UNIVERSIDAD DE LA REPÚBLICA URUGUAY

FACULTAD DE INGENIERIA

Self-Practice Sessions

CART Trees and Random Forests - Jean-Michel POGGI
Master 2 Course in Statistics

Universidad de la República – Facultad de Ingeniería, Montevideo, Uruguay
February 2025

Guide for the self-practice sessions.
Produce a report (of 10 to 15 pages) with the code to answer the questions below, with your comments. You can also add some material to introduce the different elements of the course.

1. Data

1. Load the library **`kernlab`**
2. Load the dataset **`spam`** in R and build the *dataframes* of learning and test sets (the first will be used for designing trees, the second for evaluating errors)

2. CART trees

1. Load the library **`rpart`**
2. Compute the default tree provided by **`rpart`**
3. Build a tree of depth 1 (stump) and draw it
4. Examine splits primary splits and surrogate splits
5. Build a maximal tree and draw it
6. Draw the cross-validation errors of the Breiman's sequence of the pruned subtrees of the maximal tree and interpret it
7. Find the best of them in the sense of an estimate given by the cross-validation prediction error
8. Compare the errors of the different trees obtained, both in learning and in test

3. Random Forests

1. Load the library **`randomForest`**
2. Build a RF for *mtry=p* (unpruned bagging) and calculate the gain in terms of error with respect to a single tree
3. Build a default RF
4. Calculate an estimate of the prediction error and compare it to bagging
5. Study the evolution of the OOB error with respect to *ntree* using `do.trace`

4. Variable importance for Random Forests

1. Calculate the variable importance of the `spam` variables for the default RF
2. What are the most important variables?
3. Calculate the importance of `spam` variables for stumps RF
4. Illustrate the influence of the *mtry* parameter on the OOB error and on the VI

5. Gradient boosting

1. Load the **`gbm`** library
2. Show how to estimate the model using the `adaboost` loss function.
3. Draw the curves showing the evolution of the training and test errors as a function of the number of iterations. Comment on them.
4. Calculate the error rate on the training set, then on the test set of the selected model. Comment on the results.
5. Which variables have the greatest influence on the selected model? Please comment.
6. Compare with the variables highlighted using random forests.
7. Study the effect of the different parameters: number of iterations (trees), tree depth (default 1) and regularisation parameter.
8. Compare the performance of the selected model with the analogue model using the logit loss function.