Poslovna inteligenca

3. izpitni rok

8. september 2017

Priimek in ime (tiskano): _____

Vpisna številka: _____

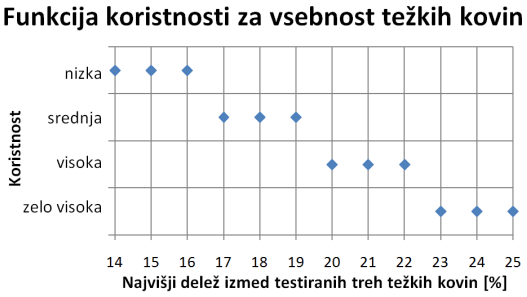
Naloga	1	2	3	4	5	Vsota
Vrednost	6	6	6	7	5	30
Točk						

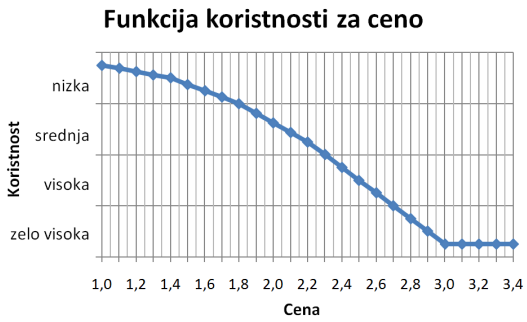
[6] 1. Našli smo test vsebnosti kancerogenih težkih kovin v škatlicah riža s prodajnih polic naših trgovcev. V tabeli so navedene povprečne vrednosti pri treh testiranih vzorcih za izbrane tri vrste riža in največje izmerjene vrednosti med vsemi testiranimi škatlicami riža.

	Vsebnost težkih kovin [ppb]			
Riž	**As**	**Pb**	**Cd**	**Cena za 1kg**
Basmati, beli, Indija	98	2,1	9,3	1,60 €
Jasminov beli, Tajska	116	2,2	4,2	2,20 €
Basmati, beli, Kalifornija	66	1,8	11,6	3,15 €
Največje vrednosti na testih	**568**	**11,7**	**80,3**	

Odločili smo se, da za vsak riž izračunamo delež posamezne težke kovine glede na največje izmerjene vrednosti na testih. Od treh deležev izberemo najmanj ugodnega kot oceno vsebnosti težkih kovin.

Skladno s podanimi funkcijami koristnosti pretvorimo oceno vsebnosti težkih kovin in ceno v razrede.





Oba kriterija združimo v končno oceno po pravilih iz programa DEXi, ki so prikazana v naslednji tabeli:

Odločitvena pravila

	Težke kovine	Cena	Riž
	50%	50%	
1	**zelo visoka**	*	**nespr**
2	*	**zelo visoka**	**nespr**
3	visoka	visoka:srednja	spr
4	visoka:srednja	visoka	spr
5	visoka	***nizka***	dobr
6	srednja	srednja	dobr
7	***nizka***	visoka	dobr
8	>=srednja	***nizka***	***odl***
9	***nizka***	>=srednja	***odl***

Kateri riž bi kupili?

Solution:

	Arzen	Svinec	Kadmij	As	Pb	Cd	
	[ppb]			[%]			max
Basmati bel, Indija	98	2,1	9,3	17,25	17,95	11,58	17,9
Jasminovbeli, Tajska	116	2,2	4,2	20,42	18,80	5,23	20,4
Basmati, beli, Kalifornija	66	1,8	11,6	11,62	15,38	14,45	15,4
max	**568**	**11,7**	**80,3**				

	Težkekovine	Cenana 1kg	**Skupnaocena**
Basmati bel, Indija	srednja	nizka	**odl**
Jasminovbeli, Tajska	visoka	srednja	**spr**
Basmati, beli, Kalifornija	nizka	zelovisoka	**nespr**

Izbrali bi basmati bel riž iz Indije.

Stran je prazna, da lahko nanjo rešujete nalogo.

[6] 2. 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	$\{a, b, c, d, e\}$

Poiščite vsa pravila, ki vključujejo vse izdelke (in nobenih drugih) nabora $\{a, b, d, e\}$ z zaupanjem vsaj 0.7. Pri tem ne računajte zaupanja za pravila, za katera veste, da bo zaupanje premajhno - to jasno označite ter na kratko argumentirajte.

Solution:

```
b e d --> a,  conf: 0.6 STOP
a e d --> b,  conf: 0.6 STOP
a b d --> e,  conf: 1.0
a b e --> d,  conf: 1.0
a b   --> e d conf: 0.75

Na desni sta lahko le e in d! Zato konec.
```

Ocenjevanje 2017: zna izračunati confidence za posamezdni element na desni (2 točki), izračunal je vse confidence, ki jih je moral (1 točka), izračunal je samo tiste, ki jih je bilo treba (3 točke)

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

Stran je prazna, da lahko nanjo rešujete nalogo.

