Machine Learning (BITS F464) - Assignment 2
Decision Tree Learning

Maximum Marks: 40
Submission Deadline: 2359Hrs 30/09/2017

The assignment has three tasks. First, you must implement the ID3 algorithm which is described in your textbook (Machine Learning by Tom Mitchell). Your code must be able to handle continuous attributes as detailed in section 3.7.2 of the textbook. You must also be able to handle missing values using any of the approaches given in section 3.7.4.

It is an established fact that decision trees learnt using ID3 tend to overfit to the training data. One of the ways to overcome overfitting is Reduced Error Pruning. This is detailed in section 3.7.1.1 of the textbook. For the second task, you must implement Reduced Error Pruning for the decision tree. For this task, you will require a validation set. This set can be obtained by splitting the training data into the actual training set and the validation set.

Though pruning can improve accuracy, one of the better ways to avoid overfitting is to construct Random Forests. A Random Forest is a bunch of decision trees, each learnt by making use of the dataset (containing N data points) that is randomly sampled from the training dataset (containing N data points). Note that in the sampled dataset, a data point may be selected more than once from the training data set or the data point may not be selected.

After obtaining the sampled data set as explained above, at each decision node, a subset of the attributes which haven't been used at previous levels is chosen. The attribute among the sampled attributes which has the highest information gain is chosen. The number of attributes sampled is either $\sqrt{p}$ or $\log p$ where p is the total number of remaining attributes.

The final output of a Random forest is the mode of the outputs of the individual trees. For the final task, you must implement Random Forests. For more details refer to the following book: "Introduction to Data Mining" by Michael Steinbach, Pang-Ning Tan, and Vipin Kumar. This book is available in reference section of our library.

Train and test each of the classifiers on the same dataset. Compare their performance in terms of accuracy, precision, recall, F-measure and training time. Interpret the results and give examples of scenarios in which a given algorithm must be used. The results and their analysis must be compiled and submitted in a separate document along with the code. Also mention the values of the various hyper-parameters you've chosen. Your code should be generic and must be able to handle any dataset.

Dataset: You will be using the UCI Census Income dataset to evaluate the learning algorithms. It can be found at https://archive.ics.uci.edu/ml/machine-learning-databases/adult/

A sample data point is:
39, State-gov, 77516, Bachelors, 13, Never-married, Adm-clerical, Not-in-family, White, Male, 2174, 0, 40, United-States, <=50K

There are 14 attributes for each data point and the learning algorithms must predict the final target attribute which indicates the income of a person (It takes one of two values - <=50K or >50K). Some

of the attributes can be missing for a data point. A ? value for an attribute indicates a missing value. You are welcome to handle these missing attributes in any way – you can either estimate the values of these attributes based on statistics as detailed in your textbook or simply ignore the missing attributes.

The adult.data file should be used for training and the adult.test file should be used while testing.

Additional information regarding the various attributes can be found at - https://archive.ics.uci.edu/ml/datasets/Census+Income

Languages allowed:- C, C++, Java. It goes without saying that you are not allowed to use any packages which implement the aforementioned algorithms or copy code.

Deliverables

1. Original code for each of the tasks. The code must be well documented and logically organised.
2. A document containing the results and their analysis.

All the deliverables must be zipped into a file and uploaded on CMS.

Contact the following Teaching Assistants for any clarification on this assignment.
Aniketh on f2014096@hyderabad.bits-pilani.ac.in
Rajitha p2015409@hyderabad.bits-pilani.ac.in