Uvod v odkrivanje znanj iz podatkov (Poslovna inteligenca)

1. izpitni rok

29. januar 2019

Priimek in ime (tiskano): ______________________________

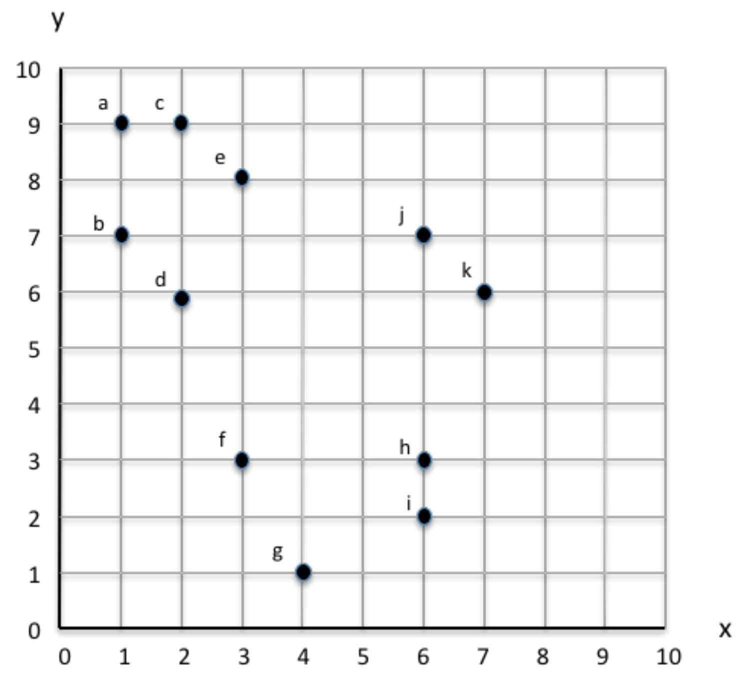
Vpisna številka: ______________________________

Naloga	1	2	3	4	5	Vsota
Vrednost	5	7	8	6	3	29
Točk						

Izjavljam, da sem nalogo rešil sam, brez kakršnekoli zunanje pomoči in brez uporabe nedovoljenih virov informacij.

Podpis (podpis te izjave je obvezen): ______________________________

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.

Solution:

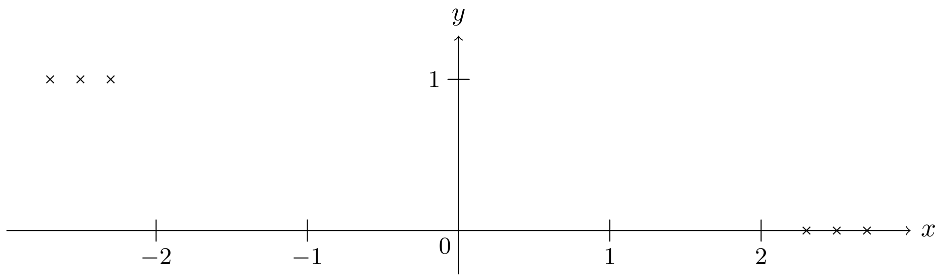
```
ac e bd | fg hi | jk
1     2    3 1    2
  3
    4         4
                 8
       12
```

(prostor za rešitve)

[3] 2. (a) Z "da" ali "ne" označi, ali so sledeče izjave glede logistične regresije resnične.

- Z regularizacijo ne moremo poslabšati rezultatov na učni množici.
- Z regularizacijo ne moremo poslabšati rezultatov na testni množici.
- Z dodajanjem novih atributov v model (npr. zmnožkov obstoječih atributov) preprečimo pretirano prilagajanje podatkov učni množici.

[2] (b) Imamo podatke s šestimi primeri in enim atributom (x, glej skico). Napovedati želimo razred y. Dvakrat uporabimo logistično regresijo: prvič z zelo majhno vrednostjo regularizacijskega koeficienta λ, drugič z zelo veliko. V koordinatni sistem vrišite krivulji, ki opisujeta napovedi logistične regresije $P(Y = 1)$ pri veliki in majhni vrednosti λ.



[2] (c) Osnovno tehniko logistične regresije uporabljamo na dvorazrednih podatkih. Kako bi lahko prilagodili postopek za podatke, kjer je razredov več (npr. pet)?

(prostor za rešitve)

3. Dana je spodnja matrika podatkov:

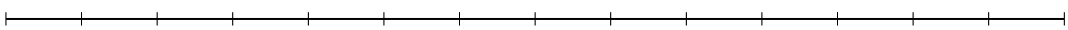
	x_1	x_2
A	1	1
B	3	1
C	5	1
D	6	2
E	-1	-1
F	-3	-1
G	-5	-3
H	-6	0
I	0	0

Podatke A do I projecirajte v eno dimenzijo tako, da bo predstavitev čim bolj verna. (Projekcijo lahko določite ročno, brez kalkulatorjev oziroma brez uporabe linearne algebre, pomagate pa si lahko z izrisom podatkov v kakšen graf. Za pomoč smo izbrali podatke tako, da so ti že centrirani).

