Domáca úloha 4

Hrebeňová regresia (Ridge regression) je upravená verzia klasickej regresie. Teda máme model

$$y_i = \mathbf{x}_i^T \beta + \varepsilon_i,$$

kde $\beta = (\beta_1, \ldots, \beta_k)^T$ sú hľadané parametre a $\mathbf{x}_i \in \mathbb{R}^k$ sú vstupné údaje o i-tom objekte. Potom optimálnu β hľadáme minimalizáciou trochu inej účelovej funkcie, než je súčet štvorcov chýb. Konkrétne minimalizujeme

$$\sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \beta)^2 + 10 \sum_{j=2}^{k} \beta_j^2 \tag{1}$$

voľbou $\beta \in \mathbb{R}^k$ (všimnite si, že druhá suma nezahŕňa prvok $j = 1$, ktorý reprezentuje intercept). Zahrnutím druhej sumy sa hrebeňová regresia snaží zabrániť pretrénovaniu: snaží sa, aby odhadnuté parametre $\widehat{\beta}$ neboli príliš veľké.

Aplikujte túto metódu na dáta o študentoch na stránke. Konkrétne:

1. Načítajte dáta zo `studenti1.txt` na stránke (hint: zvoľte správne `sep` a `header`). Prvý stĺpec (`vysledok`) značí výsledok na teste a ostatné stĺpce sú vysvetľujúce premenné (napr. vek, počet vymeškaných hodín). Dáta pochádzajú z `https://archive.ics.uci.edu/ml/datasets.php`.

2. Vektor $\mathbf{y} \in \mathbb{R}^n$ je teda prvý stĺpec z datasetu, a matica $\mathbf{X} \in \mathbb{R}^{n \times k}$ je tvorená stĺpcom jednotiek a ostatnými stĺpcami z datasetu, teda

   $$\mathbf{X} = [\mathbf{1}_n, \text{ stĺpce 2 až 12 zo studenti1.txt}]$$

   (teda máme $k = 12$). Vektory $\mathbf{x}_i^T$ v (1) sú riadky matice $\mathbf{X}$.

3. Pomocou optimalizačných funkcií v Rku pre tieto dáta nájdite $\widehat{\beta}_{HR} \in \mathbb{R}^k$, ktorá minimalizuje (1). Na nájdenie $\widehat{\beta}_{HR}$ nevyužívajte priamo funkcie riešiace hrebeňovú regresiu.

4. Porovnajte váš výsledok s výsledkom získaným klasickou lineárnou regresiou. Lineárnu regresiu môžete vyriešiť ľubovoľne (napr. použitím optimalizačných funkcií, ako sme to robili na hodine; použitím `lm`; priamym výpočtom $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$). Konkrétne:

   - Porovnajte $\widehat{\beta}_{HR}$ s odhadom $\widehat{\beta}$ lineárnou regresiou. Sú si tieto odhady podobné? Má odhad $\widehat{\beta}_{HR}$ menšiu normu ako $\widehat{\beta}$ (ako by naznačovalo (1))?
   - Porovnajte predikcie získané z týchto modelov na dátovom súbore `studenti2.txt`. Predikcia pre i-teho študenta zo `studenti2.txt` pomocou lineárnej regresie je $\widehat{y}_i = \tilde{\mathbf{x}}_i^T \widehat{\beta}$, kde $\tilde{\mathbf{x}}_i^T$ sú údaje o príslušnom študentovi, teda pre obidve regresie:

     $$\tilde{\mathbf{x}}_i^T = (1, \text{ hodnoty 2 až 12 z } i\text{-teho riadku studenti2.txt}).$$

     Podobne predikcie pomocou hrebeňovej regresie sú $\widehat{y}_{i,HR} = \tilde{\mathbf{x}}_i^T \widehat{\beta}_{HR}$. Pre `studenti2.txt` vypočítajte priemernú štvorcovú chybu

     $$MSE_{LR} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2, \quad MSE_{HR} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \widehat{y}_{i,HR})^2$$

(tu y_i sú skutočné výsledky študentov zo `studenti2.txt` na teste) a určte, ktorá chyba je menšia. **Aktualizácia 24.4.:** odhady $\widehat{\beta}$ a $\widehat{\beta}_{HR}$ používané v tomto bode sú stále tie isté ako doteraz (teda tie vypočítané zo `studenti1.txt`). Tento bod sa pýta, ako dobre modely získané v predchádzajúcej časti (aj s príslušnými odhadmi biet) predikujú nové dáta, ktoré tie modely doteraz nevideli (`studenti2.txt`).

Poznámka 1: Minimalizácia (1) má v skutočnosti analytické riešenie, na to ale teraz zabudnime. Druhá suma je tiež vo všeobecnosti prenásobená voliteľným parametrom λ a nie pevným číslom 10, čo teraz tiež neberieme do úvahy.
Poznámka 2: Ak je niečo nejasné, ozvite sa.
*Pomôcka:*Ak budete potrebovať konvertovať data frame do matice, môžete použiť napr. funkciu `data.matrix`.
Bonus (2 body): Výpočty spravte efektívne (teda vektorizovane, bez zbytočných for-cyklov). Úplne všetky výpočty sa dajú spraviť bez for-cyklov a bez funkcií typu `apply`.