Homework 4

CS 57300

(Taking one-day extension)

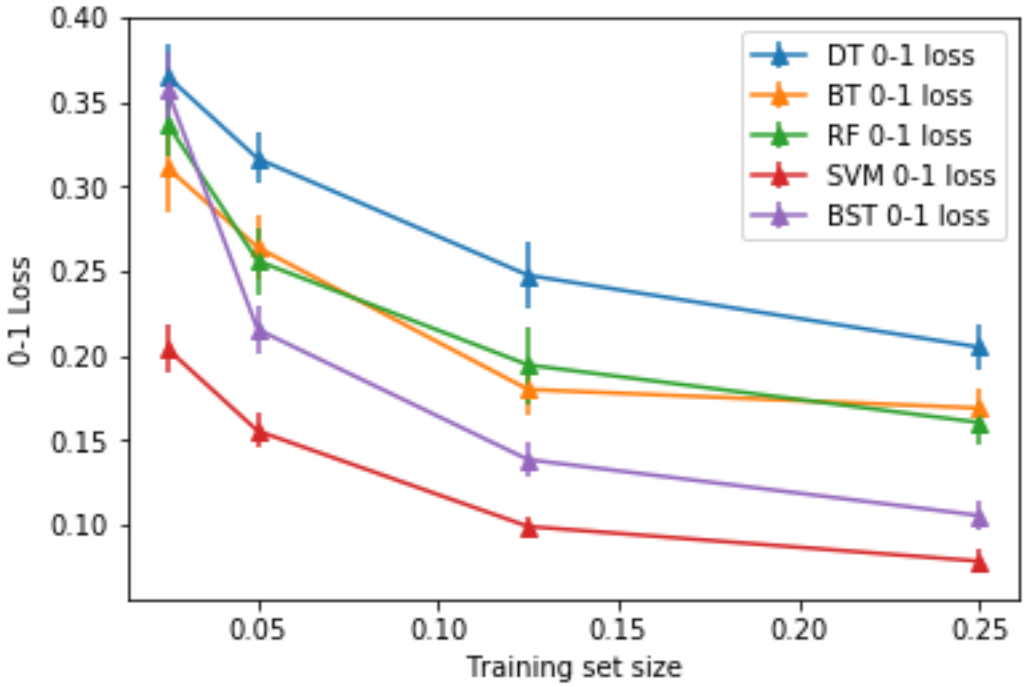
Submitted by:-

Name : Rashmi Soni

PUID: 0029136200

Results include “Boosting” also.

Solution 1a)



Solution 1b)

Performance between one of the ensembles (RF) and SVM

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

Null Hypothesis (H0): μRF = μSVM

Alternative Hypothesis (H1): μRF > μSVM

From the graph, we see that SVM model has a lower 0/1 loss for all training set sizes compared to RF model. This difference is significant because the standard error bars of RF do not overlap and are sufficiently far away from that of the SVM.

Alternatively, we can perform a one-tailed paired t-test. We will choose our significance $\alpha = 0.05$. We can reject the null hypothesis if the p-value is less than $\alpha/4 = 0.0125$.

Training Set percentages	p-Value
0.025	0.000163
0.05	0.000178
0.125	0.001463
0.25	0.000107

From the above table, we can see that we can reject the null hypothesis for all training set percentages. And we can conclude that SVM performs better than RF and the difference between the performances are significant.

Performance between SVM and Boosting:

Let μSVM refer to the mean zero-one loss of the SVM model and μBST refer to the mean zero-one loss of the Boosting model.

