CS534 – Machine Learning

Homework Assignment 3*

Fall 2018

Ali Raza 933620601 razaa@oregonstate.edu
Sheekha Jariwala 933620505 jariwalas@oregonstat.edu
Vivswan Shitole 933591615 shitolev@oregonstate.edu

Oregon State University
November 29, 2018

* Due 11:59PM Nov 30th, 2018

We worked as a team of three students, and contribution of each member is 1/3. We coded our implementation in Python 3.6. The effects of tree depth, number of trees in the Random Forest and Adaboost are discussed in the following parts.

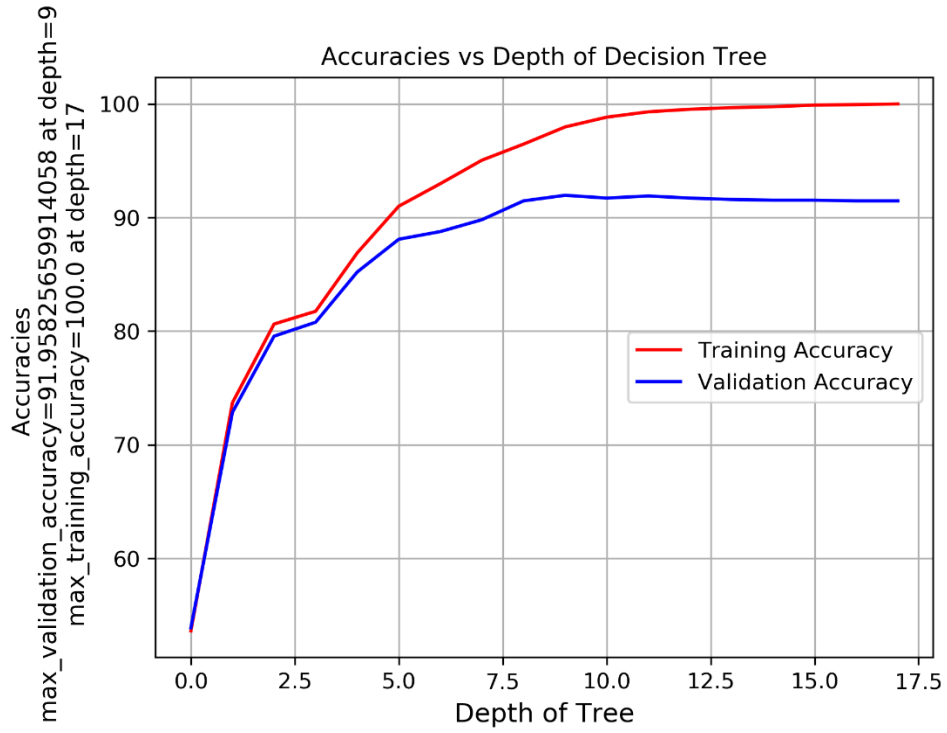
Part 1

a)

Decision tree with maximum depth of 20 has been created with the source code attached with the assignment.

b)

Required plot of training accuracy and validation accuracy vs depth is given below:



c)

Behavior of train/validation performance against depth: As seen from the graph, training and validation accuracy of the decision tree increases with increase in depth of the decision

tree. However, as training accuracy starts reaching 100%, overfitting happens and validation accuracy starts decreasing.

d)

Depth that gives best validation accuracy:
The validation accuracy reaches its maximum at DT depth = 9, after which it starts decreasing due to overfitting.
Best validation accuracy – 91.95%
Depth – 9

