Nathalie Vialaneix
Année 2018/2019

M1 in Economics and Economics and Statistics
Applied multivariate Analysis - Big data analytics
Worksheet 2 - Bagging

This worksheet illustrates the use of bagging for classification. Bagging is used in conjunction with different types of models and compared to the original classifier. The simulations are carried out on datasets coming from the UCI Machine Learning repository[1], which are available in the R package **mlbench**:

```
library(mlbench)
```

Exercice 1 Bagging CART

This exercise illustrates the use of bagging for predicting the presence of some type of structure in the ionosphere from radar data. Data are contained in the dataset `Ionosphere`:

```
data(Ionosphere)
```

and will be analyzed using classification trees (package **rpart**):

```
library(rpart)
```

1. Using

   ```
   help(Ionosphere)
   ```

   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. Using the function `sample`, split the dataset into a training and a test datasets, of respective size 201 and 150.

3. Using the training dataset, define a classification tree for predicting the class from all the other variables (with default parameters of the function `rpart` and no pruning). Print a summary and a graphical representation of this model and find its training error.

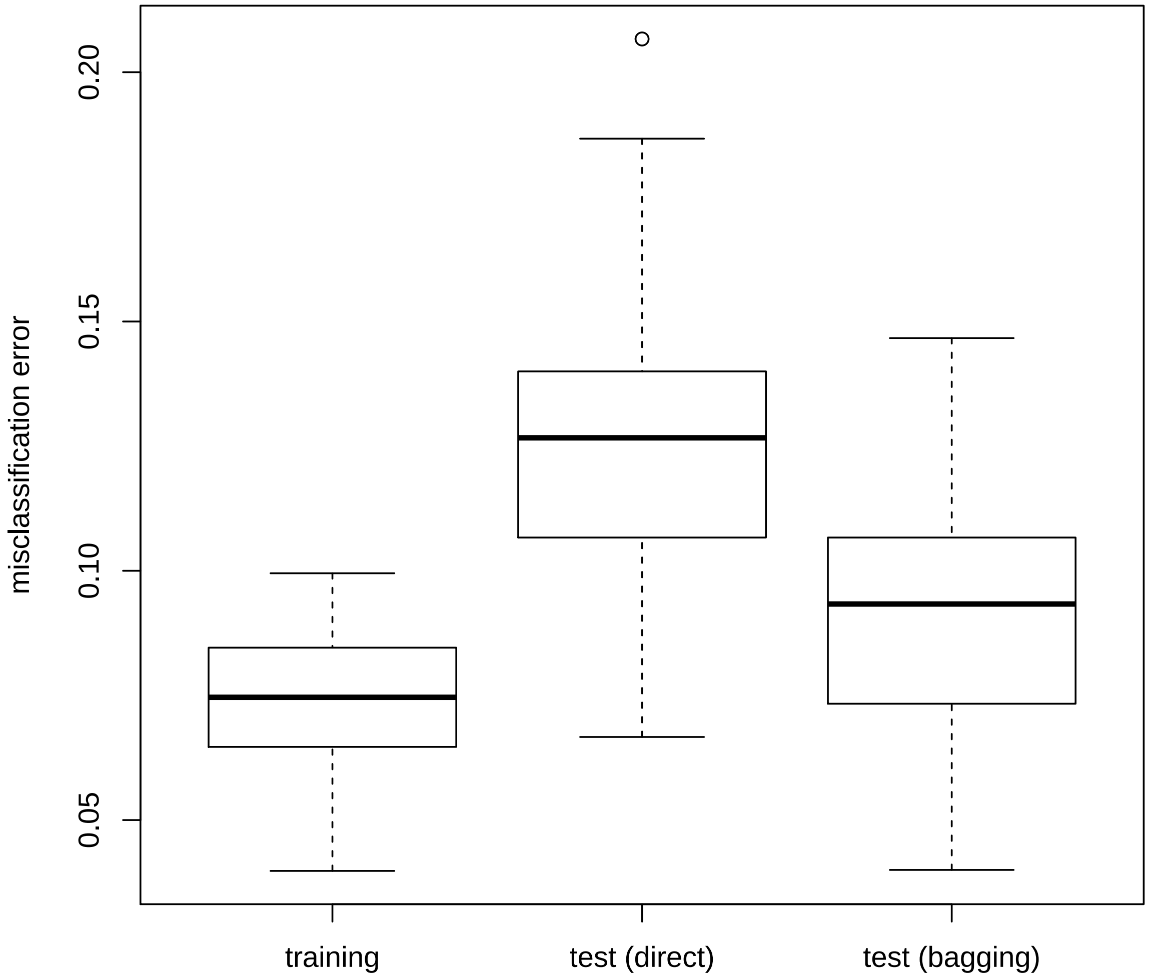
4. Find the test error for this model. What can you say comparing the two errors?

