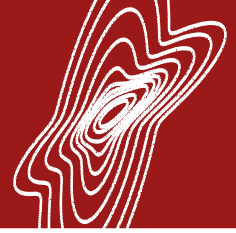


METODI STATISTICI PER LA BIOINGEGNERIA (B)

SIMULAZIONE ESAME, PARTE DI TEORIA

A.A. 2025-2026

Prof. Martina Vettoretti

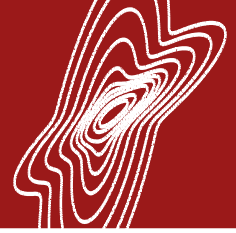


MODALITA' D'ESAME



Esame scritto composto da 3 parti:

- Parte 1 (durata 20 minuti)
 - 10 domande a risposta multipla (1 sola risposta giusta) da svolgere con Moodle Esami
 - 1 punto per ogni risposta giusta, -0.33 per ogni errore, 0 per ogni risposta non data
 - Punteggio massimo: 10 punti
 - Sbarramento: si passa alla parte 2 se si prendono almeno 4.5 punti nella parte 1
- Parte 2 (durata 30 minuti)
 - 2 domande aperte
 - Punteggio massimo: 11 punti
- Parte 3 (durata 60 minuti)
 - 4 esercizi Matlab da svolgere al calcolatore
 - Punteggio massimo: 11 punti
- Voto finale: somma dei punteggi ottenuti nella parte 1, 2 e 3 (max 32).

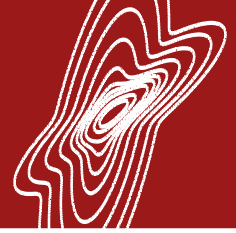


APPELLI D'ESAME



- Primo appello: 22 gennaio 2025
 - 1° turno alle ore 13:00
 - 2° turno alle ore 15:30
- Secondo appello: 10 febbraio 2025
 - 1° turno alle ore 13:00
 - 2° turno alle ore 15:30
- Terzo appello: TBD
- Quarto appello: TBD

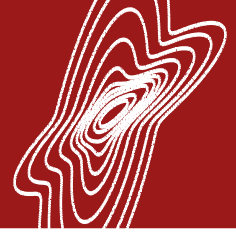
Tutti gli appelli si svolgeranno nelle aule Ue, Te, le e Da.



QUIZ 1



1. Se X , Y e Z sono variabili aleatorie indipendenti, e a e b sono delle quantità deterministiche diverse da zero, allora la covarianza $Cov(X, aY + bZ)$ è:
- A. un valore diverso da 0 pari a: $a \cdot Cov(X, Y) + b \cdot Cov(X, Z)$
 - B. uguale a 0
 - C. un valore diverso da 0 pari a: $a \cdot b \cdot Cov(X, Y + Z)$
 - D. un valore diverso da 0 pari a: $a \cdot b \cdot Cov(X \cdot Y, X \cdot Z)$



QUIZ 2

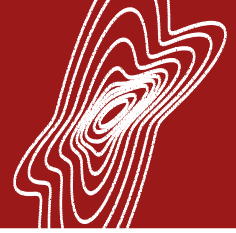


2. Dato il seguente insieme di valori osservati per la variabile X

$\{71 \ 76 \ 80 \ 80 \ 90 \ 100 \ 110 \ 125 \ 130 \ 135\}$

la mediana risulta:

- A. Mediana = 99.7
- B. Mediana = 90
- C. Mediana = 100
- D. Mediana = 95

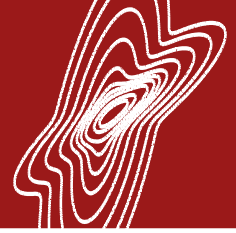


QUIZ 3



3. L'indice di asimmetria o skewness:

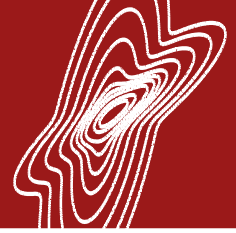
- A. È definito come rapporto tra il momento centrale di ordine 3 e il cubo della deviazione standard.
- B. È definito come rapporto tra il momento centrale di ordine 4 e il quadrato della varianza.
- C. È pari a 3 per una variabile aleatoria normale.
- D. È positivo se la distribuzione è simmetrica.



