CS 573 Data Mining Homework 4

By: Parag Guruji, pguruji@purdue.edu

Date: 15 Apr 2017 – USING TWO (2nd & 3rd) LATE DAYS

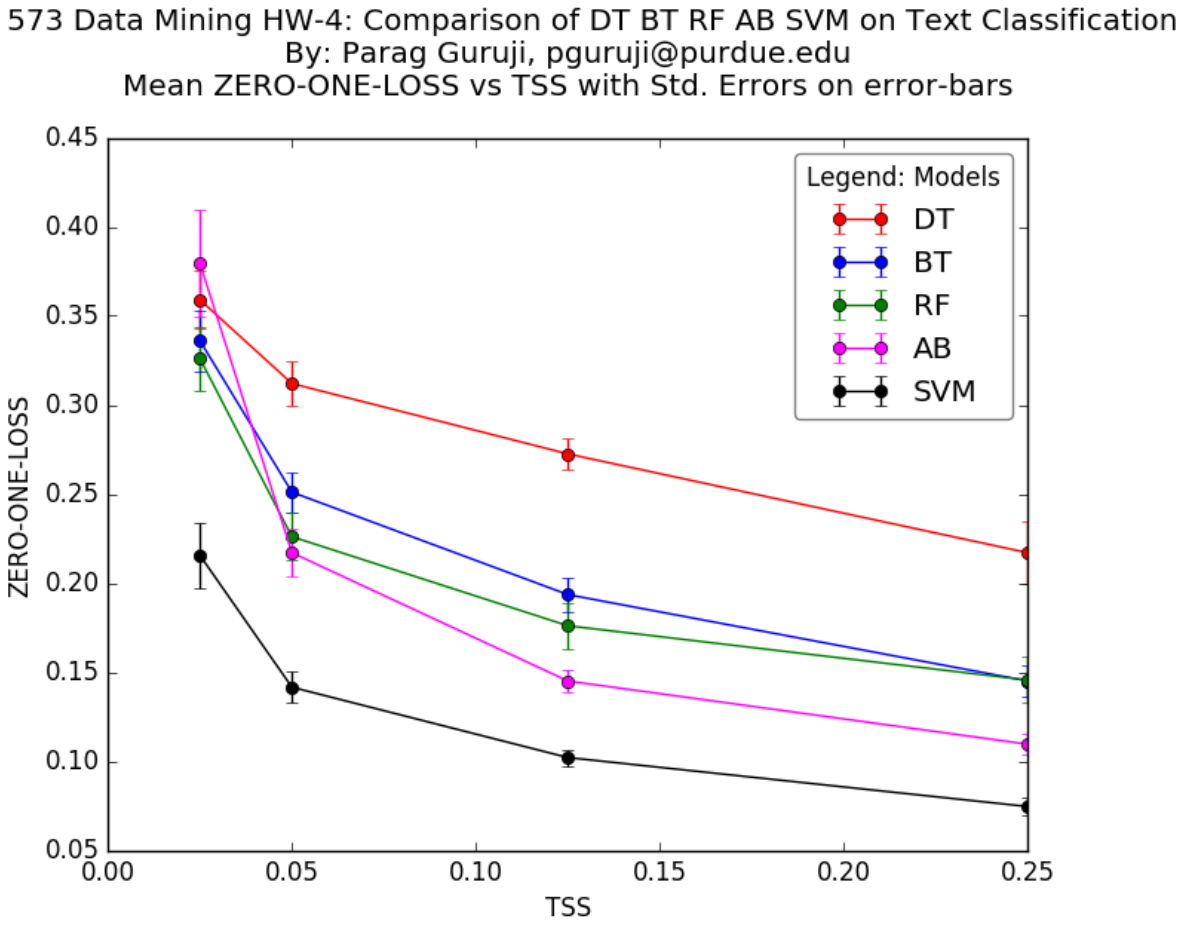
Model	Short-form used
Single Decision Tree	DT
Bagging Ensemble of Decision Trees	BT
Random Forest Ensemble of Decision Trees	RF
Adaptively Boosted (Ada-Boost) Decision Tree	AB
Support Vector Machine	SVM

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.

