MACHINE LEARNING ASSIGNMENT-1

BITS F464

BITS-PILANI Hyderabad Campus

Team Members:

NAME	ID Number
G V Sandeep	2014A7PS106H
Snehal Wadhwani	2014A7PS430H
Tanmaya Dabral	2014A7PS138H
Kushagra Agarwal	2014AAPS334H

Pre-processing:

1. The data contains both numeric and string values. So the data was first read row by row and an ArrayList of DataSet objects was created. Each DataSet object contains the data as it is.
2. The rows containing missing values were ignored as they constitute only 7% of the total data.
3. Since there are some continuous attributes in the data they were split into two categories based on split which gives minimum entropy.
4. We then iterate through the ArrayList of DataSet objects and populate a **matrix** with the corresponding integer value. We get the corresponding numeric value from a class called DataRef where every attribute column is given a number and the possible values it can take are associated with an integer.
    a. For Eg : attribute[1][0] = “Private”
        majorRef[1] = “workClass” means that 1st column of the matrix stores values corresponding to the attribute workClass and “Private” in work Class corresponds to 0.

Note: All those values which are less than the calculated split are given the value 0 and those greater than the split are given the value 1.

Comparing Results:

	ID3	Random Forest	AdaBoost
Running Time	9 seconds	250 seconds	150 seconds
Accuracy	81.32%	83.45%	84%

Random Forest

1. No of trees chosen : 300
2. No of Attributes chosen : 4

AdaBoost

1. No of trees chosen : 5000
2. No of rows chosen for each tree : 10

Some Observations:

- The accuracy of the simple ID3 takes a dip from 81.32% to around 80% at half the training data whereas, methods like Random Forest and AdaBoost still maintains more or less the same accuracy. This shows that these techniques make the decision tree more robust.
- Even at $1/4^{th}$ of the training data, Random Forest and AdaBoost give an accuracy of around 81% whereas the simple ID3 gives an accuracy of only77%.
- Random Forest and AdaBoost do not give the exact same accuracy every time. They tend to vary within a small limit. When we increase the number of trees in each case the limit becomes smaller and smaller.
- Only 10 examples are considered for AdaBoost at one time because, AdaBoost works on weak - classifiers.
- When we increase the number of rows (to 1000) we see that although the accuracy still remains more than the simple ID3, it takes a dip to around 82.5%.