Poslovna inteligenca

2. izpitni rok

17. februar 2017

Priimek in ime (tiskano): _____

Vpisna številka: _____

Naloga	1	2	3	4	5	Vsota
Vrednost	5	5	6	6	5	27
Točk						

1. Dana je funkcija $y(\theta_0, \theta_1) = (\theta_0 - 3)^2 + (\theta_1 - 5)^2$.

[1] (a) Izračunaj gradient funkcije $y(\theta_0, \theta_1)$ (podaj enačbo).

[1] (b) Izračunaj gradient funkcije $y(\theta_0, \theta_1)$ v točki $[1, 1]^T$ (podaj številsko vrednost gradienta).

[2] (c) Z gradientnim sestopom iščemo vrednosti parametrov funkcije $y(\theta_0, \theta_1)$, pri katerih ima ta funkcija minimum. Začetne vrednosti parametrov nastavimo na $[\theta_0, \theta_1]^T = [1, 1]^T$. Stopnjo učenja nastavimo na 0.1. Kakšna je vrednost parametrov po prvem koraku gradientnega sestopa, torej po tem, ko z gradientnim sestopom prvič osvežimo vrednost parametrov.

[1] (d) Kakšna je vrednost parametrov po drugem koraku gradientnega sestopa?

Solution: $2(\theta_0 - 3), 2(\theta_0 - 5)$, $[-4, -8]^T$, $[1.4, 1.8]^T$, gradient $[-3.2, -6.4]^T$ parametri pa $[1.72, 2.44]^T$

[5] 2. Metka sodeluje s kadrovskim podjetjem. Tam z 42 atributi opišejo zaposlene in pogoje v firmi, kjer so zaposleni. Cilj je napovedati, ali bodo zaposleni v firmi zdržali več kot pol leta. Za to analizo so zbrali zgodovinske podatke o 1500 zaposlenih, med katerimi je prej kot v pol leta po zaposlitvi dalo odpoved 450 anketiranih, 1050 pa jih je zaposlitev obdržalo oziroma so v firmi ostali dlje kot pol leta.

Metka, ki je pri analizi sodelovala še z Alešem in Majo, je pri študiji naredila kar nekaj poskusov in si vestno beležila klasifikacijske točnosti. Različne vrednosti klasifikacijskih točnosti si je zapisala na listke:

Listek A) 100%
Listek B) 95%
Listek C) 70%
Listek D) 58%
Listek E) 20%
Listek F) 0%

Na drugih listkih je opisala poskuse, listke pa označila s številkami:

1. Deset-kratno prečno preverjanje z logistično regresijo. Prvi poskusi kažejo na obetavne rezultate.
2. Kolono z razredom naključno premešam (za vse podatke). Tako spremenjen nabor podatkov dam Alešu in Maji (oba torej prejmeta isti nabor podatkov). Aleš zgradi model z metodo naključnega gozda, s tem modelom pa pomaga Maji klasificirati primere iz njenega nabora podatkov.
3. Izberem naključnih 65% primerov podatkov. Te dam Alešu. Vse ostale podatke dam Maji. Alešu naročim, da vrednost razreda napove tako, da je njegova napoved vedno en sam razred, ki je večinski razred v njegovih podatkih. S takimi napovedmi Aleš pomaga Maji klasificirati primere iz njenega nabora podatkov.
4. Izberem naključnih 50% primerov. Te dam Alešu, ostalim primerom pa pomešam kolono z vrednostjo razredne spremenljivke in jih dam Maji. Aleš na podatkih zgradi model z metodo naključnih gozdov. Ta model potem uporabi pri napovedovanju razredov za primere iz Majinega nabora.

Listek, kjer je za vsak opis poskusa (številka) zapisala tudi ustrezno oznako rezultato oziroma točnosti (črka) je zgubila. Pomagaj! Za vsak od zgornjih poskusov povej, kakšno točnost je Metka najverjetneje dobila. (Naloga je enostavna za 1, 2 in 3, za 4 pa bo potrebno malo bolj globoko premisliti. Za pravilen rezultat pri poskusu 4 dobite točko, drugo točko pa, če pravilno pokažete, kako ste pri tem poskusu ocenjeno klasifikacijsko točnost izračunali).

Solution: 1B, 2A, 3C, 4D (.7*.7+.3*.3)

3. V Pythonu smo v nekem programu zapisali spodnjo funkcijo:

```
def j(theta, x, y, reg=0.1):
    return -(y.dot(np.log(h(theta, x))) + (1-y).dot(np.log(1-h(theta, x))))
```

[1] (a) Pri kateri tehniki analize podatkov smo to funkcijo uporabljali?

[2] (b) Kaj ta funkcija počne (kaj so vhodni podatki in kaj je izhod)?

[3] (c) Funkcija že vključuje argument za stopnjo regularizacije, a ga v funkciji nismo uporabili. Dopolni funkcijo tako, da dodaš člen z regularizacijo.

Solution:

```
    return -(y.dot(np.log(h(theta, x))) + (1-y).dot(np.log(1-h(theta, x))) -
             reg * sum(theta[1:]**2))
```

4. V kadrovski službi so pripravili drevo kriterijev za izbor novega sodelavca. Del drevesne structure je prikazan v naslednji tabeli. Vsi trije kriteriji imajo po tri zaloge vrednosti: manj primeren, primeren, zelo primeren.

