COL774: Machine Learning- Assignment 3
Decision Trees and Random Forests

Rajat Jaiswal (2017CS50415)

20th March 2020

Decision Trees

(a) **Decision Tree Construction:** In this part, ID3 algorithm was used for constructing the decision tree. Data was splitted on the basis on median attribute value. The best attribute was selected by computing the information gain. At any step, the attribute which results in highest mutual information by splitting on its median value is chosen.

By doing so the fully grown tree which was obtained had **10005 decision nodes** and a **height of 52**.
The accuracy obtained over **Training set** was **90.458%**.
The accuracy obtained over **Validation set** was **77.327%**.
The accuracy obtained over **Test set** was **77.794%**.

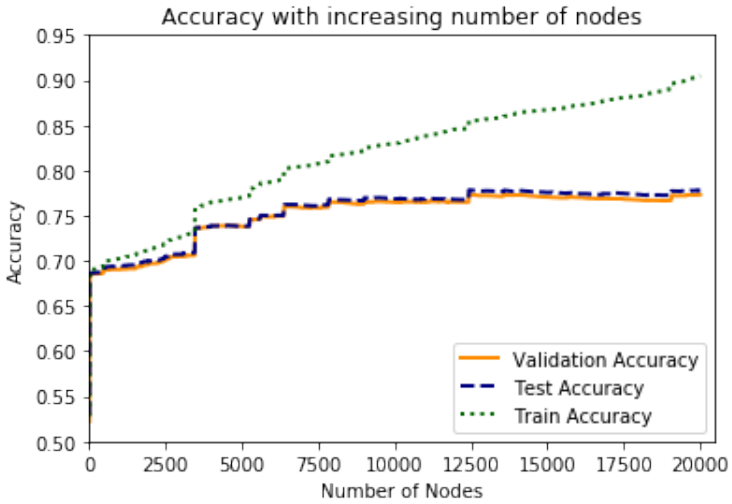

Fig 1: Accuracy v/s Number of total nodes as the tree grows.

The accuracy increases gradually as we go on adding nodes to the tree using our decision tree learning algorithm. There is a rise at around 12500 nodes as now the tree also has a right branch contrasting to before where it was only getting expanded in the left subtree of the root. The increase at around 3000 nodes tells us that there is a subtree which is getting a lot of data to classify and hence is very crucial. The attribute used at this decision node is crucial for this subset of data.
The difference between accuracy over training set and validation or test set increases as number of nodes increase, possibly because the tree is already getting sufficiently dense and further increase leads to some overfitting.

(b) **Decision Tree Post Pruning:** In this part, the decision tree was fully grown and then was post-pruned based on a validation set. In post-pruning, I greedily pruned the nodes of the tree (and sub-tree below them) by iteratively picking a node to prune so that resultant tree gives maximum increase in accuracy on the validation set. In other words, among all the nodes in the tree, I pruned the node such that pruning it(and sub-tree below it) results in maximum increase in accuracy over the validation set. This was repeated until any further pruning leads to decrease in accuracy over the validation set.

By doing so the pruned-tree which was obtained had **2579 decision nodes** and a **height of 47**.
The accuracy obtained over **Training set** was **82.378%**.
The accuracy obtained over **Validation set** was **80.702%**.
The accuracy obtained over **Test set** was **79.408%**.
The accuracy over **Validation set increased by 3.375%**, and the accuracy over **Test set increased by 1.614%** compared to accuracies in part(a).

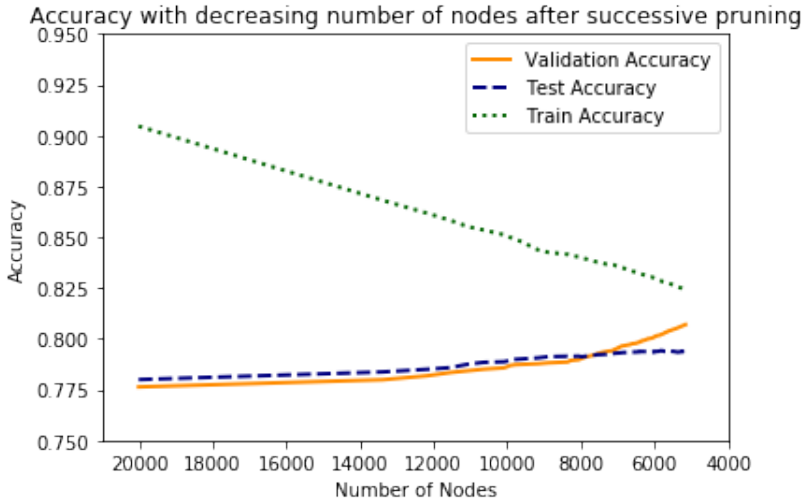

Fig 2: Accuracy v/s Number of total nodes as the tree is pruned successively.

As I go on post pruning the tree increasing the accuracy on validation tree with each node that I prune, the accuracy for test set increases as well while that of training set decreases from a previously excellent value. This helps in reducing the overfitting and even though it performs somewhat poorly on the training set, it will perform better on new test sets. Thus, it is a better model than obtained in part (a).

