G0O02a: Statistical Data Analysis: Project Academic year 2020 – 2021

Mia Hubert, Cécile Adam

The project consists of analyzing the Dry Bean data set. A computer vision system was developed to distinguish different registered varieties of dry beans with similar features (12 dimensions and 4 shape forms) in order to obtain uniform seed classification. This data set contains the features such as form, shape, type, and structure by the market situation of 3 varieties of dry beans. The 16 variables of the data set contain these measurement data (`Area`, `Perimeter`, ...). The last variable `Class` indicates what type of Dry Bean each observation is. The following table gives you the description of the variables:

Area (A)	The area of a bean zone and the number of pixels within its boundaries.
Perimeter (P)	Bean circumference is defined as the length of its border.
Major axis length (L)	The distance between the ends of the longest line that can be drawn from a bean.
Minor axis length (l)	The longest line that can be drawn from the bean while standing perpendicular to the main axis.
Aspect ratio (K)	L/l
Eccentricity (Ec)	Eccentricity of the ellipse having the same moments as the region.
Convex area (C)	Number of pixels in the smallest convex polygon that can contain the area of a bean seed.
Equivalent diameter (Ed)	The diameter of a circle having the same area as a bean seed area.
Extent (Ex)	The ratio of the pixels in the bounding box to the bean area.
Solidity (S)	The ratio of the pixels in the convex shell to those found in beans: A/C.
Roundness (R)	$(4\pi A)/(P^2)$
Compactness (CO)	Ed/L
ShapeFactor1 (SF1)	L/A
ShapeFactor2 (SF2)	l/A
ShapeFactor3 (SF3)	$\frac{A}{\frac{L^2}{4}\pi}$
ShapeFactor4 (SF4)	$\frac{A}{\frac{L}{2}\frac{l}{2}\pi}$
Class	Seker, Dermosan and Sira

Each group of 1 or 2 students draws an individual data set of random 200 beans from each class. The training data set consists of the first 100 observations from each class, the test set of the remaining 100 ones. You use the following code, where you change 0012345 by one of your student numbers.

```
set.seed(0012345)
library(readxl)
Dry_Bean_Dataset <- read_excel("Dry_Bean_Dataset.xlsx")
sample1 <- sample(which(Dry_Bean_Dataset[,17]=='DERMASON'),200)
```

```
sample2 <- sample(which(Dry_Bean_Dataset[,17]=='SIRA'),200)
sample3 <- sample(which(Dry_Bean_Dataset[,17]=='SEKER'),200)
mydatafull <- data.frame(Dry_Bean_Dataset[c(sample1,sample2,sample3),])
mytrainingdata <- mydatafull[c(1:100,201:300,401:500),]
mytestdata <- mydatafull[c(101:200,301:400,501:600),]
```

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 12 pages (12pt font size). Only report results and interpretations, do not repeat theory from the course! Additionally a separate file with the full R script should be provided.

One single folder containing your report and R script should be uploaded on Toledo before **April 30, 2021, 23h**. This project is graded on 10 points.

Good luck!

1 Exploratory analysis and transformation to normality

1. Perform an exploratory analysis on the **full** data set. State your main findings.

2. Now consider only the continuous variables. If there are variables whose distribution is very different from a normal distribution, try to transform it so that they are closer to a normal distribution. Briefly report which variables you transform, why and how you perform the transformation. For the following exercises you continue working with these, whether or not transformed, continuous variables.

2 PCA

1. Perform a PCA analysis on your **training** data. The data matrix should only consist of the continuous measurements. Please argue why you base the analysis on the correlation or covariance matrix of the data. Explain how you choose the number of components.

2. Make biplots of the first three scores (or only the first two if you have kept 2 components). Can you recognize the different types of beans (provided in the variable `Class`)?

3. Discuss whether the training data set contains PCA outliers. Compare your analysis with a robust PCA analysis.

4. Continue with the PCA analysis you find most appropriate. Consider now the observations from the **test** set, and compute their scores and predicted values. Compute the mean predicted area (A) for the 3 groups of the test set. Compare with the mean area for these 3 varieties.

5. Make an outlier map with the observations from both the training and the test set (use a different symbol or color). Discuss the result.

3 Clustering

Only consider the continuous variables of the **training** data set.

1. Perform a *partitioning* cluster analysis on the observations. Discuss your choice of the method, and the choice of the number of clusters. You are not allowed to use the `Class` information to make a decision.

2. Perform a *hierarchical* cluster analysis on the observations. Discuss your choice of the method, and the choice of the number of clusters. You are not allowed to use the `Class` information to make a decision.

3. Perform a *hierarchical* cluster analysis on the variables. Discuss your choice of the method.

4. Consider the clustering of the observations that corresponds mostly to the different types of beans (available in the `Class` variable). Make a heatmap where the observations and variables are sorted according to the selected clusterings. What are your findings?