Segmentation

Wael Dhouib

October 1, 2018

Objectifs

Classification des données en utilisant les algorithmes de K-means et CAH.

Application de K-means sur les données :

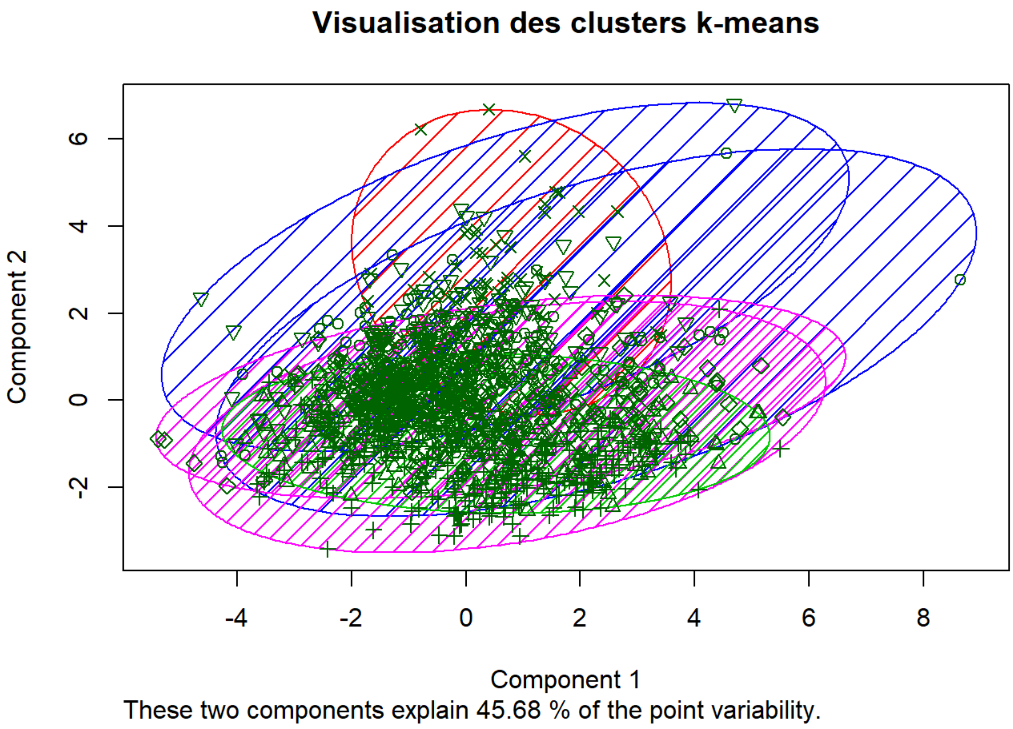
```
setwd("C:/")
redwine=read.table("winequality-red.csv", header=T, sep=";", dec=".", na.string="")
summary(redwine)
```

```
##  fixed.acidity   volatile.acidity  citric.acid    residual.sugar  
##  Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900  
##  1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900  
##  Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200  
##  Mean   : 8.32   Mean   :0.5278   Mean   :0.271   Mean   : 2.539  
##  3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600  
##  Max.   :15.90   Max.   :1.5800   Max.   :1.000   Max.   :15.500  
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide
##  Min.   :0.01200   Min.   : 1.00       Min.   :  6.00      
##  1st Qu.:0.07000   1st Qu.: 7.00       1st Qu.: 22.00      
##  Median :0.07900   Median :14.00       Median : 38.00      
##  Mean   :0.08747   Mean   :15.87       Mean   : 46.47      
##  3rd Qu.:0.09000   3rd Qu.:21.00       3rd Qu.: 62.00      
##  Max.   :0.61100   Max.   :72.00       Max.   :289.00      
##     density             pH          sulphates         alcohol     
##  Min.   :0.9901   Min.   :2.740   Min.   :0.3300   Min.   : 8.40  
##  1st Qu.:0.9956   1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50  
##  Median :0.9968   Median :3.310   Median :0.6200   Median :10.20  
##  Mean   :0.9967   Mean   :3.311   Mean   :0.6581   Mean   :10.42  
##  3rd Qu.:0.9978   3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10  
##  Max.   :1.0037   Max.   :4.010   Max.   :2.0000   Max.   :14.90  
##     quality     
##  Min.   :3.000  
##  1st Qu.:5.000  
##  Median :6.000  
##  Mean   :5.636  
##  3rd Qu.:6.000  
##  Max.   :8.000  
```

```
rw1=redwine[,1:11]
km1=kmeans(rw1,6)
table(redwine$quality,km1$cluster)
```

