k-means & CAH

systeme informatique décisionnelle et data mining

Réalisé par : EL RHARROUBI Mohamed amine

Encadré par : Pr. Abdelhadi FENNAN

1-Importation des données, description

Hide

```
#modifier le répertoire par défaut
setwd("C:/Users/Asus/Desktop/DM")
#charger les données - attention aux options
fromage <- read.table(file="fromage.txt",header=T,row.names=1,sep="\t",dec=".")
print(fromage)
```

	calories <int>	sodium <dbl>	calcium <dbl>	lipides <dbl>	retinol <dbl>	folates <dbl>	proteines <dbl>	cholesterol <int>	magnesium <int>
CarredelEst	314	353.5	72.6	26.3	51.6	30.3	21.0	70	20
Babybel	314	238.0	209.8	25.1	63.7	6.4	22.6	70	27
Beaufort	401	112.0	259.4	33.3	54.9	1.2	26.6	120	41
Bleu	342	336.0	211.1	28.9	37.1	27.5	20.2	90	27
Camembert	264	314.0	215.9	19.5	103.0	36.4	23.4	60	20
Cantal	367	256.0	264.0	28.8	48.8	5.7	23.0	90	30
Chabichou	344	192.0	87.2	27.9	90.1	36.3	19.5	80	36
Chaource	292	276.0	132.9	25.4	116.4	32.5	17.8	70	25
Cheddar	406	172.0	182.3	32.5	76.4	4.9	26.0	110	28
Comte	399	92.0	220.5	32.4	55.9	1.3	29.2	120	51

1-10 of 29 rows Previous **1** 2 3 Next

```
#afficher les premières lignes
print(head(fromage))
```

	calories <int>	sodium <dbl>	calcium <dbl>	lipides <dbl>	retinol <dbl>	folates <dbl>	proteines <dbl>	cholesterol <int>	magnesium <int>
CarredelEst	314	353.5	72.6	26.3	51.6	30.3	21.0	70	20
Babybel	314	238.0	209.8	25.1	63.7	6.4	22.6	70	27
Beaufort	401	112.0	259.4	33.3	54.9	1.2	26.6	120	41
Bleu	342	336.0	211.1	28.9	37.1	27.5	20.2	90	27
Camembert	264	314.0	215.9	19.5	103.0	36.4	23.4	60	20
Cantal	367	256.0	264.0	28.8	48.8	5.7	23.0	90	30

