

Statistical Learning

Prova d'esame

14 Luglio 2022

Tempo a disposizione: 180 minuti

Problema 1

Si considerino i seguenti dati:

Player	n_i	s_i	pi_i
Baines	90	26	0.289
Barfield	90	22	0.256
Bell	90	23	0.265
Biggio	90	25	0.287
Bonds	90	27	0.297

di $p = 5$ giocatori di baseball, dove n_i e s_i indicano rispettivamente il numero di volte a battuta e il numero di battute valide, mentre π_i indica la vera media battuta (calcolata su tutta la carriera di ciascun giocatore).

Sia Z_i la variabile aleatoria Binomiale(n_i, π_i)/ n_i , e si supponga che Z_1, \dots, Z_p siano indipendenti.

Si consideri valida la seguente approssimazione

$$Z_i \approx N(\pi_i, \sigma_0^2)$$

dove $\sigma_0^2 = \frac{\bar{p}(1-\bar{p})}{90}$ con \bar{p} pari alla media dei valori (s_i/n_i).

- Sia $\hat{\pi}^{\text{MLE}}$ la stima di massima verosimiglianza per $\pi = (\pi_1, \dots, \pi_p)$. Riportare il valore della stima per Bell.
- Sia $\hat{\pi}^{\text{JS}}$ la stima secondo James-Stein per π . Riportare il valore della stima per Bell.
- Calcolare

$$\frac{\sum_{i=1}^p (\hat{\pi}_i^{\text{JS}} - \pi_i)^2}{\sum_{i=1}^p (\hat{\pi}_i^{\text{MLE}} - \pi_i)^2}$$

Problema 2

Si consideri il dataset `cement` presente nella libreria `MASS`. La variabile risposta è y , i predittori sono le rimanenti variabili.

- Sia X la matrice del disegno. Calcolare il *condition number* di $(X^T X + \lambda I_p)$ per il modello di regressione *ridge* con $\lambda = 10$.
- Calcolare la stima dell'errore di previsione tramite *leave-one-out cross-validation*

$$\text{LOO}_\lambda = \frac{1}{n} \sum_{i=1}^n (y_i - x_i^t \hat{\beta}_\lambda^{(-i)})^2$$

dove $\hat{\beta}_\lambda^{(-i)}$ è la stima *ridge* calcolata su $(n - 1)$ osservazioni escludendo l'osservazione (x_i, y_i) per λ pari a $0, 1, 2, \dots, 99$. Riportare il valore minimo di LOO_λ .

- c. Calcolare la stima dell'errore di previsione tramite *generalized cross-validation* con il valore di λ che ha minimizzato LOO_λ nel punto precedente.

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** il risultati richiesti dalle domande a) e b).

Si consideri il dataset `Boston` presente nella libreria `MASS`. La variabile risposta è `medv`, i predittori sono le rimanenti variabili.

Si consideri come *training set* le prime 505 osservazioni (righe) del dataset `Boston`, e come *test point* l'ultima osservazione (riga 506).

Si utilizzi l'algoritmo *split conformal* considerando come *Learning set* le osservazioni del training set con indici pari, i.e. $L = \{2, 4, \dots, 504\}$ e come *Inference set* le osservazioni del training set con indici dispari, i.e. $I = \{1, 3, \dots, 505\}$.

Si costruisca l'intervallo di previsione per il test point a livello $1 - \alpha$ con $\alpha = 25/254$. L'algoritmo da utilizzare è la regressione *forward*, impostata in modo da selezionare 6 variabili (per la regressione *forward* è obbligatorio utilizzare la funzione `step` presente nella libreria `stats`).

Riportare

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

Problema 4

Si consideri il seguente esperimento (ipotetico) con tecnologia *microarray*. Ci sono $n = 400$ partecipanti che entrano nello studio uno per giorno, e ricevono il Trattamento o il Placebo a giorni alterni (il giorno 1 il soggetto 1 riceve il Trattamento, il giorno 2 il soggetto 2 riceve il Placebo, il giorno 3 il soggetto 3 riceve Trattamento, etc.). A ciascun soggetto viene misurata la risposta di $p = 200$ geni. La matrice dei dati X è di dimensione 400×200 , e ciascun elemento è indipendente con distribuzione gaussiana, ovvero

$$X_{ij} \stackrel{\text{ind}}{\sim} \mathcal{N}(\mu_{ij}, 1)$$

Si supponga che per $j = 30, 48, 57, 65, 84, 92, 113, 128, 143, 195$

$$\mu_{ij} = 2 \quad i = 1, 3, \dots, 317, 319 \text{ (Treatment)} \quad \mu_{ij} = -2 \quad i = 2, 4, \dots, 318, 320, \text{ (Placebo)}$$

e per tutto il resto $\mu_{ij} = 0$. Si consideri l'utilizzo della Foresta Casuale (*Random Forest*)

- con suddivisione in training set di 320 soggetti e test set di 80 soggetti in maniera casuale (RF-I);
- con suddivisione in training set con i primi 320 soggetti e test set con gli ultimi 80 soggetti (RF-II);

Si risponda alle seguenti domande (motivare la risposta per ciascun quesito.)

- Ti aspetti che l'errore di previsione valutato sul training set con RF-I sia minore/uguale/maggiore a quello con RF-II?
- Ti aspetti che l'errore di previsione valutato sul test set con RF-II sia minore/uguale/maggiore al 50%?
- Quante variabili (geni) ti aspetti di trovare con elevato punteggio di *Variable Importance* (calcolato sul training set) se utilizzi RF-II?