Analisi Esplorativa (Analisi Statistica Multivariata)

28 Febbrario 2022

Tempo a disposizione: 120 minuti

Modalità di consegna: svolgere gli esercizi di teoria (parte A) riportando le soluzioni sul foglio protocollo, e consegnare il foglio protocollo assieme al testo della prova d'esame. Successivamente, accedere alla piattaforma esamionline tramite computer e svolgere gli esercizi di analisi dei dati (parte B). In questo caso la consegna si svolge tramite piattaforma esamionline. Il tempo da dedicare alla parte A e alla parte B è a discrezione dello studente.

Compilare con nome, cognome e numero di matricola.

NOME: COGNOME: MATRICOLA:

PARTE A: esercizi di teoria

Esercizi da svolgere sul foglio protocollo senza l'ausilio di R/Rstudio.

Problema 1

1. Si supponga che la matrice dei dati standardizzati $\underset{n\times 2}{Z}$ consista di due colonne $\underset{n\times 1}{z_1}$ e $\underset{n\times 1}{z_2}$ con la seguente matrice di correlazione $R = \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}$ dove $-1 \leq r \leq 1$. Si consideri la matrice dei dati trasformati $\underset{n\times 2}{Y}$ dove $\underset{n\times 1}{y_1} = \underset{n\times 1}{z_1}$ e $\underset{n\times 1}{y_2} = b\underset{n\times 1}{z_2}$ per una costante $b > 0$.
   a. Calcolare la varianza totale per i dati Y.
   b. Calcolare la varianza generalizzata per i dati Y.
   c. Calcolare l'indice relativo di variabilità per i dati Y.
2. Si consideri il modello fattoriale con le seguenti matrici di pesi fattoriali e varianze specifiche:

$$\Lambda = \begin{bmatrix} 4 & 1 \\ 7 & 2 \\ -1 & 6 \\ 1 & 8 \end{bmatrix}, \qquad \Psi = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix}$$

   a. Calcolare la covarianza tra la prima variabile e il primo fattore comune, i.e. $\mathbb{C}\text{ov}(x_1, f_1)$
   b. Calcolare la varianza delle prima variabile, i.e. $\mathbb{V}\text{ar}(x_1)$
   c. Calcolare la covarianza tra le prime due variabili, i.e. $\mathbb{C}\text{ov}(x_1, x_2)$
   d. Calcolare la correlazione tra le prime due variabili, i.e. $\mathbb{C}\text{orr}(x_1, x_2)$
3. Si consideri la seguente matrice dei dati $X = \begin{bmatrix} 0 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}$.

a. Calcolare la matrice $D = \{d^2(u_i, u_l)\}$ delle distanze Euclidee al quadrato.

b. Calcolare la distanza totale $T = \frac{1}{2n}\sum_{i=1}^{n}\sum_{l=1}^{n} d^2(u_i, u_l)$.

c. Si consideri l'algoritmo delle K-medie con $K = 2$. Calcolare la distanza entro i gruppi $W = \sum_{k=1}^{K} W(G_k)$ dove $W(G_k) = \dfrac{1}{2n_k}\displaystyle\sum_{i:u_i\in G_k}\sum_{l:u_l\in G_k} d^2(u_i, u_l)$ per la partizione $G_1 = \{u_1\}$ e $G_2 = \{u_2, u_3\}$.

PARTE B: esercizi di analisi dei dati

Esercizi da svolgere con il computer sulla piattaforma esamionline con l'ausilio di R/Rstudio.

Problema 2

Si consideri la matrice dei dati $\underset{n\times 3}{X}$, a cui è associata la matrice di varianze/covarianze S. Sapendo che i) la prima riga di X è pari a $u_1' = (20.6, 87, 77)$ e il baricentro è pari a $\bar{x}' = (13.9, 76.2, 32.9)$; ii) la seconda colonna di V è $v_2 = (0.1, -1, 0.2)'$, dove V è la matrice degli autovettori relativa a S; la varianza della prima variabile è $s_{11} = 9.8$, mentre il secondo autovalore di S è $\lambda_2 = 26.5$. (nel vostro esercizio il valori potrebbero essere diversi).

1. Calcolare $d_\infty(u_1, \bar{x})$. Riportare il risultato arrotondando al **terzo decimale** (si ricordi l'uso della **virgola** per i decimali).
2. Calcolare il valore del primo elemento di $\underset{n\times 1}{y_2} = \tilde{X}v_2$, dove $\tilde{X}$ è la matrice dei dati centrati e y_2 è il vettore dei punteggi della seconda componente principale. Riportare il risultato arrotondando al **terzo decimale** (si ricordi l'uso della **virgola** per i decimali).
3. Calcolare la correlazione tra la prima colonna $\tilde{x}_1$ di $\tilde{X}$ e i punteggi y_2 della seconda componente principale. Riportare il risultato arrotondando al **terzo decimale** (si ricordi l'uso della **virgola** per i decimali).

Problema 3

Si consideri il dataset `quakes` presente nella libreria `datasets`. Si tratta di 1000 osservazioni misurate su 5 variabili: `lat` Latitude of event; `long` Longitude; `depth` Depth (km); `mag` Richter Magnitude; `stations` Number of stations reporting. Sia X la matrice 999×4 corrispondente al dataset `quakes`, escludendo la riga 1 (nel vostro esercizio il numero di riga potrebbe essere diverso) e la variabile `stations` .

1. Calcolare la matrice dei dati standardizzati Z e riportare il numero di osservazioni anomale verificando se la distanza di Mahalanobis al quadrato di ciascuna osservazione di Z dal baricentro 0 è superiore alla soglia s, dove s corrisponde al quantile 0.95 di una variabile casuale χ^2_p (dove p è il numero di colonne di Z).
2. Rimuovere da Z le osservazioni anomale individuate al punto precedente e su questi dati utilizzare l'algoritmo delle K-medie i) inizializzando i centrodi con le prime K unità statistiche; ii) impostando il numero di iterazioni (argomento `iter.max`) a 100; iii) utilizzando l'algoritmo di Lloyd (argomento `algorithm`). Decidere il numero di gruppi K ottimale secondo l'indice CH, per $K = 2, \ldots, 12$. Riportare la distanza entro i gruppi W corrispondente alla scelta effettuata, arrotondando al **terzo decimale** (si ricordi l'uso della **virgola** per i decimali).
3. Calcolare il valore medio della *silhouette* corrispondente alla raggruppamento selezionato al punto precedente, utilizzando la matrice delle distanze Euclidee. Riportare il risultato arrotondando al **terzo decimale** (si ricordi l'uso della virgola per i decimali).