Teorijska pitanja

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

Jan Šnajder, v1.0

2 Osnovni koncepti

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

   A Različitih primjera iz $\mathcal{X} \times \mathcal{Y}$ potencijalno ima beskonačno mnogo, pa pogrešku računamo na uzorku $\mathcal{D}$ za koji pretpostavljamo da je reprezentativan

   B Ne možemo izračunati očekivanje gubitka jer nam nije poznata distribucija primjera iz $\mathcal{X} \times \mathcal{Y}$, no pretpostavljamo da je $\mathcal{D}$ reprezentativan uzorak iz te distribucije

   C Očekivanje gubitka ne možemo izračunati jer primjera iz $\mathcal{X} \times \mathcal{Y}$ ima potencijalno beskonačno, stoga pogrešku računamo na temelju skupa $\mathcal{D}$ za koji pretpostavljamo da je konačan

   D Funkciju gubitka jednostavnije je definirati nego funkciju pogreške, a aproksimacija je točna uz pretpostavku i.i.d.

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

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

   Što to zapravo znači?

   A Da model sadrži beskonačno mnogo funkcija h čija konkretna definicija ovisi o vrijednostima parametara $\boldsymbol{\theta}$

   B Da različite funkcije h imaju različite parametre $\boldsymbol{\theta}$, i da su sve one sadržane u modelu, to jest za sve njih vrijedi $h \in \mathcal{H}$

   C Da za različite parametre $\boldsymbol{\theta}$ dobivamo različite funkcije h, i da su sve one sadržane u modelu, to jest za sve njih vrijedi $h \in \mathcal{H}$

   D Da su funkcije h definirane sa slobodnim parametrima $\boldsymbol{\theta}$ i da broj različitih funkcija odgovara broju parametara

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

   A Parametre optimira algoritam strojnog učenja, dok optimizacija hiperparametara nije u nadležnosti tog algoritma

   B Hiperparametri određuju jačinu regularizacije, a parametri stupanj nelinearnosti hipoteze

   C Parametri određuju iznos empirijske pogreške na skupu za učenje, a hiperparametri iznos te pogreške na skupu za provjeru

   D Hiperparametri mogu biti diskretni ili kontinuirani, dok su parametri uvijek kontinuirani

4. (T) U strojnom učenju, model je skup funkcija $\mathcal{H}$ indeksiran parametrima $\boldsymbol{\theta}$. **Što to znači?**

   A Svaki $\boldsymbol{\theta}$ jednoznačno određuje funkciju koja primjer $\mathbf{x}$ preslikava u oznaku y u ovisnosti o parametrima $\boldsymbol{\theta}$

   B Svaki skup funkcija $\mathcal{H}$ ima svoj vektor parametara $\boldsymbol{\theta}$ i svaki vektor parametara $\boldsymbol{\theta}$ određuje skup funkcija $\mathcal{H}$

   C Svaki $\mathbf{x}$ određuje parametar $\boldsymbol{\theta}$ kojim se oznaka y preslikava u primjer $\mathbf{x}$

   D Svaka funkcija koja primjeru $\mathbf{x}$ dodjeljuje oznaku y jednoznačno određuje točku $\boldsymbol{\theta}$ u višedimenzijskome prostoru parametra

5. (T) Hipoteza h je funkcija koja primjerima iz $\mathcal{X}$ pridjeljuje oznake iz $\mathcal{Y}$. Za h kažemo da je definirana "do na parametre $\boldsymbol{\theta}$". **Što to znači?**

   A Funkcija h jednoznačno određuje parametre $\boldsymbol{\theta}$ iz skupa svih mogućih parametara, koji nazivamo prostor parametara

   B Svaka vrijednost parametara $\boldsymbol{\theta}$ daje jednu konkretnu funkciju h koja se razlikuje od svih drugih funkcija u modelu $\mathcal{H}$

   C Funkcija h definirana je bez parametara, i njih treba odrediti naknadno postupkom odabira modela $\mathcal{H}$

   D Različite vrijednosti za $\boldsymbol{\theta}$ mogu dati različite funkcije h, a skup svih takvih različitih funkcija definira model $\mathcal{H}$

6. (T) Svaki algoritam strojnog učenja ima neku induktivnu pristranost. Bez induktivne pristranosti, učenje na temelju podataka ne bi imalo smisla, odnosno algoritam bez induktivne pristranosti ne bi mogao ništa naučiti. **Zašto strojno učenje bez induktivne pristranosti nije moguće?**

   A Model bi bio prejednostavan te ne bi postojala hipoteza s empirijskom pogreškom nula

   B Primjeri ne bi nužno bili linearno odvojivi

   C Oznaka niti jednog neviđenog primjera ne bi bila jednoznačno određena

   D Prostor parametara bio bi neograničen, tj. postojalo bi beskonačno mnogo vektora parametara

7. (T) Modeli strojnog učenja općenito su različite složenosti. S porastom složenosti modela raste vjerojatnost da model bude prenaučen. Ta vjerojatnost raste s količinom šuma u podatcima. **Zašto šum u podatcima za učenje može dovesti do prenaučenosti klasifikacijskog modela?**

   A Zbog šuma granica između klasa izgleda nelinearnijom nego što ona to zapravo jest, pa primjeri blizu granice znatno više doprinose pogrešci učenja nego primjeri koji su udaljeni od granice

   B Efekt šuma je slučajan, pa će hipoteza koja se previše prilagodi šumu na skupu za učenje očekivano imati veliku pogrešku na ispitnom skupu gdje je šum drugačiji ili ga nema

   C Povećanjem količine šuma granica između klasa postaje sve nelinearnija, pa raste i složenost modela te dobivena hipoteza očekivano neće odgovarati granici između klasa na ispitnom skupu

   D Zbog šuma su oznake nekih primjera u skupu za učenje pogrešne, pa sve hipoteze iz modela imaju na tom skupu pogrešku koja je veća od nula, a još veća na ispitnom skupu

8. (T) Svaki model strojnog učenja ima neku induktivnu pristranost. **Što je induktivna pristranost?**

   A Kriterij koji, na temelju modela, jednoznačno određuje hipotezu sa minimalnom empirijskom pogreškom

   B Svaka pretpostavka koja jednoznačno određuje model na temelju hipoteze i skupa za učenje

   C Odstupanje procjene parametra na temelju podataka u odnosu na pravu vrijednost parametra u populaciji

   D Minimalan skup pretpostavki koje, uz skup za učenje, jednoznačno određuju klasifikaciju svakog primjera

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

   A Da, primjerice ako $\mathcal{X} = \mathbb{R}^n$ i $h(\mathbf{x};\boldsymbol{\theta}) = \boldsymbol{\theta}^{\mathrm{T}}\mathbf{x}$

   B Ne, jer je skup primjera $\mathcal{D}$ uvijek konačan, neovisno o dimenzionalnosti ulaznog prostora n

   C Da, primjerice ako je $\mathcal{X} = \{0,1\}^n$ i $h(\mathbf{x};\boldsymbol{\theta}) = \mathbf{1}\{\boldsymbol{\theta}^{\mathrm{T}}\mathbf{x} \geqslant 0\}$

   D Ne, jer za beskonačan skup $\mathcal{H}$ optimizacijski problem $\operatorname{argmax}_{h\in\mathcal{H}} E(h|\mathcal{D})$ nije definiran

3 Linearna regresija

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

   A Hipoteza h je linearna kombinacija težina $\mathbf{w}$ i značajki $\mathbf{x}$

   B Težine $\mathbf{w}$ maksimiziraju iznos $\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$

   C Težine $\mathbf{w}$ minimiziraju iznos $\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$

   D Hipoteza h je funkcija iz $\mathbb{R}^n$ u $\mathbb{R}$

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

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

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

   A Ako je rang matrice $\mathbf{X}$ jednak $N+1$, a složenost izračuna dominantno ovisi o N

   B Ako je rang matrice $\mathbf{X}^T\mathbf{X}$ jednak N, a složenost izračuna dominantno ovisi o n

   C Ako je rang matrice $\mathbf{X}$ jednak $n+1$, a složenost izračuna dominantno ovisi o n

   D Ako je matrica $\mathbf{X}^T\mathbf{X}$ kvadratna i punog ranga, a složenost izračuna dominantno ovisi o N

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

   A Algoritmi imaju različite pristranosti, i to različitu pristranost preferencije jer regularizirana regresija preferira jednostavnije hipoteze, a onda i različitu pristranost jezika jer je model neregularizirane regresije nadskup modela regularizirane regresije

   B Oba algoritma imaju isti model, definiran kao linearnu kombinaciju značajki i težina, pa dakle imaju istu pristranost jezika, ali se razlikuju u pristranosti preferencije jer imaju različito definiranu empirijsku pogrešku (osim ako je regularizacijski faktor jednak nuli)

   C Algoritmi se ne razlikuju po pristranosti preferencijom budući da koriste istu funkciju gubitka (kvadratni gubitak), međutim regularizirana regresija ima jaču induktivnu pristranost jezika od regularizirane regresije budući da prvi model uključuje drugi model

