Nathalie Villa-Vialaneix
Année 2015/2016

M1 in Economics and Economics and Statistics
Applied multivariate Analysis - Big data analytics
Final exam

General recommandations

The project must be returned as a PDF file with a separate R script (.R file). Examples of expected answers are provided at http://www.nathalievilla.org/doc/pdf/ex_answers_M1SE.pdf (PDF file) and http://www.nathalievilla.org/doc/R/ex_answers_M1SE.R (R script).

Exercice 1 Example exercise

This section is an example of the kind of answers expected from you. The typical solution is provided at http://www.nathalievilla.org/doc/pdf/ex_answers_M1SE.pdf (PDF file) and http://www.nathalievilla.org/doc/R/ex_answers_M1SE.R (R script). Do not include it in your own essay!

Using the public iris dataset (available with `data(iris)`), sample at random 50 observations and make a barplot to represent the distribution of the different Species in this dataset. Use the random seed `set.seed(1609)` at the begining of your script. The final document must contain a title, your names and a short description of the dataset. The answer to your question must be commented and you must provide the R script which allowed to obtain it.

Exercice 2 Presentation of the dataset

The final exam is based on the data described at http://archive.ics.uci.edu/ml/datasets/Online+News+Popularity. Download the zip file in “Data Folder” and uncompress it. It contains two files, one with the data set and the other file describind it. Write an introduction with a short description of the data.

Exercice 3 Bootstrap

This section aims at giving an estimate of the mean of the number of shares using only a small subsample of the whole data set.

First import the data set in R and find the true mean of the number of shares. Then, take a sample of 5000 articles and use a bootstrap approach to provide a 95% confidence interval of the mean of the number of shares based on this small subsample. Finally, print a histogram of the mean distribution as estimated by the bootstrap approach with the true mean highlighted by a vertical line. Use 5000 bootstrap samples and start your script by setting the random seed by: `set.seed(1700)`.

Exercice 4 Bagging

1. Create a new variable which is a factor with two values: `"above"` for articles which have a number of shares equal to or above the median (1400 shares) and `"below"` for articles which have a number of shares below the median. Then remove from the dataset non informative variables (the first two columns) and columns related to the original target variable (`shares`). Finally, split the data set into a train data set that contains the 5000 observations sampled in the previous exercise and a training data set that contains the remaining observations.

2. Use the package **ipred** to obtain a bagging of regression trees which predicts the new variable (with values "above" and "below") from all the other informative variables. Use only the training test to train the model with $B = 100$ bootstrap samples. What is the OOB error of this model? What is its test error? Also report the computational time needed to train the model. Start your script by setting the random seed `set.seed(1213)`.

Exercice 5 Random forest

Train a random forest on the same training data set and with 500 trees (5 times more than in the previous exercise). What is its OOB error and its test error (with the same test set as in the previous exercise)? What is the computational time needed to train this forest? Start your script by setting the random seed `set.seed(1459)`.

Exercice 6 Parallel computing

Train the previous forest in parallel and compare the computational times. What is the test error of this forest? (Do not set a random seed for this exercise.)