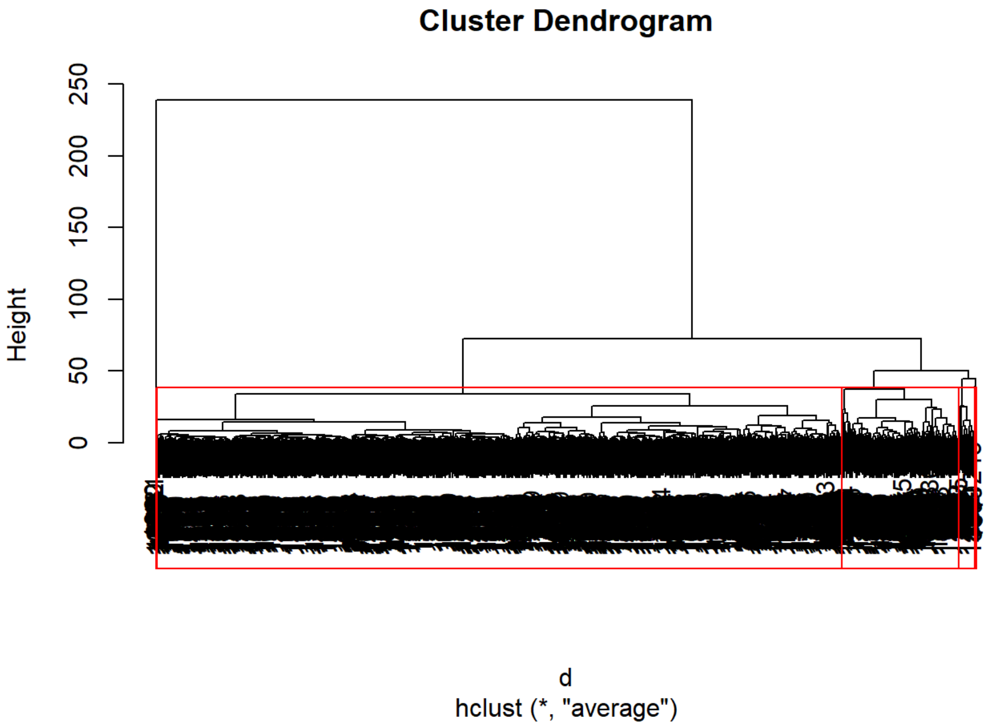
```
##    
##       1   2   3   4   5   6
##   3   0   1   6   0   3   0
##   4   7  12  21   1   8   4
##   5 114 138 133  60 120 116
##   6 105 179 154   6 148  46
##   7  15  52  77   2  42  11
##   8   1   3   9   0   3   2
```

```
library(cluster)
clusplot(rw1,km1$cluster,lines=0,color=T,shade=T,main=paste('Visualisation des clusters k-means'))
```



Application de la classification ascendante hiérarchique sur les données :

