Machine Learning (BITS F464)

Assignment 1

(Decision Tree, Random Forest, Boosting Techniques)

Submission Date: 2330Hrs on 10th Nov 2016
Max Marks: 30

Languages allowed:- C, C++, Java**.** You are not allowed to use any packages for the assignment, all the functionalities you have to code on your own and do proper documentation so that it is readable.

Definitions:
ID3 (Decision Tree Learning)**:-** ID3 algorithm takes greedy approach to find the best hypothesis. It does an incomplete search on complete hypothesis space. This algorithm selects a feature as root at that level which has maximum information gain (minimum entropy). Refer to lecture slides and class discussion for more details.

Random Forest :- This is an ensemble method for classification and regression task, it work by constructing many decision trees (say *n*) at training time and outputting the class during testing on the basis of majority vote. For making a tree (also called a split) a random sample of *m* features in drawn from the set of all features and only those *m* features are considered for splitting.
Generally $m = \sqrt{p}$ or $\log_2 p$ where p is the total number of features. For more details visit this link. For classifying an instance majority or mode of all the outputs given by all decision trees is taken.

Boosting Methods :- These are methods to convert set of weak learners(like decision trees) to strong ones. There are many ways to do Boosting like AdaBoosting(Adaptive Boosting), xgboost etc.

Assignment: - Download the classification dataset from here. You have to implement all the three algorithms

I. Decision tree
II. Random Forest using decision trees and
III. Choose and implement any of the boosting techniques like Adaboosting or xgboost etc.

Apply all the three algorithms on the same dataset and compare the results like accuracy, training time, etc. For each algorithm clearly mention all the parameters that you have chosen like no. of trees in random forest and so on.
You are allowed to code only in C, C++ or Java and not allowed to use any packages. Proper documentation of the code is required. All the team members are expected to contribute equally and all should have complete understanding of the code and algorithms.

Report:

- Team Members.
- Mention the pre-processing applied on dataset.
- Compare all the three algorithms on the basis of Accuracy , learning time, and other factors that you find interesting for comparison.

Evaluation:

- Final results
- Understanding of results.
- Ability to reason the derived results.
- Final report and demo.

Submission should be through **CMS** only.

Contact the following Teaching Assistants for any clarification on this assignment.
Rajitha <2015409@hyderabad.bits-pilani.ac.in >
DEVENDRA SINGH SHEKHAWAT <f2013204@hyderabad.bits-pilani.ac.in>