   D Za razliku od neregularizirane regresije, regularizirana regresija preferira jednostavnije hipoteze, međutim pristranosti su im identične jer su oba algoritma definirana kao linearna kombinacija značajki i težina te oba koriste identičan optimizacijski postupak (pseudoinverz matrice dizajna)

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

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

A $p(\mathbf{x}|y) = \mathcal{N}(f(\mathbf{x}), \sigma^2)$

B $p(y|\mathbf{x}) = \mathcal{N}(0, \sigma^2)$

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

D $p(y) = \mathcal{N}(0, \sigma^2)$

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

A Pretpostavka $P(\mathbf{y}|\mathbf{w}) = \prod_{i=1}^{N} P(y^{(i)}|\mathbf{w})$

B Minimizacija iznosa $\|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2$

C Odabir linearnog modela $h(\mathbf{w}; \mathbf{y}) = \mathbf{w}^{\mathrm{T}}\mathbf{x}$

D Maksimizacija iznosa $-\ln \mathcal{L}(\mathbf{w}|\mathbf{y})$

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

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

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

A Kada je broj primjera veći od broja značajki

B Kada je broj značajki manji od broja primjera i nema multikolinearnosti

C Kada je broj primjera jednak broju značajki plus jedan i nema multikolinearnosti

D Kada nema multikolinearnosti i matrica dizajna je dobro kondicionirana

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

A Primjeri $(\mathbf{x}, y)$ u skupu $\mathcal{D}$ nezavisno su uzorkovani iz zajedničke distribucije $P(\mathbf{x}, y)$

B Funkcija pogreške $E(\mathbf{w}|\mathcal{D})$ je neprekidna i unimodalna

C Matrica dizajna $\mathbf{X}$ nije singularna ili je regularizacijski faktor λ veći od 0

D Oznaka y primjera $(\mathbf{x}, y)$ je normalna varijabla sa srednjom vrijednošću $\mathbf{w}^{\mathrm{T}}\mathbf{x}$

4 Linearna regresija II

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

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

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

A Dodavanje vrijednosti λ na dijagonale Gramove matrice povećava njezin rang

B Dodavanje vrijednosti λ na dijagonale Gramove matrice povećava normu težina $\|\mathbf{w}\|$

C Minimizacija norme težina $\|\mathbf{w}\|$ čini Gramovu matricu kvadratnom i singularnom

D Minimizacija norme težina $\|\mathbf{w}\|$ povećava multikolinearnost Gramove matrice i smanjuje složenost modela

2. (T) Kao regularizacijski faktor kod modela linearne regresije tipično se koristi neka p-norma vektora težina, $\|\mathbf{w}\|_p$. **Na kojoj se činjenici temelji korištenje norme kao regularizacijskog izraza?**

   A Ako su težine hipoteze velike magnitude, model je prenaučen

   B Ako je model prenaučen, hipoteza će imati velike magnitude težina

   C Ako je model optimalne složenosti, hipoteza će imati male magnitude težina

   D Ako su težine hipoteze male magnitude, model je podnaučen

3. (T) L_1-regularizacija ili LASSO kao regularizacijski izraz koristi prvu normu vektora težina, $\|\mathbf{w}\|_1$. **Što je prednost a što nedostatak L1-regularizacije?**

   A Prednost je da L_1-regulariziranu pogrešku možemo minimizirati gradijentnim spustom, a nedostatak je da rezultira rijetkim modelima

   B Prednost je da izbacuje značajke iz modela, a nedostatak je da L_1-regularizirana pogreška nema minimizator u zatvorenoj formi

   C Prednost je da zadržava sve značajke u modelu, a nedostatak je da Gramova matrica može biti blizu singularne ako u podatcima postoji multikolinearnost

   D Prednost je da postoji rješenje u zatvorenoj formi (pseudoinverz), a nedostatak da izračun L_1-regulariziranog pseudoinverza ovisi o broju značajki ali i o broju primjera

4. (T) Optimizacijom regularizirane funkcije pogreške smanjuje se prenaučenost modela. **Kako je definirana L_2-regularizirana pogreška kod linearne regresije?**

   A Zbroj očekivanja funkcije gubitka unakrsne entropije i druge norme vektora težina

   B Zbroj prosjeka kvadratnog gubitka na svim primjerima i kvadrata druge norme vektora težina bez težine w_0

   C Zbroj neregularizirane pogreške i izraza propocionalnog s kvadratom norme vektora težina bez težine w_0

   D Zbroj funkcije gubitka po svim primjerima i neregularizirane pogreške bez težine w_0

5. (T) Optimizacijom regularizirane funkcije pogreške smanjuje se prenaučenost modela. **Kakve parametre modela nalazi optimizacija L_2-regularizirane pogreške?**

   A Parametre koji uz što veću magnitudu vrijednosti daju što manje očekivanje gubitka na skupu za ispitivanje

   B Parametre koji uz što manju magnitudu vrijednosti daju što manje očekivanje gubitka na skupu za učenje

   C Parametre koji uz što veću magnitudu vrijednosti daju što veće očekivanje gubitka na skupu za učenje

   D Parametre koji uz što manju magnitudu vrijednosti daju što veće očekivanje gubitka na skupu za ispitivanje

6. (T) Multikolinearnost značajki jedan je od problema koji može nastupiti kod primjene modela regresije na stvarnim podatcima. Efekt multikolinearnosti i savršene multikolinearnosti dobro je uočljiv kod optimizacijskoga postupka običnih najmanjih kvadrata (OLS) kada se on provodi izračunom pseudoinverza matrice dizajna. Neka je m broj značajki, Φ je matrica dizajna i $\mathbf{G} = \Phi^{\mathrm{T}}\Phi$ je Gramova matrica. **Koji je efekt savršene multikolinearnosti kod postupka OLS?**

   A $\mathrm{rang}(\Phi) < m + 1$, $\mathbf{G}$ nema puni rang i nema inverz, no ima pseudoinverz koji nije numerički stabilan

   B Φ nema puni rang, $\mathrm{rang}(\mathbf{G}) < m + 1$ i $\mathbf{G}$ nema pseudoinverz

   C Φ ima puni rang, $\mathrm{rang}(\mathbf{G}) > m$ i $\mathbf{G}$ ima inverz, ali s visokim kondicijskim brojem

   D $\mathrm{rang}(\Phi) = N$, no $\mathrm{rang}(\mathbf{G}) < N$, pa $\mathbf{G}$ ima pseudoinverz, ali nema numerički stabilan inverz

5 Linearni diskriminativni modeli

1. (T) Na predavanjima smo za klasifikaciju pokušali upotrijebiti algoritam regresije. Zaključili smo da to ne funkcionira, tj. da algoritam linearne regesije jednostavno nije klasifikacijski algoritam. **Koje bismo minimalne preinake trebale učiniti u algoritmu linearne regresije, a da dobijemo algoritam koji dobro funkcionira kao klasifikacijski algoritam?**

   A Promijeniti funkciju gubitka

   B Promijeniti model i funkciju gubitka

   C Promijeniti funkciju gubitka i optimizacijski postupak

   D Promijeniti model, funkciju gubitka i optimizacijski postupak

2. (T) Algoritam strojnog učenja idealno bi minimizirao gubitak 0-1. Međutim, funkciju gubitka 0-1 u praksi ne možemo koristiti za optimizaciju parametra modela. **Zašto gubitak 0-1 ne možemo koristiti za optimizaciju?**

   A Gradijent gubitka 0-1 svugdje je nula osim za $h(\mathbf{x}) = 0$, pa funkcija pogreške ima zaravni po kojima se gradijentni spust ne može spuštati

   B Gubitak 0-1 pored neispravno klasificiranih primjera kažnjava i ispravno klasificirane primjere, i to tim više što su oni dalje od granice između klasa

   C Funkcija gubitka 0-1 nije diferencijabilna, pa ne postoji rješenje u zatvorenoj formi i ne postoji gradijent

   D Funkcija gubitka 0-1 nije konveksna, pa ni funkcija pogreške nije konveksna već ima lokalne minimume te ne postoji minimizator u zatvorenoj formi

3. (T) Algoritam linearne regresije može se pokušati primijeniti na klasifikacijski problem, kao što smo pokušali na predavanjima, međutim to nije dobro funkcioniralo. Razmotrite tri komponente algoritma linearne regresije: model (M), funkciju gubitka (G) i optimizacijski postupak (O). Također, prisjetite se algoritma logističke regresije, koji je dobar klasfikacijski algoritam. Želimo preinačiti komponente algoritma linearne regresije tako da iz njega dobijemo nov algoritam koji klasifikaciju radi bolje od linearnog modela regresije, ali koji je drugačiji od logističke regresije. **Uz koje tri komponente bismo dobili takav algoritam?**

   A M: $h(\mathbf{x}) = \sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x})$; G: $L(y, h(\mathbf{x})) = -y \ln h(\mathbf{x}) - (1-y) \ln\big(1 - h(\mathbf{x})\big)$, O: $\mathbf{w}^* = \mathbf{X}^+\mathbf{y}$

