Poslovna inteligenca

3. izpitni rok

31. avgust 2015

Priimek in ime (tiskano): ______________________________

Vpisna številka: ______________________________

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

[6] 1. 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 natanko vse izdelke 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.

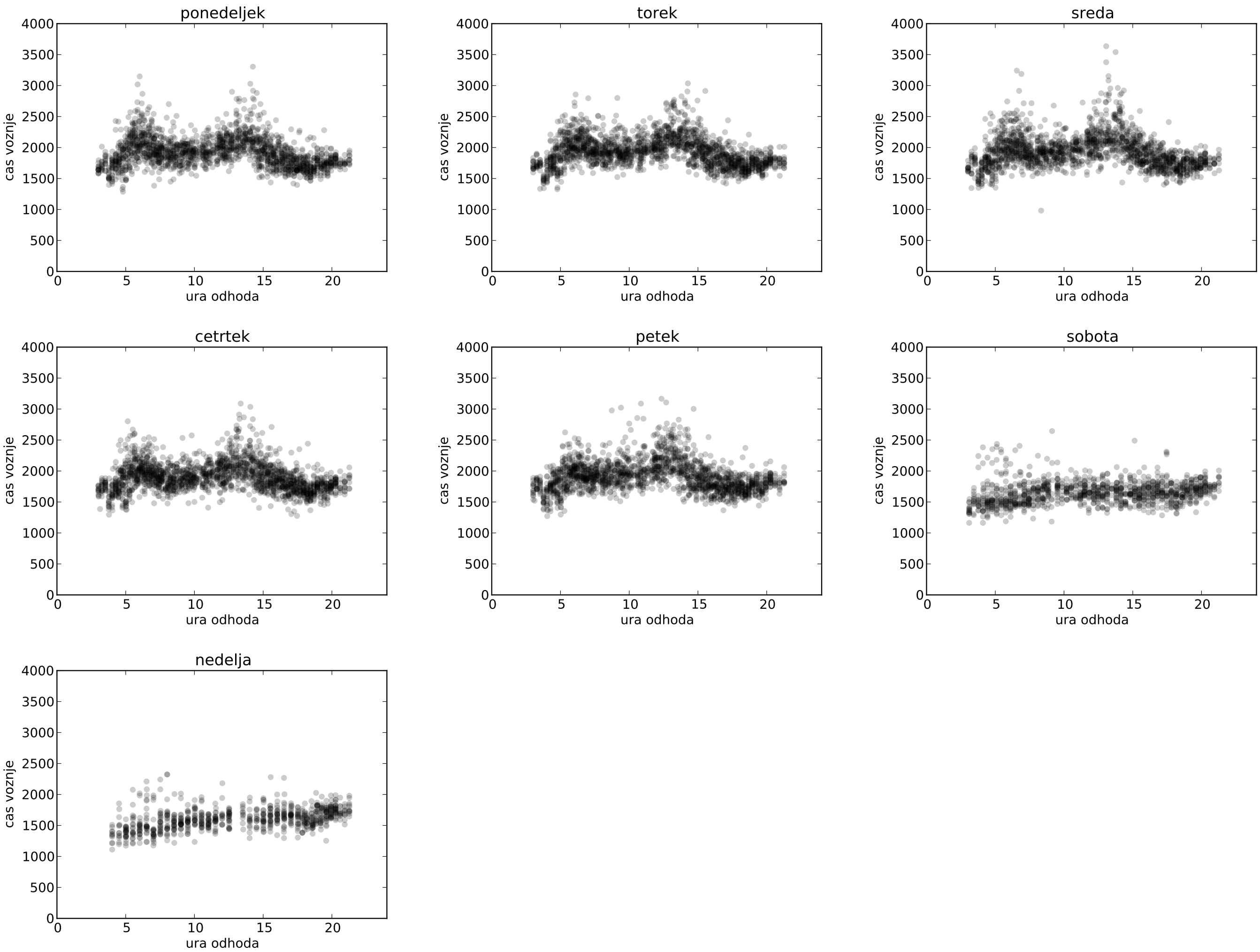
$\sigma(X) = |\{t_i | X \subseteq t_i,\ t_i \in T\}|$ $\quad s(X \rightarrow Y) = \sigma(X \cup Y)/N$ $\quad c(X \rightarrow Y) = \sigma(X \cup Y)/\sigma(X)$

Stran je prazna, da lahko nanjo rešujete nalogo.

2. 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?



