COL774 Assignment 3A

Siddhant Mago, 2017CS50419

March 2020

1 Decision Tree Construction

The splitting attribute was selected on the basis of the mutual information gain metric. Mutual information gain between A and B is the amount of information we learn about A by knowing B. It is the change in entropy and is equal to:

$$IG(T; a) = H(T) - H(T|a)$$

where IG is the information gain in T by knowing attribute a.
H is the information entropy.

- The nodes were expanded in a way that attribute values > median go to the right subtree and attribute values <= median go to the left subtree. The median was considered for splitting because the attributes are integer valued.
- If at any node, the mutual information gain was 0 for all attributes, that node was not expanded further and gave a majority prediction based on the training entries that reached that node.

The accuracy plot as the tree was grown is:

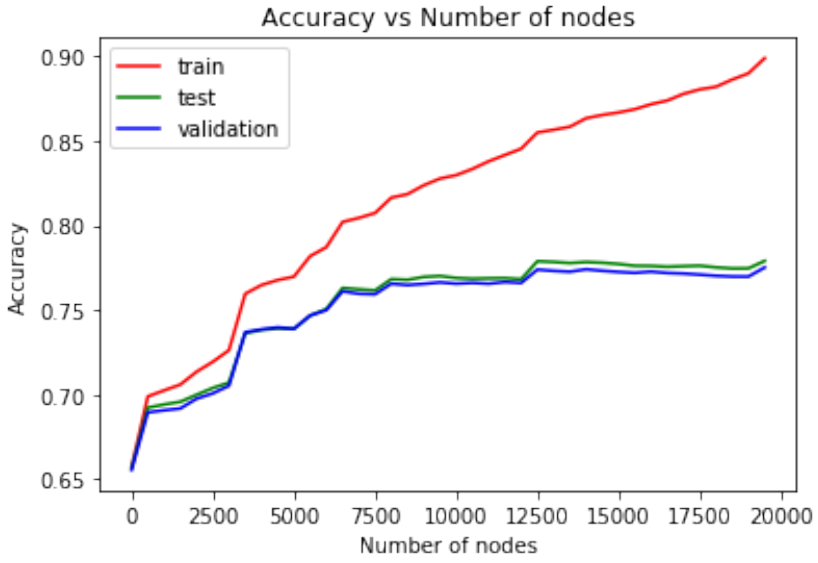


Figure 1: Accuracy as the tree grows

The final accuracy on fully grown tree was:

Training: 90.44%
Test: 77.98%
Validation: 77.62%

The number of nodes in the fully grown tree was approximately 19900. I observed that as the number of nodes in the tree increased the accuracy improves but the test and validation accuracy do not increase as much as the training accuracy. Towards the end of the growth of tree, it is visible that test and validation accuracy saturate and even decrease slightly, possibly due to over-fitting on the training data.

2 Decision tree post pruning:

Post pruning is performed on the fully grown tree to prevent over-fitting (which was observed in the last case) on the training data. The nodes are greedily removed based on the accuracy obtained on validation data. This removal of nodes is carried out until no further improvement in observed.
The plot as the tree was pruned is:

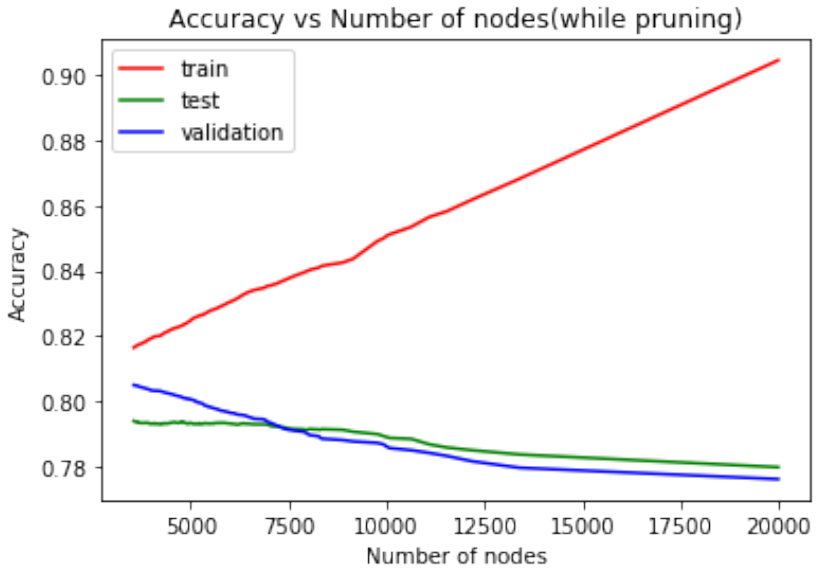


Figure 2: Accuracy as the tree is pruned

The final accuracy were:

Training: 81.65%
Test: 79.39%
Validation: 80.49%

I observe that as the tree is pruned (number of nodes decrease), the validation and test accuracy increase, which is expected, as we're removing nodes greedily, whereas the training accuracy reduces (as we're reducing overfitting). The final tree obtained has 3500 nodes compared to 19900 nodes in the initial tree. Thus, it can be concluded that pruning has been able to reduce over-fitting to some extent.

