CS57300: Homework 4

Praneeth Sai Valiveti – PUID: 0029203008

NOTE: Main Question and Extra Credit have been solved separately. Extra Credit starts from page 9.

NOTE: One late day has been used for the submission of this homework.

1. ***Assess whether ensembles improve performance.***

   (a) **Plot the learning curves for the three models plus SVM (in the same plot), including error bars that indicate ±1 standard error, from the evaluation based on incremental CV.**

   Ans. The following are the result tables for each model on all 10 validation cases:

Decision

	1	2	3	4	5	6	7	8	9	10	Means	STD
0.025	0.3400	0.3650	0.4300	0.3300	0.3700	0.3600	0.3000	0.3100	0.3350	0.3700	0.3510	0.0370
0.05	0.3650	0.3150	0.3000	0.2200	0.4200	0.2450	0.3050	0.3850	0.2750	0.4250	0.3255	0.0673
0.125	0.3100	0.2200	0.2200	0.2300	0.3150	0.1600	0.2100	0.2900	0.3000	0.3150	0.2570	0.0525
0.25	0.1800	0.2150	0.3050	0.2550	0.2500	0.1650	0.2000	0.1800	0.2350	0.1950	0.2180	0.0410

Bagged

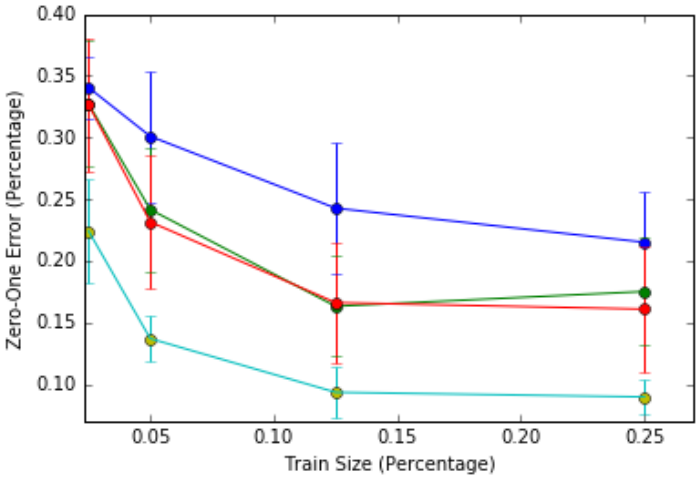
	1	2	3	4	5	6	7	8	9	10	Means	STD
0.025	0.2500	0.3850	0.3100	0.3200	0.4000	0.3100	0.2600	0.2700	0.3000	0.3350	0.3140	0.0470
0.05	0.2800	0.2350	0.3150	0.1550	0.2400	0.2400	0.2000	0.2350	0.1450	0.1850	0.2230	0.0503
0.125	0.1950	0.2200	0.2250	0.1500	0.1600	0.1450	0.1350	0.2450	0.2100	0.2000	0.1885	0.0363
0.25	0.1750	0.1650	0.2150	0.1750	0.1750	0.1650	0.1350	0.1250	0.1900	0.1300	0.1650	0.0267

Random

	1	2	3	4	5	6	7	8	9	10	Means	STD
0.025	0.1900	0.3950	0.2950	0.3100	0.4000	0.3200	0.2100	0.3300	0.3050	0.3500	0.3105	0.0649
0.05	0.3050	0.2500	0.3050	0.1850	0.2050	0.1900	0.2000	0.1900	0.1450	0.2100	0.2185	0.0497
0.125	0.1600	0.2250	0.1800	0.1400	0.1650	0.1350	0.1200	0.1900	0.2150	0.2150	0.1745	0.0348
0.25	0.1500	0.1550	0.2150	0.1750	0.1350	0.1900	0.1200	0.1300	0.1800	0.1550	0.1605	0.0280

SVM

	1	2	3	4	5	6	7	8	9	10	Means	STD
0.025	0.1150	0.2100	0.1950	0.2300	0.2500	0.1950	0.1700	0.1950	0.1400	0.2450	0.1945	0.0413
0.05	0.1750	0.1750	0.1900	0.1350	0.1300	0.1200	0.0800	0.1350	0.1300	0.1150	0.1385	0.0313
0.125	0.0650	0.0950	0.1450	0.0950	0.1200	0.0950	0.1100	0.0500	0.1050	0.0550	0.0935	0.0282
0.25	0.0800	0.0700	0.0800	0.0650	0.0750	0.1100	0.0850	0.0600	0.0700	0.0500	0.0745	0.0154



From the above plot, we can see that the SVM model (Teal) performs best by a significant margin, followed by Random Forest (Red), Bagged (Green) and last by the single Decision Tree (Blue).

At first glance, we can see that there is a significant difference between the performance between Single tree, RF and Bagged, and the SVM.

Also, we see that on an average, the performance betters with selection of more train size. For all the models.

(b) **Formulate a hypothesis about the performance difference you observe for one of the ensembles compared to the SVM. Discuss how the observed data support the hypothesis (i.e., are the observed differences significant).**

Ans.

Hypothesis:

H_0: The average performance of the Random Forest classifier is the same as that of the SVM classifier.

H_1: The average performance of the Random Forest classifier is less than that of the SVM classifier.

