TP5 : Classification Non Supervisée (Unsupervised Machine Learning) : Méthode des centres mobiles (K-MEANS CLUSTERING)

DJEBALI Wissam

3 mars 2018

Méthode des centres mobiles (K-MEANS)

Packages R : stats, factoextra

L'algorithme de classification par méthode des centres mobiles (KMEANS) de MacQueen, variante de l'algorithme de Forgy/Lloyd est l'une des plus connues et des plus utilisées.

Principe de l'algorithme K-means clustering :

On a des données d'individus qu'on souhaite classer en K groupes, tel que les individus dans un même groupe soient les plus similaires (forte similarité intra-classe), et les individus de groupes différents soient les plus dissimilaires (faible similarité inter-classe). Chaque groupe sera représentés par son centre(centroïd) qui correspond à la moyenne des points assignés au groupe.

On veut définir les groupes tel que la variance intra-groupe (total within-cluster variation) soit minimale.

variation intra-groupe : $W(C_k) = \sum_{x_i \in C_k}(x_i - \mu_k)^2$ où x_i point représentant un individu appartenant au groupe C_k et μ_k moyenne assigné au groupe C_k

Chaque individu (x_i) est assigné à un groupe tel que la somme des variances intra-groupes(total within-cluster variation) soit minimale.

$tot.withinss = \sum_{k=1}^{K} W(C_k) = \sum_{k=1}^{K} \sum_{x_i \in C_k}(x_i - \mu_k)^2$

K-means algorithm can be summarized as follow :

1)Specify the number of clusters (K) to be created (by the analyst)

2)Select randomly k objects from the dataset as the initial cluster centers or means

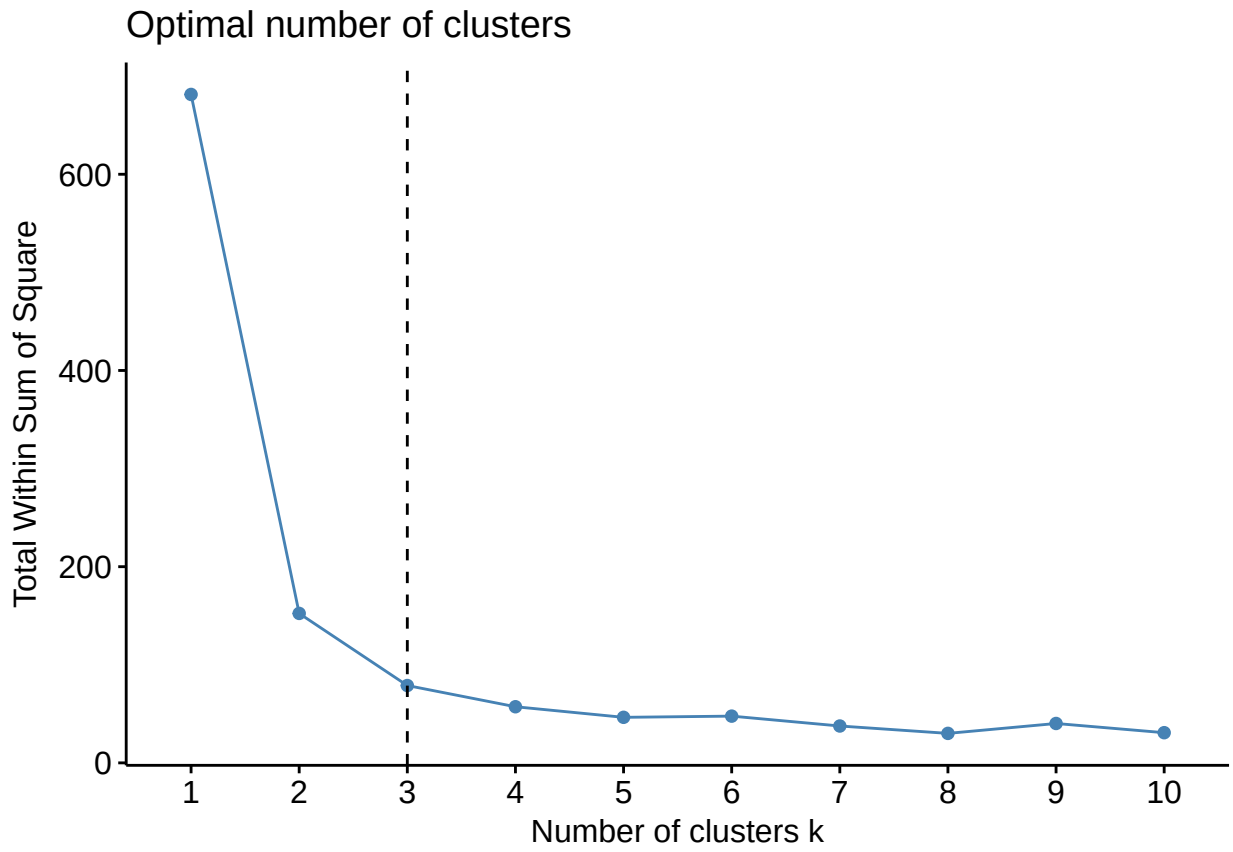
3)Assigns each observation to their closest centroid, based on the Euclidean distance between the object and the centroid

