CS573 HOMEWORK 4

Mohit Gupta

March 31, 2019

1 Preprocessing

The code is given in preprocess_assg4.py file.

2 Implement Decision Trees, Bagging and Random Forests

The code is in trees.py. Example usage is given below:
python trees.py trainingSet.csv testSet.csv 1

The scores which I get for the 3 models are given below:
Decision Tree
Training Accuracy DT: 0.78
Testing Accuracy DT: 0.7

Bagging
Training Accuracy BT: 0.8
Testing Accuracy BT: 0.75

Random Forests
Training Accuracy RF: 0.78
Testing Accuracy RF: 0.75

3 The Influence of Tree Depth on Classifier Performance

3.1 Learning Curves

We can see from the learning curves in figure 1 that with the present hyperparameters as given in the homework problem statement Random Forests and Bagging outperform Decision Tree. The accuracy for bagging and random forests increase with increase in depth limit which contrasts with the decrease in test

accuracy observed for decision trees.
The reason for that could be decision trees overfitting on the training dataset when the depth limit is increased from 3 to 9.

3.2 Hypothesis Testing

We run Paired t-test for Bagging and Decision Tree models.
The test results are given below in Figure 2 for each depth i.e. 3, 5, 7, 9.

4 Compare Performance of Different Models

4.1 Learning Curves

We can see from the learning curves in figure 3 that with the present hyper-parameters as given in the homework problem statement Random Forests and Bagging outperform Decision Tree.

