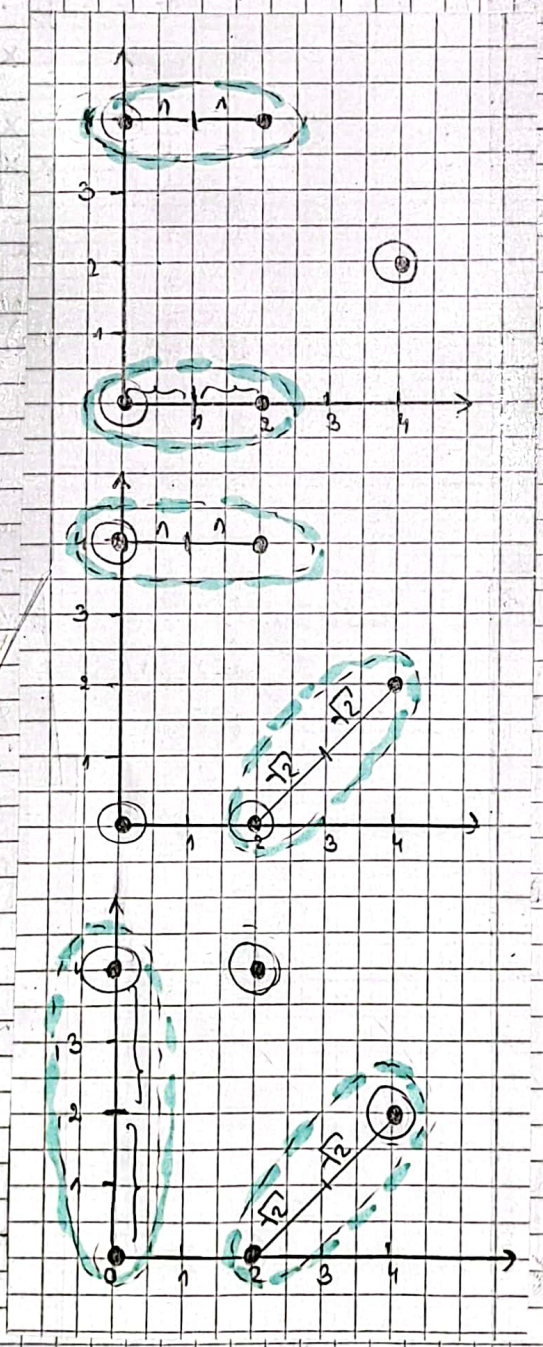
$N = 5$
$k = 3$

• početna središta nasumično odabrana iz D

$$J = \sum_{k=1}^{3} \sum_{i=1}^{5} b_k^i \|\vec{x}^i - \vec{\mu}_j\|^2$$

J^+ = lokalni minimum
J^* = globalni minimum

$J^+ - J^* = ?$



$J = 4 \cdot 1^2 = 4 = J^*$

$J = 2 \cdot 1^2 + 2 \cdot (\sqrt{2})^2 = 6$

$J = 2 \cdot 2^2 + 2 \cdot (\sqrt{2})^2 = 12 = J^+$

$J^+ - J^* = 12 - 4 = 8$ //

2.3.

$D = \{\vec{x}^i\} = \{(0,0), (0,4), (2,4), (4,2)\}$

$K = 2$

K-means – pretražuje K^N particija

$K^N = 2^4 = 16$

K-means

Primjeri	1	2	3	4	
1, 0 → grupa	0	0	0	0	
	0	0	0	1	
	0	0	1	0	
	0	0	1	1	
	0	1	0	0	
	0	1	0	1	⇒ ova grupa nije moguća (križanje)
	0	1	1	0	
	0	1	1	1	
ostalo sim. →	1	0	0	0	

$|S(A_1)| = 7$

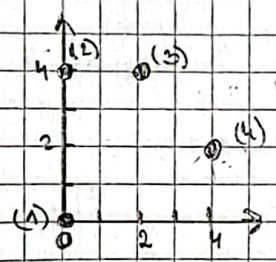
K-means ++

	1	2	3	4	1	2	3	4
1	×	×			0	1	1	0/1
2	×		×		0	1	1	1
3	×			×	0	0	1	1
4		×	×		0	0	1	1
5		×		×	0	0	0	1
6			×	×	0/1	0	0	1

(1 2 3 4: početna središta primjeri; 1 2 3 4: klasifikacija)

inicijalna – $\binom{4}{2} = 6$ moguć

$|S(A_2)| = 4$



$|S(A_1)| - |S(A_2)| = 3$ //

2.4.

K-medoida

$N = 5$

$K = 2$

$\vec{\mu}_1 = \vec{x}^1$

$\vec{\mu}_2 = \vec{x}^5$

	$\vec{x}^1$	$\vec{x}^2$	$\vec{x}^3$	$\vec{x}^4$	$\vec{x}^5$
$\vec{x}^1$	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

$$b_k^i = \begin{cases} 1 & \text{ako } k = \operatorname{argmin}_j v(\vec{x}^i, \vec{\mu}_j) \\ 0 & \text{inače} \end{cases}$$

$$\vec{\mu}_k = \operatorname{argmin}_{j \in D \setminus \{\vec{\mu}_k\}} \sum b_k^i v(\vec{x}^i, \vec{\mu}_j)$$

	$\vec{b}^i$
$\vec{x}^1$	(1, 0)
$\vec{x}^2$	(1, 0)
$\vec{x}^3$	(0, 1)
$\vec{x}^4$	(1, 0)
$\vec{x}^5$	(0, 1)

Kandidati za $\vec{\mu}_1 \in \{\vec{x}^1, \vec{x}^2, \vec{x}^3, \vec{x}^4\}$

	$v(\vec{x}^i, \vec{x}^1)$	$v(\vec{x}^i, \vec{x}^2)$	$v(\vec{x}^i, \vec{x}^3)$	$v(\vec{x}^i, \vec{x}^4)$
$\vec{x}^1$	0	0.2	0.9	0.7
$\vec{x}^2$	0.2	0	0.9	0.1
$\vec{x}^4$	0.7	0.1	0.7	0
	0.9	0.3	2.7	0.8

$\downarrow$

$\boxed{\vec{\mu}_1 = \vec{x}^2}$

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
	1.5	0.3	1.5	0.3

$\boxed{\vec{\mu}_2 = \vec{x}^3 \text{ ili } \vec{x}^5}$ - odluka ovisi o preferenciji pretraživ. algoritma

D $\vec{x}^2$ i $\vec{x}^5$

2.5.

$N = 1000$
$K = 3$ $\binom{3}{2}$ $\binom{3}{2}$ $\binom{3}{2}$

$\{\{1,1,0\}, \{1,0,1\}, \{0,2,1,2\}\}$

$a = 1 + 1 + 1 = 3$

$b = 2 \cdot 1 + 1 \cdot 2 + 2 \cdot 1 + 2 \cdot 2 + 1 \cdot 1 + 1 \cdot 2 + 2 \cdot 1 + 2 \cdot 2 + 1 \cdot 2 + 1 \cdot 1$
$= 2 + 2 + 2 + 4 + 1 + 2 + 2 + 4 + 2 + 1$
$= 10 + 1 + 10 + 1$
$= 22$

$R = \frac{a+b}{\binom{N}{2}} = \frac{22+3}{\binom{10}{2}} = \frac{5}{9} \approx 0.5556$ B

2.6.

$N = 1000$

Parovi 1, 2 i 3 → parovi iz iste grupe
4, 5, 6, 7 i 8 → parovi iz različ. grupe

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}

Parovi	true ista grupa?	K=3 ista grupa?	K=4 ista grupa?	K=5 ista grupa?
1-1	1	1	1	1
2-2	1	0	0	1
3-3	1	0	0	0
4-4	0	0	1	0
5-5	0	1	0	1
6-6	0	1	1	0
7-7	0	0	1	1
8-8	0	0	0	0

$RI(K) = \frac{TN+TP}{\binom{N}{2}}$

$RI(K=3) = \frac{3+1}{8} = \frac{4}{8} = 0.5$

$RI(K=4) = \frac{1+2}{8} = \frac{3}{8} = 0.375$

$RI(K=5) = \frac{2+3}{8} = \frac{5}{8} = 0.625$

Optimalni $RI(K^* = 5) = 0.625$

B

2.7

$N = 1000$
$Y = \{1,2,3,4\}$

$K = 3$: $\{ \{1,2,4,4\}, \{2,3,3\}, \{1,1,3,4\} \}$
$K = 4$: $\{ \{1,2,2,4,4\}, \{3,3\}, \{1,1\}, \{3,4\} \}$

Za $K = 3$

$$a = 1+1+1 = 3$$
$$b = 1+2+2+2+4 + 1+1+2+1+1+4+2 + 2+1+1+4+2$$
$$b = 11+12+10 = 33$$
$$R = \frac{a+b}{\binom{N}{2}} = \frac{36}{\binom{11}{2}} = \frac{36}{55}$$

Za $K=4$

$$a = 1+1+1+1 = 4$$
$$b = 2+4+4 + 4+4 + 1+1+2+2+2 + 4+2 + 2+2$$
$$b = 10+8+8+6+4 = 36$$
$$R = \frac{a+b}{\binom{N}{2}} = \frac{40}{\binom{11}{2}} = \frac{8}{11}$$

$$\text{razlika} = \frac{8}{11} - \frac{36}{55} = \frac{4}{55} \approx 0{,}07273 \Rightarrow \text{C}$$

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-izglednost 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.