Paired T-Test and CI: C1, C8

```
Paired T for C1 - C8

             N    Mean   StDev  SE Mean
C1          10  0.3105  0.0684   0.0216
C8          10  0.1945  0.0435   0.0138
Difference  10  0.1160  0.0446   0.0141

95% lower bound for mean difference: 0.0902
T-Test of mean difference = 0 (vs > 0): T-Value = 8.23  P-Value = 0.000
```

Paired T-Test and CI: C2, C9

```
Paired T for C2 - C9

             N    Mean   StDev  SE Mean
C2          10  0.2185  0.0524   0.0166
C9          10  0.1385  0.0330   0.0104
Difference  10  0.0800  0.0357   0.0113

95% lower bound for mean difference: 0.0593
T-Test of mean difference = 0 (vs > 0): T-Value = 7.09  P-Value = 0.000
```

Paired T-Test and CI: C3, C10

```
Paired T for C3 - C10

             N    Mean   StDev  SE Mean
C3          10  0.1745  0.0367   0.0116
C10         10  0.0935  0.0297   0.0094
Difference  10  0.0810  0.0523   0.0165

95% lower bound for mean difference: 0.0507
T-Test of mean difference = 0 (vs > 0): T-Value = 4.90  P-Value = 0.000
```

Paired T-Test and CI: C4, C11

```
Paired T for C4 - C11

             N     Mean    StDev  SE Mean
C4          10  0.16050  0.02948  0.00932
C11         10  0.07450  0.01624  0.00513
Difference  10  0.08600  0.02933  0.00927

95% lower bound for mean difference: 0.06900
T-Test of mean difference = 0 (vs > 0): T-Value = 9.27  P-Value = 0.000
```

Case	0.025	0.05	0.125	0.25
p-Value	0	0	0	0

From the Paired T-Tests, we see that the P-Value of all the cases is less than 0.05 (95% confidence). Hence, we can reject the null hypothesis in favor of the alternate that the RF is indeed less performing than the SVM.

2. ***Assess whether the number of features affects performance.***
Fix the training set size at 500 (0.25%) and vary the number of features: [200,500,1000,1500].

(a) **Plot the learning curves for the three models plus SVM (in the same plot), including error bars that indicate ±1 standard error, from the evaluation based on incremental CV.**

Ans.

Decision

	1	2	3	4	5	6	7	8	9	10	Means	STD
200	0.3050	0.2100	0.2250	0.3450	0.2300	0.2550	0.3450	0.3200	0.2850	0.3100	0.2830	0.0475
500	0.1800	0.3200	0.1750	0.3250	0.2250	0.1950	0.2050	0.2450	0.1700	0.1800	0.2220	0.0550
1000	0.1800	0.2150	0.1850	0.2450	0.2400	0.2400	0.3000	0.2000	0.1950	0.2150	0.2215	0.0342
1500	0.1900	0.2200	0.1300	0.3200	0.1750	0.2400	0.2250	0.2300	0.3100	0.2150	0.2255	0.0542

Bagged

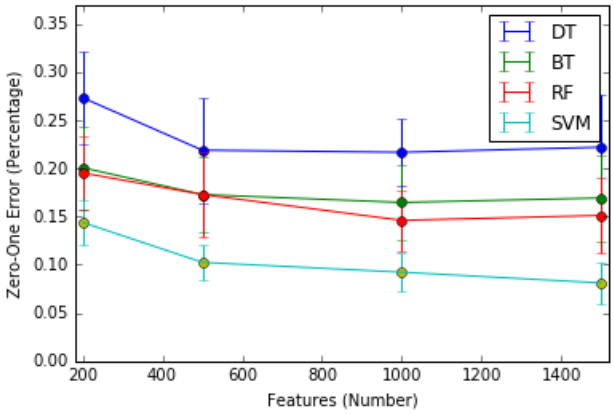
	1	2	3	4	5	6	7	8	9	10	Means	STD
200	0.1900	0.1450	0.1800	0.2900	0.1850	0.2500	0.2550	0.1850	0.1700	0.1800	0.2030	0.0434
500	0.1850	0.2300	0.1450	0.2450	0.1600	0.1650	0.1950	0.1400	0.1400	0.1200	0.1725	0.0388
1000	0.1400	0.1900	0.1450	0.2450	0.1200	0.2050	0.1750	0.1500	0.1150	0.1500	0.1635	0.0383
1500	0.2050	0.1900	0.1400	0.2800	0.1350	0.1650	0.1650	0.1450	0.1250	0.1300	0.1680	0.0448

Random

	1	2	3	4	5	6	7	8	9	10	Means	STD
200	0.1650	0.1550	0.1750	0.2650	0.2000	0.2200	0.2600	0.1950	0.1550	0.1850	0.1975	0.0378
500	0.1900	0.2050	0.1350	0.2800	0.1850	0.1250	0.1800	0.1400	0.1400	0.1400	0.1720	0.0447
1000	0.1500	0.1850	0.1200	0.2050	0.1100	0.1650	0.1350	0.1450	0.1150	0.1050	0.1435	0.0317
1500	0.1950	0.1550	0.1300	0.2450	0.1250	0.1100	0.1350	0.1150	0.1300	0.1450	0.1485	0.0394

SVM

	1	2	3	4	5	6	7	8	9	10	Means	STD
200	0.1450	0.1550	0.1300	0.1800	0.1250	0.1800	0.1350	0.1300	0.1350	0.1050	0.1420	0.0226
500	0.0850	0.1150	0.0750	0.1300	0.1150	0.0900	0.0850	0.0800	0.0800	0.1100	0.0965	0.0182
1000	0.1000	0.1000	0.0800	0.1200	0.0600	0.0950	0.0700	0.1000	0.0550	0.0700	0.0850	0.0200
1500	0.0650	0.0650	0.0400	0.0950	0.0550	0.1050	0.0950	0.0500	0.0600	0.0950	0.0725	0.0217



From the plot, we can see that on an average, as the number of features increases, the model performs slightly better, and there is still the significant difference between the performance of each model as described above.

Here, we see that the Random Forest and Bagged Trees intertwine, with RF performing better on an average.

(b) **Formulate a hypothesis about the performance difference you observe for one of the ensembles compared to the SVM. Discuss how the observed data support the hypothesis.**

Ans.

Hypothesis:

H_0: The average performance of the Random Forest classifier is the same as that of the SVM classifier.

H_1: The average performance of the Random Forest classifier is less than that of the SVM classifier.

