METODI STATISTICI PER LA BIOINGEGNERIA (B)

SIMULAZIONE ESAME, PARTE DI TEORIA

A.A. 2024-2025

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 20 minuti)
  - 2 domande aperte
  - 6 punti per ogni domanda
  - Punteggio massimo: 12 punti
- Parte 3 (durata 60 minuti)
  - 4 esercizi Matlab da svolgere al calcolatore
  - 3 punti per ogni esercizio
  - Punteggio massimo: 12 punti
- Voto finale: somma dei punteggi ottenuti nella parte 1, 2 e 3 (max 34).

APPELLI D'ESAME

- Primo appello: 24 gennaio 2025
  - 1° turno alle ore 13:00
  - 2° turno alle ore 15:30
- Secondo appello: 14 febbraio 2025
  - 1° turno alle ore 13:00
  - 2° turno alle ore 15:30
- Terzo appello: 9 luglio 2025
  - Unico turno alle ore 14:00
- Quarto appello: 17 settembre 2025
  - Unico turno alle ore 10:00

Tutti gli appelli si svolgeranno nelle aule Ue, Te, Ie 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 = 95
C. Mediana = 90
D. Mediana = 100

QUIZ 3

3. Il campo di variazione o range di un insieme di dati:
   A. è un valore sempre negativo
   B. è un indice di posizione
   C. è pari alla differenza tra il valore massimo e il valore minimo
   D. è un indice insensibile alla presenza di outlier.

QUIZ 4

4. Sia X una variabile aleatoria discreta che assume n valori distinti $x_1, x_2, \ldots, x_n$. Siano f(x) la densità discreta di X e F(x) la funzione di ripartizione di X. Quale delle seguenti condizioni deve essere soddisfatta?
   A. $f(x_i) > 1 \; \forall i = 1, \ldots, n$
   B. $F(x_n) = +\infty$
   C. $\sum_{i=1}^{n} F(x_i) = 1$
   D. $\sum_{i=1}^{n} f(x_i) = 1$

QUIZ 5

5. 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 6

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

QUIZ 7

7. 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 t-test di Welch
   B. Il test di Wilcoxon Mann-Whitney
   C. Il test dei segni
   D. Il log-rank test

QUIZ 8

8. 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
   C. Scegliere il numero di componenti principali
   D. Tutte le precedenti

QUIZ 9

9. 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 10

10. 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.

DOMANDE APERTA 1

a. 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 la funzione obiettivo che viene utilizzata nel caso di utilizzo della distanza euclidea, specificando il significato di tutti i termini utilizzati.

b. Supponendo che i dati vengano suddivisi in K cluster, quanti saranno i centroidi e che dimensione avranno?

SOLUZIONE

a. L'algoritmo K-means divide le osservazioni in K cluster disgiunti, $C_i, i = 1, \dots, K$, con K prestabilito. La partizione ottima viene scelta minimizzando la seguente funzione obiettivo:

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

dove $W(C_k)$ è la variabilità intra-cluster per il cluster k definita come:

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

con $d(\boldsymbol{x_i}, \boldsymbol{x_j})$ distanza euclidea tra le osservazioni x_i e x_j (righe i e j della matrice $\boldsymbol{X}$).

SOLUZIONE

a. In pratica la soluzione viene trovata con il seguente algoritmo iterativo:
   1. Ogni osservazione viene assegnata ad uno dei K cluster in modo casuale.
   2. Si calcola il centroide per ciascun cluster mediando le variabili 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.

b. I centroidi saranno K e avranno dimensione 1xM.

DOMANDA APERTA 2

a. 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.

b. Si scriva la formula dell'odds ratio associato alla variabile X_1.

c. Spiegare come settando una soglia si può risolvere un problema di classificazione binaria tramite regressione logistica.

SOLUZIONE

a. 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 l'outcome sia pari a 1
- X_1: variabile indipendente
- β_0: intercetta
- β_1: coefficiente associato alla variabile X_1. Rappresenta di quanto aumenta la funzione logit di p quando la variabile X_1 aumenta di una unità.

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

c. Si sceglie una soglia th. Con il modello di regressione logistica si calcola la probabilità della classe 1, p. Se $p < th$ si classifica il dato come appartenente alla classe 0, se $p \geq th$ il dato viene assegnato alla classe 1.