   (a) Što je prednost, a što nedostatak, algoritma maksimizacije očekivanja primijenjenog na GMM u odnosu na algoritam K-sredina?

   (b) Napišite izraz za gustoću $p(\mathbf{x})$ za model miješane gustoće (bez latentnih varijabli) i izraz za pripadnu (nepotpunu) log-izglednost.

   (c) 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?

   (d) Definirajte E-korak i M-korak algoritma maksimizacije očekivanja primijenjenog na Gaussovu mješavinu.

   (e) 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 jednostruke i potpune povezanosti.***] Jednako kao i algoritam K-medoida, algoritam hijerarhijskog aglomerativnog grupiranja može se primijeniti u slučajevima kada primjeri nisu prikazani kao vektori značajki te kada umjesto mjere udaljenosti između vektora raspolažemo općenitjom mjerom sličnosti (ili različitosti). Neka je *sličnost* primjera iz $\mathcal{D}$ definirana sljedećom matricom sličnosti:

$$S = \begin{array}{c} \\ a \\ b \\ c \\ d \\ e \end{array} \begin{array}{c} \begin{array}{ccccc} a & b & c & d & e \end{array} \\ \begin{pmatrix} 1.00 & 0.26 & 0.15 & 0.20 & 0.17 \\ 0.26 & 1.00 & 0.24 & 0.31 & 0.31 \\ 0.15 & 0.24 & 1.00 & 0.20 & 0.50 \\ 0.20 & 0.31 & 0.20 & 1.00 & 0.24 \\ 0.17 & 0.31 & 0.50 & 0.24 & 1.00 \end{pmatrix} \end{array}$$

   (a) Izgradite dendrogram uporabom jednostrukog povezivanja. Kada bi bilo potrebno napraviti particiju grupa, na kojoj biste razini presjekli taj dendrogram?

   (b) Izgradite dendrogram uporabom potpunog povezivanja. Kada bi bilo potrebno napraviti particiju grupa, na kojoj biste razini presljekli taj dendrogram?

3. [***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.***]

   (a) 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):

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

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 Zadatci 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:

$$\begin{array}{ll} \mathcal{H}_1: & K = 2 \text{ grupa, puna kovarijacijska matrica} \\ \mathcal{H}_2: & K = 2 \text{ grupa, izotropna kovarijacijska matrica} \\ \mathcal{H}_3: & K = 3 \text{ grupe, izotropna kovarijacijska matrica} \end{array}$$

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 $\mathcal{H}_\alpha$, to jest vjerojatnost pripadanja tog primjera najvjerojatnijoj grupi dobivenoj grupiranjem pomoću modela $\mathcal{H}_\alpha$. **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 inicijaliziranima algoritmom K-sredina

$\mathcal{H}_2$: Model sa $K = 50$ središta inicijaliziranima 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 inicijaliziranima 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 $\mathcal{H}_\alpha$ 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 $LL_2^0 \geqslant LL_4^0$, $LL_1^* \geqslant LL_2^* \geqslant LL_3^*$

B $LL_3^0 \geqslant LL_4^0$, $LL_1^* \geqslant LL_3^* \geqslant LL_4^*$

C $LL_2^0 \geqslant LL_4^0 \geqslant LL_3^0$, $LL_1^* \geqslant LL_2^*$

D $LL_2^0 \leqslant LL_4^0$, $LL_2^* \leqslant LL_1^* \geqslant LL_3^*$

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 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 $\mathcal{H}_\alpha$ 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 $LL_1^0 \geqslant LL_2^0$, $LL_2^* \geqslant LL_3^*$, $k_1 \geqslant k_2$

B $LL_1^0 \geqslant LL_3^0$, $LL_1^* \geqslant LL_3^*$, $k_2 \geqslant k_1$

C $LL_2^0 \geqslant LL_3^0$, $LL_3^* \geqslant LL_2^*$, $k_3 \geqslant k_2$

D $LL_3^0 \geqslant LL_2^0$, $LL_3^* \geqslant LL_2^*$, $k_2 \geqslant k_3$

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:

$$p(\mathbf{x}) = \sum_{j=1}^{3} \frac{1}{3}\mathcal{N}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \quad \text{gdje} \quad \mu_1 = (5,5), \mu_2 = (5,10), \mu_3 = (-10,-10), \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_3 = 2\mathbf{I}$$

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 dijagonalna 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):

$$\begin{array}{c} \\ \mathbf{x}^{(1)} \\ \mathbf{x}^{(2)} \\ \mathbf{x}^{(3)} \\ \mathbf{x}^{(4)} \\ \mathbf{x}^{(5)} \end{array} \begin{array}{c} \begin{array}{ccccc} \mathbf{x}^{(1)} & \mathbf{x}^{(2)} & \mathbf{x}^{(3)} & \mathbf{x}^{(4)} & \mathbf{x}^{(5)} \end{array} \\ \begin{pmatrix} 1 & 0.4 & 0.5 & 0.7 & 0.5 \\ & 1 & 0.9 & 0.3 & 0.6 \\ & & 1 & 0.7 & 0.1 \\ & & & 1 & 0.8 \\ & & & & 1 \end{pmatrix} \end{array}$$

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:

$$\begin{pmatrix} 1.0 & 0.2 & 0.8 & 0.1 & 0.4 \\ 0.2 & 1.0 & 0.9 & 0.3 & 0.7 \\ 0.8 & 0.9 & 1.0 & 0.6 & 0.5 \\ 0.1 & 0.3 & 0.6 & 1.0 & 0.4 \\ 0.4 & 0.7 & 0.5 & 0.4 & 1.0 \end{pmatrix}$$

Provedite 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 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 su operacije unije i presjeka definirane nad skupovima slova od kojih se riječi sastoje. Npr., $\kappa(\text{“water”}, \text{“watering”}) = 5/8 = 0.625$. Provedite 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 primjenjenog na GMM
- probabilistički izlaz (meko grupiranje)

Nedostatak EM
- složeniji od K-means
- nema rješenje u zatvorenoj formi

b) model miješane gustoće: $p(\vec{x}) = \sum_{k=1}^{K} p(\vec{x}, y=k)$

$$= \sum_{k=1}^{K} P(y=k)\, p(\vec{x} \mid y=k)$$

$$= \sum_{k=1}^{K} \pi_k\, p(\vec{x} \mid \vec{\theta}_k)$$

nepotpuna log-izglednost

$$\ln \mathcal{L}(\vec{\theta} \mid D) = \ln P(D \mid \vec{\theta}) = \ln \prod_{i=1}^{N} p(\vec{x}^i)$$

$$= \ln \prod_{i=1}^{N} \sum_{k=1}^{K} \pi_k \cdot p(\vec{x}^i \mid \vec{\theta}_k)$$

$$= \sum_i \ln \sum_k \pi_k\, p(\vec{x}^i \mid \vec{\theta}_k)$$

c) model miješane gustoće s latentnim varijablama

$$p(\vec{x}, \vec{z}) = P(\vec{z})\, P(\vec{x} \mid \vec{z}, \vec{\theta})$$

$$P(\vec{z}^i = e_k) = \prod_{k=1}^{K} \pi_k^{z_k}$$

$$p(\vec{x}, \vec{z}) = \prod_{k=1}^{K} \pi_k^{z_k} \prod_{k=1}^{K} p(\vec{x} \mid \vec{\theta}_k)^{z_k} = \prod_{k=1}^{K} \pi_k^{z_k}\, p(\vec{x} \mid \vec{\theta}_k)^{z_k}$$

potpuna log-izglednost

$$\ln \mathcal{L}(\vec{\theta} \mid D, Z) = \ln P(D, Z \mid \vec{\theta}) = \ln \prod_{i=1}^{N} p(\vec{x}^i, \vec{z}^i)$$

$$= \ln \prod_i \prod_k \pi_k^{z_k^i}\, p(\vec{x}^i \mid \vec{\theta}_k)^{z_k^i}$$

$$= \sum_{i=1}^{N} \sum_{k=1}^{K} z_k^i \left(\ln \pi_k + \ln p(\vec{x}^i \mid \vec{\theta}_k) \right)$$