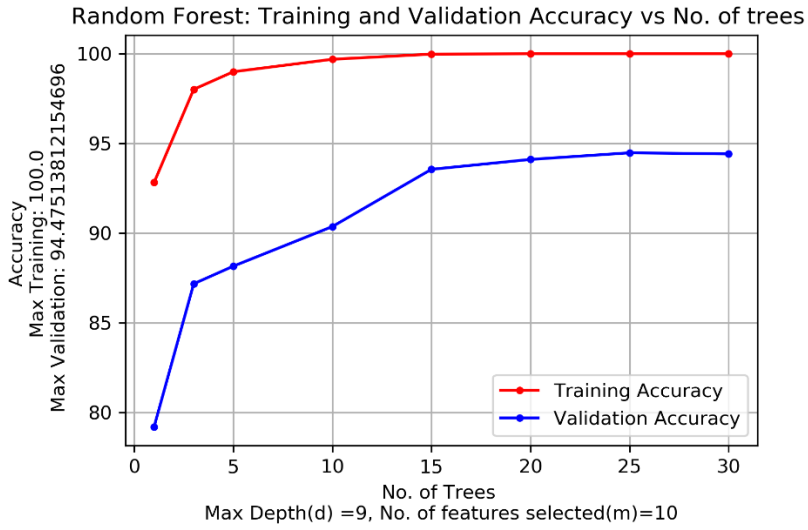
Part 2

a)

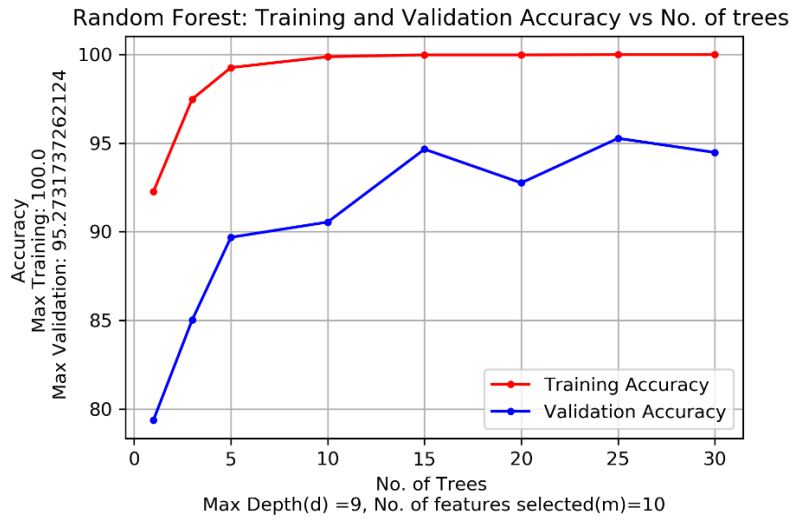
Random Forest is implemented in Python 3.6. Source Code is included with the assignment.

b)

Due to the stochastic nature of the Random Forest, result of two runs are given below:



Run 1



Run=2

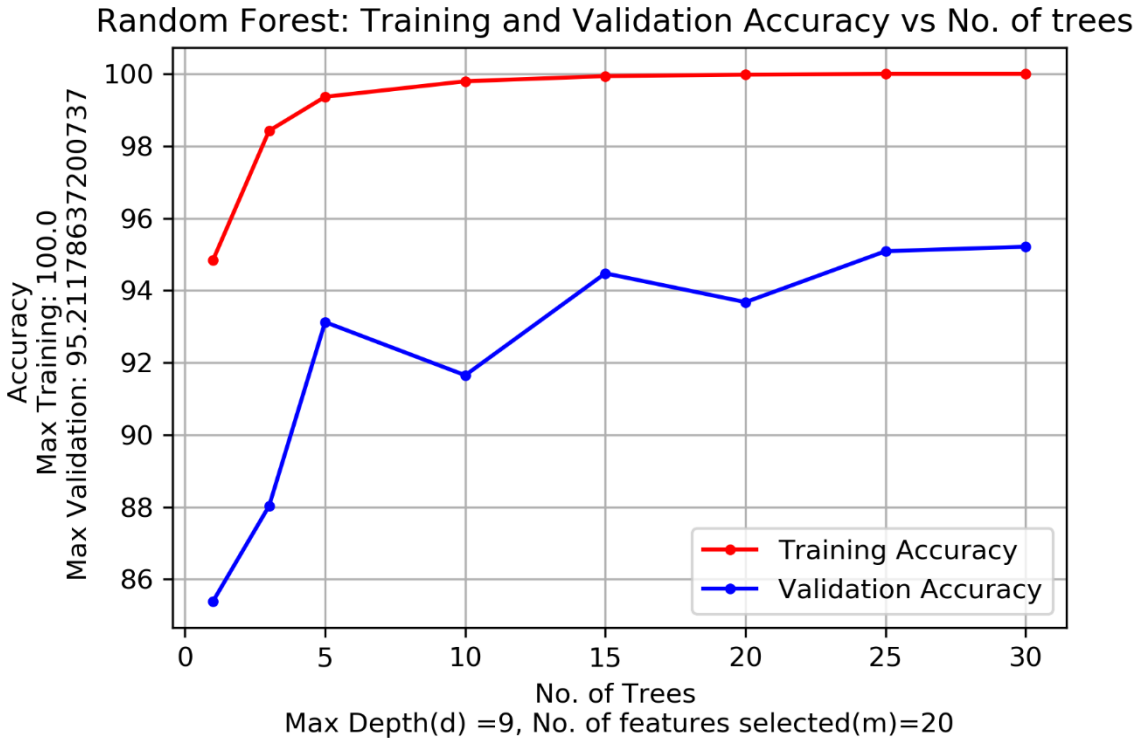
c)

By adding more trees into the forest, training accuracy is always increasing, whereas, validation accuracy is generally increasing. It is important to note that after training accuracy reaches 100%, validation accuracy continues to increase. It means that Random Forest is robust to overfitting (not always).

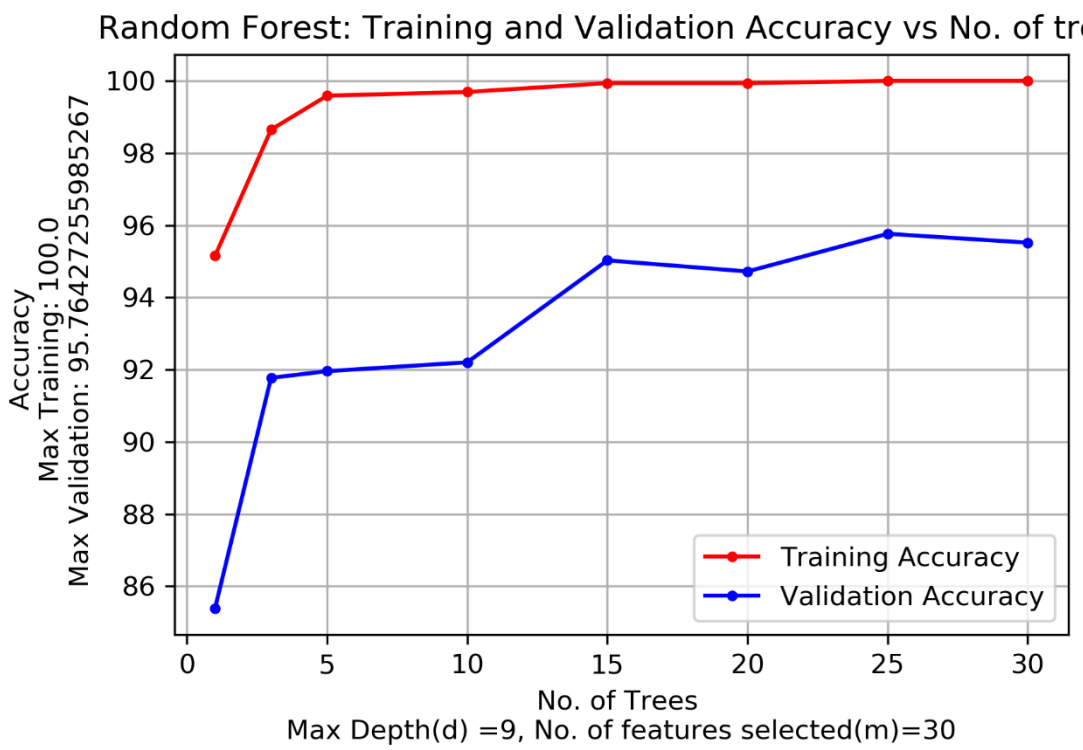
d)

