Poslovna inteligenca

3. izpitni rok

22. avgust 2016

Priimek in ime (tiskano): ______________________________

Vpisna številka: ______________________________

Naloga	1	2	3	4	5	6	Vsota
Vrednost	7	3	5	6	6	6	33
Točk							

1. 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

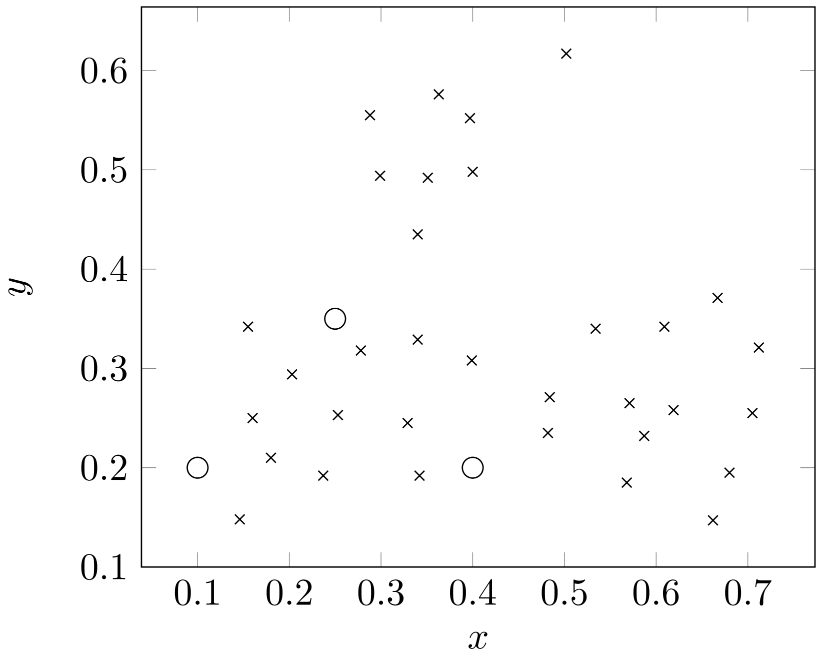
$$-\frac{\lambda}{2} \sum_{j=1}^{n} \Theta_j^2$$

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

[3] 2. Časopisno hišo, ki objavlja novice na spletnih straneh, zanima model, ki bi na podlagi besedila novice ocenil, ali bo ta dobro brana. Za naš pilotni projekt so nam pripravili zbirko 10.000 novic in pri vsaki označili, ali je bila dobro ali slabo brana. Odločili smo se, da bomo za potrebe modeliranja novice predstavili z vektorjem prisotnosti besed. Vseh 10.000 novic skupaj uporablja 13.345 različnih besed. Da bi zadevo poenostavili, smo zato izbrali manjši nabor 1.000 besed tako, da smo vsako predstavili kot atribut (prisotnost besede v novici), stopnjo povezanosti atributa z razredom pa ocenili na podlagi informacijskega prispevka. Izbrali smo 1.000 besed z najvišjim informacijskim prispevkov. Na tako dobljeni podatkovni množici (10.000 novic, vsaka opisana z vektorjem prisotnosti 1.000 besed) smo potem ovrednotili uporabo logistične regresije ter na prečnem preverjanju izmerili AUC, ki je znašal 0.95. Časopisno hišo smo obvestili, da smo na njihovem vzorcu dobili izjemno visoko točnost in da je logistična regresija primerna metoda za gradnjo modelov za napovedovanje branosti novic.

