Simulazione d'esame 13/01/2025 - Prova MATLAB
(punti 12)

Contrassegnare il proprio canale di afferenza: [A] [B]
Cognome e nome: _____
Numero di matricola: _____
AULA: _____
Numero PC: _____

Consegnare il foglio con il testo del compito e un unico script MATLAB (file .m) da intitolare:
letteraMaiuscolaDelCanale_Cognome_Nome_simulazione.m
Esempio: A_Bertoldo_Alessandra_simulazione.m

Caricamento dati

Scaricare il file simulazione_20250113.mat e incollarlo nella cartella di lavoro.
Il file simulazione_20250113.mat contiene le seguenti variabili

- data: matrice [1000 soggetti x 7 variabili] con le seguenti 7 colonne
    1. Età in anni (**age**)
    2. Peso in kg (**weight**)
    3. Altezza in cm (**height**)
    4. Frequenza cardiaca in bpm (**heart_rate**)
    5. Pressione sistolica alla baseline in mmHg (**systolic_blood_pressure_baseline**)
    6. Assunzione di un farmaco (**medication**) codificata come un indicatore sì/no tale che "1" significa che il paziente assume il farmaco e "0" che non lo assume.
    7. Pressione sistolica al follow-up in mmHg (**systolic_blood_pressure_followup**)

- labels: cell array di 7 elementi, ciascuno contenente l'etichetta della colonna corrispondente nella matrice data (ovvero le stringhe tra parentesi e in font monospace sopra riportate).
- units: cell array di 7 elementi, ciascuno contenente l'unità di misura della colonna corrispondente nella matrice data.
- signal_data: matrice [250 segnali x 240 istanti di tempo] contenente 250 segnali di concentrazione di un farmaco nel sangue in mg/dL.
- signal_time: vettore di lunghezza 240 contenente gli istanti di tempo a cui sono stati campionati i segnali in signal_data (unità di misura: minuti).

Esercizio 1. Pulizia dati. Data la matrice di dati `data`, portare a termine la procedura di pulizia dati sotto descritta.

1. Considerando l'intera matrice, individuare e generare le seguenti variabili:
    a. N_nan = numero complessivo di NaN
    b. N_neg = numero complessivo di valori negativi

2. Eliminare le colonne di data con più del 20% (>20%) di valori mancanti (NaN) o non fisiologici (ossia valori negativi) e salvare la matrice così ottenuta in una variabile chiamata data_reduced.

3. Eliminare da data_reduced le righe che hanno almeno un valore mancante o non fisiologico. Salvare la matrice così ottenuta in una variabile chiamata data_reduced_filtered.

Scrivere:

N_nan = _____

N:neg = _____

Dimensione di data_reduced_filtered = _____

2. Test statistici. Data la matrice di dati `data`, considerarne la prima colonna (age).

1. Generare le seguenti variabili con i rispettivi valori:
   Media_age
   Mediana_age
   Moda_age
   Curtosi_age
   Diff_curtosi = differenza tra il valore riportato in curtosi_age e il valore della curtosi per una variabile gaussiana

2. Applicare un test statistico per testare la gaussianità della variabile age (usare un livello di significatività pari $\alpha = 0.05$). Salvare il p_value nella variabile:

   p_gauss

3. Scrivere quale è l'ipotesi nulla testata e scrivere il motivo per cui è accetta o rifiutata

   H0: _____

   _____

3. Regressione lineare. Data la matrice di dati data, considerare le colonne 1, 5, 6 (age, systolic_blood_pressure_baseline, medication) come variabili indipendenti X e la colonna 7 (systolic_blood_pressure_followup) come variabile dipendente Y. Si vuole descrivere la variabile dipendente tramite in modello di regressione lineare utilizzando le colonne 1,5 e 6 come predittori

1. Dopo aver deciso se inserire o meno l'intercetta nel modello, stimare i parametri del modello (beta_hat) senza usare funzioni Matlab (ossia usando la formula esplicita vista a lezione)

2. ottenere le precisioni di stima dei parametri espresse come SE sapendo che la varianza dell'errore di misura è da stimare (σ^2)

3. Calcolare e riportare di seguito l'intervallo di confidenza al 95% della stima del parametro del modello relativo alla variabile medication

Scrivere:

stima del parametro relativo alla variabile medication = _____

SE del parametro relativo alla variabile medication = _____

Min dell'intervallo di confidenza del punto 3 (2 cifre decimali) = _____

4. Clustering. Considerando la matrice di segnali signal_data (unità di misura: mg/dL), acquisiti ai tempi signal_time (unità di misura: minuti), si vuole determinare se 2 oppure 3 sia il numero di cluster K migliore con cui implementare l'algoritmo k-means.

1. Disegnare, in uno stesso grafico, tutte sovrapposte, le tracce contenute nella matrice signal_data. Visualizzare l'unità di misura dell'asse X e dell'asse Y.

2. Inserire rng(42) nel codice ; (per agevolare la correzione)

3. Utilizzare due volte l'algoritmo k-means: la prima con k fissato pari a 2; la seconda con K fissato pari a 3. Entrambe le volte, fare in modo che
    a. La distanza di riferimento sia la distanza euclidea
    b. Il numero di repliche (inizializzazioni diverse) sia 10
    c. Il numero massimo di iterazioni sia 100

4. Riportare i valori di silhouette medi ottenuti per

a. K fissato pari a 2: _____
b. K fissato pari a 3: _____

5. Qual è il numero di cluster migliore tra K=2 e K=3? _____