Random Forests

(c) **Random Forests - Grid Search:** In this part, I used the scikit-learn library of Python to grow a Random Forest. I grew different forests by playing around with various parameter values. I varied *n_estimators* from 50 to 450 in steps of 100, *max_features* from 0.1 to 1.0 in steps of 0.2, and *min_samples_split* from 2 to 10 in steps of 2. I performed a grid search over this space of parameters and an out-of-bag accuracy was used to tune to the optimal values for these parameters.

By doing so the optimal set of parameters obtained were ***n_estimators*: 350**, ***max_features*: 0.30**, and ***min_samples_split*: 10**.
The accuracy obtained over **Training set** in the optimal model was **88.109%**.
The **out-of-bag accuracy** obtained in the optimal model was **80.906%**.
The accuracy obtained over **Validation set**in the optimal model was **80.767%**.
The accuracy obtained over **Test set** in the optimal model was **80.817%**.

The test set accuracy is about 1.4% higher than in (b) but the validation set accuracy is almost similar. However, the training set accuracy is much greater than in (b) and this even performs better on the test set so this is indeed a better model.

(d) **Random Forests - Parameter Sensitivity Analysis:** In this part, I varied one of the parameters (in a range) while fixing others to their optimum.

Observations:

- Test set accuracy is consistently higher than Validation set accuracy for all the models.
- The accuracy is most sensitive for *min_samples_split*, then *max_features* and least sensitive to *n_estimators*. This is because each splitting and features help reduce overfitting and *n_estimators* just increases the number of trees in the forest.
- The accuracy is not very sensitive to *n_estimators* for either test set or validation set. However, an although small but consistent increase in training set accuracy was observed with the increase in *n_estimators*.

- Accuracy decreases with increasing fraction of features to be considered due to overfitting. At *max_features* = 0.10, accuracy is less than at the optimal value due to underfitting. However, more is the fraction, better the training set accuracy we get.
- Accuracy increases with increasing *min_samples_split* to be considered, due to reduction in overfitting. However, more is the fraction worse the training accuracy we get. It is the most sensitive parameter.

Number of Estimators	Validation Set Accuracy(in %)	Test Set Accuracy(in %)
50	80.558	80.608
150	80.609	80.766
250	80.512	80.719
350	80.767	80.817
450	80.591	80.752

Table 1: Validation and Test set accuracy with varying number of estimators

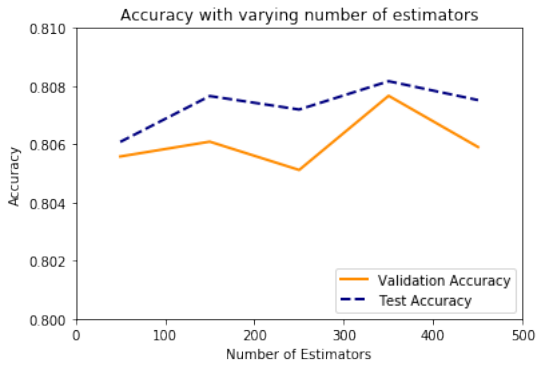

Fig 3: Validation and Test set accuracy with varying number of estimators

Maximum % of features	Validation Set Accuracy(in %)	Test Set Accuracy(in %)
0.10	80.665	80.752
0.30	80.767	80.817
0.50	80.591	80.673
0.70	80.609	80.673
0.90	80.489	80.483

Table 2: Validation and Test set accuracy with varying maximum percentage of features

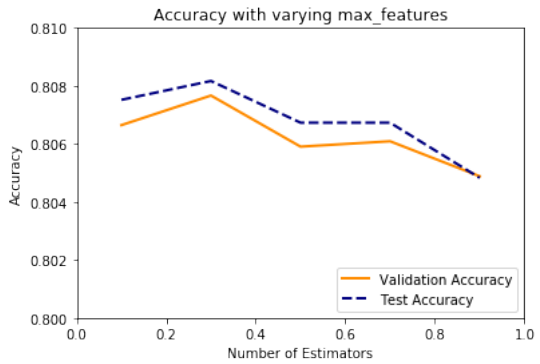

Fig 4: Validation and Test set accuracy with varying maximum percentage of features

Minimum No. of Samples split	Validation Set Accuracy(in %)	Test Set Accuracy(in %)
2	79.835	79.917
4	80.136	80.312
6	80.317	80.534
8	80.465	80.738
10	80.767	80.817

Table 3: Validation and Test set accuracy with varying minimum number of samples split

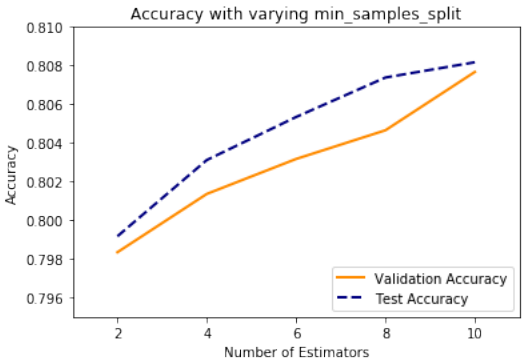

Fig 5: Validation and Test set accuracy with varying minimum number of samples split