Uvod v odkrivanje znanj iz podatkov (Poslovna inteligenca)

3. izpitni rok

10. september 2018

Priimek in ime (tiskano): _____

Vpisna številka: _____

Naloga	1	2	3	4	5	Vsota
Vrednost	5	9	8	6	8	36
Točk						

1. Metka sodeluje s kadrovskim podjetjem. Tam z 42 atributi opišejo zaposlene in pogoje v firmi, kjer so zaposleni. Cilj je napovedati, ali bodo zaposleni v firmi zdržali več kot pol leta. Za to analizo so zbrali zgodovinske podatke o 1500 zaposlenih, med katerimi je prej kot v pol leta po zaposlitvi dalo odpoved 450 anketiranih, 1050 pa jih je zaposlitev obdržalo oziroma so v firmi ostali dlje kot pol leta.

   Metka, ki je pri analizi sodelovala še z Alešem in Majo, je pri študiji naredila kar nekaj poskusov in si vestno beležila klasifikacijske točnosti. Različne vrednosti klasifikacijskih točnosti si je zapisala na listke:

   Listek A) 100%
   Listek B) 95%
   Listek C) 70%
   Listek D) 58%
   Listek E) 20%
   Listek F) 0%

   Na drugih listkih je opisala poskuse, listke pa označila s številkami:

[1] (a) Deset-kratno prečno preverjanje z logistično regresijo. Prvi poskusi kažejo na obetavne rezultate.

[1] (b) Kolono z razredom naključno premešam (za vse podatke). Tako spremenjen nabor podatkov dam Alešu in Maji (oba torej prejmeta isti nabor podatkov). Aleš zgradi model z metodo naključnega gozda, s tem modelom pa pomaga Maji klasificirati primere iz njenega nabora podatkov.

[1] (c) Izberem naključnih 65% primerov podatkov. Te dam Alešu. Vse ostale podatke dam Maji. Alešu naročim, da vrednost razreda napove tako, da je njegova napoved vedno en sam razred, ki je večinski razred v njegovih podatkih. S takimi napovedmi Aleš pomaga Maji klasificirati primere iz njenega nabora podatkov.

[2] (d) Izberem naključnih 50% primerov. Te dam Alešu, ostalim primerom pa pomešam kolono z vrednostjo razredne spremenljivke in jih dam Maji. Aleš na podatkih zgradi model z metodo naključnih gozdov. Ta model potem uporabi pri napovedovanju razredov za primere iz Majinega nabora.

   Listek, kjer je za vsak opis poskusa (številka) zapisala tudi ustrezno oznako rezultato oziroma točnosti (črka) je zgubila. Pomagaj! Za vsak od zgornjih poskusov povej, kakšno točnost je Metka najverjetneje dobila. (Naloga je enostavna za (a), (b) in (c), za (d) pa bo potrebno malo bolj globoko premisliti. Za pravilen rezultat pri poskusu (d) dobite točko, drugo točko pa, če pravilno pokažete, kako ste pri tem poskusu ocenjeno klasifikacijsko točnost izračunali).

Solution: 1B, 2A, 3C, 4D (.7*.7+.3*.3)

(prostor za rešitve)

2. Dani so podatki v enodimenzionalnem prostoru.

[1] (a) Predlagaj razvrstitev primerov v dve skupini. Ne poslužuj se nobenega algoritma, razvrstitev oceni samo na podlagi hitrega ogleda zgornje slike.

[1] (b) Za centroida $\{18, 45\}$ ustvari dve skupini primerov tako, da vsako točko prirediš najbližjemu centroidu. Poročaj o dobljeni razvrstitvi primerov (naštej primere, ki pripadajo prvi in primere, ki pripadajo drugi skupini).

[1] (c) Za centroida $\{15, 40\}$ ustvari skupini tako, da vsako točko prirediš najbližjemu centroidu. Tudi tu poročaj o dobljeni razvrstitvi.

[2] (d) Za para centroidov iz podnaloge (b) in (c) simuliraj metodo voditeljev do konvergence. Dobiš v obeh primerih enake rezultate?

