Domáca úloha 10 - Moderná aplikovaná regresia

Prosím odovzdať do budúceho týždňa.

1 (60b)

Vezmite dáta `fat` so `siri` ako odozvou a ostatné regresory ako prediktory okrem `brozek` a `density`. V rozumnom pomere náhodne rozdeľte dátovú vzorku na testovaciu a trénovaciu. Fitnite nasledujúce modely, kde hyper parametre (odborný termín pre parametre metód ako napríklad λ pri ridge regresion) volíte podľa možností

(a) Lineárny model so všetkými prediktormi,

(b) Lineárny model s prediktormi, ktoré zvolíte vy. (model selection alebo podľa oka a pod.),

(c) PCA regresiu,

(d) Ridge regresiu,

(e) Lasso regresiu,

Každý model použite na predikciu testovacej vzorky a porovnajte. Napište záver, ktorý z modelov bol podla vás najlepší z hľadiska predikcie a prečo.

2 (40b)

Majme dáta `kanga` obsahujúce data o historických lebkách kengúr.

(a) Použite PCA na 18 mier lebky. Pozor, dataset nie je "čistý". Koľko percent variancie vysvetluje prvý komponent ?

(b) Vytiahnite z PCA prvú lineárnu kombináciu (loading). Ktoré premenné majú v prvom komponente výrazné zastúpenie?

(c) Zopakujte (a) a (b), ale na preškálovanej matici premenných, popíšte rozdiely a dôvod prečo sú (možno) rozdielne.

(d) Interpretujte druhý hlavný komponent.

(e) Spočítajte Mahalanobisove vzdialenosti na základe ktorých nájdite outlierov,

(f) Vykreslite obrázok prvej a druhej komponenty za použitia rôznych znakov na rozlíšenie pohlavia kengúr. Myslíte si, že tieto dva komponenty postačujú na identifikáciu pohlavia ?

3 (Bonus)

Majme ohadovaciu metódu založenu na minimalizovani

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p}|\beta_j|\le s$$

pre nejaké s. V nasledujúcich otázkach (a) až (d) vyberte jednu z možností (i) - (v) a svoju odpoveď overte simulačne.

(a) Ak začneme zvyšovať hodnotu s z 0, tak trénovacia RSS bude:

(i) Počiatočne bude stúpať a potom eventuálne začne klesať (ako otočené písmeno U)

(ii) Počiatočne bude klesať a potom eventuálne začne stúpať (ako písmeno U)

(iii) Stále klesať

(iv) Stále stúpať

(v) Zostane konštantná

(b) Zopakujte (a) pre testovaciu RSS

(c) Zopakujte (a) pre varianciu odhadu parametrov,

(d) Zopakujte (a) pre (štvorec) biasu,