Cognome: .... Nome: .... Matricola: ....

Tipologia d'esame: ☐ 12 CFU ☐ 15 CFU

Prova scritta di ASM 12CFU e 15CFU - Modulo Analisi Esplorativa del 21.04.2017

La durata della prova è di 75 minuti.
Si svolgano gli esercizi 1, 2 e 3 riportando il risultato dove indicato.

Esercizio 1 (Punti: 9)

Si consideri la seguente matrice di correlazione $\underset{3\times 3}{R} = \begin{bmatrix} 1 & 1/2 & 1/2 \\ 1/2 & 1 & 2/3 \\ 1/2 & 2/3 & 1 \end{bmatrix}$.

1.a) Riportare l'indice di variabilità relativo (arrotondare al secondo decimale)

$= \ldots\ldots\ldots\ldots$

1.b) Sapendo che $s_{11} = 4$, $s_{22} = 9$ e $s_{33} = 1$, determinare la matrice di varianze/covarianze

$$\underset{3\times 3}{S} = \begin{bmatrix} \ldots\ldots & \ldots\ldots & \ldots\ldots \\ \ldots\ldots & \ldots\ldots & \ldots\ldots \\ \ldots\ldots & \ldots\ldots & \ldots\ldots \end{bmatrix}.$$

1.c) Sia $\underset{n\times 3}{\tilde{X}}$ la matrice dei dati centrati a cui corrisponde la matrice di correlazione $\underset{3\times 3}{R}$ riportata nel testo dell'esercizio. Determinare l'angolo (espresso in gradi) tra $\underset{n\times 1}{\tilde{x}_1}$ e $\underset{n\times 1}{\tilde{x}_2}$:

$= \ldots\ldots\ldots\ldots$

1.d) Determinare gli autovalori λ_1, λ_2 e λ_3 associati alla matrice di correlazione $\underset{3\times 3}{R}$ (arrotondare al secondo decimale)

$$\lambda_1 = \ldots\ldots, \quad \lambda_2 = \ldots\ldots, \quad \lambda_3 = \ldots\ldots$$

1.e) Riportare la proporzione di varianza spiegata dalle prime due componenti principali calcolate a partire dalla matrice di correlazione $\underset{3\times 3}{R}$ (arrotondare al secondo decimale)

$= \ldots\ldots\ldots\ldots$

1.f) Sia $\underset{n\times 3}{Z}$ la matrice dei dati standardizzati a cui corrisponde la matrice di correlazione $\underset{3\times 3}{R}$ riportata nel testo dell'esercizio. Calcolare la correlazione tra la prima colonna $\underset{n\times 1}{\tilde{z}_1}$ di $\underset{n\times 3}{Z}$ e i punteggi $\underset{n\times 1}{y_1}$ della prima componente principale di $\underset{n\times 3}{Z}$ (arrotondare al secondo decimale):

$= \ldots\ldots\ldots$

Esercizio 2 (Punti: 7)

Si consideri la seguente matrice di distanze relativa a tre unità statistiche a, b e c:

$\underset{3\times 3}{D} =$

	(a)	(b)	(c)
(a)	0		
(b)	7	0	
(c)	6	5	0

2.a) Se utilizziamo un algoritmo gerarchico agglomerativo, le unità (a) e (b) vengono messe assieme nel gruppo (a, b). Aggiornare la matrice delle distanze utilizzando il metodo del legame singolo:

	(a,b)	(c)
(a,b)	0	
(c)	...	0

2.b) Aggiornare la matrice delle distanze utilizzando il metodo del legame medio:

	(a,b)	(c)
(a,b)	0	
(c)	...	0

Si consideri la seguente matrice dei dati relativa a 4 unità statistiche

$$\underset{4\times 2}{X} = \begin{bmatrix} 2.5 & 2.5 \\ 4.5 & 8.5 \\ 6.5 & 1.5 \\ 3.5 & 3.5 \end{bmatrix}$$

2.c) Si calcoli la matrice delle distanze $\underset{4\times 4}{D}$ per la matrice $\underset{4\times 2}{X}$ riportata sopra utilizzando la metrica di Lagrange:

$$\underset{4\times 4}{D} = \begin{bmatrix} \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \end{bmatrix}$$

2.d) Data una generica matrice $\underset{n\times p}{X}$ con vettore delle medie $\underset{p\times 1}{\bar{x}}$ e matrice di varianze/covarianze $\underset{p\times p}{S}$, si riporti la definizione della distanza di Mahalanobis $d_M(u_i, \bar{x})$ tra l'i-sima unità statistica $\underset{1\times p}{u_i'}$ e il baricentro $\underset{1\times p}{\bar{x}'}$.

$d_M(u_i, \bar{x}) =$

Esercizio 3 (Punti: 10)

Si consideri il dataset `mtcars` presente nella libreria `datasets`, che contiene $n = 32$ unità statistiche (automobili) relative alle seguenti 11 variabili:

- *mpg* Miles/(US) gallon
- *cyl* Number of cylinders
- *disp* Displacement (cu.in.)
- *hp* Gross horsepower
- *drat* Rear axle ratio
- *wt* Weight (1000 lbs)
- *qsec* 1/4 mile time
- *vs* V/S
- *am* Transmission (0 = automatic, 1 = manual)
- *gear* Number of forward gears
- *carb* Number of carburetors

3.a) Si consideri la matrice $\underset{32\times 6}{X}$ che contiene solo le seguenti 6 variabili: *mpg*, *disp*, *hp*, *drat*, *wt* e *qsec*. Per ciascuna unità statistica, si calcoli la distanza di Mahalanobis dal baricentro e si riporti il nome delle due marche di automobili con distanza di Mahalanobis superiore a 3.5:

...

...

3.b) Partendo da $\underset{32\times 6}{X}$, calcolare la matrice dei dati standardizzati $\underset{32\times 6}{Z}$. Calcolare l'indice di Calinski and Harabasz (CH) per un numero di gruppi K da 2 a 8, impostando per ciascun valore di K `set.seed(123)` prima di eseguire l'algoritmo delle K-medie (specificando `algorithm = Lloyd`). Riportare per ciascun valore di K il rispettivo valore dell'indice CH (arrotondando al secondo decimale).

K	2	3	4	5	6	7	8
Indice CH	....	....	....	....	....	....	....

3.c) Sulla base di $\underset{32\times 6}{Z}$, calcolare la matrice delle distanze $\underset{32\times 32}{D}$ utilizzando la metrica Euclidea, ed effettuare l'analisi dei cluster gerarchica utilizzando il legame completo, ricavandone 3 gruppi. Calcolare, arrotondando al secondo decimale, il valore medio della silhouette per i tre gruppi individuati (utilizzando il comando `silhouette` presente nella libreria `cluster`).

	Valore medio della Silhouette
Gruppo 1	
Gruppo 2	
Gruppo 3	