SOLUZIONE Simulazione 07/01/2026 - Prova MATLAB
(punti 11)

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.m

Esempio: `A_Rossi_Mario.m`

Lo script dovrà iniziare con un commento riportante

- Nome e cognome
- Numero di matricola
- Canale di afferenza

Esempio:

```
% Mario Rossi
% 999999
% canale A
clc
…
```

Per consegnare utilizzare il commando `consegna('nome_file.m')` nella command window di MATLAB (non nello script). Attenzione ad includere nella stringa l'estensione `.m` del file.

Il comando consegna richiede una login. Indifferentemente, potete usare (ma usate sempre la stessa per tutta la durata del compito):

- **Il vostro account DEI.**
- **Oppure l'account temporaneo teXX, ueXX, ieXX, daXX corrispondente alla vostra postazione, con la password provvisoria (come in laboratorio).**

Caricamento dati

Scaricare il file `simulazione_20260107.mat` e incollarlo nella cartella di lavoro.

Il file `simulazione_20260107.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).

1. Regressione lineare. Data la matrice di dati `data`, considerare le colonne 1, 5, 6 (`age`, `systolic_blood_pressure_baseline`, `medication`) come variabili indipendenti e la colonna 7 (`systolic_blood_pressure_followup`) come variabile dipendente.

1. Stimare, senza usare funzioni Matlab (ossia usando la formula esplicita vista a lezione per lo stimatore ai minimi quadrati lineari), i parametri del modello, intercetta inclusa, inserirli in una variabile `beta_hat` e riportarne i valori di seguito
    a. Stima del parametro relativo a `age`: 1.9914
    b. Stima del parametro relativo a `systolic_blood_pressure_baseline`: 1.0163
    c. Stima del parametro relativo a `medication`: -0.8661
    d. Stima dell'intercetta: -126.3797

2. Calcolare e riportare la stima della varianza a posteriori $\hat{\sigma}^2$: 23.7484

3. Calcolare e riportare di seguito l'intervallo di confidenza al 95% intorno alla stima del parametro relativo alla variabile `medication`: [-1.47, -0.26] (utilizzando il termine moltiplicativo 1.96; va bene anche utilizzare 2)

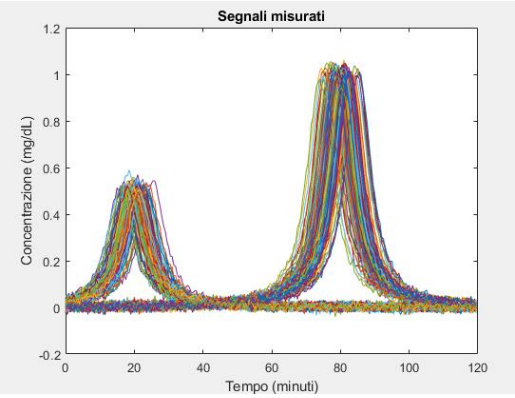
2. Test statistici. Data la matrice di dati `data`, considerarne la prima colonna (`age`).

1. Riportarne la skewness e la curtosi e dire se sarebbero compatibili con la gaussianità di `age` (rispondere semplicemente "sì" o "no").
    a. Skewness di `age`: -0.0167. Compatibile? Sì.
    b. Curtosi di `age`: 3.0585. Compatibile? Sì.

2. A prescindere da quanto concluso al punto precedente, applicare l'opportuno test statistico per confrontare la sua media con il valore medio dell'età in Italia, ovvero 46.6 anni (usare un livello di significatività pari $\alpha = 0.05$).
    a. Riportare il p-value ottenuto: 0
    b. È possibile rifiutare l'ipotesi nulla? Giustificare brevemente la risposta.
    Sì, è possibile perché il p-value restituito da MATLAB è così piccolo da essere stato approssimato a 0 che è < 0.05 (bonus: si può concludere che la media della variabile age è significativamente diversa dal valore 46.6 anni).

3. 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`.



2. Inserire l'istruzione `rng(42);` (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: 0.8954
   b. K fissato pari a 3: 0.7866

5. Qual è il numero di cluster migliore tra K=2 e K=3? 2 (bonus: perché il valore di silhouette medio è più alto / più vicino a 1).

4. 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 riportare il numero di
   a. valori mancanti: 464
   b. valori negativi: 102

2. Eliminare le colonne con più del 20% di valori mancanti o negativi e salvare la matrice così ottenuta in una variabile chiamata `data_reduced`.
   a. Dimensioni di `data_reduced`: 1000x5

3. Eliminare da `data_reduced` le righe che hanno almeno un valore mancante o negativo. Salvare la matrice così ottenuta in una variabile chiamata `data_reduced_filtered`.
   a. Dimensioni di `data_reduced_filtered`: 938x5