Nathalie VIALANEIX
Année 2018/2019

M1 in Economics and Economics and Statistics

Applied multivariate Analysis - Big data analytics

Worksheet 3 - Random Forest

This worksheet's aim is to learn how to train and use random forests with R. The method is implemented in the package **randomForest** which is to be installed and loaded. The method is illustrated on examples taken from the UCI Machine Learning repository[1], which are available in the R package **mlbench**:

```
library(randomForest)

## randomForest 4.6-14
## Type rfNews() to see new features/changes/bug fixes.

library(mlbench)
```

Exercice 1 Random forest: training, interpretation and tuning

This exercise illustrates the use of random forest to discriminate sonar signals bounced off a metal cylinder from those bounced off a roughly cylindrical rock. Data are contained in the dataset `Sonar`:

```
data(Sonar)
```

1. Using

   ```
   help(Sonar)
   ```

   read the description of the dataset. How many potential predictors are included in the dataset and what are their types? How many observations are included in the dataset?

2. Randomly split the data into a training and a test sets of size 100 and 108, respectively.

3. Use the function `randomForest` to train a random forest on the training set with default parameters (use the "formula" syntax described in `help(randomForest)` and the options `keep.inbag=TRUE` and `importance=TRUE`). What is the OOB error? Make a contingency tables between OOB predictions and real classes.

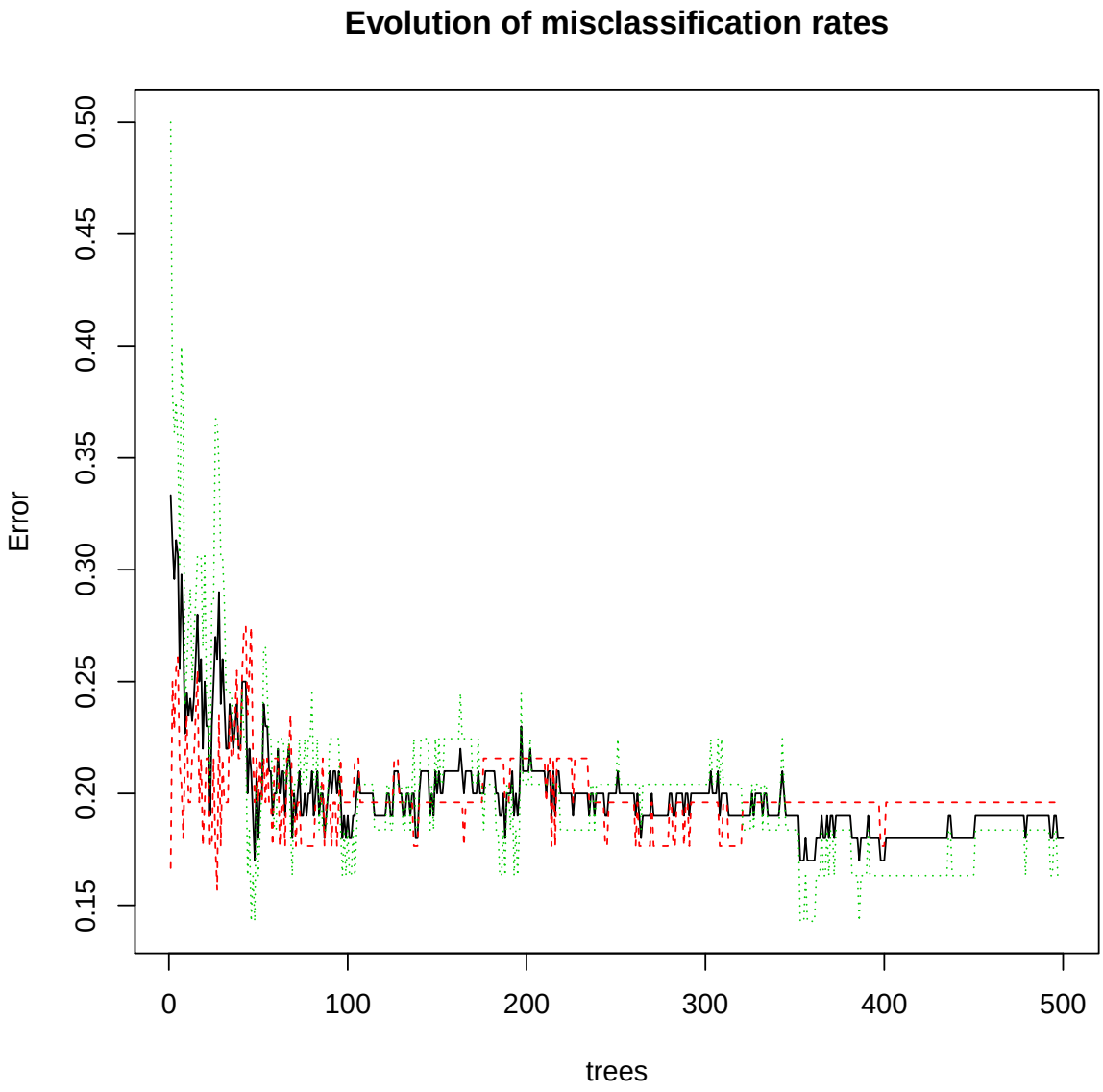
4. Using the help, describe what is in `predicted`, `err.rate`, `confusion`, `votes`, `oob.times`, `importance`, `ntree`, `mtry` and `inbag` in the random forest object trained in the previous question.

