Lab 12 Cluster Analysis

Lusine Zilfimian

May 06 (Wednesday), 2020

Contents

- Libraries
- K-Means
- Hierarchical clustering

Needed packages

```r
library(dplyr)
library(ggplot2)
```

K-Means

- Perfectly separated two clusters

```
set.seed(2708)
x <- matrix(rnorm(100), ncol=2)
x[1:25,1] = x[1:25,1] + 3
x[1:25,2] = x[1:25,2] - 4
km.out <- kmeans(x = x, centers = 2)
```

- Cluster Membership for each record

```
km.out$cluster
```

```
##  [1] 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 1 1 1 1 1 1 1 1 1
## [36] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

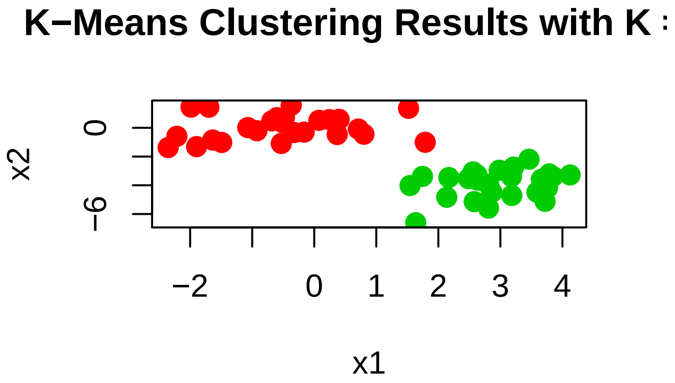
- Cluster means/centers of each variable for each cluster

```
km.out$centers
```

```
##         [,1]        [,2]
## 1 -0.5210669  0.03040455
## 2  2.9246016 -3.92630843
```

Visualization

```
plot(x, col = (km.out$cluster + 1),
  main = "K-Means Clustering Results with K = 2",
  xlab = "x1", ylab = "x2", pch = 20, cex = 2)
```



Changing arguments

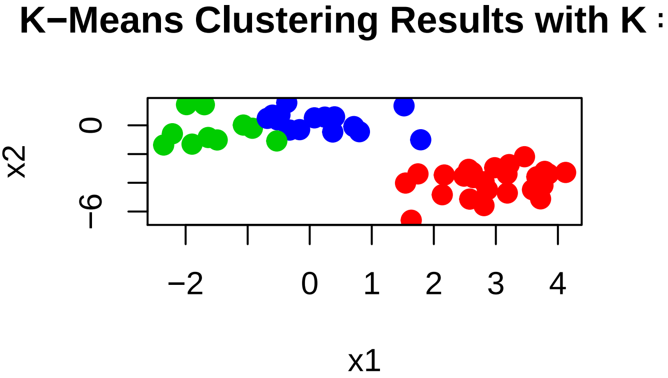
- kmeans() function has nstart option that attempts multiple initial configurations and reports on the best one.

```
km.out <- kmeans(x, 3, nstart = 20)
km.out
```

```
## K-means clustering with 3 clusters of sizes 25, 10, 15
##
## Cluster means:
##         [,1]       [,2]
## 1  2.9246016 -3.9263084
## 2 -1.5801166 -0.3524770
## 3  0.1849663  0.2856589
##
## 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 3 2 2 2 3 3 3 3 2
## [36] 2 3 3 2 3 2 3 3 3 3 3 2 3 2 2
##
## Within cluster sum of squares by cluster:
## [1] 36.57173 12.89579 15.19079
##  (between_SS / total_SS =  85.0 %)
##
```

Visualization

```r
plot(x, col = (km.out$cluster + 1),
  main = "K-Means Clustering Results with K = 3",
  xlab = "x1", ylab = "x2", pch = 20, cex = 2)
```



Variances (TSS, BSS, WSS)

The idea behind K-means clustering is that a good clustering is one for which the **within-cluster** variation is as **small** as possible. With this steps K-means algorithm maximizes **between** group sum of squares and minimizes **within group** sum of squares.

