

Cognome: Nome: Matricola:

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 e 2 riportando il risultato dove indicato.

Esercizio 1 (Punti: 9)

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

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

=

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

$$S_{3 \times 3} = \begin{bmatrix} & & \\ \dots & \dots & \dots \\ \dots & \dots & \dots \\ \dots & \dots & \dots \end{bmatrix}.$$

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

$\begin{matrix} n \times 1 & n \times 1 \end{matrix}$

=

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

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

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

=

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

$\begin{matrix} n \times 1 & n \times 3 & n \times 1 \end{matrix}$

=

[1] 0.39

```
##      [,1] [,2] [,3]
## [1,]    4    3    1
## [2,]    3    9    2
## [3,]    1    2    1

## [1] 60

## [1] 2.12 0.55 0.33

## [1] 0.89

## [1] -0.78
```

Esercizio 2 (Punti: 7)

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

$$D_{3 \times 3} =$$

	(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

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

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

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

```
##      u1 u2 u3
## u2    6
## u3    4  7
## u4    1  5  3
```

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

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

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 $X_{32 \times 6}$ 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:

...

...

```
##      Merc 230 Maserati Bora
##           9             31
```

3.b) Partendo da $X_{32 \times 6}$, calcolare la matrice dei dati standardizzati $Z_{32 \times 6}$. 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

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7]
## KS  2.00  3.00  4.00  5.0  6.0  7.00  8.00
##      33.92 24.85 23.33 17.8 17.1 16.35 13.93
```

3.c) Sulla base di $Z_{32 \times 6}$, calcolare la matrice delle distanze $D_{32 \times 32}$ 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`).

```
## Loading required package: cluster

##      1      2      3
## 0.22 0.26 0.40
```

	Valore medio della Silhouette
Gruppo 1	
Gruppo 2	
Gruppo 3	