CSC6515 – Machine Learning for Big Data

Assignment 1
Oct. 9, 2013
Due date: Monday October 26, 2015 - 11:55pm

In this assignment you will practice decision trees, random forests and naïve Bayes classifier using R. You will also practice cross-validation as an evaluation technique and a statistical significance test.

Deliverables:

- A report file in PDF format that includes your results, accuracies, confusion matrices, plots, discussions and any other explanations that you might have for the tasks defined below.
- Your R source code.

Submissions:
Please upload your completed assignment as a single zip file to the moodle (courses.cs.dal.ca). The filename MUST include your last name and your banner number (e.g. A1_HajiSoleimani_B00444444.zip).

Dataset:
Use the provided wine quality dataset. It is a classification task, so don't forget to use factor on the target variable. The target variable is the last column in the CSV file. We simplified the dataset by removing underrepresented classes.

- Number of instances: 4873
- Number of features: 11
- Number of classes: 5

Useful R packages:
You can use the "rpart" package in R for training a decision tree. It has an "rpart" function for growing the tree and a "prune" function for pruning. "rpart" package is generally capable of growing classification and regression trees. Therefore, you have to specify that you need a classification tree. You may find the following link useful:
http://www.statmethods.net/advstats/cart.html
Package "DMwR" also has an "rpartXse" function that grows and prunes a decision tree in one step. It also has a "prettyTree" function for visualizing the tree. For the random forest you can use "randomForest" package in R. This package also builds classification and regression models. For the naïve Bayes classifier you can use "klaR" package. These packages are just suggestions but you can use any other package that you want.

Your task:

(a) Split the data randomly into a training set and a testing set (e.g. 65%-35%). Train a decision tree classifier using the train data. Report the confusion matrix and accuracy for both train and test data. Compare the train and test accuracy. Is there a big difference between train and test accuracy? Why? Finally, visualize the tree.

(b) Repeat (a) and after training the classifier, prune the decision tree. Report the confusion matrix and accuracy for train and test data using the pruned tree and compare it to those obtained from (a). Visualize the tree and discuss the effect of pruning on generalization ability of the model. How the train and test accuracy changes by pruning? Why?

(c) Using 10-fold cross-validation, train and evaluate a random forest and a naïve Bayes classifier. Compare the accuracy of the two methods in terms of mean (μ) and standard deviation (σ) of accuracy in 10 folds. Eventually use a statistical significance test (e.g. student's t-test) and determine whether the two methods are significantly different or not. Use $\alpha = 0.05$ as the significance threshold.

Please mention that whenever you are asked to compare the results, you have to discuss on the results and provide acceptable reasons that justifies your results.

Also, please feel free to use any kind of plots (e.g. bars, boxplots, …) in order to visualize your results.