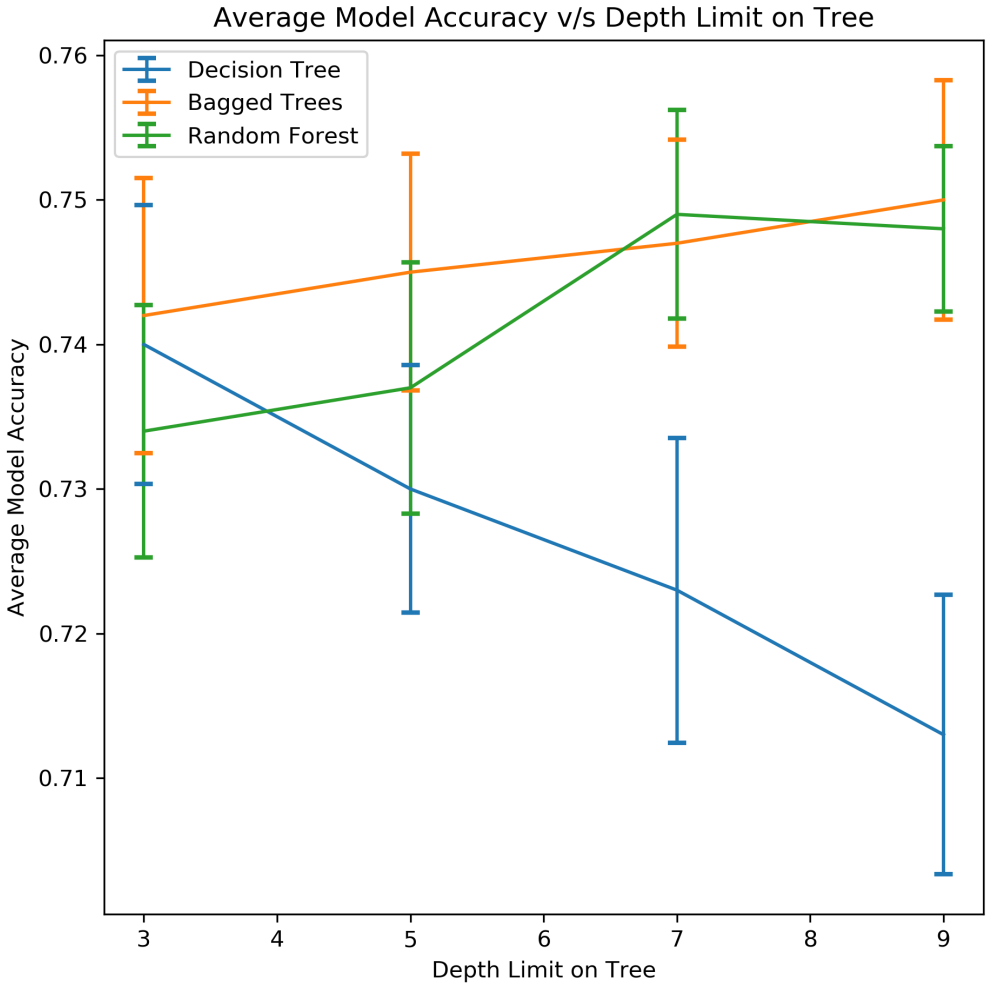

Figure 1: Learning Curve Average Test Accuracy v/s Tree Depth Limit

```
For Depth 3 run Paired t-test for Decision Tree and Bagging models

Null Hypothesis H0: Decision Tree Average Accuracy = Bagging Average Accuracy
Alternate Hypothesis H1: Decision Tree Average Accuracy != Bagging Average Accuracy
[0.73, 0.77, 0.73, 0.69, 0.79, 0.75, 0.75, 0.76, 0.73, 0.7]
[0.73, 0.78, 0.73, 0.7, 0.79, 0.75, 0.75, 0.76, 0.73, 0.7]
Paired T-test Statistics are: t_statistic= 1.5 pvalue= 0.16785065605707486

Accepting Null Hypothesis H0 since pvalue is greater than 0.05

For Depth 5 run Paired t-test for Decision Tree and Bagging models

Null Hypothesis H0: Decision Tree Average Accuracy = Bagging Average Accuracy
Alternate Hypothesis H1: Decision Tree Average Accuracy != Bagging Average Accuracy
[0.7, 0.75, 0.72, 0.69, 0.75, 0.72, 0.72, 0.78, 0.75, 0.72]
[0.74, 0.78, 0.75, 0.71, 0.77, 0.73, 0.75, 0.77, 0.75, 0.7]
Paired T-test Statistics are: t_statistic= 2.4227185592617446 pvalue= 0.0384388708080151

Rejecting Null Hypothesis H0 since the pvalue is less than 0.05

For Depth 7 run Paired t-test for Decision Tree and Bagging models

Null Hypothesis H0: Decision Tree Average Accuracy = Bagging Average Accuracy
Alternate Hypothesis H1: Decision Tree Average Accuracy != Bagging Average Accuracy
[0.69, 0.75, 0.71, 0.69, 0.75, 0.71, 0.71, 0.79, 0.74, 0.69]
[0.74, 0.78, 0.75, 0.73, 0.77, 0.73, 0.73, 0.77, 0.76, 0.71]
Paired T-test Statistics are: t_statistic= 4.0 pvalue= 0.0031104283103858535

Rejecting Null Hypothesis H0 since the pvalue is less than 0.05

For Depth 9 run Paired t-test for Decision Tree and Bagging models

Null Hypothesis H0: Decision Tree Average Accuracy = Bagging Average Accuracy
Alternate Hypothesis H1: Decision Tree Average Accuracy != Bagging Average Accuracy
[0.69, 0.72, 0.7, 0.69, 0.7, 0.7, 0.72, 0.79, 0.73, 0.69]
[0.75, 0.77, 0.74, 0.73, 0.78, 0.74, 0.74, 0.79, 0.76, 0.7]
Paired T-test Statistics are: t_statistic= 4.9591142728573505 pvalue= 0.0007814718410754046

Rejecting Null Hypothesis H0 since the pvalue is less than 0.05
```

