MS – E 2112 – Multivariate statistical analysis – Home exercise 10

Srikanth Gadicherla , 546195

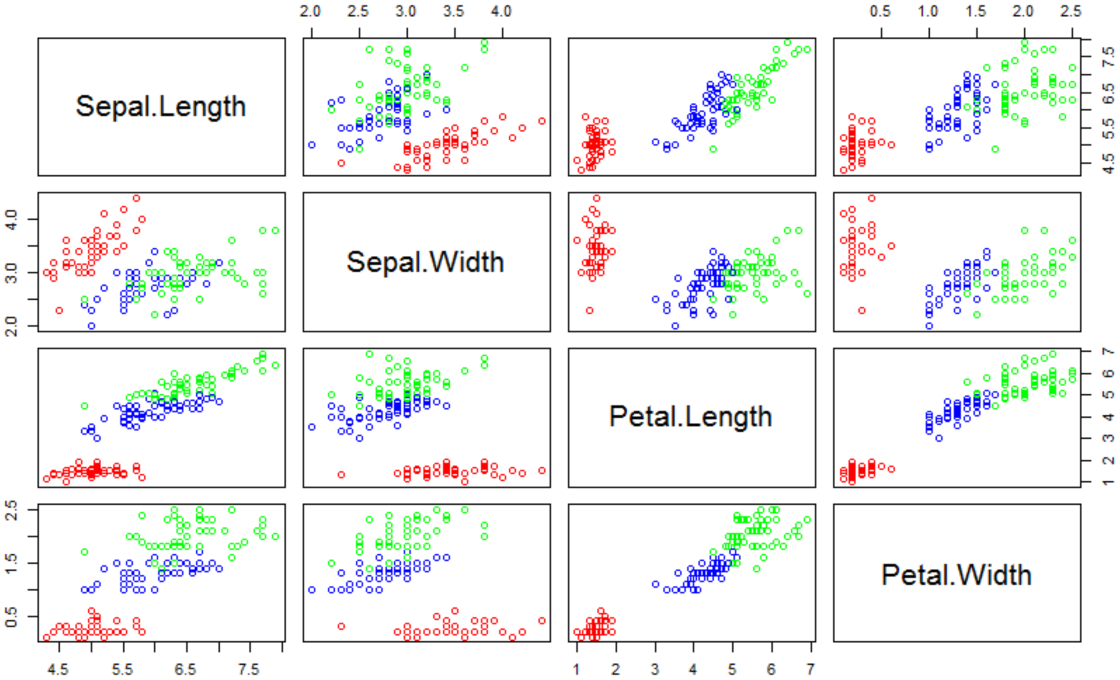
March 25 *th* ,2016

Problem 1

Data: The data set is the popular iris data which consists of information of flowers Petal length, Petal Width, Sepal length and Sepal Width. There are classes (species) of data namely setosa, versicolor and virginica with 50 belonging to each group giving a total of 150 samples.

a) **Scatter Plot**

We were asked to scatter plot the variables. The plot is given below. The red dots belong to Setosa, blue belongs to Versicolor, and green belongs to Virginica.



In most of the univariate scatter plots above, we can make three distinct groups based on the data.

b) **Calculate the Euclidean distance between the species of flower?**

It was accomplished using R dist method with method "euclidean".