5. What is the OOB prediction for the $50th$ observation in your training sample? What is the true class for this observation? How many times has it been included in a tree and in which trees has it been included? How many trees voted for the prediction "R" for this observation?

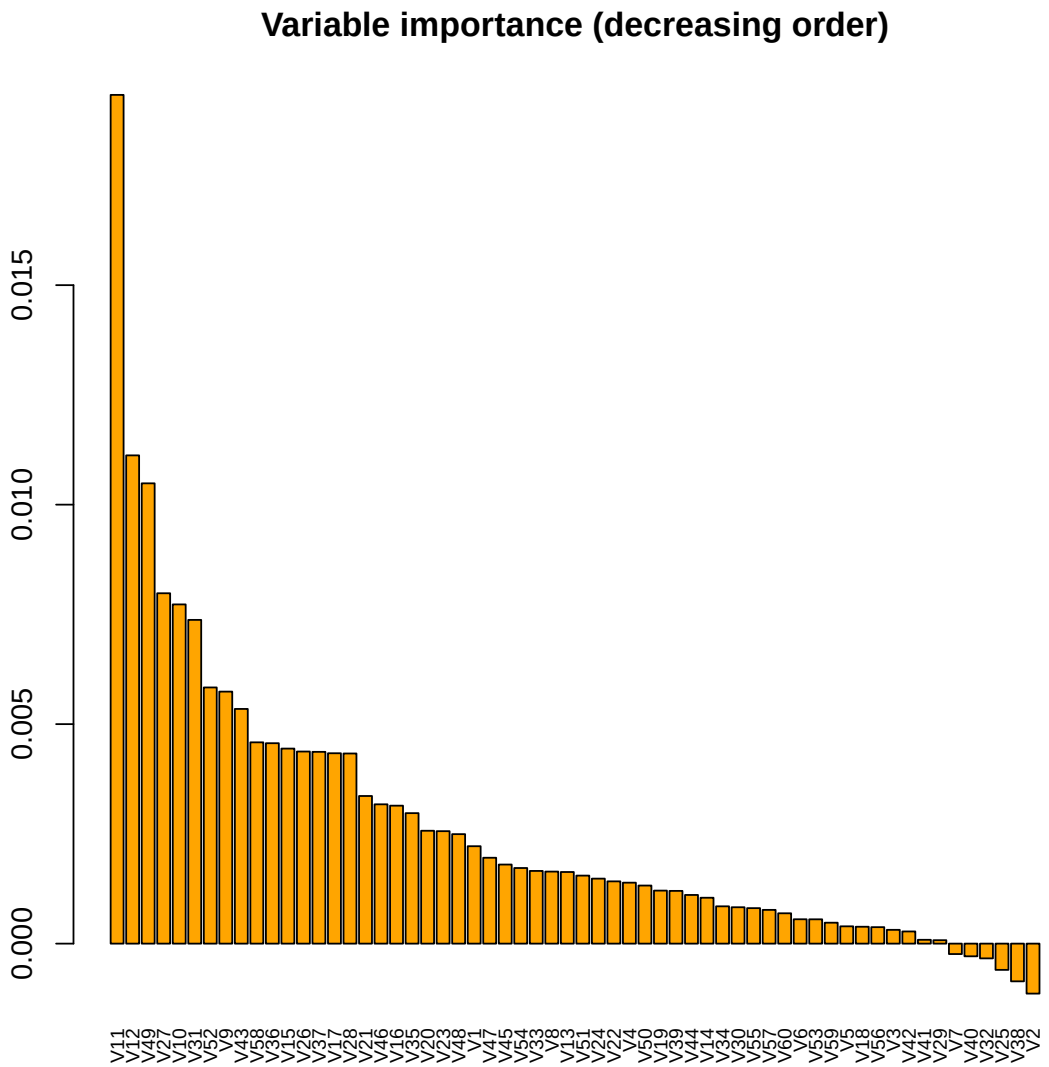
6. Using the `plot.randomForest` function, make a plot that displays the evolution of the OOB misclassification rate when the number of trees in the forest increases. What is represented by the different colors? Add a legend and comment.