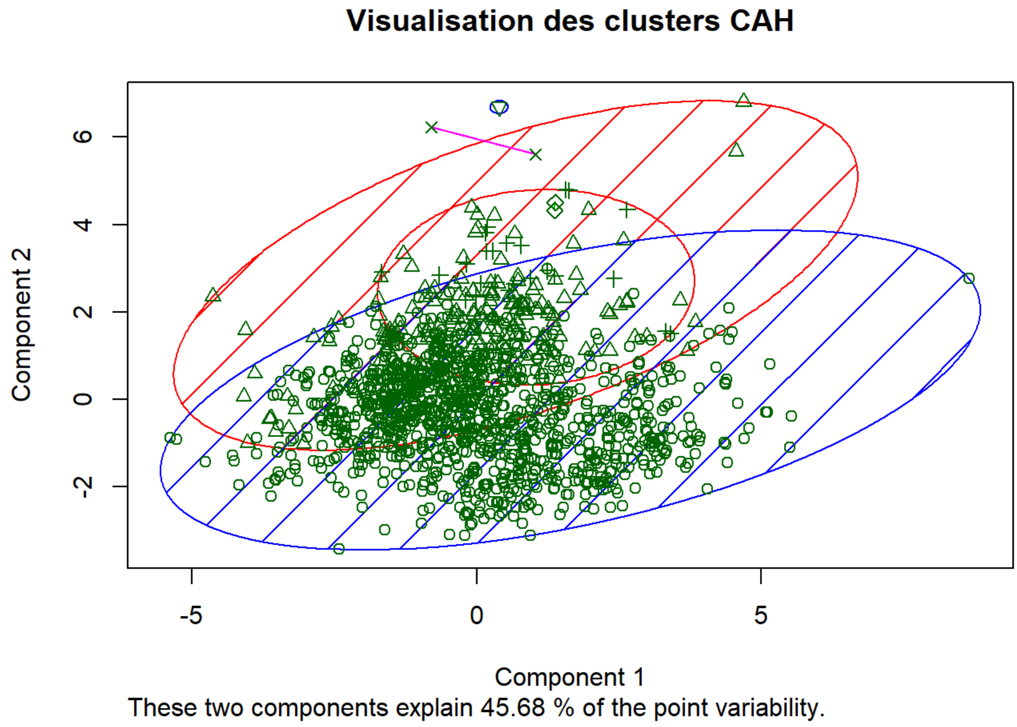
```
d=dist(rw1,"euclidean")
hc=hclust(d, method="average")
plot(hc)
rect.hclust(hc, k=6)
```



```
groupes=cutree(hc,6)
table(redwine$quality,groupes)
```

```
##     groupes
##        1   2   3   4   5   6
##   3  10   0   0   0   0   0
##   4  48   5   0   0   0   0
##   5 497 155  26   3   0   0
##   6 580  52   5   0   0   1
##   7 184  13   0   0   2   0
##   8  16   2   0   0   0   0
```

```
library(cluster)
clusplot(rw1, groupes, lines = 0, color= T, shade= T, main = paste('Visualisation des clusters CAH'))
```