Figure 2: Hypothesis Test Results

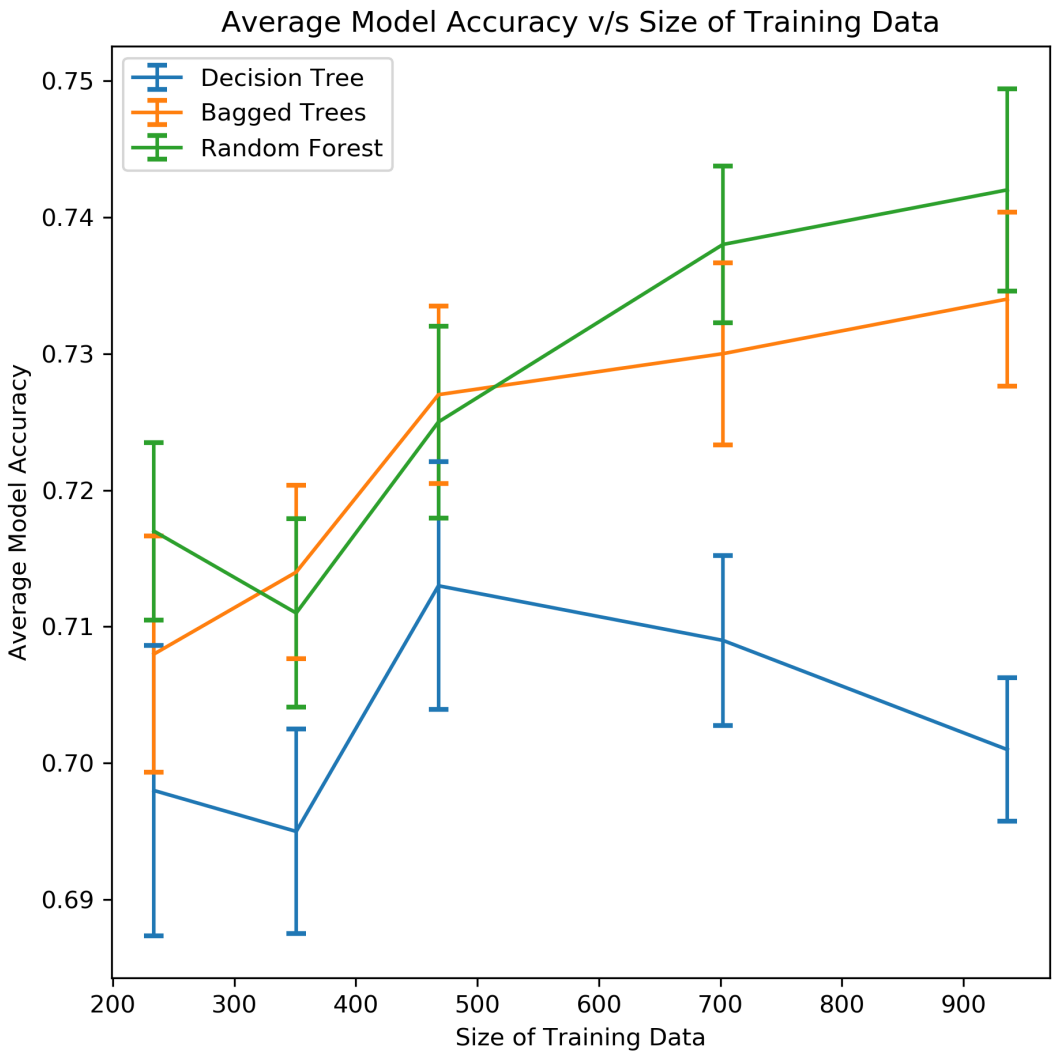

Figure 3: Learning Curve Average Test Accuracy v/s Size of Training Data

4.2 Hypothesis Testing

We run Paired t-test for Bagging and Decision Tree models.
The test results are given below in Figure 4 for each t_frac i.e. 0.05,0.075,0.1,0.15,0.2.

