Poslovna inteligenca

1. izpitni rok

2. februar 2015

Priimek in ime (tiskano): _____

Vpisna številka: _____

Naloga	1	2	3	4	5	Vsota
Vrednost	10	6	5	6	6	33
Točk						

[10] 1. Odločamo se za nakup manjšega rabljenega avta. Pri nakupu si pomagamo z modelom, ki upošteva ceno avta, ceno vzdrževanja, udobje in varnost. Vsak kriterij je lahko ocenjen z vrednostmi "nizek" ali "visok". Naša preferenca je seveda nizka cena ter čim večja stopnjo varnosti in udobja. Uporabimo naslednji hierarhični model:

cena avta	cena vzdrževanja	stroški
nizka	nizka	nizki
nizka	visoka	srednji
visoka	nizka	srednji
visoka	visoka	visoki

udobje	varnost	kvaliteta
nizko	nizka	nizka
nizko	visoka	nizka
visoko	nizka	srednja
visoko	visoka	visoka

Skupna ocena avta je sestavljena iz povprečja stroškov in kvalitete, pri čemer upoštevamo to, da želimo čim nižje stroške in čim višjo kvaliteto. Skupna ocena ima lahko vrednosti "dober", "srednji" ali "slab". Povprečje zaokrožimo navzgor (na boljšo vrednost kriterija).

Ovrednotite naslednje variante.

	cena avta	cena vzdrževanja	udobje	varnost
clio	nizka	nizka	visoko	0.5 niz : 0.5 vis
punto	nizka	visoka	nizko	visoka
mini	0.2 niz : 0.8 vis	visoka	0.6 niz : 0.4 vis	0.4 niz : 0.6 vis
yaris	visoka	0.6 niz : 0.4 vis	visoko	*

Pri neznanih vrednostih (označenih z "*") predpostavite, da so vse vrednosti tega kriterija enako verjetne.

Solution:

Stran je prazna, da lahko nanjo rešujete nalogo.

2. 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.

Solution:

- Regularizacija zmanjša vrednosti parametrov, predvsem tistih, ki bi bili brez regularizacije visoki.
- Manjše vrednosti odvodov, bolj gladka funkcija.
- $$\mathrm{J}(\Theta) = \frac{1}{2m}\sum_{i=1}^{m}(h_\Theta(x^{(i)}) - y^{(i)})^2 + \lambda\sum_{j=1}^{n}\Theta_j^2$$
- $$\Theta_j \leftarrow \Theta_j(1 - \frac{\alpha\lambda}{m}) - \frac{\alpha}{m}\sum_{i=1}^{m}(h_\Theta(x^{(i)}) - y^{(i)})x_j^{(i)}$$