Paired T-Test and CI: C1, C8

```
Paired T for C1 - C8

             N    Mean   StDev  SE Mean
C1          10  0.1975  0.0399   0.0126
C8          10  0.1420  0.0238   0.0075
Difference  10  0.0555  0.0376   0.0119

95% lower bound for mean difference: 0.0337
T-Test of mean difference = 0 (vs > 0): T-Value = 4.67  P-Value = 0.001
```

Paired T-Test and CI: C2, C9

```
Paired T for C2 - C9

             N    Mean   StDev  SE Mean
C2          10  0.1720  0.0471   0.0149
C9          10  0.0965  0.0192   0.0061
Difference  10  0.0755  0.0357   0.0113

95% lower bound for mean difference: 0.0548
T-Test of mean difference = 0 (vs > 0): T-Value = 6.69  P-Value = 0.000
```

Paired T-Test and CI: C3, C10

```
Paired T for C3 - C10

             N     Mean    StDev  SE Mean
C3          10   0.1435   0.0334   0.0106
C10         10   0.0850   0.0211   0.0067
Difference  10  0.05850  0.01765  0.00558

95% lower bound for mean difference: 0.04827
T-Test of mean difference = 0 (vs > 0): T-Value = 10.48  P-Value = 0.000
```

Paired T-Test and CI: C4, C11

```
Paired T for C4 - C11

             N    Mean   StDev  SE Mean
C4          10  0.1485  0.0416   0.0131
C11         10  0.0725  0.0229   0.0072
Difference  10  0.0760  0.0422   0.0133

95% lower bound for mean difference: 0.0516
T-Test of mean difference = 0 (vs > 0): T-Value = 5.70  P-Value = 0.000
```

Case	200	500	1000	1500
p-Value	0.001	0	0	0

From the Paired T-Tests, we see that the P-Value of all the cases is less than 0.05 (95% confidence). Hence, we can reject the null hypothesis in favor of the alternate that the RF is indeed less performing than the SVM.

3. ***Assess whether the depth of the tree affects performance.***
Fix the training set size at 500 and vary the depth limit on the decision trees: [5,10,15,20].

(a) **Plot the learning curves for the three tree models (in the same plot), including error bars that indicate ±1 standard error, from the evaluation based on incremental CV.**

Ans.

Decision

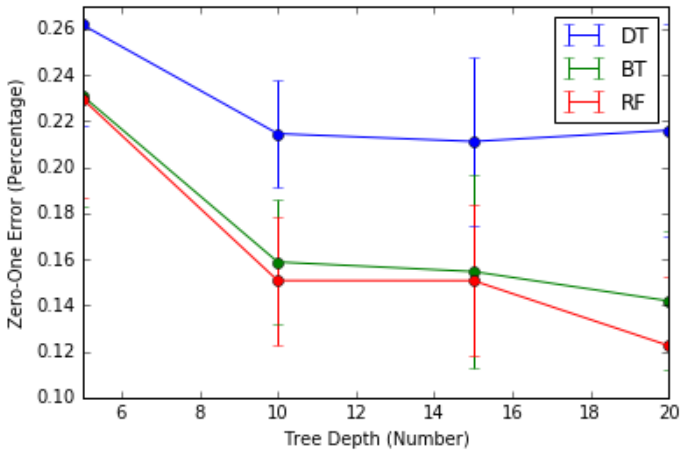
	1	2	3	4	5	6	7	8	9	10	Means	STD
5	0.2500	0.2500	0.2950	0.2600	0.2950	0.2550	0.2300	0.3050	0.3650	0.2000	0.2705	0.0437
10	0.1950	0.1900	0.2050	0.2000	0.2550	0.2550	0.2050	0.2400	0.2250	0.2300	0.2200	0.0231
15	0.2100	0.2450	0.2000	0.2200	0.2400	0.1900	0.1900	0.2200	0.2900	0.1450	0.2150	0.0369
20	0.1650	0.2700	0.1600	0.2650	0.2850	0.2000	0.2200	0.1500	0.2400	0.2400	0.2195	0.0463

Bagged

	1	2	3	4	5	6	7	8	9	10	Means	STD
5	0.2200	0.2000	0.2300	0.2300	0.2850	0.2400	0.1500	0.2850	0.3250	0.1950	0.2360	0.0484
10	0.1850	0.1750	0.1750	0.1700	0.1200	0.1900	0.1700	0.1450	0.1450	0.1050	0.1580	0.0269
15	0.1500	0.2050	0.1100	0.1050	0.1950	0.1300	0.1050	0.1550	0.2300	0.1350	0.1520	0.0421
20	0.0900	0.1850	0.1400	0.1600	0.1350	0.1350	0.1650	0.1600	0.1350	0.0850	0.1390	0.0301

Random

	1	2	3	4	5	6	7	8	9	10	Means	STD
5	0.2500	0.1850	0.2550	0.2300	0.1850	0.2550	0.1850	0.3300	0.2600	0.2150	0.2350	0.0431
10	0.1750	0.1750	0.1650	0.1450	0.1600	0.1500	0.1450	0.1650	0.1350	0.0750	0.1490	0.0277
15	0.1200	0.1750	0.1000	0.1400	0.1450	0.1450	0.1300	0.1800	0.2200	0.1300	0.1485	0.0328
20	0.0750	0.1550	0.0950	0.1550	0.1150	0.1100	0.1300	0.1550	0.1150	0.0700	0.1175	0.0300



From the above plot, we can see a drastic increase in performance resulting from increase in depth of the ensemble trees. As expected, the Single decision tree also improves on an average.

The random forest is slightly better than the Bagged Tree, and these two are significantly better than the Single Decision Tree.

(b) **Formulate a hypothesis about the performance difference you observe between two of the models. Discuss how the observed data support the hypothesis.**

Ans.

Hypothesis:

H_0: The average performance of the Random Forest classifier is the same as the Bagged Tree classifier.

H_1: The average performance of the Random Forest classifier is better than the Bagged Tree classifier.