5 The Influence of Number of Trees on Classifier Performance

5.1 Learning Curves

We can see from the learning curves in figure 5 that with the present hyper-parameters as given in the homework problem statement Random Forests and Bagging perform similarly with Random Forests performing slightly better than Bagging with increase in number of trees.

```
For t_frac 0.05 run Paired t-test for Decision Tree and Bagging models

Null Hypothesis H0: Decision Tree Average Accuracy = Bagging Average Accuracy
Alternate Hypothesis H1: Decision Tree Average Accuracy != Bagging Average Accuracy
[0.66, 0.7, 0.66, 0.71, 0.69, 0.7, 0.73, 0.73, 0.75, 0.65]
[0.67, 0.71, 0.67, 0.72, 0.75, 0.71, 0.72, 0.71, 0.74, 0.68]
Paired T-test Statistics are: t_statistic= 1.3987572123604706 pvalue= 0.19538901886445117

Accepting Null Hypothesis H0 since pvalue is greater than 0.05

For t_frac 0.075 run Paired t-test for Decision Tree and Bagging models

Null Hypothesis H0: Decision Tree Average Accuracy = Bagging Average Accuracy
Alternate Hypothesis H1: Decision Tree Average Accuracy != Bagging Average Accuracy
[0.64, 0.72, 0.7, 0.68, 0.7, 0.71, 0.68, 0.7, 0.72, 0.7]
[0.68, 0.73, 0.71, 0.7, 0.74, 0.72, 0.71, 0.72, 0.74, 0.69]
Paired T-test Statistics are: t_statistic= 3.942772444036624 pvalue= 0.0033916702267946077

Rejecting Null Hypothesis H0 since the pvalue is less than 0.05

For t_frac 0.1 run Paired t-test for Decision Tree and Bagging models

Null Hypothesis H0: Decision Tree Average Accuracy = Bagging Average Accuracy
Alternate Hypothesis H1: Decision Tree Average Accuracy != Bagging Average Accuracy
[0.69, 0.7, 0.67, 0.69, 0.73, 0.73, 0.72, 0.73, 0.77, 0.7]
[0.69, 0.74, 0.71, 0.73, 0.74, 0.73, 0.75, 0.73, 0.75, 0.7]
Paired T-test Statistics are: t_statistic= 2.0397003109502547 pvalue= 0.07179988272763554

Accepting Null Hypothesis H0 since pvalue is greater than 0.05

For t_frac 0.15 run Paired t-test for Decision Tree and Bagging models

Null Hypothesis H0: Decision Tree Average Accuracy = Bagging Average Accuracy
Alternate Hypothesis H1: Decision Tree Average Accuracy != Bagging Average Accuracy
[0.71, 0.69, 0.69, 0.71, 0.73, 0.69, 0.72, 0.7, 0.75, 0.7]
[0.72, 0.75, 0.75, 0.74, 0.74, 0.7, 0.72, 0.72, 0.76, 0.7]
Paired T-test Statistics are: t_statistic= 2.973153822552955 pvalue= 0.01562222768428982

Rejecting Null Hypothesis H0 since the pvalue is less than 0.05

For t_frac 0.2 run Paired t-test for Decision Tree and Bagging models

Null Hypothesis H0: Decision Tree Average Accuracy = Bagging Average Accuracy
Alternate Hypothesis H1: Decision Tree Average Accuracy != Bagging Average Accuracy
[0.69, 0.72, 0.7, 0.72, 0.71, 0.7, 0.69, 0.67, 0.72, 0.69]
[0.73, 0.75, 0.72, 0.76, 0.76, 0.72, 0.72, 0.73, 0.75, 0.7]
Paired T-test Statistics are: t_statistic= 6.98292159700013 pvalue= 6.445216659500229e-05

Rejecting Null Hypothesis H0 since the pvalue is less than 0.05
```

