

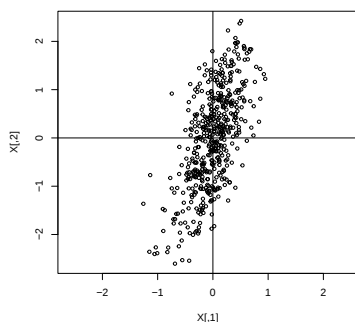
Zopár príkladov písomkového typu

Tieto príklady sú iba z prvých 3 týždňov, lebo som ich vymyslel v tom čase. Na písomke, samozrejme, budú aj ďalšie preberané témy.

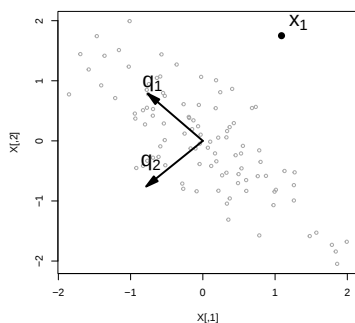
Poznámka: Pre výberové otázky (typu a), b), c)...) je správna vždy práve jedna odpoveď. Za správnu odpoveď v takých otázkach je $+x$ bodov a za nesprávnu odpoveď $-x$ bodov.

1. Načrtnite hustotu premennej, pre ktorú platí, že jej priemer je jasne väčší ako jej medián.

2. Na obrázku nižšie sú dáta $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^2$. Označme vlastné hodnoty výberovej kovariančnej matice $\mathbf{S}$ ako $\lambda_1 > \lambda_2 > 0$, a príslušné vlastné vektory ako $\mathbf{q}_1, \mathbf{q}_2$. Do obrázku dokreslite $\mathbf{q}_2$.



3. Na obrázku sú dáta spolu s vlastnými vektormi matice $\mathbf{S}$ a realizujeme metódu hlavných komponentov. Veľká čierna bodka je jedno konkrétne pozorovanie $\mathbf{x}_1 \in \mathbb{R}^2$.



Vyberte, ktorý z nasledujúcich vektorov je vektor skóre $\mathbf{y}_1$ pre toto pozorovanie (t.j. $\mathbf{y}_1$ je “PCA reprezentácia bodu $\mathbf{x}_1$ ”):

a) $(0.5, 2)^T$

c) $(0.5, -2)^T$

b) $(-0.5, 2)^T$

d) $(-0.5, -2)^T$

e) $(8, 2)^T$

g) $(8, -2)^T$

f) $(-8, 2)^T$

h) $(-8, -2)^T$

4. Máme dáta o futbalistoch, pričom o každom futbalistovi sú vo vektore $\mathbf{x}_i$ postupne údaje: 1) počet striel, 2) počet odobratí lopty, 3) počet faulov, 4) počet úspešných vedení lopty, 5) počet gólov, pričom všetky údaje sú uvádzané “na zápas”. Prvé dva stĺpce matice rotácie $\mathbf{Q}$ v metóde hlavných komponentov sú:

$$\mathbf{Q} = \begin{bmatrix} 0.46 & -0.03 \\ 0.35 & 0.72 \\ 0.46 & 0.27 \\ 0.34 & -0.55 \\ 0.57 & -0.3 \end{bmatrix}.$$

(i) Od ktorých premenných výrazne závisí druhý hlavný komponent (PC2)? Teda zvýšenie hodnoty PC2 zodpovedá výraznému nárastu ktorých a výraznému poklesu ktorých premenných?

(ii) Skúste slovne popísať PC2 (t.j. ako by ste charakterizovali futbalistov, ktorí majú vysokú hodnotu PC2?).

Otázky (i), (ii) zodpovedzte aj pre PC1.

5. Máme údaje o 150 firmách a lineárnou regresiou modelujeme zisk firmy pomocou nasledovných premenných: obrat firmy, kvadratický vplyv obratu firmy, celková výška dlhu firmy, typ odvetvia v ktorom firma podniká (3 kategórie: poľnohospodárstvo, priemysel, služby).

(i) Zapište príslušný model lineárnej regresie (teda typu $zisk_i \approx \dots$).

(ii) Napíšte maticu $\mathbf{X}$ lineárnej regresie (pre tvar $\mathbf{y} \approx \mathbf{X}\beta$), vrátane určenia jej rozmerov.

6. Máme model

$$zisk_i \approx \beta_0 + \beta_1 obrat_i + \beta_2 dlh_i + \beta_3 \ln(dlh_i), \quad (1)$$

kde dáta sú ako v predchádzajúcom príklade. Parametre β sme odhadli:

$$\hat{\beta} = (-1, 0.5, -0.1, -2)^T.$$

Pre novú firmu sme získali jej údaje: obrat = 4, dlh = 2.72. Vypočítajte predikciu zisku tejto firmy modelom (1).

7. Na nasledovnom obrázku máme vrstevnice (úrovňové množiny) súčtu štvorcov chýb $J(\beta)$ pre lineárnu regresiu $\mathbf{y} \approx \mathbf{X}\beta$. Sú tam tiež dva konkrétne vektory parametrov $\beta^{(1)}, \beta^{(2)}$.

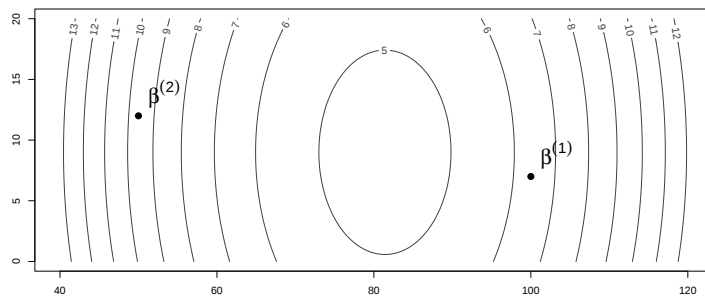
(i) Napíšte matematický predpis $J(\beta)$. (Pozor, nepýtam sa tu na vzorec pre $\hat{\beta}$.)

(ii) Určte, ktoré z tvrdení o $J(\beta^{(1)})$, $J(\beta^{(2)})$ je pravdivé:

a) $J(\beta^{(1)}) \leq J(\beta^{(2)})$

b) $J(\beta^{(1)}) \geq J(\beta^{(2)})$

c) nevieme povedať, ktoré z $J(\beta^{(1)})$, $J(\beta^{(2)})$ je väčšie



8. Dátový súbor sme si rozdelili na trénovaciu, validačnú a testovaciu časť, a pre regresný predikčný model sme vypočítali $MSE_{\text{train}} = 21.3$, $MSE_{\text{val}} = 152.4$.

(i) O tomto modeli môžeme povedať, že pravdepodobne ide o

a) podučenie (underfit)

b) preučenie (overfit)

(ii) Pre MSE_{test} potom nevyhnutne platí, že leží v intervale:

a) $(-\infty, 21.3]$

c) $[152.4, \infty)$

b) $[21.3, 152.4]$

d) nevieme povedať