Training Set Size (Proportion of Whole Data) - 0.025, 0.05, 0.125, 0.25 CV Folds: 10

Feature Count: 1000, Max Tree Depth: 10, #Trees/Ensemble: 50,



Discussion:
The learning curves clearly show a decrease in the zero-one loss as the training set size increases, indicating that the classifiers learn better with higher TSS. We also observe that the Ensembles outperform the single decision tree whereas SVM outperforms all other models at all given TSS values. Also, the ensembles learn faster than single D-tree with increasing training set size, as shown by the steep fall in their learning curves.

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).

Let μBT refer to mean zero-one loss of the Bagging Decision Tree Classifier and μSVM refer to the mean zero-one loss of the SVM Classifier.

Null Hypothesis – $H0: \mu BT = \mu SVM$
Alternative Hypothesis – $H1: \mu SVM < \mu BT$

From the graph in (a) above, we see that the Bagging Decision Tree Classifier has a higher $0/1$ loss for all training set sizes compared to SVM. This difference is significant because the standard error bars of SVM do not overlap and are sufficiently far away from that of the Bagging Decision Tree Classifier.

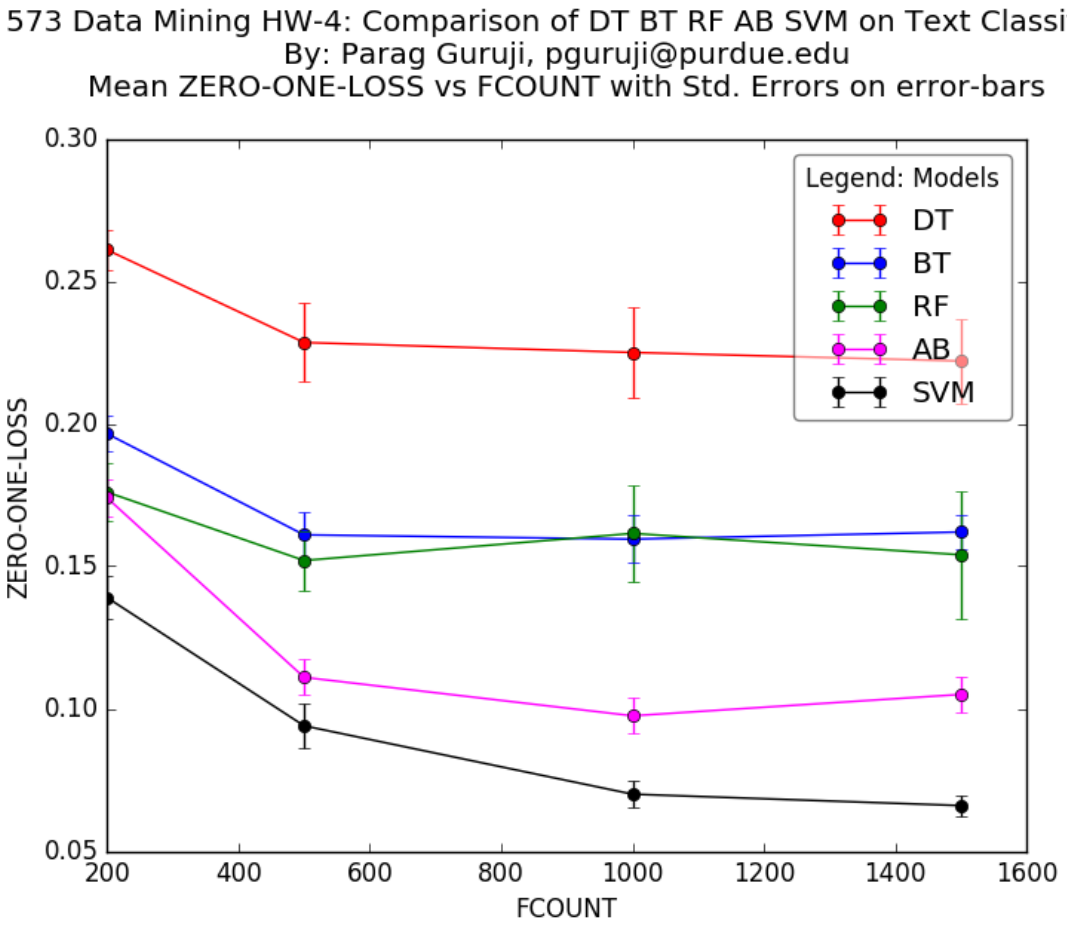
Alternatively, we can perform a one-tailed two sample t-test to compare the means of zero one losses of BT & SVM for each training set size. We will choose our significance $\alpha = 0.05$. To correct for testing multiple hypotheses, we apply Bonferroni's correction. We reject the null hypothesis if the one tailed p-value i.e. 0.5(two-tailed p-value) is less than $\alpha/4 = 0.0125$.