   B M: $h(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x}$; G: $L(y, h(\mathbf{x})) = -y \ln h(\mathbf{x}) - (1-y) \ln\big(1 - h(\mathbf{x})\big)$ O: $\mathbf{w}^*$ izračunat gradijentnim spustom

   C M: $h(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x}$; G: $L(y, h(\mathbf{x})) = (y - h(\mathbf{x}))^2$, O: $\mathbf{w}^* = \mathbf{X}^+\mathbf{y}$

   D M: $h(\mathbf{x}) = \sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x})$; G: $L(y, h(\mathbf{x})) = (y - h(\mathbf{x}))^2$, O: $\mathbf{w}^*$ izračunat gradijentnim spustom

4. (T) Višeklasni problem može se riješiti binarnim klasifikatorom uz primjenu sheme OVO ili sheme OVR. Obje sheme imaju svoje prednosti i nedostatke. Pretpostavite da raspolažemo sa K klasa i da svaka klasa ima N/K primjera, gdje je N ukupan broj primjera u skupu za učenje. **Što su prednosti odnosno nedostatci OVO i OVR sheme u takvom slučaju?**

   A OVR treba K puta više klasifikatora nego OVO, ali su kod OVO pozitivne klase $K/2$ puta manje zastupljene nego OVR

   B OVO iziskuje $(K-1)/2$ puta više parametara nego OVR, ali svaki OVR klasifikator ima $K-1$ puta manje pozitivnih primjera nego negativnih

   C OVR iziskuje $K-1$ puta više klasifikatora od sheme OVO, ali kod OVO pozitivne klase imaju $K-1$ puta manje primjera nego kod OVR

   D OVO svaki klasifikator trenira s $K/2$ puta manje primjera nego OVR, ali pozitivne klase kod OVR imaju K puta manje primjera nego kod OVO

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

   A Svaki primjer nanosi gubitak, ali manji za točno klasificirane primjere nego za netočno klasificirane primjere

   B Gubitak za sve točno klasificirane primjere je nula, a za netočno klasificirane može biti manji od 1

   C Točno klasificirani primjer nanosi gubitak manji od 1 te gubitak opada što je primjer bliže granici

   D Gubitak netočno klasificiranih primjera raste linearno s udaljenošću od granice

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

   A Za ispravno klasificirane primjere gubitak zglobnice je manji od gubitka za neispravno klasificirane primjere

   B Gubitak zglobnice je nula za primjere koji su ispravno klasificirani i daleko od granice

   C Za neispravno klasificirane primjere gubitak zglobnice raste linearno s udaljenošću od hiperravnine

   D Gubitak zglobnice kažnjava sve primjere koji se nalaze unutar margine, čak i one koji su ispravno klasificirani

6 Logistička regresija

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

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

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

   A $P(y|\mathbf{x}) = (y - h(\mathbf{x}))\mathbf{x}$

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

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

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

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

   A Da postupak uvijek konvergira, neovisno o odabranoj stopi učenja η i početnim težinama $\mathbf{w}$

   B Da postupak uvijek konvergira, pod uvjetom da su primjeri linearno neodvojivi ili da regulariziramo sa $\lambda > 0$

   C Da postupak ne može zaglaviti u lokalnome minimumu, pod uvjetom da u skupu $\mathcal{D}$ nema multikolinearnosti

   D Da postupak konvergira brže, pod uvjetom da su primjeri linearno odvojivi i da stopa učenja nije η prevelika

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

   A Empirijska pogreška logističke regresije smanjuje se s porastom broja primjera

   B Ispravno klasificirani primjeri koji su vrlo udaljeni od granice nanose malen gubitak

   C S porastom norme vektora težina gubitak na točnim primjerima teži prema nuli

   D Netočno klasificirani primjeri nanose gubitak koji je proporcionalan normi vektora težina

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

   A Optimizacijski algoritam će konvergirati prema parametrima koji minimiziraju pogrešku na skupu za učenje, ali neće doseći minimimum

   B Gradijentni spust neće krivudati u prostoru parametara jer optimizacijski algoritam koristi informaciju o zakrivljenosti površine pogreške

   C Optimizacijski algoritam će konvergirati do minimuma, ali nema garancije da će to biti globalni minimum funkcije pogreške

   D Neovisno o inicijalizaciji, optimizacijski će algoritam pronaći parametre koji minimiziraju pogrešku na skupu za učenje

7 Logistička regresija II

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

   A Za razliku od Newtonovog postupka, gradijentni spust može se koristiti za L2-regulariziranu logističku regresiju, no ako je stopa učenja prevelika, postupak može divergirati, dok Newtonov postupak nema taj problem

   B Za razliku od Newtonovog postupka, gradijentni spust može se koristiti za "online" (pojedinačno) učenje, no može krivudati i zato sporije konvergirati od Newtonovog postupka

   C Newtonov postupak brže konvergira, ali se može koristiti samo za konveksnu funkciju pogreške, dok gradijentni spust nema tog ograničenja, ali može zaglaviti u lokalnom optimumu

   D Gradijentni spust znatno je računalno jednostavniji od Newtonovog postupka, no za razliku od Newtonovog postupka kod L2-regularizirane regresije ne konvergira ako primjeri nisu linearno odvojivi

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

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

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

   A Značajke moraju biti linearno zavisne

   B Funkcija pogreške mora biti konveksna

   C U podatcima ne smije biti multikolinearnosti

   D Broj primjera mora biti veći od broja značajki plus jedan

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

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

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

   A MLR ima više parametara od BLR-OVR, ali nije osjetljiva na neuravnoteženost broja primjera po klasama

   B MLR i BLR-OVR imaju manje parametara od BLR-OVO, no jedino za MLR vrijedi $\sum_k P(y = k|\mathbf{x}) = 1$

   C Za razliku od BLR-OVR i BLR-OVO, kod MLR ne postoje područja u ulaznom prostoru za koje klasifikacijska odluka nije određena

   D Za razliku od BLR-OVO i BLR-OVR, MLR koristi funkciju softmax, pa minimizacija L1-regularizirane pogreške ima rješenje u zatvorenoj formi

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

   A Optimizaciju parametara linearne regresije također možemo napraviti gradijentnim spustom po empirijskoj pogrešci, ali to ne radimo jer izračun pseudoinverza ima manju računalnu složenost

   B Zbog nelinearnosti logističke funkcije, kod logističke regresije izračun pseudoinverza matrice dizajna nije moguće napraviti u zatvorenoj formi

   C Maksimizacija log-izglednosti oznaka logističke regresije kao rješenje za parametre ne daje izraz u zatvorenoj formi koji sadržava pseudoinverz matrice dizajna

   D Optimizaciju možemo provesti izračunom pseudoinverza matrice dizajna, međutim, za razliku od gradijentnog spusta, taj postupak ne funkcionira kada je matrica dizajna singularna

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

   A Minimizirati $\sum_{i=1}^{N} \ln h(\mathbf{x}^{(i)}; \mathbf{w})$, gdje je $h(\mathbf{x}; \mathbf{w}) = f(\mathbf{w}^{\mathrm{T}}\mathbf{x})$

   B Minimizirati $-\sum_{i=1}^{N} \ln y^{(i)} h(\mathbf{x}^{(i)}; \mathbf{w})$, gdje je $h(\mathbf{x}; \mathbf{w}) = \mathbf{w}^{\mathrm{T}}\mathbf{x}$

   C Maksimizirati $\sum_{i=1}^{N} \ln p(y^{(i)}|\mathbf{x}^{(i)})$, gdje je $\mathbb{E}[p(y^{(i)}|\mathbf{x}^{(i)})] = f(\mathbf{w}^{\mathrm{T}}\mathbf{x})$

   D Maksimizirati $\prod_{i=1}^{N} \ln p(y^{(i)}|\mathbf{x}^{(i)})$, gdje je $\mathbb{E}[p(y^{(i)}|\mathbf{x}^{(i)})] = \mathbf{w}^{\mathrm{T}}\mathbf{x}$

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

   A NN i LR imaju istu funkciju pogreške, ali se samo LR može optimirati Newtonovim postupkom jer funkcija gubitka NN nije konveksna

   B Jezgreni stroj s Gaussovom jezgrenom funkcijom istovjetan je NN sa L_2-regulariziranom funkcijom pogreške

   C Model LR istovjetan je modelu dvoslojne NN sa sigmoidnim prijenosnim funkcijama i pogreškom unakrsne entropije

   D Model dvoslojne NN istovjetan je modelu LR s poopćenim linearnim modelima sa sigmoidnim funkcijama kao baznim funkcijama

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

   A Razlika je to veća što je stopa učenja η bliža jedinici

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

   C Razlika je to veća što je izlaz modela $h(\mathbf{x})$ bliži nuli

   D Razlika je to manja što je oznaka y bliže jedinici

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

   A Multinomijalna logistička regresija s aktivacijskom funkcijom softmax istovjetna je dvoslojnoj neuronskoj mreži sa više neurona u izlaznom sloju

