CS-573 Data Mining HW4

Mijanur Palash (mpalash@purdue.edu)

November 17, 2020

My python version 3.7.7

Q1- Prepossessing

Preprocessing steps are performed and the resulting files are saved with names "trainingSet.csv" and "testSet.csv" as suggested.

Q2-Implement Decision Trees, Bagging, and Random Forests

```
(deep_37) mpalash@scholar-fe00:~/573/Assignment 4 $ python trees.py "trainingSet.csv" "testSet.csv" 1
Training Accuracy DT: 77.85
Testing Accuracy DT: 71.62
(deep_37) mpalash@scholar-fe00:~/573/Assignment 4 $ python trees.py "trainingSet.csv" "testSet.csv" 2
Training Accuracy BT: 75.90
Testing Accuracy BT: 74.31
(deep_37) mpalash@scholar-fe00:~/573/Assignment 4 $ python trees.py "trainingSet.csv" "testSet.csv" 3
Training Accuracy RF: 72.76
Testing Accuracy RF: 75.54
(deep_37) mpalash@scholar-fe00:~/573/Assignment 4 $
```

Figure 1: Terminal output lines for Decision Tree (DT), Bangging (BT) and Random Forrest (RF) models

Q3-The Influence of Tree Depth on Classifier Performance

(a)- Plotting the average accuracy for 10-fold cross validation with error bars

The relation between the Tree Depth and the Classifier Performance is shown in the figure 2 below:

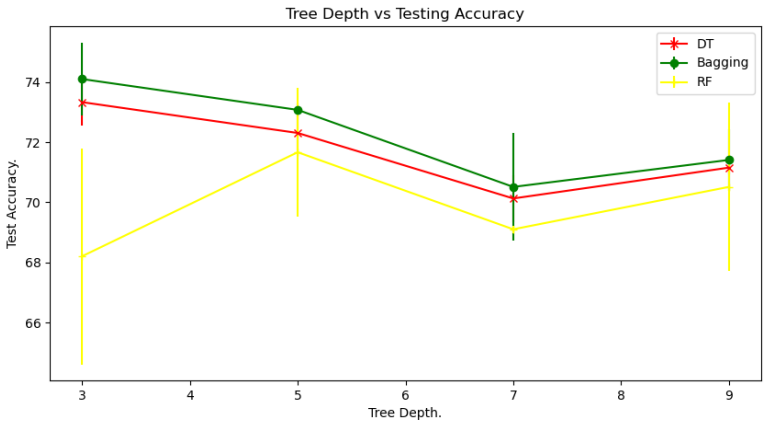

Figure 2: Testing accuracy vs tree depth.

(b)-Hypothesis Testing

Hypothesis Formulation

My hypothesis deal with DT and BT classifiers. The null and alternate hypothesis is as such:

- Null Hypothesis: Performance difference between DT and BT is trivial
- Alternate Hypothesis: Performance difference between DT and BT is significant
- The significance will be calculated with a paired t-test and associated *pvalue*. This test measures whether the average score differs significantly across samples. The input to this function is the mean accuracies over all the folds for each value of the tree depth.
- The selected significance level for the test is $\alpha = 5\%$ i.e. 0.05

Hypothesis Testing Result

From the test I get following result: $Ttest\_relResult(statistic = -1.0182663499532, pvalue = 0.38351791475499175)$

- The $pvalue$ returned by the test is $pvalue = 0.38$
- $pvalue = 0.38$ is greater than our significance level 0.05
- So we accept the null hypothesis and reject the alternate hypothesis-" Performance difference between DT and BT is trivial"

Q4-Compare Performance of Different Models-The Influence of tfrac on Classifier Performance

(a)- Plotting the average accuracy for 10-fold cross validation with error bars

The relation between the average accuracy vs the fraction of the training data used during the training process are given in the fig 3