Null Hypothesis (H0): μSVM = μBST

Alternative Hypothesis (H1): μBST > μSVM

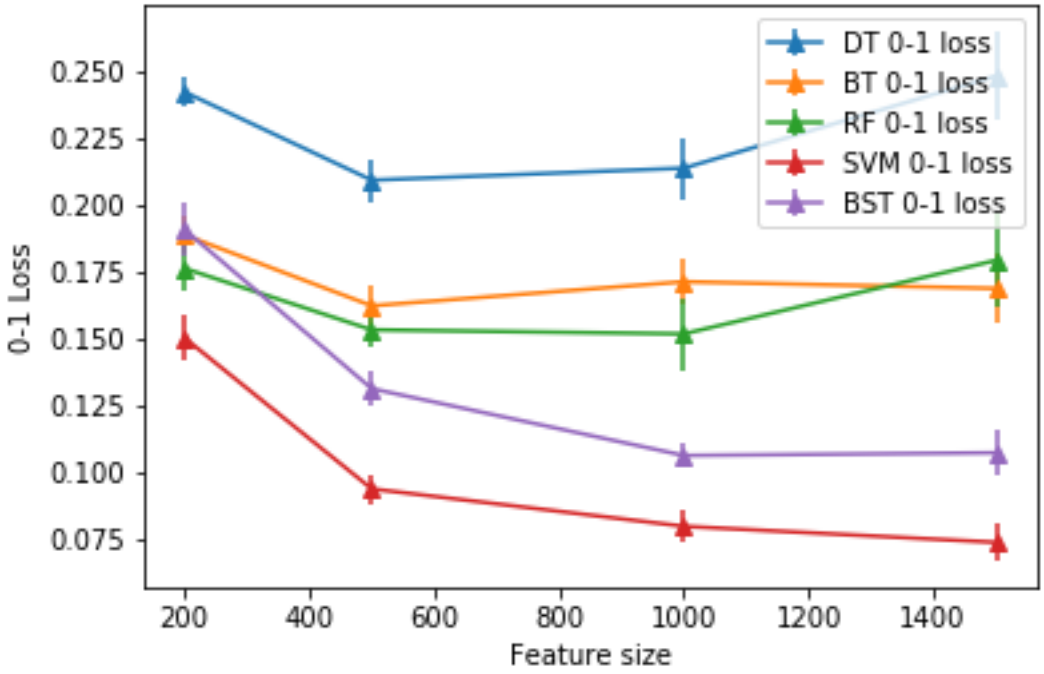
From the graph, we see that SVM model has a lower 0/1 loss for all training set sizes compared to Boosting model. This difference is significant because the standard error bars of Boosting do not overlap and are sufficiently far away from that of the SVM.

Alternatively, we can perform a one-tailed paired t-test. We will choose our significance $\alpha = 0.05$. We can reject the null hypothesis if the p-value is less than $\alpha/4 = 0.0125$.

Training Set percentages	p-Value
0.025	0.000292
0.05	0.000651
0.125	0.000986
0.25	0.002791

From the above table, we can see that we can reject the null hypothesis for all training set percentages. And we can conclude that SVM performs better than Boosting as SVM takes into account all the features while boosting considers only a subset of features. We can also see that since it is a bag of words representation where 'words' are the features so may be boosting could not perform better than SVM.

Solution 2a)



Solution 2b)

Performance between one of the ensembles (RF) and SVM

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

Null Hypothesis (H0): μRF = μSVM

Alternative Hypothesis (H1): μRF > μSVM

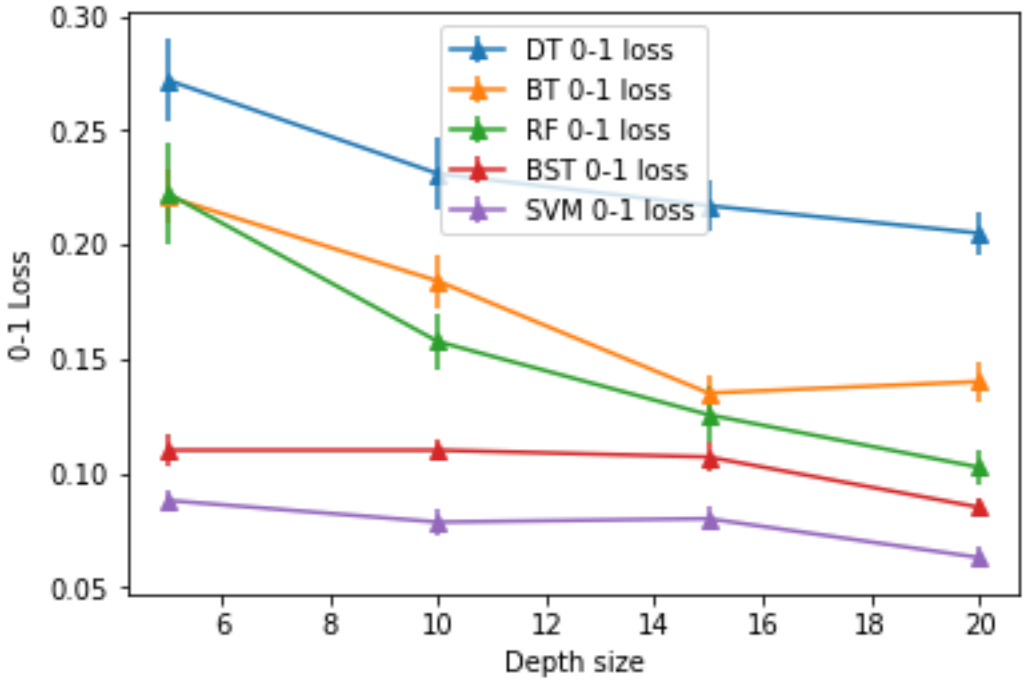
From the graph, we see that SVM model has a lower 0/1 loss for all feature set sizes compared to RF model. This difference is significant because the standard error bars of RF do not overlap and are sufficiently far away from that of the SVM.

