CS57300: Assignment 4

Due date: Wednesday March 31, 11:59 pm (submit via turnin)

1 Implement Decision Trees, Bagging and Random Forests (10 points)

```
$python trees.py trainingSet.csv testSet.csv 1
Training Accuracy DT 0.7790384615384616
Testing Accuracy DT 0.7038461538461538

$python trees.py trainingSet.csv testSet.csv 2
Training Accuracy BT 0.7967307692307692
Testing Accuracy BT 0.7623076923076924

$python trees.py trainingSet.csv testSet.csv 3
Training Accuracy RF 0.7763461538461538
Testing Accuracy RF 0.7376923076923076
```

2 The Influence of Tree Depth on Classifier Performance (10 points)

(a) Plot shown in Fig 1.

(b) Let m_{BT} refer to mean accuracy of Bagging and m_{RF} refer to the mean accuracy of the Random Forest Classifier.
Null Hypothesis $(H_0) : m_{BT} = m_{RF}$
Alternative Hypothesis $(H_1) : m_{RF} > m_{BT}$

Perform a one-tailed paired t-test for each depth limit of tree:

Depth limit of tree	p-value
3	0.51311707541818752
5	0.52163265790483293
7	0.38877259797547326
9	0.71516870619486994

We choose our significance $\alpha = 0.05$. We reject the null hypothesis if the p-value is less than α. From the above table, we fail to reject the null hypothesis for any depth limit of tree. Hence, the performance (accuracy obtained) with respect to depth limit of tree of both Bagging and Random Forests is similar, for the dataset under consideration.

3 Compare Performance of Different Models (10 points)

(a) Plot shown in figure 2.