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

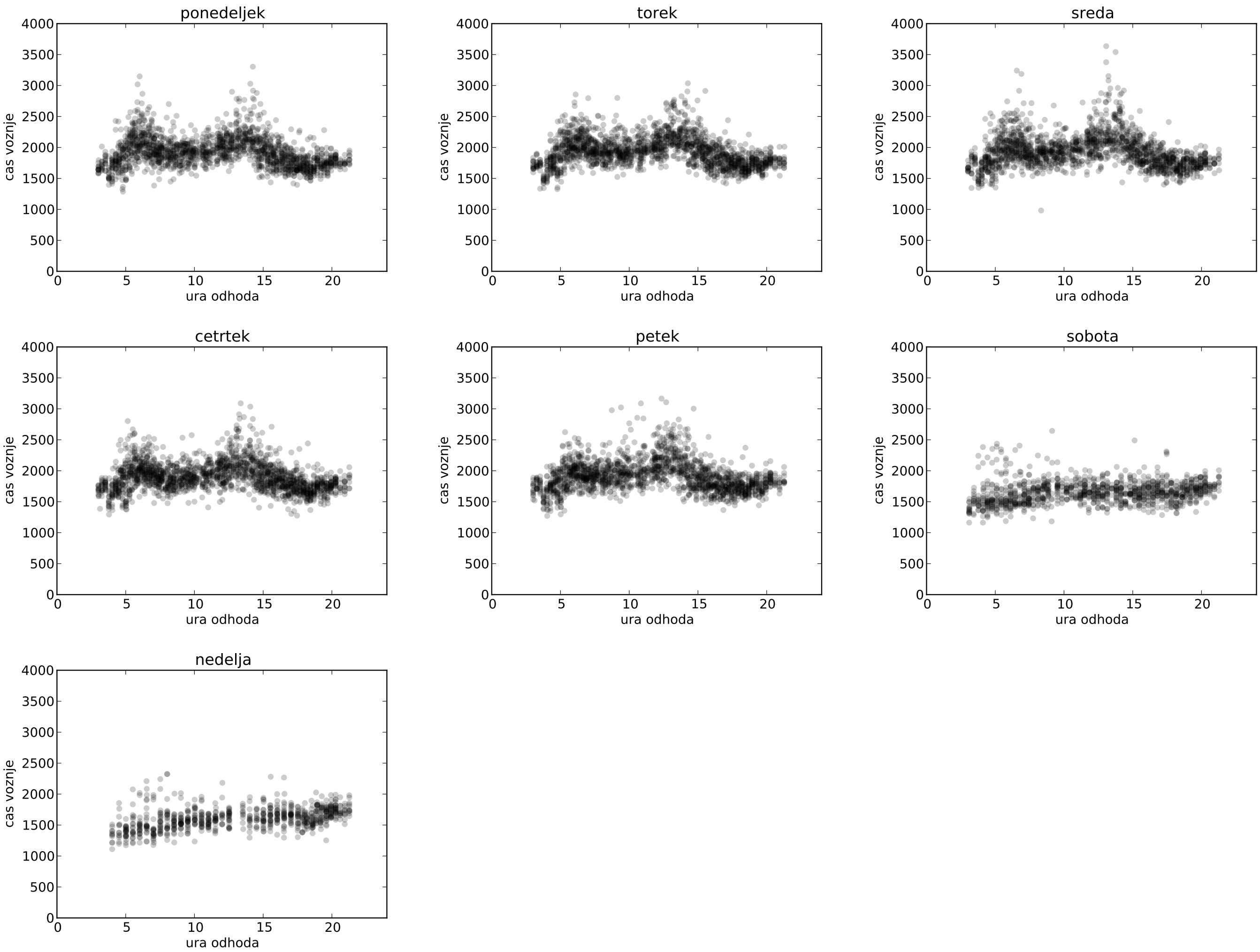
[5] 3. Dani so podatki, ki smo jih izrisali kot točke v evklidskem prostoru (križci). Tri voditelje v tem prostoru smo označili s krogci. Kam se prestavijo voditelji po eni iteraciji tehnike razvrščanja v skupine z metodo voditeljev (angl. *k-means*)? Odgovor utemeljite, tako da jasno opišete oba koraka,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



4. Spodnji razsevni diagrami prikazujejo podatke o vožnjah avtobusa številka 9. Atributa sta dva, dan (ponedeljek, torek, sreda, četrtek, petek, sobota, nedelja) in ura odhoda z začetne postaje, ciljna spremenljivka pa je čas vožnje do končne postaje. Ker iz razsevnih diagramov vidimo, da odvisnosti med uro odhoda (ali dnevom) in časom vožnje niso linearne, želimo uporabiti polinomsko regresijo.

[4] (a) Predlagajte, kako naj predelamo izvorna atributa, da bomo lahko za učenje modela polinomske regresije uporabili knjižnico za linearno regresijo. Vaš predlog tudi utemeljite.

[2] (b) Kako naj predelamo izvorna atributa, da bo knjižnica za linearno regresijo hkrati upoštevala dan in uro in ne zgolj ločeno (kot da bi bila neodvisna) določala uteži zanje?



Solution: Oboje kodiramo kot številke in dodamo potence.

Za hkratno obravnavo še pomnožimo potence.

5. 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.

6. Podana je tabela dobičkov, 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 a1 ali a2, ki imata enakovredno najugodnejšo pričakovano kor

Če omeni samo eno, dam 1.5 točke.

b) Odločili bi se za varianto a1.

c) Odločili bi se za varianto a2.

d) u(a1) = d * max(150;190;120) + (1-d) * min(150;190;120) = 0,3 * 190 + 0,7 * 120 = 14
u(a2) = d * max(180;160;150) + (1-d) * min(180;160;150) = 0,3 * 180 + 0,7 * 150 = 159
u(a3) = d * max(130;140;170) + (1-d) * min(130;140;170) = 0,3 * 170 + 0,7 * 130 = 142
Odločili bi se za varianto a2.

Če je zamešal kam paše d je 169 171 158 -> a2. Tudi priznam.
```