[1] http://archive.ics.uci.edu/ml

The figure should look like this one:



7. Make a barplot of the variable importances ranked in decreasing order. Comment.
   The figure should look like this one:



8. With the function `predict.randomForest`, predict the output of the random forest on the test set and find the test misclassification rate. Comment.

9. Use the package **e1071** and the function `tune.randomForest` to find, with a 10-fold cross validation performed on the training set only, which pair of parameters is the best among `ntree=c(500, 1000, 1500, 2000)` and `mtry=c(5, 7, 10, 15, 20)`. Set the option `importance` to `TRUE` to obtain the variable importances for the best model. Use the functions `summary.tune` and `plot.tune` to interpret the result.

```
library(e1071)
```

10. From the previous question, extract the best model (in the object obtained with the function `tune.randomForest`, it is in `$best.model`). Compare its OOB error and its evolution, its most important variables and its test error with the random forest obtained with default parameters. Comment.

Exercice 2 Comparison between random forest and CART boosting

This exercise's aim is to compare random forest and CART boosting from an accuracy point of view. The notion of importance is also introduced for CART boosting. The comparison is illustrated on the dataset `Vehicle`

```
data(Vehicle)
```

(package `mlbench`) which purpose is to classify a given silhouette as one of four types of vehicle, using a set of features extracted from the silhouette.