4)For each of the k clusters update the cluster centroid by calculating the new mean values of all the data points in the cluster. The centoid of a Kth cluster is a vector of length p containing the means of all variables for the observations in the kth cluster ; p is the number of variables.

5)Iteratively minimize the total within sum of square. That is, iterate steps 3 and 4 until the cluster assignments stop changing or the maximum number of iterations is reached. By default, the R software uses 10 as the default value for the maximum number of iterations.

```
ir<-iris[,-5]
species = iris$Species

# Choix de K pour le clustering
fviz_nbclust(ir, kmeans, method = "wss") +
    geom_vline(xintercept = 3, linetype = 2)
```




```r
# K-means avec K=3
set.seed(123)
km.res <- kmeans(ir, 3, nstart = 25)

# Print the results of kmeans
print(km.res)
```

```
## K-means clustering with 3 clusters of sizes 50, 38, 62
##
## Cluster means:
##   Sepal.Length Sepal.Width Petal.Length Petal.Width
## 1     5.006000    3.428000     1.462000    0.246000
## 2     6.850000    3.073684     5.742105    2.071053
## 3     5.901613    2.748387     4.393548    1.433871
##
## Clustering vector:
##   [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [36] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3 3 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
##  [71] 3 3 3 3 3 3 3 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 2 3 2 2 2
## [106] 2 3 2 2 2 2 2 2 3 3 2 2 2 2 3 2 3 2 3 2 2 3 3 2 2 2 2 2 3 2 2 2 2 3 2
## [141] 2 2 3 2 2 2 3 2 2 3
##
## Within cluster sum of squares by cluster:
## [1] 15.15100 23.87947 39.82097
##  (between_SS / total_SS =  88.4 %)
##
```

```
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
## [5] "tot.withinss" "betweenss"    "size"         "iter"
## [9] "ifault"
```

```
# Moyennes des variables dans les 3 groupes
aggregate(ir, by=list(cluster=km.res$cluster), mean)
```

```
##   cluster Sepal.Length Sepal.Width Petal.Length Petal.Width
## 1       1     5.006000    3.428000     1.462000    0.246000
## 2       2     6.850000    3.073684     5.742105    2.071053
## 3       3     5.901613    2.748387     4.393548    1.433871
```

```
# Tableau des individus avec leur moyenne par variable et leur groupe
dd <- cbind(ir, cluster = km.res$cluster)
head(dd)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width cluster
## 1          5.1         3.5          1.4         0.2       1
## 2          4.9         3.0          1.4         0.2       1
## 3          4.7         3.2          1.3         0.2       1
## 4          4.6         3.1          1.5         0.2       1
## 5          5.0         3.6          1.4         0.2       1
## 6          5.4         3.9          1.7         0.4       1
```

```
# Groupe de chaque observations
km.res$cluster
```

```
##   [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [36] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3 3 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
##  [71] 3 3 3 3 3 3 3 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 2 3 2 2 2 2
## [106] 2 3 2 2 2 2 2 2 3 3 2 2 2 2 3 2 3 2 3 2 2 3 3 2 2 2 2 2 3 2 2 2 2 3 2
## [141] 2 2 3 2 2 2 3 2 2 3
```