TSS	BT MEAN	BT ERROR	SVM MEAN	SVM ERROR	T	P (2-tailed)	P/2 < α/4
0.025	0.336	0.016923	0.2155	0.018089	4.86447	0.000126	TRUE
0.05	0.251	0.011375	0.1415	0.008833	7.602975	7.36E-07	TRUE
0.125	0.1935	0.009355	0.102	0.004539	8.799501	7.70E-07	TRUE
0.25	0.145	0.008544	0.0745	0.00517	7.059714	4.16E-06	TRUE

From the above table, we see that we can reject the null hypothesis for all TSS values.

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.

Training Set Size (Proportion of Whole Data): 0.25 Feature Count: 200, 500, 1000, 1500 Max Tree Depth: 10, #Trees/Ensemble: 50, CV Folds: 10



Discussion:
Initial rise in feature count improves performance for all models. SVM continues to gain from increase in feature count, however, the improvement decays. The Decision Tree shows only marginal gain in performance after the feature count of 500 and becomes nearly saturated. BT and RF show tiny loss and gains as the feature count increases. Ada-boost continues to gain till 1000 features after which it performance starts to fall.

(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.

Let μRF refer to mean zero-one loss of the Random Forest Tree Classifier and μSVM refer to the mean zero-one loss of the SVM Classifier.

Null Hypothesis – $H0: \mu RF = \mu SVM$
Alternative Hypothesis – $H1: \mu SVM < \mu RF$

From the graph in (a) above, we see that the Random Forest Tree Classifier has a higher $0/1$ loss for all values of number of features compared to SVM. This difference is significant because the standard error bars of SVM do not overlap and are sufficiently far away from that of the Random Forest Tree Classifier.