3. Kriterijska funkcija, ki jo želimo minimizirati pri linearni regresiji, je

$$\mathrm{J}(\Theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\Theta(x^{(i)}) - y^{(i)})^2$$

kjer je funkcija h_Θ linearna kombinacija vhodnih spremenljivk (atributov). Z uporabo metode gradientnega spusta lahko izpeljemo pravilo za iterativni popravek i-tega parametra linearne kombinacije:

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

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

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

[1] (b) Zakaj bi se tako dobljen model manj prilegal učnim podatkom?

[2] (c) V zgornjo enačbo za kriterijsko funkcijo dodaj člen z regularizacijo.

[2] (d) Kako se z regularizacijo spremeni 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 linerarni funkciji h_Θ.

4. Je čas kosila in odločamo se, kam bi šli jest. V bližini so 4 gostilne: “Pr’ Metki”, “Pr’ Janezu”, “Pr’ Lojzki” in “Pr’ Petru”. Odločili se bomo na podlagi okusa, cene obroka, hitrosti postrežbe in prijaznosti natakarjev.

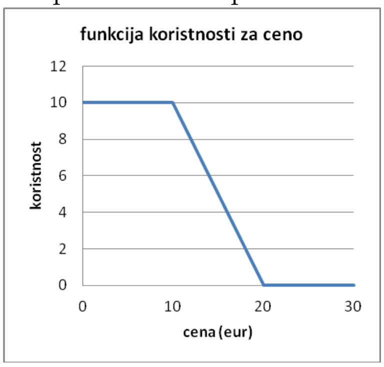
V Gurmanskem vodiču sta okus in prijaznost natakarjev za izbrane gostilne ocenjena takole (z ocenami 1-10):

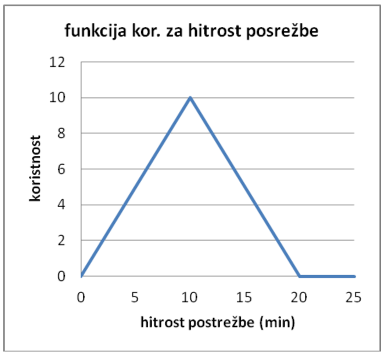
	okus	prijaznost
Pr’ Metki	10	9
Pr’Janezu	6	7
Pr’Lojzki	10	8
Pr’Petru	7	10

Iz preteklih izkušenj vemo, kakšne so cene obroka v posamezni gostilni in koliko časa je treba v povprečju čakati na postrežbo. Podatki so zbrani v spodnji tabeli.

	cena	hitrost
Pr’ Metki	9 EUR	12 min
Pr’Janezu	12 EUR	13 min
Pr’Lojzki	15 EUR	9 min
Pr’Petru	8 EUR	10 min

Za ceno in hitrost postrežbe sta podani naslednji funkciji koristnosti:





Pri odločanju imajo kriteriji naslednje uteži: cena 30, hitrost postrežbe 20, okus 40 in prijaznost natakarjev 10.

[4] (a) Katera gostilna najbolj ustreza našim zahtevam? Kakšen je vrstni red gostiln glede na podani model in ocene?

[2] (b) Poiščite pare pareto-optimalnih in sub-optimalnih gostiln glede na (cena, hitrost).

Stran je prazna, da lahko nanjo rešujete nalogo.

5. V matriki ocen $R \in \mathbb{R}^{m \times n}$ vsaka vrstica predstavlja enega od m uporabnikov, vsak stolpec pa enega od n predmetov (ali izdelkov). Matrika R je redka matrika, kar pomeni, da večina njenih vrednosti ni določenih. Matriko R približno predstavimo z matrikama $P \in \mathbb{R}^{m \times k}$ in $Q \in \mathbb{R}^{k \times n}$ (tako, da je $r_{ui} \approx \hat{r}_{ui} = p_u q_i^T$). Naj bodo konkretne vrednosti teh matrik:

$$P = \begin{bmatrix} 1 & 0 \\ 2 & 2 \\ 2 & 1 \\ 1 & 2 \end{bmatrix}$$

$$Q = \begin{bmatrix} 2 & 0 & 1 & 2 & 1 \\ 1 & 2 & 0 & 2 & 1 \end{bmatrix}$$

[1] (a) Kaj predstavlja matrika P?

[1] (b) Kaj predstavlja matrika Q?

[2] (c) V priporočilnih sistemih $\hat{r}_{ui}$ uporabimo kot napovedano oceno. Izračunajte napovedane ocene za vse uporabnike in vse predmete – torej celo matriko $\hat{R}$.

[2] (d) Algoritem ISMF vsako iteracijo matriki P in Q spremeni tako, da dobimo boljši približek matrike R. Kako merimo kakovost razcepa matrike R v matriki P in Q? Opišite z besedami ali podajte kriterijsko funkcijo.