   B Logistička regresija koja kao adaptivne bazne funkcije koristi logističku regresiju istovjetna je neuronskoj mreži sa sigmoidnom aktivacijskom funkcijom

   C Neuronska mreža optimirana algoritmom propagacije pogreške unazad istovjetna je logističkoj regresiji optimiranoj stohastičkim gradijentnim spustom

   D Logistička regresija s linearnim jezgrenim funkcijama istovjetna je neuronskoj mreži sa linearnom aktivacijskom funkcijom i kvadratnom funkcijom pogreške

8 Stroj potpornih vektora

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

   A Što je vektor $\mathbf{w}$ kraći, to je manja vrijednost $h(\mathbf{x})$, pa primjeri moraju biti što dalje da bi vrijedilo $h(\mathbf{x}) = \pm 1$, a to znači da je margina to šira

   B Što je vektor $\mathbf{w}$ kraći, to je manja udaljenost d primjera od hiperravnine, a to efektivno znači da je margina to šira jer je margina fiksna a udaljenosti d se smanjuju

   C Što je vektor $\mathbf{w}$ kraći, to je manja vrijednost $h(\mathbf{x})$, ali je težina w_0 konstantna, pa se udaljenosti između hiperravnine i primjera povećavaju, što znači da se margina širi

   D Što je vektor $\mathbf{w}$ kraći, to je veća udaljenost d primjera od hiperravnine, što znači da se potporni vektori udaljavaju od hiperravnine, a to znači da margina postaje šira

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

   A $\operatorname{argmin}_{\boldsymbol{\alpha}} \max_{\mathbf{w},w_0} L(\mathbf{w}, w_0, \boldsymbol{\alpha})$

   B $\operatorname{argmax}_{\mathbf{w},w_0} \min_{\boldsymbol{\alpha}} L(\mathbf{w}, w_0, \boldsymbol{\alpha})$

   C $\operatorname{argmin}_{\mathbf{w},w_0} \max_{\boldsymbol{\alpha}} L(\mathbf{w}, w_0, \boldsymbol{\alpha})$

   D $\operatorname{argmax}_{\boldsymbol{\alpha}} \min_{\mathbf{w},w_0} L(\mathbf{w}, w_0, \boldsymbol{\alpha})$

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

SVM (tvrda margina) razlikuje od induktivne pristranosti algoritma perceptrona?

A SVM ima pristranost preferencijom kojom maksimizira marginu, dok perceptron nema induktivnu pristranost preferencijom već samo pristranost jezika

B Imaju istu pristranost preferencijom, a to je da primjeri moraju biti linearno odvojivi, no SVM ima dodatnu pristranost ograničenjem u vidu optimizacijskih ograničenja

C Razlikuju se po pristranost preferencijom, jer perceptron ne maksimizira marginu, premda se može dogoditi da pronađe rješenje koje maksimizira marginu

D Imaju istu pristranost jezika, a pristranost preferencijom također će biti ista ako se oba optimiraju gradijentnim spustom s istim početnim težinama i istom stopom učenja

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

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

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

A Uvjet zadovoljava beskonačno mnogo hipoteza, međutim samo za jednu vrijedi $yh(\mathbf{x}) = 1$ za najbliže primjere, i to je hipoteza koju odabire SVM

B Uvjet zadovoljava konačan broj hipoteza koje su linearno odvojive, a SVM između njih odabire onu jednu koja minimizira kvadrat vektora težina

C Uvjet zadovoljava beskonačno mnogo hipoteza, a SVM odabire onu jednu koja minimizira kvadrat vektora težina te koja ispravno klasificira sve primjere, uz uvjet da $h(\mathbf{x})$ nije u intervalu $(-1, +1)$

D Uvjet zadovoljava konačan broj hipoteza koje su linearno odvojive, no one se razlikuju samo po faktoru koji množi težine $(\mathbf{w}, w_0)$, pa SVM odabire onu jednu za koju vrijedi $yh(\mathbf{x}) \geqslant 1$ za sve primjere

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

A Ako je linearna kombinacija značajki iz $\mathbf{x}$ s pozitivnim težinama veća ili jednaka linearnoj kombinaciji značajki iz $\mathbf{x}$ s negativnim težinama

B Ako je vektor $\mathbf{x}$ po skalarnom umnošku sličniji potpornim vektorima s pozitivnom oznakom nego potpornim vektorima s negativnom oznakom

C Ako je skalarni umnožak vektora $\mathbf{x}$ i vektora oznaka $\mathbf{y}$ veća od praga definiranog parametrom w_0

D Ako većina od ukupno α primjera iz skupa za učenje koji su po euklidskoj udaljenosti najbliži primjeru $\mathbf{x}$ ima pozitivnu oznaku

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

A Primjeri iz obje klase trebaju činiti konveksne skupove u ulaznom prostoru

B Najbliži primjeri iz suprotnih klasa moraju biti u vrhovima konveksnih ljusaka

C Barem jedna spojnica između primjera jedne klase treba biti unutar konveksne ljuske druge klase

D Konveksne ljuske dviju klasa ne smiju se preklapati (trebaju biti disjunktne)

9 Stroj potpornih vektora II

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

   A $2N$ ograničenja i $2n$ varijabli

   B N ograničenja i $2N + 1$ varijabli

   C N ograničenja i $n + 1$ varijabli

   D $2N$ ograničenja i $N + n + 1$ varijabli

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

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

   Što možemo zaključiti na temelju ovog uvjeta?

   A Da se primjeri koji nisu potporni vektori sigurno nalaze izvan margine

   B Da se potporni vektori nalaze na margini ili izvan nje, a na pravoj strani granice

   C Da se primjeri koji nisu potporni vektori nalaze na margini ili unutar margine

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

3. (T) Problem meke margine SVM-a s u primarnoj se formulaciji svodi na rješavanje sljedećeg optimizacijskog problema:

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

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

   A Što je veća vrijednost $\|\mathbf{w}\|^2$, to je margina uža i tim manje primjera ulazi u marginu, pa je tim manji zbroj od ξ_i

   B Što je veća vrijednost $\|\mathbf{w}\|^4$ to je model složeniji, no tim je veća nelinearnost granice i to je veći hiperparametar C

   C Što je manji zbroj od ξ_i, to više primjera može ući u marginu i tim je veća vrijednost $\|\mathbf{w}\|^2$ te je model manje složenosti

   D Što je veći zbroj od ξ_i, to više primjera ulazi u marginu i tim je manja vrijednost $\|\mathbf{w}\|^2$ te je model veće složenosti

4. (T) Optimizacijski problem algoritma SVM može se postaviti u formulaciji meke ili tvrde margine te u primarnoj ili dualnoj formulaciji. Ovisno o formulaciji, kvadratni program sadrži različit broj varijabli po kojima optimiramo (optimizacijske varijable). **Ako matrica dizajna ima više redaka nego stupaca, koja formulacija ima najmanje optimizacijskih varijabli?**

   A Primarni problem meke margine

   B Primarni problem tvrde margine

   C Dualni problem meke margine

   D Dualni problem tvrde margine

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

   A Kada se koristi RBF jezgra sa Mahalanobisovom udaljenošću, jer ta udaljenost uzima u obzir varijancu značajki

   B Kada se koristi linearna jezgra i značajke su centrirane oko nule, jer se onda ne računa skalarni produkt između značajki

   C Kada se koristi dualna formulacija SVM-a i algoritam SMO, jer se tada implicitno provodi L1-regularizacija

   D Kada se koristi Gaussova jezgrena funkcija, jer ta jezgra implicitno inducira beskonačnodimenzijski prostor značajki

10 Jezgrene metode

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

   A Ne, jer u dualnoj (neparametarskoj) formulaciji problema maksimalne margine nemamo vektor značajki

   B Ne, jer granica između klasa u prostoru značajki općenito ne mora biti hiperravnina, već može biti hiperpovršina

   C Da, ako nismo koristili Gaussovu jezgru ili neku složeniju jezgru koja koristi Gaussovu jezgru kao gradivni blok

   D Da, ako smo koristili linearnu jezgru, odnosno ako je ulazni prostor jednak prostoru značajki

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

   A Zato što takva jezgra odgovara skalarnom produktu u nekom prostoru značajki, a to je nužno da bismo mogli primijeniti jezgreni trik

   B Zato što takva jezgra inducira Hilbertov prostor, tj. prostor beskonačnodimenzijskih značajki, što nam daje potencijalno vrlo složene modele

   C Zato što takva jezgra omogućava da, umjesto da vektoriziramo primjere, klasifikaciju određujemo na temelju sličnost između primjera i prototipnih primjera

   D Zato što takva jezgra nužno daje nenegativne vrijednosti sličnosti između parova primjera, što je nužno kako gubitak ne bi bio negativan

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

   A Što je vrijednost γ manja, to je manja vrijednost $\kappa(\mathbf{x}_1, \mathbf{x}_2)$, i to je model manje nelinearan

   B Što je vrijednost γ veća, to je veća vrijednost $\kappa(\mathbf{x}_1, \mathbf{x}_2)$, i to je model manje nelinearan