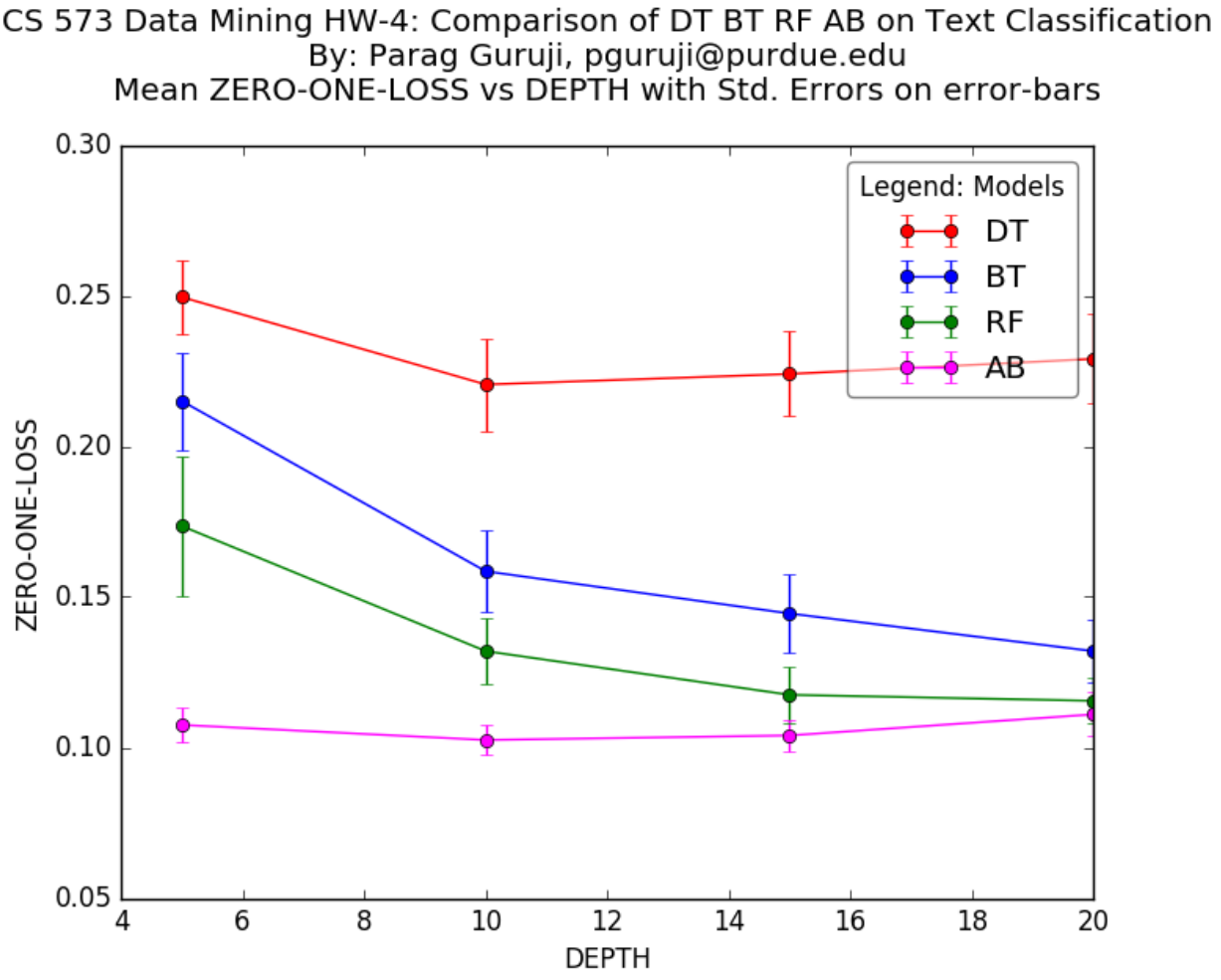
Alternatively, we can perform a one-tailed two sample t-test to compare the means of zero one losses of RF & SVM for each value of number of features used. We will choose our significance $\alpha = 0.05$. To correct for testing multiple hypotheses, we apply Bonferroni's correction. We reject the null hypothesis if the one tailed p-value i.e. 0.5(two-tailed p-value) is less than $\alpha/4 = 0.0125$.

FCOUNT	RF MEAN	RF ERROR	SVM MEAN	SVM ERROR	T	P (2-tailed)	P/2 < α/4
200	0.176	0.01007	0.139	0.007477	2.950104	0.009126	TRUE
500	0.152	0.010397	0.094	0.007576	4.508468	0.000334	TRUE
1000	0.1615	0.016897	0.07	0.004796	5.209259	0.000344	TRUE
1500	0.154	0.022425	0.066	0.003661	3.872858	0.003421	TRUE

From the above table, we see that we can reject the null hypothesis for all Feature Counts.

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.

Training Set Size (Proportion of Whole Data): 0.25 Feature Count: 500
Max Tree Depth: 5, 10, 15, 20 #Trees/Ensemble: 50, CV Folds: 10



Discussion:
The initial increase in depth from 5 to 10 improves all models. After the apparently optimal depth of 10, the performance of single decision tree starts worsening – evidently because, all the true important features are already considered till depth 10 and those which are added after it, hamper more than they help in accurate classification. On the other hand, the ensembles continue to gain from increased number of features as they combine multiple weak tree-learners to generate their prediction.

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

Let μBT refer to mean zero-one loss of the Bagging Decision Tree Classifier and μRF refer to the mean zero-one loss of the Random Forest Decision Tree Classifier.

Null Hypothesis – $H0: \mu BT = \mu RF$
Alternative Hypothesis – $H1: \mu RF < \mu BT$

From the graph in (a) above, we see that the Bagging Decision Tree Classifier has a higher 0/1 loss for all training set sizes compared to RF. This difference is not undoubtedly significant because the standard error bars of RF are very close to that of the Bagging Decision Tree Classifier, even if they do not overlap clearly.