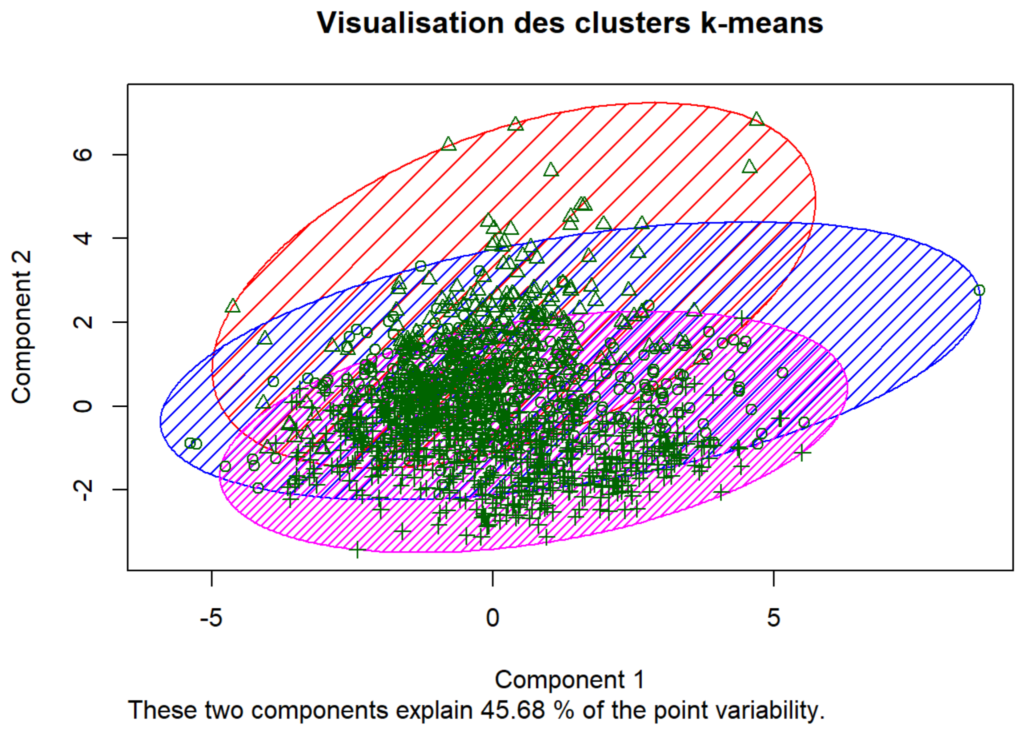


Application de K-means sur les données avec 3 groupes :

```
rw2=redwine
rw2[which(rw2$quality>6),"review"]='fine'
rw2[which(rw2$quality %in% c(5,6)),"review"]='average'
rw2[which(rw2$quality<5),"review"]='bad'
km2=kmeans(rw1,3)
table(rw2$review,km2$cluster)
```

```
##          
##              1   2   3
##   average 470 220 629
##   bad      18   5  40
##   fine     56  16 145
```

```
library(cluster)
clusplot(rw1,km2$cluster,lines=0,color=T,shade=T,main=paste('Visualisation des clusters k-means'))
```



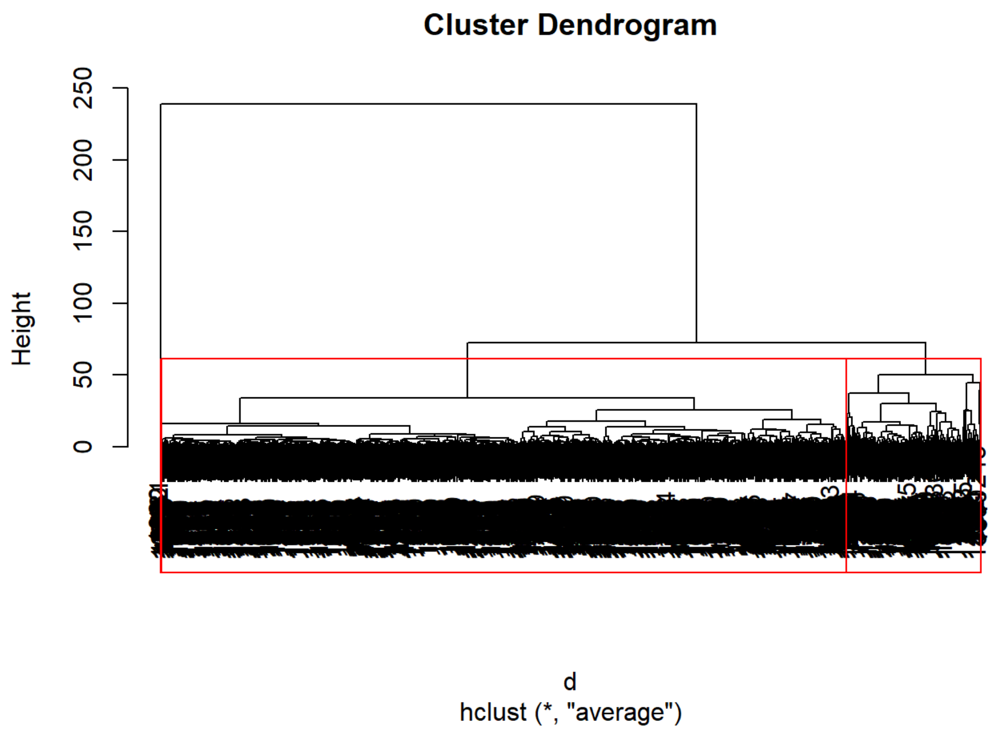
These two components explain 45.68 % of the point variability.

