Uvod v odkrivanje znanj iz podatkov (Poslovna inteligenca)

1. izpitni rok

31. januar 2020

Priimek in ime (tiskano): ______________________________

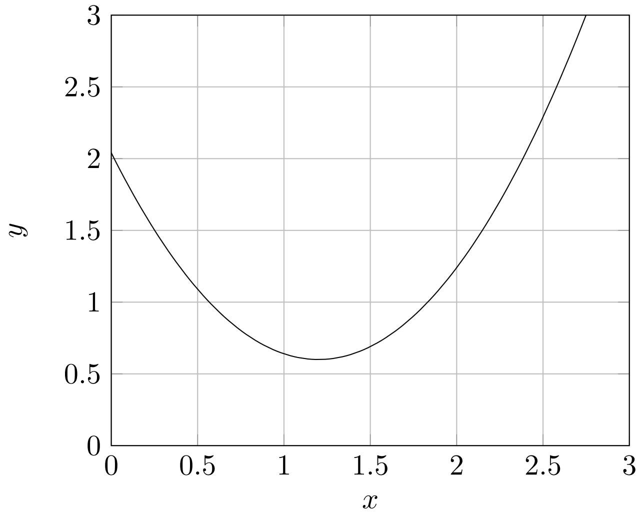
Vpisna številka: ______________________________

Naloga	1	2	3	4	5	Vsota
Vrednost	6	5	5	5	5	26
Točk						

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

Podpis (podpis te izjave je obvezen): ______________________________

1. Podana je funkcija $y = f(x)$ kot jo prikazuje spodnji graf.



[1] (a) Oceni vrednost gradienta funkcije v točki $x = 2$.

[1] (b) Z gradientnim sestopom bi želeli poiskati minimum funkcije. Trenutno rešitev zapišimo z x_i. Zapiši enačbo za izračun naslednje vrednosti približka rešitve x_{i+1}, v kateri naj bo λ stopnja učenja.

[1] (c) Gradientni sestop prični v točki $x_0 = 2$. Stopnja učenja naj bo $\lambda = 0.25$. Izračunaj x_1.

[2] (d) Nadaljuj postopek iz zgornje točke in izračunaj x_2.

[1] (e) Kaj bi se zgodilo, če bi za stopnjo učenja pri zgornjem postopku in začetni točki $x_0 = 2$ uporabil $\lambda = 2$?

Solution: $y = (x - 1.2)^2 + 0.6$,
1) 1.6
2) $x_{i+1} = x_1 - \lambda f'(x_i)$
3) $x_1 = 2 - 0.4 = 1.6$
4) $y'(1.6) = 0.8$, $x_2 = 1.6 - 0.25 * 0.8 = 1.4$
5) gradientni sestop ne bi skonvergiral v rešitev

Stran je prazna, da lahko nanjo rešujete nalogo.

2. Dana je spodnja množica primerov, opisanih z atributima x_1 in x_2.

x_1	x_2
1	3
2	2
3	2
4	4
5	4
7	6
6	7

Dopolni zgornjo tabelo tako, da vsakemu primeru pripišeš koordinati v novem koordinatnem sistemu, ki ga določajo glavne komponente (PCA) podatkov. Smeri glavnih komponent oceni (ne računaj) tako, da

[1] (a) točke predstaviš grafično v originalnem koordinatnem sistemu,

[2] (b) oceniš smeri glavnih koordinat in izrišeš nov koordinatni sistem s prvo in drugo glavno komponento

[2] (c) ter jasno označiš projekcije točk iz originalnega koordinatnega sistema v nov koordinatni sistem.

Namig: premisli, kje je koordinatno izhodišče novega koordinatnega sistema.

Solution:

t_1	t_2
-2.9	1.2
-2.8	-0.2
-2.1	-0.9
0.0	0.0
0.8	-0.7
3.6	-0.5
3.5	1.0

Stran je prazna, da lahko nanjo rešujete nalogo.

3. V spodnji tabeli je dana učna množica s tremi atributi (Outlook, Company, Sailboat) in razredno spremenljivko (Sail).

	Outlook	Company	Sailboat	Sail
1	sunny	big	small	yes
2	sunny	med	small	yes
3	sunny	med	big	yes
4	sunny	no	small	yes
5	sunny	big	big	yes
6	rainy	no	small	no
7	rainy	med	small	yes
8	rainy	big	big	yes
9	rainy	no	big	no
10	rainy	med	big	no

[1] (a) Kakšna je stopnja nedoločenosti (entropija) razredne spremenljivke Sail?

[2] (b) Kakšen je informacijski prispevek atributa Outlook?
(Namig: najprej izračunaj residualno entropijo $H(\text{Sail}|\text{Outlook})$, potem pa še informacijski prispevek. Residualna entropija ti pove, koliko znaša entropija razreda pri tem, če veš, kakšno vrednost je zavzela spremenljivka Outlook.)