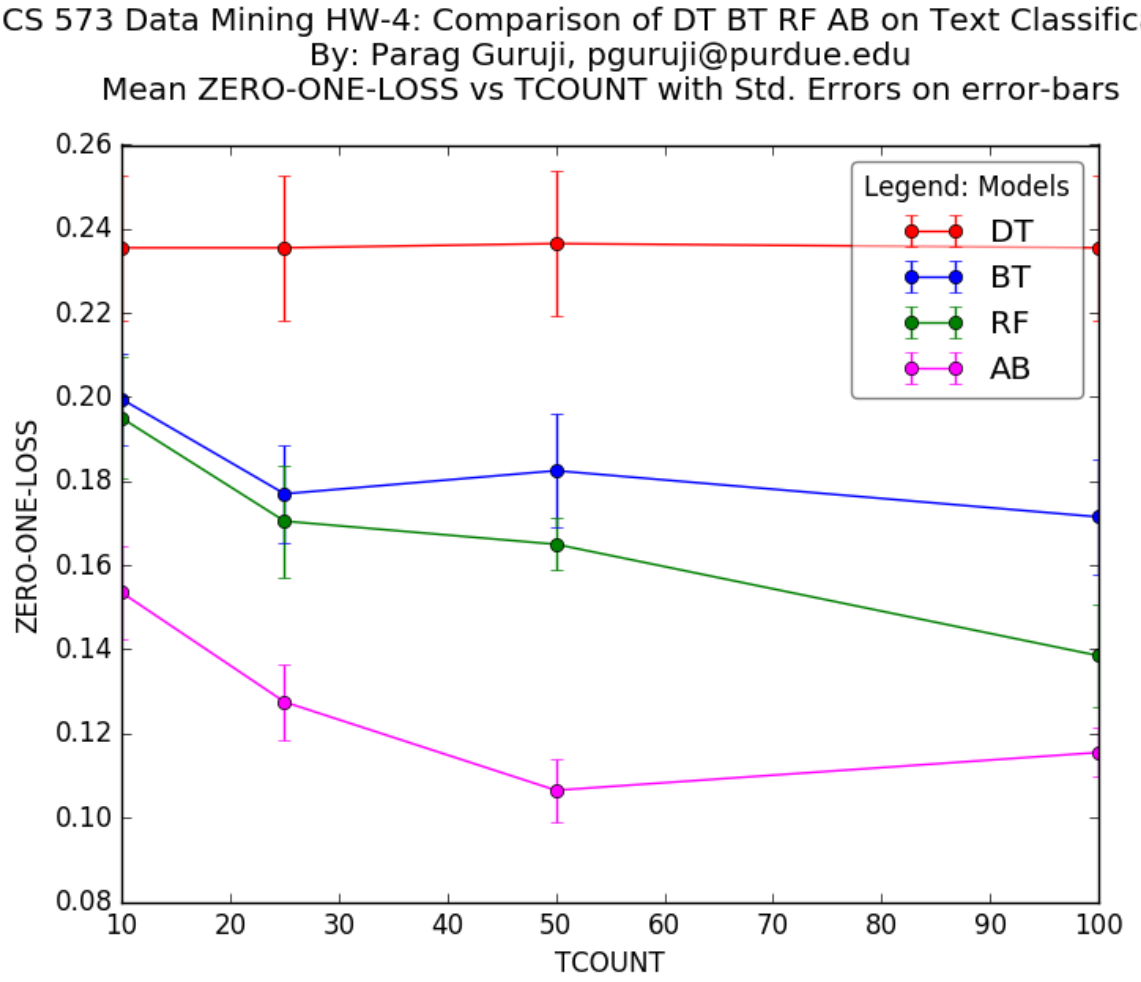
Alternatively, we can perform a one-tailed two sample t-test to compare the means of zero one losses of BT & RF for each value of tree depth. We will choose our significance $\alpha = 0.05$. To correct for testing multiple hypotheses, we apply Bonferroni's correction. We reject the null hypothesis if the one tailed p-value i.e. 0.5(two-tailed p-value) is less than $\alpha/4 = 0.0125$.

DEPTH	BT MEAN	BT ERROR	RF MEAN	RF ERROR	T	P (2-tailed)	P/2 < α/4
5	0.215	0.016047	0.1735	0.023185	1.47183	0.160447	FALSE
10	0.1585	0.013417	0.132	0.010821	1.537362	0.142368	FALSE
15	0.1445	0.013142	0.1175	0.009226	1.681435	0.111926	FALSE
20	0.132	0.010325	0.1155	0.007498	1.293074	0.213875	FALSE

From the above table, we see that we FAIL TO REJECT the null hypothesis for all Depth values.

