Međuispit iz Strojnog učenja (ak. god. 2020./2021.)

– NEKORIGIRANA VERZIJA –

*Ispit se sastoji od **24 pitanja** i ukupno nosi **24 bodova** (koji se skaliraju na 35% bodova na kolegiju). Sva pitanja nose po 1 bod (1/3 boda oduzima se za pogrešan odgovor). Trajanje ispita je **135 minuta**. Primjerak ispita morate predati zajedno sa svojim rješenjima.*

Cjelina 1: Osnove i regresija

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

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

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

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

2 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 polinom drugog stupnja s kvadratnim, interakcijskim i linearnim značajkama. Od interakcijskih značajki uzimamo samo interakcije parova i interakcije trojki svih značajki. **Koliko minimalno primjera za učenje trebamo imati, a da bi rješenje bilo stabilno i bez regularizacije?**

A 38 B 75 C 48 D 63

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

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

Koliko iznosi $|\mathcal{H}|$?

A 28 B 56 C 42 D ∞

4 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 naučenih iz ovih modela 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 $\mathcal{H}_{5,100}$ će bolje generalizirati od hipoteze iz $\mathcal{H}_{2,0}$ i imat će manju pogrešku na skupu za učenje

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 $\mathcal{H}_{2,0}$ imati će veću pogrešku na skupu za učenje od hipoteze $\mathcal{H}_{5,0}$, ali mogu podjednako loše generalizirati

5 Zadan je sljedeći skup označenih primjera iz $\mathbb{R}^3$:

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

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 iznose najveća donja i najmanja gornja ograda tako definirane empirijske pogreške $E(h|\mathcal{D})$?**

A $0.5 \leq E(h|\mathcal{D}) \leq 4.5$ B $0 \leq E(h|\mathcal{D}) \leq 1.5$ C $\frac{1}{8} \leq E(h|\mathcal{D}) \leq 4$ D $1.5 \leq E(h|\mathcal{D}) \leq 4$

6 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 Optimalan model je onaj s hiperparametrom iz intervala $[\alpha_1, \alpha_2]$

B Model $\mathcal{H}_{\alpha_2}$ je prenaučen

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

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

7 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 0.29 C 1.58 D 2.69

8 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:

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

S tako pogrešno definiranom funkcijom gubitka postupkom najmanjih kvadrata treniramo naš model 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 $(\frac{1}{2}, -1)$ B $(2, -4)$ C $(2, -\frac{1}{2})$ D $(1, -2)$

Cjelina 2: Linearni klasifikacijski modeli

9 Na skupu označenih primjera $\mathcal{D}$ trenirali smo model logističke regresije. Dobili smo neki vektor težina $\mathbf{w}$ i pomak $w_0 = 3.15$. Tako naučenom modelu neki primjer $\mathbf{x}$, čija je oznaka u skupu primjera $y = 0$, nanosi gubitak unakrsne entropije od $L(0, h(\mathbf{x})) = 0.5$. **Koliki gubitak unakrsne entropije bi nanosio primjer x kada bismo njegove značajke pomnožili sa dva i promijenili mu oznaku**?

A 2.54 B 4.03 C 7.11 D 1.19

10 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 $\approx$ 1:405 B OVO:OVR $\approx$ 16:25 C OVO:OVR $\approx$ 32:50 D OVO:OVR $\approx$ 4:5

11 Na skupu označenih primjera treniramo tri modela: (1) model neregularizirane logističke regresije (NR), (2) model L2-regularizirane logističke regresije (L2R) i (3) perceptron s funkcijom preslikavanja. Za sva tri algoritma promatramo iznos empirijske pogreške učenja kroz iteracije optimizacijskog postupka. Za algoritam perceptrona opažamo da empirijska pogreška učenja ne konvergira. **Kako se u ovom slučaju ponaša empirijska pogreška učenja kroz iteracije za dva spomenuta modela logističke regresije, NR i L2R?**

A Pogreške učenja modela NR i modela L2R obje stagniraju nakon određenog broja iteracija, ali pogreška modela NR doseže manju vrijednost

B Pogreška učenja modela NR asimptotski teži nuli, dok pogreška učenja modela L2R nakon određenog broja iteracija stagnira

C Pogreška učenja modela NR nakon određenog broja iteracije doseže nulu, dok pogreška učenja modela L2R najprije pada pa raste

D Pogreške učenja modela NR i modela L2R dosežu nulu, ali modelu L2R za to treba više iteracija

12 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 "online" (pojedinačno) učenje, no može krivudati i zato sporije konvergirati od Newtonovog postupka

B 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

C 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

D 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

13 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 $K = 2$ klase u 100-dimenzijskome ulaznom prostoru. Za taj klasifikacijski problem koristimo binarnu logističku regresiju, ali s adaptivnim baznim funkcijama. Svaka adaptivna bazna funkcija ϕ_j je i sama jedan model logističke regresije, kao što smo radili na predavanjima. Naš je model onda definiran ovako:

$$h(\mathbf{x}) = \sigma\big(\sum_{j=0}^{5} w_j \phi_j(\mathbf{x})\big)$$

Koliko ukupno parametara ima ovaj model?