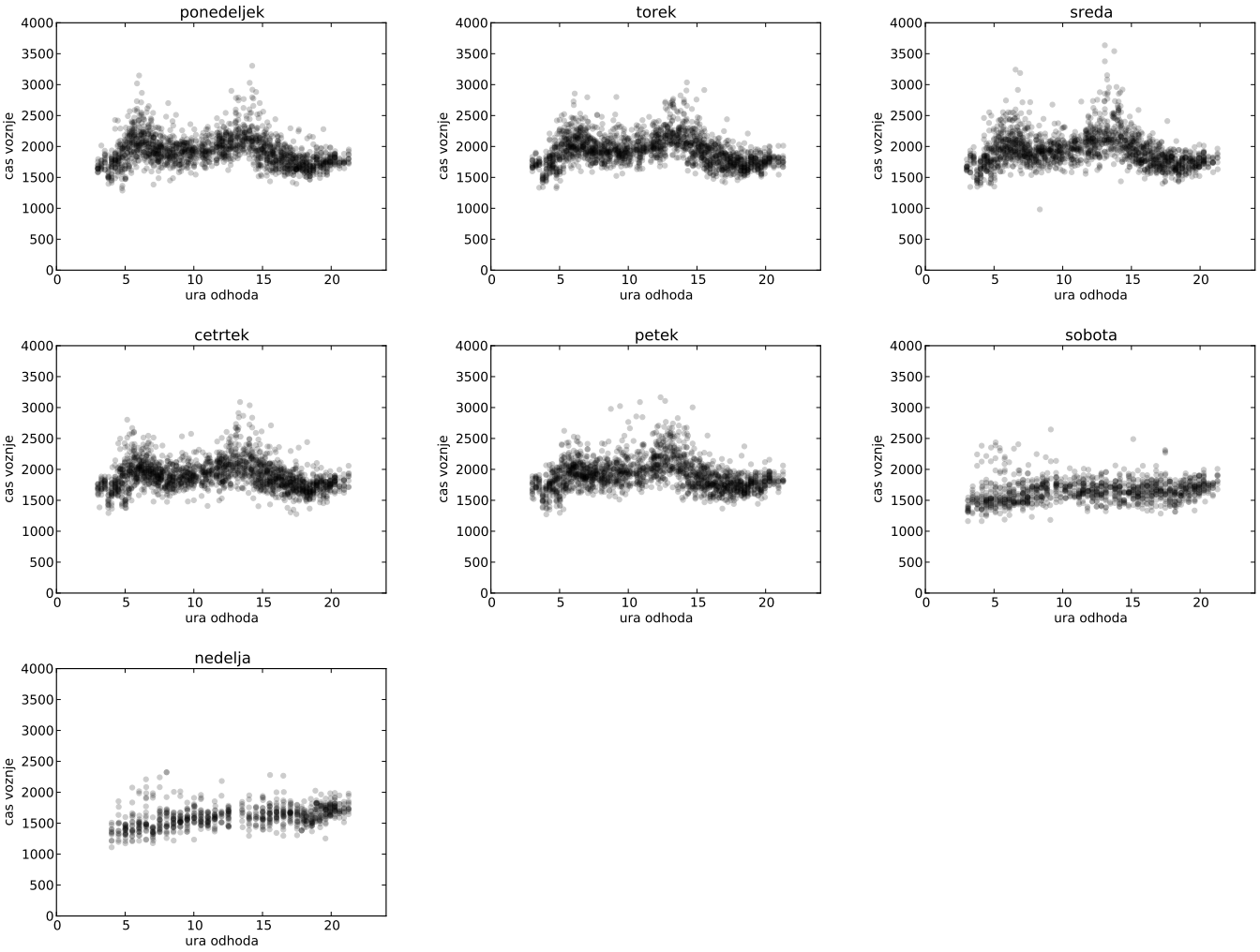
[5] 3. Miha, vzdrževalec portala večjega trgovskega podjetja, želi uporabnikom portala ob nakupovanju ponuditi izdelke, ki bi jih najverjetneje zanimali in bi jih uporabniki lahko dodali v nakupovalno košarico. Ker je pred leti napeto spremljal "NetFlix Prize", se je spomnil na način analize z matrično faktorizacijo. Za analizo ima na voljo bazo s 1500 uporabniki, 1000 izdelki ter 50000 transakcijami. Na spletu najde nek algoritem za matrično faktorizacijo ter ga požene na celotnem naboru podatkov. Poskusi z 10 latentnimi komponentami ($k = 10$), uspešnost postopka pa meri z mero RMSE na celotnem naboru podatkov. Rezultate primerja z algoritmom, ki za danega uporablnika in izdelek oceni verjetnost transakcije iz podatkov 100 najbolj podobnih uporabnikov. Ugotovi, da je RMSE matrične faktorizacije slabši. Zato na enak način izmeri uspešnost postopka za vse k od 1 do 150. Najboljši rezultat je pri $k = 150$, kjer je RMSE veliko nižji kot RMSE algoritma, ki deluje na osnovi ocenjevanja podobnosti uporabnikov. Ta rezultat bo kot dokaz uspešnosti postopka ter način izbora primerne stopnje faktorizacije vključil v poročilo ostalim članom skupine, ki upravlja s portalom.

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

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?



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.