Paired T-Test and CI: C1, C8

```
Paired T for C1 - C8

             N    Mean   StDev  SE Mean
C1          10  0.2360  0.0510   0.0161
C8          10  0.2350  0.0455   0.0144
Difference  10  0.0010  0.0470   0.0148

95% lower bound for mean difference: -0.0262
T-Test of mean difference = 0 (vs > 0): T-Value = 0.07  P-Value = 0.474
```

Paired T-Test and CI: C2, C9

```
Paired T for C2 - C9

             N     Mean    StDev  SE Mean
C2          10  0.15800  0.02830  0.00895
C9          10  0.14900  0.02923  0.00924
Difference  10  0.00900  0.02413  0.00763

95% lower bound for mean difference: -0.00499
T-Test of mean difference = 0 (vs > 0): T-Value = 1.18  P-Value = 0.134
```

Paired T-Test and CI: C3, C10

```
Paired T for C3 - C10

             N     Mean    StDev  SE Mean
C3          10   0.1520   0.0444   0.0140
C10         10   0.1485   0.0346   0.0109
Difference  10  0.00350  0.02809  0.00888

95% lower bound for mean difference: -0.01278
T-Test of mean difference = 0 (vs > 0): T-Value = 0.39  P-Value = 0.351
```

Paired T-Test and CI: C4, C11

```
Paired T for C4 - C11

             N     Mean    StDev  SE Mean
C4          10   0.1390   0.0317   0.0100
C11         10   0.1175   0.0316   0.0100
Difference  10  0.02150  0.01270  0.00402

95% lower bound for mean difference: 0.01414
T-Test of mean difference = 0 (vs > 0): T-Value = 5.35  P-Value = 0.000
```

Case	5	10	15	20
p-Value	0.474	0.134	0.351	0

From the Paired T-Tests, we see that the P-Value is less than 0.05 (95% confidence) for only 1 case, and hence can be rejected only for that one case. Hence, by majority vote, we fail to reject the null hypothesis. Hence, we can reject the null hypothesis in favor of the alternate that the RF is indeed the same as that of Bagged Tree. But, a point to be noted here is that there is p-value of 0 for tree depth of trees. This might mean that there is a monotonic increase starting from that point onwards.

4. ***Assess whether the number of trees affects performance.***
Fix the training set size at 500 and vary the number of trees in the ensembles: [10,25,50,100].

(a) **Plot the learning curves for the ensemble models (in the same plot), including error bars that indicate ±1 standard error, from the evaluation based on incremental CV.**

Ans.

Decision

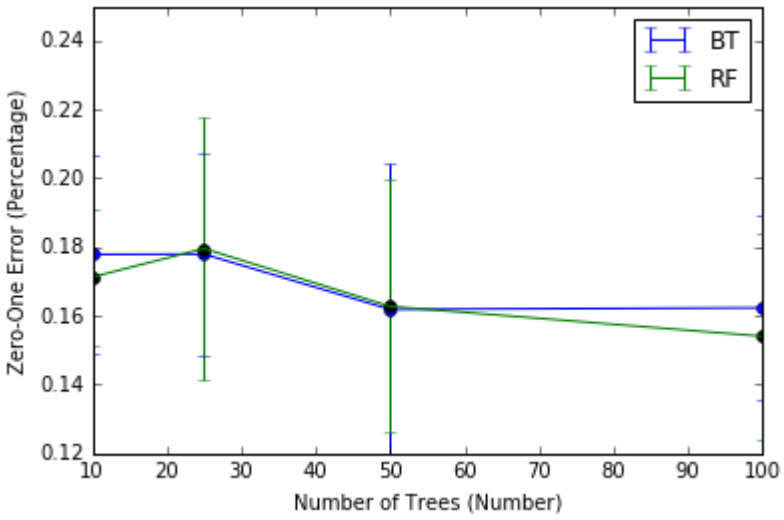
	1	2	3	4	5	6	7	8	9	10	Means	STD
10	0.2050	0.1950	0.2750	0.2850	0.1850	0.2400	0.3000	0.1950	0.2400	0.2450	0.2365	0.0388
25	0.2850	0.1850	0.2700	0.2050	0.2050	0.1700	0.2150	0.2250	0.1850	0.2900	0.2235	0.0412
50	0.2550	0.2850	0.2050	0.2050	0.1450	0.2150	0.2050	0.2000	0.2350	0.2700	0.2220	0.0385
100	0.2250	0.2650	0.1650	0.2150	0.1850	0.1750	0.2350	0.3350	0.2400	0.2000	0.2240	0.0474

Bagged

	1	2	3	4	5	6	7	8	9	10	Means	STD
10	0.1900	0.1850	0.2100	0.2300	0.1600	0.1800	0.2150	0.1650	0.1550	0.2000	0.1890	0.0236
25	0.1700	0.1900	0.2400	0.1050	0.1250	0.1750	0.1850	0.1300	0.1550	0.2100	0.1685	0.0390
50	0.2300	0.1600	0.1600	0.1700	0.1650	0.1600	0.2000	0.1350	0.1300	0.1800	0.1690	0.0278
100	0.1750	0.1900	0.1500	0.1800	0.1750	0.1700	0.1750	0.2400	0.1500	0.1500	0.1755	0.0252

Random

	1	2	3	4	5	6	7	8	9	10	Means	STD
10	0.2150	0.2000	0.2050	0.2450	0.1800	0.1300	0.1850	0.1850	0.1800	0.1450	0.1870	0.0312
25	0.1650	0.1700	0.2050	0.0950	0.1450	0.1700	0.1800	0.1550	0.1250	0.1700	0.1580	0.0290
50	0.2250	0.1300	0.1400	0.1900	0.1550	0.1700	0.1750	0.1550	0.1350	0.2000	0.1675	0.0291
100	0.1850	0.2000	0.1300	0.1150	0.1500	0.1600	0.1950	0.2050	0.0950	0.1250	0.1560	0.0371



As we can see, the performance of the two models seems to have an erratic nature with the Random Forest out-performing the Bagged Trees slightly.

(b) **Formulate a hypothesis about the performance difference you observe for one of the ensembles compared to the single decision tree. Discuss how the observed data support the hypothesis.**

Ans.