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.

Training Set Size (Proportion of Whole Data): 0.25 Feature Count: 500
Max Tree Depth: 10 #Trees/Ensemble: 10, 25, 50, 100 CV Folds: 10



Discussion:
Single Decision Tree performance is constant as tree count is irrelevant w.r.t. to it. RF improves faster than the bagging ensemble, which shows a slight loss in performance at 50 trees. The boosting improves significantly faster till it has 50 trees in its ensembles, after which at 100 trees, it performs worse as too many weak learners try to divide data space in more complicated (overfit) decision boundaries.

(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.

Let μDT refer to mean zero-one loss of the Single Decision Tree Classifier and μRF refer to the mean zero-one loss of the Random Forest Decision Tree Classifier.

Null Hypothesis – $H0: \mu DT = \mu RF$
Alternative Hypothesis – $H1: \mu RF < \mu DT$

From the graph in (a) above, we see that the Single Decision Tree Classifier has a higher $0/1$ loss for all training set sizes compared to Random Forest Classifier. This difference is significant because the standard error bars of RF do not overlap and are sufficiently far away from that of the Single Decision Tree Classifier.

Alternatively, we can perform a one-tailed two sample t-test to compare the means of zero one losses of DT & RF for each value of number of trees used. We will choose our significance $\alpha = 0.05$. To correct for testing multiple hypotheses, we apply Bonferroni's correction. We reject the null hypothesis if the one tailed p-value i.e. 0.5(two-tailed p-value) is less than $\alpha/4 = 0.0125$.

TCOUNT	DT MEAN	DT ERROR	RF MEAN	RF ERROR	T	P (2-tailed)	P/2 < α/4
10	0.2355	0.017153	0.195	0.014387	1.809	0.087692	FALSE
25	0.2355	0.017153	0.1705	0.013313	2.993612	0.008183	TRUE
50	0.2365	0.017191	0.165	0.006245	3.909235	0.002305	TRUE
100	0.2355	0.017153	0.1385	0.012043	4.62824	0.000273	TRUE

From the above table, we see that we can reject the null hypothesis for all Tree Counts except for the Tree Count value 10.

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.

Let f(x) be the true function we need to predict, t can be defined as

$$t = f(x) + \in$$

where ∈ be the noise contained in the true value.
Assumption: ∈ follows a normal distribution with mean 0 and variance σ^2.

Let $\hat{f}(x)$ be the predicted value.

Then Squared Error is given by -

SE = (Actual_value – Predicted_value)^2

$$SE = (t - \hat{f}(x))^2 \quad \text{_____ (1)}$$

To prove:

$$MSE = E[(t - \hat{f}(x))^2] = (Bias[\hat{f}(x)]) + Var[\hat{f}(x)] + \sigma^2$$

$$Bias[\hat{f}(x)] = E[\hat{f}(x) - f(x)]$$

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

$$E(\in) = 0 \quad and \quad Var(\in) = \sigma^2$$

Since, $f(x)$ is a deterministic function, $E[f(x)] = f(x)$

So,

$$E[t] = E[f(x) + \in] = E[f(x)] + E[\in] = E[f(x)] = f(x) \quad \text{_____(2)} \because E[\in] = 0$$

$$Var(t) = E[(t - E[t])^2] = E[(t - f(x))^2)] = E[(f(x) + \in - f(x))^2] = E[\in^2]$$

$$Var(t) = Var(\in) + E[\in]^2 = \sigma^2 \quad \text{_____(3)}$$

Now,

MSE

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

$= E[t^2 + \hat{f}(x)^2 - 2E[\, t \times \hat{f}(x)\,]\,]$

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

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

$= \sigma^2 + f(x)^2 + Var\left(\hat{f}(x)\right) + E\left[\hat{f}(x)\right]^2 - 2f(x)E\left[\hat{f}(x)\right]$ _____$from(2\ \&\ 3)$

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

$= \sigma^2 + Var\left(\hat{f}(x)\right) + (E\left[\left(f(x) - \hat{f}(x)\right)\right])^2$ _____$Since\ f(x)\ is\ deterministic$