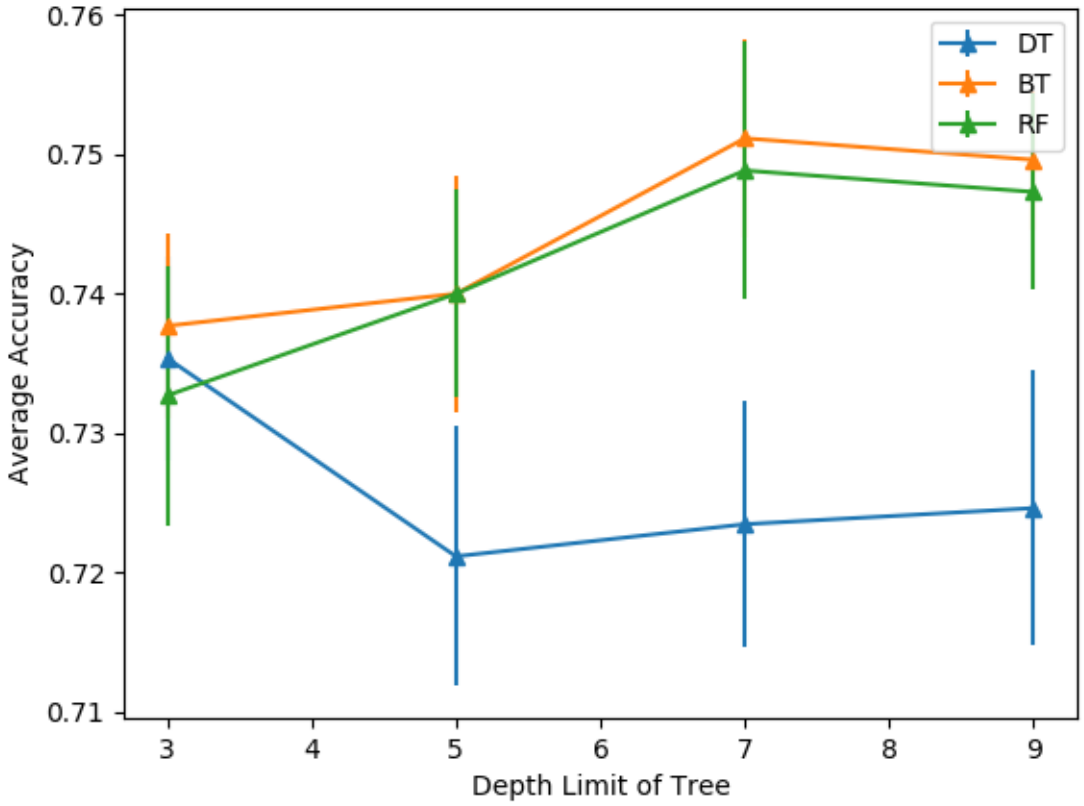

Figure 1: Accuracy versus Depth

(b) Let m_{BT} refer to mean accuracy of Bagging and m_{RF} refer to the mean accuracy of Random Forests.
Null Hypothesis $(H_0) : m_{BT} = m_{RF}$
Alternative Hypothesis $(H_1) : m_{BT} > m_{RF}$

Perform a one-tailed paired t-test for each training set size:

Fraction of Data	p-value
0.05	0.37039570316976234
0.075	0.82161658170073038
0.1	0.68703201515862578
0.15	0.6081713276760834
0.2	0.90789772175510897

We will choose our significance $\alpha = 0.05$. We can see that the p-value is greater than 0.05 for all proportions of datasets. Hence, we fail to reject the null hypothesis, indicating that the performance (accuracy obtained) with respect to fraction of data used for training of both Bagging and Random Forests is similar, for the dataset under consideration.

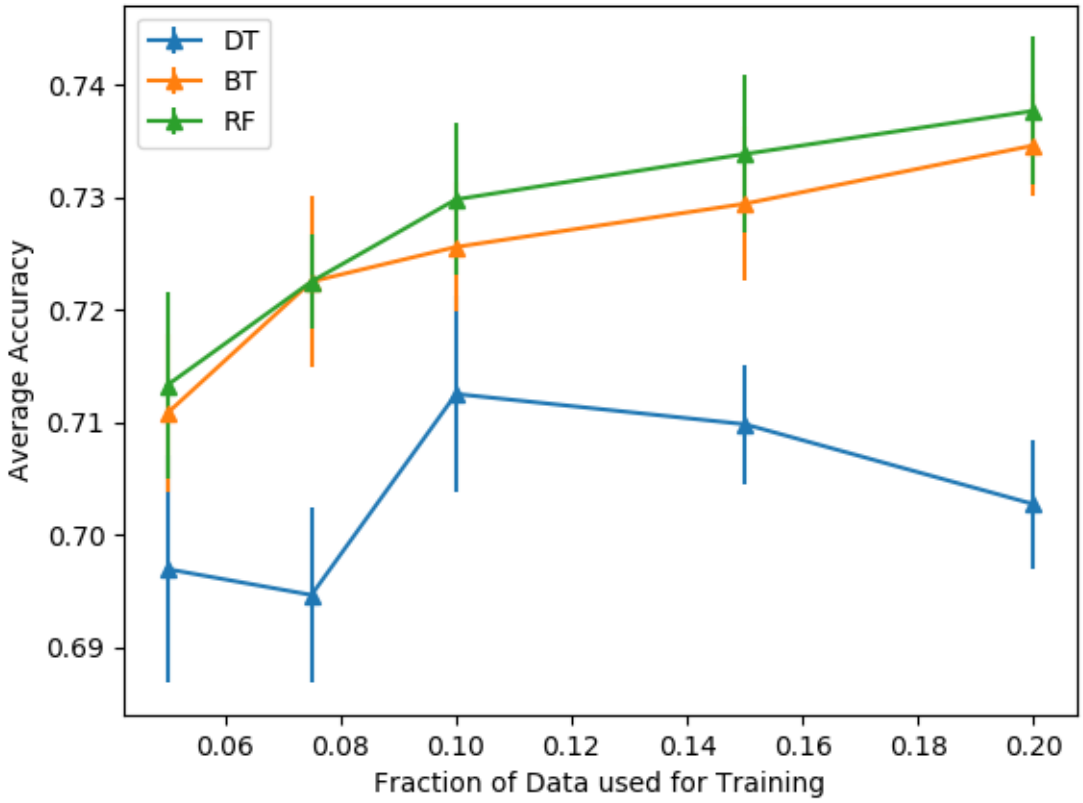

Figure 2: Accuracy versus Fraction of Training Data

4 The Influence of Number of Trees on Classifier Performance (10 points)

(a) Plot shown in figure 3.