Hypothesis:

H_0: The average performance of the Random Forest classifier is the same as the Decision Tree classifier.

H_1: The average performance of the Random Forest classifier is better than the Decision Tree classifier.

Paired T-Test and CI: C1, C6

```
Paired T for C1 - C6

             N    Mean   StDev  SE Mean
C1          10  0.2365  0.0409   0.0129
C6          10  0.1870  0.0329   0.0104
Difference  10  0.0495  0.0485   0.0153

95% lower bound for mean difference: 0.0214
T-Test of mean difference = 0 (vs > 0): T-Value = 3.23  P-Value = 0.005
```

Paired T-Test and CI: C2, C7

```
Paired T for C2 - C7

             N    Mean   StDev  SE Mean
C2          10  0.2235  0.0434   0.0137
C7          10  0.1580  0.0306   0.0097
Difference  10  0.0655  0.0419   0.0132

95% lower bound for mean difference: 0.0412
T-Test of mean difference = 0 (vs > 0): T-Value = 4.95  P-Value = 0.000
```

Paired T-Test and CI: C3, C8

```
Paired T for C3 - C8

             N    Mean   StDev  SE Mean
C3          10  0.2220  0.0406   0.0128
C8          10  0.1675  0.0307   0.0097
Difference  10  0.0545  0.0467   0.0148

95% lower bound for mean difference: 0.0274
T-Test of mean difference = 0 (vs > 0): T-Value = 3.69  P-Value = 0.002
```

Paired T-Test and CI: C4, C9

```
Paired T for C4 - C9

             N    Mean   StDev  SE Mean
C4          10  0.2240  0.0499   0.0158
C9          10  0.1560  0.0391   0.0124
Difference  10  0.0680  0.0440   0.0139

95% lower bound for mean difference: 0.0425
T-Test of mean difference = 0 (vs > 0): T-Value = 4.89  P-Value = 0.000
```

Case	10	25	50	100
p-Value	0.005	0	0.002	0

From the Paired T-Tests, we see that the P-Value is less than 0.05 (95% confidence) for all cases, and the null hypothesis is rejected for all 4 cases. Hence, by majority vote, we reject the null hypothesis in favor of the alternate that the performance RF is indeed better than that of a Single Decision Tree.

5. ***Prove that the expected squared loss for a single example can be decomposed into bias/variance/noise.*** **Show the decomposition, and identify the bias, variance, and noise terms.**

Ans.

$$E(x) = E\left[\left(f(x) - \hat{f}(x)\right)^2\right]$$

$$= E[\left(f(x) - E[\hat{f}(x)] + E[\hat{f}(x)] - \hat{f}(x)\right)^2]$$

$$= E\left[\left(f(x) - E[\hat{f}(x)]\right)^2\right] + E\left[\left(E[\hat{f}(x)] - [\hat{f}(x)]\right)^2\right] + Noise$$

$$= Bias + Variance + Noise$$

As we know, there always exists a break-even between the Bias and Variance, and even if we manage to obtain zero Bias and Variance, we would still not be able to reduce the Noise, hence this is also called as "Irreducible Error".

Bonus (15 pts)

Implement boosted decision trees, using the same parameters as for bagging (i.e. same depth limit, same number of trees).

Experiment 1:

Bagged

	1	2	3	4	5	6	7	8	9	10	Means	STD
0.025	0.4450	0.3700	0.2850	0.3850	0.3600	0.2550	0.3400	0.3550	0.3050	0.3250	0.3425	0.0512
0.05	0.1250	0.2800	0.2100	0.2850	0.2600	0.2600	0.2500	0.2500	0.3250	0.2300	0.2475	0.0505
0.125	0.1350	0.1600	0.2050	0.1550	0.2200	0.1200	0.1850	0.2100	0.1350	0.0900	0.1615	0.0406
0.25	0.2550	0.1400	0.1950	0.1800	0.1200	0.2000	0.0950	0.2000	0.1850	0.1750	0.1745	0.0434

Random

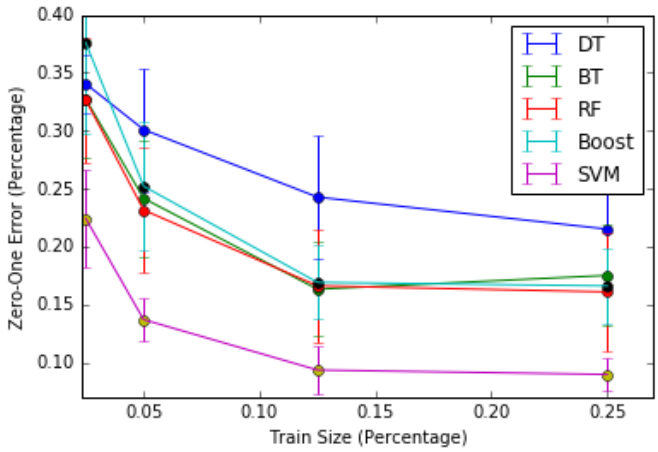
	1	2	3	4	5	6	7	8	9	10	Means	STD
0.025	0.3900	0.3400	0.2900	0.4150	0.4000	0.2450	0.3100	0.3950	0.2950	0.3350	0.3415	0.0540
0.05	0.1150	0.2450	0.2250	0.2700	0.2550	0.2150	0.2100	0.2500	0.3450	0.2300	0.2360	0.0544
0.125	0.0750	0.1400	0.2350	0.1600	0.2200	0.1550	0.1550	0.2100	0.1900	0.1000	0.1640	0.0487
0.25	0.2700	0.1450	0.1700	0.1550	0.1250	0.1850	0.0650	0.1950	0.1400	0.1300	0.1580	0.0508

