Uvod v odkrivanje znanj iz podatkov (Poslovna inteligenca)

1. izpitni rok

22. januar 2018

Priimek in ime (tiskano): _____

Vpisna številka: _____

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

1. Dani so trije nabori podatkov.

Nabor A:

x_1	x_2
0.5	0.6
0.7	0.5
1.2	1.3
1.4	1.6
1.7	1.5
1.9	1.8
2.4	2.2

Nabor B:

x_1	x_3
0.5	1.1
0.7	0.3
1.2	1.5
1.4	1.8
1.7	1.3
1.9	2.4
2.4	1.7

Nabor C:

x_1	x_4
0.5	1.2
0.7	2.1
1.2	0.2
1.4	0.3
1.7	3.1
1.9	0.1
2.4	1.5

[1] (a) Za kateri nabor podatkov bo prva glavna komponenta dobljena po metodi glavnih komponent (PCA) pojasnila največ variance v podatkih? Za kateri nabor bo ta komponenta pojasnila najmanj variance v podatkih?

[1] (b) Zakaj?

[1] (c) V razsevnem diagramu izriši podatkovni nabor, kjer prva od komponent PCA pojasni največ variance. Uporabi ravnilo, primerno označi koordinatni osi in opremi koordinatni osi z oznakami za vrednosti obeh spremenljivk. Izriši premico, za katero misliš, da je njena smer enaka smeri glavne komponente (smeri glavne komponente ne rabiš izračunati).

[1] (d) Izračunaj masni center podatkov, ga s točko M označi v razsevnem diagramu in na tem grafu prikaži, kje je njegova projekcija na premico, s katero si označil glavno komponento.

[1] (e) Na glavno komponento proiciraj vse točke (primere) iz podatkovnega nabora in glede na projekcijo masnega centra izračunaj nove koordinate točk oziroma vrednosti, ki bi jih dobil s projekcijo PCA, kjer upoštevaš samo glavno komponento. Te vrednosti dopiši v novi koloni, ki jo dodaš podatkovnemu naboru (torej tvoji izbrani tabeli, ki si jo upodobil na razsevnem diagramu).

Solution: a, b. M: (1.40, 1.36). $[-1.175, -1.096, -0.186, 0.165, 0.317, 0.668, 1.306]$

Stran je prazna, da lahko nanjo rešujete nalogo.

2. Dana je funkcija $y(\theta_1, \theta_2) = 2\theta_1^2 + (3 - \theta_0\theta_1)^2$.

[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 $y(\theta_0, \theta_1)$, pri katerih ima ta funkcija minimum. Začetne vrednosti parametrov nastavimo na $[\theta_0, \theta_1]^T = [1, 2]^T$. 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: $[4\theta_1 - 2(3 - \theta_1\theta_2)\theta_2]$, $[0, -2]^T$, enako, po prvem koraku $[1, 2.2]$, po drugem koraku $[0.952, 2.36]$.

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:

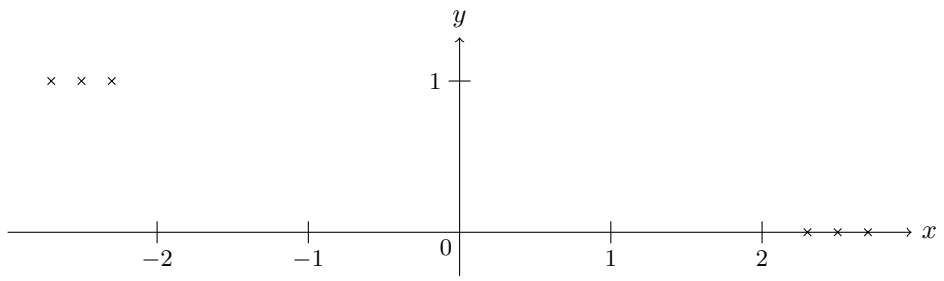
- Regularizacija zmanjša vrednosti parametrov, predvsem tistih, ki bi bili brez regularizacije visoki.
- Manjše vrednosti odvodov, bolj gladka funkcija.
- $$\mathrm{J}(\Theta) = \frac{1}{2m}\sum_{i=1}^{m}(h_\Theta(x^{(i)}) - y^{(i)})^2 + \lambda\sum_{j=1}^{n}\Theta_j^2$$
- $$\Theta_j \leftarrow \Theta_j(1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

Stran je prazna, da lahko nanjo rešujete nalogo.

[1] 4. (a) Označi, ali so sledeče izjave glede logistične regresije resnične ali ne.

- Z regularizacijo ne moremo poslabšati rezultatov na učni množici.
- Z regularizacijo ne moremo poslabšati rezultatov na testni množici.
- Z dodajanjem novih atributov v model (npr. zmnožkov obstoječih atributov) preprečimo pretirano prilagajanje podatkov učni množici.

[2] (b) Imamo podatke s šestimi primeri in enim atributom (x, glej skico). Napovedati želimo razred y. Dvakrat uporabimo logistično regresijo: prvič z zelo majhno vrednostjo regularizacijskega koeficienta λ, drugič z zelo veliko. V koordinatni sistem vrišite krivulji, ki opisujeta napovedi logistične regresije $P(Y = 1)$ pri veliki in majhni vrednosti λ.



[2] (c) Osnovno tehniko logistične regresije uporabljamo na dvorazrednih podatkih. Kako bi lahko prilagodili postopek za podatke, kjer je razredov več (npr. pet)?

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 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.