Metodo delle K-medie

Dati Forza Lavoro

1. Importare i dati ed ottenere la matrice X che contiene tutte le variabili ad esclusione di `Country` e `blocco`. Applicare l'algoritmo delle K-medie (argomento `algorithm = Lloyd`) sui dati X specificando $K = 3$ gruppi, inizializzando i centroidi con le osservazioni di riga 1, 25 e 26. Riportare le numerosità dei gruppi ottenuti e la tabella a doppia entrata che incrocia i gruppi ottenuti con la variabile `blocco`.

```
load("/Users/aldosolari/Dropbox/Oldbox/TSC/esercitazioni tsc/slides/paesi.Rdata")

X = paesi[,-c(1,11)]
n = nrow(X)
p = ncol(X)

# K-medie
km = kmeans(X, centers = X[c(1,25,26),], algorithm = "Lloyd")

# numerosità dei gruppi
table(km$cluster)
```

```
##
##  1  2  3
## 14  9  3
```

```
# tabella gruppi e blocco
table(km$cluster, paesi$blocco)
```

```
##
##      e  n  w
##   1  1  1 12
##   2  6  0  3
##   3  0  2  1
```

2. Determinare i centroidi, le somme dei quadrati *within* W e *between* B, il valore dell'indice CH di Calinski and Harabasz.

```
# centroidi
km$centers
```

```
##         Agr       Min      Man       Pow      Con       Ser      Fin
## 1  8.235714 0.9857143 29.08571 0.9571429 8.428571 16.278571 5.000000
## 2 25.022222 1.7777778 28.07778 0.9333333 8.722222  9.544444 2.133333
## 3 52.300000 0.9333333 14.10000 0.6000000 5.266667  7.700000 4.933333
##        SPS       TC
## 1 24.17143 6.864286
## 2 17.11111 6.688889
## 3  9.40000 4.633333
```

```
# somme dei quadrati W e B
( W = km$tot.withinss )
```

```
## [1] 2362.602
```

```
( B = km$betweenss )
```

```
## [1] 6936.988
```

```r
# indice CH
K = 3
(B/(K-1)) / (W/(n-K))
```

```
## [1] 33.76588
```

3. Costruire il grafico *silhouette* basato sulla distanza Euclidea con la funzione `silhouette` presente nella libreria `cluster`, e commentare i risultati.

```r
# Silhouette
library(cluster)

# matrice delle distanze euclidee
D = dist(X, method="euclidean")

# silhouette
sil <- silhouette(x=km$cluster, dist=D)
row.names(sil) <- paesi$Country
plot(sil)
```