Required graphs are shown below:

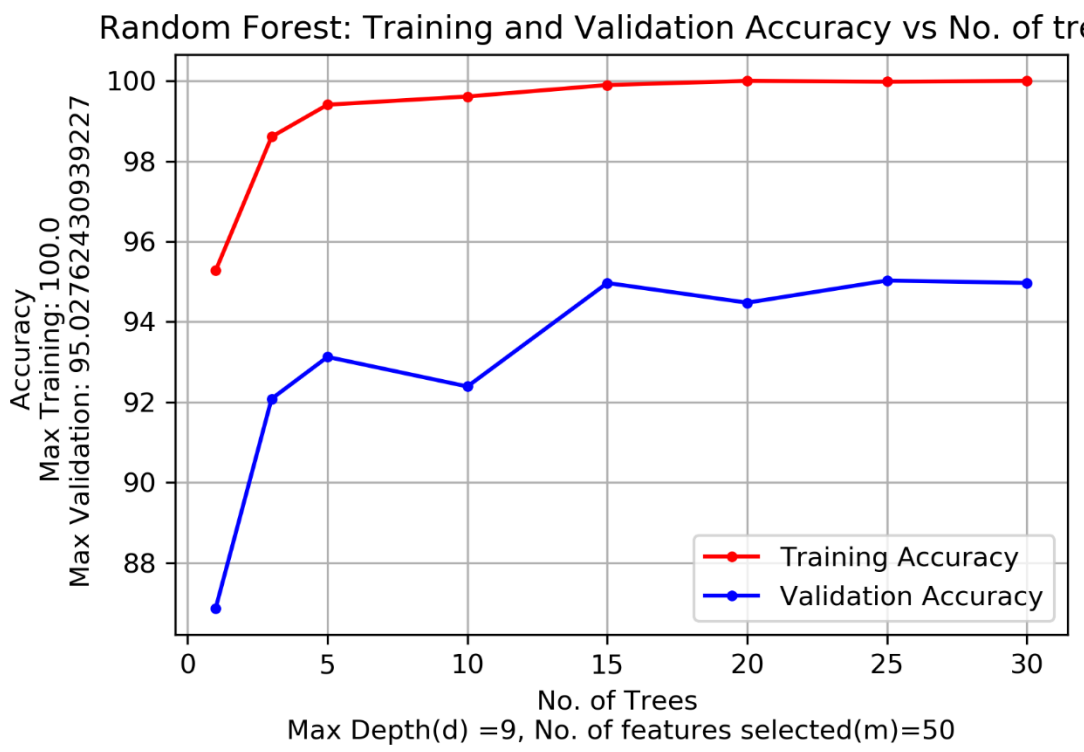
By increasing the value of m (for given range), there is no noticeable change in the behavior of training accuracy and validation accuracy. But as m starts approaching total number of features, the training validation accuracy will start dropping as advantage of sampling features diminishes.



m = 20



$\mathrm{m} = 30$



$\mathrm{m} = 50$

Part 3

a)

We have implemented Weak learner as a decision tree with given requirements.

b)

We have implemented Adaboost algorithm in Python 3.6. Source code is included with the assignment.