↳ ne možemo dalje raditi izravno s tom izglednošću jer ne znamo vrijednosti latentnih varijabli

d) E-korak (prosjek očekivanja potpune log-izgl.)

$$Q(\theta|\theta^t) = \mathbb{E}_{Z|D,\theta^t}\left[\ln \mathcal{L}(\vec{\theta}|D,Z)\right]$$

$$= \mathbb{E}\left[\sum_{i=1}^{N}\sum_{k=1}^{K} z_k^i \left(\ln \pi_k + \ln p(\vec{x}^i|\vec{\theta}_k)\right)\right]$$

$$= \sum_{i=1}^{N}\sum_{k=1}^{K} \mathbb{E}\left[z_k^i|\theta^t, D\right]\left(\ln \pi_k + \ln p(\vec{x}^i|\vec{\theta}_k)\right)$$

→ očekivanje potpune izgled: uz fiks. parametre $\vec{\theta}^t$

$$\left[\mathbb{E}\left[z_k^i|D,\vec{\theta}^t\right] = \mathbb{P}(z_k^i = 1|\vec{x}^i, \theta^t) = \frac{p(\vec{x}^i|\vec{\theta}_k^t)\pi_k^t}{\sum_{j=1}^{K} p(\vec{x}^i|\vec{\theta}_j^t)\pi_j^t} = h_k^i\right]$$

odgovornost!

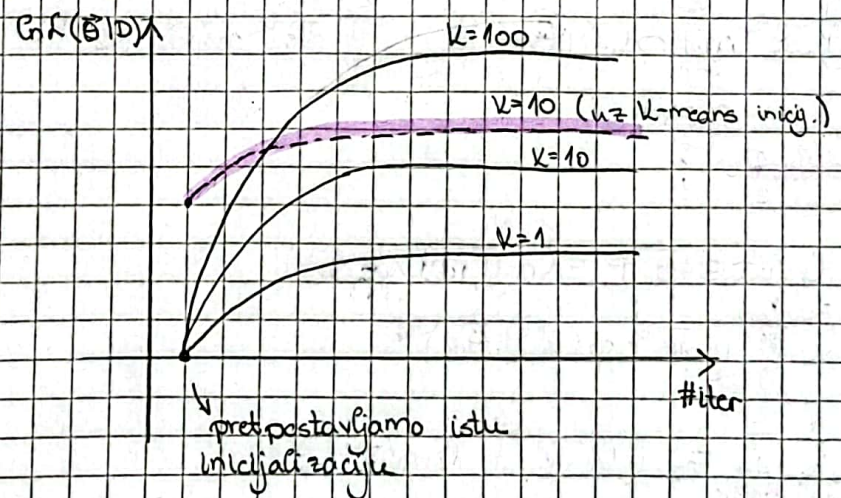
M-korak (maksimiz. potpune izglednosti po π_k i $\vec{\theta}_k$)

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

$$\nabla_\theta Q(\theta|\theta^t) = 0$$

$$\hookrightarrow \left[\begin{aligned} \pi_k^{t+1} &= \frac{1}{N}\sum_{i=1}^{N} h_k^i \\ \vec{\mu}_k^{t+1} &= \frac{\sum h_k^i \vec{x}^i}{\sum h_k^i} \\ \Sigma_k^{t+1} &= \frac{\sum_i h_k^i (\vec{x}^i - \vec{\mu}_k^{t+1})(\vec{x}^i - \vec{\mu}_k^{t+1})^T}{\sum_i h_k^i} \end{aligned}\right]$$

e)

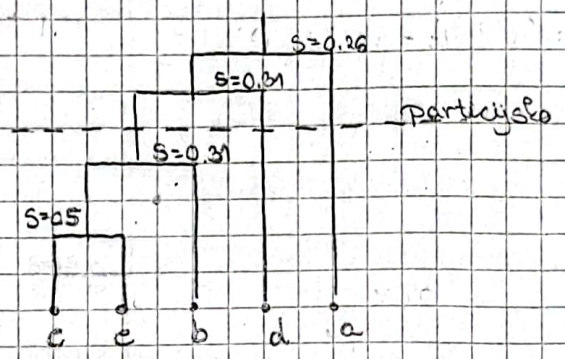


1.2. HAC

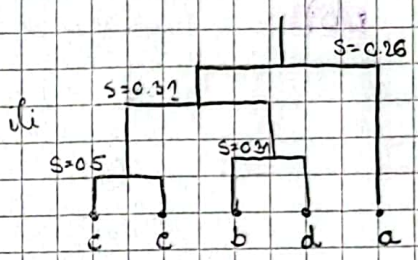
$$S = \begin{array}{c} \\ a \\ b \\ c \\ d \\ e \end{array} \begin{array}{c} \begin{array}{ccccc} a & b & c & d & e \end{array} \\ \begin{bmatrix} 1 & 0.26 & 0.15 & 0.2 & 0.17 \\ 0.26 & 1 & 0.24 & 0.31 & 0.31 \\ 0.15 & 0.24 & 1 & 0.2 & 0.5 \\ 0.2 & 0.31 & 0.2 & 1 & 0.24 \\ 0.17 & 0.31 & 0.5 & 0.24 & 1 \end{bmatrix} \end{array}$$

a) Dendogram dobiven jednostrukim povezivanjem

$$d_{min}(G_i, G_j) = \min_{\vec{x} \in G_i, \vec{y} \in G_j} d(\vec{x}, \vec{y}) = \max S(\vec{x}, \vec{y})$$



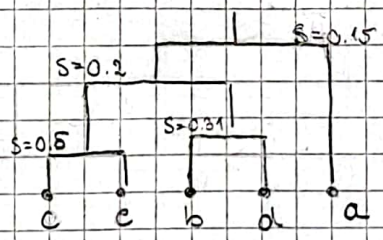
ili



b) Dendogram dobiven potpunim 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 listove spajaš koji su najbliže



1. grupa $\{c\}$ najbliža $\{e\}$

2. $\{c,e\} - b = 0.24$
   $\{c,e\} - d = 0.2$
   $\{c,e\} - a = 0.15$

   $b - d = 0.31 \longrightarrow \{b\}$ i $\{d\}$ najbliži
   $b - a = 0.26$

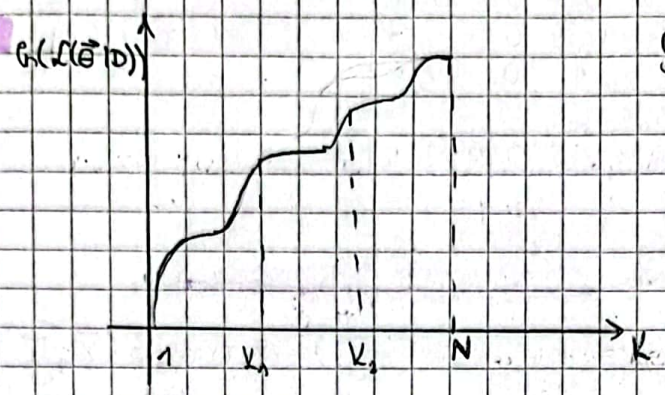
   $d - a = 0.2$

3. $\{c,e\} - \{b,d\} = 0.2 \rightarrow \{c,e\}$ i $\{b,d\}$ najbliže
   $\{c,e\} - a = 0.15$
   $\{b,d\} - a = 0.2$

1.3.

$$\boxed{\ln \mathcal{L}(\vec{\theta} \mid D) \propto -J}$$

a)



log-izglednost raste kako broj grupa K raste jer vrijedi veza $\ln(\mathcal{L}(\vec{\theta} \mid D)) \propto -J$. Također raste složenost modela pa ima smisla da izglednost raste.

