Uvod v odkrivanje znanj iz podatkov (Poslovna inteligenca)

2. izpitni rok

10. februar 2020

Priimek in ime (tiskano): _____

Vpisna številka: _____

Naloga	1	2	3	4	5	Vsota
Vrednost	7	7	5	6	7	32
Točk						

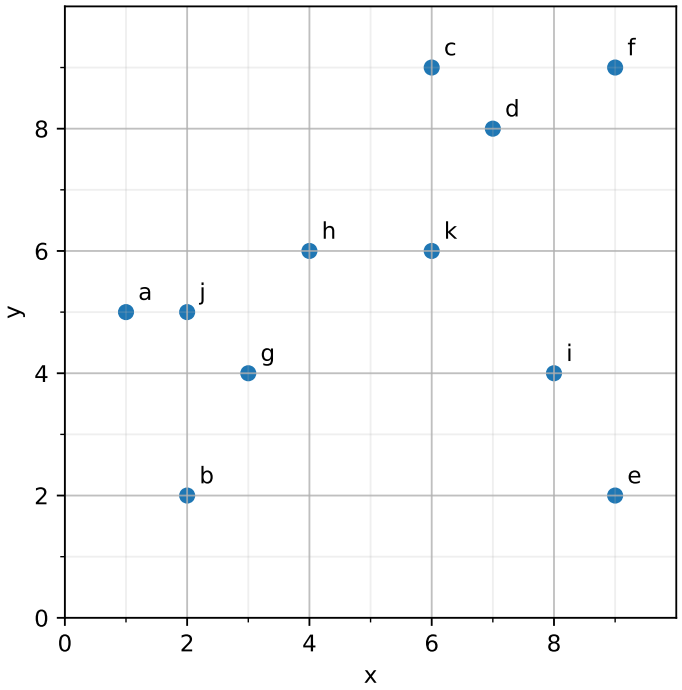
Izjavljam, da nalogo rešujem sam brez kakršnekoli zunanje pomoči.

Prepišite zgornjo izjavo in dodajte svoj podpis: _____

_____

_____

1. Dana je spodnja množica učnih primerov, ki smo jih opisali z dvema zveznima atributoma x in y in jih lahko predstavimo kot točke v Evklidski ravnini:



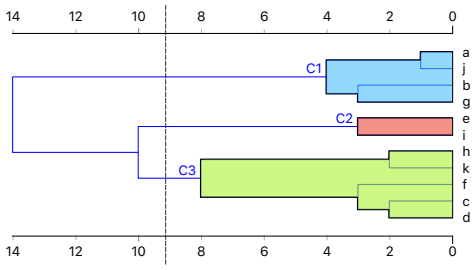
[4] (a) Izriši dendrogram, ki ga dobiš s hierarhičnim razvrščanjem točk v skupine. Kot mero za podobnost uporabi Manhattansko razdaljo, kjer je razdalja med primeroma i in j določena kot $d_{ij} = |x_i - x_j| + |y_i - y_j|$. Podobnost med dvema skupinama meri s tehniko maksimalne razdalje med paroma točk iz različnih skupin (t. im. *complete linkage*).

[1] (b) Uporabi izrisani dendrogram in na podlagi njega predlagaj razdelitev primerov v tri skupine (na dendrogramu izriši vertikalo, ki točke razdeli v tri skupine). Izpiši, kateri primeri pripadajo posamezni skupini.

[1] (c) Kateri primer ima najmanjšo vrednost silhuete in koliko ta znaša? Tudi za izračun silhiete uporabi Manhattansko razdaljo.

[1] (d) Kateri primer ima najvišjo vrednost silhuete in koliko ta znaša?

$s_i = \frac{b_i - a_i}{\max(a_i, b_i)}$ $\quad a(i) = \frac{1}{|C_i|-1} \sum_{j \in C_i, i \neq j} d(i,j)$ $\quad b(i) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i,j)$



Solution:
h (-0.20), j (+0.71)

Stran je prazna, da lahko nanjo rešujete nalogo.

