Statistical Learning

Prova d'esame

24 Giugno 2022

Tempo a disposizione: 180 minuti

Problema 1

Si risponda alle seguenti domande.

a) Si consideri un modello di regressione con matrice del disegno X e vettore di risposta y dati da

$$X = \begin{bmatrix} 10^{-9} & -1 \\ -1 & 10^{-5} \end{bmatrix}, \quad y = c(3 \times 10^9, -2.99997)$$

Si calcoli la stima dei minimi quadrati $\hat{\beta}$. Si riporti il primo elemento di $\hat{\beta}$.

b) Siano $X_1, \ldots, X_n$ i.i.d. ad una variabile X con distribuzione $N(\mu, 1)$. Si consideri lo stimatore media campionaria $\bar{X}$. Quanto vale il rapporto tra errore di previsione atteso ed errore di stima atteso, ovvero

$$\frac{\mathrm{E}[(\bar{X} - X)^2]}{\mathrm{E}[(\bar{X} - \mu)^2]}$$

per $n = 5$? Riportare il risultato. Potete rispondere alla domanda in modo analitico oppure utilizzando una simulazione con un gran numero di ricampionamenti (e.g. 10^5).

c) Sia

$$x = (0.24, 1.61, 4.61, 3.95)^t$$

la realizzazione di una variabile casuale X con distribuzione Gaussiana p-variata $N_p(\mu, \Sigma)$ con vettore delle medie μ incognito e matrice di varianze/covarianze

$$\Sigma = \begin{bmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{bmatrix}$$

con $\rho = -1/4$. La stima di massima verosimiglianza per μ è data da $\hat{\mu}_{\mathrm{MLE}} = x$. Calcolare la stima secondo James-Stein $\hat{\mu}_{\mathrm{JS}}$ (con $c = 1$). Riportare il valore del primo elemento di $\hat{\mu}_{\mathrm{JS}}$.

Problema 2

Si risponda alle seguenti domande.

a) Si consideri il modello lineare semplice con intercetta e una covariata x_i. I dati sono

$$\{(y_i, x_i)_{i=1}^4\} = \{(1.4, 0), (1.4, -2), (0.8, 0), (0.4, 2)\}$$

Riportare il valore di λ che corrisponde alla stima ridge $\hat{\beta}_\lambda = (1, -1/24)^t$ senza penalizzare l'intercetta.

b) Si supponga di avere una matrice del disegno X tale che $X^tX = I_p$. Sia $X^ty = (-0.56, -0.23, 1.56, 0.07)^t$. Calcolare la stima $\tilde{\beta}_\lambda$ dello stimatore *best subset selection*

$$\tilde{\beta}_\lambda = \min_\beta \frac{1}{2}\|y - X\beta\|^2 + \lambda\|\beta\|_0$$

con il valore di $\lambda = 0.1$. Riportare il valore del primo elemento di $\tilde{\beta}_\lambda$.

c) Con i dati del punto precedente, calcolare la stima James-Stein $\hat{\beta}_{\text{JS}}$ per β, ipotizzando di conoscere il vero valore di $\sigma = 1$. Riportare il valore del primo elemento di $\hat{\beta}_{\text{JS}}$.

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

Si consideri come *training set* le prime 15 osservazioni (righe) del dataset `longley`, e come *test point* la sedicesima osservazione (anno 1962).

Si utilizzi l'algoritmo *split conformal* considerando come *Learning set* le osservazioni del training set con indici pari, i.e. $L = \{2, 4, 6, 8, 10, 12, 14\}$ e come *Inference set* le osservazioni del training set con indici dispari, i.e. $I = \{1, 3, 5, 7, 9, 11, 13, 15\}$. Si consideri la regressione *ridge* per $\lambda = 1$, utilizzando la funzione `lm.ridge` presente nella libreria `MASS`.

Si costruisca l'intervallo di previsione a livello $1 - \alpha$ con $\alpha = 1/3$ per il test point. Riportare

a) gli estremi dell'intervallo di previsione;
b) il valore `TRUE` se il test point si trova all'interno dell'intervallo di previsione, `FALSE` altrimenti.

Problema 4

a) Sia X una variabile causale con distribuzione $N_p(\mu, \Sigma)$. La matrice Σ ha elementi diagonali pari a 1 ed elementi fuori dalla diagonale pari a $r > 0$. In questo caso l'autovettore più grande di Σ è pari a

$$\lambda_1 = 1 + (p - 1)r$$

Per quale valori di p è preferibile utilizzare lo stimatore di James-Stein come alternativa allo stimatore di massima verosimiglianza? Cosa succede se $r = 0.5$?

b) Per gli intervalli di previsione, qual è la relazione tra *conditional coverage* e *marginal coverage*?

c) Si consideri un insieme di n punti (x_i, y_i), $i = 1, \ldots, n$. Cosa succede se stimiamo una natural cubic spline con un n nodi in corrispondenza dei valori $x_1, \ldots, x_n$?