SVM

	1	2	3	4	5	6	7	8	9	10	Means	STD
0.025	0.2700	0.2650	0.2350	0.2200	0.3000	0.2150	0.1450	0.2500	0.2000	0.1900	0.2290	0.0426
0.05	0.1150	0.1150	0.1500	0.1250	0.1550	0.1350	0.1400	0.1700	0.1100	0.1300	0.1345	0.0185
0.125	0.0650	0.0800	0.0900	0.0800	0.0850	0.1250	0.0700	0.1050	0.0500	0.1100	0.0860	0.0213
0.25	0.0500	0.0900	0.0850	0.0900	0.0650	0.0950	0.1000	0.0850	0.0850	0.0800	0.0825	0.0140

Boosted

	1	2	3	4	5	6	7	8	9	10	Means	STD
0.025	0.48	0.48	0.38	0.38	0.34	0.32	0.34	0.48	0.28	0.28	0.3760	0.07933
0.05	0.25	0.23	0.28	0.17	0.21	0.28	0.34	0.33	0.23	0.2	0.2520	0.05533
0.125	0.14	0.14	0.14	0.21	0.14	0.2	0.18	0.14	0.19	0.21	0.1690	0.03178
0.25	0.21	0.14	0.19	0.18	0.19	0.12	0.2	0.12	0.16	0.15	0.1660	0.03272



From the above plot, we can see that the Boosted tree performs very like the Random Forest in terms of its predictive accuracy.

Experiment 2:

Decision

	1	2	3	4	5	6	7	8	9	10	Means	STD
200	0.2450	0.1950	0.2400	0.2550	0.2750	0.2650	0.2650	0.2900	0.2350	0.2850	0.2550	0.0266
500	0.2600	0.2300	0.2400	0.2000	0.1650	0.2000	0.2900	0.2150	0.2650	0.2850	0.2350	0.0385
1000	0.2500	0.3000	0.2300	0.2000	0.2250	0.1950	0.1850	0.2950	0.2600	0.2300	0.2370	0.0376
1500	0.2000	0.2100	0.2600	0.2700	0.2350	0.2100	0.2800	0.2000	0.2400	0.2150	0.2320	0.0281

Bagged

	1	2	3	4	5	6	7	8	9	10	Means	STD
200	0.1550	0.1800	0.2500	0.2150	0.2300	0.1700	0.1750	0.1900	0.1500	0.2000	0.1915	0.0307
500	0.2000	0.1800	0.1900	0.1950	0.1700	0.1300	0.1150	0.1100	0.1850	0.2400	0.1715	0.0392
1000	0.1400	0.1900	0.2150	0.1800	0.1850	0.1250	0.1450	0.1600	0.1100	0.1400	0.1590	0.0311
1500	0.1950	0.2000	0.1250	0.2000	0.1700	0.1650	0.1950	0.1550	0.1900	0.1550	0.1750	0.0239

Random

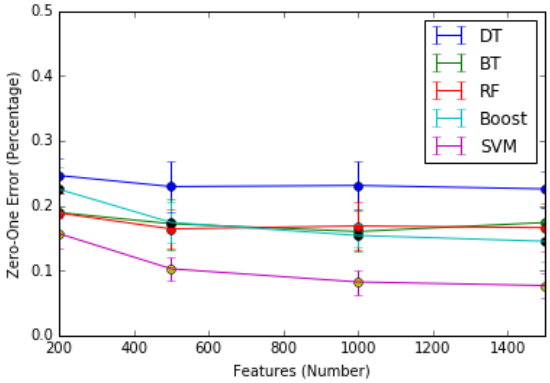
	1	2	3	4	5	6	7	8	9	10	Means	STD
200	0.1350	0.1750	0.2400	0.1950	0.2250	0.2000	0.1800	0.2100	0.1450	0.2000	0.1905	0.0312
500	0.1950	0.1900	0.2100	0.1700	0.1550	0.1300	0.1650	0.1000	0.1450	0.1750	0.1635	0.0310
1000	0.1400	0.2300	0.2300	0.1800	0.1850	0.1500	0.1300	0.1700	0.1450	0.1200	0.1680	0.0368
1500	0.2000	0.1850	0.1000	0.2200	0.1250	0.1400	0.1550	0.1600	0.2150	0.1500	0.1650	0.0373

SVM

	1	2	3	4	5	6	7	8	9	10	Means	STD
200	0.1200	0.1450	0.1850	0.1850	0.1750	0.1400	0.1300	0.1350	0.1600	0.1850	0.1560	0.0239
500	0.0600	0.1050	0.1000	0.1050	0.1000	0.1150	0.1150	0.0650	0.0900	0.1100	0.0965	0.0184
1000	0.0250	0.0900	0.0900	0.0600	0.0800	0.0750	0.0850	0.0650	0.0850	0.0850	0.0740	0.0189
1500	0.0400	0.0500	0.0600	0.0800	0.0950	0.0700	0.0950	0.0450	0.0800	0.0650	0.0680	0.0186

Boosted

	1	2	3	4	5	6	7	8	9	10	Means	STD
200	0.24	0.22	0.24	0.16	0.24	0.25	0.26	0.17	0.2	0.27	0.2250	0.03528
500	0.14	0.21	0.17	0.11	0.21	0.2	0.18	0.2	0.17	0.15	0.1740	0.03136
1000	0.16	0.17	0.12	0.14	0.15	0.14	0.17	0.17	0.14	0.18	0.1540	0.018
1500	0.1	0.17	0.14	0.18	0.15	0.15	0.1	0.12	0.14	0.2	0.1450	0.03106



From the plot, we see that even in this case, the Boosted Tree performs like the Random Forest.

