Uvod v odkrivanje znanj iz podatkov (Poslovna inteligenca)

3. izpitni rok

29. avgust 2019

Priimek in ime (tiskano): _____

Vpisna številka: _____

Naloga	1	2	3	4	5	Vsota
Vrednost	6	3	6	5	5	25
Točk						

Izjavljam, da sem nalogo rešil sam, brez kakršnekoli zunanje pomoči in brez uporabe nedovoljenih virov informacij.

Podpis (podpis te izjave je obvezen): _____

[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 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)

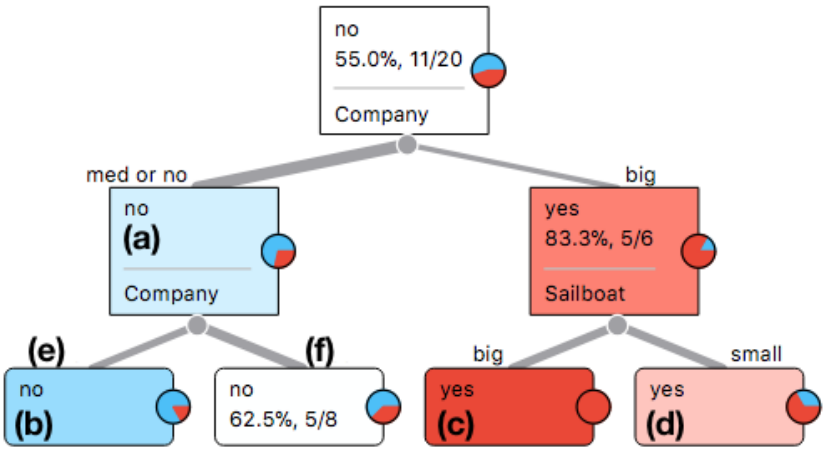
$\sigma(X) = |\{t_i | X \subseteq t_i,\ t_i \in T\}|$ $\quad s(X \to Y) = \sigma(X \cup Y)/N$ $\quad c(X \to Y) = \sigma(X \cup Y)/\sigma(X)$

Stran je prazna, da lahko nanjo rešujete nalogo.

2. Dobili smo spodnje podatke, ki poročajo, ali bo prijateljica šla jadrat glede na podatke o vremenski napovedi, velikosti družbe, ki bi šla zraven in velikosti barke, ki je na voljo. V koloni "Sail" je označeno, ali je na jadranje šla ("yes") ali ne ("no").

	Sail	Outlook	Company ▲	Sailboat
1	yes	rainy	big	big
2	yes	rainy	big	small
5	yes	sunny	big	big
6	yes	sunny	big	small
17	yes	sunny	big	big
18	no	sunny	big	small
3	no	rainy	med	big
4	no	rainy	med	small
7	yes	sunny	med	big
8	yes	sunny	med	big
9	yes	sunny	med	small
12	no	rainy	med	big
19	no	sunny	med	big
20	no	sunny	med	big
10	yes	sunny	no	small
11	no	sunny	no	big
13	no	rainy	no	big
14	no	rainy	no	big
15	no	rainy	no	small
16	no	rainy	no	small

Iz podatkov smo zgradili klasifikacijsko drevo.



Nekaj podatkov v drevesu manjka: označili smo jih z (a) do (f). Dopolni manjkajoče skladno z ostalimi oznakami v drevesu in skladno s podano učno množico. Odgovore zapiši na spodnje alineje:

[½] (a) _____

[½] (b) _____

[½] (c) _____

[½] (d) _____

[½] (e) _____

[½] (f) _____

Solution: **a**: 71.4%, 10/14, **b**: 83.3%, 5/6, **c**: 100%, 3/3, **d**: 66.7%, 2/3, **e**: no, **f**: med.

Stran je prazna, da lahko nanjo rešujete nalogo.

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.

Solution:

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

Stran je prazna, da lahko nanjo rešujete nalogo.

[5] 4. Priporočilni sistem, ki deluje na podlagi matričnega razcepa, smo pognali na podatkih o ocenah knjig. V podatkih se pojavlja 2.000 uporabnikov, ki so z ocenami med 0 in 10 ocenili (nekatere od) 15.000 knjig. Skupno imamo v učni množici 100.000 ocen. Število latentnih faktorjev smo nastavili na $k = 50$, regularizacije pa ne uporabljamo.

Skicirajte koren srednje kvadratne napake (RMSE) v odvisnosti od števila iteracij gradientnega sestopa: narišite koordinatni sistem s številom iteracij na osi x in RMSE na osi y ter vanj vrišite dve krivulji, eno za RMSE na učnih, drugo za RMSE na testnih podatkih. Vaš graf seveda ne more biti natančen, a iz njega naj bo razvidna razlika med rezultati učne in testne množice.

Solution:

Krivulji začneta zelo blizu. Na obeh RMSE na začetku pada, vendar na učni množici hitreje kot na testni. RMSE na učni množici vedno pada (sicer vedno počasneje), RMSE na testni množici pa nekje začne naraščati.

[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.