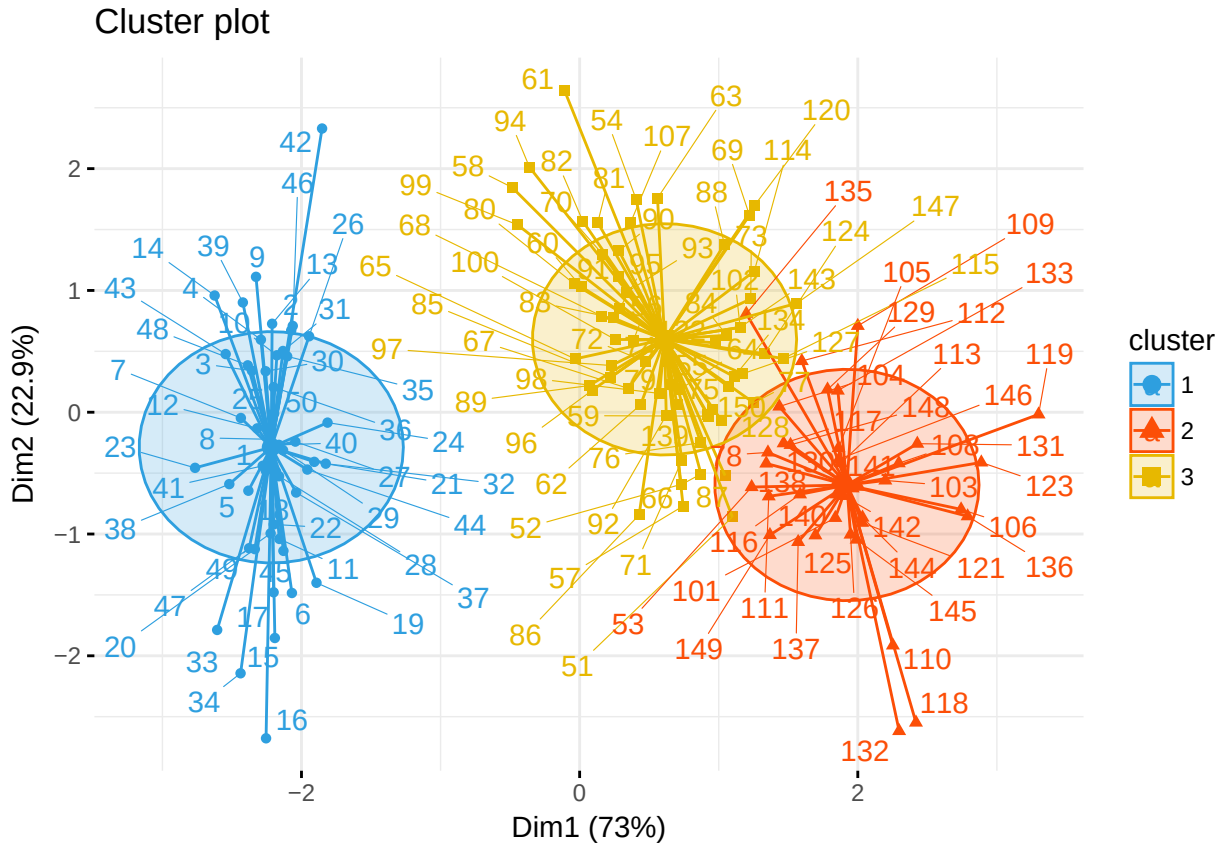
```
# Effectif des groupes
km.res$size
```

```
## [1] 50 38 62
```

```
# Centres des groupes
km.res$centers
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width
## 1     5.006000    3.428000     1.462000    0.246000
## 2     6.850000    3.073684     5.742105    2.071053
## 3     5.901613    2.748387     4.393548    1.433871
```

```
# Visualisation des individus en fonction de leur groupes
fviz_cluster(km.res, data = ir,
             palette = c("#2E9FDF", "#FC4E07", "#E7B800" ),
             ellipse.type = "euclid", # Concentration ellipse
             star.plot = TRUE, # Add segments from centroids to items
             repel = TRUE, # Avoid label overplotting (slow)
             ggtheme = theme_minimal()
             )
```




```
# Etude des pétales
# Petal
kmoy3<-kmeans(ir[,c(3,4)],3,nstart=4)
kmoy3
```

```
## K-means clustering with 3 clusters of sizes 50, 52, 48
##
## Cluster means:
##   Petal.Length Petal.Width
## 1     1.462000    0.246000
## 2     4.269231    1.342308
## 3     5.595833    2.037500
##
## Clustering vector:
##   [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [36] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
##  [71] 2 2 2 2 2 2 2 3 2 2 2 2 2 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3
## [106] 3 2 3 3 3 3 3 3 3 3 3 3 3 2 3 3 3 3 3 2 3 3 3 3 3 3 3 3 3 2 3
## [141] 3 3 3 3 3 3 3 3 3
##
## Within cluster sum of squares by cluster:
## [1]  2.02200 13.05769 16.29167
##  (between_SS / total_SS =  94.3 %)
##
## Available components:
##
```

```
## [1] "cluster"      "centers"      "totss"        "withinss"
## [5] "tot.withinss" "betweenss"    "size"         "iter"
## [9] "ifault"
```

```
table(kmoy3$cluster,species)
```

```
##    species
##     setosa versicolor virginica
##   1     50          0         0
##   2      0         48         4
##   3      0          2        46
```

```
par(mfrow=c(1,2))
plot(ir[c("Petal.Length", "Petal.Width")], col=kmoy3$cluster)
points(kmoy3$centers[,c("Petal.Length", "Petal.Width")],
       col=1:3, pch=23, cex=3)
plot(ir[c("Petal.Length", "Petal.Width")],col=iris$Species)
```