Experiment 3:

Decision

	1	2	3	4	5	6	7	8	9	10	Means	STD
200	0.2550	0.2800	0.2100	0.2650	0.2650	0.2400	0.2800	0.2350	0.1900	0.2550	0.2475	0.0279
500	0.1850	0.2400	0.2300	0.2400	0.1900	0.1600	0.1100	0.2950	0.2250	0.2150	0.2090	0.0482
1000	0.2950	0.2250	0.1600	0.2150	0.2050	0.1550	0.1750	0.2950	0.1950	0.2100	0.2130	0.0465
1500	0.1600	0.2200	0.2600	0.1900	0.2350	0.3250	0.1600	0.2400	0.2100	0.1850	0.2185	0.0476

Bagged

	1	2	3	4	5	6	7	8	9	10	Means	STD
200	0.2350	0.2750	0.2000	0.3050	0.2650	0.3100	0.3250	0.1300	0.2250	0.1550	0.2425	0.0627
500	0.1600	0.1500	0.1600	0.1350	0.1950	0.1600	0.0750	0.2450	0.2000	0.2050	0.1685	0.0439
1000	0.0900	0.1450	0.1150	0.1450	0.1650	0.1450	0.1350	0.1200	0.1500	0.1600	0.1370	0.0216
1500	0.1400	0.1350	0.1500	0.1550	0.1200	0.1450	0.1000	0.1450	0.1700	0.1400	0.1400	0.0182

Random

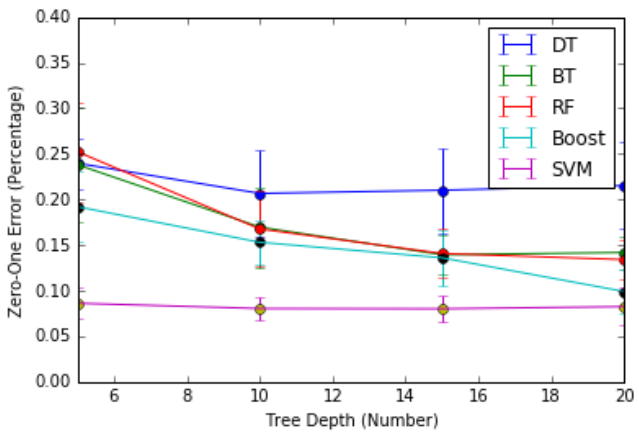
	1	2	3	4	5	6	7	8	9	10	Means	STD
200	0.2200	0.2250	0.1750	0.3100	0.2750	0.3350	0.3250	0.2900	0.2550	0.1800	0.2590	0.0547
500	0.1500	0.1650	0.1400	0.1600	0.1900	0.1400	0.0800	0.2300	0.2000	0.2100	0.1665	0.0410
1000	0.0900	0.1400	0.1400	0.1750	0.1600	0.1300	0.1050	0.1800	0.1300	0.1300	0.1380	0.0269
1500	0.1350	0.1300	0.1400	0.1450	0.0900	0.1600	0.0900	0.1450	0.1400	0.1350	0.1310	0.0219

SVM

	1	2	3	4	5	6	7	8	9	10	Means	STD
200	0.0600	0.0850	0.0700	0.0750	0.0950	0.0900	0.1100	0.0500	0.0850	0.0650	0.0785	0.0170
500	0.0650	0.0800	0.0750	0.0550	0.0600	0.0650	0.0700	0.1000	0.0800	0.0750	0.0725	0.0121
1000	0.0650	0.0800	0.0850	0.0450	0.0900	0.0600	0.0550	0.0900	0.0800	0.0700	0.0720	0.0147
1500	0.0300	0.0550	0.0850	0.0650	0.0850	0.1000	0.0650	0.1050	0.0700	0.0800	0.0740	0.0210

Boosted

	1	2	3	4	5	6	7	8	9	10	Means	STD
200	0.21	0.19	0.14	0.17	0.18	0.14	0.24	0.21	0.26	0.18	0.1920	0.0391
500	0.2	0.13	0.14	0.17	0.16	0.14	0.12	0.16	0.14	0.17	0.1530	0.02359
1000	0.14	0.14	0.12	0.17	0.15	0.14	0.1	0.18	0.14	0.08	0.1360	0.02988
1500	0.11	0.07	0.1	0.09	0.12	0.1	0.08	0.1	0.15	0.07	0.0990	0.02424



From the above plot, we notice that the Boosted Tree consistently performs slightly better than the Random Forest Model.

Experiment 4:

Bagged

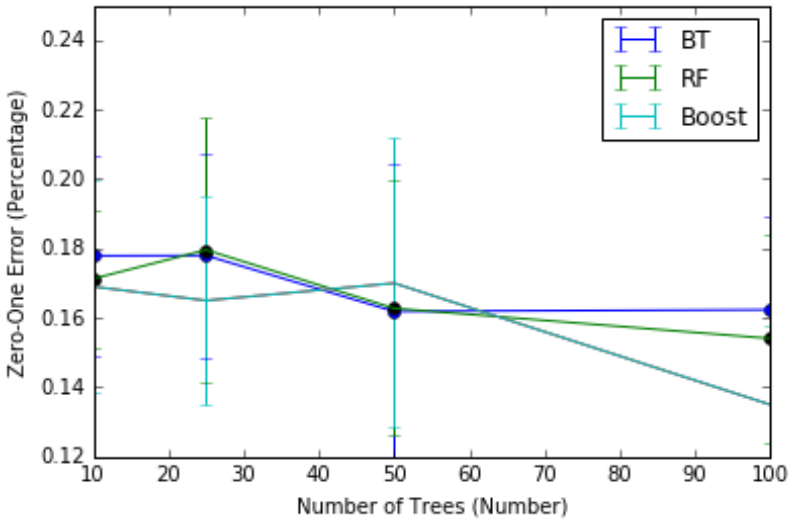
	1	2	3	4	5	6	7	8	9	10	Means	STD
10	0.1700	0.1550	0.1600	0.1700	0.1900	0.1800	0.2500	0.1400	0.1750	0.2000	0.1790	0.0287
25	0.1850	0.1750	0.1900	0.1650	0.1900	0.2300	0.2100	0.1150	0.1550	0.1750	0.1790	0.0296
50	0.1650	0.1950	0.0800	0.1750	0.1800	0.1600	0.2400	0.1050	0.1550	0.1450	0.1600	0.0424
100	0.1700	0.1900	0.1250	0.1350	0.2050	0.1950	0.1500	0.1400	0.1700	0.1400	0.1620	0.0266

