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

Prova scritta di ASM - Modulo Analisi Esplorativa del 14.02.2017

La durata della prova è di 90 minuti.
Si svolgano gli esercizi A e B riportando il risultato dove indicato.

Esercizio A (Punti: 14)

1. *Decomposizione Spettrale e Analisi delle Componenti Principali*

Alla matrice di varianze/covarianze relativa a $\underset{n\times p}{X}$ sono associati i seguenti autovalori e autovettori normalizzati: $\lambda_1 = 9$, $\lambda_2 = 6$, $\underset{2\times 1}{v_1} = \begin{bmatrix} 1/\sqrt{5} \\ 2/\sqrt{5} \end{bmatrix}$ e $\underset{2\times 1}{v_2} = \begin{bmatrix} 2/\sqrt{5} \\ -1/\sqrt{5} \end{bmatrix}$.

a. Determinare la matrice di varianze/covarianze $\underset{p\times p}{S} = \begin{bmatrix} \ldots\ldots & \ldots\ldots \\ \ldots\ldots & \ldots\ldots \end{bmatrix}$

b. Riportare

- varianza totale = ... e generalizzata = ...
- l'indice di variabilità relativo (<u>arrotondare al secondo decimale</u>) = ...

c. Determinare, <u>arrotondando al secondo decimale</u>, $\underset{p\times p}{S^{1/2}} = \begin{bmatrix} \ldots\ldots & \ldots\ldots \\ \ldots\ldots & \ldots\ldots \end{bmatrix}$

d. Calcolare la correlazione tra la seconda colonna $\underset{n\times 1}{\tilde{x}_2}$ di $\underset{n\times p}{\tilde{X}}$ e i punteggi $\underset{n\times 1}{y_2}$ della seconda componente principale, <u>arrotondando al secondo decimale</u>.

= ....

```
##      [,1] [,2]
## [1,]  6.6  1.2
## [2,]  1.2  8.4
```

```
## [1] 54
```

```
## [1] 15
```

```
## [1] 0.97
```

```
##      [,1] [,2]
## [1,] 2.56 0.22
## [2,] 0.22 2.89
```

```
## [1] -0.38
```

2. *Distanze e Cluster Analysis*

a. Per una generica matrice di dati $\underset{n\times p}{X}$, si riporti la definizione di distanza di Minkowski $d_m(u_i, u_l)$ di ordine $m \geq 1$ tra due unità statistiche $\underset{1\times p}{u_i'}$ e $\underset{1\times p}{u_l'}$.

$d_m(u_i, u_l) =$

b. Per una generica matrice di distanze $\underset{n\times n}{D}$ con elemento di posizione (i,l) pari a $d(u_i, u_l)$, si riporti la definizione di legame medio tra due gruppi G_1 e G_2.

$d(G_1, G_2) =$

c. Si calcoli il valore dell'indice di similarità di Jaccard per il seguente esempio:

$$\begin{bmatrix} 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} u_1' \\ u_2' \end{bmatrix}$$

$s_J(u_1, u_2) =$

3. *Analisi Fattoriale*

a. Si riportino le assunzioni del modello fattoriale con k fattori $\underset{p\times 1}{x} = \underset{p\times k}{\Lambda}\underset{k\times 1}{f} + \underset{p\times 1}{u}$.

- $\mathbb{E}(\underset{p\times 1}{x}) =$

- $\mathbb{E}(\underset{k\times 1}{f}) =$, $\mathbb{C}\text{ov}(\underset{k\times 1}{f}) =$

- $\mathbb{E}(\underset{p\times 1}{u}) =$, $\mathbb{C}\text{ov}(\underset{p\times 1}{u}) =$

- $\mathbb{C}\text{ov}(\underset{p\times 1}{u}, \underset{k\times 1}{f}) =$

b. La seguente tabella riporta la stima della matrice di pesi fattoriali $\underset{5\times 2}{\hat{\Lambda}}$ di un modello fattoriale a due fattori ottenuta a partire dalla matrice di correlazione $\underset{5\times 5}{R}$.

$$\underset{5\times 2}{\hat{\Lambda}} = \begin{bmatrix} .56 & ? \\ .78 & -.53 \\ .65 & .75 \\ .94 & -.10 \\ ? & -.54 \end{bmatrix} \begin{array}{c} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{array}$$