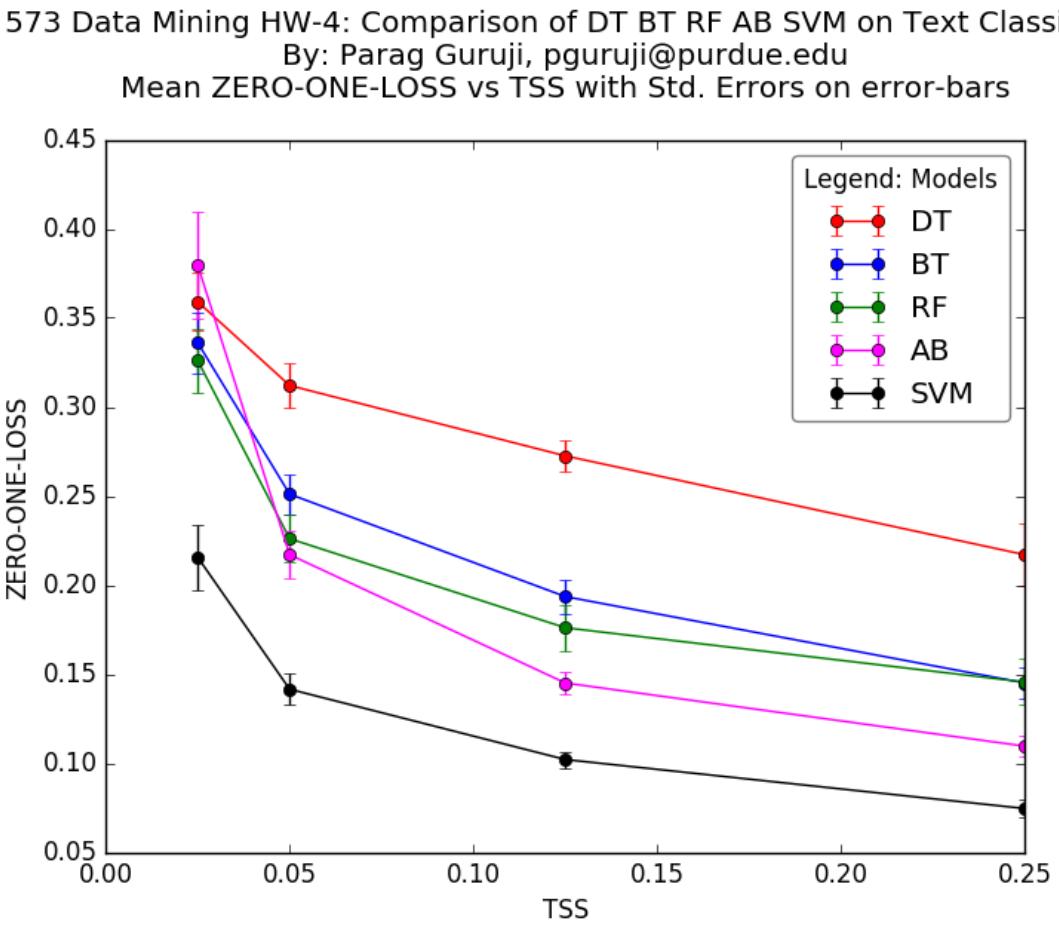
$= Noise^2 + Var\left(\hat{f}(x)\right) + (Bias\left(\hat{f}(x)\right))^2$

Hence Proved.

Bonus

Implement boosted decision trees, using the same parameters as for bagging (i.e. same depth limit, same number of trees). Include boosting results in all the experiments above. Formulate at least two hypotheses w.r.t. your boosting results:
(1) compare boosting to SVMs, and (2) compare boosting to one of the other ensembles. Discuss how the observed data support the hypothesis.