Temeljem ove krivulje može se pokušati odrediti optimalni K metodom lakta

→ kandidati su slike K_1 i K_2

→ ako vidimo stagnaciju dodatno dijelimo prirodne grupe

b) Akaikeov informacijski kriterij (AIC)

$$K^* = \underset{K}{\arg\min} \left(-2\ln \mathcal{L}(K) + 2g(K) \right)$$

$g(K)$ = broj parametara modela s K grupa

$K = 5$

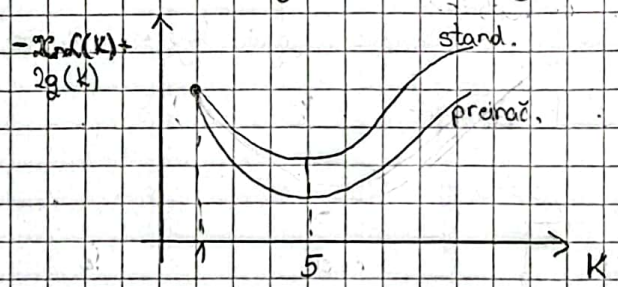
- standardni alg.
- preinačeni - dijeljena Σ

ovaj se više kažnjava

$$g(K) = K \cdot \frac{n}{2}(n+1) + K - 1 + nK = 5n + 4 + 5\frac{n}{2}(n+1)$$

$$g(K) = \frac{n}{2}(n+1) + K - 1 + nK$$

$$= \frac{n}{2}(n+1) + 5n + 4$$

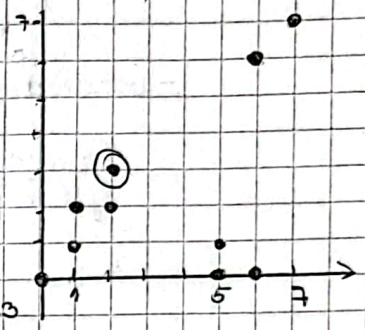


II Zadaci s ispita

2.1.

$N = 10$

$D = \{(0,0), (1,1), (1,2), (2,2), (2,3), (5,0), (5,1), (6,0), (6,6), (7,7)\}$



$\mathcal{H}_1$: $K=2$, puna Σ
$\mathcal{H}_2$: $K=2$, izotropna Σ
$\mathcal{H}_3$: $K=3$, izotropna Σ
} nedijelj.

↳ očekujemo da je odgovornost veća za manji broj klasa (A)

$h_{\alpha_1} > h_{\alpha_2} > h_{\alpha_3}$

↳ usporedivi modeli

OBLIK IZOKONTURA

- $\mathcal{H}_1 - \mathcal{H}_2$
  → $\mathcal{H}_1$ bolje opisuje dani skup podataka ⇒ $h_{\alpha_1} > h_{\alpha_2}$
  ⇒ zbog tranzitivnosti: $h_{\alpha_1} > h_{\alpha_2} > h_{\alpha_3}$
- $\mathcal{H}_2 - \mathcal{H}_3$
  ↳ zbog broja klasa $h_{\alpha_2} \geq h_{\alpha_3}$ (A)

2.2

$\mathcal{H}_1$: $K=50$, incij. K-means
$\mathcal{H}_2$: $K=50$, inicij K-means i dijelj. Σ
$\mathcal{H}_3$: $K=50$, inicij. sluč. i dijelj. Σ
$\mathcal{H}_4$: $K=10$, inicij. K-means i dij. Σ

K-means ubrzava konvergenciju

- usporedivi modeli
  - $\mathcal{H}_1 - \mathcal{H}_2$: $LL_1^0 \geq LL_2^0$
    - očekujemo da će nedijeljena Σ bolje procijeniti parametre grupa
    $LL_1^* \geq LL_2^*$
  - $\mathcal{H}_2 - \mathcal{H}_3$: $LL_2^0 \geq LL_3^0$
    - očekujemo da će inicijalizacija K-meansom dati bolje rezultate
    $LL_2^* \geq LL_3^*$ ⇒ ovo bi moglo konv. u isto
  - $\mathcal{H}_2 - \mathcal{H}_4$: $LL_2^0 \geq LL_4^0$, $LL_2^* \geq LL_4^*$ } veća klasa bolja izglednost

Zaključci
$LL_2^0 \geq LL_4^0$
$LL_1^* \geq LL_2^* \geq LL_3^*$ (A)

2.3.

$\mathcal{H}_1$: $K=25$, inicij. K-means++
$\mathcal{H}_2$: $K=50$, slučajna inicij. i dij. Σ
$\mathcal{H}_3$: $K=50$, sred. inicij. K-means, dij. Σ

k_K = broj iteracija, (ekviv. koji brže konvergira)

- usporedivi modeli
  - $\mathcal{H}_2 - \mathcal{H}_3$
    - očekujemo da K-means daje bolje rezultate i to u manjem broju iteracija

$$\left.\begin{array}{l} k_2 \geq k_3 \\ LL_3^{\circ} \geq LL_2^{\circ} \\ LL_3^{*} \geq LL_2^{*} \end{array}\right\} \text{D}$$

- model $\mathcal{H}_1$ hipotetski jako teško možemo usporediti sa $\mathcal{H}_2$ i $\mathcal{H}_3$

2.4.

$$p(\vec{x}) = \sum_{j=1}^{3} \frac{1}{3} \mathcal{N}(\vec{\mu}_j, \Sigma_j)$$

$\Sigma_1 = \Sigma_2 = \Sigma_3 = 2I$ → matrice u stvarnosti izotropne i dijeljenje
↳ izokonture su kružnice

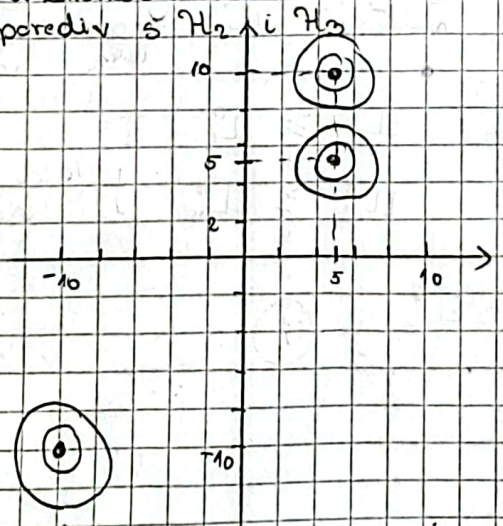
$\vec{\mu}_1 = (5,5)$
$\vec{\mu}_2 = (5,10)$
$\vec{\mu}_3 = (-10,-10)$

S obzirom na dane modele i dani skup možemo zaključiti: $\boxed{\begin{array}{l} \mathcal{L}_2 > \mathcal{L}_3 \\ \mathcal{L}_2 > \mathcal{L}_1 \end{array}}$ D

$\boxed{K=2}$

$\mathcal{H}_2$ sož. od $\mathcal{H}_3$ $\left\{\begin{array}{l} \mathcal{H}_1 - \text{dij. i puna } \Sigma \\ \mathcal{H}_2 - \text{nedij. i dija. } \Sigma \\ \mathcal{H}_3 - \text{nedij. i izotr. } \Sigma \end{array}\right.$

- u kontekstu složenosti $\mathcal{H}_1$ ne usporediv s $\mathcal{H}_2$ i $\mathcal{H}_3$



- inicijalizacija K-means++

$\mathcal{L}_i$ = izgled. modela $\mathcal{H}_i$

$$\mathcal{L}(\vec{\theta}|D) = p(D|\vec{\theta}) = \prod_{i=1}^{N} p(\vec{x}^i|\vec{\theta}) = \prod_{i=1}^{N} \sum_{k=1}^{K} \pi_k p(\vec{x}^i|\vec{\theta}_k)$$

$\mathcal{H}_1$: puna $\Sigma_1 = \Sigma_2$
→ omogućava zakošene elipse u izokonturama
→ ne pomaže baš s obzirom da izokonture moraju biti iste za obje grupe

$\mathcal{H}_2$: izokonture elipse i to različite za iste grupe

$\mathcal{H}_3$: izokonture kružnice, ali iste za obje grupe

2.5.

