G0O02a: Statistical Data Analysis: Project 2

Prof. Mia Hubert

April 2022

The project consists of analyzing the Melon data set. The data set contains the result of a spectroscopy experiment conducted on $n = 2158$ cantaloupe melons of four different cultivars. Each of the spectra was measured on 256 wavelengths. The last variable `y` indicates the groups number of the cultivar, whereas the variable `cultivar_levels` contains their name.

You answer the questions by performing an appropriate analysis with `R`. The discussion of the results and the necessary figures are reported in a written text that consists of a maximum of 5 pages (12pt font size). This file should be named "LastName_FirstName_Project2.pdf". Only report results and interpretations, do not repeat theory from the course! Include the figures when they are discussed, not at the end of the report. Additionally a separate file with the full `R` script should be provided. It should be named "LastName_FirstName_Project2.R".

One single folder containing your report and `R` script should be uploaded on Toledo no later than **May 15, 23h**. This project is graded on 3.5 points.

You first draw an individual data set of random 50 spectra from **each** random class. You use the following code, where you change 0012345 by your student number. When you make figures, always use `colors.gr` to plot observations (spectra) from the known groups, and `colors.cl` to color the clusters.

```
library(dplyr)
load("Melon.rdata")
set.seed(0012345)
alldata <- data.frame(cbind(X,y))
mydata <- alldata %>% group_by(y) %>% sample_n(50)
X <- mydata[, -257]
group <- mydata$y
# colors for the 4 known groups
colors.gr <- c("blue","red","darkgreen","orange")
# colors for the clusters
colors.cl <- c("cadetblue","firebrick","darkolivegreen3","gold","brown","purple")
```

1. For each group, make a plot of the spectra. Also make a heatmap of all spectra. State your main findings.

2. Partition the spectra (X) by means of k-medoids, imposing 2 up to 6 clusters.

   (a) Discuss the quality of the resulting clusterings without using the `group` information.

   (b) Compare each clustering with the known groups, by means of (1) a contingency table, (2) a mosaic plot, (3) plots of the spectra colored according to the clustering, (4) a cluster plot where the observations are colored according to their group number and (5) a heatmap of the data sorted according to the pam clustering. Can you relate the obtained clusters with the groups in the raw data?

   To draw the cluster plot, you can use the following code. Of course the first line depends on the cluster method you apply.

```
X.pam.3 <- pam(X, 3)
f_clust_pam <- fviz_cluster(X.pam.3, geom="point", ellipse.type="convex",
                            palette=colors.cl, ellipse.border.remove = T)
f_clust_pam[["layers"]][[1]][["data"]][["cluster"]] <- as.factor(group)
f_clust_pam[["layers"]][[2]][["data"]][["cluster"]] <- as.factor(X.pam.3$cluster)
f_clust_pam +
  scale_colour_manual(values=colors.gr)
```

   (c) Explain which number of clusters k_{opt} you find most appropriate for your data.

   In your report, only include the most relevant figures. The `R` file should contain the full code.

3. Perform `agnes` with average and complete linkage. Consider the resulting hard clusterings with k_{opt} clusters. Compare the results with the k-medoids clustering with k_{opt} clusters.

4. Conclude which cluster method yields the best results.