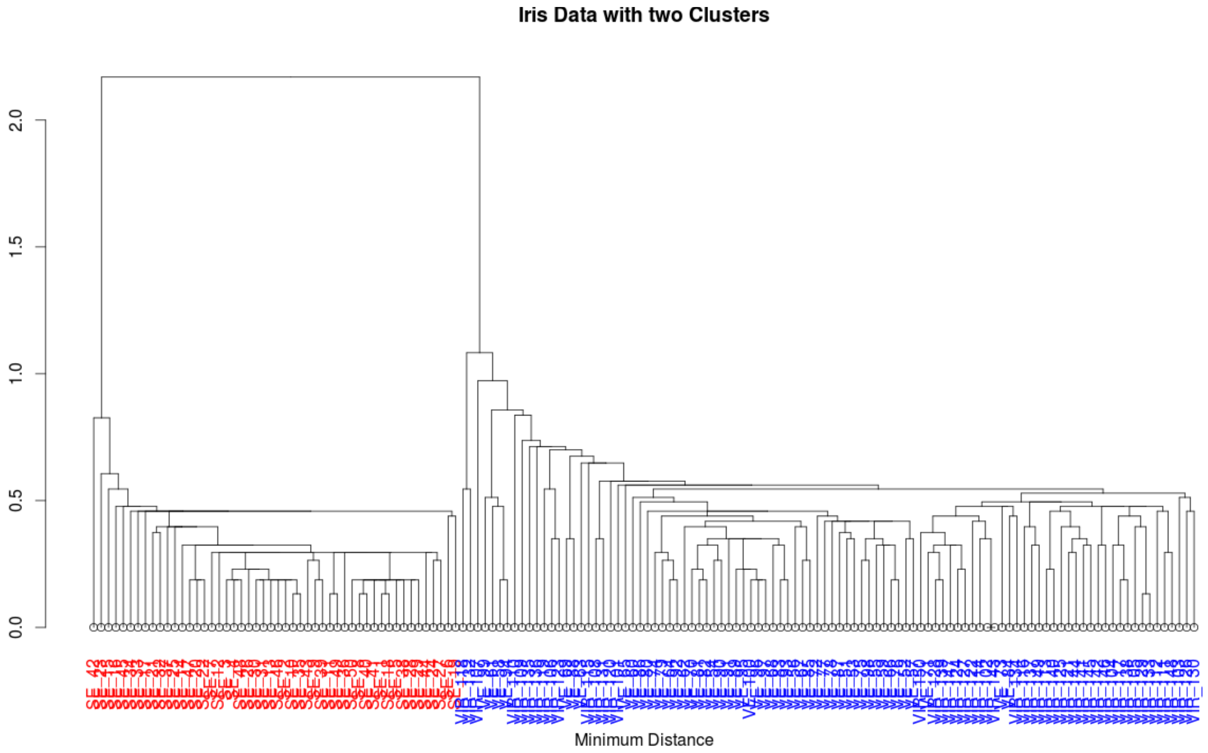
iris.dist <- dist(iris,method="euclidean")

The distance values were then rounded to two digits and then sorted to see the minimum distances.

d,e,f) **Hierarchical Clustering using hclust():**

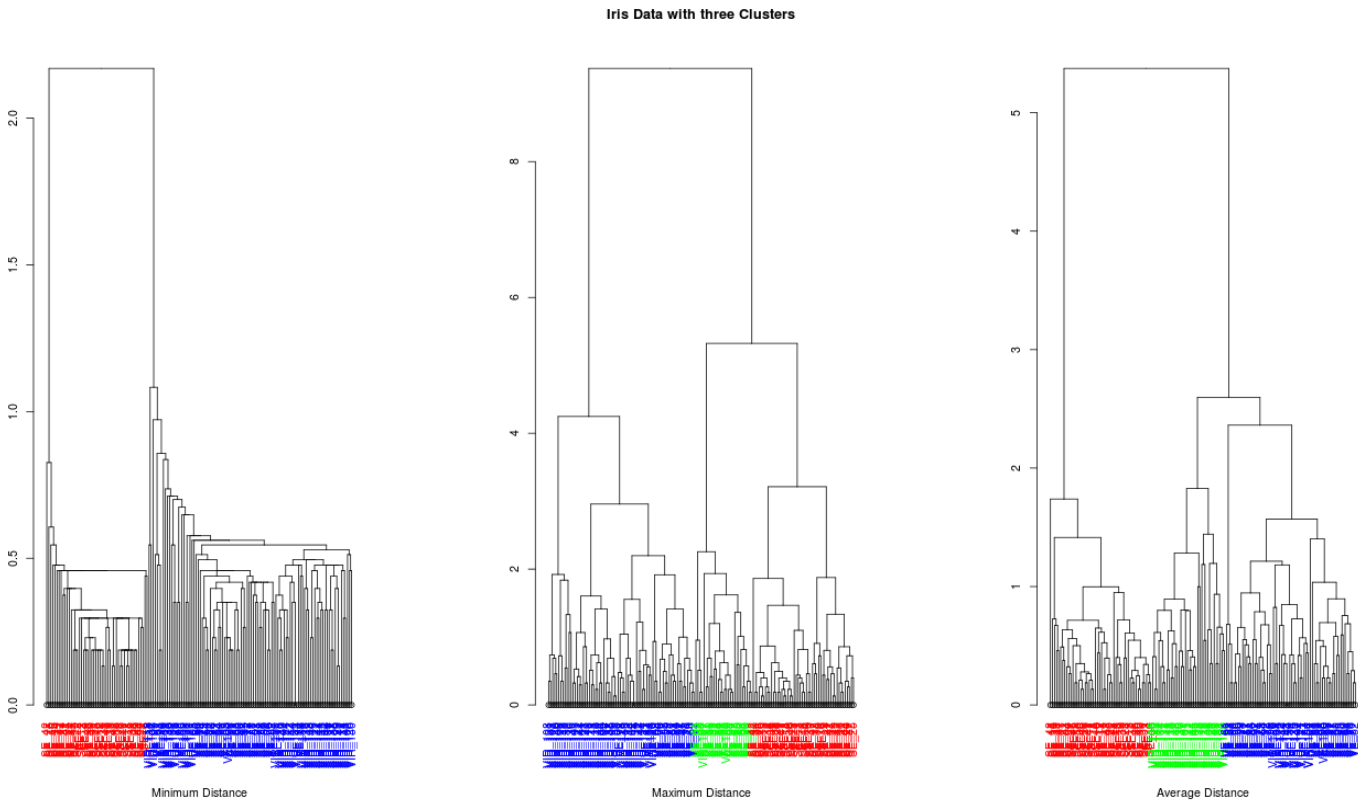
The hclust function was used from stats package. The cut trees with two and three clusters were plotted.

