Poslovna inteligenca

1. izpitni rok

31. januar 2012

Ime in priimek: _____

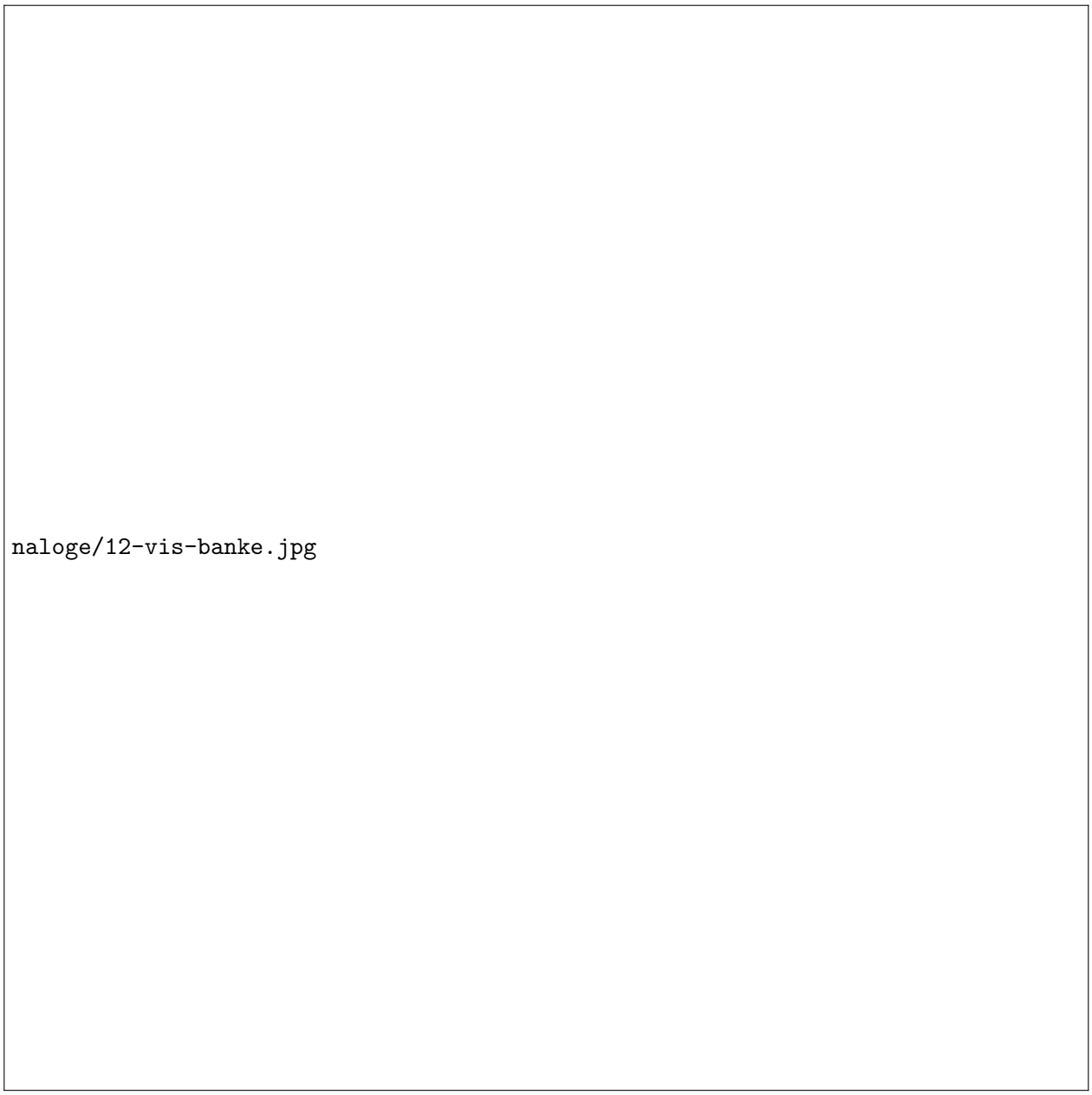
Vpisna številka: _____

Naloga	1	2	3	4	5	6	Vsota
Vrednost	10	10	5	10	5	5	45
Točk							

[10] 1. Spodnja slika kaže ceno bank (koliko milijard dolarjev bi potrebovali, da bi kupili vse delnice banke) pred gospodarsko krizo (sredina 2007) in v času krize (januar 2009).