6 rows

Hide

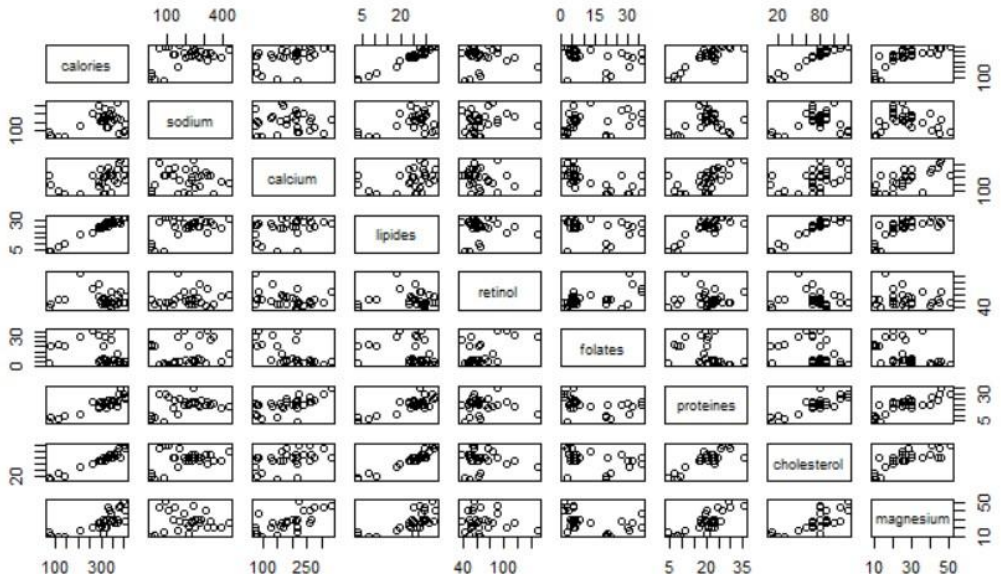
```
#stat. descriptives
print(summary(fromage))
```

```
    calories        sodium          calcium          lipides          retinol          folates          proteines     
 Min.   : 70   Min.   : 22.0   Min.   : 72.6   Min.   : 3.40   Min.   : 37.10   Min.   : 1.20   Min.   : 4.10  
 1st Qu.:292   1st Qu.:140.0   1st Qu.:132.9   1st Qu.:23.40   1st Qu.: 51.60   1st Qu.: 4.90   1st Qu.:17.80  
 Median :321   Median :223.0   Median :202.3   Median :26.30   Median : 62.30   Median : 6.40   Median :21.00  
 Mean   :300   Mean   :210.1   Mean   :185.7   Mean   :24.16   Mean   : 67.56   Mean   :13.01   Mean   :20.17  
 3rd Qu.:355   3rd Qu.:276.0   3rd Qu.:220.5   3rd Qu.:29.10   3rd Qu.: 76.40   3rd Qu.:21.10   3rd Qu.:23.40  
 Max.   :406   Max.   :432.0   Max.   :334.6   Max.   :33.30   Max.   :150.50   Max.   :36.40   Max.   :35.70  
  cholesterol       magnesium     
 Min.   : 10.00   Min.   :10.00  
 1st Qu.: 70.00   1st Qu.:20.00  
 Median : 80.00   Median :26.00  
 Mean   : 74.59   Mean   :26.97  
 3rd Qu.: 90.00   3rd Qu.:30.00  
 Max.   :120.00   Max.   :51.00  
```

2- description graphiques

Hide

```
#graphique - croisement deux à deux
pairs(fromage)
```



3- CAH (HCLUST)

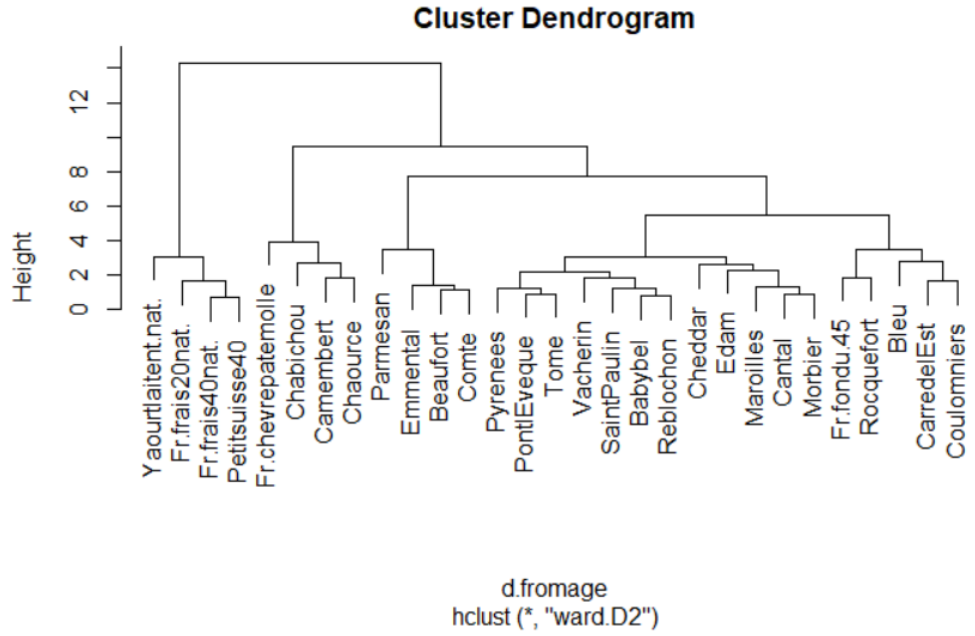
```
#centrage réduction des données
fromage.cr <- scale(fromage,center=T,scale=T)
#distance entre individus
d.fromage <- dist(fromage.cr)
```