Application de la classification ascendante hiérarchique sur les données avec 3 groupes :

```
d=dist(rw1,"euclidean")
hc=hclust(d, method="average")
hc
```

```
##
## Call:
## hclust(d = d, method = "average")
##
## Cluster method   : average
## Distance         : euclidean
## Number of objects: 1599
```

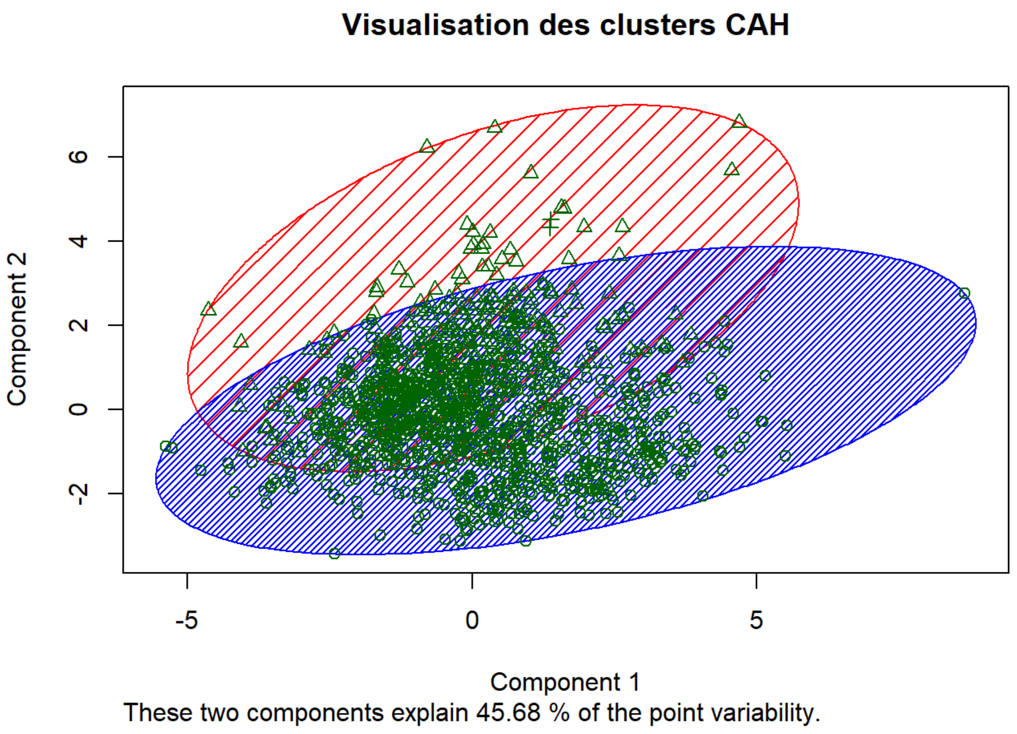
```
plot(hc)
rect.hclust(hc, k=3)
```



```
groupes=cutree(hc,3)
table(rw2$review,groupes)
```

```
##          groupes
##              1    2    3
##   average 1077  242    0
##   bad       58    5    0
##   fine     200   15    2
```

```
library(cluster)
clusplot(rw1, groupes, lines = 0, color= T, shade= T, main = paste('Visualisation des clusters CAH'))
```


Visualisation des clusters CAH
Component 2
Component 1
These two components explain 45.68 % of the point variability.