Ensimag – Grenoble INP – UGA
Statistical Analysis and Document Mining
Pedro L. C. Rodrigues
Jacopo Iollo

Year 2022-2023

pedro.rodrigues@inria.fr
jacopo.iollo@inria.fr

⚠ General guidelines for TPs

Each team shall upload its report on Teide before the deadline indicated at the course website. Please **include the name of all members of the team** on top of your report.

The report should contain graphical representations. For each graph, axis names should be provided as well as a legend when it is appropriate. Figures should be explained by a few sentences in the text. Answer to the **questions in order and refer to the question number in your report**. Computations and graphics have to be performed with `R`.

The report should be written using the `Rmarkdown` format. This is a file format that allows users to format documents containing text and `R` instructions. You should include all of the `R` instructions that you have used in the `rmd` document so that it may be possible to replicate your results. From your `rmd` file, you are asked to generate an `html` file for the final report. In Teide, you are asked to submit both the `rmd` and the `html` files. In the `html` file, you should limit the displayed `R` code to the most important instructions.

TP 3: Classification with genetic markers

The goal of this practical session is to use genetic markers to predict the geographical origin of an individual, which can be "North America", "Central America", or "South America". We will compare the results of classification using three approaches: multinomial regression (i.e. logistic regression with more than two classes), liner discriminant analysis (LDA), and the naive Bayes classifier.

Important. Set the seed of `R` at the beginning of your script so to ensure the reproducibility of your results. You can do this using `set.seed(0)` at the beginning of your script.

▶ Part 1: Data

Take the dataset `NAm2.txt` from Chamilo, and load it using `NAm2 = read.table("NAm2.txt", header=TRUE)`. Each row is an individual. Check that columns have explicit names. The third column contains the source populations of the individuals. Columns 7 and 8 contain the latitude and the longitude. Each column from 9th onwards is a genetic marker. To create a vector containing the continent of origin of each individual, run the following code:

```
cont <- function(x){
if(x %in% c('Canada'))
  cont <- 'NorthAmerica'
else if(x %in% c('Guatemala', 'Mexico', 'Panama', 'CostaRica'))
  cont<-'CentralAmerica'
else
  cont<-'SouthAmerica'
return(factor(cont))
}
contID<-sapply(as.character(NAm2[,4]),FUN=cont)
```

▶ Part 2: Multinomial regression

Load the `nnet` package via `library("nnet")`.

(a) Use the `multinom` function to estimate the parameters of a multinomial regression model for predicting the continent of individuals based on their genetic markers. To do so, you can create a table containing the contineng of origin in the first column followed by the genetic markers:

```
NAcont <- cbind(contID=contID, NAm2[,-(1:8)])
NAcont[,1] <- factor(NAcont[,1])
```

What happens? Why? You may avoid this problem by setting `MaxNWts = 18000` and `maxit = 200` when calling the `multinom` function. Can you explain why we now avoid the problems reported just before?

(b) Compute a PCA on the genetic markers (using `scale = FALSE}`) and use the maximal number of principal components as predictors of a new multinomial regression model. Using the `table` function, compute the confusion matrix between predicted and true classes and comment your results.

(c) Select an optimal number of principal components (PCs) using a 10-fold cross-validation procedure and use the mean classification error on the test set for selection.

Remark: You may start with a step of 10 PCs (i.e. increasing the number of PCs by 10 on each cross-validation) to find an initial guess $\hat{P}$ and then refine your results around $\hat{P} - 9$ and $\hat{P} + 9$. Plot the curve of the mean test classification error in terms of the increasing number of PCs represent their confidence bounds as vertical segments. Comment your results.

(d) Repeat question (b) now using the number of PCs found in question (c). Represent the confusion matrix between predicted and true classes and compare it to your result in (b). Comment your results and the reasons for which you might have had better/worse results now.

▶ Part 3: Linear discriminant analysis

Load the `MASS` and `class` packages.

(a) Repeat Questions (b) to (d) from Part 2 but using the `lda` classifier.
(b) Compare the results obtained with LDA versus multinomial logistic regression.

▶ Part 4: Naive Bayes classifier

Load the `naivebayes` package.

(a) Repeat Questions (b) to (d) from Part 2 but using now the Naive Bayes classifier with a Bernoulli distribution. You may want to check the documentation of the `naivebayes` pacakge.
(b) Compare your new results with those using multinomial regression and LDA. Comment the differences between the performance of each classifier and try to explain why one of them seems superior.