1. Using

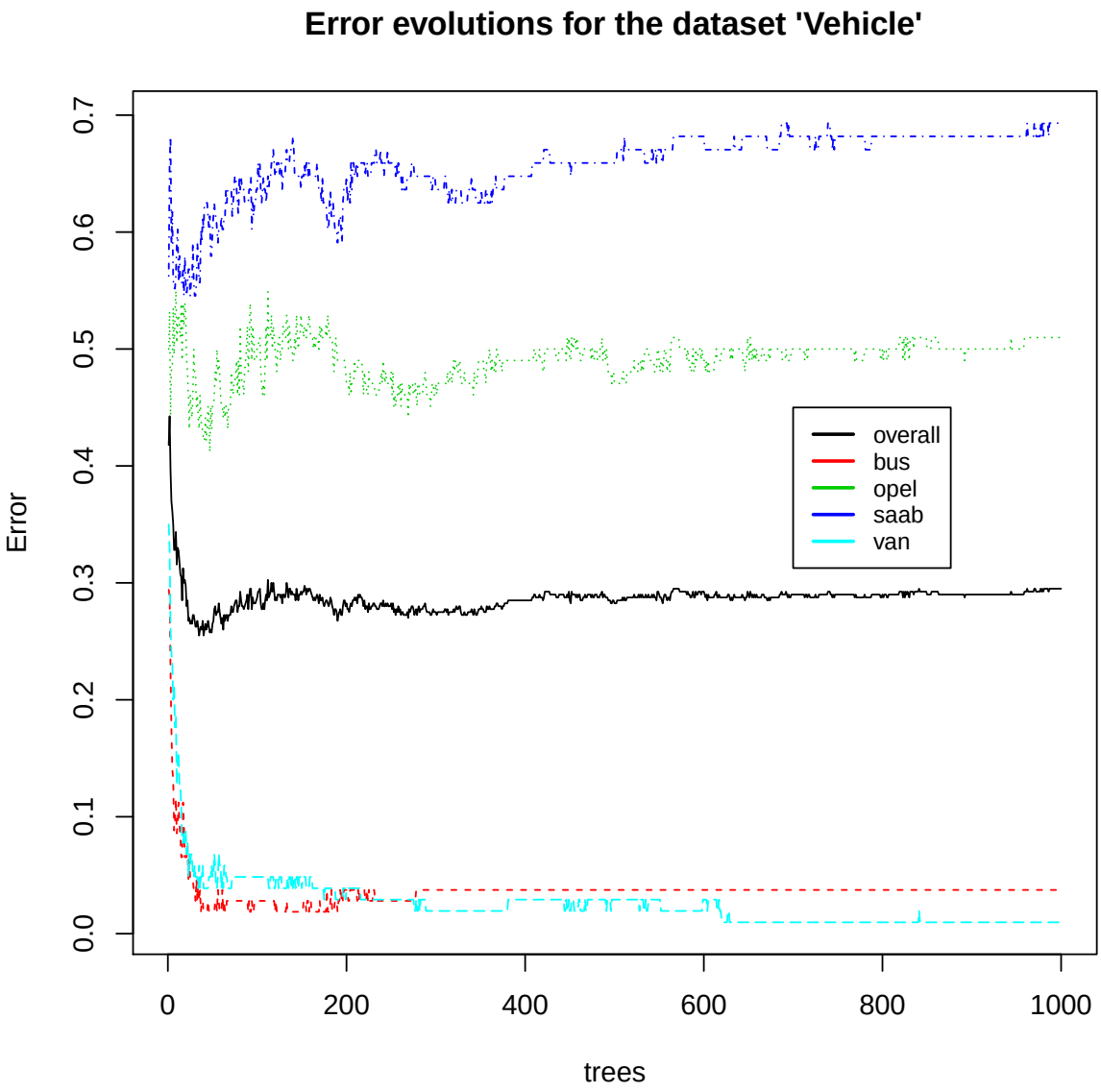
   ```
   help(Vehicle)
   ```

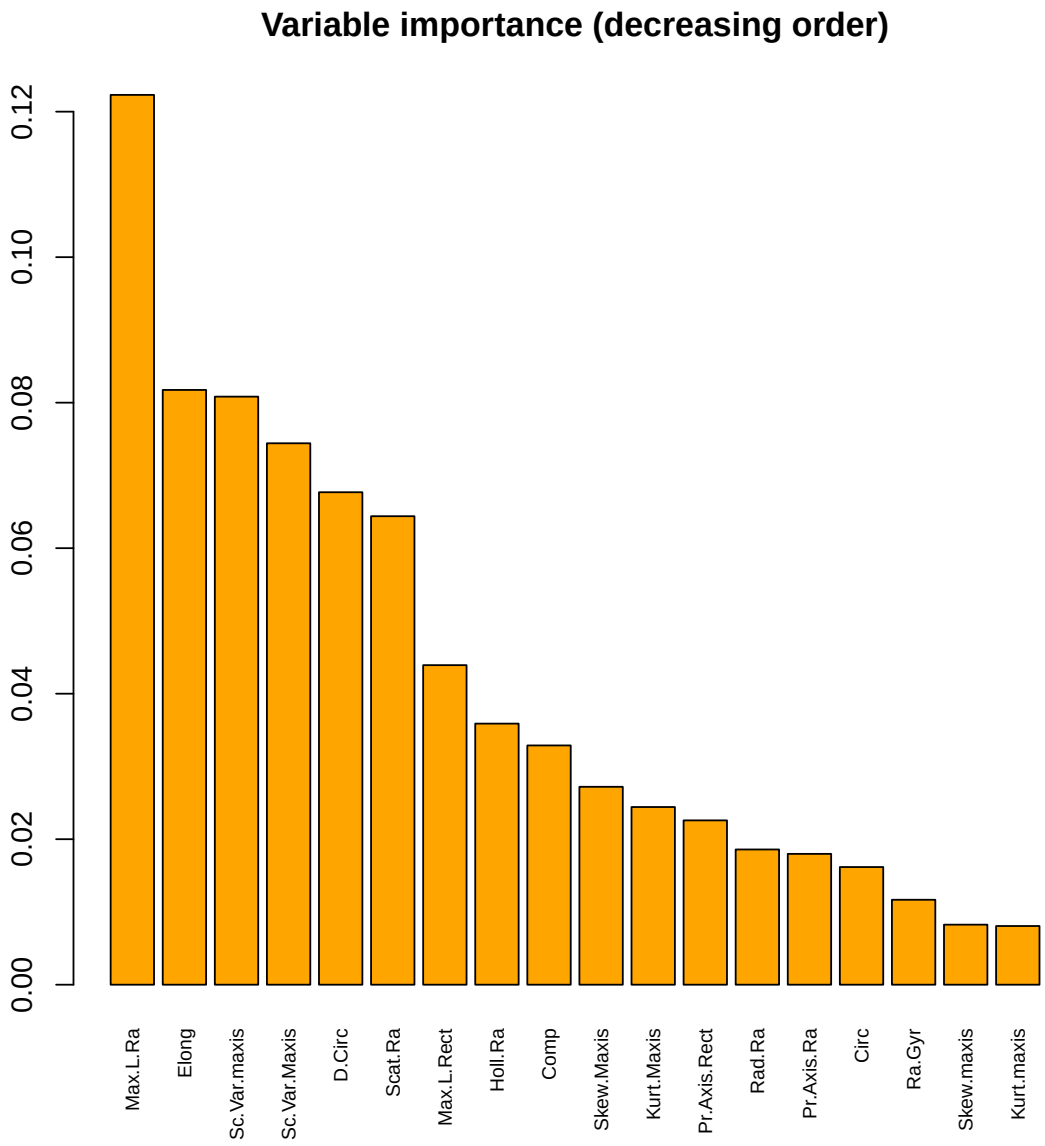
   read the description of the dataset. How many potential predictors are included in the dataset and what are their types? How many observations are included in the dataset?

2. Split the data set into a training and a test sets, having respective sizes 400 and 446.

3. Using the function `tune.randomForest`, train a random forest with possible number of trees 500, 1000, 2000 and 3000 and possible number of variables selected to every split 4, 5, 6, 8, 10. Keep the variable importances and use the options `xtest` and `ytest` to obtain the test misclassification error directly. Analyze the results of the tuning process.
   Warning: When using the options `xtest` and `ytest`, the forest is not kept unless setting `keep.forest=TRUE` which means that it cannot be used to make predictions with new data. For using it with the method `tune`, you must thus set `keep.forest=TRUE` or the function will return an error while trying to compute the CV error.

4. Analyze the best forest obtained with the tuning process in the previous question in term of out-of-bag error, test error and their evolutions. What are the most important variables in this model?
   The figures expected for this question should look like the following ones:


Error evolutions for the dataset 'Vehicle'
Error
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
200
400
600
800
1000
trees
overall
bus
opel
saab
van



5. This question aims at building a simple bagging of trees to compare with the previous the random forest. The package `boot` and `rpart`

```
library(boot)
library(rpart)
```

will be used: to do so, answer the following questions:

(a) Define a function

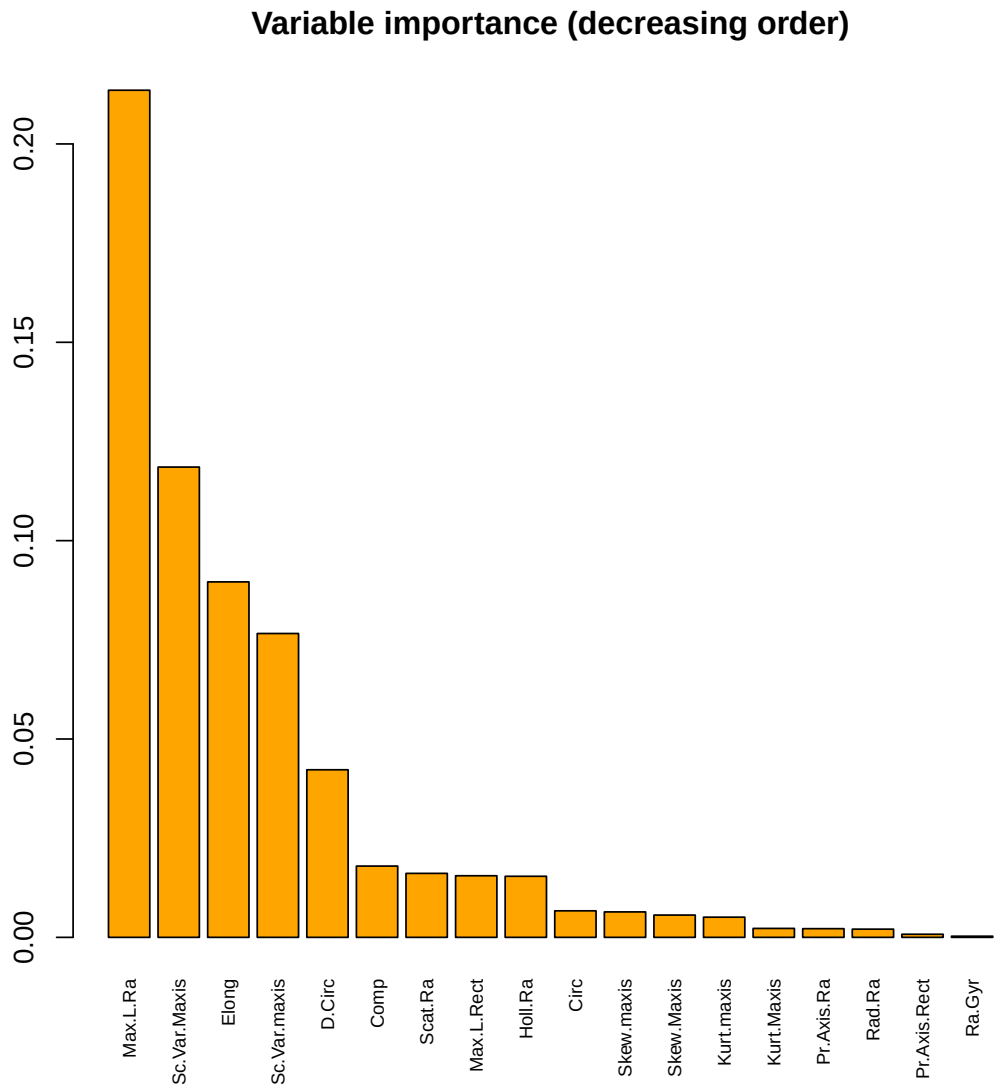
```
boot.vehicle <- function(x, d) {
  ...
}
```

that returns 3 values:

- the OOB misclassification rate for the classification tree trained with the subsample `x[d, ]`;
- the test misclassification rate for the same classification tree;
- the misclassification rates obtained with the same classification tree on OOB observations when one of the variable has been permuted (it is thus a vector of size 18, one value for each variable).

(b) Use the function defined in the previous question with `boot` to obtain bootstrap estimates of the OOB misclassification rate (computed only from the OOB observations in the training data set) and of the test misclassification rate for a bagging of R trees with R being the optimal parameter found in question 3.

(c) Use the result of the previous function to compute and represent the variable importances.
The result should look like the following figure:



6. Analyze the results (OOB and test errors, important variables) of the bagging of classification trees by comparing them with the ones obtained from a random forest.