Hide

```
#CAH - critère de Ward
cah.ward <- hclust(d.fromage,method="ward.D2")
```

Hide

```
#affichage dendrogramme
plot(cah.ward)
```




```
#découpage en 4 groupes
groupes.cah <- cutree(cah.ward,k=4)

#liste des groupes
print(sort(groupes.cah))
```

4- K-MEANS

```
#k-means avec les données centrées et réduites
#center = 4 - nombre de groupes demandés
#nstart = 5 - nombre d'essais avec différents individus de départ
groupes.kmeans <- kmeans(fromage.cr,centers=4,nstart=5)
#affichage des résultats
print(groupes.kmeans)
```

```
K-means clustering with 4 clusters of sizes 5, 4, 14, 6

Cluster means:
    calories     sodium    calcium     lipides    retinol    folates  proteines cholesterol  magnesium
1  0.8395372 -0.7332260  1.2856329  0.65210487 -0.1242419 -0.8436457  1.2861074   0.9705456  1.6287198
2 -2.1572744 -1.5213272 -0.7167418 -2.19980413 -0.5136787  0.2955348 -1.8634139  -1.9945017 -1.3884943
3  0.3726429  0.5276310  0.1925511  0.41101185 -0.3108901 -0.4505349  0.1522469   0.3181087  0.0156683
4 -0.1309315  0.3941009 -1.0428188 -0.03591228  1.1713977  1.5572630 -0.1847229  -0.2213739 -0.4681630

Clustering vector:
       CarredelEst            Babybel           Beaufort               Bleu          Camembert             Cantal
                 4                  3                  1                  3                  4                  3
         Chabichou           Chaource            Cheddar              Comte        Coulomniers               Edam
                 4                  4                  3                  1                  4                  1
          Emmental Fr.chevrepatemolle        Fr.fondu.45     Fr.frais20nat.     Fr.frais40nat.          Maroilles
                 1                  4                  3                  2                  2                  3
           Morbier           Parmesan      Petitsuisse40        PontlEveque           Pyrenees          Reblochon
                 3                  1                  2                  3                  3                  3
        Rocquefort        SaintPaulin               Tome           Vacherin Yaourtlaitent.nat.
                 3                  3                  3                  3                  2

Within cluster sum of squares by cluster:
[1]  9.871039  6.446342 28.737063 25.431001
 (between_SS / total_SS =  72.0 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"
[8] "iter"         "ifault"
```

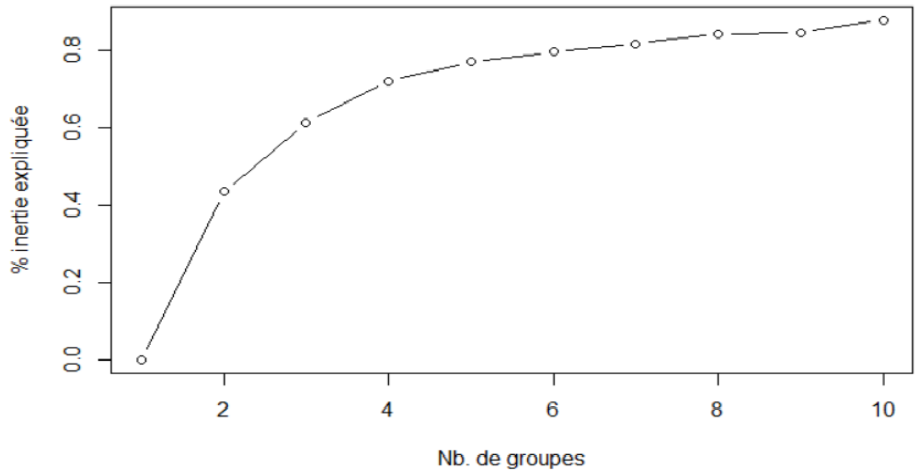
```
#correspondance avec les groupes de la CAH
print(table(groupes.cah,groupes.kmeans$cluster))
```

```
groupes.cah  1  2  3  4
          1  1  0 14  2
          2  4  0  0  0
          3  0  0  0  4
          4  0  4  0  0
```