Kriterij	Zaloga vrednosti
Kandidat	manj prim.; prim.; *z_prim.*
├─Os.lastn.	manj prim.; prim.; *z_prim.*
└─Znanja	manj prim.; prim.; *z_prim.*

Priložili so tudi del agregiranih pravil funkcije koristnosti za kriterij Kandidat, ki žal ni popoln, saj manjka nekaj vrstic. Ne vemo, katere vrstice manjkajo, niti koliko jih manjka.

Os.lastn.	Znanja	Kandidat
50%	50%	
manj prim.	*	manj prim.
prim.	>=prim.	prim.
z_prim.	*z_prim.*	*z_prim.*

[4] (a) Izpolnite tabelo osnovnih pravil funkcije koristnosti za kriterij Kandidat:

Os.lastn.	Znanja	Kandidat
manj prim.	manj prim.	
manj prim.	prim.	
manj prim.	*z_prim.*	
prim.	manj prim.	
prim.	prim.	
prim.	*z_prim.*	
z_prim.	manj prim.	
z_prim.	prim.	
z_prim.	*z_prim.*	

[2] (b) Izračunajte končno oceno za naslednja kandidata. Katerega kandidata bi izbrali? Utemeljite svoj odgovor.

Kandidat	Os.lastn.	Znanja	Kandidat
A	*z_prim.*	manj prim.	
B	prim.	manj prim.	

Solution:

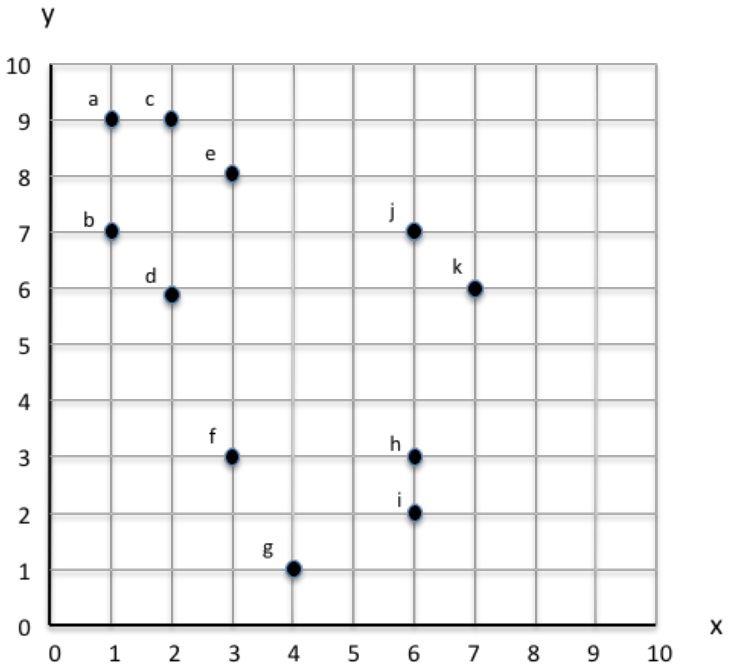
	Os.lastn.	Znanja	Kandidat
1	**manj prim.**	**manj prim.**	**manj prim.**
2	**manj prim.**	prim.	**manj prim.**
3	**manj prim.**	***z_prim.***	**manj prim.**
4	prim.	**manj prim.**	**manj prim.**
5	prim.	prim.	prim.
6	prim.	***z_prim.***	prim.
7	***z_prim.***	**manj prim.**	**manj prim.**
8	***z_prim.***	prim.	prim.
9	***z_prim.***	***z_prim.***	***z_prim.***

1 točka za pravila 1, 2, 3, 5, 6 in 9
3 točke za pravila 4, 7 in 8

Kandidat	Os.lastn.	Znanja	Kandidat
A	***z_prim.***	**manj prim.**	**manj prim.**
B	prim.	**manj prim.**	**manj prim.**

Čeprav sta oba kandidata dosegla enaka končno oceno, je primernejši kandidat A, ki ima boljše osebne lastnosti, v ničemer pa ni slabši od kandidata B.
1 točka za pravilen odgovor
1 točka za utemeljitev

5. Dana je spodnja množica učnih primerov, ki smo jih opisali z dvema zveznima atributoma x in y in jih lahko predstavimo kot točke v Evklidski ravnini:



[4] (a) Izriši dendrogram, ki ga dobiš z hierarhičnim razvrščanjem točk v skupine. Kot mero za podobnost uporabi Manhattansko razdaljo, kjer je razdalja med primeroma i in j določena kot $d_{ij} = |x_i - x_j| + |y_i - y_j|$. Podobnost med dvema skupinama meri s tehniko maksimalne razdalje med paroma točk iz različnih skupin (t. im. *complete linkage*).

[1] (b) Uporabi izrisani dendrogram in na podlagi njega predlagaj razdelitev primerov v tri skupine (na dendrogramu izriši vertikalo, ki točke razdeli v tri skupine). Izpiši, kateri primeri pripadajo posamezni skupini.

Solution:

```
ac e bd | fg hi | jk
1     2    3 1    2
  3
    4        4
               8
       11
```

Stran je prazna, da lahko nanjo rešujete nalogo.

Stran je prazna, da lahko nanjo rešujete nalogo.