Decision Trees Report

GROUP: 0015
Members : Suchintan Pati (18CS10064) , D. Vamshidhar Reddy (18CS30017)

Procedure

Step 1

1.1 Preprocessing Dataset

The Breast Cancer dataset had 285 entries with 9 features and 1 target class (recurrence-event , no-recurrence-event). Some of the values in certain rows were missing (?) which were removed from the dataset.

The dataset was then split into training data, validation data, testing data in the ratio 60:20:20. 10 random splits were considered.

1.2 Building Tree

The decision tree was built using the ID3 algorithm. The ID3 function takes in the training data and the max depth limit and builds a decision tree. The decision tree was built recursively by selecting the feature upon which tree should be further grown. The feature was selected by computing Information Gain of all the features and selecting that feature which had the highest Information Gain. The base case for recursion is reached when one of the following occur

- The max depth limit is reached
- All the data points in that particular node have the same target class
- There are too few data points in that node to do a further grow the tree

At each node the number of recurrence-events and no-recurrence-events was also computed and stored in each node.

1.3 Predicting Accuracy for a Maximum Depth Limit

For each split (60:20:20 into training data, validation data, testing data), the training data is used to build a tree for the given Depth limit (to be provided by the user). The tree is then used to predict the target class on the testing data and the results are compared with the actual target class and the accuracy is computed. The average over the 10 splits is displayed to the user.

Step 2

2.1 Plotting Accuracy vs Depth

Since there are 9 features, the max depth of the tree can be at most 9. For each depth the average accuracy is calculated. The graph of Average Accuracy vs Max Depth is plotted using Matplotlib.

Step 3

3.1 Post Pruning

The depth for which we got the maximum Average Accuracy is considered for pruning. Post Pruning is done using the Reduced Error Pruning algorithm. A validation set is considered for this operation. Each node in the decision tree is temporarily removed and the accuracy on the validation set is found. The node which caused the greatest increase in validation accuracy is permanently removed from the decision tree. This process is repeated till there is no node in the tree which when removed causes an increase in validation set accuracy.

Results

Step 2

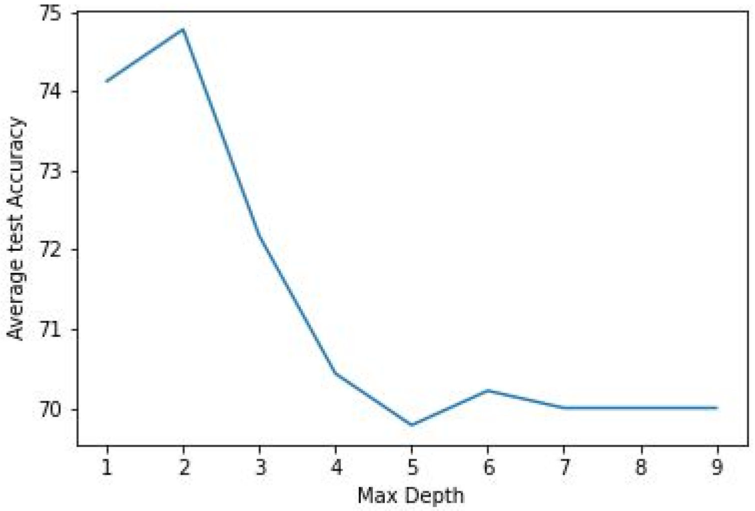

Fig 1 : Average Test Accuracy vs Max Depth

The best possible depth limit to be used is **2**.

Step 3

Depth of the tree chosen	Validation accuracy before pruning	Validation accuracy after pruning	Test accuracy before pruning	Test accuracy after pruning	Nodes pruned
2	72.0	78.0	73.91	69.56	3

Since we have obtained the maximum depth as 2, pruning the tree though increases validation accuracy, but reduces the test accuracy expectedly(underfitting). However, if we apply pruning on higher depths, it usually results in a better test accuracy.