CS 573: Assignment 4

Tiantu Xu

March 31, 2019

1. **Preprocessing**

   In Question 1, data pre-processing is run by the command line below:

   ```
   $ python preprocess-assg4.py
   ```

   The output is trainingSet.csv and testSet.csv.

2. **Implement Logistic Regression and Linear SVM**

   To train and test decision tree, specify `sys.argv[3] = 1`:

   ```
   $ python trees.py trainingSet.csv testSet.csv 1
   ```

   The output from my code is

   ```
   Training Accuracy DT: 0.77
   Test Accuracy DT: 0.72
   ```

   To train and test bagging, specify `sys.argv[3] = 2`:

   ```
   $ python trees.py trainingSet.csv testSet.csv 2
   ```

   The output from my code is

   ```
   Training Accuracy BT: 0.78
   Test Accuracy BT: 0.75
   ```

   To train and test random forest, specify `sys.argv[3] = 3`:

   ```
   $ python trees.py trainingSet.csv testSet.csv 3
   ```

   The output from my code is

   ```
   Training Accuracy RF: 0.76
   Test Accuracy RF: 0.73
   ```

3. **The Influence of Tree Depth on Classifier Performance**

   (a) K-fold cross-validation is run on the command line below.

   ```
   $ python cv_depth.py
   ```

The model performance on 3 models (DT, BT, and RF) is shown below.

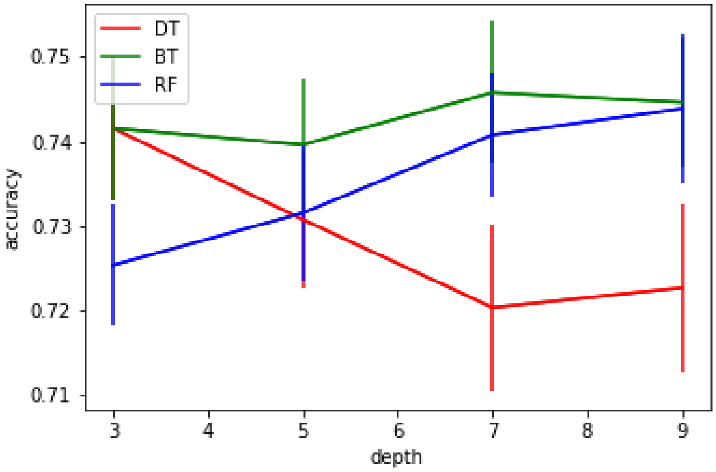

Figure 1: The model performance of DT, BT, and RF

(b) The hypothesis testing is formulated as

H_0: `DT and RF model performances do not differ significantly.`
H_1: `DT and RF model performances differ significantly.`

Assume I have a significance level of $\alpha = 0.05$. ttest is run on the performance numbers obtained in the above cross-validation. The output from the ttest is shown below

```
Ttest_indResult(statistic=-1.0235888525051595, pvalue=0.345514217128905 45)
```

It turns out that p-value $> \alpha$, so that we **fail to reject** the null hypothesis H_0 that DT and RF performances do not differ significantly.

4. **Compare Performance of Different Models**

(a) K-fold cross-validation is run on the command line below.

```
$ python cv_frac.py
```

The model performance on 3 models (DT, BT, and RF) is shown below.

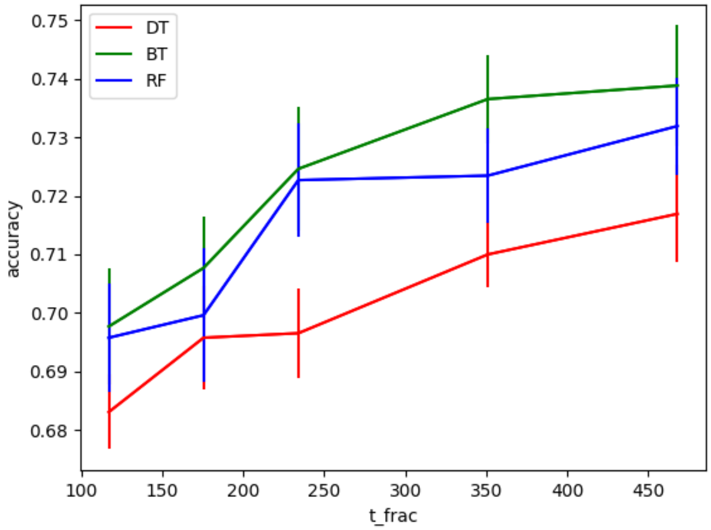

Figure 2: The model performance of DT, BT, and RF

(b) The hypothesis testing is formulated as

H_0: BT and RF model performances do not differ significantly.
H_1: BT and RF model performances differ significantly.

Assume I have a significance level of $\alpha = 0.05$. ttest is run on the performance numbers obtained in the above cross-validation. The output from the ttest is shown below

```
Ttest_indResult(statistic=0.4967647134373131, pvalue=0.6327165975663145)
```

It turns out that p-value $> \alpha$, so that we **fail to reject** the null hypothesis H_0 that BT and RF performances do not differ significantly.

5. The Influence of Number of Trees on Classifier Performance

(a) K-fold cross-validation is run on the command line below.

```
$ python cv_numtrees.py
```

The model performance on 3 models (DT, BT, and RF) is shown below.

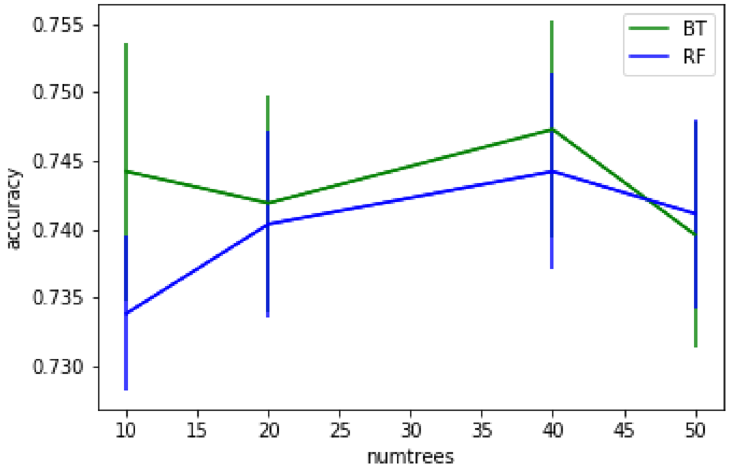

Figure 3: The model performance of DT, BT, and RF

(b) The hypothesis testing is formulated as

H_0: `BT and RF model performances do not differ significantly.`
H_1: `BT and RF model performances differ significantly.`

Assume I have a significance level of $\alpha = 0.05$. ttest is run on the performance numbers obtained in the above cross-validation. The output from the ttest is shown below

```
Ttest_indResult(statistic=1.2315497620248153, pvalue=0.26419356424794865)
```

It turns out that p-value $> \alpha$, so that we **fail to reject** the null hypothesis H_0 that BT and RF performances do not differ significantly.