```
km.out$totss
```

```
## [1] 429.8973
```

```
km.out$withinss
```

```
## [1] 36.57173 12.89579 15.19079
```

```
km.out$betweenss
```

```
## [1] 365.239
```

```
sum(km.out$withinss) + km.out$betweenss
```

```
## [1] 429.8973
```

```
km.out$betweenss/km.out$totss
```

```
## [1] 0.8495959
```

- The number of points in each cluster

```
km.out$size
```

```
## [1] 25 10 15
```

- The number of iterations

```
km.out$iter
```

```
## [1] 2
```

Comparison

```
set.seed(270895)
km.out <- kmeans(x, 3, nstart = 1)
km.out$tot.withinss
```

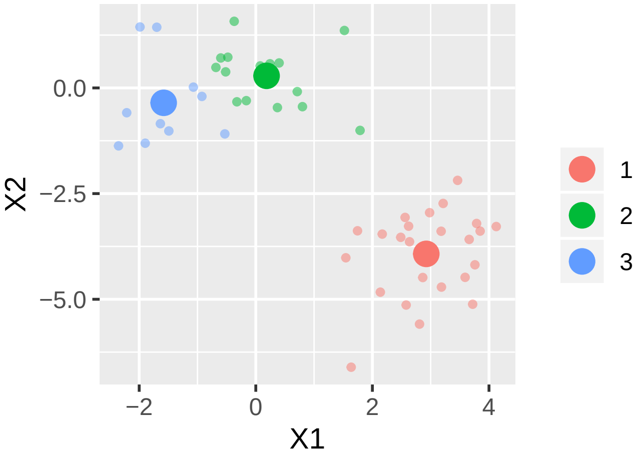
```
## [1] 68.65732
```

```
km.out <- kmeans(x, 3, nstart = 20)
km.out$tot.withinss
```

```
## [1] 64.65832
```

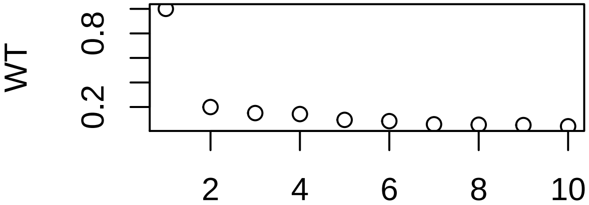
Visualization

```
ggplot(data.frame(x), aes(x = X1, y = X2, col = factor(km.out$cluster))) +
  geom_point(size = 1, alpha = 0.5) +
  geom_point(data.frame(km.out$centers, cl = factor(1:3)),
    mapping = aes(X1, X2, col = cl), size = 4) +
  labs(col = "")
```



Make simple loop for WithinSS/TotalSS

```r
WT <- c()
for(i in 1:10){
  set.seed(2708)
  km <- kmeans(x, i)
  WT[i] <- km$tot.withinss/km$totss
  }
plot(WT)
```



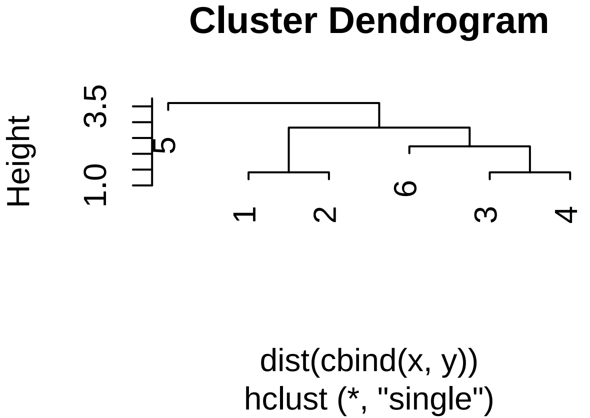
Hierarchical Clustering

```
x <- c(1,2,3,4,5,6)
y <- c(4,5,8,7,3,8)
dist(cbind(x,y))
```

```
##          1        2        3        4        5
## 2 1.414214
## 3 4.472136 3.162278
## 4 4.242641 2.828427 1.414214
## 5 4.123106 3.605551 5.385165 4.123106
## 6 6.403124 5.000000 3.000000 2.236068 5.099020
```

Dendogram

```
plot(hclust(dist(cbind(x,y)), method = "single"))
```



Example with Real Data

```r
index <- read.csv("index2017.csv")
colnames(index)
```

```
##  [1] "CountryID"                       "Country.Name"
##  [3] "Abbr"                            "Region"
##  [5] "World.Rank"                      "Region.Rank"
##  [7] "X2017.Score"                     "Property.Rights"
##  [9] "Judical.Effectiveness"           "Government.Integrity"
## [11] "Tax.Burden"                      "Gov.t.Spending"
## [13] "Fiscal.Health"                   "Business.Freedom"
## [15] "Labor.Freedom"                   "Monetary.Freedom"
## [17] "Trade.Freedom"                   "Investment.Freedom"
## [19] "Financial.Freedom"               "Tariff.Rate"
## [21] "Income.Tax.Rate"                 "Corporate.Tax.Rate"
## [23] "Tax.Burden.perc.of.GDP"          "Gov.t.Expenditure.perc.of.GDP"
## [25] "Population_Millions"             "GDP.Billions.PPP"
## [27] "GDP.Growth.Rate"                 "GDP.per.Capita.PPP"
## [29] "Unemployment"                    "Inflation.Perc"
## [31] "FDI.Inflow.Millions"             "Public.Debt.Perc.of.GDP"
```