5- Méthode des centres mobiles

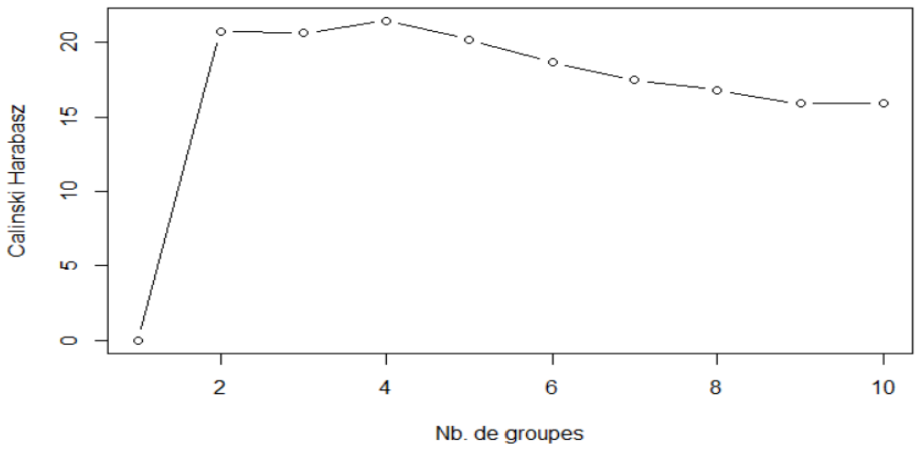
```r
#(1)évaluer la proportion d'inertie expliquée
inertie.expl <- rep(0,times=10)
for (k in 2:10){
  clus <- kmeans(fromage.cr,centers=k,nstart=5)
  inertie.expl[k] <- clus$betweenss/clus$totss
}

#graphique
plot(1:10,inertie.expl,type="b",xlab="Nb. de groupes",ylab="% inertie expliquée")
```




```r
#(2) indice de Calinski Harabasz
#utilisation du package fpc
library(fpc)

#évaluation des solutions
sol.kmeans <- kmeansruns(fromage.cr,krange=2:10,criterion="ch")
#graphique
plot(1:10,sol.kmeans$crit,type="b",xlab="Nb. de groupes",ylab="Calinski Harabasz")
```



IRIS FLOWER SPECIES

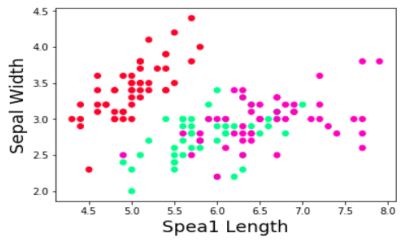
Entrée [1]:
```
from sklearn import datasets
import matplotlib.pyplot as plt
import pandas as pd
from sklearn.cluster import KMeans
```

Entrée [2]:
```
iris = datasets.load_iris()
```

Entrée [3]:
```
X = iris.data[:, :2]
y = iris.target
```

Entrée [4]:
```
plt.scatter(X[:,0], X[:,1], c=y, cmap='gist_rainbow')
plt.xlabel('Spea1 Length', fontsize=18)
plt.ylabel('Sepal Width', fontsize=18)
```

Out[4]:
```
Text(0, 0.5, 'Sepal Width')
```



Entrée [5]:
```
km = KMeans(n_clusters = 3, n_jobs = 4, random_state=21)
km.fit(X)
```