   (a) Komentirajte sliko: kaj je narobe z njo?

   (b) Predlagajte boljšo vizualizacijo teh podatkov!



[10] 2. Odločamo se za nakup novega avta. Pri nakupu si pomagamo z modelom, ki upošteva ceno avta, ceno vzdrževanja, udobje in varnost. Vsak kriterij je lahko ocenjen z vrednostmi "nizek" ali "visok". Želimo nizke cene ter visoko stopnjo varnosti in udobja. Uporabimo naslednji hierarhični model:

cena avta	cena vzdrževanja	stroški
nizka	nizka	nizki
nizka	visoka	srednji
visoka	nizka	srednji
visoka	visoka	visoki

udobje	varnost	kvaliteta
nizko	nizka	nizka
nizko	visoka	nizka
visoko	nizka	srednja
visoko	visoka	visoka

Skupna ocena avta je sestavljena iz povprečja stroškov in kvalitete, pri čemer upoštevamo to, da želimo čim nižje stroške in čim višjo kvaliteto. Skupna ocena ima lahko vrednosti "dober", "srednji" ali "slab". Povprečje zaokrožimo navzgor (na boljšo vrednost kriterija).

Ovrednotite naslednje variante.

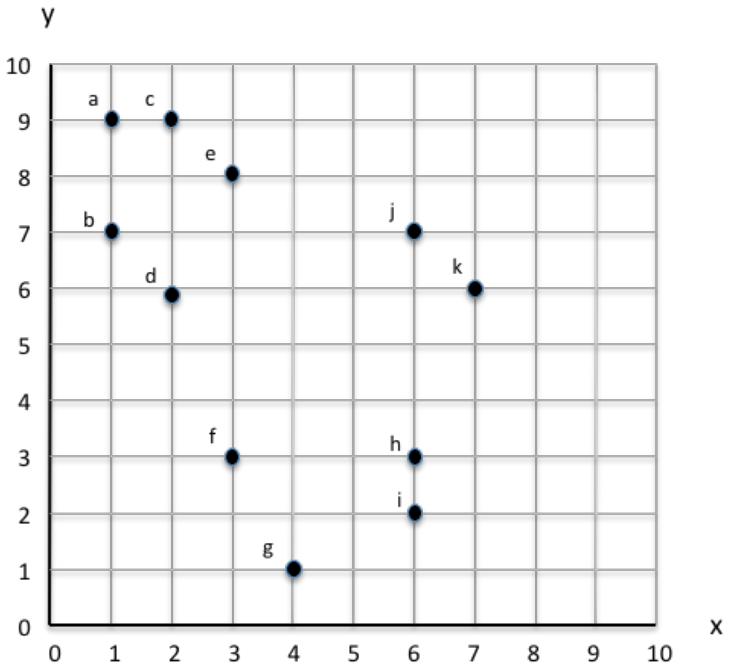
	cena avta	cena vzdrževanja	udobje	varnost
clio	nizka	nizka	visoko	0.5 niz : 0.5 vis
punto	nizka	visoka	nizko	visoka
mini	0.2 niz : 0.8 vis	visoka	0.6 niz : 0.4 vis	0.4 niz : 0.6 vis
yaris	visoka	0.6 niz : 0.4 vis	visoko	*

Pri neznanih vrednostih (označenih z "*") predpostavite, da so vse vrednosti tega kriterija enako verjetne.