3. 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

4. Kriterijska funkcija, ki jo želimo maksimizirati pri logistični regresiji, je

$$l(\Theta) = \sum_{i=1}^{m} y^{(i)} \log h_\Theta(x^{(i)}) + (1 - y^{(i)}) \log(1 - h(x^{(i)}))$$

kjer je funkcija $h_\Theta(x) = g(\Theta^T x)$ logistična funkcija linearne kombinacije vhodnih spremenljivk (atributov). Z uporabo metode gradientnega spusta lahko izpeljemo pravilo za iterativni popravek i-tega parametra linearne kombinacije:

$$\Theta_j \leftarrow \Theta_j + \alpha \sum_{i=1}^{m} (y^{(i)} - h_\Theta(x^{(i)})) x_j^{(i)}$$

Problem opisanega postopka je preveliko prileganje učnim podatkom. Zato uvedemo regularizacijo.

[1] (a) Kako vpliva regularizacija na vrednost parametrov Θ?

[1] (b) Ali je točna trditev: večja je stopnja regularizacije, manjša je klasifikacijska točnost na učnih podatkih?

[1] (c) Ali je točna trditev: večja je stopnja regularizacije, večja je klasifikacijska točnost na testnih podatkih?

[2] (d) V zgornjo enačbo za kriterijsko funkcijo $l(\Theta)$ dodaj člen z regularizacijo (uporabi tako enačbo oziroma tako regularizacijo, ki jo boš znal odvajati).

[2] (e) Kako se z regularizacijo spremeni enačba za iterativni popravek? Zapiši novo enačbo popravka, ki upošteva regularizacijo. (Ne pričakujemo, da znaš enačbo na pamet. Še najbolj enostavno boš rešitev dobil z odvodom kriterijske funkcije).

Pri odgovorih skušaj upoštevati, da je med parametri Θ parameter Θ_0 uporabljen kot konstantni člen v linearni funkciji h_Θ.

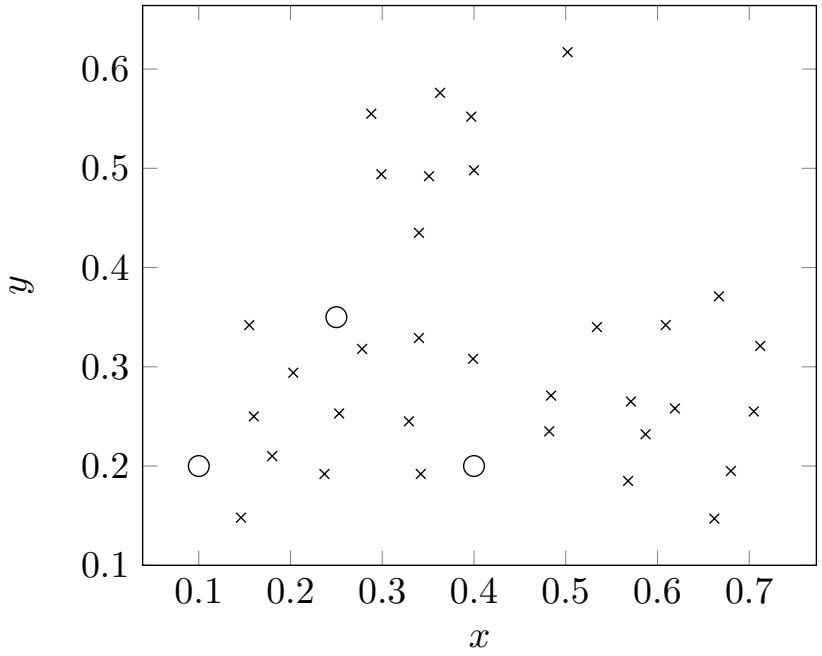
Solution:

- Regularizacija zmanjša vrednosti parametrov, predvsem tistih, ki bi bili brez regularizacije visoki.
- Da.
- Ne.
- Ker $l(\Theta)$ maksimiziramo, želimo pa, da bi bili parametri čim manjši, dodamo člen

$$-\frac{\lambda}{2} \sum_{j=1}^{n} \Theta_j^2$$

- Namesto Θ_j na desni strani imamo $\Theta_j(1 - \alpha\lambda)$

[5] 5. 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.



Stran je prazna, da lahko nanjo rešujete nalogo.

Stran je prazna, da lahko nanjo rešujete nalogo.