‘From Statistics to Data Mining’ — Computer Lab Session Exam
Master 1 CPS^2
Saint-Étienne, France

Fabrice Muhlenbach

Preliminary Remarks

90mn-exam in ‘From Statistics to Data Mining’ using R *Project for Statistical Computing*. Documents (lecture and lab sessions notes) are allowed, R codes too, but not Internet searches, forums, e-mail, web chat or other kind of communication. The total score of all the exercises is equal to 100 points. Write your answers to the questions in your R code as a comment, in a sentence preceded by a sharp symbol (#). Use your name for naming your file (so your file will be FamilyName_FirstName.R) and send your file via the *Claroline* platform called ‘XAMFS2DMCPS2’ (*Exam From Statistics to Data Mining CPS2*).

1 Italian wines (50 points)

The “Wine recognition data” is a dataset with the the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines: the class of the cultivar, alcohol by volume, malic acid, ash, alcalinity of ash, magnesium, total phenols, flavanoids, nonflavanoid phenols, proanthocyanins, color intensity, hue, OD280/OD315 of diluted wines, and proline.

We want to find out if there are features –among the 14 features of the wine dataset– that follow a Normal distribution. For this, you will compute in the following questions for each feature how far the ratio of elements having a smaller value is from the probability of being smaller to the value computed using a Normal approximation.

Questions

5 pts (1) Load the wine.csv file (found on Claroline platform) with the read.csv function or “Import Dataset” in RStudio. Print the summary of the dataset.

10 pts (2) Write a code in R that returns the ratio (between 0 and 1) of elements in a column of the wine dataset that are smaller or equal to a given value. Using this code on the column “Class” should give a ratio 0 with a value 0, ratio 0.3314607 with value 1, ratio 0.7303371 with value 2 and ratio 1 with value 3 (note: you can make the code generic by writing a function).

10 pts (3) Write another code in R that returns the probability of having in a column of the wine dataset a value smaller or equal to a given threshold. This can be done by computing the mean and standard deviation of the given column, and then, using the pnorm function. Using this code on the column “Alcohol” with a value of 13 should give a probability of 0.4996963, and with a value of 14 a probability of 0.890844 (note: you can also make a function for doing this computation).

15 pts (4) Now for each of the 14 features of the wine dataset: (i) first generate a sequence of 50 elements between the minimum and maximum value in the feature (hint: seq function), (ii) then compute and display the mean for each value in the sequence of the absolute value between the difference of the ratio computed with the code from question 2 and the probability computed with the code from question 3. Which feature has the smallest mean value and what is the corresponding value?

10 pts (5) Finally, display an histogram of the feature you just found. Compute the mean and the standard deviation of this feature and plot with them a Normal distribution between the min and max of the feature as a red line on top of the histogram (without erasing it and without duplicating text on the axes).

2 Minke whales in Iceland (50 points)

An Icelander biologist wants to study the Minke whales (a type of baleen whale) caught in his country. He was able to collect 12 variables on 190 whales hunted between 2003 and 2007 even if some information was missing (written 'NA' for *Not Available* in the dataset). The variables of the file whale.csv are:

1. **whale.id**: unique identifier for the whale (not relevant)
2. **date.caught**: the date when the whale was caught
3. **lat**: latitude
4. **lon**: longitude
5. **area**: derived from location (North/South)
6. **length**: length of the whale
7. **weight**: weight of the whale
8. **age**: age of the whale
9. **sex**: Male or Female
10. **maturity**: maturity status of the whale
11. **stomach.volume**: volume (in liters) of the stomach content
12. **stomach.weight**: weight of the stomach content
13. **year**: the year when the whale was caught

Questions

10 pts (1) Download the file from the *Claroline* platform and load the CSV file using the function read.csv or with "Import Dataset" in RStudio. Print the summary. What are the variables with missing values? Which one cannot be used for doing statistical analysis? Remove this variable from the dataset. Then clean the data by removing the observations with missing values with the function na.omit.

20 pts (2) Install and load lubridate package (for working easier with dates). Extract the numbers of the month and the week from the variable "date" (functions month and week of lubridate package). Plot the histogram of the months and the weeks (you can use gglopt2 package for a better figure). By looking at these plots, when is whaling allowed? At what season do the whale hunters catch the most whales?

10 pts (3) Create a new data.frame having only numerical variables (with the "month" variable) by binding these variables by column with the cbind function. Compute a PCA on the dataset and plot the result on a biplot. What are the two pairs of attributes the most associated? How can you explain these associations?

10 pts (4) What are the two most negatively correlated variables? Print the table of the relation between these two attributes. How can you explain this negative association?