   C Što je vrijednost γ manja, to je manja vrijednost $\kappa(\mathbf{x}_1, \mathbf{x}_2)$, i to je model više nelinearan

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

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

   A $\forall \mathbf{x}_1 \forall \mathbf{x}_2.\ \mathbf{x}_1^{\mathrm{T}} \mathbf{x}_2 = \exp(-\gamma \|\boldsymbol{\phi}(\mathbf{x}_1) - \boldsymbol{\phi}(\mathbf{x}_2)\|^2)$

   B $\forall \mathbf{x}_1 \forall \mathbf{x}_2.\ \boldsymbol{\phi}(\mathbf{x}_1)^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_2) = \exp(-\gamma \Delta^2)$ gdje $\Delta^2 = (\mathbf{x}_1 - \mathbf{x}_2)^{\mathrm{T}} (\mathbf{x}_1 - \mathbf{x}_2)$

   C $\forall \mathbf{x}_1 \forall \mathbf{x}_2.\ \boldsymbol{\phi}(\mathbf{x}_1^{\mathrm{T}} \mathbf{x}_2) = \exp(-\gamma \Delta^2)$, gdje $\Delta^2 = (\mathbf{x}_1 - \mathbf{x}_2)^{\mathrm{T}} (\mathbf{x}_1 - \mathbf{x}_2)$

   D $\forall \mathbf{x}_1 \forall \mathbf{x}_2.\ (\mathbf{x}_1 \neq \mathbf{x}_2) \Rightarrow \left(\exp(-\gamma \|\boldsymbol{\phi}(\mathbf{x}_1) - \boldsymbol{\phi}(\mathbf{x}_2)\|^2) = \mathbf{x}_1^{\mathrm{T}} \mathbf{x}_2 \right)$

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

   A Jezgrenim trikom implicitno se ostvaruje nelinearnost prostora značajki, što povećava kapacitet modela i povećava njegovu točnost

   B Umjesto vektorizacije primjera, dovoljno je defnirati nelinearnu funkciju preslikavanja iz ulaznog prostora u prostor značajki

   C Jezgrena funkcija mjeri sličnost između primjera, čime se primjeri efektivno preslikavaju u beskonačnodimenzijski prostor značajki

   D Jezgrena funkcija može biti mjera sličnost između nevektoriziranih primjera, što implicitno inducira vektorski prostor značajki

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

   A Zbog L1-regularizacije, mnoge težine modela bit će pritegnute na nulu

   B Dimenzija prostora značajki ne može biti veća od broja primjera

   C Broj parametara modela jednak je dimenziji prostora značajki

   D Prototipni primjeri odabiru se u okviru optimizacijskog postupka

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

   A Klasifikacija primjera definirana je na temelju umnoška jezgrene funkcije za taj primjer i sve druge primjere u skupu za učenje

   B Jezgrena funkcija između primjera u prostoru značajki jednaka je skalarnom produktu tih primjera u ulaznom prostoru

   C Broj dimenzija prostora značajki implicitno ovisi o broju klasa u ulaznom prostoru te može biti beskonačna

   D Vrijednost jezgrene funkcije nad parom vektora jednaka je skalarnom produktu tih vektora nakon preslikavanja u prostor značajki

11 Neparametarske metode

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

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

A Što je C manji, to je neparametarski model rjeđi, ali to nema utjecaja na rijetkost parametarskog modela jer on nema potporne vektore

B Što je C veći, to je neparametarski model manje rijedak, dok je parametarski to rjeđi jer λ pada

C Što je C manji, to je neparametarski model rjeđi, a također je to rjeđi i parametarski model jer λ raste

D Što je C veći, to je neparametarski model manje rijedak, dok parametarski model nije rijedak jer ima L_2-regularizaciju a ne L_1-regularizaciju

2. (T) Problem prokletstva dimenzionalnosti (engl. *course of dimensionality*) pojavljuje se kod algoritama koji rade u visokodimenzijskome vektorskom prostoru i manifestira se na različite načine. **Kako se problem prokletstva dimenzionalnosti u visokodimenzijskim prostorima manifestira kod algoritma k-NN?**

A Udaljenosti između primjera se smanjuju i model k-NN postaje sve složeniji

B Povećava se broj susjeda u okolini svakog primjera i model k-NN postaje sve jednostavniji

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

D Broj susjeda k nekog primjera se smanuje i gube se granice između klasa

3. (T) Algoritmi strojnog učenja mogu biti parametarski ili neparametarski. **Što je karakteristika neparametarskih algoritama strojnog učenja?**

A Pretpostavljaju vjerojatnosnu distribuciju podataka

B Broj parametara ovisi o broju primjera

C Eksplicitno modeliraju granicu između primjera

D Svaki primjer ima globalan utjecaj na izgled hipoteze

4. (T) Nalaženje najbližih susjeda kod algoritma k-NN predstavlja izazov zbog računalne složenosti. Algoritam stabla lopti (engl. *ball tree*) jedan je od pristupa za smanjenje računalne složenosti dohvaćanja susjeda u visokodimenzijskom vektorskom prostoru. **Na koji način funkcionira algoritam stabla lopti?**

A Koristi preslikavanje osjetljivo na lokalne promjene u vektoru kojim se bliske točke pohranjuju u iste pretince

B Koristi pretraživanje duž pravca u vektorskome prostoru u smjeru suprotnome od gradijenta funkcije pogreške

C Koristi brzo pretraživu binarnu indeksnu strukturu za particioniranje prostora primjera u preklapajuće regije

D Koristi jezgreni trik za izračun euklidske udaljenosti između točke upita i svih drugih točaka u skupu primjera

Rješenja

	1	2	3	4	5	6	7	8	9
2. Osnovni koncepti	B	B	A	A	D	C	B	D	A
3. Linearna regresija	C	C	B	C	B	A	B	D	
4. Linearna regresija II	A	B	B	C	B	A			
5. Linearni diskriminativni modeli	C	A	D	B	B	D			
6. Logistička regresija	B	B	B	C	D				
7. Logistička regresija II	B	C	A	B	C	C	D	B	B
8. Stroj potpornih vektora	A	D	C	C	B	D			
9. Stroj potpornih vektora II	D	D	A	B	A				
10. Jezgrene metode	C	A	D	B	D	D	D		
11. Neparametarske metode	C	C	B	C					

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:

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

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

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

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:

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

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 $E_1 = E_2 = E_3 > E_4$

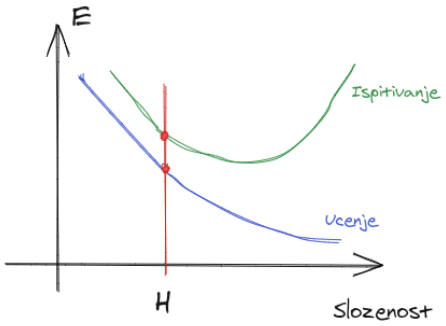
10. (P) Razmatramo klasifikacijski problem u ulaznome prostoru $\mathcal{X} = \mathbb{Z}^2$. Skup označenih primjera je $\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\} = \big\{((0,0),1), ((-1,-1),0), ((1,1),0)\big\}$. Razmatramo sljedeće modele $\mathcal{H}$ i funkcije preslikavanja $\boldsymbol{\phi} : \mathbb{Z}^2 \to \mathbb{Z}^2$, 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 $\mathcal{H}_{\alpha_2}$

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

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

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 $5\times$ manje, što je u redu jer to ionako nitko ne čita. Svaki novinski članak prikazujemo kao vektor riječi, gdje su komponente vektora broj pojavljivanja pojedine riječi. Problem rješavamo algoritmom 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 $2\times$ puta manje značajki od OVR, ali bi mogao raditi bolje na člancima iz kategorije "znanost"

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

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

   D OVR će imati $5\times$ manje značajki od OVO, ali bi mogao raditi bolje na člancima iz kategorije "znanost"

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

   A 2 B 3 C 4 D 5

4. (N) Raspolažemo sljedećim skupom za učenje u 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

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. Demonstrirati 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:

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

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

   A P+$\boldsymbol{\phi_0}$ 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 $\boldsymbol{\phi}(\mathbf{x}) = (1, x_1, x_2, x_1 x_2)$. U jednoj iteraciji treniranja modela vektor parametara jednak je $\mathbf{w} = (0.2, 0.5, -1.1, 2.7)$. **Koliko u toj iteraciji iznosi promjena težine w_1 za primjer $(\mathbf{x}, y) = ((-1,2),1)$?**

   A -12 B -5 C -2 D $+22$

7. Logistička regresija II

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

Jan Šnajder, vježbe, v1.11

1 Zadatci za učenje

1. [***Svrha: Znati izvesti algoritam multinomijalne logističke regresije.***]

   (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:

   $$\mathbf{w}^* = \underset{\mathbf{w}}{\text{argmax}} \ln P(\mathcal{D}|\mathbf{w}) = \underset{\mathbf{w}}{\text{argmin}} E(\mathbf{w}|\mathcal{D})$$

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:

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

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

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 $d(\text{LINR}) < d(\text{LR}) < d(\text{SVM})$

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

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:

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

   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

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 $\mathbb{R}^3 \times \{0,1\}$:

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

   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 $y = 1$

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.