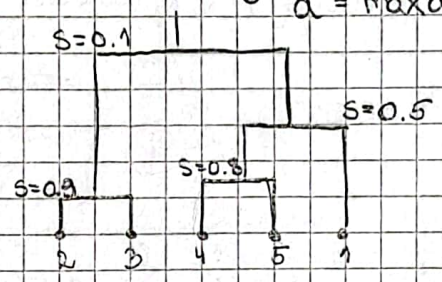
$N=5$

• sličnost

$$\begin{array}{c|ccccc} & \vec{x}^1 & \vec{x}^2 & \vec{x}^3 & \vec{x}^4 & \vec{x}^5 \\ \hline \vec{x}^1 & 1 & 0.4 & 0.5 & 0.7 & 0.5 \\ \vec{x}^2 & 0.4 & 1 & 0.9 & 0.3 & 0.6 \\ \vec{x}^3 & 0.5 & 0.9 & 1 & 0.7 & 0.1 \\ \vec{x}^4 & 0.7 & 0.3 & 0.7 & 1 & 0.8 \\ \vec{x}^5 & 0.5 & 0.6 & 0.1 & 0.8 & 1 \end{array}$$

• potpuno povezivanje

$$d = \max d(\vec{x}, \vec{y}) = \min S(\vec{x}, \vec{y})$$



I najsliče $\vec{x}^2$ i $\vec{x}^3$

II
$\{2,3\} - 1 = 0.4$
$\{2,3\} - 4 = 0.3$
$\{2,3\} - 5 = 0.1$
$1 - 4 = 0.7$
$1 - 5 = 0.5$
$4 - 5 = 0.8$

II
$\{2,3\} - 1 = 0.4$
$\{2,3\} - \{4,5\} = 0.1$
$\{4,5\} - 1 = 0.5$

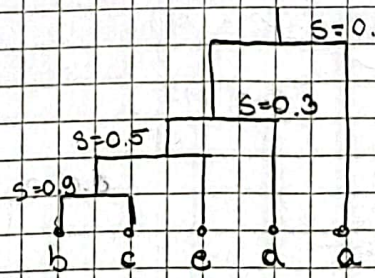
C $((\vec{x}^2, \vec{x}^3), ((\vec{x}^4, \vec{x}^5), \vec{x}^1))$

2.6.

$N=5$

$$\begin{array}{c|ccccc} & a & b & c & d & e \\ \hline a & 1 & 0.2 & 0.8 & 0.1 & 0.4 \\ b & 0.2 & 1 & 0.9 & 0.3 & 0.7 \\ c & 0.8 & 0.9 & 1 & 0.6 & 0.5 \\ d & 0.1 & 0.3 & 0.6 & 1 & 0.4 \\ e & 0.4 & 0.7 & 0.5 & 0.4 & 1 \end{array}$$

• potpuno poveziv.
$d = \min S$



I $\{b\}$ i $\{c\}$

II
$\{b,c\} - a = 0.2$
$- d = 0.3$
$- e = 0.5$
$a - d = 0.1$
$- e = 0.4$
$d - e = 0.4$

III
$\{b,c,e\} - a = 0.2$
$- d = 0.3$
$a - d = 0.1$

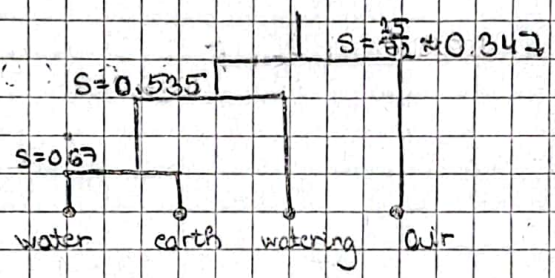
$0.9 + 0.5 + 0.3 + 0.1$
$= 1.4 + 0.4 = 1.8$ A

2.7.

$$K(\vec{x}_1, \vec{x}_2) = \frac{|\vec{x}_1 \cap \vec{x}_2|}{|\vec{x}_1 \cup \vec{x}_2|}$$

	water	watering	earth	air
water	1	$5/8 = 0.625$	$4/6 = 0.67$	$2/6 = 0.33$
watering		1	$4/9 = 0.44$	$3/8 = 0.375$
earth			1	$2/6 = 0.33$
air				1

• <u>prosječno povezivanje</u> $d(G_i, G_j) = \frac{1}{N_i \cdot N_j} \sum\sum d(\vec{x}^i, \vec{x}^j)$



(I) water, earth

(II) $\{w, e\}$ - wing $= \frac{77}{144} \approx 0.535$

$\{w, e\}$ - a $= 1/3 \approx 0.33$

wing - a $= 3/8 = 0.375$

(C) 0.535

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:

   $$\left\{\left(y^{(i)}, h(\mathbf{x}^{(i)})\right)\right\}_{i=1}^{11} = \left\{(1,1),(0,2),(2,2),(1,2),(1,1),(0,0),(1,1),(2,1),(0,1),(2,0),(2,1)\right\}.$$

   (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 oznake 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
$h_{\text{LR}}(\mathbf{x})$	0.8	0.6	0.8	0.6	0.8	0.8	0.8	0.2	0.2	0.2
$h_{\text{NB}}(\mathbf{x})$	0.3	0.8	0.3	0.5	0.8	0.3	0.8	0.5	0.3	0.5
$h_{\text{SVM}}(\mathbf{x})$	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 alarma).

   (a) Skicirajte krivulje ROC za ova tri klasifikatora, linearno interpolirajući između točaka točaka dobivenih na temelju gornjeg uzorka.

   (b) Izračunajte mjere 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 $C \in \{2^{-5}, 2^{-4}, \ldots, 2^{15}\}$ i $\gamma \in \{2^{-15}, 2^{-14}, \ldots, 2^{3}\}$.

   (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 petlji umjesto peterostruke unakrsne provjere 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):

$$\begin{array}{c} \\ y=1 \\ y=2 \\ y=3 \end{array} \begin{array}{c} \begin{array}{ccc} y=1 & y=2 & y=3 \end{array} \\ \begin{pmatrix} 15 & 3 & 1 \\ 6 & 5 & 4 \\ 4 & 2 & 23 \end{pmatrix} \end{array}$$

   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):

$$\begin{array}{c} \\ y=1 \\ y=2 \\ y=3 \end{array} \begin{array}{c} \begin{array}{ccc} y=1 & y=2 & y=3 \end{array} \\ \begin{pmatrix} 30 & 18 & 3 \\ 11 & 25 & 2 \\ 4 & 2 & 5 \end{pmatrix} \end{array}$$

   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 vrednujemo 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:

$$\left\{(y^{(i)}, h(\mathbf{x}^{(i)})\right\}_{i=1}^{10} = \left\{(1, 0.8), (0, 0.2), (0, 0.6), (0, 0.6), (1, 0.8), (0, 0.8), (1, 0.6), (1, 0.2), (0, 0.6), (1, 0.8)\right\}$$

   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:

$$h_1 : (0.4, 0.2), (0.6, 0.2), (0.9, 0.4) \qquad h_3 : (0.1, 0.5), (0.6, 0.6), (0.7, 0.8)$$
$$h_2 : (0.3, 0.1), (0.5, 0.4), (0.8, 0.6) \qquad h_4 : (0.3, 0.8), (0.4, 0.9), (0.6, 1.0)$$

Skicirajte odgovarajuće krivulje ROC, linearno interpolirajući između izmjerenih 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 popravljenu 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 $h_1 \succ h_2$ označava da naučeni model h_1 striktno dominira nad naučenim modelom h_2 prema krivulji ROC, tj. da vrijedi

$$\forall\theta.\Big(\big(\mathrm{FPR}_\theta(h_1) = \mathrm{FPR}_\theta(h_2)\big) \Rightarrow \big(\mathrm{TPR}_\theta(h_1) > \mathrm{TPR}_\theta(h_2)\big)\Big)$$

gdje su $\mathrm{FPR}_\theta(h)$ i $\mathrm{TPR}_\theta(h)$ vrijednosti stope lažnih pozitiva odnosno stvarnih pozitiva hipoteze h s pragom θ. Nadalje, neka $h_1 > h_2$ označava da vrijedi $\mathrm{AUC}(h_1) > \mathrm{AUC}(h_2)$. **Što od sljedećeg vrijedi?**

A $h_1 > h_2 \Rightarrow h_1 \succ h_2$

B $h_1 > h_2 \Rightarrow h_2 \succ h_1$

C $h_1 \succ h_2 \Rightarrow h_1 > h_2$

D $h_1 \succ h_2 \Rightarrow h_2 > h_1$

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 $C \in \{2^{-15}, 2^{-14}, \ldots, 2^{15}\}$ i $\gamma \in \{2^{-15}, 2^{-14}, \ldots, 2^{15}\}$. 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 Zadaci za učenje

1.1.

