Uvod v odkrivanje znanj iz podatkov (Poslovna inteligenca)

2. izpitni rok

13. februar 2019

Priimek in ime (tiskano): _____

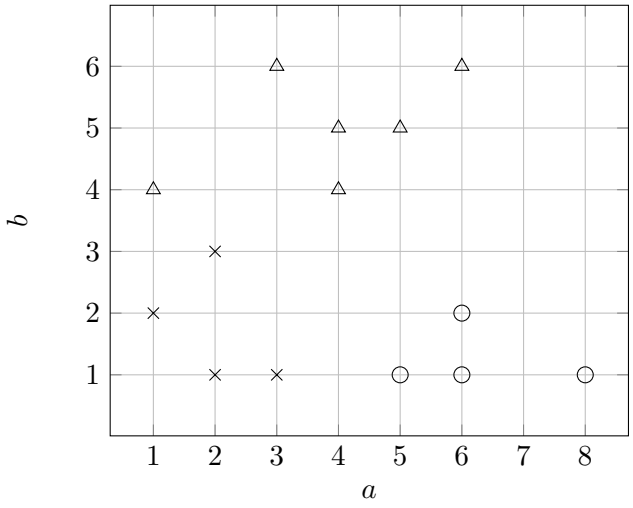
Vpisna številka: _____

Naloga	1	2	3	4	5	Vsota
Vrednost	8	6	3	6	7	30
Točk						

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

Podpis (podpis te izjave je obvezen): _____

1. Podatke, ki vključujejo 14 primerov in so opisani z atributoma a in b, prikazuje spodnje slika. Primeri so razvrščeni v tri skupine, katerih pripadnost za posamezne primere smo na sliki prikazali z simbolom, s katerim je označen primer (križec, trikotnik, krogec). Primer $(a, b) = (3, 1)$ tako na primer pripada skupini križcev, v katerih so skupaj štirje primeri.



Odločimo se, da bomo razdaljo med primeri meri z Manhattansko razdaljo.

[1] (a) Naj bosta x_1 in x_2 dva primera, vrednosti njunih atributov pa označimo z $x_{1,a}$ in $x_{1,b}$ za prvi primer in podobno za drugega. Zapiši enačbo, s katero izračunamo Manhattansko razdaljo med primeroma x_1 in x_2.

[1] (b) Kakšna je Manhattanska razdalja med primeroma $(2, 3)$ in $(6, 1)$?

[1] (c) Kakšna je vrednost silhuete za primer $(5, 1)$?

[2] (d) Kateri izmed primerov iz skupine trikotnikov, to je primeri, ki so na sliki označeni s tem likom, ima najmanjšo silhueto. Kolikšna je ta?

[1] (e) Kaj merimo s silhueto in kakšna je njena zaloga vrednosti?

[1] (f) Kaj lahko rečemo o razvrstitvi primera z negativno vrednostjo silhuete?

[1] (g) Za izračun silhuete uporabimo isto mero razdalj, kot smo jo uporabili pri razvrščanju v skupine. Kakšno zalogo vrednosti pričakujemo za to silhueto? Je ta zaloga vrednosti enaka kot v prejšnjem vprašanju? Zakaj?

$s_i = \frac{b_i - a_i}{\max(a_i, b_i)}$

Solution: a) $d(x_1, x_2) = |x_{1,a} - x_{2,a}| + |x_{1,b} - x_{2,b}|$
b) 6
c) $a = (1 + 2 + 3) = 6/3 = 2$, $b = (2 + 3 + 5 + 5)/4 = 15/4 = 3.75$, $s = 1.75/3.75 = 0.46$
d) primer $(1, 4)$, $a = (3 + 4 + 4 + 5 + 7) = 23/5 = 4.6$, $b = (2 + 2 + 4 + 5) = 13/4 = 3.25$, $s = (3.25 - 4.6)/4.6 = -0.29$
e) centralnost primerov v skupini glede na najbližjo tujo skupino, $[-1, 1]$
f) primer bi moral biti razvrščen v drugo skupino
g) $[-1, 1]$, $[0, 1]$. Ni, pričakujemo, da bodo primeri najbližji skupini, kamor so razvrščeni.

(prostor za rešitve)

2. 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 predstavljajo argumenti te funkcije in kaj ta funkcija vrača? Za argumente funkcije in vrnjene vrednosti napiši tudi, ali gre za skalar, vektor, ali matriko.

[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))
```

[3] 3. Na nekih podatkih z dvema razredoma smo z Naivnim Bayesom zgradili model. Model nam za vsak primer vrne verjetnost p_1, da ta primer pripada prvemu razredu. S privzeto mejo za ločevanje med razredoma 0.5 (če je verjetnost večja kot 0.5 vrni prvi razred) smo izmerili AUC 0.8 in klasifikacijsko točnost 0.6. Nato smo ugotovili, da se klasifikacijska točnost poveča na 0.7, če mejo prestavimo na 0.3. Kakšen bo AUC pri meji 0.3? Če je sprememba odvisna od nabora podatkov, razloži vse možnosti.

4. Dana je funkcija $y(\theta_1, \theta_2) = 2\theta_1^2 + (3 - \theta_2\theta_1)^2$.

[1] (a) Izpelji gradient funkcije $y(\theta_1, \theta_2)$ (podaj enačbo gradienta).

[1] (b) Izračunaj gradient funkcije $y(\theta_1, \theta_2)$ v točki $[\theta_1, \theta_2]^T = [1, 2]^T$.

[1] (c) Preveri zgoraj izračunano vrednost gradienta tako, da zanj izračunaš numerični približek. Pri tem uporabi samo enačbo funkcije in ne izpeljano enačbo gradienta. Dober približek na primer dobiš z uporabo $\epsilon = \pm 0.01$.

[2] (d) Z gradientnim sestopom iščemo vrednosti parametrov funkcije $y(\theta_1, \theta_2)$, pri katerih ima ta funkcija minimum. Začetne vrednosti parametrov nastavimo na $[\theta_1, \theta_2]^T = [1, 2]^T$. Uporabimo stopnjo učenja 0.1. Kakšna je vrednost parametrov po prvem koraku gradientnega sestopa, torej po tem, ko z gradientnim sestopom prvič osvežimo vrednost parametrov.

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

Solution:
$[4\theta_1 - 2(3 - \theta_1\theta_2)\theta_2, -2(3 - \theta_2\theta_1)\theta_1]$
$[0, -2]^T$
enako
po prvem koraku $[1, 2.2]$,
po drugem koraku $[0.952, 2.36]$

(prostor za rešitve)

5. 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 ustrezno 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, $\hat{y} = \bar{y} = 4$

(prostor za rešitve)