Dubinska analiza i otkrivanje znanja iz podataka — Ispit 1.1

23. 06. 2026.

Napomene. Sva rješenja i pomoćne račune pišete isključivo na papirima koje će vam dati dežurni asistent/ica. Dozvoljeno je korištenje isključivo pribora za pisanje i brisanje! Mobitele isključite i pospremite; nisu dozvoljeni niti kao zamjena za sat. Dozvoljeni načini zapisivanja algoritama u zadacima: bilo koji proceduralni jezik, uključivo i pseudokod. Svi argumenti funkcija trebaju biti navedeni i objašnjeni; petlje moraju biti korektno naznačene; poželjno je pisati komentare.

1. (10) Neka je zadan skup podataka s predviđanjima ($\mathcal{P}_{M_1}$ i $\mathcal{P}_{M_2}$) dva klasifikacijska modela (M_1, M_2), te ciljnom klasom `Cilj` kao u doljnjem prikazu:

E	...	$\mathcal{P}_{M_1}$	$\mathcal{P}_{M_2}$	Cilj
e_1	...	1	1	1
e_2	...	0	1	0
e_3	...	0	0	0
e_4	...	0	1	1
e_5	...	1	1	1
e_6	...	1	1	1
e_7	...	0	1	0
e_8	...	0	1	1
e_9	...	0	0	0
e_{10}	...	1	1	1

- Ukoliko zadani skup podataka predstavlja skupinu kandidata za kredit banke, a klasifikacijski model odlučuje kome banka treba dati kredit, koji od modela M_1 ili M_2 bi trebalo koristiti? Detaljno objasnite i potkrijepite svoj odgovor izabirom i računom podskupa odgovarajućih evalucijskih mjera (od niže navedenih). Obavezno navedite sve korake u računu. Obrazložite učinak primjene modela na dostupnim podacima na odluke banke i dostupnost kredita kandidatima.

- Ukoliko zadani skup podataka predstavlja skupinu osoba čija medicinska mjerenja nam trebaju otkriti koju osobu poslati na detaljnu obradu specijalistu onkologu (specijalizacija za proučavanje, detekciju i liječenje tumora), koji od modela M_1 ili M_2 bi trebalo koristiti? Detaljno objasnite i potkrijepite svoj odgovor izabirom i računom podskupa odgovarajućih evalucijskih mjera (od niže navedenih). Obavezno navedite sve korake u računu. Obrazložite učinak primjene modela na dostupnim podacima na odluke bolnice i život potencijalnih bolesnika.

- **Preciznost** - definira se kao: $P(\mathcal{M}, D) = \dfrac{SP}{SP + LP}$. $\mathcal{M}$ označava klasifikacijski model, a D skup podataka.

OKRENITE!

- **Odziv** (eng. *recall* ili *sensitivity* ili *true positive rate*) - definira se kao: $R(\mathcal{M}, D) = \dfrac{SP}{SP + LN}$.
- **Točnost** (eng. *accuracy*) - definira se kao: $Toc(\mathcal{M}, D) = \dfrac{SP + SN}{SP + SN + LP + LN}$.
- **Specifičnost** (eng. *specificity* ili *true negative rate*) - definira se kao: $Spec(\mathcal{M}, D) = \dfrac{SN}{SN + LP}$.
- **Ispadanje** (eng. *fall-out* ili *false positive rate*) - definira se kao: $Isp(\mathcal{M}, D) = \dfrac{LP}{LP + SN}$.
- **Omjer promašaja** (eng. *miss rate* ili *false negative rate*) - definira se kao: $Prom(\mathcal{M}, D) = \dfrac{LN}{LN + SP}$.

2. (10) Neka je zadan skup podataka kao u tablici.

E	...	C_1	C_2	C_3
e_1	...	1	1	1
e_2	...	1	0	0
e_3	...	0	0	0
e_4	...	0	0	1
e_5	...	1	1	1
e_6	...	1	1	1
e_7	...	1	1	0
e_8	...	0	1	1
e_9	...	0	0	1
e_{10}	...	1	1	1

$C_1,\ C_2,\ C_3$ predstavljaju ciljne varijable.

- Kako se zove prediktivni zadatak koji rješava navedeni problem predviđanja?
- Navedite ime algoritma strojnog učenja, baziranog na strukturi stabla, koji bi mogao stvoriti prediktivni model učenjem iz opisanih podataka. Model treba za nove entitete e_k predviđati vrijednost ciljnih varijabli C_1, C_2 i C_3.
- Napišite formule mjera koje možete koristiti za računanje točaka dijeljenja unutar navedenog algoritma.
- Opišite glavne prednosti navedenog algoritma.
- Opišite kriterij zaustavljanja dijeljenja tog algoritma.

OKRENITE!

3.
(15)

- Što je particijsko klasteriranje?
- Napišite i objasnite pseudokod k-means algoritma.
- Napišite formulu funkcije koju optimira **standardna verzija** navedenog algoritma.
- Koji je najbolji izbor za centroid ukoliko optimiramo funkciju kao u gornjoj točki. Dokažite da je navedeni izbor najbolji.
- Kada k-means algoritam staje s izvršavanjem?

4.
(15)

- Definirajte FP-stablo.
- Napišite pseudokod algoritma za stvaranje FP-stabla. Obavezno objasnite funkcije svih funkcija u pseudokodu.
- O čemu ovise dubina i veličina FP-stabla? Dokažite.

5.
(20)

Pretpostavimo da je zadan skup podataka kao dolje:

E	$\dots$	A_{k_1}	E	$\dots$	B_{k_2}	E	$\dots$	C_{k_3}
e_1	$\dots$	$v_1^{A_{k_1}}$	e_1	$\dots$	$v_1^{B_{k_2}}$	e_1	$\dots$	$v_1^{C_{k_3}}$
e_2	$\dots$	$v_2^{A_{k_1}}$	e_2	$\dots$	$v_2^{B_{k_2}}$	e_2	$\dots$	$v_2^{C_{k_3}}$
e_3	$\dots$	$v_3^{A_{k_1}}$	e_3	$\dots$	$v_3^{B_{k_2}}$	e_3	$\dots$	$v_3^{C_{k_3}}$
e_4	$\dots$	$v_4^{A_{k_1}}$	e_4	$\dots$	$v_4^{B_{k_2}}$	e_4	$\dots$	$v_4^{C_{k_3}}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$e_{\lvert E\rvert}$	$\dots$	$v_{\lvert E\rvert}^{A_{k_1}}$	$e_{\lvert E\rvert}$	$\dots$	$v_{\lvert E\rvert}^{B_{k_2}}$	$e_{\lvert E\rvert}$	$\dots$	$v_{\lvert E\rvert}^{C_{k_3}}$

v_i^A označava vrijednost atributa A za entitet e_i.

Pretpostavimo da atributi A_i opisuju tjelesne aktivnosti osobe, atributi B_i unos različitih prehrambenih proizvoda, a atributi C_i rezultate zdravstvenih pretraga.

- Navedite i definirajte zadatak dubinske analize podataka kojim možemo proučiti ovisnosti navedene tri skupine atributa na temelju dostupnih podataka.
- Navedite i detaljno objasnite pseudokod algoritma (njegovih komponenata kao i pseudokodove sastavnih dijelova) navedenog zadatka koji može ponuditi navedena objašnjenja. Objasnite sve korištene funkcije i korake algoritma.

REZULTATI: ponedjeljak, 29. 06. 2026., na webu,
uvidi: utorak, 30. 06. 2026. u 10 sati, ured 226. Matej Mihelčić