Strojno učenje – pismeni ispit

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

27. kolovoza 2020.

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

1. (*5 bodova*) Osnovni koncepti.

   (a) Ukratko objasnite tri osnovne komponente svakog algoritma strojnog učenja te ih na primjeru modela logističke regresije povežite s konceptima pristranosti jezikom i pristranosti preferencijom.

   (b) Raspolažemo sljedećim primjerima za učenje:

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

   Skicirajte predvidive hipoteze (1) linearne regresije, (2) perceptrona, (3) logističke regresije i (4) stroja potpornih vektora. Učinite isto (na novoj skici) za skup podataka u koji je dodan primjer $\big((8,1),1\big)$ i komentirajte razlike naspram prve skice.

2. (*5 bodova*) Linearna regresija.

   (a) Kod linearne regresije empirijska je pogreška definirana kao:

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

   Pokažite da je minimizacija ovog izraza istovjetna maksimizaciji log-izglednosti $\ln P(\mathcal{D}|\mathbf{w})$ (tj. minimizaciji negativne log-izglednosti) uz pretpostavku normalno distribuiranog šuma $\mathcal{N}(h(\mathbf{x}|\mathbf{w}), \sigma^2)$.

   (b) Linearnom višestrukom regresijom modeliramo ovisnost prihoda (zavisna varijabla) o dobi, godinama radnog staža i broju djece. Na ovom primjeru objasnite problem multikolinearnosti te njenu vezu s rangom matrice dizajna $\boldsymbol{\Phi}$. Koliko najmanje primjera trebamo imati da bi rješenje bilo stabilno?

3. (*5 bodova*) Logistička regresija.

   (a) Skicirajte regularizirane pogreške učenja i ispitivanja u ovisnosti o broju iteracija za $\lambda = 0$ i $\lambda = 100$ za (1) linearno odvojiv problem i (2) linearno nedvojiv problem (dva grafikona sa po 4 krivulje).

   (b) Napišite model multinomijalne logističke regresije i objasnite interpretaciju izlaza modela.

4. (*4 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) Primjenjujemo model k-NN na skup podataka D koji se sastoji tri klase, pri čemu je 40 primjera jedne klase, 20 druge i 70 treće. Kolika će biti pogreška učenja ovog modela na skupu D ako koristimo $k = 1$ i ako koristimo $k = 130$? Objasnite.

5. (*7 bodova*) Procjenitelji, Bayesov klasifikator i probabilistički grafički modeli.

   (a) Navedite prednosti i nedostatke probabilističkih modela, a zatim objasnite vezu između učenja probabilističkih modela i procjene parametara.

   (b) Izvedite model naivnog Bayesovog klasifikatora krenuvši od nefaktorizirane izglednosti. Napišite sve korištene pretpostavke.

   (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 $s \in \{muški, ženski\}$, $p \in \{\bot, \top\}$, $t \in \{1, 3, 5\}$ i $r \in \{nizak, umjeren, visok\}$. Zajednička razdioba faktorizirana je kao $P(S, T, P, R) = P(S)P(P)P(T|S, P)P(R|T)$. 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. (*4 boda*) 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 makro-F_2.

   (b) Napišite pseudokôd ugniježđene višestruke unakrsne provjere (engl. *nested k-fold CV*) i objasnite kada moramo koristiti taj postupak, a ne običnu višestruku unakrsnu provjeru?

7. (*5 bodova*) Grupiranje.

   (a) Navedite dva načina odabira početnih središta u algoritmu k-sredina. Kako odabir početnih središta utječe na konvergenciju algoritma, a kako na sâmu brzinu konvergencije?

   (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.