```r
rownames(index) <- index$Abbr
```

Choosing the subset of observations and features to show the dendogram

```
(index1 <- index[1:7, c("Unemployment", "GDP.per.Capita.PPP")])
```

```
##     Unemployment GDP.per.Capita.PPP
## AFG          9.6               1947
## ALB         17.3              11301
## DZA         10.5              14504
## AGO          7.6               7344
## ARG          6.7              22554
## ARM         16.3               8468
## AUS          6.3              47389
```

```
index1 <- na.omit(index1)
```

HC

- Calculating distances

```
(d <- dist(index1, method = "euclidian"))
```

```
##           AFG       ALB       DZA       AGO       ARG       ARM
## ALB  9354.003
## DZA 12557.000  3203.007
## AGO  5397.000  3957.012  7160.001
## ARG 20607.000 11253.005  8050.001 15210.000
## ARM  6521.003  2833.000  6036.003  1124.034 14086.003
## AUS 45442.000 36088.002 32885.000 40045.000 24835.000 38921.001
```

```
(cl <- hclust(d, method = "complete"))
```

```
##
## Call:
## hclust(d = d, method = "complete")
##
## Cluster method   : complete
## Distance         : euclidean
## Number of objects: 7
```

Merging

- At first stage 4 and 6 are merged (Armenia and Angola)

```
cl$merge
```

```
##      [,1] [,2]
## [1,]   -4   -6
## [2,]   -2   -3
## [3,]   -1    1
## [4,]   -5    2
## [5,]    3    4
## [6,]   -7    5
```

```
min(d)
```

```
## [1] 1124.034
```

Merging

```
d
```

```
##           AFG       ALB       DZA       AGO       ARG       ARM
## ALB  9354.003
## DZA 12557.000  3203.007
## AGO  5397.000  3957.012  7160.001
## ARG 20607.000 11253.005  8050.001 15210.000
## ARM  6521.003  2833.000  6036.003  1124.034 14086.003
## AUS 45442.000 36088.002 32885.000 40045.000 24835.000 38921.001
```

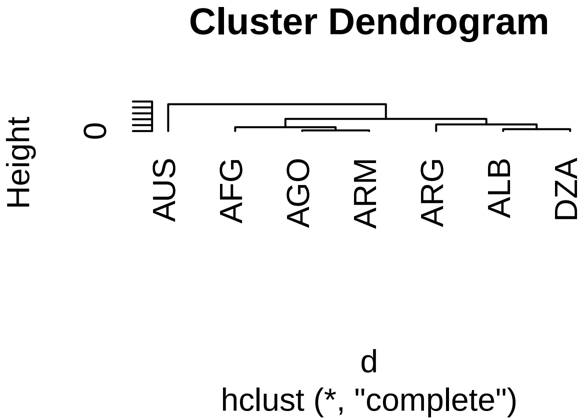
- At the second step ALB and DZA (Albania and Algeria) are merged
- At the third step AFG is merged with the cluster from the first step (ARM and Angola)
- The height is going to show the distance between clusters that are merged

```
cl$height
```

```
## [1]  1124.034  3203.007  6521.003 11253.005 20607.000 45442.000
```

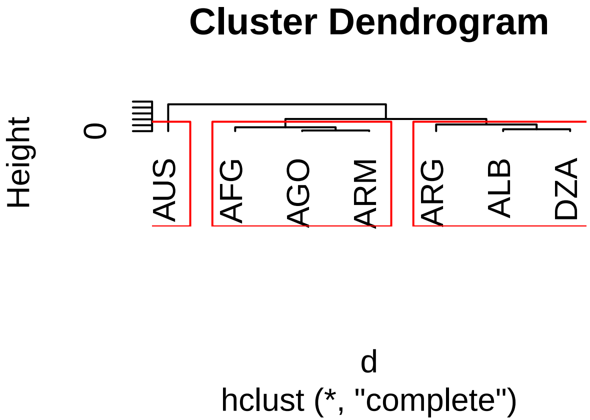
Dendrogram

```
plot(cl, hang = -1)
```



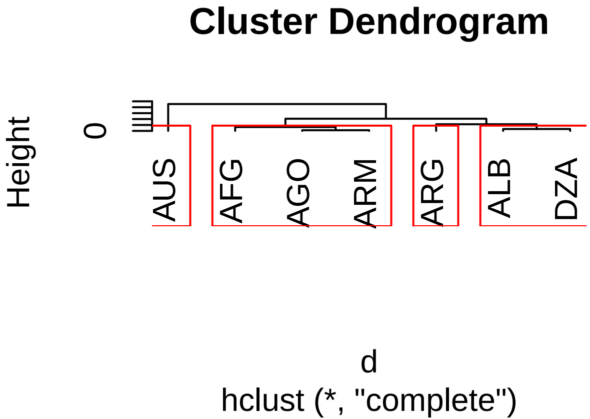
How many clusters to chose?