[5] 3. Peter kot podatkovni analitik sodeluje z večjim telekomunikacijskim podjetjem. Od njih je ravno prejel podatke, ki 200 naključno izbranih uporabnikov storitev opišejo s 34983 atributi, ter uporabnike razvrstijo v dva razreda glede načina uporabe telefona ("redno", 20 uporabnikov, in "občasno", 180 uporabnikov). Petrova začetna naloga je ugotoviti, ali lahko na podlagi danih atributov poiščemo model, ki bi nove uporabnike uvrstil v enega od teh dveh razredov. Peter se je naloge lotil tako, da je za vse atribute ocenil njihov informacijski prispevek, nato pa izbral 10 najbolje ocenjenih atributov. Na tako dobljenih podatkih (180 uporabnikov, 10 atributov, informacija o razredu) je z 10-kratnim prečnim preverjanjem ocenil napovedno točnost naivnega Bayesa. Dobil je visoko povprečno klasifikacijska točnost, ki je bila enaka 97%. Podjetju je sporočil, da lahko na osnovi njihovih podatkov zgradi model, ki bo zelo zanesljivo nove uporabnike razvrstil v ciljne razrede.

Komentiraj primernost Petrovega izbora postopkov ter upravičenost njegovega zaključka. Če se s kakšnim delom opisanega postopka ne strinjaš, predlagaj alternativno rešitev.

4. 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:



[8] (a) Izriši dendrogram, ki ga dobiš z 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*).

[2] (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
       11
```

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

ID	kupljeni izdelki
1	$\{a, b, d, e\}$
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	$\{a, b, c\}$
9	$\{a, d, e\}$
10	$\{b, d\}$

Za spodnja pravila poišči njihovo podporo in zaupanje:

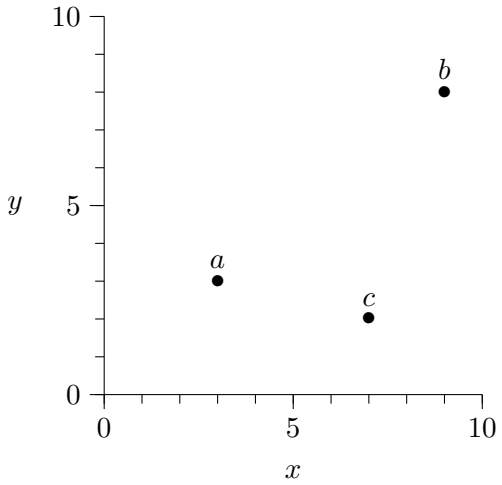
- $\{a\} \rightarrow \{e, d\}$
- $\{a, e\} \rightarrow \{d\}$
- $\{c\} \rightarrow \{d\}$

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

Solution: support, confidence

```
0.400   0.800   a -> e d
0.400   1.000   e a -> d
0.400   0.800   c -> d
```

6. Lastnosti treh besedilnih dokumentov a, b in c smo opisali z dvema neodvisnima numeričnima atributoma, x in y. Med dokumenti bi na podlagi njihovih atributov radi določili njihovo razdaljo.



[1] (a) Oceni, katera dva dokumenta sta si najbližja, če uporabiš evklidsko razdaljo. To razdaljo označi na zgornjem grafu.

[2] (b) Katera dva pa sta si najbližja, če uporabiš kosinusno podobnost. To podobnost označi na zgornjem grafu.

[2] (c) Zakaj imamo na področju analitike besedil raje kosinsno podobnost oziroma kakšna je njena prednost pred evklidsko razdaljo?

Opomba: enačb pri tej nalogi namenoma nismo podali, saj želimo samo, da razdalje oceniš oziroma samo odgovoriš, kateri par dokumentov je z ozirom na izbrano mero najbližji.

Solution: a) ab (nariši črte, kot), b) ac, c) dolžina besedila ne sme vplivati na oceno bližine