Strojno učenje – pismeni ispit

UNIZG FER, ak. god. 2019./2020.

7. rujna 2020.

spit traje 150 minuta i nosi 35 bodova. Svaki zadatak rješavajte na zasebnoj stranici. Pišite uredno i čitko. Jemojte pretpostavljati da je nešto očito; Vaše znanje može se ocijeniti samo na temelju onog što napišete. Kod kica grafikona, označite osi, budite uredni i precizni te označite ekstreme krivulja, ako postoje.

1. (*5 bodova*) Osnovni koncepti.

   (a) Skup označenih primjera je $\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_i = \{((0,0),1),((0,1),1),((1,1),0)\}$. Razmatramo linearan klasifikacijski model $\mathcal{H}$ s parametrima $\boldsymbol{\theta} \in \mathbb{R}^3$. Odredite $|\mathcal{H}|$ i $|\mathrm{VS}_{\mathcal{H},\mathcal{D}}|$ za slučajeve (1) $\mathcal{X} = \{0,1\}^2$, (2) $\mathcal{X} = \mathbb{Z}^2$ i (3) $\mathcal{X} = \mathbb{Z}^2$ uz proširenje skupa $\mathcal{D}$ primjerom $((-1,1),0)$.

   (b) Objasnite što su to induktivna pristranost jezika i pretraživanja. Na primjeru objasnite je li moguće deduktivno odrediti oznaku svakog primjera x iz $\mathcal{D} = \left\{(x^{(i)}, y^{(i)})\right\}_{i=1}^{N}$ služeći se samo pristranošću jezika. A samo pristranošću pretraživanja?

2. (*4 boda*) Linearna regresija.

   (a) Model regresije treniramo na podatcima koji su generirani funkcijom $f(x) = 3\cdot(x-2)^2+1$. Koristimo funkciju preslikavanja $\phi(x) = (1, x, x^2)$.. Skicirajte izokonture neregularizirane funkcije pogreške u ravnini $\mathbb{R}^2$ koju definiraju parametri w_1 i w_2 i izokonture L2-regularizacijskog izraza. Ako je faktor λ odabran tako da se jednak značaj pridaje složenosti modela i minimizaciji pogreške, skicirajte (otprilike) vektor optimalnih težina (w_1^*, w_2^*).

   (b) Ukratko opišite kako se algoritam linearne regresije može upotrijebiti za binarnu klasifikaciju. Objasnite koji je glavni nedostatak takvog postupka kroz primjer i kroz skicu funkcije gubitka tog algoritma (graf od L u ovisnosti o $yh(\mathbf{x})$).

3. (*4 boda*) Logistička regresija.

   (a) Izvedite pogrešku unakrsne entropije i objasnite sve korake te korištene pretpostavke.

   (b) Koristimo model multinomijalne logističke regresije za klasifikaciju primjera u $K = 4$ klase. Uz danu matricu težina $\mathbf{W}$, je li vjerojatnije da je primjer $\mathbf{x} = (2,5)$ član prve ili treće klase? Napišite čitav postupak kojim ste došli do zaključka.

$$\mathbf{W} = \begin{pmatrix} 0.3 & 0.4 & 0.5 \\ 0.1 & 0.9 & 0.9 \\ 2.5 & 1.0 & 0.5 \\ 1.0 & 1.2 & 0.4 \end{pmatrix}$$

4. (*8 boda*) Stroj potpornih vektora, jezgrene i neparametarske metode.

   (a) Neka su potporni vektori linearnog SVM-a $\mathbf{x}^{(1)} = (-2, 3, 5, 5)$ i $\mathbf{x}^{(2)} = (6, 4, 3, 1)$. Prvi primjer je negativan, a drugi pozitivan. Dualni parametri su $\alpha_1 = 0.2$ i $\alpha_2 = 0.5$, a pomak je $w_0 = -2$. Napišite izraz za gubitak zglobnice i odredite gubitak hipoteze za primjer $\mathbf{x}^{(3)} = (1, 1, 1, 1)$, ako $y^{(3)} = -1$.

   (b) Skicirajte pogreške učenja i ispitivanja kao funkcije od k za model k-NN.

5. (*8 bodova*) Procjenitelji, Bayesov klasifikator i probabilistički grafički modeli.

   (a) Izvedite, korak po korak, ML-procjenitelj parametra μ Bernoullijeve razdiobe. Na primjeru ilustrirajte problem prenaučenosti kod izvedenog procjenitelja te kako biste taj problem ublažili.

   (b) Definirajte KL-divergenciju i uzajamnu informaciju, te objasnite kako biste ih iskoristili kao kriterij za odabir modela kod polunaivnog Bayesovog klasifikatora.

   (c) Bayesovom mrežom modeliramo vjerojatnost oboljenja od kardiovaskularnih bolesti. Mreža sadrži četiri varijable: spol osobe (S), koliko često osoba tjedno odlazi u teretanu (T), je li osoba pušač (P), te varijablu koja govori o kakvom se riziku radi (R). Pritom vrijedi $s \in \{\textit{muški}, \textit{ženski}\}$, $p \in \{\bot, \top\}$, $t \in \{1, 3, 5\}$ i $r \in \{\textit{nizak}, \textit{umjeren}, \textit{visok}\}$. Zajednička razdioba faktorizirana je kao $P(S, T, P, R) = P(S)P(P)P(T|S, P)P(R|T)$. Primjenom (Laplaceovog) MAP-procjenitelja procijenite $P(T|S, P)$. Pritom je dan skup podataka $\mathcal{D}$:

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*

6. (*6 bodova*) Vrednovanje klasifikatora i odabir značajki.

   (a) Od $N = 1000$ primjera, klasifikator je za prvu, drugu i treću klasu ispravno klasificirao njih 590, 146 odnosno 134. Od preostalih 130 neispravno klasificiranih primjera, 30 ih je klasificirano u drugu klasu umjesto u prvu, 60 u drugu umjesto u treću, a 40 u treću umjesto u prvu klasu. Izračunajte mikro-F_2.

   (b) Pretpostavimo skup podataka $\mathcal{D}$ s $N = 100$ primjera, pri čemu je 60 primjera jedne klase, a 40 druge. Provodimo binarnu klasifikaciju korištenjem modela većinskog glasanja. Ako koristimo stratificiranu 10-struku unakrsnu provjeru za procjenu pogreške, hoće li se procjena pogreške promijeniti ako ponovimo isti postupak koristeći po 50% primjera obje klase? Ako da, hoće li biti veća ili manja? Detaljno obrazložite.

   (c) Opišite metodu "izdvoji" jednog i navedite njene nedostatke.

7. (*5 bodova*) Grupiranje.

   (a) Napišite pseudokôd algoritma k-medoida. Koje su prednosti, a koje mane ovog algoritma nad algoritmom k-sredina?

   (b) Raspolažemo manjim skupom od 7 primjera $(x_1, x_2, \ldots, x_7)$. Referentno grupiranje ovih primjera grupe definirano je vektorom pridjeljivanja primjera grupama $x_i \mapsto j$: $(1, 2, 1, 3, 4, 4, 1)$. Algoritmom k-medoida dobiveno je grupiranje $(2, 2, 1, 1, 2, 3, 1)$. (Dakle, četvrti primjer je u referentnom grupiranju član grupe 3, dok je u dobivenom grupiranju član grupe 1.) Izračunajte Randov indeks dotičnog grupiranja.