Poslovna inteligenca

2. izpitni rok

9. februar 2016

Priimek in ime (tiskano): _____

Vpisna številka: _____

Naloga	1	2	3	4	5	Vsota
Vrednost	5	6	5	6	6	28
Točk						

1. V spodnji tabeli je dana učna množica s tremi atributi (Outlook, Company, Sailboat) in razredno spremenljivko (Sail).

	Outlook	Company	Sailboat	Sail
1	sunny	big	small	yes
2	sunny	med	small	yes
3	sunny	med	big	yes
4	sunny	no	small	yes
5	sunny	big	big	yes
6	rainy	no	small	no
7	rainy	med	small	yes
8	rainy	big	big	yes
9	rainy	no	big	no
10	rainy	med	big	no

[1] (a) Kakšna je stopnja nedoločenosti (entropija) razredne spremenljivke Sail?

[2] (b) Kakšen je informacijski prispevek atributa Outlook?
(Namig: najprej izračunaj residualno entropijo $H(\text{Sail}|\text{Outlook})$, potem pa še informacijski prispevek. Residualna entropija ti pove, koliko znaša entropija razreda pri tem, če veš, kakšno vrednost je zavzela spremenljivka Outlook.)

[2] (c) Informacijski prispevek atributa Sailboat je $0,035$, atributa Company pa $0,281$. Kateri atribut bi postopek gradnje odločitventega drevesa postavil v koren drevesa? Naj bo to tudi edino notranje vozlišče drevesa (vse ostalo so listi). Skiciraj, kako izgleda tako odločitveno drevo.

$$H(x) = -\sum_{i \in D(x)} p(x=i) \log_2 p(x=i)$$

$$H(y|x) = \sum_{i \in D(x)} p(x=i) H(y|x=i)$$

$$IG(x;y) = H(y) - H(y|x)$$

kjer so x spremenljivka (razred ali atribut), $D(x)$ zaloga vrednosti spremenljivke x, in y razredna spremenljivka

Solution:

(a) $0.3 \times \log_2(0.3) + 0.7 \times \log_2(0.7) = 0.88$

(b) $H(y|o) = 0.5 \times 0.0 + 0.5 \times (0.4 \times \log_2(0.4) + 0.6 \times \log_2(0.6)) = 0.49$,
$IG(o;y) = H(y) - H(y|o) = 0.88 - 0.49 = 0.39$

(c) Outlook

Stran je prazna, da lahko nanjo rešujete nalogo.

2. 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).

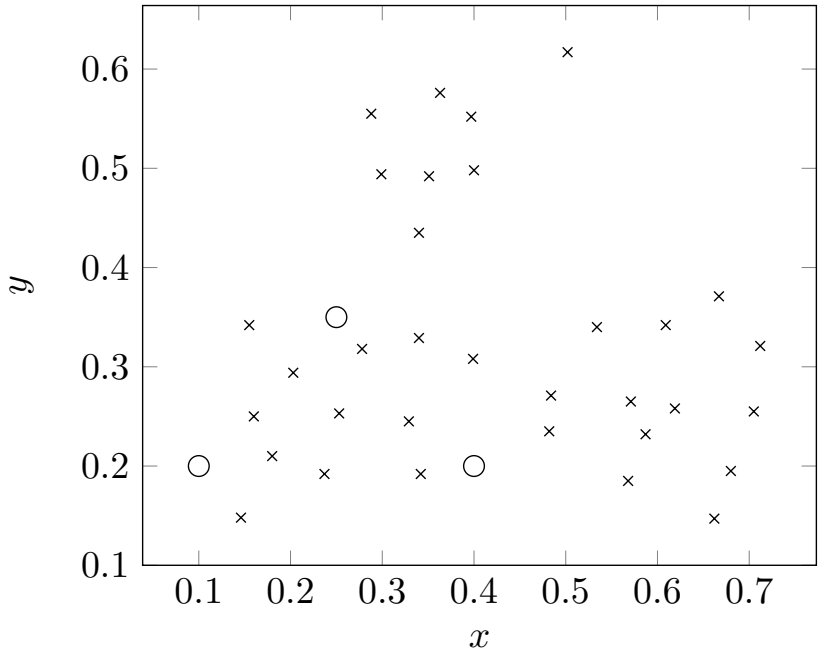
[1] (d) Regularizacijo znižamo na $\lambda = 0.01$. Klasifikacijska točnost na testnih podatkih se izboljša (poveča).

[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] 3. Dani so podatki, ki smo jih izrisali kot točke v evklidskem prostoru (križci). Tri voditelje v tem prostoru smo označili s krogci. Kam se prestavijo voditelji po eni iteraciji tehnike razvrščanja v skupine z metodo voditeljev (angl. *k-means*)? Odgovor utemeljite, tako da jasno opišete oba koraka, ki sta za to potrebna in potrebne podatke za premik voditeljev ustrezno označite na sliki.



4. V matriki ocen $R \in \mathbb{R}^{m \times n}$ vsaka vrstica predstavlja enega od m uporabnikov, vsak stolpec pa enega od n ali 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 $r_{ui} \approx \hat{r}_{ui} = p_u q_i^T$).

[2] (a) Kako znotraj algoritma ISMF 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 ime 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.

5. Podana je **tabela izgub**, ki zajema tri stanja (S_i) in tri alternative (a_j). Tabela vključuje verjetnosti nastopa posameznih stanj.

	Verjetnost	Alternative		
Stanje	$p(S_i)$	a_1	a_2	a_3
S_1	0,2	150	180	130
S_2	0,5	190	160	140
S_3	0,3	120	150	170

[2] (a) Izračunajte pričakovane koristnosti za vse alternative. Za katero alternativo bi se odločili?

[1] (b) Kako bi se odločili po kriteriju optimista, če verjetnosti nastopa posameznih stanj ne bi poznali?

[1] (c) Kako bi se odločili po kriteriju pesimista, če verjetnosti nastopa posameznih stanj ne bi poznali?

[2] (d) Kako bi se odločili po Hurwitzovem kriteriju, če verjetnosti nastopa posameznih stanj ne bi poznali in bi za vrednost koeficienta tveganja d vzeli 0,3?

Solution:

```
a) u(a1) = 0,2 * 150 + 0,5 * 190 + 0,3 * 120 = 161

u(a2) = 0,2 * 180 + 0,5 * 160 + 0,3 * 150 = 161

u(a3) = 0,2 * 130 + 0,5 * 140 + 0,3 * 170 = 147

Odločili bi se za varianto a3, ki ima najmanjšo pričakovano izgubo.

b) Odločili bi se za varianto a1.

c) Odločili bi se za varianto a3.

d) u(a1) = d * min(150;190;120) + (1-d) * max(150;190;120) = 0,3 * 120 + 0,7 * 190 = 169
u(a2) = d * min(180;160;150) + (1-d) * max(180;160;150) = 0,3 * 150 + 0,7 * 180 = 171
u(a3) = d * min(130;140;170) + (1-d) * max(130;140;170) = 0,3 * 130 + 0,7 * 170 = 158

Odločili bi se za varianto a3, ki ima najmanjšo pričakovano izgubo.
```