5. Design a function for using with the function `boot` of the package **boot** which takes two arguments `x` and `d` and returns the prediction for the test set of a CART model trained from a bootstrap sample with identifiers `d` coming from the dataset `x`. Test your function with `x` being the training dataset obtained in question 2 and a bootstrap sample manually generated using the function `sample`.

6. Use the function written in the previous question to obtain $B = 100$ predictions from bootstrap samples for the test dataset. What is the final prediction for the test dataset obtained from a bagging approach? What is the test error obtained from the bagging approach? Compare it with the test error obtained for the direct approach.

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

7. Repeat the whole procedure for 100 different splits of the dataset into training and test sets to obtain 100 train, test and bagging test errors (use a `for` loop). Compare the distributions of these errors with boxplots.



Exercice 2 Using the R package ipred

This exercise will illustrate the use of the package **ipred** for bagging. The dataset used in this exercise is the dataset `PimaIndiansDiabetes` :

```
data(PimaIndiansDiabetes)
```

which goal is to predict diabete from description of people in the Pima indians population.

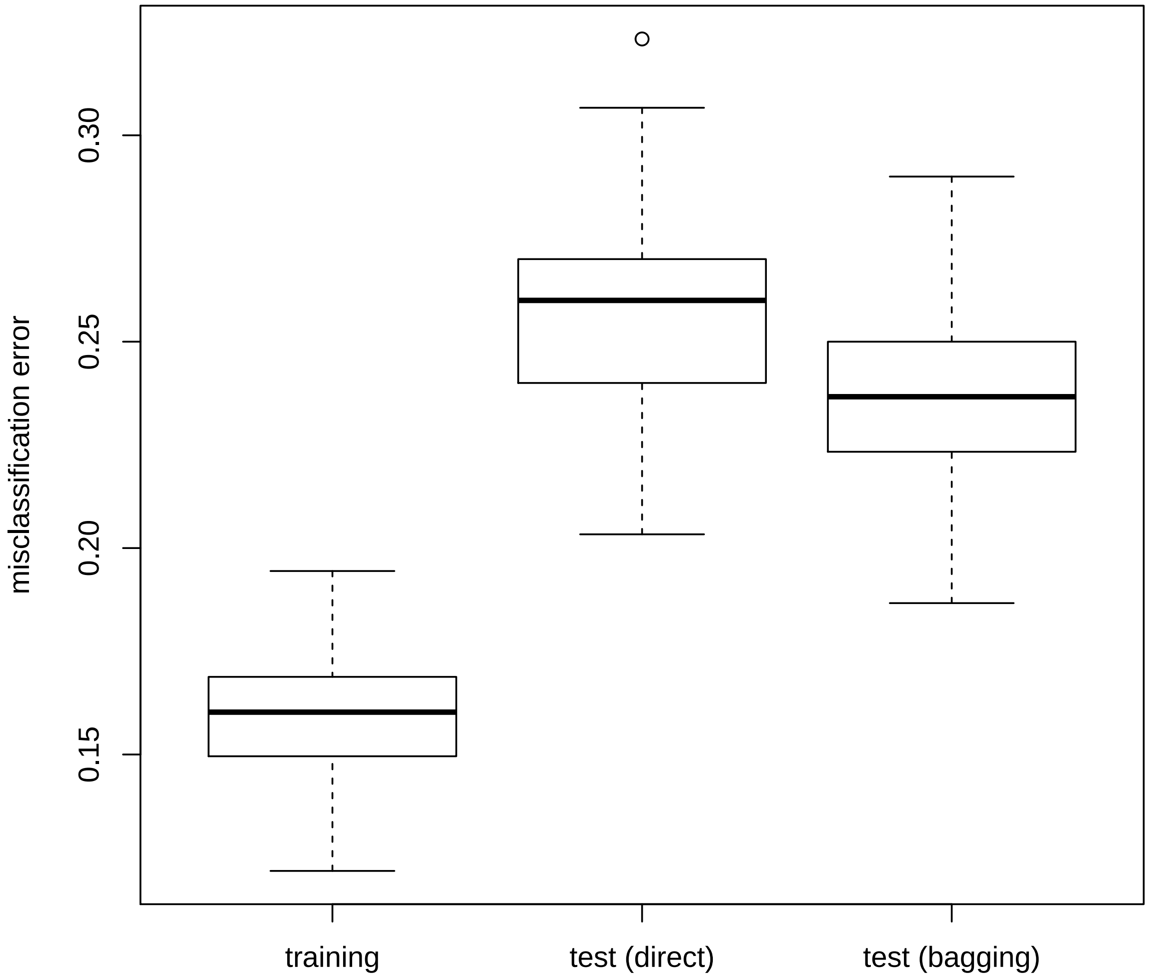
1. Using

   ```
   help(PimaIndiansDiabetes)
   ```

   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. Using the same approach than in the previous exercise (with the package **boot**, a test sample size equal to 300 and $B = 100$ bootstrap samples and 100 replicates of the splitting between training and test sets), find the distribution of

training, test and test with bagging errors. Display the distributions with boxplots. What is the computational time of the approach?