14. Procjena parametara II

Strojno učenje 1, UNIZG FER, ak. god. 2022./2023.

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 $\hat{\boldsymbol{\theta}}_{\text{ML}}$.

   (b) Izvedite MLE-procjenitelj $\hat{\mu}_{\text{ML}}$ za parametar μ Bernoullijeve razdiobe $P(x|\mu)$.

   (c*) Izvedite MLE-procjenitelj $\hat{\mu}_{k,\text{MLE}}$ 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 $\hat{\mu}_{\text{ML}}$ 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 $\hat{\mu}_{\text{ML}}$ nepristran, a $\hat{\sigma}^2_{\text{ML}}$ 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:

$$\begin{aligned} \mathbf{x}^{(1)} &= (9.5, -0.7, -2.8) & \mathbf{x}^{(4)} &= (2.3, 0.3, 1.2) \\ \mathbf{x}^{(2)} &= (8.8, -0.8, -3.2) & \mathbf{x}^{(5)} &= (2.2, \quad 0, \quad 0) \\ \mathbf{x}^{(3)} &= (6.5, -0.2, -0.8) & \mathbf{x}^{(6)} &= (3.6, 0.3, 1.2) \end{aligned}$$

   (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 $\hat{\boldsymbol{\theta}}_{\text{MAP}}$ i objasnite zašto je on bolji od MLE-procjenitelja $\hat{\boldsymbol{\theta}}_{\text{ML}}$.

   (b) Objasnite što je to (1) konjugatna distribuicija 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 $\hat{\mu}_{\text{MAP}}$ i $\hat{\mu}_{\text{ML}}$ 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. (T) Funkcija izglednosti $\mathcal{L}(\boldsymbol{\theta}|\mathcal{D})$ nije isto što i vjerojatnost. **Po čemu se izglednost razlikuje od vjerojatnosti?**

   A Funkcija izglednosti $\mathcal{L}(\boldsymbol{\theta}|\mathcal{D})$ jednaka je gustoći vjerojatnosti $p(\mathcal{D}|\boldsymbol{\theta})$, samo što je izglednost funkcija parametara $\boldsymbol{\theta}$, dok je $p(\mathcal{D}|\boldsymbol{\theta})$ funkcija uzorka $\mathcal{D}$

   B Funkcija izglednosti $\mathcal{L}(\boldsymbol{\theta}|\mathcal{D})$ jednaka je gustoći vjerojatnosti $p(\boldsymbol{\theta}|\mathcal{D})$, ali, za razliku od gustoće vjerojatnosti, nije odozgo ograničena sa 1

   C Ako su podatci diskretni (kategoričke značajke), onda je funkcija izglednosti parametara $\boldsymbol{\theta}$ isto što i zajednička vjerojatnost uzorka $\mathcal{D}$ i parametara $\boldsymbol{\theta}$

   D Za razliku od vjerojatnosti, funkcija izglednosti $\mathcal{L}(\boldsymbol{\theta}|\mathcal{D})$ je simetrična, u smislu da vrijedi $\mathcal{L}(\boldsymbol{\theta}|\mathcal{D}) = p(\mathcal{D}|\boldsymbol{\theta})$

2. (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 $+0.48$

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 $\mathcal{L}_0 > \mathcal{L}_1$, $\mathcal{L}_2 \geqslant \mathcal{L}_3$

   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 $\mathcal{L}_0 \geqslant \mathcal{L}_1$, $\mathcal{L}_0 > \mathcal{L}_2 > \mathcal{L}_3$

4. (T) MAP-procjenitelj definiramo kao $\hat{\boldsymbol{\theta}}_{\text{MAP}} = \text{argmax}_{\boldsymbol{\theta}} p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$. Pri odabiru apriorne distribucije $p(\boldsymbol{\theta})$, nastojimo da je to neka standardna teorijska distribucija i da je konjugatna distribucija

za izglednost $\mathcal{L}(\boldsymbol{\theta}|\mathcal{D})$. **Što to znači i zašto to želimo?**

A To znači da će umnožak izglednosti i apriorne distribucije dati distribuciju koja je iste vrste kao i apriorna distribucija, a ako je riječ o standardnoj teorijskoj distribuciji iz eksponencijalne familije, njezin mod (maksimizator) postoji u zatvorenoj formi, što nam omogućava da procjenitelj izračunamo analitički

B To znači da je apriorna distribucija ista vrsta distribucije kao i vjerojatnost podataka uz dane parametre, tj. izglednost parametara, pa će njihov umnožak biti distribucija koja je proporcionalna aposteriornoj distribuciji i čiji ćemo maksimum moći izračunati Bayesovim pravilom

C To znači da je apriorna distribucija upravljana hiperparametrima kojima možemo ugoditi distribucija parametara koji procjenjujemo, tj. parametri apriorne distribucije i parametri izglednosti su identični, što nam omogućava da te dvije distribucije pomnožimo i zatim nađemo maksimizator

D To znači da je aposteriorna distribucija parametara ista kao izglednost parametara, pa primjenom Bayesovog pravila možemo izračunati apriornu vjerojatnost parametara te, nakon zanemarivanja nazivnika koji je za fiksiran skup podataka konstantan, pronaći parametre koji maksimiziraju aposterionu vjerojatnost

5. (T) Kod MAP-procjenitelja, apriorna distribucija parametra $p(\boldsymbol{\theta})$ tipično se odabire tako da bude konjugatna za funkciju izglednosti $p(\mathcal{D}|\boldsymbol{\theta})$. Pretpostavimo da MAP-procjenitelj izračunavamo heurističkom metodom (npr., gradijentnim usponom). Što se događa ako za apriornu distribuciju parametra upotrijebimo distribuciju koja *nije* konjugatna funkciji izglednosti?

   A Zajedničku distribuciju $p(\mathcal{D}, \boldsymbol{\theta})$ ne možemo izvesti u zatvorenoj formi, pa MAP nije definiran

   B Aposteriornu distribuciju $p(\boldsymbol{\theta}|\mathcal{D})$ ne možemo izvesti u zatvorenoj formi, ali MAP možemo izračunati heurističkom optimizacijom

   C Ako je apriorna distribucija $p(\boldsymbol{\theta})$ iz eksponencijalne familije, onda je aposteriona distribucija $p(\boldsymbol{\theta}|\mathcal{D})$ u zatvorenoj formi i MAP je izračunljiv

   D Neovisno o apriornoj distribuciji parametra $p(\boldsymbol{\theta})$, MAP je izračunljiv optimizacijom drugog reda (npr., Newtonovim postupkom)

6. (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

7. (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 $0 < \mu_1 < \frac{1}{3}$, $\frac{1}{3} < \mu_2 < 1$, $0 < \mu_3 < \mu_2 < 1$

8. (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 $\boldsymbol{\alpha} = (1, 1, 1, 1, 3, 1)$

15. Bayesov klasifikator

Strojno učenje 1, UNIZG FER, ak. god. 2022./2023.

Jan Šnajder, vježbe, v3.1

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

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

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

A Kategoričku distribuciju za $P(y)$ i Gaussovu distribuciju za $P(\mathbf{x}|y)$

B Bernoullijevu distribuciju za $P(y)$ i Gaussovu distribuciju za $P(\mathbf{x}|y)$

C Kategoričku distribuciju za $P(y)$ i za $P(\mathbf{x}|y)$

D Gaussovu distribuciju za $P(y)$ i multinulijevu distribuciju za $P(\mathbf{x}|y)$

2. (T) Bayesov klasifikator definiran je kao

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

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

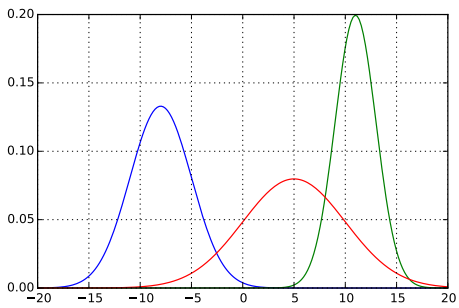
A Modelira vjerojatnost primjera i oznaka, budući da je, na temelju pravila umnoška, umnožak $p(\mathbf{x}|y)P(y)$ jednak zajedničkoj vjerojatnosti $p(\mathbf{x}, y)$

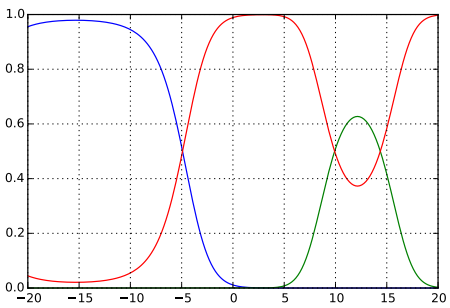
B Zajedničku vjerojatnost primjera i oznaka, $p(\mathbf{x}|y)P(y)$, faktorizira u dva faktora te zanemaruje nazivnik $p(\mathbf{x})$, koji je ionako konstantan za svaku klasu y