Alternatively, we can perform a one-tailed paired t-test. We will choose our significance $\alpha = 0.05$. We can reject the null hypothesis if the p-value is less than $\alpha/4 = 0.0125$.

Number of features	p-Value
200	0.020611
500	1.4E-05
1000	0.000247
1500	0.000169

From the above table, we can see that we can reject the null hypothesis for all sizes of features except for features=200. So, SVM performs better than RF in varying the number of features and the performance between the two models are significant.

Solution 3a)



Solution 3b)

Performance between DT and RF

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

Null Hypothesis (H0): μDT = μRF

Alternative Hypothesis (H1): μDT > μRF

From the graph, we see that RF model has a lower 0/1 loss for all depth limit compared to DT model. This difference is significant because the standard error bars of DT do not overlap and are sufficiently far away from that of the RF.

Alternatively, we can perform a one-tailed paired t-test. We will choose our significance $\alpha = 0.05$. We can reject the null hypothesis if the p-value is less than $\alpha/4 = 0.0125$.

Depth Limit	p-Value
5	0.059212
10	0.002418
15	2.4E-05
20	3.2E-05

From the above table, we can see that we can reject the null hypothesis for all limits of depth except for depth limit=5. So, we can say that RF performs better than DT in varying the depth limit and the performances between the two models are significant.

Performance between BT and Boosting

Let μBT refer to the mean zero-one loss of the Bagging model and μBST refer to the mean zero-one loss of the Boosting model.

Null Hypothesis (H0): μBT = μBST

Alternative Hypothesis (H1): μBT > μBST

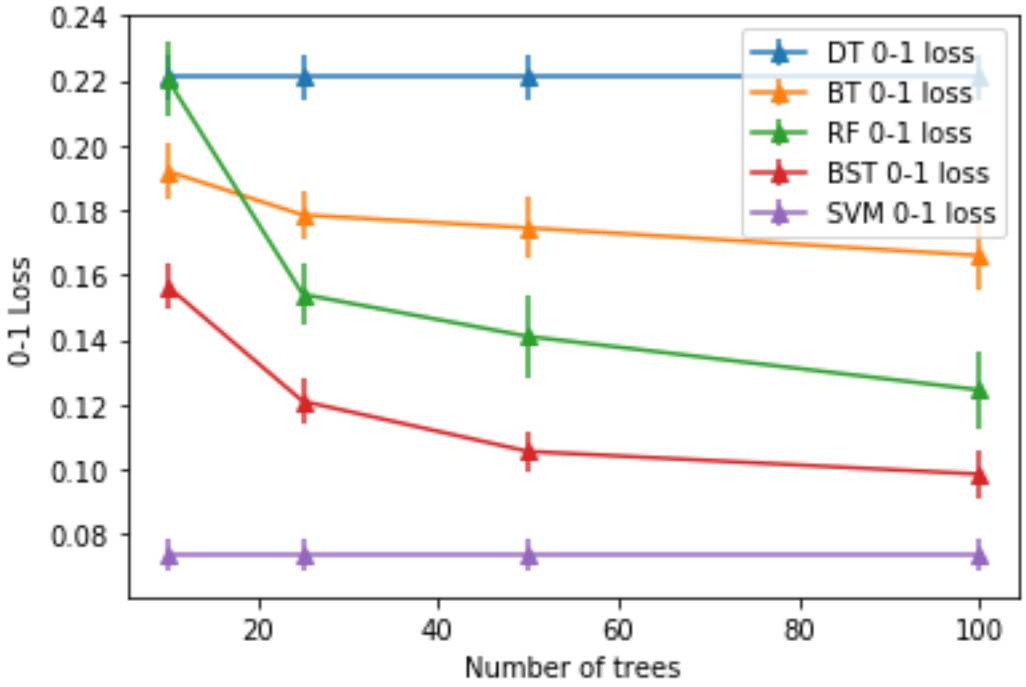
From the graph, we see that BST model has a lower 0/1 loss for all depth limit as compared to BT model. This difference is significant because the standard error bars of BT do not overlap and are far away from that of the BST.