QUIZ 4



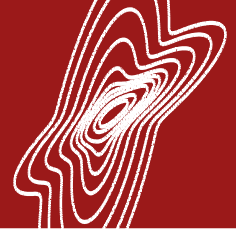
4. Dati due campioni estratti da popolazioni aventi distribuzione normale a media e varianza incognite e potenzialmente diverse, vogliamo testare l'ipotesi nulla H_0 : le medie delle due popolazioni sono uguali. Che test statistico è più opportuno usare?
- A. Il test di Wilcoxon Mann-Whitney
 - B. Il test dei segni
 - C. Il t test di Welch
 - D. Il t test a due campioni per campioni omoschedastici



QUIZ 5



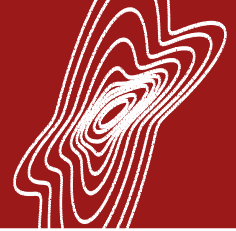
5. Quale dei seguenti test statistici consente il confronto di due campioni appaiati?
- A. Il test ANOVA
 - B. Il t test a due campioni per campioni omoschedastici
 - C. Il test di Wilcoxon Mann-Whitney
 - D. Il test dei ranghi con segno



QUIZ 6



6. Consideriamo lo stimatore ai minimi quadrati lineari $\hat{\beta} = (X^T X)^{-1} X^T Y$, dove come di consueto Y è il vettore dei valori dell'outcome da predire e X è la matrice delle variabili indipendenti. Sia $\hat{\sigma}^2$ il valore della varianza dell'errore stimato a posteriori. La matrice di covarianza di $\hat{\beta}$ può essere calcolata come:
- A. $Cov(\hat{\beta}) = \hat{\sigma}^2 X^T X$
 - B. $Cov(\hat{\beta}) = (X^T X)^{-1}$
 - C. $Cov(\hat{\beta}) = \hat{\sigma}^2 (X^T X)^{-1}$
 - D. Nessuna delle precedenti



QUIZ 7

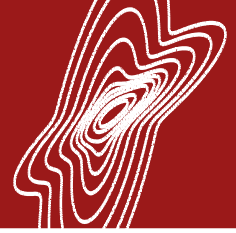


7. Si consideri l'indice Akaike Information Criterion (AIC) dato dall'equazione:

$$AIC = n \cdot \log \left(\frac{SSE}{n} \right) + 2 \cdot p$$

dove n è il numero di osservazioni, p il numero di parametri e SSE la somma dei quadrati dei residui. All'aumentare della complessità del modello, tipicamente:

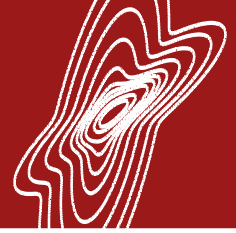
- A. Il primo addendo aumenta, il secondo diminuisce
- B. Il primo addendo diminuisce, il secondo aumenta
- C. Entrambi gli addendi aumentano
- D. Entrambi gli addendi diminuiscono



QUIZ 8



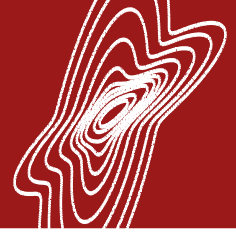
8. La regolarizzazione elastic net:
- A. Azzera sempre almeno uno dei coefficienti del modello di regressione.
 - B. Non azzera mai nessuno dei coefficienti del modello di regressione.
 - C. Combina le proprietà delle regolarizzazioni Ridge e LASSO.
 - D. Non ha nessuna delle precedenti proprietà.



QUIZ 9



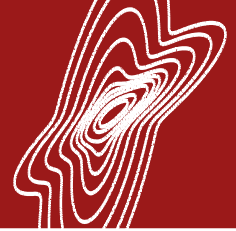
9. Il C-index di un modello di Cox rappresenta la probabilità che:
- A. Risk score maggiori siano assegnati a soggetti aventi minor tempo di sopravvivenza
 - B. Risk score maggiori siano assegnati a soggetti aventi maggior tempo di sopravvivenza
 - C. Risk score minori siano assegnati a soggetti aventi minor tempo di sopravvivenza
 - D. Risk score minori siano assegnati a soggetti censurati