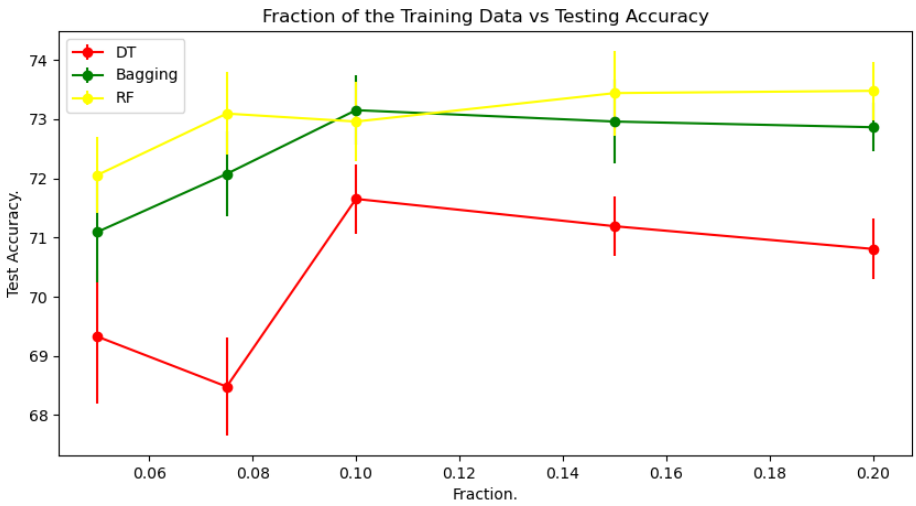

Figure 3: Testing accuracy vs training data fraction (tfrac).

(b)-Hypothesis Testing

Hypothesis Formulation

My hypothesis deal with DT and BT classifiers. The null and alternate hypothesis is as such:

- Null Hypothesis: Performance difference between DT and BT is trivial
- Alternate Hypothesis: Performance difference between DT and BT is significant
- The significance will be calculated with a paired t-test and associated *pvalue*. This test measures whether the average score differs significantly across samples. The input to this function is the mean accuracies over all the folds for each value of the tree depth.
- The selected significance level for the test is $\alpha = 5\%$ i.e. 0.05

Hypothesis Testing Result

From the test I get following result:$Ttest\_relResult(statistic = -5.703389110033904, pvalue = 0.004671640649517098)$

- The *pvalue* returned by the test is $pvalue = 0.0047$
- $pvalue = 0.0047$ is smaller than our significance level 0.05
- So we reject the null hypothesis and accept the alternate hypothesis-" Performance difference between DT and BT is significant"

Q5-The Influence of Number of Trees on Classifier Performance

(a)- Plotting the average accuracy for 10-fold cross validation with error bars

The relation between the Number of Trees during the training and the classifier Performance is shown in the fig 4

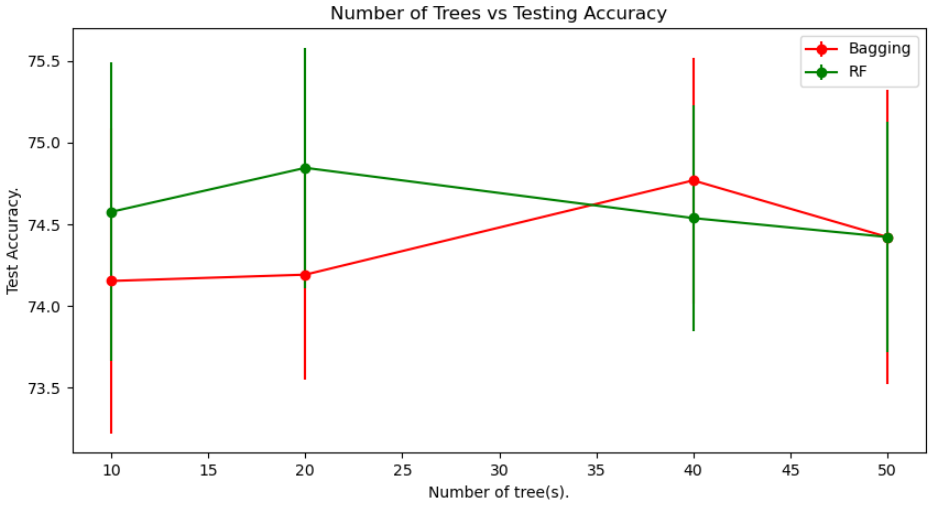

Figure 4: Testing accuracy vs Number of Trees.