Random

	1	2	3	4	5	6	7	8	9	10	Means	STD
10	0.1650	0.1700	0.1550	0.1700	0.1700	0.1750	0.2250	0.1450	0.1700	0.1800	0.1725	0.0199
25	0.1700	0.1850	0.1700	0.1450	0.2150	0.2200	0.2150	0.1000	0.1550	0.2250	0.1800	0.0381
50	0.1150	0.1950	0.1150	0.1700	0.1550	0.1650	0.2400	0.1250	0.1800	0.1550	0.1615	0.0367
100	0.1650	0.2100	0.1150	0.1250	0.1850	0.1700	0.1250	0.1200	0.1650	0.1450	0.1525	0.0300

Boosted

	1	2	3	4	5	6	7	8	9	10	Means	STD
10	0.17	0.21	0.22	0.19	0.17	0.12	0.16	0.16	0.14	0.15	0.1690	0.03071
25	0.15	0.2	0.2	0.13	0.11	0.17	0.18	0.15	0.19	0.17	0.1650	0.0299
50	0.17	0.17	0.14	0.12	0.16	0.18	0.17	0.27	0.19	0.13	0.1700	0.041633
100	0.12	0.12	0.11	0.15	0.15	0.14	0.17	0.16	0.13	0.1	0.1350	0.02273



Include boosting results in all the experiments above. Formulate at least two hypotheses w.r.t. your boosting results:

(1) compare boosting to SVM:

Ans.

Hypothesis:

H_0: The average performance of the Boosted Tree classifier is the same as the SVM classifier for different Train Sizes.

H_1: The average performance of the Boosted Tree classifier is less that the SVM classifier for different Train Sizes.

Paired T-Test and CI: C1, C8

```
Paired T for C1 - C8

             N     Mean   StDev  SE Mean
C1          10   0.2290  0.0450   0.0142
C8          10   0.3760  0.0793   0.0251
Difference  10  -0.1470  0.0658   0.0208

95% upper bound for mean difference: -0.1089
T-Test of mean difference = 0 (vs < 0): T-Value = -7.07  P-Value = 0.000
```

Paired T-Test and CI: C2, C9

```
Paired T for C2 - C9

             N     Mean   StDev  SE Mean
C2          10   0.1345  0.0195   0.0062
C9          10   0.2520  0.0553   0.0175
Difference  10  -0.1175  0.0486   0.0154

95% upper bound for mean difference: -0.0893
T-Test of mean difference = 0 (vs < 0): T-Value = -7.64  P-Value = 0.000
```

Paired T-Test and CI: C3, C10

```
Paired T for C3 - C10

             N     Mean   StDev  SE Mean
C3          10   0.0860  0.0225   0.0071
C10         10   0.1690  0.0318   0.0100
Difference  10  -0.0830  0.0354   0.0112

95% upper bound for mean difference: -0.0625
T-Test of mean difference = 0 (vs < 0): T-Value = -7.40  P-Value = 0.000
```

Paired T-Test and CI: C4, C11

```
Paired T for C4 - C11

             N     Mean   StDev  SE Mean
C4          10   0.0825  0.0148   0.0047
C11         10   0.1660  0.0327   0.0103
Difference  10  -0.0835  0.0415   0.0131

95% upper bound for mean difference: -0.0594
T-Test of mean difference = 0 (vs < 0): T-Value = -6.36  P-Value = 0.000
```

Case	0.025	0.05	0.125	0.25
p-Value	0	0	0	0

From the Paired T-Tests, we see that the P-Value is less than 0.05 (95% confidence) for all cases, and the null hypothesis is rejected for all 4 cases. Hence, by majority vote, we reject the null hypothesis in favor of the alternate that the performance SVM is indeed better than that of a Boosted Tree.

(2) compare boosting to one of the other ensembles. Discuss how the observed data support the hypothesis.

Ans.

Hypothesis:

H_0: The average performance of the Boosted Tree classifier is the same as the Random Forest classifier for different Train Sizes.

H_1: The average performance of the Boosted Tree classifier is less that the Random Forest classifier for different Train Sizes.

Paired T-Test and CI: C1, C8

```
Paired T for C1 - C8

             N     Mean   StDev  SE Mean
C1          10   0.3415  0.0570   0.0180
C8          10   0.3760  0.0793   0.0251
Difference  10  -0.0345  0.0713   0.0225

95% upper bound for mean difference: 0.0068
T-Test of mean difference = 0 (vs < 0): T-Value = -1.53  P-Value = 0.080
```

Paired T-Test and CI: C2, C9

```
Paired T for C2 - C9

             N     Mean   StDev  SE Mean
C2          10   0.2360  0.0573   0.0181
C9          10   0.2520  0.0553   0.0175
Difference  10  -0.0160  0.0898   0.0284

95% upper bound for mean difference: 0.0361
T-Test of mean difference = 0 (vs < 0): T-Value = -0.56  P-Value = 0.293
```

Paired T-Test and CI: C3, C10

```
Paired T for C3 - C10

             N     Mean   StDev  SE Mean
C3          10   0.1640  0.0514   0.0162
C10         10   0.1690  0.0318   0.0100
Difference  10  -0.0050  0.0679   0.0215

95% upper bound for mean difference: 0.0344
T-Test of mean difference = 0 (vs < 0): T-Value = -0.23  P-Value = 0.411
```

Paired T-Test and CI: C4, C11

```
Paired T for C4 - C11

             N     Mean   StDev  SE Mean
C4          10   0.1580  0.0536   0.0169
C11         10   0.1660  0.0327   0.0103
Difference  10  -0.0080  0.0642   0.0203

95% upper bound for mean difference: 0.0292
T-Test of mean difference = 0 (vs < 0): T-Value = -0.39  P-Value = 0.351
```

Case	0.025	0.05	0.125	0.25
p-Value	0.08	0.0284	0.411	0.351

From the p-values, we can reject the null-hypothesis in one case (0.05 Train Size), but fail to reject the null in 3 cases. Hence, we fail to reject the null hypothesis.