```
plot(cl, hang = -1)
rect.hclust(cl, 3)
```



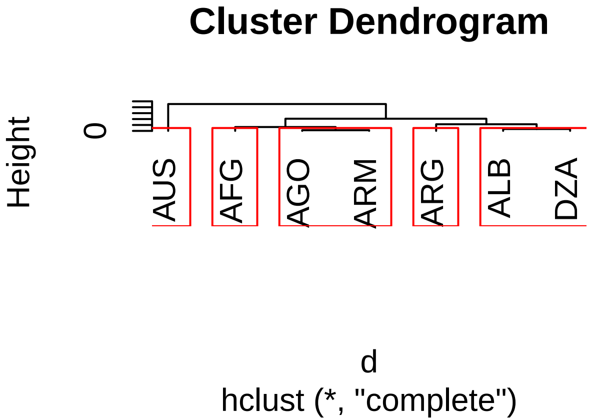
4 clusters

```
plot(cl, hang = -1)
rect.hclust(cl, 4)
```



5 clusters

```
plot(cl, hang = -1)
rect.hclust(cl, 5)
```



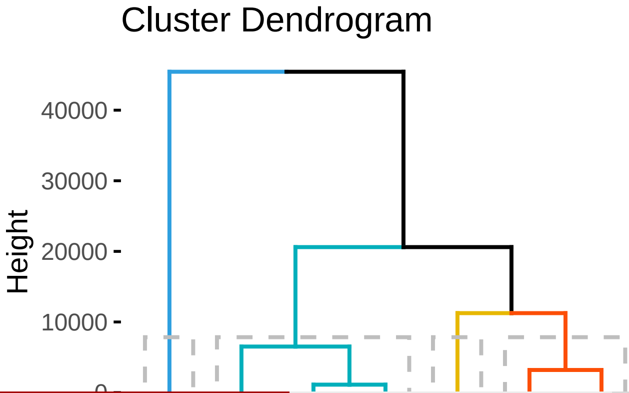
Cluster membership

```
index1$cl.memb <- cutree(cl, k=3); index1
```

```
##     Unemployment GDP.per.Capita.PPP cl.memb
## AFG          9.6               1947       1
## ALB         17.3              11301       2
## DZA         10.5              14504       2
## AGO          7.6               7344       1
## ARG          6.7              22554       2
## ARM         16.3               8468       1
## AUS          6.3              47389       3
```

Visualization

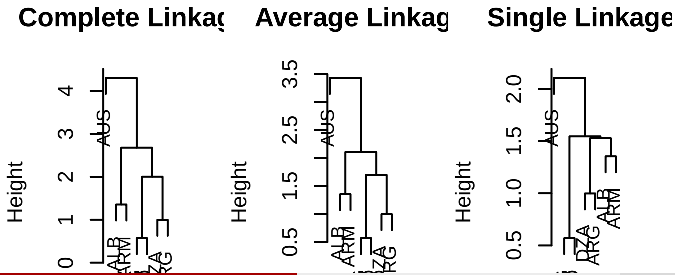
```r
factoextra::fviz_dend(cl, k = 4, # Cut in four groups
  cex = 0.5, # label size
  k_colors = c("#2E9FDF", "#00AFBB", "#E7B800", "#FC4E07"),
  color_labels_by_k = TRUE, # color labels by groups
  rect = TRUE # Add rectangle around groups
  )
```



Comparison

```
index1 <- as.data.frame(scale(index1, center = T, scale = T))
hc.complete <- hclust(dist(index1), method="complete")
hc.average <- hclust(dist(index1), method ="average")
hc.single <- hclust(dist(index1), method ="single")

par(mfrow=c(1,3))
plot(hc.complete ,main = "Complete Linkage ", xlab = "", sub = "", cex=.9)
plot(hc.average , main = "Average Linkage", xlab = "", sub = "", cex=.9)
plot(hc.single, main = "Single Linkage ", xlab = "", sub = "", cex=.9)
```



Comparison

- Determine the cluster labels

```r
cutree(hc.complete, 3)
```

```
## AFG ALB DZA AGO ARG ARM AUS
##   1   2   1   1   1   2   3
```

```r
cutree(hc.average, 3)
```

```
## AFG ALB DZA AGO ARG ARM AUS
##   1   2   1   1   1   2   3
```

```r
cutree(hc.single, 3)
```

```
## AFG ALB DZA AGO ARG ARM AUS
##   1   2   2   1   2   2   3
```