

Poslovna inteligenca

1. izpitni rok

2. februar 2017

Priimek in ime (tiskano): _____

Vpisna številka: _____

Naloga	1	2	3	4	5	Vsota
Vrednost	7	5	6	7	7	32
Točk						

1. Spodnja tabela podaja učno množico, kjer je x_1 vhodna spremenljivka oziroma atribut, y pa razred, ki bi ga želeli napovedati.

x_1	y
1	3
2	3
3	5
4	5

- [1] (a) Odvisnost med x_1 in y prikaži v razsevnem diagram (nariši, lično inženirsko, z ravnilom, ne skiciraj).
- [1] (b) Predlagaj linearni model povezave med y in x_1 . Model zapiši kot enačbo in ga grafično (z ustrežno premico) predstavi v razsevnem diagramu.
- [1] (c) Napiši splošno enačbo kriterijske funkcije $J(\Theta)$, s katero lahko oceniš kvaliteto tvojega modela pri danih parametrih modela.
- [1] (d) Model, ki si ga predlagal, ovrednoti. V zgornji tabeli dodaj kolono za \hat{y} (napoved modela) in ϵ (napako napovedi), napake grafično predstavi v razsevnem diagramu (z ustreznimi črtami, ki označujejo velikost napake) in izračunaj vrednost kriterijske funkcije.
- [1] (e) Predlagaj slabši model, torej tak, katerega vrednost kriterijske funkcije je večja (slabša) od zgoraj predlaganega modela. Tudi tokrat podaj enačbo modela, izračunaj napovedi, napake in vrednost kriterijske funkcije, model pa grafično predstavi v razsevnem diagramu.
- [1] (f) V enačbo za kriterijsko funkcijo dodaj regularizacijo in parameter regularizacije označi z λ .
- [1] (g) Kakšen je model, ki se najbolj prilega učnim podatkom pri maksimalni regularizaciji. Model zapiši z enačbo in ga grafično predstavi v razsevnem diagramu.

Solution:

- razsevni diagram s štirimi točkami
- primer modela je $y = 2 + 0.8 \times x_1$
- $J(\Theta) = \sum_{i=1}^4 (\theta_0 + \theta_1 x_1^{(i)} - y^{(i)})^2 = \sum_{i=1}^4 (2 + 0.8x_1^{(i)} - y^{(i)})^2$
- regularizacija: doda se člen $\lambda \times \theta_1^2$, saj θ_0 ne regulariziramo
- maksimalna regularizacija, $\hat{y} = \bar{y} = 4$

Stran je prazna, da lahko nanjo rešujete nalogo.

2. Pridobili smo manjši vzorec podatkov o naročnikih telekomunikacijskih uslug in njihovih morebitnih odpovedih naročniškega razmerja. Naročniki so opisani s 150 binarnimi atributi, podatke pa imamo za 40 naročnikov, od katerih je 8 odpovedalo naročniško razmerje. Podatke naključno razdelimo na dve enako veliki množici z enako porazdelitvijo razredov, ter tako dobimo učno množico U in testno množico T . Stolpec z razredno spremenljivko v množici U naključno premešamo in tako dobimo množico U' .

- [1] (a) Na množici U' preverjamo točnost klasifikacijskih dreves, tako da na celotni množici U' zgradimo model in na njej točnost tudi ocenimo. Ocenjena klasifikacijska točnost je 0.95. Zakaj je ta tako visoka, kljub temu, da smo razrede naključno premešali?
- [1] (b) Kaj bi lahko bil razlog, da zgoraj ocenjena točnost ni enaka 1.0?
- [1] (c) Klasifikacijsko drevo, ki smo ga zgradili na U' , testiramo na T . Kakšno klasifikacijsko točnost pričakujemo?
- [1] (d) Ko klasifikacijsko točnost drevesa, zgrajenega na U , preverjamo na T , dobimo 0.9. Kaj lahko rečeš o velikosti drevesa, zgrajenega U in tistega, ki je zgrajen na U' - katero drevo pričakuješ, da je večje? Zakaj?
- [1] (e) Zakaj smo pri zgornji nalogi omenili, da je točnost na T enaka 0.9? Bi se kaj v tvojem odgovoru spremenilo, če bi zapisali, da je točnost na T enaka 0.8?

Solution:

- Testiramo na učni množici, drevo si je zapomnilo vse primere.
- Dva primera imata enako vrednost atributov a različen razred.
- 0.8.
- Drevo zgrajeno na U' je večje.
- 0.8 je točnost napovedi z večinskim razredom. Podatki za tako točnost so lahko tudi naključni, in zato bi tam drevo lahko bilo večje.

[6] 3. 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 posamezni 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)

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

4. Dani so podatki v enodimenzionalnem prostoru.



- [1] (a) Za centroida $\{18, 45\}$ ustvari skupini tako, da vsako točko prirediš najbližjemu centroidu. Definiraj skupino in izračunaj vsoto kvadratov napak (SSE, sum of squared errors) za obe skupini.
- [1] (b) Za centroida $\{15, 40\}$ ustvari skupini tako, da vsako točko prirediš najbližjemu centroidu. Definiraj skupino in izračunaj vsoto kvadratov napak za obe skupini.
- [2] (c) Za para centroidov iz podnaloge (a) in (b) simuliraj metodo voditeljev do konvergence. Dobiš v obeh primerih enake rezultate?
- [2] (d) Kateri skupini dobimo pri hierarhičnem razvrščanju z metodo "single linkage"?
- [1] (e) Katera od obeh metod (pri metodi voditeljev vzemite par začetnih voditeljev, ki vodi do manjšega SSE) pa vodi do "bolj naravnih" skupin v tem primeru?

5. Prijatelji se odločajo kam na morje za poletne počitnice. Izbirajo med 4-imi variantami, ki so jih ocenili na absolutni skali med 0 in 100:

	Kreta	Malta	Lošinj	Pag
Mojca	57	81	58	57
Maja	42	32	56	39
Jernej	63	62	61	61
Anže	77	60	59	45

- [2] (a) Poiščite pare pareto-optimalnih in sub-optimalnih destinacij.
 [1] (b) Kam bi potovali, če bi se odločalo po metodi Harsany-ja?
 [1] (c) Katera destinacija bi bila najbolj primerna po metodi Nash-a?
 [2] (d) Kaj je prednost Nash-eve tehnike napram Harsany-jevi?
 [1] (e) Če bi dekleti imeli vsaka po 30% utež, fanta pa po 20%, katera destinacija bi bila zmagovalna?

Solution: a) Pag je pareto-sub-optimalen glede na Kreta in Lošinj.

b) Kreta

c) Lošinj

d) Produkt vs seštevek. Zelo slaba ocena enega praktično prepreči izbiro te variante.

(Ocenjevanje 2017: če omeni le kako žrtvovanje, dam 1 točko in označim, da je to nejasno)

e) Malta

weigth Kreta Malta Lošinj Pag

30 Mojca 57 81 58 57

30 Maja 42 32 56 39

20 Jernej 63 62 61 61

20 Anže 77 60 59 45

Harsany 239 235 234 202

Nash 11.613.294,00

9.642.240,00

11.689.552,00

6.102.135,00

Uteži 5770 5830 5820 5000

Stran je prazna, da lahko nanjo rešujete nalogo.

Stran je prazna, da lahko nanjo rešujete nalogo.

Stran je prazna, da lahko nanjo rešujete nalogo.