

Statistical Learning

Prova d'esame

26 Aprile 2022

Tempo a disposizione: 180 minuti

Problema 1

Si risponda alle seguenti domande.

a) Sia

$$X^t = \begin{bmatrix} 2 & 1 & -2 \end{bmatrix}, y^t = \begin{bmatrix} -1 & -1 & 1 \end{bmatrix}, X^t X = \begin{bmatrix} 9 \end{bmatrix}, X^t y = \begin{bmatrix} -5 \end{bmatrix},$$

Riportare la varianza dello stimatore *ridge regression*, i.e. $\text{Var}(\hat{\beta}_\lambda)$ per $\lambda = 1$, sostituendo al valore incognito σ^2 la stima $\hat{\sigma}^2 = n^{-1} \|y - X\hat{\beta}_\lambda\|^2$.

b) Sia $y = X\beta + \epsilon$, dove

$$X = \begin{bmatrix} -0.5 & -0.5 \\ -0.5 & 0.5 \\ 0.5 & -0.5 \\ 0.5 & 0.5 \end{bmatrix}, \quad \beta = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \epsilon \sim N \left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \right)$$

Riportare il valore di λ che minimizza $\text{MSE}(\hat{\beta}_\lambda)$ per lo stimatore *ridge regression*

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

(non è richiesto di standardizzare y e/o le colonne di X e non è presente l'intercetta).

c) Sia $y = (-1.6, -0.2, 1.6, 1.1)^t$ una realizzazione del modello specificato al punto precedente. Calcolare la stima $\tilde{\beta}_\lambda$ dello stimatore *lasso*

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

con il valore di λ determinato al punto precedente (non è richiesto di standardizzare y e/o le colonne di X e non è presente l'intercetta). Riportare il valore del primo elemento di $\tilde{\beta}_\lambda$.

Problema 2

Si consideri il dataset `longley` presente nella libreria `datasets`. La variabile risposta è `Employed`, i predittori sono `GNP.deflator`, `GNP`, `Unemployed`, `Armed.Forces`, `Population` e `Year`.

a) Per questi dati, si stimi la regressione *Best Subset Selection* scegliendo il *Best Subset* finale con il criterio BIC. Riportare la stima $\hat{\beta}_{GNP}$ per la variabile `GNP`.

b) Si suddivida il dataset in *Learning set* con osservazioni con indici in $L = \{1, 3, 5, 7, 9, 11, 13, 15\}$ e *Inference set* con osservazioni con indici in $I = \{2, 4, 6, 8, 10, 12, 14, 16\}$. Sulla base del *Learning set*, stimare $\hat{S} = \{j : \hat{\beta}_j \neq 0\}$, dove $\hat{\beta}_j$ sono le stime della regressione *Best Subset Selection* scegliendo

il *Best Subset* finale con il criterio BIC. In altre parole, \hat{S} contiene le variabili selezionate da *Best Subset Selection* stimato sul *Learning set*. Sulla base dell'*Inference set*, calcolare i p -values del modello lineare con le variabili selezionate (e l'intercetta). Riportare il p -value relativo alla variabile **Unemployed** aggiustato con il metodo di Bonferroni, che tiene conto della molteplicità della selezione.

- c) Sia $\hat{\mu}_L(x) = \hat{\mu}(x; (x_l, y_l), l \in L)$ il modello *Best Subset Selection* stimato sul *Learning set* al punto precedente. Calcolare i residui in valore assoluto $R_i = |y_i - \hat{\mu}_L(x_i)|$ per $i \in I$, e ordinare $\{R_i, i \in I\}$ in senso crescente, i.e. $R_{(1)} \leq \dots \leq R_{(m)}$. Riportare il valore critico $R_\alpha = R_{(k)}$ con $k = \lceil (1 - \alpha)(m + 1) \rceil$, dove $m = 8$ e $\alpha = 1/3$.

Problema 3

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

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 3 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** e un ordine delle differenze pari a 2. Si determini il valore di λ tra i seguenti valori

```
lambdas = 10 ^ seq(from = -4, to = 2, by = .1)
```

in modo da minimizzare l'errore di convalida incrociata *Leave-One-Out*, ovvero

$$\text{LOO}(\lambda) = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i^{(-i)}(\lambda))^2$$

dove $\hat{y}_i^{(-i)}(\lambda)$ è la stima per y_i ottenuta con la regressione *P-splines*(λ) rimuovendo l' i -sima osservazione.

Riportare

- il valore λ^* che minimizza $\text{LOO}(\lambda)$
- il valore $\text{LOO}(\lambda^*)$,
- i gradi di libertà effettivi corrispondenti a λ^* .

Problema 4

Si risponda alle seguenti domande :

- a) Si consideri il metodo *Stability Selection*. Se si vuole garantire

$$\mathbb{E}(|\hat{S}_{\text{stab}} \cap N|) \leq 10$$

con $p = 2000$ e $q = \mathbb{E}(|\hat{S}_{n/2}|) = 10$, quanto deve valere la soglia τ ?

- 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 $\hat{\mu} = (1, 2, 3)$ per $\mu = (\mu_1, \mu_2, \mu_3)$ è ammissibile? Si motivi la risposta.
- Si considerino i dati `longley` del Problema 2. Si supponga di voler utilizzare il metodo *fixed-X knockoffs* per selezionare le variabili, controllando il *False Discovery Rate* al livello $\alpha = 0.1$. Prima di analizzare i dati, vi confrontate con una vostra amica, che vi consiglia di lasciar perdere perché con $\alpha = 0.1$ il metodo non selezionerà nessuna variabile con probabilità 1. La vostra amica ha ragione? Motivare la risposta.