(b) Let m_{BT} refer to mean accuracy of Bagging and m_{RF} refer to the mean accuracy of the Random Forest Classifier.
Null Hypothesis $(H_0) : m_{BT} = m_{RF}$
Alternative Hypothesis $(H_1) : m_{RF} > m_{BT}$

Perform a one-tailed paired t-test for each depth limit of tree:

Number of trees	p-value
10	0.57646748703478412
20	0.76759379799606875
40	0.82489774090062973
50	0.24571184777273516

We will choose our significance $\alpha = 0.05$. We can see that the p-value is greater than 0.05 for all values of number of trees. Hence, we fail to reject the null hypothesis, indicating that the

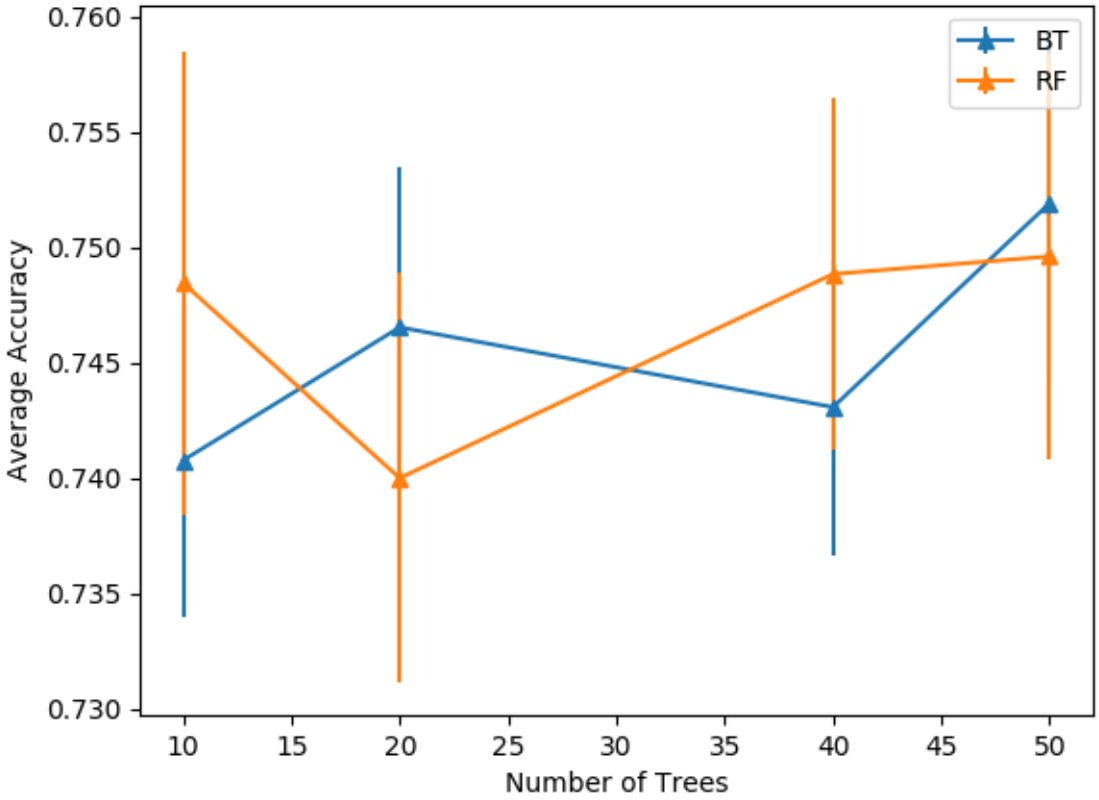

Figure 3: Accuracy versus Number of Trees

performance (accuracy obtained) with respect to number of trees of both Bagging and Random Forests is similar, for the dataset under consideration.