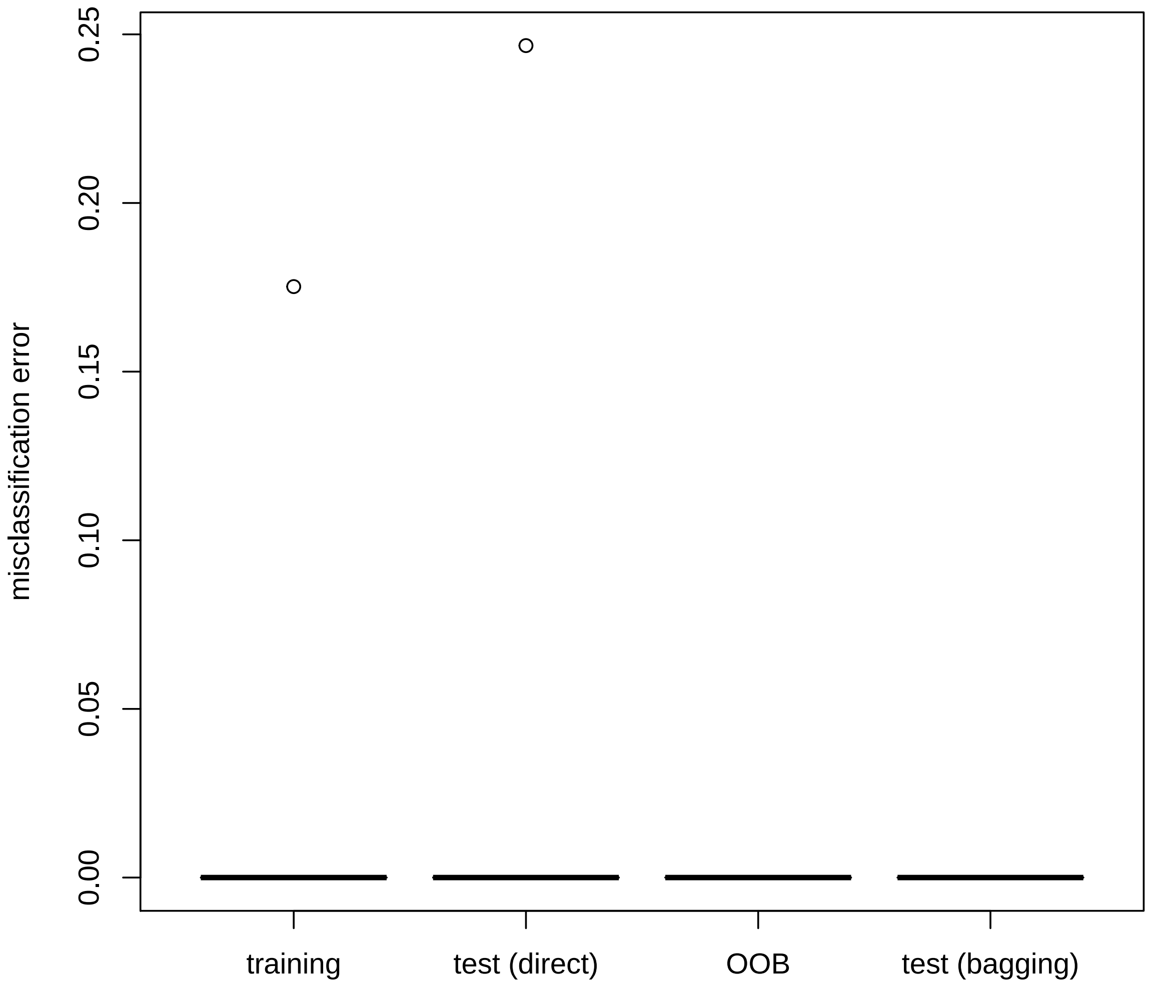


3. In this question, the package **ipred** is used to obtain a bagging estimate of the prediction.

```r
library(ipred)
```

Split the dataset into test and training sets of the same size than in the previous question and use the function `bagging` to obtain a bagging classification function. Obtain the out-of-bag and bagging test errors (for the first one, check the help page of the function and for the other one, use the function `predict`).

4. Repeat this method 100 times to obtain the distribution of the OOB and bagging test errors. Also obtain the direct training and test errors for CART. What is the computational time with this function? Display the four error distributions with boxplots.



Remark: The package **ipred** can be used directly for bagging trees but it provides a simple way to use bagging with other methods. For an example, see

```
help(ipredknn)
```

Exercice 3 Bagging with k-nearest neighbors

This exercise will compare direct and bagging approach for the k-nearest neighbors algorithm[2] and the CART algorithm. The k-nearest neighbors algorithm is known to be rather stable. The comparison will be carried out on a letter recognition problem:

```
data(LetterRecognition)
```

which goal is to predict handwritten letters from a description of the images.

1. Using

[2] See http://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm if you do not know this method.

```r
help(LetterRecognition)
```

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 dataset into a test and training datasets, of respective size 6,000 ad 14,000.

3. Use the functions `rpart` and `bagging` to obtain direct and bagging test errors ($B = 100$) for a classification tree.

4. Use the function `tune.knn` of the package **e1071** for tuning a k-nearest neighbors classifier with an optimal number of neighbors chosen between 1 and 10 with a 10-fold CV strategy. Check the help pages

```r
help(tune.knn)
help(tune.control)
```

and in particular the arguments `k`, `tunecontrol` (for `tune.knn`) and `sampling` and `cross` (for `tune.control`). What is the optimal number of neighbors found by CV? Use this value to train a k-nearest neighbor algorithm with the function `knn`, using the argument `test` to compute the predictions for the test dataset. What is the classification error of this model?

```r
library(class)
library(e1071)
```

5. Make a bagging of k-nearest neighbors models with k equal to the optimal number found in the previous question (use the `boot` package). What is the test error for this method?