C Parametre distribucija $p(\mathbf{x}|y)$ i $P(y)$, a time indirektno i parametre aposteriorne distribucije $P(y|\mathbf{x})$, računa MAP-procjeniteljem, čime sprječava prenaučenost

D Primjer $\mathbf{x}$ klasificira prema MAP-hipotezi, dakle u klasu koja maksimizira aposteriornu vjerojatnost oznake, $p(y|\mathbf{x})$, koja je proporcionalna zajedničkoj vjerojatnosti primjera i oznaka, $p(\mathbf{x}, y)$

3. (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 $P(y=1)=P(y=2)=0.1, P(y=3)=0.8$

4. (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 $\mathcal{H}_3 > \mathcal{H}_1 > \mathcal{H}_2$, $\mathcal{H}_3 \supset \mathcal{H}_2 \supset \mathcal{H}_1$

5. (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:

$$p(\mathbf{x}|y=j) = \frac{1}{(2\pi)^{n/2}|\mathbf{\Sigma}_j|^{1/2}} \exp\Big\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_j)^{\mathrm{T}}\mathbf{\Sigma}_j^{-1}(\mathbf{x}-\boldsymbol{\mu}_j)\Big\}$$

   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

16. Bayesov klasifikator II

Strojno učenje 1, UNIZG FER, ak. god. 2022./2023.

Jan Šnajder, vježbe, v3.1

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 model 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 neke 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 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

5. (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 $P(y)P(x_1|y)P(x_2|y)P(x_3|y)$

17. Probabilistički grafički modeli

Strojno učenje 1, UNIZG FER, ak. god. 2022./2023.

Jan Šnajder, vježbe, v2.2

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 $s \in \{muški, ženski\}$, $p \in \{\bot, \top\}$, $t \in \{1, 3, 5\}$ i $r \in \{nizak, umjeren, visok\}$. 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. (T) Za Bayesovu mrežu kažemo da je generativni i parametarski model. **Zašto?**

   A Generativni jer definira zajedničku vjerojatnost svih varijabli, i opaženih i skrivenih, a parametarski jer se parametri modela mogu dobiti MLE-procjenom za svaki čvor Bayesove mreže zasebno, budući da se log-izglednost dekomponira po strukturi mreže

   B Generativni jer se može koristiti za generiranje skupa primjera na temelju zajedničke distribucije, a parametarski jer su broj čvorova mreže i njihovo povezivanje (dakle graf) definirani parametrima koji se mogu ugađati na skupu za učenje, čime se mogu dobiti različite strukture mreže

   C Generativni jer svaki čvor odgovara uvjetnoj vjerojatnosti koja je, na temelju Markovljevog uređajnog svojstva, generirana distribucijama čvorova roditelja, a parametarski jer Bayesova mreža zapravo definira zajedničku distribuciju koja je opisana skupom parametara

   D Generativni jer opisuje postupak kojim se mogu generirati podatci koji se pokoravaju određenoj zajedničkoj vjerojatnosnoj distribuciji, a parametarski jer svaki čvor Bayesove mreže definira uvjetnu vjerojatnost preko teorijske distribucije koja je opisana svojim parametrima

2. (T) Bayesove mreže na sažet način prikazuju zajedničku distribuciju te kodiraju uvjetne stohastičke nezavisnosti između varijabli. No, kao i svaki model strojnog učenja, tako se i Bayesove mreže mogu prenaučiti. **Koja je veza između uvjetnih nezavisnosti varijabli u Bayesovoj mreži i opasnosti od prenaučenosti?**

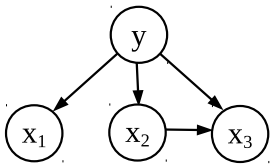
   A Uvođenje pretpostavki o uvjetnoj nezavisnosti pojednostavljuje strukturu Bayesove mreže i smanjuje broj parametara, čime se smanjuje i mogućnost prenaučenosti

   B Uvođenjem pretpostavki o uvjetnoj nezavisnosti povećava se broj čvorova mreže, a time i broj parametara, što model čini složenijim i time sklonijim prenaučenosti

   C Uvjetne nezavisnosti određuju stukturu mreže na način da definiraju koji su čvorovi mreže međusobno povezani, međutim to nema utjecaja na složenost modela niti na sklonost prenaučenosti

   D Pretpostavke o uvjetnoj nezavisnosti čine induktivnu pristranost modela, pa što je više uvjetnih nezavisnosti, to je veća pristranost i model je lako prenaučiti

3. (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

4. (P) Razmotrite Bayesovu mrežu koja zajedničku vjerojatnost faktorizira na sljedeći način:

$$P(w, x, y, z) = P(w)P(y)P(x|w, y)P(z|w)$$

Odredite topološki uređaj varijabli. Ako postoji više mogućih topoloških uređaja, izaberite onaj koji po leksičkom poretku dolazi prvi (npr. x, y, z dolazi prije x, z, y). Zatim primijenite uređajno Markovljevo svojstvo te izvedite sve uvjetne nezavisnosti koje su kodirane u ovoj Bayesovoj mreži. **Koje sve uvjetne nezavisnosti vrijede u ovoj Bayesovoj mreži?**

A $w \perp y,\ z \perp \{x, y\}|w$ B $x \perp y|z,\ z \perp w|y$ C $w \perp y,\ z \perp x,\ x \perp w|\{z, y\}$ D $y \perp w,\ y \perp x|\{w, z\}$

5. (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:

$$v \perp w \qquad w \perp x|v \qquad v \perp y|\{w, x\} \qquad \{v, w\} \perp z|\{x, y\}$$

Izvedite faktorizaciju zajedničke distribucije koja odgovara ovoj Bayesovoj mreži. **Koliko parametara ima dotična Bayesova mreža?**

A 10 B 22 C 25 D 27

6. (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 $P(T = 1|V = 1, P = 1) > P(T = 1|V = 1)$

18. Probabilistički grafički modeli II

Strojno učenje 1, UNIZG FER, ak. god. 2022./2023.

Jan Šnajder, vježbe, v2.2

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*

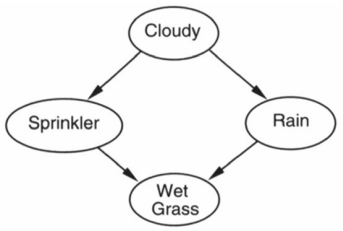
(b) Korištenjem egzaktnog zaključivanja izvedite izraz za vjerojatnost visokog rizika oboljenja osobe koja je pušač i posjećuje teretanu pet puta tjedno. Za svaku od četiri varijable naznačite radi li se o varijabli upita, opaženoj varijabli ili varijabli smetnje.

(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. (T) Procjena parametara Bayesove mreže temelji se na maksimizaciji log-izglednosti parametara pod modelom. Procjena parametara može biti bitno drugačija za slučaj potpunih podataka, gdje su sve varijable opažene, u odnosu na slučaj nepotpunih podataka, gdje u model trebamo uključiti skrivene ili latentne varijable. **Što je prednost procjene parametara kod potpunih podataka (modela bez skrivenih varijabli) u odnosu na nepotpune podatke (modela sa skrivenim varijablama)?**

A Kod potpunih podataka minimizacija funkcije log-izglednosti ima rješenje u zatvorenoj formi, ali funkcija nije konkavna, pa može imati više lokalnih optimuma, za razliku od modela sa skrivenim varijablama koji ima više parametara, ali konkavnu funkciju log-izglednosti

B Kod potpunih podataka maksimizacija log-izglednosti ima rješenje u zatvorenoj formi, ali samo ako su opažene varijable na početku niza po topološkom uređaju čvorova, za razliku od modela sa skrivenim varijablama kod kojega MLE procjenitelj ne postoji u zatvorenoj formi

C Kod potpunih podataka log-izglednost se dekomponira po strukturi mreže, pa parametre svake uvjetne distribucije možemo procijeniti nezavisno od drugih čvorova i u zatvorenoj formi, međutim parametara može biti više nego kod modela sa skrivenim varijablama

D Kod potpunih podataka MLE procjena parametara ima rješenje u zatvorenoj formi, dok MAP procjena nema, za razliku od modela sa skrivenim varijablama kod kojeg je situacija obrnuta, a k tome taj model ima još više parametara od modela bez skrivenih varijabli

19. Grupiranje

Strojno učenje 1, UNIZG FER, ak. god. 2022./2023.