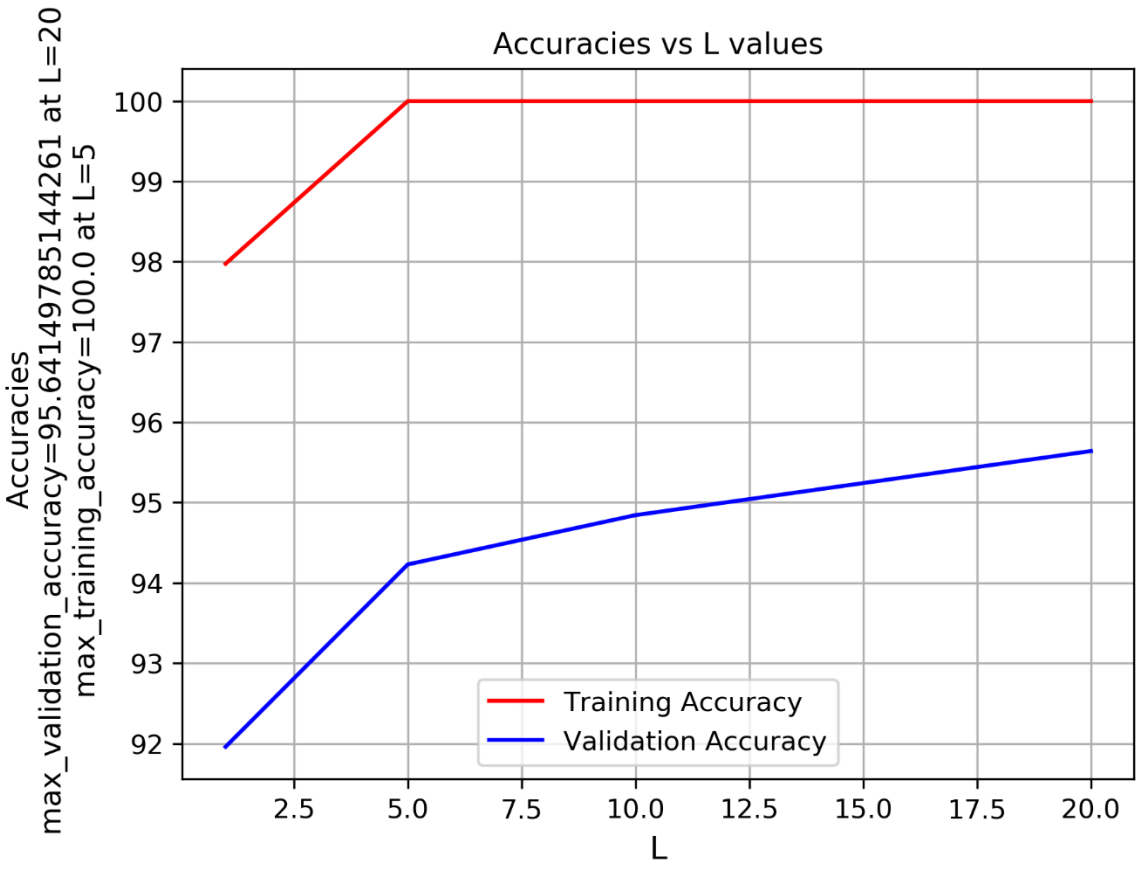
c)

Training and Validation accuracy for different values of L are given below:

L	Training Accuracy	Validation Accuracy
1	97.97	91.96
5	100	94.23
10	100	94.84
10	100	95.64

d)

As L increases, we are calculating predictions over larger number of decision trees. Thus, the training and validation accuracies increase with L. There is no drop in validation accuracy even after the training accuracy reaches 100 since there is no over fitting given the large number of decision trees who collectively yield the prediction. Therefore, Adaboost is robust to overfitting.



Random Forest is implemented in Python 3.6. Source Code is included with the assignment.

Conclusion

We generated a Decision Tree model and explored Bootstrap Aggregating and Boosting. In addition, we experimented with different depths of the Decision Tree, number of trees in the Random Forest and number of different learners in Adaboosting. Source file and report are included with the assignment