2. V matriki ocen $R \in \mathbb{R}^{m \times n}$ vsaka vrstica predstavlja enega od m uporabnikov, vsak stolpec pa enega od n predmetov (ali izdelkov). Matrika R je redka matrika, kar pomeni, da večina njenih vrednosti ni določenih. Matriko R približno predstavimo z matrikama $P \in \mathbb{R}^{m \times k}$ in $Q \in \mathbb{R}^{k \times n}$ (tako, da je $r_{ui} \approx \hat{r}_{ui} = p_u q_i^T$). Naj bodo konkretne vrednosti teh matrik:

$$P = \begin{bmatrix} 1 & 0 \\ 2 & 2 \\ 2 & 1 \\ 1 & 2 \end{bmatrix}$$

$$Q = \begin{bmatrix} 2 & 0 & 1 & 2 & 2 \\ 1 & 2 & 2 & 1 & 0 \end{bmatrix}$$

[1] (a) Kaj predstavlja vrstica z vrednostmi $[2, 2]$ v matriki P?

[1] (b) Kaj predstavlja kolona z vrednostmi $[2, 0]^T$ v matriki Q?

[2] (c) Izdelke po vrsti označimo z i_j, kjer je i oznaka za izdelek in je $j = 1 \ldots n$ zaporedna številka izdelka. Na osnovi podanega predlagaj dve skupini izdelkov tako, da navedeš, kateri izdelki so v prvi in kateri so v drugi skupini.

[1] (d) V priporočilnih sistemih $\hat{r}_{ui}$ uporabimo kot oceno, ki jo je napovedal naš model. Izračunajte napovedane ocene za vse uporabnike in vse predmete – torej, izračunajte vse elemente matrike $\hat{R}$.

[1] (e) Metoda, ki priporočilni model gradi z matrično faktorizacijo (npr. algoritem ISMF) v vsaki iteraciji spremeni matriki P in Q tako, da dobimo boljši približek podatkov v matriki R. Kako merimo kakovost razcepa matrike R v njena faktorja P in Q? Opišite z besedami in podajte funkcijo, ki meri kakovost razcepa na učnih podatkih.

[1] (f) Izriši dvodimenzionalni razsevni diagram izdelkov tako, da so pozicije posameznih izdelkov v grafu smiselne in da te pozicije pridobiš s čimmanj dodatnega (nepotrebnega) računanja. Na grafu označi izdelek z njegovo oznako (glej točko c).

Solution:

a) predstavitev drugega uporabnika v latentnem prostoru

b) predstavitev zadnjega izdeleka v latentnem prostoru

c) $(1, 4, 5), (2, 3)$

d)

$$P = \begin{bmatrix} 2 & 0 & 1 & 2 & 2 \\ 6 & 4 & 6 & 6 & 4 \\ 5 & 2 & 4 & 5 & 4 \\ 4 & 4 & 5 & 4 & 2 \end{bmatrix}$$

e) za podani primer, kvadrat razlike napovedi in znane vrednosti

f) pozicije izdelkov določa matrika Q

Stran je prazna, da lahko nanjo rešujete nalogo.

[5] 3. Priporočilni sistem, ki deluje na podlagi matričnega razcepa, smo pognali na podatkih o ocenah knjig. V podatkih se pojavlja 2000 uporabnikov, ki so z ocenami med 0 in 10 ocenili (nekatere od) 15000 knjig. Skupno imamo v učni množici 100000 ocen. S pomočjo validacijske množice in poskušanjem smo našli najboljše parametre: $k = 13$ latentnih faktorji in ustavitev po 20 iteracijah, ker dobimo najboljši rezultat na validacijski množici RMSE = 1.81. Regularizacije pa ne uporabljamo.

Skicirajte koren srednje kvadratne napake (RMSE) v odvisnosti od števila iteracij gradientnega sestopa: narišite koordinatni sistem s številom iteracij na osi x (v intervalu $[0, 50]$) in RMSE na osi y (med $[0, 10]$). Vanj čim bolj natančno vrišite in označite tri krivulje (rišite jih na celotnem območju iteracij; brez ustavljanja):

- RMSE na učni množici za število latentnih faktorjev $k = 13$
- RMSE na učni množici za število latentnih faktorjev $k = 2$
- RMSE na učni množici za število latentnih faktorjev $k = 90$

Solution:

(1 točka) Vse krivulje začnejo več ali manj na isti na isti točki (0.5 točke), za smiselno inicializacijo največ na RMSE = 5 (0.5 točke).

(1 točka) Vse krivulje vedno padajo, RMSE > 0.

(1 točka) Večji kot je k, nižjo vrednost doseže krivulja.

(2 točki) Krivulji za $k = 13$ in $k = 90$ morata imeti pri 20 iteracijah vrednost nižji od 1.81.

