Statistical Learning

Prova d'esame

14 Settembre 2022

Tempo a disposizione: 150 minuti

Problema 1

Si consideri il modello $y = X\beta + \epsilon$, dove

$$y = \begin{bmatrix} -1.3 \\ 0.2 \\ -0.5 \\ -0.6 \end{bmatrix}, \quad X = \frac{1}{2} \begin{bmatrix} 1 & -1 \\ 1 & 1 \\ -1 & -1 \\ -1 & 1 \end{bmatrix}$$

β è un vettore di p parametri incogniti e $\epsilon \sim N(0, \sigma^2 I_p)$.

a. Calcolare la stima OLS $\hat{\beta}$. Riportare il valore del primo elemento di $\hat{\beta}$.

b. Si consideri lo stimatore *ridge regression*

$$\hat{\beta}_\lambda = \min_\beta \|y - X\beta\|^2 + \lambda \|\beta\|_2^2$$

Riportare il valore di λ che corrisponde la stima $\hat{\beta}_\lambda = 0.5\hat{\beta}$, ovvero il valore che rende la stima *rigde* pari alla stima OLS dimezzata.

c. Si consideri lo stimatore *lasso*

$$\tilde{\beta}_\lambda = \min_\beta \frac{1}{2}\|y - X\beta\|^2 + \lambda \|\beta\|_1$$

Calcolare la stima lasso per il valore di λ determinato al punto precedente. Riportare il valore del primo elemento di $\tilde{\beta}_\lambda$.

Problema 2

Si consideri il dataset `mcycle`, presente nella libreria `MASS`, dove `accel` è la variabile risposta e `times` il predittore.

Costruire una base *B-splines* `B` di grado 2 con 50 intervalli equidistanti (il *range* da dividere è da `min(times)` a `max(times)`). Si consideri la regressione *P-splines* che utilizza la base `B`, un ordine delle differenze pari a 3 e un valore di λ pari a 10.

a. Calcolare il *mean squared error* $\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, dove $\hat{y}_i$ è la stima per y_i ottenuta con la regressione *P-splines*.

b. Calcolare l'errore di convalida incrociata *Leave-One-Out*, ovvero $\text{LOO} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i^{(-i)})^2$, dove $\hat{y}_i^{(-i)}$ è la stima per y_i ottenuta con la regressione *P-splines* rimuovendo l'i-sima osservazione.

Problema 3

Si consegni il file .R che produce le risposte alle domande richieste. Il codice deve essere **riproducibile** e, se eseguito, deve stampare in output **solo** il risultati richiesti dalle domande a) e b).

Si consideri il dataset `Boston` presente nella libreria `MASS`, utilizzando come variabile risposta `medv` e le rimanenti variabili come predittori.

Calcolare le frequenze relative di selezione $\hat{\pi}_j$, $j = 1, \ldots, p$ dell'algoritmo *Complementary Pairs Stability Selection*, utilizzando $B = 50$ ricampionamenti.

Utilizzare come metodo per calcolare $\hat{S}_{n/2}$ la regressione *Best Subset Selection*, impostata in modo da selezionare $q = 6$ variabili, utilizzando la funzione `regsubsets` presente nella libreria `leaps`.

a. Riportare l'insieme di predittori stabili $\hat{S}_{\text{stab}}$ utilizzando la soglia $\tau = 0.95$.

b. Per i predittori stabili ottenuti al punto precedente, calcolare il limite superiore del numero atteso di errori di I tipo.

Problema 4

a. Per quale motivo è preferibile utilizzare una *B-spline basis* rispetto alla *truncated power basis*?

b. Cosa afferma il Teorema di Theobald (1974)?

c. Siano X_1, X_2, X_3 variabili aleatorie indipendenti con $X_i \sim N(\mu_i, 1)$ per $i = 1, 2, 3$. Lo stimatore $\hat{\mu} = (3, 2, 1)$ per $\mu = (\mu_1, \mu_2, \mu_3)$ è ammissibile? Si motivi la risposta.