$N = 11$
$K = 3$

y^i	1	0	2	1	1	0	1	2	0	2	2
$h(\vec{x}^i)$	1	2	2	2	1	0	1	1	1	0	1

a) $Acc = ?$

$$Acc = \frac{1}{N}\sum_{i=1}^{N} 1\{y^i = h(\vec{x}^i)\} = \frac{1}{11}\cdot 5 = \frac{5}{11} \approx 0.455$$

b)

pred \ true	0	1	2
0	1	0	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	$y=1$	$y\neq 1$
$y=1$	3	3
$y\neq 1$	1	4

pred \ true	$y=2$	$y\neq 2$
$y=2$	1	2
$y\neq 2$	3	5

Mikro progjene

$\sum TP = 5 \qquad \sum FP = 6$

$\sum FN = 6 \qquad \sum TN = 16$

$$P^{\mu} = \frac{TP}{TP+FP} = \frac{5}{11}$$

$$R^{\mu} = \frac{TP}{TP+FN} = \frac{5}{11}$$

$$F^{\mu} = \frac{2PR}{P+R} = \frac{5}{11}$$

Makro progjene

$P_1 = \frac{1}{2} \qquad R_1 = \frac{1}{3} \qquad F_{1,1} = \frac{2}{5}$

$P_2 = \frac{1}{2} \qquad R_2 = \frac{3}{4} \qquad F_{1,2} = \frac{3}{5}$

$P_3 = \frac{1}{3} \qquad R_3 = \frac{1}{4} \qquad F_{1,3} = \frac{2}{7}$

$$P^M = \frac{1}{3}\sum P_i = \frac{4}{9} = 0.4\dot{4}$$

$$R^M = \frac{1}{3}\sum R_i = \frac{4}{9} = 0.4\dot{4}$$

$$F_1^M = \frac{1}{3}\sum F_{1,i} = \frac{3}{7} \approx 0.429$$

1.2.

$N = 1000$

$TP_1 = 620$
$TP_2 = 146$
$TP_3 = 134$

$FN^{pred,tr}_{2,1} = 50$
$FN_{2,3} = 20$
$FN_{3,2} = 30$

pred \ true	1	2	3
1	620	0	0
2	50	146	20
3	0	30	134

• dekomporicija

$$\begin{array}{c|cc} & y=1 & y\neq 1 \\ \hline y_p=1 & 620 & 0 \\ y_p\neq 1 & 50 & 330 \end{array} \qquad \begin{array}{c|cc} & y_t=2 & y_t\neq 2 \\ \hline y_p=2 & 146 & 70 \\ y_p\neq 2 & 30 & 754 \end{array} \qquad \begin{array}{c|cc} & y_{true}=3 & y_{true}\neq 3 \\ \hline y_{pred}=3 & 134 & 30 \\ y_{pred}\neq 3 & 20 & 816 \end{array}$$

$$P_1 = \frac{TP}{TP+FP} = 1 \qquad P_2 = \frac{146}{216} \qquad P_3 = \frac{134}{164}$$

$$R_1 = \frac{TP}{TP+FN} = \frac{620}{670} \qquad R_2 = \frac{146}{176} \qquad R_3 = \frac{134}{154}$$

$$F_{1,1} = \frac{2P_1R_1}{P_1+R_1} = \frac{124}{129} \qquad F_{1,2} = \frac{73}{98} \qquad F_{1,3} = \frac{134}{159}$$

$$F_1^M = \frac{1}{3}\sum_{i=1}^{3} F_{1,i} = 0.8496$$

1.3

$N = 10$
- 3 binarna klasifikatora

h_{LR}
h_{NB}
h_{SVM}

	1	2	3	4	5	6	7	8	9	10
y^i	1	1	0	0	1	1	1	0	0	1
$h_{LR}(\vec{x}^i)$	0.8	0.6	0.8	0.6	0.8	0.8	0.8	0.2	0.2	0.2
$h_{NB}(\vec{x}^i)$	0.3	0.8	0.3	0.5	0.8	0.3	0.8	0.5	0.3	0.5
$h_{SVM}(\vec{x}^i)$	0.6	0.1	0.7	0.6	0.1	0.7	0.7	0.6	0.1	0.7

a) $TPR = \frac{TP}{TP+FN}$ $\qquad FPR = \frac{FP}{FP+TN}$

• za različite pragove treba provesti klasifikaciju i izračunati pripadne TPR i FPR
=> radimo programski - python

• izračun ćemo provesti za pragove: 1, 0.8, 0.6, 0.4, 0.2, 0

$$\begin{array}{cc} & \text{true} \\ \text{pred} & \begin{array}{c|cc} & 1 & 0 \\ \hline 1 & TP & FP \\ 0 & FN & TN \end{array} \end{array}$$

Logistička regresija

i	1	2	3	4	5	6	7	8	9	10
$h_{LR}(\tilde{x}^i)$	0.8	0.6	0.8	0.6	0.8	0.8	0.8	0.2	0.2	0.2
y^i	1	1	0	0	1	1	1	0	0	1
pragovi: 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

Pragovi 0.6 i 0.4 te pragovi 0.2 i 0 rezultiraju istim klasifikacijama pa će vrijednosti TPR i FPR biti jednake za korespondentne pragove.

prag	$FPR = \frac{FP}{FP+TN}$	$TPR = \frac{TP}{TP+FN}$
1	$\frac{0}{0+4} = 0$	$\frac{0}{0+6} = 0$
0.8	$\frac{1}{1+3} = \frac{1}{4}$	$\frac{4}{4+2} = \frac{4}{6} = \frac{2}{3}$
0.6	$\frac{2}{2+2} = \frac{1}{2}$	$\frac{5}{5+1} = \frac{5}{6}$
0.2	$\frac{4}{4+0} = 1$	$\frac{6}{6+0} = 1$

Naivan Bayesov klasifikator

i	1	2	3	4	5	6	7	8	9	10
$h_{NB}(\tilde{x}^i)$	0.3	0.8	0.3	0.5	0.8	0.3	0.8	0.5	0.3	0.5
y^i	1	1	0	0	1	1	1	0	0	1
1	0	0	0	0	0	0	0	0	0	0
ista cif: 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
ista cif: 0.2	1	1	1	1	1	1	1	1	1	1
0	1	1	1	1	1	1	1	1	1	1

prag	$FPR = \frac{FP}{FP+TN}$	$TPR = \frac{TP}{TP+FN}$
1	$\frac{0}{0+4} = 0$	$\frac{0}{0+6} = 0$
0.8	$\frac{0}{0+4} = 0$	$\frac{3}{3+3} = 0.5$
0.4	$\frac{2}{2+2} = \frac{1}{2}$	$\frac{4}{4+2} = \frac{2}{3}$
0.2	$\frac{4}{4+0} = 1$	$\frac{6}{6+0} = 1$

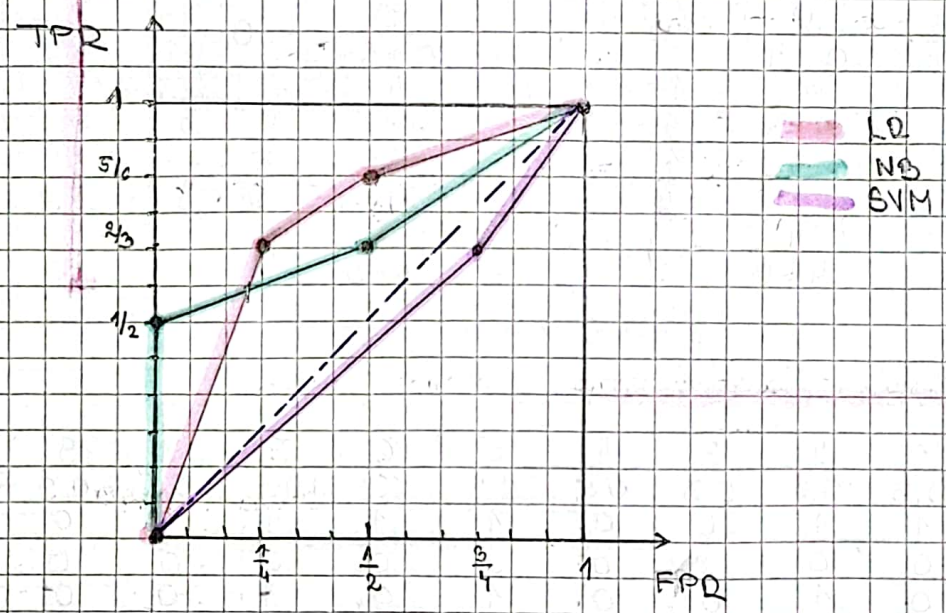
SVM

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	0.6	0.1	0.7
$y^{(i)}$	1	1	0	0	1	1	1	0	0	1
1	0	0	0	0	0	0	0	0	0	0
0.8	0	0	0	0	0	0	0	0	0	0
0.6	1	0	1	1	0	1	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

lista cef

lista cef

prag	$FPR = \frac{FP}{FP+TN}$	$TPR = \frac{TP}{TP+FN}$
1	$\frac{0}{0+4} = 0$	$\frac{0}{0+6} = 0$
0.6	$\frac{3}{3+1} = \frac{3}{4}$	$\frac{4}{4+2} = \frac{2}{3}$
0	$\frac{4}{4+0} = 1$	$\frac{6}{6+0} = 1$



b) AUC = računamo kao površinu ispod ROC krivulje
=> zbroj površina poligona

$$ROC_{LR} = \frac{2}{3}\cdot\frac{1}{4}\cdot\frac{1}{2} + \frac{1}{2}\cdot\frac{\frac{2}{3}+\frac{5}{6}}{2} + \frac{\frac{5}{6}+1}{2}\cdot\frac{1}{2} = \frac{35}{48} \approx 0.729$$

$$ROC_{NB} = \frac{1}{2}\cdot\frac{\frac{1}{2}+\frac{2}{3}}{2} + \frac{1}{2}\cdot\frac{\frac{2}{3}+1}{2} = \frac{17}{24} \approx 0.708$$

$$ROC_{SVM} = \frac{3}{4}\cdot\frac{2}{3}\cdot\frac{1}{2} + \frac{1}{4}\cdot\frac{\frac{2}{3}+1}{2} = \frac{11}{24} \approx 0.458$$

c) ROC krivulja nasumičnog klasifikatora odgovara pravcu koji spaja točke (0,0) do (1,1).