(b)-Hypothesis Testing

Hypothesis Formulation

The null and alternate hypothesis is as such:

- Null Hypothesis: Performance difference between BT and RF is trivial
- Alternate Hypothesis: Performance difference between BT and RF is significant
- The significance will be calculated with a paired t-test and associated *pvalue*. This test measures whether the average score differs significantly across samples. The input to this function is the mean accuracies over all the folds for each value of the tree depth.
- The selected significance level for the test is $\alpha = 5\%$ i.e. 0.05

Hypothesis Testing Result

From the test I get following result:

- The *pvalue* returned by the test is $pvalue = 0.077$
- $pvalue = 0.77$ is greater than our significance level 0.05
- So we accept the null hypothesis and reject the alternate hypothesis-" Performance difference between BT and RF is trivial"

Bonus Question-Perceptron classifier

(a)- Model description

As part of the bonus work I implemented single layer perceptron classifier. Here I used same 10-fold strategy during training as we did on the previous sections of DT, BT and RF. As suggested the test accuracy was not considered during the training process. The algorithm I followed is given below. I shuffled training data and test data appropriately. I added one bias dimension with each training example with value initialized to 1. The initial value of the weights w is randomly initialized between $(0, 1)$.

Algorithm: Perceptron Learning Algorithm

$P \leftarrow inputs \quad with \quad label \quad 1$;
$N \leftarrow inputs \quad with \quad label \quad 0$;
Initialize **w** randomly;
while !*convergence* **do**
  Pick random $\mathbf{x} \in P \cup N$;
  if $\mathbf{x} \in P \quad and \quad \mathbf{w}.\mathbf{x} < 0$ **then**
    $\mathbf{w} = \mathbf{w} + \mathbf{x}$;
  end
  if $\mathbf{x} \in N \quad and \quad \mathbf{w}.\mathbf{x} \geq 0$ **then**
    $\mathbf{w} = \mathbf{w} - \mathbf{x}$;
  end
end
//the algorithm converges when all the inputs are classified correctly

(b)-Hyper-parameters:

For this classifier I have experimented with multiple hyper-parameters:

- Learning rate, α
- Number of epochs, N

(c)- Plotting the average accuracy for different number of epochs

This relation shown in the figure 5

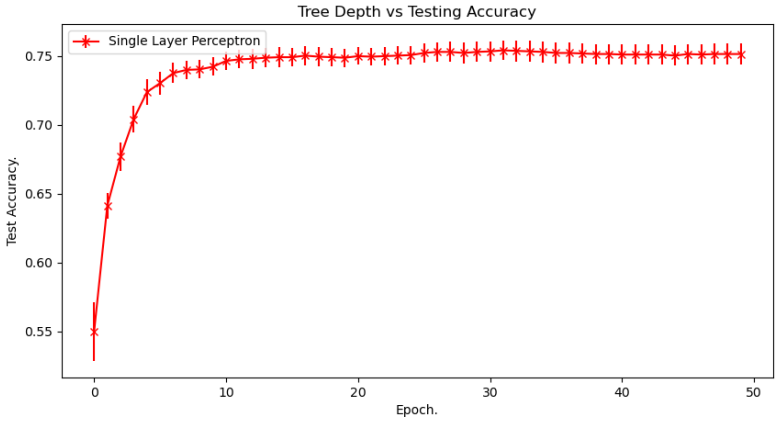

Figure 5: Testing accuracy vs epoch.

(c)- Plotting the average accuracy for different values of learning rate *alpha*

:

This relation shown in the figure 6

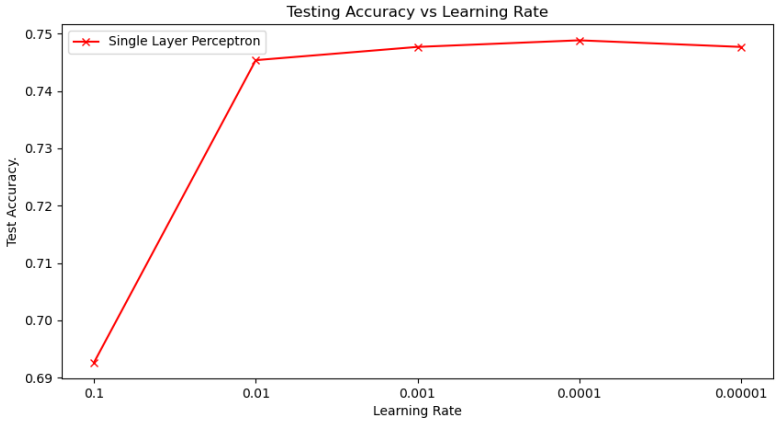

Figure 6: Testing accuracy vs LR for fixed epoch 20.

(d)-Best Hyper-parameters values:

From the figure 5 and 6 the best parameter values are

- Learning rate, $\alpha = 0.0001$
- Number of epochs, $N = 50$

With these values I got best accuracy 75% on the test data set.