[1] (a) Kaj pomeni, da so podatki centrirani? Odgovori tako, da zapišeš matematični izraz, ki mora biti pravilen, če so podatki centrirani.

[3] (b) Zapišite predpis $y = f(x_1, x_2)$, ki za podatek iz osnovnega prostora (x_1, x_2) izračuna vrednost njegove enodimenzionalne projekcije y.

[1] (c) Na spodnji številski osi označi (številčne vrednosti na osi označite sami), kam se projicirajo podatki iz zgornje tabele (to je, na osi označi, kam se transformirajo podatki A do I).



[2] (d) Kaj smo mislili z izrazom "čim bolj verna". Opredelite ta pojem v stavku in s cenilno funkcijo.

[1] (e) Kako imenujemo matematični postopek, ki nam lahko služi za reševanje te naloge in s katerim pridobimo transformacijski predpis?

Solution: a) $\sum_{i=1}^{N} x_j^{(i)} = 0$ za $j \in \{1, 2\}$
b) $p = (4, 1) = (0.97, 0.24)$, $y = x^T p$
c) H G F E I A B C D
d) Maksimiziramo varianco, ali pa skušamo ohraniti razdalje med primeri.
e) Metoda glavnih komponent.

(prostor za rešitve)

4. V matriki ocen $R \in \mathbb{R}^{m \times n}$ vsaka vrstica predstavlja enega od m uporabnikov, vsak stolpec pa enega od n izdelkov. Matrika R je redka matrika, kar pomeni, da večina njenih vrednosti ni določenih. V našem primeru je

$$R = \begin{bmatrix} 3.5 & 4 & & 2.5 \\ 4 & & & \\ & & 3 & \\ 2.5 & & 1.5 & \\ & 3 & 2 & 2 \end{bmatrix}$$

Matriko R lahko približno predstavimo z matrikama $P \in \mathbb{R}^{m \times k}$ in $Q \in \mathbb{R}^{k \times n}$ (tako, da je $r_{ui} \approx \hat{r}_{ui} = p_u q_i$).

[2] (a) Kako merimo kakovost razcepa matrike R v matriki P in Q? Opišite z besedami ali podajte kriterijsko funkcijo.

[3] (b) R nam je brez napake uspelo faktorizirati v matriki P in Q (zgornja kriterijska funkcija ima vrednost 0). Žal smo matriko Q izgubili. Določite izgubljeno Q, če poznamo

$$P = \begin{bmatrix} 1.5 & 1 \\ 1 & 1.5 \\ 1.5 & 1.5 \\ 0.5 & 1 \\ 1 & 1 \end{bmatrix}.$$

[1] (c) Glede na matriki P in Q rangirajte predmete za tretjega uporabnika.

Solution:

```
a = np.array([[ 1.5, 1, 1.5, 0.5, 1] , [ 1, 1.5, 1.5, 1., 1 ]]).T
b = np.array([[ 1, 2, 1, 1.], [ 2, 1, 1., 1, ]])
>>> a.dot(b)
```

```
>>> a.dot(b)
array([[ 3.5,  4. ,  2.5,  2.5],
       [ 4. ,  3.5,  2.5,  2.5],
       [ 4.5,  4.5,  3. ,  3. ],
       [ 2.5,  2. ,  1.5,  1.5],
       [ 3. ,  3. ,  2. ,  2. ]])
```

Stran je prazna, da lahko nanjo rešujete nalogo.

[3] 5. Dani so transakcijski podatki v obliki nakupovalnih košaric:

ID	kupljeni izdelki
1	$\{c, b, d, e\}$
2	$\{b, c, d\}$
3	$\{a, b, d, e\}$
4	$\{a, c, d, e\}$
5	$\{b, c, d, e\}$
6	$\{b, d, e\}$
7	$\{c, d\}$
8	$\{a, b, e\}$

Za spodnja pravila poišči njihovo podporo in zaupanje:

- $\{e\} \rightarrow \{d, b\}$
- $\{e, b\} \rightarrow \{d\}$
- $\{c\} \rightarrow \{d\}$

$\sigma(X) = |\{t_i | X \subseteq t_i,\ t_i \in T\}| \qquad s(X \rightarrow Y) = \sigma(X \cup Y)/N \qquad c(X \rightarrow Y) = \sigma(X \cup Y)/\sigma(X)$

Solution: support, confidence

```
0.500   0.667   e -> d b
0.500   0.800   e b -> d
0.625   1.000   c -> d
```