$\Rightarrow$ Nasumični klasifikator slučajno odabere $k\%$ primjera od svih poz. i proglašava ih pozitivnim, tj. $TPR = \frac{TP}{TP+FN} = k\%$.

Također nasumični klasifikator slučajno odabire $k\%$ primjera od ukupnog broja negativnih primjera i proglašava ih negativnim, tj. $FPR = \frac{FP}{FP+TN} = k\%$.

d) S obzirom na dobivene ROC krivulje bolji 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 naivan Bayesov klasifikator, a inače logistička regresija.

1.4.

$N = 1000$ primjera

ugniježđena unakrsna provjera 5×5

$C \in \{2^{-5}, 2^{-4}, \dots, 2^{15}\}$

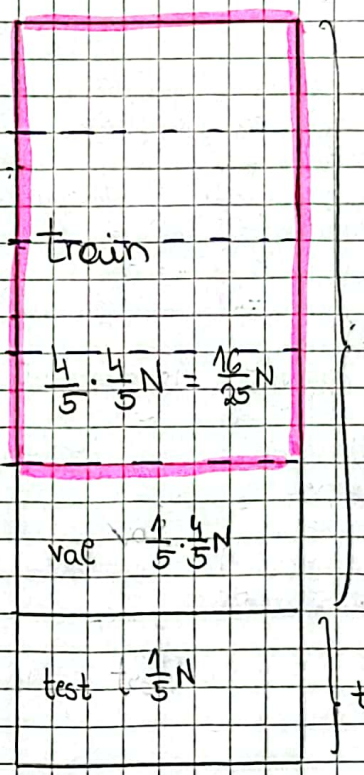
$|C| = 15 + 5 + 1 = 21$

$\gamma \in \{2^{-15}, 2^{-14}, \dots, 2^{3}\}$

$|\gamma| = 15 + 3 + 1 = 19$

Ukupno $21 \cdot 19 = 399$ kombinacija hiperparametara

a) Koliko puta ćemo ukupno trenirati model?



k = vanjski prekloپi

l = unutarnji prekloپi

$$\underbrace{k \cdot (\underbrace{l \cdot |C| \cdot |\gamma|}_{\text{optimizacija hiperparam.}} + \underbrace{1}_{\text{model za optim hiperparametre}})}_{\text{izmjena vanjskih prekloپa}}$$

Ukupni broj treniranja

$k(l \cdot |C| \cdot |\gamma| + 1)$

$= 5 \cdot (5 \cdot 21 \cdot 19 + 1)$

$= 9980$

situacija za $k = 1$

b) Koliko ćemo primjera koristiti u svakoj iteraciji koristiti za treniranje, koliko za provjeru, a koliko za ispitivanje?

$N = 1000$

$\frac{4}{5}N = 800$

train: $\frac{1}{5}N = \frac{1}{5} \cdot 1000 = 200$

train: $\frac{4}{5} \cdot \frac{4}{5}N = 640$

val: $\frac{1}{5} \cdot \frac{4}{5}N = \frac{1}{5} \cdot 800 = 160$

Za treniranje koristimo 640 primjera, za validaciju 160 primjera, a za ispitivanje 200 primjera

c) U vanjskoj petlji koristimo 1000V, tj. $k = N$

Ukupan broj treniranja: $k(e \cdot |C| \cdot |V| + 1)$
$= 1000 \cdot 5 \cdot 21 \cdot 19 + 1000$
$= 1.996\,000$

<u>U svakoj iteraciji</u>

$N(\text{test}) = 1$
$N(\text{val}) = \frac{1}{5} \cdot 999 = 200$
$N(\text{train}) = \frac{4}{5} \cdot 999 = 799$

d) <u>SVM</u>

- standardizacija značajki podrazumijeva skaliranje značajki na interval [0, 1] na sljedeći način $\vec{x}' = \frac{\vec{x} - \vec{\mu}_x}{\sigma_x}$

- SVM je osjetljiv na skale jer promatra euklidsku udaljenost između primjera pa značajke većih skala više utječu na udaljenost

- u postupku ugniježđene unakrsne provjere standardizacija bi se mogla ugraditi neposredno prije treniranja/validacije/testiranja modela (kao gjevovod)

e) U postupku ugniježđene unakrsne provjere bi postupak

- odabira značajki ugradio prije vanjske petlje, tj. prije podjele na vanjske preklope

- optimizacije praga po mjeri AUC ugradio u unutarnju petlju, tj. u optimizaciju hiperparametara

II Zadaci s ispita

2.1.

$K=3$

$$\begin{array}{c|ccc} & & \text{true} & \\ & 1 & 2 & 3 \\ \hline \text{pred}\ 1 & 15 & 3 & 1 \\ \text{pred}\ 2 & 6 & 5 & 4 \\ \text{pred}\ 3 & 4 & 2 & 23 \end{array}$$

$$\begin{array}{c|cc} & \text{true}\ i & \text{true}\ \neq i \\ \hline \text{pred}\ i & TP & FP \\ \text{pred}\ \neq i & FN & TN \end{array}$$

$F_1^{\mu} - F_1^{M} = ?$

$$F_1^{\mu} = Acc = \frac{1}{N}\sum_{i=1}^{N} 1\{y_{pred}^{i} = y_{true}^{i}\}$$

$$= \frac{1}{63}(15+5+23) = \frac{43}{63}$$

$$F_1 = \frac{2PR}{P+R}$$

Dekomponicija po klasama

$y=1$

$$\begin{bmatrix} 15 & 4 \\ 10 & 34 \end{bmatrix}$$

$P_1 = 15/19$

$R_1 = 15/25$

$F_1^1 = \frac{15}{22}$

$y=2$

$$\begin{bmatrix} 5 & 10 \\ 5 & 43 \end{bmatrix}$$

$P_2 = 5/15$

$R_2 = 5/10$

$F_1^2 = \frac{2}{5}$

$y=3$

$$\begin{bmatrix} 23 & 6 \\ 5 & 29 \end{bmatrix}$$

$P_3 = 23/29$

$R_3 = 23/28$

$F_1^3 = \frac{46}{57}$

$$P = \frac{TP}{TP+FP}$$

$$R = \frac{TP}{TP+FN}$$

$$F_1^{M} = \frac{1}{K}\sum_{k=1}^{K} F_1^{k} \approx 0.6286$$

$$F_1^{\mu} - F_1^{M} = 0.0529 \Rightarrow B$$

2.2

MLR

$K=3$

$$\begin{array}{c|ccc} & & \text{true} & \\ & 1 & 2 & 3 \\ \hline \text{pred}\ 1 & 30 & 18 & 3 \\ \text{pred}\ 2 & 11 & 25 & 2 \\ \text{pred}\ 3 & 4 & 2 & 5 \\ \hline & \downarrow & \downarrow & \downarrow \\ & 45 & 45 & 10 \\ \Rightarrow & 0.45 & 0.45 & 0.1 \end{array}$$

$$F_1^{\mu}(MLR) = Acc = \frac{1}{N}\sum_{i=1}^{N} 1\{y_{pred}^{i} = y_{true}^{i}\} = \frac{3}{5}$$

