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

Tipologia d'esame: ☐ 12 CFU ☐ 15 CFU

Prova scritta di ASM 12CFU e 15CFU - Modulo Analisi Esplorativa del 27.06.2017

La durata della prova è di 70 minuti.
Si svolgano gli esercizi 1, 2 e 3 riportando il risultato dove indicato.

Esercizio 1 (Punti: 10)

Si consideri la seguente matrice di varianze/covarianze $\underset{3\times 3}{S} = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 9 & 0 \\ 0 & 0 & 1 \end{bmatrix}$.

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

$$= \ldots\ldots\ldots\ldots$$

1.b) Determinare la matrice di correlazione

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

1.c) Sia $\underset{n\times 3}{\tilde{X}}$ la matrice dei dati centrati a cui corrisponde la matrice di correlazione $\underset{3\times 3}{R}$ calcolata al punto precedente. 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 varianze/covarianze $\underset{3\times 3}{S}$ (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 $\underset{3\times 3}{S}$ (arrotondare al secondo decimale)

$$= \ldots\ldots\ldots\ldots$$

1.f) *Dimostrazione*

Dimostrare che la matrice di centramento $\underset{n\times n}{H}$ è idempotente, giustificano tutti i passaggi.

```
## [1] 1

##      [,1] [,2] [,3]
## [1,]    1    0    0
## [2,]    0    1    0
## [3,]    0    0    1

## [1] 90

## [1] 9 4 1

## [1] 0.93
```

Esercizio 2 (Punti: 8)

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

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

a. 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 Manhattan, riportando la definizione di distanza di Manhattan $d_1(u_i, u_l)$ tra due unità statistiche $\underset{1\times p}{u_i'}$ e $\underset{1\times p}{u_l'}$ per una generica matrice di dati $\underset{n\times p}{X}$.

$$\underset{4\times 4}{D} = \begin{bmatrix} \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix} \qquad d_1(u_i, u_l) =$$

```
##    u1 u2 u3
## u2  8
## u3  5  9
## u4  2  6  5
```

b. Utilizzando il metodo del legame completo con riferimento alla matrice delle distanze $\underset{4\times 4}{D}$ calcolata al punto a., si riportino i gruppi (*clusters*) identificati tagliando il dendogramma ad altezza $c = 6$.

```
## u1 u2 u3 u4
##  1  2  1  1
```

c. Con riferimento a due gruppi (*clusters*) di osservazioni G_I e G_L, si riporti la definizione di distanza tra i due gruppi secondo il metodo del legame medio.

$d_{medio}(G_I, G_L) =$

d. Si completi la seguente tabella:

Legame	Inversione (Si/No)	Trasformazioni monotone (Invariante/Non invariante)	Interpretazione taglio (Si/No)
Singolo			
Completo			
Medio			
Centroide			

Esercizio 3 (Punti: 8)

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 4 marche di automobili con distanza di Mahalanobis inferiore a 1.5:

...

...

...

...

```
##  Datsun 710  Merc 450SL Merc 450SLC   Fiat X1-9
##           3          13          14          26
```

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	....	....	....	....	....	....	....

```
##      [,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 $\underset{32\times 6}{Z}$, calcolare la matrice delle distanze $\underset{32\times 32}{D}$ utilizzando la metrica Euclidea, ed effettuare l'analisi dei cluster gerarchica utilizzando il legame singolo, 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
## -0.02  0.00  0.82
```

	Valore medio della Silhouette
Gruppo 1	
Gruppo 2	
Gruppo 3	