Figure 4: Hypothesis Test Results

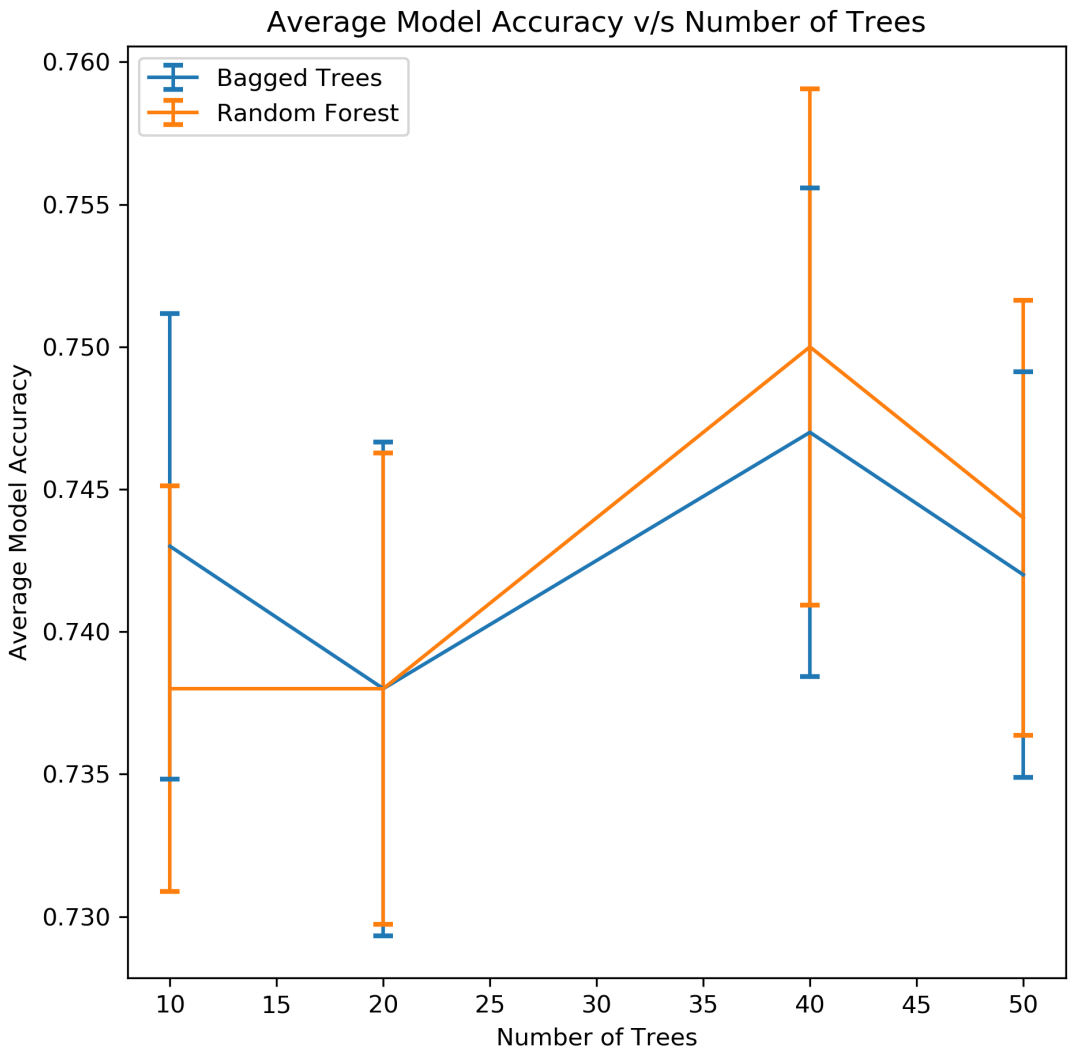

Figure 5: Learning Curve Average Test Accuracy v/s Number of Trees

5.2 Hypothesis Testing

We run Paired t-test for Bagging and Random Forest models.
The test results are given below in Figure 6 for each num_trees i.e. 10, 20, 40, 50.

6 Bonus Question

I have implemented Boosted Decision Trees using Adaboost algorithm. The code is present in bonus.py.

I have tuned the hyperparameters and the best results are:

Training Accuracy Boosted Decision Tree: 0.8

Testing Accuracy Boosted Decision Tree: 0.77

I tuned my model by varying the depth limit starting from 1 i.e. boosted decision stumps and then increased it till 4. Also, I varied the number of trees from 30 to 70. The final tuned parameters are depth limit of 3 and the number of trees are 50. The accuracy which I get on testSet.csv with these tuned parameters is 0.77.