[2] (c) Informacijski prispevek atributa Sailboat je $0,035$, atributa Company pa $0,281$. Kateri atribut bi postopek gradnje odločitventega drevesa postavil v koren drevesa? Naj bo to tudi edino notranje vozlišče drevesa (vse ostalo so listi). Skiciraj, kako izgleda tako odločitveno drevo.

$$H(x) = -\sum_{i \in D(x)} p(x=i) \log_2 p(x=i)$$

$$H(y|x) = \sum_{i \in D(x)} p(x=i) H(y|x=i)$$

$$IG(x;y) = H(y) - H(y|x)$$

kjer so x spremenljivka (razred ali atribut), $D(x)$ zaloga vrednosti spremenljivke x, in y razredna spremenljivka

Solution:

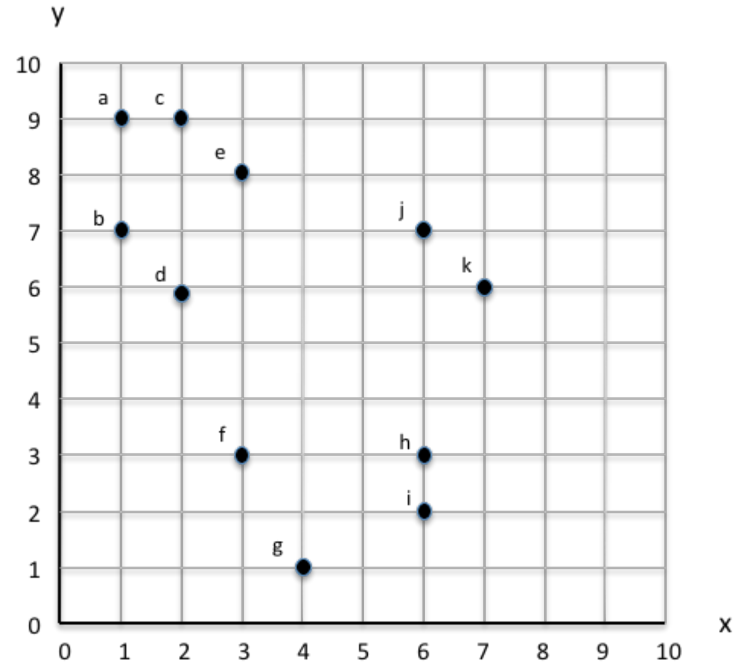
(a) $0.3 \times \log_2(0.3) + 0.7 \times \log_2(0.7) = 0.88$

(b) $H(y|o) = 0.5 \times 0.0 + 0.5 \times (0.4 \times \log_2(0.4) + 0.6 \times \log_2(0.6)) = 0.49,$
$IG(o;y) = H(y) - H(y|o) = 0.88 - 0.49 = 0.39$

(c) Outlook

Stran je prazna, da lahko nanjo rešujete nalogo.

4. 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 s 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
       12
```

Stran je prazna, da lahko nanjo rešujete nalogo.

[5] 5. Priporočilni sistem, ki deluje na podlagi matričnega razcepa, smo pognali na podatkih o ocenah knjig. V podatkih se pojavlja 2000 uporabnikov, ki so z ocenami med 0 in 10 ocenili (nekatere od) 15000 knjig. Skupno imamo v učni množici 100000 ocen. S pomočjo validacijske množice in poskušanjem smo našli najboljše parametre: $k = 13$ latentnih faktorji in ustavitev po 20 iteracijah, ker dobimo najboljši rezultat na validacijski množici RMSE = 1.81. 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 (v intervalu $[0, 50]$) in RMSE na osi y (med $[0, 10]$). Vanj čim bolj natančno vrišite in označite tri krivulje:

- RMSE na validacijski množici za število latentnih faktorjev $k = 13$
- RMSE na validacijski množici za število latentnih faktorjev $k = 2$
- RMSE na validacijski množici za število latentnih faktorjev $k = 90$

Solution:

(1 točka) Najnižjo vrednost 1.81 bo dosegla krivulja za $k = 13$ pri 20 iteracijah. Drugi dve krivulji.

(1 točka) Vse krivulje najprej padajo in naraščajo po minimumu. Le tista za $k = 2$ lahko ne narašča.

(1 točka) Vse krivulje začnejo več ali manj na isti na isti točki (0.5 točke), za smiselno inicializacijo največ na RMSE = 5 (0.5 točke).

(2 točki) Krivulja za $k = 90$ mora po svojem minimumu naraščati hitreje kot tista za $k = 13$, ki mora naraščati hitreje kot tista za $k = 2$.