3 Luglio 2018 - Analisi Esplorativa

Cognome: .....

Nome: .....

Matricola: .....

Tipologia d'esame: ☐ 12 CFU ☐ 15 CFU

Prova scritta - versione B

Si svolgano gli esercizi riportando il risultato dove indicato. Durata: 60 minuti

---

Esercizio 1 (5 punti)

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*).

Si consideri la matrice $\underset{150\times 4}{X}$ che contiene le seguenti variabili: *Sepal.Length*, *Sepal.Width*, *Petal.Length* e *Petal.Width*. Sulla base di questa matrice, 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:

gruppo 1 = ... , gruppo 2 = ... , gruppo 3 = ...

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

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

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.

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
```

Esercizio 2 (6 punti)

a. Riportare le seguenti definizioni (in forma matriciale), specificando tutte le quantità coinvolte:

Vettore delle medie

Matrice di centramento

Matrice dei dati centrati

Matrice di varianze/covarianze

Matrice dei dati standardizzati

Matrice dei dati ortogonalizzati

b. Si dimostri che $\text{tr}(S) = \sum_{j=1}^{p} \lambda_j$, dove $\lambda_1, \ldots, \lambda_p$ sono gli autovalori di S.

c. Elencare le proprietà di una metrica.

3 Luglio 2018 - Analisi Esplorativa

Cognome: .....

Nome: .....

Matricola: .....

Tipologia d'esame: ☐ 12 CFU ☐ 15 CFU

Prova scritta - versione B

Si svolgano gli esercizi riportando il risultato dove indicato. Durata: 60 minuti

---

Esercizio 3 (6 punti)

Data la matrice dei dati $X = \begin{bmatrix} 5.3 & 2.0 \\ 3.9 & 2.8 \\ -2.3 & 2.6 \\ 2.6 & 3.6 \\ 2.5 & 3.3 \end{bmatrix}$

si calcolino

a. la proporzione di varianza spiegata dalla prima componente principale basata sui dati centrati, arrotondando il risultato alla terza cifra decimale:

.....

```
## [1] 0.956
```

b. la proporzione di varianza spiegata dalla prima componente principale basata sui dati standardizzati, arrotondando il risultato alla terza cifra decimale:

.....

```
## [1] 0.582
```

c. i punteggi delle cinque unità statistiche per la prima e la seconda componente principale basata sui dati centrati, arrotondando il risultato alla seconda cifra decimale:

prima	seconda
.....	.....
.....	.....
.....	.....
.....	.....
.....	.....

```
##      Comp.1 Comp.2
## [1,]   2.93  -0.75
## [2,]   1.50   0.00
## [3,]  -4.69  -0.44
## [4,]   0.17   0.75
## [5,]   0.08   0.44
```

d. il vettore delle medie per la matrice che ha come colonne i due vettori di punteggi relativi alle prime due componenti principali:

.....

e. la matrice di varianza/covarianza per la matrice che ha come colonne i due vettori di punteggi relativi alle prime due componenti principali, arrotondando il risultato alla seconda cifra decimale:

. . . .
. . . .

```
##       [,1]  [,2]
## [1,] 6.569 0.000
## [2,] 0.000 0.302
```

f. si stabilisca se il vettore dei punteggi della prima componente principale ha maggiore correlazione (in valore assoluto) con la prima o la seconda colonna di X:

. . . . . .

```
## [1] 0.9999679
```

Esercizio 4 (2 punti)

Presi i punti $u_1' = (5, 0)$, $u_2' = (6, 2)$ e $u_3' = (8, 4)$, si verifichi che il quadrato della distanza Euclidea non soddisfa la diseguaglianza triangolare.

Esercizio 5 (7 punti)

Si considererì la seguente matrice di correlazione calcolata sulla base di $n = 50$ osservazioni:

	Murder	Assault	UrbanPop	Rape
Murder	1.000	0.802	0.070	0.564
Assault	0.802	1.000	0.259	0.665
UrbanPop	0.070	0.259	1.000	0.411
Rape	0.564	0.665	0.411	1.000

a. Sulla base dalla matrice di correlazione, si stimi il modello fattoriale con $k = 1$ fattori utilizzando il metodo della massima verosimiglianza senza effettuare alcuna rotazione. Arrontondando al terzo decimale, si riportino le stime dei pesi fattoriali:

	Stima punteggi fattoriali
Murder	
Assault	
UrbanPop	
Rape	

```
##       [,1]
## [1,] 0.818
## [2,] 0.979
## [3,] 0.262
## [4,] 0.683
```

b. Si determini il punteggio fattoriale con il metodo di Thompson (arrotondando alla quarta cifra decimale) per l'unità statistica "Arizona" sapendo che i suoi valori nelle quattro variabili standardizzate sono

	Murder	Assault	UrbanPop	Rape
Arizona	1.2426	0.7828	-0.5209	-0.0034

.....

```
##        [,1]
## [1,] 0.7903
```

c. Si riporti la stima delle comunalità e delle varianze specifiche per le quattro variabili, arrotondando al quarto decimale:

	Murder	Assault	UrbanPop	Rape
Comunalità				
Varianza specifica				

```
##        [,1]  [,2]  [,3]  [,4]
## [1,] 0.669 0.958 0.069 0.466
## [2,] 0.331 0.042 0.931 0.534
```

d. Si valuti l'opportunità di stimare un modello a 2 fattori, motivando la risposta.