RAND

$$\begin{array}{c|ccc} & 1 & 2 & 3 \\ \hline 1 & 0.45\cdot 45 & 0.45\cdot 45 & 0.45\cdot 10 \\ 2 & 0.45\cdot 45 & 0.45\cdot 45 & 0.45\cdot 10 \\ 3 & 0.1\cdot 45 & 0.1\cdot 45 & 0.1\cdot 10 \end{array} = \begin{bmatrix} 20.25 & 20.25 & 4.5 \\ 20.25 & 20.25 & 4.5 \\ 4.5 & 4.5 & 1 \end{bmatrix}$$

$$F_1^{\mu}(RAND) = \frac{1}{100}(20.25 + 20.25 + 1) = \frac{83}{200}$$

$$F_1^{\mu}(MLR) - F_1^{\mu}(RAND) = \frac{37}{200} = 0.185 \Rightarrow C$$

2.3.

i	1	2	3	4	5	6	7	8	9	10
$h(\vec{x}^i)$	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

$$\begin{bmatrix} TP & FP \\ FN & TN \end{bmatrix}$$

$$TPR = \frac{TP}{TP+FN}$$

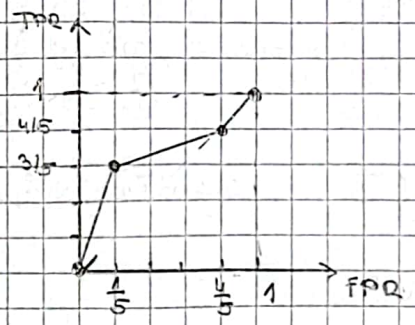
$$FPR = \frac{FP}{FP+TN}$$

$p = 0.8$ $\begin{bmatrix} 3 & 1 \\ 2 & 4 \end{bmatrix}$

$p = 0.6$ $\begin{bmatrix} 4 & 4 \\ 1 & 1 \end{bmatrix}$

$p = 0.2$ $\begin{bmatrix} 5 & 5 \\ 0 & 0 \end{bmatrix}$

prag	FPR	TPR
0.8	$\frac{1}{5}$	$\frac{3}{5}$
0.6	$\frac{4}{5}$	$\frac{4}{5}$
0.2	1	1



$AUC(RAND) = 0.5$

$$AUC(LR) = \frac{1}{2}\cdot\frac{1}{5}\cdot\frac{3}{5} + \frac{3}{5}\cdot\frac{\frac{3}{5}+\frac{4}{5}}{2} + \frac{1}{5}\cdot\frac{1+\frac{4}{5}}{2}$$

$$= \frac{33}{50} = 0.66$$

$AUC(LR) - AUC(RAND) = 0.16$

10

2.4.

(FPR, TPR)

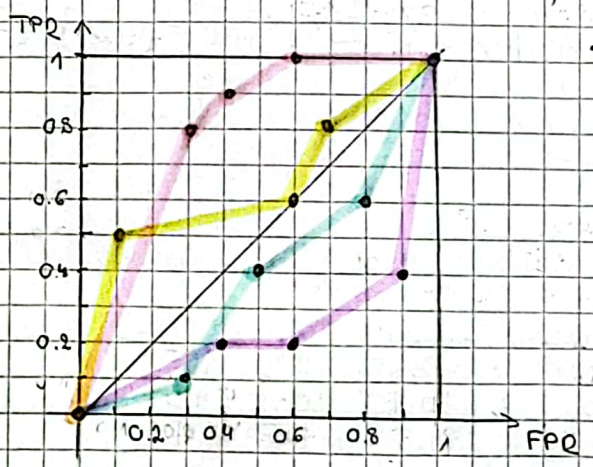
h_1: (0.4, 0.2), (0.6, 0.2), (0.9, 0.4)

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, 0.9), (0.6, 1)

• binarni klasifikatori $K=2$



• za h_1 i h_2 trebamo razmotriti klasif. $1-h_1$ i $1-h_2$

↳ komplementarni klasifikatori

$TP_h \rightarrow FN_{1-h}$

$FN_h \rightarrow TP_{1-h}$

$TN_h \rightarrow FP_{1-h}$

$FP_h \rightarrow TN_{1-h}$

$$TPR_h = \frac{TP_h}{TP_h+FN_h} = \frac{FN_{1-h}}{FN_{1-h}+TP_{1-h}} = 1 - TPR_{1-h}$$

$$FPR_h = \frac{FP_h}{TN_h+FP_h} = \frac{TN_{1-h}}{TN_{1-h}+FP_{1-h}} = 1 - FPR_{1-h}$$

$$\begin{bmatrix} TP & FP \\ FN & TN \end{bmatrix}$$

$FPR_{1-h} = 1 - FPR_h$

$TPR_{1-h} = 1 - TPR_h$

(FPR, TPR)

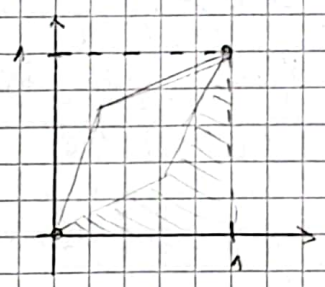
$1-h_1$: (0.6, 0.8), (0.4, 0.8), (0.1, 0.6)
$1-h_2$: (0.7, 0.8), (0.5, 0.6), (0.2, 0.4)
h_3 : (0.1, 0.5), (0.6, 0.6), (0.7, 0.8)
h_4 : (0.3, 0.8), (0.4, 0.9), (0.6, 1)

(B) h_1 i h_4

2.5.

$h_1 \succ h_2$
$\hookrightarrow \forall \theta \left((FPR_\theta(h_1) = FPR_\theta(h_2)) \Leftrightarrow (TPR_\theta(h_1) > TPR_\theta(h_2)) \right)$

$h_1 > h_2$
$\hookrightarrow AUC(h_1) > AUC(h_2)$



Ono što sigurno vrijedi jest $h_1 \succ h_2 \Leftrightarrow h_1 > h_2$, tj. ako 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.

$N = 1000$

Broj paramet. (kombo.)
$31^2 + 31 = 992$

3 hiperparametra
- jezgra - linearna ili RBF
- C - $|C| = 15 + 15 + 1 = 31$
- γ - $|\gamma| = 15 + 15 + 1 = 31$

Koliko će puta svaki primjer biti iskorišten za treniranje?

$k = 10$ $\quad$ $\ell = 5$ $\quad$ $\ell = 1$

train	$\frac{4}{5} \cdot \frac{9}{10} N = \frac{18}{25} N$
val	$\frac{1}{5} \cdot \frac{9}{10} N$
test	$\frac{1}{10} N$

$\frac{9}{10} N$

$(k-1) \cdot ((\ell - 1) \cdot 992 + 1)$
$= 9 \cdot (4 \cdot 992 + 1)$
$= 35\,721$

(A)

2.7.

$N = 1000$

SVM

$k = 5$

$\ell = 5$

train	$\frac{k-1}{k} \cdot \frac{\ell-1}{\ell} \cdot N$
val	$\frac{k-1}{k} \cdot \frac{1}{\ell} N$
test	$\frac{1}{k} \cdot N$

Svaki par modela nad kojima se optimiraju hiperparametri dijele većinu primjera. Razlikuju se samo u 2 preklopa - validacijski preklop je u drugom modelu dio skupa za treniranje. Drugim riječima svaki par modela dijeli $\frac{k-1}{k} \cdot \frac{\ell-2}{\ell} \cdot N$ primjera.

Ako je $k=5$, $\ell=5$ → par modela dijeli

$$\frac{4}{5} \cdot \frac{3}{5} \cdot N = 480 \text{ primjera}$$

Ako je $k=5$, $\ell=10$ → par modela dijeli

$$\frac{4}{5} \cdot \frac{8}{10} \cdot N = 640 \text{ primjera}$$

Razlika u broju primjera koji se dijele je onda $640 - 480 = 160$ primjera

C

2.8.

$k*\ell$ unakrsna provjera

- optimiramo reg. faktor λ

λ_1 = prosjek optimalnih vrijed. reg. faktora
F_1^1 = prosječna F_1 mjera na ispitnom skupu

↑ unutarnja petlja → validacija uvijek na 1. preklopu

λ_2 = prosjek optim. vrij. reg. faktora
F_1^2 = prosjek F_1 mjere na ispitnom skupu

optim. model treniran samo na Dtrain

λ_3 = prosjek optim. vrij. reg. faktora
F_1^3 = prosjek F_1 mjere na ispitnom skupu

$$\left.\begin{array}{l} \lambda_2 = \lambda_3 > \lambda_1 \\ F_1^3 > F_1^2 > F_1^1 \end{array}\right\}$$

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 $F_1^3 > F_1^1$ i $F_1^2 > F_1^1$
$F_1^3 > F_1^2$ zbog pogreške u tren. optimalnog modela