Nathalie VILLA-VIALANEIX
Année 2016/2017

M1 in Economics and Economics and Statistics
Applied multivariate Analysis - Big data analytics
Final exam - 1h30

This exam uses the dataset available at `http://www.nathalievilla.org/teaching/m1se/fbpopularity.csv`. This dataset is a text file (columns are separated by commas) that is described at this link: `https://archive.ics.uci.edu/ml/datasets/Facebook+Comment+Volume+Dataset`. In particular, the purpose of the dataset is to describe the last column which corresponds to the number of comments related to a post. The post is described in the previous variables that are provided in the previous columns of the file (the variables are all described at the end of the web page cited above). Some variables are factors encoded with numeric codes (you have to be careful with that when using the dataset).

More information are given in this article:

Singh, K., Sandhu, R.K. and Kumar, D. (2015) Comment volume prediction using neural networks and decision trees. *In Proceedings of the 17th International Conference on Computer Modelling and Simulation (UKSim2015)*, Cambridge, UK.

Answer the questions below. The answers must include comments (if requested), R script and output of the script. Most questions are independant so do not stay too long on one question if you don't know how to answer it. You are strongly advised to use RMarkdown file. Answers must be sent by email at `mailto:nathalie@nathalievilla.org` (PDF or HTML files only). You are responsible to check that I have received your email properly before leaving the exam room.

```
## Loading required package: foreach
## Loading required package: iterators
## Loading required package: parallel
## randomForest 4.6-12
## Type rfNews() to see new features/changes/bug fixes.
```

Exercice 1 Data importation and preparation

1. Import the dataset and convert to factors all variables that are supposed to be factors (they are described as "Value Encoding" or "Binary Encoding" on the webpage. Transform the last variable (target variable) with a log transformation (after adding the value 1 to avoid NA produced by posts having zero comments).
2. How many variables and observations does the dataset contain?
3. Create a training dataset with 10000 observations (sampled at random) and put the remaining observations into a test dataset.

Exercice 2 Bootstrap

This section aims at giving an estimate of the mean of the (log transformed) target variable using only the data in the training dataset. If you're unsure about your importation of the dataset in the previous exercise, you can load the data at `http://www.nathalievilla.org/teaching/m1se/fbpopularity.rda`. In this file, `df` is the full dataset, `train_df` the training dataset and `test_df` the test dataset.

1. Compute the true mean of the log-transformed target variable in the whole dataset (its name is `V54` if you are using my data).

2. Use the function `boot` from the package `boot` to obtain a 95% confidence interval of the mean from the training dataset by a bootstrap approach. Use $B = 5000$ bootstrap samples and save the computational time.

3. Using a `foreach` loop run in parallel with the maximum number of cores minus 1, obtain a bootstrap estimate of the 95% confidence interval for the mean from the whole dataset. Save the computational time: how does it compare with the previous one?

Exercice 3 Bagging

1. Use the package **ipred** to obtain a bagging of regression trees which predicts the target variable from all the other variables. Use only the training dataset to train the model with $B = 100$ bootstrap samples. What is the OOB error of this model? What is its test error (compute the mean squared error, *i.e.*, the average of the squared difference between the predicted and the true value)? Also report the computational time needed to train the model.

Exercice 4 Random forest

1. Remove variable `V4` from the training and test datasets.

2. Train a random forest on the new training dataset with 100 trees. Save the computational time, compute the OOB and test errors. How do these results compare with the ones obtained in the previous exercise?