CS 57300 Data Mining Assignment 4

Shuang Wu (wu1716@purdue.edu)

Environment

See README.md. Generally, the Python we use with Poetry is ~3.9 for better type hinting.

In [ ]:
```
TEST_SET = 'testSet.csv'
TRAINING_SET = 'trainingSet.csv'
MODE_DT = 1
MODE_BT = 2
MODE_RF = 3
```

Preprocessing

In [ ]:
```
%run -i preprocess-assg4.py
```

Implement Decision Trees, Bagging and Random Forests

In [ ]:
```
%run -i trees.py {TRAINING_SET} {TEST_SET} {MODE_DT}
```

```
Training Accuracy DT: 0.77
Testing Accuracy DT: 0.71
```

In [ ]:
```
%run -i trees.py {TRAINING_SET} {TEST_SET} {MODE_BT}
```

```
Training Accuracy BT: 0.79
Testing Accuracy BT: 0.73
```

In [ ]:
```
%run -i trees.py {TRAINING_SET} {TEST_SET} {MODE_RF}
```

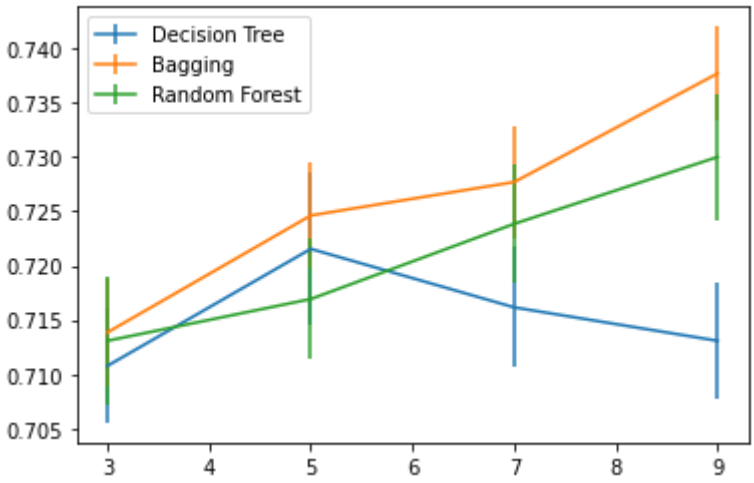
```
Training Accuracy RF: 0.78
Testing Accuracy RF: 0.74
```

The Influence of Tree Depth on Classifier Performance

In [ ]:
```
%run -i cv_depth.py
```

```
Depth: 3
[Decision Tree] Test Accuracy: 0.7107692307692308
[Decision Tree] CV Average Accuracy: 0.7328846153846154
[Decision Tree] CV Standard Error: 0.005227445210803523
[Bagging] Test Accuracy: 0.7138461538461538
[Bagging] CV Average Accuracy: 0.7355769230769231
[Bagging] CV Standard Error: 0.004912322990171074
[Random Forest] Test Accuracy: 0.713076923076923
[Random Forest] CV Average Accuracy: 0.7261538461538461
[Random Forest] CV Standard Error: 0.005878453284474705
```

```
Depth: 5
[Decision Tree] Test Accuracy: 0.7215384615384616
[Decision Tree] CV Average Accuracy: 0.7271153846153846
[Decision Tree] CV Standard Error: 0.007047230770910058
[Bagging] Test Accuracy: 0.7246153846153847
[Bagging] CV Average Accuracy: 0.7344230769230771
[Bagging] CV Standard Error: 0.004822668980416813
[Random Forest] Test Accuracy: 0.7169230769230769
[Random Forest] CV Average Accuracy: 0.7350000000000001
[Random Forest] CV Standard Error: 0.005538995700736822
Depth: 7
[Decision Tree] Test Accuracy: 0.7161538461538461
[Decision Tree] CV Average Accuracy: 0.7217307692307692
[Decision Tree] CV Standard Error: 0.005547335305719567
[Bagging] Test Accuracy: 0.7276923076923076
[Bagging] CV Average Accuracy: 0.7438461538461538
[Bagging] CV Standard Error: 0.005215759006303332
[Random Forest] Test Accuracy: 0.7238461538461538
[Random Forest] CV Average Accuracy: 0.7403846153846153
[Random Forest] CV Standard Error: 0.005452864184137627
Depth: 9
[Decision Tree] Test Accuracy: 0.713076923076923
[Decision Tree] CV Average Accuracy: 0.7196153846153845
[Decision Tree] CV Standard Error: 0.005390107089647641
[Bagging] Test Accuracy: 0.7376923076923076
[Bagging] CV Average Accuracy: 0.7432692307692308
[Bagging] CV Standard Error: 0.004310867648357675
[Random Forest] Test Accuracy: 0.73
[Random Forest] CV Average Accuracy: 0.7476923076923078
[Random Forest] CV Standard Error: 0.005787151653478162
```



We formulate the following hypothesis.