QUIZ 10



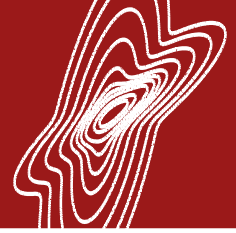
10. Data una matrice di dati X in cui ogni colonna rappresenta una variabile, quali delle seguenti operazioni è necessario fare prima di effettuare l'analisi delle componenti principali (PCA)?
- A. Standardizzare le colonne di X
 - B. Sottrarre la media a ciascuna colonna di X , se queste non sono già a media nulla
 - C. Scegliere il numero di componenti principali
 - D. Tutte le precedenti



DOMANDA APERTA 1 (6 PUNTI)



1. Data una matrice $\mathbf{X} \in \mathbb{R}^{N \times M}$, dove N è il numero di osservazioni ed M il numero di variabili, si descriva il metodo di clustering K-means riportando in particolare:
 - a. la funzione obiettivo che viene utilizzata nel caso di utilizzo della distanza euclidea, specificando il significato di tutti i termini utilizzati
 - b. gli step principali dell'algoritmo iterativo che consente di minimizzare tale funzione obiettivo (non occorre riportare la formula dei centroidi).
2. Supponendo che i dati vengano suddivisi in K cluster, quanti saranno i centroidi e che dimensione avranno?



SOLUZIONE



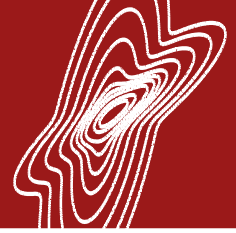
1.a La funzione obiettivo è:

$$\sum_{k=1}^K W(C_k)$$

dove $C_k, k = 1, \dots, K$ sono i K cluster in cui vengono suddivisi i dati e $W(C_k)$ è la variabilità intra-cluster per il cluster k -esimo definita come:

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,j \in C_k} d(x_i, x_j)^2$$

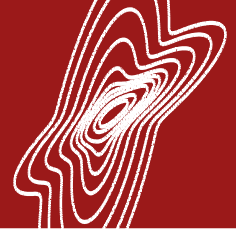
con $d(x_i, x_j)$ distanza euclidea tra le osservazioni x_i e x_j (righe i e j della matrice X) e $|C_k|$ pari alla cardinalità del cluster k -esimo.



SOLUZIONE



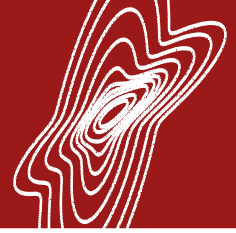
- 1.b Algoritmo iterativo per trovare il minimo della funzione obiettivo:
 1. Ogni osservazione viene assegnata ad uno dei K cluster in modo casuale.
 2. Si calcola il centroide per ciascun cluster come media delle osservazioni che appartengono al cluster.
 3. Si calcola la distanza delle osservazioni dai K centroidi. Ogni osservazione viene assegnata al cluster corrispondente al centroide più vicino.
 4. Si iterano i passi 2 e 3 finché i centroidi non cambiano più.
2. I centroidi saranno K e avranno dimensione $1 \times M$.



DOMANDA APERTA 2 (5 PUNTI)



1. Si scriva l'equazione del modello di regressione logistica univariata, ovvero la sua versione più semplice in cui compare una sola variabile indipendente X_1 , specificando il significato dei termini coinvolti.
2. Si scriva la formula dell'odds ratio associato alla variabile X_1 .
3. Come si interpretano i valori dell'odds ratio associato alla variabile X_1 ?



SOLUZIONE



1. Formula della regressione logistica con una variabile indipendente:

$$p = \frac{e^{\beta_0 + \beta_1 X_1}}{1 + e^{\beta_0 + \beta_1 X_1}}$$

- p : probabilità che la variabile di outcome binaria, Y , sia pari a 1
- X_1 : variabile indipendente
- β_0 : intercetta
- β_1 : coefficiente associato alla variabile X_1 .

2. L'odds ratio associato a X_1 è: e^{β_1}

3. Interpretazione dell'odds ratio:

- se $e^{\beta_1} = 1 \rightarrow$ la variabile X_1 non ha effetto sulla probabilità di Y
- Se $e^{\beta_1} > 1 \rightarrow$ all'aumentare di X_1 aumenta la probabilità che Y sia 1
- se $e^{\beta_1} < 1 \rightarrow$ all'aumentare di X_1 diminuisce la probabilità che Y sia 1