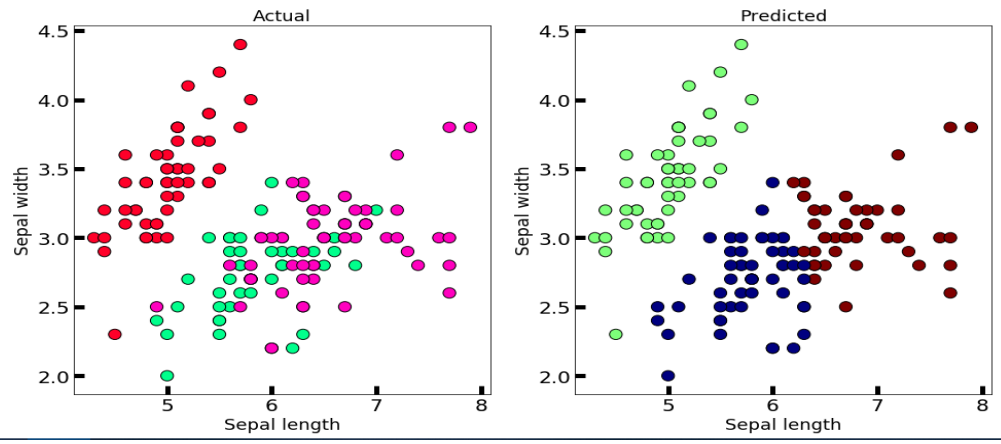
Out[5]:
```
KMeans(algorithm='auto', copy_x=True, init='k-means++', max_iter=300,
    n_clusters=3, n_init=10, n_jobs=4, precompute_distances='auto',
    random_state=21, tol=0.0001, verbose=0)
```

Entrée [7]:
```
centers = km.cluster_centers_
print(centers)
```

```
[[5.77358491 2.69245283]
 [5.006      3.428     ]
 [6.81276596 3.07446809]]
```

Entrée [8]:
```
#this will tell us to which cluster does the data observations belong.
new_labels = km.labels_
# Plot the identified clusters and compare with the answers
fig, axes = plt.subplots(1, 2, figsize=(16,8))
axes[0].scatter(X[:, 0], X[:, 1], c=y, cmap='gist_rainbow',
edgecolor='k', s=150)
axes[1].scatter(X[:, 0], X[:, 1], c=new_labels, cmap='jet',
edgecolor='k', s=150)
axes[0].set_xlabel('Sepal length', fontsize=18)
axes[0].set_ylabel('Sepal width', fontsize=18)
axes[1].set_xlabel('Sepal length', fontsize=18)
axes[1].set_ylabel('Sepal width', fontsize=18)
axes[0].tick_params(direction='in', length=10, width=5, colors='k', labelsize=20)
axes[1].tick_params(direction='in', length=10, width=5, colors='k', labelsize=20)
axes[0].set_title('Actual', fontsize=18)
axes[1].set_title('Predicted', fontsize=18)
```

Out[8]:
```
Text(0.5, 1.0, 'Predicted')
```



```
newiris <- iris
newiris$Species <- NULL
(kc <- kmeans(newiris, 3))
```

```
K-means clustering with 3 clusters of sizes 33, 96, 21

Cluster means:
  Sepal.Length Sepal.Width Petal.Length Petal.Width
1     5.175758    3.624242     1.472727   0.2727273
2     6.314583    2.895833     4.973958   1.7031250
3     4.738095    2.904762     1.790476   0.3523810

Clustering vector:
  [1] 1 3 3 3 1 1 1 1 3 3 1 1 3 3 1 1 1 1 1 1 1 1 1 3 3 1 1 1 3 3 1 1 1 3 1 1 1 3 1 1 3 3 1 1 3 1 3 1 1 2 2 2 2 2 2 2
 [58] 3 2 2 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 2 2 2 2 3 2 2 2 2 2 2 2 2 2 2 2 2
[115] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2

Within cluster sum of squares by cluster:
[1]   6.432121 118.651875  17.669524
 (between_SS / total_SS =  79.0 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"
[8] "iter"         "ifault"
```

```
table(iris$Species, kc$cluster)
```

```
            1  2  3
  setosa     33  0 17
  versicolor  0 46  4
  virginica   0 50  0
```

Hide

```
 plot(newiris[c("Sepal.Length", "Sepal.Width")], col=kc$cluster)
points(kc$centers[,c("Sepal.Length", "Sepal.Width")], col=1:3, pch=8, cex=2)
```