Machine Learning Assignment-2

Mudit Pandey 2014A7PS017H
Vedic Sharma 2014B4A7424H
Sagar Gupta 2014A7PS0030H

October 2, 2017

Contents

1 Languages used

The entire code was written in C++.

2 Pre-processing

We process the data as follows:

1. All missing values for an attribute are replaced by the most freequent value of the attribute. In case of missing values for continous values, we replace the missing value with the average.
2. We handle continous attributes by dividing them into two sets based on a threshold which maximizes information gain.

3 Compilation and Execution

Run the following commands in the terminal:-

1. g++ -std=c++11 main.cpp
2. ./a.out

4 Result

ID3 Algorithm

1. Training time: 0.826s
2. Accuracy: 0.8306
3. Precision: 0.668
4. Recall: 0.560
5. F-measure: 0.609

ID3 followed by Reduced Error Pruning

Note: We use 30% of the training data as the validation set.

1. Before Pruning

(a) Accuracy: 0.826

(b) Precision: 0.659

(c) Recall: 0.550

(d) F-measure: 0.600

2. After Pruning:

(a) Training time: 10.087 s

(b) Accuracy: 0.845

(c) Precision: 0.720

(d) Recall: 0.563

(e) F-measure: 0.632

Random Forest

Note: We have built a random forest consisting of 100 trees

1. Training time: 82.301 s
2. Accuracy: 0.844
3. Precision: 0.7129
4. Recall: 0.5778
5. F-measure: 0.6383