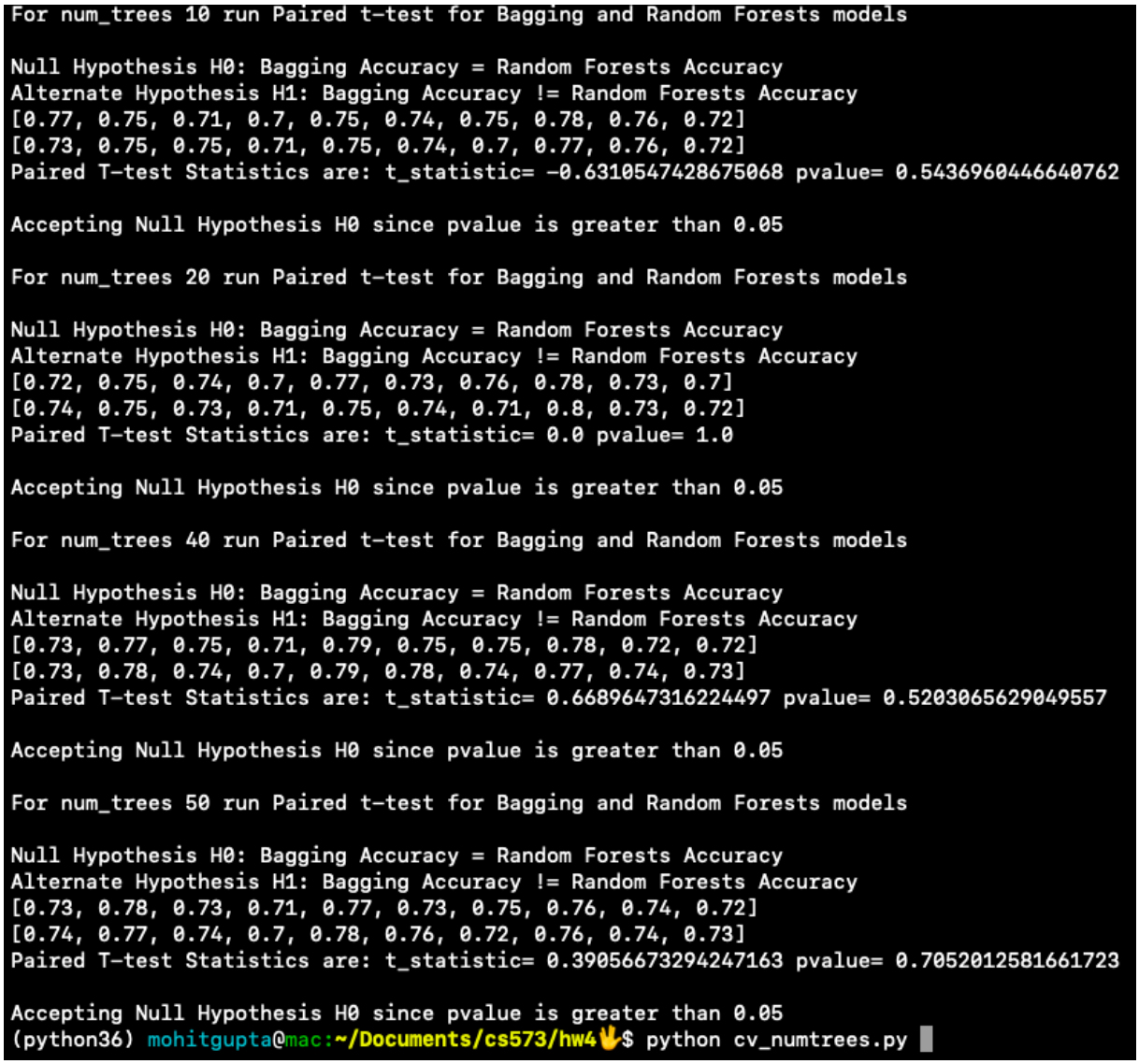

Figure 6: Hypothesis Testing Results