3 Random Forests:

I used sklearn's `RandomForestClassifier` to learn a random forest. I also implemented the grid search from scratch and used out of bag score to compare the different models to find the optimal hyper-parameters. The optimal hyper-parameters obtained were:

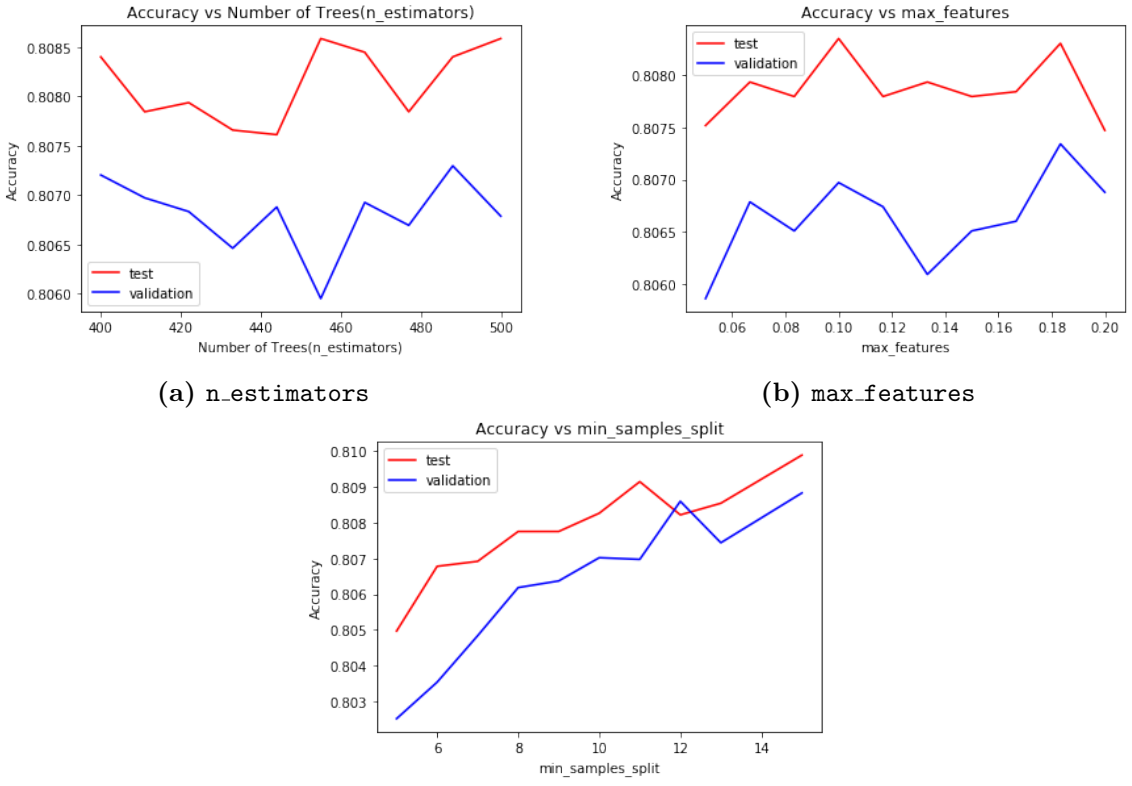
$$\text{n\_estimators} = 450$$
$$\text{max\_features} = 0.1$$
$$\text{min\_samples\_split} = 10$$

The accuracies for the optimal set of parameters were:

Dataset	Full grown tree	Pruned Tree	Random Forest
Train	90.44 %	81.65 %	87.31 %
Test	77.98 %	79.39 %	80.9 %
Validation	77.62 %	80.49 %	80.65 %

The out of bag score for the Random forest was observed to be **81.09 %**.
All the three test, train and validation accuracy increase in random forests. This is the expected behaviour as a random forest is stronger than a decision tree and thus leads to better results.

4 Parameter Sensitivity Analysis:



(a) `n_estimators` **(b)** `max_features`

(c) `min_samples_split`

Figure 3: Parameter sensitivity analysis

- For `n_estimators` and `max_features`, the accuracy variation is **random** and **small**, this is possibly because we are already at the optimal value and thus observe no clear trend of change.
- For `min_samples_split`, the accuracy is increasing, this means that as we increase `min_samples_split` the accuracy increases and the optimal value of `min_samples_split` lies somewhere beyond 10.