[2] (e) Kateri skupini dobimo pri hierarhičnem razvrščanju z metodo "single linkage"?

[1] (f) Primerjaj razvrstitve iz podnalog (d) in (e). Kater pristop vodi na teh podatkih do bolj "naravnih" skupin, torej skupin, ki so take, kot si jih dobil v nalogi (a)?

[1] (g) Imamo podatke o 10.000 naročnikih na telekomunikacijske usluge, ki so opisani z 150 atributi. Še preden se podatkov lotimo s kakšnim od algoritmov za razvrščanje, bi radi na hitro pogledali, ali imamo kakšne "naravne" skupine v podatkih. Kako bomo to storili?

Solution:

a) $\{6, 12, 18, 24, 30\}$, $\{42, 48\}$
b) $\{6, 12, 18, 24, 30\}$, $\{42, 48\}$
c) $\{6, 12, 18, 24\}$, $\{30, 42, 48\}$
d) postopek se zaključi po prvi iteraciji, ne dobim enake rezultate
e) $\{6, 12, 18, 24, 30\}$, $\{42, 48\}$
f) pravilni voditelji so (18, 45) oziroma b), enako za hierarhično metodo
g) projekcija podatkov v dvodimenzionalni prostor, vizualizacija (PCA, MDS)

(prostor za rešitve)

3. Pet prijateljev smo povprašali, katere športe imajo najraje. Odgovore podaja spodnja tabela.

	tango	salsa	plezanje	smučanje	bordanje	plavanje	kolesarjenje
Agnieszka			♡		♡	♡	♡
Marko	♡		♡		♡	♡	♡
Nejc		♡		♡			♡
Sara			♡	♡	♡		♡
Urška		♡			♡		

[4] (a) Agnieszka želi poskusiti nekaj novega. Pomagajte ji: razvrstite tango, salso in smučanje glede na njene pričakovane preference. Matematično in numerično utemeljite izbran postopek (izberite mero podobnosti, uporabite to mero tako, da odgovorite na vprašanje).

[4] (b) Bi z večjo gotovostjo priporočili smučanje Agnieszki ali Urški? Zakaj? Tudi tu odgovor utemeljite s primernimi izračuni.

$$s_{\mathrm{c}}(u,u') = \frac{\mathbf{r}_u \mathbf{r}_{u'}}{|\mathbf{r}_u|\ |\mathbf{r}_{u'}|} \qquad d_{\mathrm{e}}(u,u') = |\mathbf{r}_u - \mathbf{r}_{u'}| \qquad s_{\mathrm{e}}(u,u') = \frac{1}{1+d_{\mathrm{e}}(u,u)} \qquad s_{\mathrm{J}} = \frac{||\mathbf{r}_u \cap \mathbf{r}_{u'}||}{||\mathbf{r}_u \cup \mathbf{r}_{u'}||}$$

Solution:

1) Če delim z vsoto vseh podobnosti

a) Jaccardova podobnost med Agnieszko in Markom = 0.8, Nejcem = 0.1666, Saro 0.6, Urško: 0.2 (skuaj 1.76). Tango je zato 0.8, salsa 0.366, smučanje 0.7666 (vse deljeno z 1.76). Končne ocene: tango 0.45, salsa 0.21, smučanje 0.43.

b) Jaccardova podobnost med Urško in Agnieszko 0.2, Markov 0.16666, Najcem 0.25, Saro 0.2 (skupaj 0.82). Smučanje: 0.2 + 0.25 = 0.45. Če delim s 0.82, dobim 0.55

4. V Pythonu smo v nekem programu zapisali spodnjo funkcijo:

```
def j(theta, x, y, reg=0.1):
    return -(y.dot(np.log(h(theta, x))) + (1-y).dot(np.log(1-h(theta, x))))
```

[1] (a) Pri kateri tehniki analize podatkov smo to funkcijo uporabljali?