Two Cluster with 'minimum' distance method.



It can be seen that the minimum distance method is able to cluster the two groups well. The species setosa (text with red color) is well separated from species versicolor and species virginica (text with blue color).

Three Cluster with 'minimum', 'maximum' and 'average' distance method.



Due to label clutter, the labels are not clearly visible. But the virginica with labels VIR are hanging lower than versicolor VE, so in a way we can differentiate them.

The setosa species was separated very well in all the three methods. It is seen the red text and labels starting from SE in all the three dendrogram. This can also be substantiated from the fact that the setosa species in the scatter plot was linearly separable in most of the sub-plots.

The species versicolor and virginica were misclassified. They were combined in the minimum distance method. But in maximum and average distance, the misclassification reduced.

In the maximum distance method, few virginica (green text hanging below) was classified as versicolor and substantial part of versicolor species was (blue text; not low hanging) classified as virginica.

In the average distance method, the versicolor was separated well but it also classified few virginica (low hanging blue text) as versicolor.

Conclusion: Hierarchical clustering was not able completely classify the three species so we need more sophisticated methods for this problem.

g) Where to cut the tree?

For the two clusters problem, the dendrogram has two well defined trees, so, can be cut easily but the problem complicates for three trees cut. In the maximum distance method's dendrogram, the second branch from the top has two fairly good defined branches, so it cut there, even the same goes for the average method.

This can be substantiated using the cutree function in R.

Appendix

The code for the problem solved above.

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
#setwd("//home.org.aalto.fi/gadichs1/data/Desktop/P3_4/MSA/Ex 10")
setwd("~/Documents/OneDrive/Aalto/Sem2/MSA/Ex 10")

library(ape)

# function to get color labels
colLab <- function(n) {
  if (is.leaf(n)) {
    a <- attributes(n)
    labCol <- labelColors[clusMember[which(names(clusMember) == a$label)]]
    attr(n, "nodePar") <- c(a$nodePar, lab.col = labCol)
  }
  n
}

data(iris)
View(iris)

iris["Rows"] <- 1:150
iris["RowNames"] <- NA

row <- lab <- c("SE","VE","VIR")
iris$RowNames <- row[iris$Species]

labelColors = c('red', 'blue','green')

iris <- within(iris, Rows <- paste(iris[,7], iris[,6], sep='_'))

rownames(iris) <- iris[,6]

plot(iris[,1:4])
legend("center", legend=c("Setosa","Versicolor","Virginica"),title="IrisGroup")
## install packages fpc, MASS, cluster
library(fpc)
library(MASS)

library(cluster)


lab <- c("Setosa","Versicolor","Virginica")


#plot(iris,panel = function(x,y) text(x,y,labels = lab,xpd=T))

par(mar=c(5.1, 4.1, 4.1, 8.1), xpd=TRUE)

plot(iris[,1:4], col=c("red","blue","green")[iris$Species]);

#legend('topright', lab, col=c('red', 'blue', 'green'))

legend("center", inset=c(-20,-20), legend=c("Setosa","Versicolor","Virginica"),
pch=c(1,3), title="IrisGroup")


iris.dist <- dist(iris,method="euclidean")




round(iris.dist,2)


min(iris.dist)


sort((round(iris.dist,1)))


iris.min <- hclust(iris.dist,method = "single")

clusMember = cutree(iris.min, 2)

iris.min = dendrapply(as.dendrogram(iris.min), colLab)


par(mfrow=c(1,1))

plot(iris.min,sub='Minimum Distance',main="Iris Data with two Clusters")


iris.max <- hclust(iris.dist,method = "complete")

iris.ave <- hclust(iris.dist,method = "average")






clusMember = cutree(iris.min, 3)

iris.min = dendrapply(as.dendrogram(iris.min), colLab)
```

```
clusMember = cutree(iris.max, 3)

iris.max = dendrapply(as.dendrogram(iris.max), colLab)


clusMember = cutree(iris.ave, 3)

iris.ave = dendrapply(as.dendrogram(iris.ave), colLab)


par(mfrow=c(1,3))

plot(iris.min,sub='Minimum Distance')

plot(iris.max,sub='Maximum Distance',main="Iris Data with three Clusters")

plot(iris.ave,sub='Average Distance')
```

%%