Machine Leaning Assignment-3

Lalith Satya Srinivas Kaladi
2017CS10340

March 2020

1 Question A

Training data is used to build the decision tree. The fully grown decision tree contains around 43,000 nodes. The accuracy on training data for fully grown tree is 90.8. The figure below shows the variation of training, validation and test set accuracies with number of nodes in the decision tree.

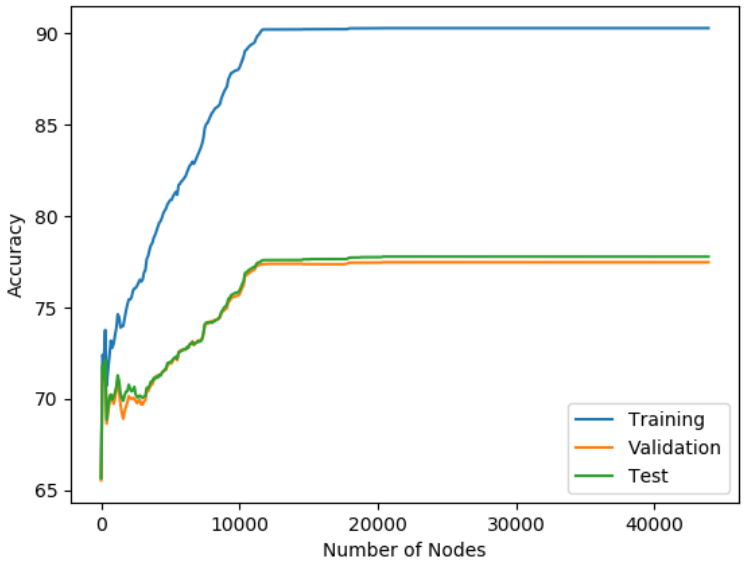

Figure 1: Accuracies obtained on Test, Validation, Training datasets Vs. Number Of Nodes

As it can be observed from the plot, accuracy increases on expanding the tree and it almost becomes constant after a certain number of nodes.

2 Question B

The fully grown decision tree is now pruned with respect to validation dataset. The number of nodes in fully grown tree is now reduced to 3600 and the accuracy on validation dataset now became 79.8 percent.

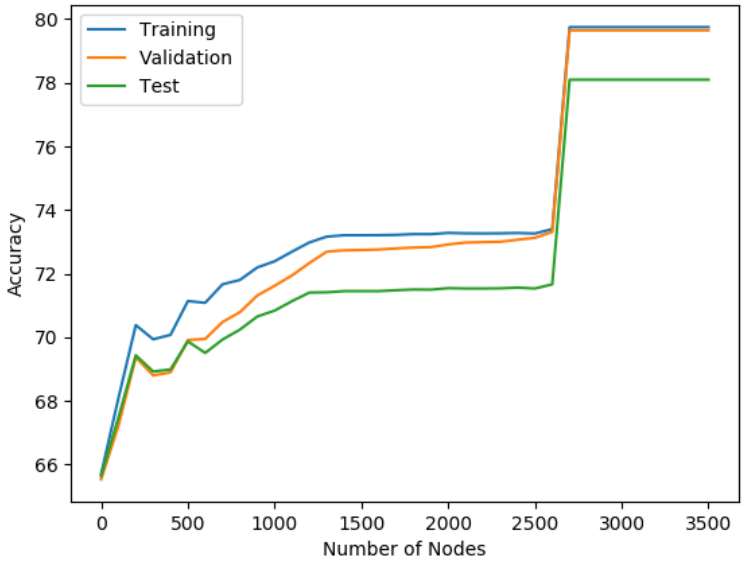

Figure 2: Accuracies obtained on Test, Validation, Training datasets Vs. Number Of Nodes after Pruning

Again, the accuracies are increasing continuously. But the training accuracy has reduced to around 80 percent from 90. But the validation dataseta cuuracy has increased. This can be explained as the tree is pruned with respect to validation dataset, which means that the nodes that are removed had reduced accuracy which are removed.

3 Question D

Each parameter is changed keeping rest as same and test, validation accuracies were observed (in the plot position refers to place that parameter holds in the below order). The order of parameter values are

$$\text{n-estimator} = \{50, 150, 250, 350, 450\}$$
$$\text{max-features} = \{0.1, 0.3, 0.5, 0.7, 0.9\}$$
$$\text{min-samples-split} = \{2, 4, 6, 8, 10\}$$

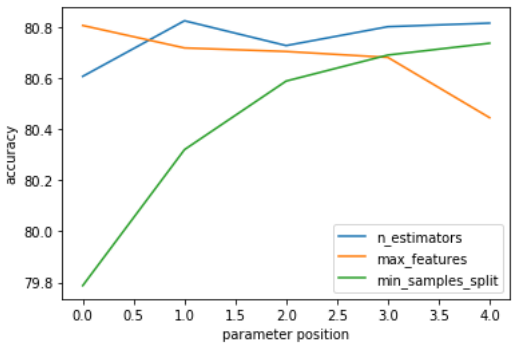

Figure 3: Accuracies obtained on Test datasets Vs. Parameter

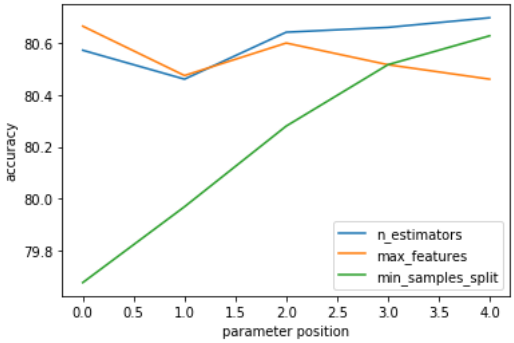

Figure 4: Accuracies obtained on Validation datasets Vs. Parameter

Both test, validation accuracies vary similarly when the parameters are changed. Accuracy increases continuously when the parameter min-sample-split is increased. The increase is almost linear in both the cases. Similarly, with increase in max-features the accuracy is decreasing but there is no co-relatio between n-estimators and accuracy as it can be observed that there is both increase and decrease in accuracy as n-estimators increases.

4 Question C

The optimal parameters that are obtained on using grid search, random forest classifier from sklearn library are n-estimators = 250, max-features = 0.1, min-samples-split=10. Following accuracies are observed.

Out-of-bag = 80.98
Test = 80.78
Training = 87.60
Validation = 80.75

The test, validation accuracies have increased when compared to that obtained in above section, where as training dataset accuracy has decreased. This miight indicate that there is less overfitting whe compared to the above part.