4. Zbrali smo podatke o uspešnosti kampanij na portalu Kickstarter tako, da smo vsako kampanijo opisali z atributi ter jo razvrstili v uspešno (kampanija je pridobila dovolj finančnih prispevkov) ali neuspešno. Podatke razdelimo na učno in testno množico. Na učni množici z metodo logistične regresije zgradimo model, ki napoveduje verjetnost, da je kampanija uspešna. Pri razvoju modela je bila stopnja učenja pri gradientnem pristopu $\alpha = 0.001$ in stopnjo regularizacije $\lambda = 0.1$. Mejo za verjetnosti, pri kateri primere razvrstimo v ciljni razred (uspešne kampanije) postavimo na 0.5. Klasifikacijsko točnost tako dobljenega modela ocenimo na učnih podatkih (0.9) in na testnih podatkih (0.8). Površina krivulje ROC na učnih podatkih je 0.8, na testnih podatkih pa (0.7).

   Oceni pravilnost trditev (zapiši "pravilna" če trditev vedno drži, ali pa "nepravilna" če trditev ne drži). Ne ugibajte: pri napačnem odgovoru pri tej nalogi bomo točke pri tem odgovoru odšteli (pri tem pa upoštevali, da je minimalno število točk pri tej nalogi enako 0).

[1] (a) Ko zvišamo stopnjo učenja na $\alpha = 0.01$, se AUC zniža.

[1] (b) Ko znižamo stopnjo učenja na $\alpha = 0.0001$ traja računski postopek učenja modela dlje.

[1] (c) Regularizacijo znižamo na $\lambda = 0.01$. Klasifikacijska točnost na učnih podatkih se izboljša (poveča).

[1] (d) Regularizacijo znižamo na $\lambda = 0.01$. Klasifikacijska točnost na testnih podatkih se izboljša (poveča).

[1] (e) Mejo za verjetnosti, pri kateri primere razvrstimo v ciljni razred (uspešne kampanije) postavimo na 0.8. Površina krivulje ROC na učnih podatkih se ne spremeni.

[1] (f) Mejo za verjetnosti, pri kateri primere razvrstimo v ciljni razred (uspešne kampanije) postavimo na 0.8. Površina krivulje ROC na testnih podatkih se zviša.

Solution: N, P, P, N, P, N

5. Označi, ali je trditev pravilna (obkroži DA) ali ne (obkroži NE). Za nepravilen odgovor se odbije 0.25 točke.

[.5] (a) DA | NE: Prednost metode voditeljev pred hierarhičnim gručenja v skupine je grafična predstavitev rezultatov gručenja.

[.5] (b) DA | NE: Za iskanje gručenja pri znanem številu gruč je metoda voditeljev hitrejša od hierarhičnega gručenja v skupine.

[.5] (c) DA | NE: Problem metode silhuete, ki jo lahko uporabimo pri ocenjevanju centralnosti primera v gruči je, da ne uporablja oddaljenosti primera do sosednje gruče.

[.5] (d) DA | NE: Osamelci v gručah imajo zelo nizo vrednost silhuete.

[.5] (e) DA | NE: Matrična faktorizacija z višjim številom latentnih komponent se lahko bolj prilagodi učnim podatkov.

[.5] (f) DA | NE: Večrazredno lestvičenje je primer tehnike, ki primere projecira v nižjedimenzionalne prostore.

[.5] (g) DA | NE: Za dvodimenzionalne podatke bo metoda glavnih komponent vrnila enake vrednosti novih pozicij primerov v dveh dimenzijah kot metoda večrazrednega lestvičenja.

[.5] (h) DA | NE: Lego primerov v nižjih dimenzijah lahko pri metodi večrazrednega lestvičenja pridobimo z gradientnim sestopom.

[.5] (i) DA | NE: Regularizacija tipično poslabša napovedno točnost na učnih primerih.

[.5] (j) DA | NE: Regularizacija lahko izboljša napovedno točnost na validacijski množici.

[.5] (k) DA | NE: Mejna ploskev, ki loči razreda pri logistični regresiji je ravnina.

[.5] (l) DA | NE: Kriterijske funkcije pri linearni regresiji ne moremo izpeljati iz verjetja.

[.5] (m) DA | NE: Evklidska razdalja je primerna za ocenjevanje razdalje med primeri, ki so opisani z veliko (recimo nad 100) značilkami.

[.5] (n) DA | NE: Vsaj ena od dimenzih matrik P in Q pri faktorizaciji, ki jo lahko uporabljamo za priporočilne sisteme, je enaka.

Solution:

```
NE DA NE DA DA
NE NE DA DA DA
DA NE NE DA
```