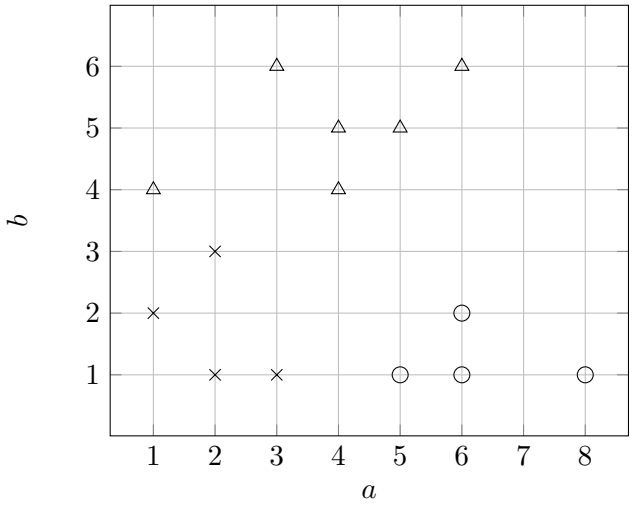
[2] (b) Kaj ta funkcija počne (kaj so vhodni podatki in kaj je izhod)?

[3] (c) Funkcija že vključuje argument za stopnjo regularizacije, a ga v funkciji nismo uporabili. Dopolni funkcijo tako, da dodaš člen z regularizacijo.

Solution:

```
    return -(y.dot(np.log(h(theta, x))) + (1-y).dot(np.log(1-h(theta, x))) -
             reg * sum(theta[1:]**2))
```

5. Podatke, ki vključujejo 14 primerov in so opisani z atributoma a in b, prikazuje spodnje slika. Primeri so razvrščeni v tri skupine, katerih pripadnost za posamezne primere smo na sliki prikazali z simbolom, s katerim je označen primer (križec, trikotnik, krogec). Primer $(a, b) = (3, 1)$ tako na primer pripada skupini križcev, v katerih so skupaj štirje primeri.



Odločimo se, da bomo razdaljo med primeri meri z Manhattansko razdaljo.

[1] (a) Naj bosta x_1 in x_2 dva primera, vrednosti njunih atributov pa označimo z $x_{1,a}$ in $x_{1,b}$ za prvi primer in podobno za drugega. Zapiši enačbo, s katero izračunamo Manhattansko razdaljo med primeroma x_1 in x_2.

[1] (b) Kakšna je Manhattanska razdalja med primeroma $(2, 3)$ in $(6, 1)$?

[1] (c) Kakšna je vrednost silhuete za primer $(5, 1)$.

[2] (d) Kateri med primeri iz skupine trikotnikov, to je primeri, ki so na sliki označeni s tem likom, ima najmanjšo silhueto. Kolika je ta?

[1] (e) Kaj merimo s slihueto in kakšna je njena zaloga vrednosti?

[1] (f) Kaj lahko rečemo o razvrstivi primera z negativo vrednostjo silhuete?

[1] (g) Kakšno zalogo vrednost pričakujemo za slihueto, ki jo izračunamo z isto mero razdalj, kot smo jo uporabili pri razvrščanju v skupine, katerega rezultate (silhuete) vrednotimo? Je ta zaloga vrednosti enaka kot v prejšnjem vprašanju? Zakaj?

$s_i = \frac{b_i - a_i}{\max(a_i, b_i)}$

Solution: a) $d(x_1, x_2) = |x_{1,a} - x_{2,a}| + |x_{1,b} - x_{2,b}|$
b) 6
c) $a = (1 + 2 + 3) = 6/3 = 2$, $b = (2 + 3 + 5 + 5)/4 = 15/4 = 3.75$, $s = 1.75/3.75 = 0.46$
d) primer $(1, 4)$, $a = (3 + 4 + 4 + 5 + 7) = 23/5 = 4.6$, $b = (2 + 2 + 4 + 5) = 13/4 = 3.25$, $s = (3.25 - 4.6)/4.6 = -0.29$
e) centralnost primerov v skupini glede na najbližjo tujo skupino, $[-1, 1]$
f) primer bi moral biti razvrščen v drugo skupino
g) $[-1, 1]$, $[0, 1]$. Ni, pričakujemo, da bodo primeri najbližji skupini, kamor so razvrščeni.

(prostor za rešitve)

(prostor za rešitve)