CSC6515 – Machine Learning for Big Data

Assignment 1
Sep. 21, 2017
Due date: Thursday, October 05, 2017 – 2:30pm

In this assignment, you will practice logistic regression, decision tree, random forest and naïve Bayes classifier using Python. 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. (Maximum 10 pages.)
- Your Python source code.

Submissions:
Please upload your completed assignment as a single zip file. The filename MUST include your last name and your banner number (e.g. A1_AmilcarSoares_B00444444.zip).

Dataset:
Use the provided Animals classification dataset. The target variable is the last column in the CSV file. Each row in the file labeled with one of the following categories:

1 Elk (52%)
2 Deer (28%)
3 Cattle (20%)

Other information about the dataset:

- Number of instances: 5135
- Number of features: 25
- Number of classes: 3

Useful Python packages:

- **Numpy:** multidimensional arrays, vector and matrix operations
- **Pandas:** data manipulation and analysis
- **Scikit-learn:** machine learning library for classification, regression, clustering, feature selection and much more

Your task:

(a) Split the data randomly into a training set and a testing set (e.g. 70%-30%). Train all classifiers (Logistic Regression, Naïve Bayes, Decision Tree and Random Forests) using the default parameters 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?

(b) Using 10-fold cross-validation, train and evaluate all classifiers. Compare the accuracy of the 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 methods are significantly different or not. Use $\alpha = 0.05$ as the significance threshold. For applying the significance test, select the classifier with the best average performance, and compare it to all the remaining classifiers.

(c) Train a Random Forest using a 10-fold cross-validation with the 10, 20, 50 and 100 trees (e.g. number of estimators in the scikit package) and report the mean accuracies. Choose one of the solutions, justify why you chose it, and compare it again with your results for Logistic Regression, Naïve Bayes, and Decision Tree using the student's t-test.

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.

For questions regarding the assignment, contact by email amilcar.soares@dal.ca