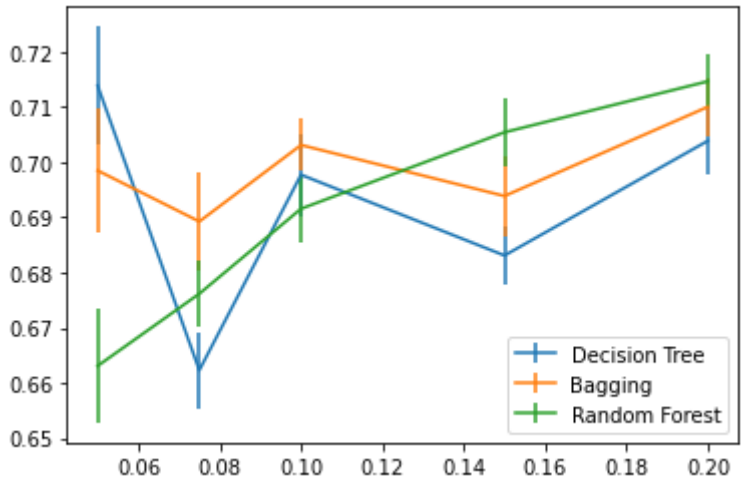
- *H0*: There is no significance difference between the accuracy of bagging and random forest models.
- *H1*: There is a significance difference between the accuracy of bagging and random forest models.

We use paired t-test to test the hypothesis. The test accuracies of bagging and random forest models against different depths are `(0.7138, 0.7246, 0.7276, 0.7377)` and `(0.7130, 0.7169, 0.7238, 0.7300)` respectively. Thus, the paired *t-value* is 2.9854. As the degree of freedom is 3, and *t-value* < 3.18, according to the t-value table, $p < 0.05$ does not hold. Thus, the test fails to reject *H0*. Namely, the difference between the accuracy of bagging and random forest models is not significant.

Compare Performance of Different Models

In [ ]:
```
%run -i cv_frac.py
```

```
t_frac: 0.05
[Decision Tree] Test Accuracy: 0.7138461538461538
[Decision Tree] CV Average Accuracy: 0.6950000000000001
[Decision Tree] CV Standard Error: 0.0106802643851769
[Bagging] Test Accuracy: 0.6984615384615385
[Bagging] CV Average Accuracy: 0.7007692307692308
[Bagging] CV Standard Error: 0.011250246545621928
[Random Forest] Test Accuracy: 0.6630769230769231
[Random Forest] CV Average Accuracy: 0.7028846153846154
[Random Forest] CV Standard Error: 0.010274523268861037
t_frac: 0.075
[Decision Tree] Test Accuracy: 0.6623076923076923
[Decision Tree] CV Average Accuracy: 0.6905769230769231
[Decision Tree] CV Standard Error: 0.00692548035204122
[Bagging] Test Accuracy: 0.6892307692307692
[Bagging] CV Average Accuracy: 0.7159615384615384
[Bagging] CV Standard Error: 0.008787640931222362
[Random Forest] Test Accuracy: 0.6761538461538461
[Random Forest] CV Average Accuracy: 0.7226923076923076
[Random Forest] CV Standard Error: 0.0058696390623183565
t_frac: 0.1
[Decision Tree] Test Accuracy: 0.6976923076923077
[Decision Tree] CV Average Accuracy: 0.7153846153846153
[Decision Tree] CV Standard Error: 0.007438107540697584
[Bagging] Test Accuracy: 0.703076923076923
[Bagging] CV Average Accuracy: 0.7319230769230769
[Bagging] CV Standard Error: 0.004704272261975585
[Random Forest] Test Accuracy: 0.6915384615384615
[Random Forest] CV Average Accuracy: 0.7184615384615385
[Random Forest] CV Standard Error: 0.006167053135712243
t_frac: 0.15
[Decision Tree] Test Accuracy: 0.683076923076923
[Decision Tree] CV Average Accuracy: 0.7140384615384615
[Decision Tree] CV Standard Error: 0.005167676620765345
[Bagging] Test Accuracy: 0.6938461538461539
[Bagging] CV Average Accuracy: 0.724423076923077
[Bagging] CV Standard Error: 0.007170008624028219
[Random Forest] Test Accuracy: 0.7053846153846154
[Random Forest] CV Average Accuracy: 0.7278846153846155
[Random Forest] CV Standard Error: 0.0061912923200020155
t_frac: 0.2
[Decision Tree] Test Accuracy: 0.7038461538461539
[Decision Tree] CV Average Accuracy: 0.7099999999999999
[Decision Tree] CV Standard Error: 0.006218411296565581
[Bagging] Test Accuracy: 0.71
[Bagging] CV Average Accuracy: 0.7323076923076923
[Bagging] CV Standard Error: 0.005215759006303328
[Random Forest] Test Accuracy: 0.7146153846153847
[Random Forest] CV Average Accuracy: 0.7325000000000002
[Random Forest] CV Standard Error: 0.00484561962867483
```