Jan Šnajder, vježbe, v2.1

1 Zadatci za učenje

1. [***Svrha: Razumjeti rad algoritma k-sredina u smislu minimizcije 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 različitu grupu. 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. (T) Konvergencija je poželjno svojstvo algoritma grupiranja. **Je li točno da algoritam k-sredina uvijek konvergira?**

   A Da, algoritam uvijek konvergira zato što je broj particija N primjera u K skupova ograničen, a optimizacijski postupak definiran je tako da se J u svakoj iteraciji smanjuje

   B Algoritam konvergira samo ako su početna središta dobro odabrana, inače se može dogoditi da algoritam oscilira između dva rješenja

   C Kako se radi o algoritmu koji grupira primjere u vektorskom prostoru, broj rješenja je neograničen, stoga algoritam ne mora konvergirati

   D Algoritam uvijek konvergira zato što je broj primjera N uvijek veći ili jednak broju grupa K, a kao mjera udaljenosti koristi se euklidska udaljenost, koja je nužno nenegativna

2. (T) Algoritmi grupiranja k-sredina i k-medoida razlikuju se, između ostaloga, i po vremenskoj računalnoj složenosti. Naime, algoritam k-medoida računalno je složeniji od algoritma k-sredina. **Zašto je algoritam k-medoida računalno složeniji od algoritma k-sredina?**

   A Za razliku od algoritma k-sredina, algoritam k-medoida je algoritam mekog grupiranja, što iziskuje provođenje dodatnih koraka unutar algoritma

   B Budući da algoritam k-medoida ne koristi centroide, nego medoide, na kraju svake iteracije mora kombinatoričkom provjerom po primjerima pronaći medoide koje minimiziraju kriterijsku funkciju J

   C Za razliku od algoritma k-sredina koji se zasniva na euklidskoj udaljenosti, čiji je izračun računalno nezahtjevan, algoritam k-medoida koristi funkcije sličnosti čije računanje iziskuje mnogo računalnih operacija

   D Kriterijska funkcija algoritma k-medoida jest mnogo složenija od one k-sredina, upravo zato što algoritam k-medoide koristi medoide, a ne centroide

3. (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

4. (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

5. (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 $\mathbf{x}^{(5)}$

6. (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

7. (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

20. Grupiranje II

Strojno učenje 1, UNIZG FER, ak. god. 2022./2023.

Jan Šnajder, vježbe, v2.2

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 algomerativnog 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. (T) Algoritam GMM, odnosno model Gaussove mješavine s algoritmom maksimizacije očekivanja kao optimizacijskim postupkom, poopćenje je algoritma K-sredina. **Uz koje uvjete algoritam GMM degenerira u algoritam K-sredina?**

   A Umjesto maksimizacije log-izglednosti, minimizira se negativna log-izglednost, a početna središta se odabiru algoritmom K-sredina

   B Koeficijenti mješavine su jednaki za sve komponente Gaussove mješavine, a kovarijacijske matrice su dijagonalne

   C Kovarijacijska matrica komponenti Gaussove mješavine je jedinična matrica, a maksimizira se negativna log-izglednost

   D Kovarijacijska matrica komponenti Gaussove mješavine je dijeljena i izotropna, a odgovornosti su zaokružene na cijeli broj

2. (T) Algoritam maksimizacije očekivanja (EM-algoritam) maksimizira očekivanje potpune log-izglednosti, što se pokazuje da dovodi i do maksimizacije nepotpune log-izglednosti. **Koja je razlika između potpune i nepotpune log-izglednosti, i zašto maksimiziramo očekivanje potpune log-izglednosti umjesto izravno log-izglednost?**

   A Potpuna log-izglednost je izglednost izračunata na svim primjerima iz neoznačenog skupa primjera, dok je nepotpuna log-izglednost izračunata samo za označene primjere koji se koriste za evaluaciju modela, a očekivanje računamo zato jer je postupak grupiranja stohastičan

   B Potpuna log-izglednost je log-izglednost s neopaženim varijablama, a u slučaju GMM-a to su centroidi i kovarijacijske matrice komponenata, koje procjenjujemo metodom MLE, koja maksimizira očekivanje log-izglednosti

   C Potpuna log-izglednost je log-izglednost modela GMM s latentnim varijablama, koje definiraju koji primjer pripada kojoj grupi, međutim kako to zapravo ne znamo, moramo računati s očekivanjem tih varijabli

   D Potpuna log-izglednost računa se za označene primjere a nepotpuna log-izglednost za neoznačene primjere, a u oba slučaja kod modela GMM računamo očekivanje log-izglednosti jer postupak za različite početne centroide može dati različite log-izglednosti

3. (T) Za procjenu parametara modela GMM tipično se koristi algoritam maksimizacije očekivanja (EM-algoritam). To je iterativan optimizacijski algoritam. **Pod kojim uvjetima EM-algoritam**

(primijenjen na model GMM) konvergira, i kamo?

A Algoritam uvijek konvergira, međutim globalni maksimum log-izglednosti parametara doseže samo ako je broj grupa postavljen na pravi broj grupa ili tako da je broj grupa jednak broju primjera

B Algoritm uvijek konvergira, i to do točke u prostoru parametara koja maksimizira log-izglednost parametara, no brzina konvergencije ovisi o toma kako su inicijalizirani parametri

C Krenuvši od nekih početnih parametara, algoritam uvijek konvergira do parametara koji maksimiziraju očekivanje log-izglednosti, međutim to ne moraju biti parametri koji maksimiziraju vjerojatnost podataka

D Algoritam konvergira samo ako su primjeri u ulaznom prostoru sferični, ako su zavisnosti između značajki linearne, i ako nema multikolinearnosti, jer u protivnom zavisnosti nije moguće modelirati kovarijacijskom matricom

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

A Za razliku od Gaussovog Bayesovog klasifikator, GMM je nenadizirani algoritam, pa log-izglednost podataka nije definirana i nije moguća maksimizacija u zatvorenoj formi

B Kod GMM-a, pored koeficijenata mješavine i vektora sredina, trebamo procijeniti i kovarijacijske matrice, za što ne postoji procjenitelj u zatvorenoj formi

C Kod GMM-a ne znamo koji primjer pripada kojoj grupi, pa je gustoća primjera jednaka zbroju gustoći komponenti, za što ne postoji maksimizator u zatvorenoj formi

D Parametri oba modela mogu se optimirati u zatvorenoj formi, međutim kod modela GMM računalno je jednostavnije koristiti EM-algoritam

5. (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:

$$\begin{aligned} \mathcal{H}_1 &: \quad K = 2 \text{ grupa, puna kovarijacijska matrica} \\ \mathcal{H}_2 &: \quad K = 2 \text{ grupa, izotropna kovarijacijska matrica} \\ \mathcal{H}_3 &: \quad K = 3 \text{ grupe, izotropna kovarijacijska matrica} \end{aligned}$$

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 $h_{\alpha_2} < h_{\alpha_1} < h_{\alpha_3}$

6. (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 $LL_2^0 \leqslant LL_4^0$, $LL_2^* \leqslant LL_1^* \geqslant LL_3^*$

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

A Model s optimalnim brojem grupa je onaj koji podatke čini najvjerojatnijima, ali to čini sa što manje parametara

B Optimalan broj grupa je onaj kod kojeg, nakon daljnjeg povećanja broja grupa, vrijednost log-izglednosti stagnira ili blago raste

C Model s optimalnim brojem grupa je onaj koji minimizira log-izglednost nepotpunih podataka, a maksimizira log-izglednost potpunih podataka

D Optimalan broj grupa je onaj koji maksimizira očekivanje log-izglednost modela, uz pretpostavku izotropne kovarijacijske matrice

8. (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 $((\mathbf{x}^{(2)}, \mathbf{x}^{(3)}), ((\mathbf{x}^{(4)}, \mathbf{x}^{(1)}), \mathbf{x}^{(5)})))$

21. Vrednovanje modela

Strojno učenje 1, UNIZG FER, ak. god. 2022./2023.

Jan Šnajder, vježbe, v1.4

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:

$$\left\{(y^{(i)}, h(\mathbf{x}^{(i)})\right\}_{i=1}^{10} = \big\{(1, 0.8), (0, 0.2), (0, 0.6), (0, 0.6), (1, 0.8), (0, 0.8), (1, 0.6), (1, 0.2), (0, 0.6), (1, 0.8)\big\}$$

   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. (T) Procjena pogreške modela metodom unakrsne provjere omogućava nam da procijenimo prediktivnu moć modela, mjerenu kao točnost modela na neviđenom skupu primjera. Daljnja razrada te ideje je ugniježđena k-struka unakrsna provjera, koja se u praksi vrlo često koristi. **Koja je motivacija za korištenje ugniježđene k-struke unakrsne provjere, umjesto obične unakrsne provjere?**

   A Razdvaja skup za učenje od skupa za ispitivanje te time osigurava da doista mjerimo prediktivnu moć modela, odnosno ispitnu pogrešku, a ne pogrešku učenja

   B Omogućava nam da odredimo točnost modela s klasifikacijskim pragom, na način da u obzir uzimamo preciznost i odziv za različite vrijednosti klasifikacijskog praga

   C Provodi optimizaciju hiperparametra modela na uniji skupa za provjeru i skupa za testiranje, čime postiže bolju točnost modela jer više primjera ostaje za treniranje

   D Omogućava nam da procijenimo prediktivnu moć modela optimalne složenosti te maksimalno iskoristimo raspoložive podatke za učenje i ispitivanje

5. (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

6. (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 $\lambda_1 = \lambda_3 < \lambda_2$, $F_1^2 < F_1^1$, $F_1^3 < F_1^2$