Sapendo che le varianze specifiche di x_1 e x_5 sono pari a $\hat{\psi}_1 = 0.02$ e $\hat{\psi}_5 = 0.07$, determinare, <u>arrotondando al secondo decimale</u>,

$$\hat{\lambda}_{12} = \ldots\ldots\ldots, \qquad \hat{\lambda}_{51} = \ldots\ldots\ldots$$

```
## [1] 0.82
```

```
## [1] 0.8
```

4. *Dimostrazione*

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

Esercizio B (Punti: 13)

Si consideri il dataset `iris` presente nella libreria `datasets` che contiene $n = 150$ unità statistiche (fiori di genere Iris) relative alle 4 variabili

- *Sepal.Length* (lunghezza dei sepali)
- *Sepal.Width* (larghezza dei sepali)
- *Petal.Length* (lunghezza dei petali)
- *Petal.Width* (larghezza dei petali)

più l'ultima colonna *Species* che specifica la specie (con modalità *setosa*, *versicolor* e *virginica*).

1. Si consideri la matrice $\underset{150\times 4}{X}$ che contiene le seguenti variabili: *Sepal.Length*, *Sepal.Width*, *Petal.Length* e *Petal.Width*. si calcoli il quadrato della distanza di Mahalanobis di ciascuna unità statistica u_i' dal baricentro $\bar{x}'$ e si riporti il valore medio e il valore massimo, arrotondando i calcoli al secondo decimale.

```
## [1] 4
```

```
## [1] 13.19
```

$$\frac{1}{150}\sum_{i=1}^{150} d_M^2(u_i,\bar{x}) = \dots\dots\dots\dots \qquad \max_{i=1,\dots,150}\{d_M^2(u_i,\bar{x})\} = \dots\dots\dots\dots$$

2. Per la matrice di dati $\underset{150\times 4}{X}$, utilizzare l'algoritmo delle K-medie (specificando `algorithm = "Lloyd"`) per formare $K = 3$ gruppi, inizializzando i centroidi con le osservazioni di riga 30, 80 e 110, ed eseguendo l'algoritmo una sola volta. Riportare

a. la numerosità dei 3 gruppi ottenuti;

b. i valori della tabella a doppia entrata che incrocia la classificazione ottenuta e la variabile *Species*;

c. il valore medio della silhouette (arrotondando al secondo decimale) per i tre gruppi (utilizzando il comando `silhouette` presente nella libreria `cluster`) considerando come matrice delle distanze quella ottenuta con la metrica Euclidea.

a. Numerosità gruppo 1 = ... , gruppo 2 = ... , gruppo 3 = ...

b.

	setosa	versicolor	virginica
gruppo 1	......	......	......
gruppo 2	......	......	......
gruppo 3	......	......	......

c. Valore medio silhouette per il gruppo 1 = ... , gruppo 2 = ... , gruppo 3 = ...

```
##
##  1  2  3
## 50 61 39
```

```
##
```

```
##     setosa versicolor virginica
##   1     50          0         0
##   2      0         47        14
##   3      0          3        36
```

```
## Loading required package: cluster
```

```
##    1    2    3
## 0.80 0.42 0.44
```

3. Partendo dalla matrice $\underset{150\times 4}{X}$ determinata al punto a., si calcoli la matrice dei dati standardizzati $\underset{150\times 4}{Z}$. Si conduca l'analisi delle componenti principali basata su $\underset{150\times 4}{Z}$ utilizzando il comando `prcomp()`, riportando (arrotondando alla seconda cifra decimale)

a. la proporzione di varianza spiegata dalle prime due componenti principali

b. l'equazione del punteggio (*score*) y_{i2} della seconda componente principale per l'i-sima unità statistica

c. la correlazione tra i punteggi della prima componente principale e la prima colonna di Z

```
## [1] 0.96
```

```
##                PC2
## Sepal.Length -0.38
## Sepal.Width  -0.92
## Petal.Length -0.02
## Petal.Width  -0.07
```

```
## [1] 0.89
```

```
## [1] 0.8901688
```

a. Proporzione di varianza spiegata dalle prime due componenti principali: ......

b. Punteggio y_{i2} della prima componente principale per l'i-sima unità statistica:

$y_{i2} = \ldots\ldots\ldots \cdot z_{i1} + \ldots\ldots\ldots \cdot z_{i2} + \ldots\ldots\ldots \cdot z_{i3} + \ldots\ldots\ldots \cdot z_{i4}$

c. Correlazione tra i punteggi della prima componente principale e la prima colonna di Z: ......

PUNTEGGI

ESERCIZIO A (14 punti): 1/1/1/2/1.5/1.5/1/1.5/1.5/2

ESERCIZIO B (13 punti): 2/1/2/2/2/2/2