Graph from Problem 1 (a) for quick reference:



(1) Comparing Boosting with SVM for each TSS:

Hypothesis for comparison:
Let µAB refer to mean zero-one loss of the Adaptive Boosting Tree Classifier and µSVM refer to the mean zero-one loss of the SVM Classifier.
Null Hypothesis – $H0$: $\mu AB = \mu SVM$
Alternative Hypothesis – $H1$: $\mu SVM < \mu AB$

From the graph in Problem 1 (a) above, we see that the Adaptive Boosting Tree Classifier has a higher 0/1 loss for all Training Set Sizes compared to SVM. However, the Boosting is converging towards SVM as the TSS grows.
The difference is significant because the standard error bars of SVM do not overlap and are sufficiently far away from that of the Random Forest Tree Classifier in the initial TSS and later, though they come closer, they don't overlap.

Alternatively, we can perform a one-tailed two sample t-test to compare the means of zero one losses of AB & SVM for each Training Set Size. We will choose our significance $\alpha = 0.05$. To correct for testing multiple hypotheses, we apply Bonferroni's correction. We reject the null hypothesis if the one tailed p-value i.e. 0.5(two-tailed p-value) is less than $\alpha/4 = 0.0125$.

TSS	AB MEAN	AB ERROR	SVM MEAN	SVM ERROR	T	P (2-tailed)	P/2 < α/4
0.025	0.3795	0.030162	0.2155	0.018089	4.663025	0.000321	TRUE
0.05	0.217	0.013494	0.1415	0.008833	4.681186	0.000271	TRUE
0.125	0.145	0.006	0.102	0.004539	5.715579	2.67E-05	TRUE
0.25	0.1095	0.005935	0.0745	0.00517	4.446798	0.000325	TRUE

From the above table, we see that we can reject the null hypothesis for all TSS values.

(2) Comparing Boosting with Bagging Tree Ensemble for each TSS:

Hypothesis for comparison:
Let μAB refer to mean zero-one loss of the Adaptive Boosting Tree Classifier and μBT refer to the mean zero-one loss of the Bagging Tree Ensemble Classifier.

Null Hypothesis – $H0: \mu BT = \mu AB$
Alternative Hypothesis – $H1: \mu AB < \mu BT$

From the graph in Problem 1 (a) above, we see that the Adaptive Boosting Tree Classifier has a higher 0/1 loss compared to Bagging Tree Ensemble for the first (smallest) TSS value (with slightly overlapping error bars). However, the boosting learns better at a higher rate as the TSS value increases. At 2nd TSS value, the Boosting nearly matches the BT's performance and continues to perform even increasingly better for 3rd and 4th TSS values and outperforms BT at TSS $= 0.125$ and 0.25.

The apparent reason for this behavior can be that the boosting classifies data by complex decision boundaries which it generates from combining several simple classifiers by assigning importance (weights) to each of the training examples. So, with very little training data, the "true" importance of an example (w.r.t.) population is not well reflected in Boosting Classifier's knowledge representation. Hence, it doesn't perform well. However, as the training set size grows, it has more examples to assign weights and thus better reflection of their "true" importance in its knowledge representation, which results in increasingly better performance.

The difference in mean zero one loss of BT and Boosting is not significant for TSS $= 0.025$ and 0.05 because the standard error bars of both the models overlap. But, for TSS $= 0.125$ and 0.25, the standard error bars of BT and AB do not overlap and are sufficiently far away from that of each other. Hence, the difference in mean zero one loss of BT and Boosting is significant at TSS $= 0.125$ and 0.25.

Alternatively, we can perform a one-tailed two sample t-test to compare the means of zero one losses of AB & BT for each Training Set Size. We will choose our significance $\alpha = 0.05$. To correct for testing multiple hypotheses, we apply Bonferroni's correction. We reject the null hypothesis if the one tailed p-value i.e. 0.5(two-tailed p-value) is less than $\alpha/4 = 0.0125$.

TSS	BT MEAN	BT ERROR	AB MEAN	AB ERROR	T	P (2-tailed)	P/2 < α/4
0.025	0.336	0.016923	0.3795	0.030162	-1.25777	0.228829	FALSE
0.05	0.251	0.011375	0.217	0.013494	1.926415	0.070454	FALSE
0.125	0.1935	0.009355	0.145	0.006	4.363795	0.00053	TRUE
0.25	0.145	0.008544	0.1095	0.005935	3.412436	0.003553	TRUE

From the above table, we see that we FAIL TO REJECT the null hypothesis for TSS $= 0.025$ and 0.05. But, we can REJECT the null hypothesis at for TSS $= 0.125$ and 0.25