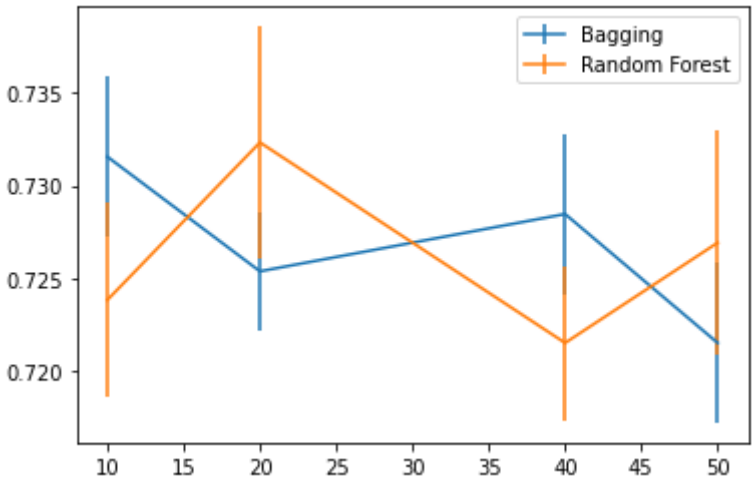
We formulate the following hypothesis.

- *H0*: There is no significance difference between the accuracy of decision tree and bagging models.
- *H1*: There is a significance difference between the accuracy of decision tree and bagging models.

We use paired t-test to test the hypothesis. The test accuracies of decision tree and bagging models against different depths are `(0.714, 0.662, 0.697, 0.683, 0.704)` and `(0.698, 0.689, 0.703, 0.693, 0.710)` respectively. Thus, the paired *t-value* is 0.9631. As the degree of freedom is 4, and *t-value* < 2.78, according to the t-value table, $p < 0.05$ does not hold. Thus, the test fails to reject *H0*. Namely, the difference between the accuracy of decision tree and bagging models is not significant.

In [ ]:
```
%run -i cv_numtrees.py
```

```
Number of Trees: 10
[Bagging] Test Accuracy: 0.7315384615384616
[Bagging] CV Average Accuracy: 0.7384615384615385
[Bagging] CV Standard Error: 0.004300130725961134
[Random Forest] Test Accuracy: 0.7238461538461538
[Random Forest] CV Average Accuracy: 0.7313461538461539
[Random Forest] CV Standard Error: 0.005196223594589669
Number of Trees: 20
[Bagging] Test Accuracy: 0.7253846153846154
[Bagging] CV Average Accuracy: 0.7465384615384616
[Bagging] CV Standard Error: 0.003145862314873621
[Random Forest] Test Accuracy: 0.7323076923076923
[Random Forest] CV Average Accuracy: 0.7426923076923077
[Random Forest] CV Standard Error: 0.0062243556735051605
Number of Trees: 40
[Bagging] Test Accuracy: 0.7284615384615385
[Bagging] CV Average Accuracy: 0.7423076923076923
[Bagging] CV Standard Error: 0.004317296984739162
[Random Forest] Test Accuracy: 0.7215384615384616
[Random Forest] CV Average Accuracy: 0.7440384615384615
[Random Forest] CV Standard Error: 0.004155359410967901
Number of Trees: 50
[Bagging] Test Accuracy: 0.7215384615384616
[Bagging] CV Average Accuracy: 0.7413461538461539
[Bagging] CV Standard Error: 0.004285053837530227
[Random Forest] Test Accuracy: 0.7269230769230769
[Random Forest] CV Average Accuracy: 0.7459615384615383
[Random Forest] CV Standard Error: 0.006016803393573954
```



We formulate the following hypothesis.

- *H0*: There is no significance difference between the accuracy of bagging and random forest models.
- *H1*: There is a significance difference between the accuracy of bagging and random forest models.

We use paired t-test to test the hypothesis. The test accuracies of bagging and random forest models against different depths are `(0.732, 0.725, 0.728, 0.722)` and `(0.724, 0.732, 0.722, 0.727)` respectively. Thus, the paired *t-value* is 0.1317. As the degree of freedom is 3, and *t-value* < 3.18, according to the t-value table, $p < 0.05$ does not hold. Thus, the test fails to reject *H0*. Namely, the difference between the accuracy of bagging and random forest models is not significant.

Bonus Question

We implemented a simple multi-layer neural network. The hyperparamters are set as the default arguments.

- Epochs: 100
- Initial Learning Rate: 0.8
- Momentum Weight: 0.1
- Network Shape: (4, 2, 1)

The model (network) is initialized with random weights. Each weight are uniformly distributed within (-1, 1). Since the training time is relatively long, and thus we try to make the network as simple as possible. Perhaps the hyperparamters are not optimal but increasing the complexity of the model will significantly increase the time to make the training converge.

In [ ]:
```
%run -i bonus.py {TRAINING_SET} {TEST_SET}
```

```
Training Accuracy MLP: 0.7534615384615385
Testing Accuracy MLP: 0.7276923076923076
```