A 256 B 506 C 511 D 261

14 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 $\mathbf{x} = (3, -4, 1, -3)$ s oznakom $\mathbf{y} = (0, 1, 0)$. **Koliko iznosi gubitak unakrsne entropije koji u ovoj iteraciji optimizacijskog postupka nanosi dotični primjer?**

A 12.02 B 0.02 C 4.02 D 6.00

15 Model logističke regresije treniramo stohastčkim 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 4.00 B 2.48 C 0.70 D 1.28

16 Treniramo linearni diskriminativni model u dvodimenzijskome prostoru primjera. 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 treniramo model koji ima induktivnu pristranost takvu da rješenje maksimizira 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 1 C 5 D -5

Cjelina 3: Jezgrene i neparametarske metode

17 Rješavamo binarni klasifikacijski problem. Raspolažemo označenim skupom primjera. Odgovarajuća matrica dizajna je sljedeća:

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 22 & 0 & -5 \\ 1 & 6 & 28 & 8 & 2 \\ 1 & -6 & -15 & -9 & -6 \\ 1 & -11 & -34 & -16 & -9 \end{pmatrix}$$

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):

$$\mathbf{w} = (-0.0880, +0.0093, +0.0493, +0.0120, +0.0013)$$

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 0.0045 C 0.0024 D 0.0089

18 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 su sljedeći:

$$(\mathbf{x}^{(1)}, y^{(1)}) = ((9, 30, 21), -1)$$
$$(\mathbf{x}^{(2)}, y^{(2)}) = ((-11, -26, -15), -1)$$
$$(\mathbf{x}^{(3)}, y^{(3)}) = ((-1, -7, -6), +1)$$

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 $\mathbf{x} = (1, 1, 25)$.**

A $+0.947$ B -0.676 C $+1.434$ D -2.330

19 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:

$$C = \{2^{-5}, 2^{-4}, \ldots, 2^4, 2^5\}$$
$$\gamma = \{10^{-5}, 10^{-4}, \ldots, 10^4, 10^5\}$$

Za model sa $C = 2^{-2}$ i $\gamma = 10^1$ utvrdili smo da je podnaučen, a za model sa $C = 2^1$ i $\gamma = 10^1$ utvrdili smo da je prenaučen. **Koliko modela još ima smisla ispitati jer su moguće optimalni?**

A 25 B 68 C 10 D 96

20 Raspolažemo sljedećim skupom označenih primjera u trodimenzijskome ulaznom prostoru:

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

Na ovom skupu primjera treniramo model SVM-a s linearnom jezgrenom funkcijom i sa $C = 0.05$. Postupak treniranja algoritmom SMO završio je s vektorom Lagrangeovih koeficijenata $\boldsymbol{\alpha} = (0.05, 0, 0.05)$. Iz ovoga se da izračunati da vrijedi $w_0 = -0.675$. Umjesto da smo koristili algoritam SMO, mogli smo upotrijebiti gradijentni spust i optimirati težine u primarnoj formulaciji problema. U tom slučaju koristili bismo empirijsku pogrešku SVM-a definiranu kao L_2-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 1.585 B 1.650 C 1.958 D 1.725

21 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 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

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

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

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

22 Raspolažemo sljedećim skupom označenih primjera u dvodimenzijskome ulaznom prostoru:

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

Na ovom skupu treniramo model SVM-a s tvrdom marginom. Međutim, naknadno smo utvrdili da je primjer $(2,0)$ 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 $\frac{5}{4}\sqrt{2}$ puta B $\frac{1}{2}\sqrt{10}$ puta C $\frac{1}{2}\sqrt{13}$ puta D $\frac{1}{2}\sqrt{17}$ puta

23 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, a također je to rjeđi i parametarski model jer λ raste

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

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

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

24 Na 500 primjera sa 80 značajki treniramo rijetki jezgreni stroj s Gaussovim jezgrama. Sve Gaussove jezgre imaju istu varijancu. Nakon treniranja, dobivamo model koji ima 38 prototipa. **Koliko parametara moramo optimirati te koliko parametara ima naučeni model?**

A Optimiramo 501 parametara, a naučeni model ima 3040 parametara

B Optimiramo 501 parametara, a naučeni model ima 3541 parametar

C Optimiramo 81 parametar, a naučeni model ima 3079 parametara

D Optimiramo 501 parametara, a naučeni model ima 3079 parametara