Alternatively, we can perform a one-tailed paired t-test. We will choose our significance $\alpha = 0.05$. We can reject the null hypothesis if the p-value is less than $\alpha/4 = 0.0125$.

Depth Limit	p-Value
5	0.00001
10	0.000151
15	0.008356
20	0.000564

From the above table, we can see that we can reject the null hypothesis for all limits of depth. Here, boosting performs better than BT while varying the depth limit and the performances between the two models are significant.

Solution 4a)



Performance between DT and BT

Let μBT refer to the mean zero-one loss of the Bagging model and μDT refer to the mean zero-one loss of the Decision Tree model.

Null Hypothesis (H0): $\mu BT = \mu DT$

Alternative Hypothesis (H1): $\mu DT > \mu BT$

From the graph, we see that BT model has a lower 0/1 loss for all number of trees as compared to DT model. This difference is significant because the standard error bars of BT do not overlap and are significantly away from that of the DT.

Alternatively, we can perform a one-tailed paired t-test. We will choose our significance $\alpha = 0.05$. We can reject the null hypothesis if the p-value is less than $\alpha/4 = 0.0125$.

Number of trees	p-Value
10	0.013672
25	0.000433
50	0.001049
100	0.001417

From the above table, we can see that we can reject the null hypothesis for all number of trees values except for number of trees=10. So, here we can see that the performance of bagging is better than the single decision tree and the differences are significant.

Solution 5)

We know that for any random variable X, we have the formula:

$Var[X] = E[X^2] – E[X]^2$

Or,

$E[X^2] = Var[X] + E[X]^2$

Say we have a training data with examples x1,x2,…,xn and real values y1,y2,…,yn respectively. Let's say there is a function with noise $y= f(x) + \varepsilon$, where the noise , ε, has zero mean and variance σ^2. Here, f(x) represents the true function. Suppose the predicted value or the function that approximates the true function be g(x).

We want to compute the mean squared error between y and g(x). This can be depicted as:

$E[(y - g(x))^2] = E[y^2 + g(x)^2 - 2\, y\, g(x)]$ ……eq (1)

Since we know that f(x) is a deterministic function, so its expected value should be equal to its true value, i.e.

$E[f(x)] = f(x)$

Since, $y = f(x) + \varepsilon$

$E[y] = E[f(x) + \varepsilon]$

$= E[f(x)]$ {As $E[\varepsilon]=0$}

$= f(x)$ ……eq(2)

$Var[y] = E[(y - E[y])^2]$

$= E[(y - f(x))^2]$ {Using eq(2) }

$= E[(f(x) + \varepsilon - f(x))^2]$

$= E[\varepsilon^2]$

$= Var[\varepsilon] + E[\varepsilon^2]$

$= \sigma^2$ {Since $Var[\varepsilon] = \sigma^2$ and $E[\varepsilon]=0$} ……..eq(3)

Now, from eq(1);

$E[(y - g(x))^2] = E[y^2 + g(x)^2 - 2\, y\, g(x)]$

$= E[y^2] + E[g(x)^2] - 2\, E[y\, g(x)]$

$= Var[y] + E[y]^2 + Var[g(x)] + E[g(x)]^2 - 2\, f(x)\, E[g(x)]$

$= Var[y] + Var[g(x)] + (f(x)^2 - 2\, f(x)\, E[g(x)] + E[g(x)]^2)$

$= Var[y] + Var[g(x)] + (f(x) - E[g(x)])^2$

$= Var[y] + Var[g(x)] + E[f(x) - g(x)]^2$

$$= \sigma^2 + \text{Var}[g(x)] + E[f(x) - g(x)]^2 \quad \text{.....eq(4)}$$

$$= \sigma^2 + \text{Var}[g(x)] + \text{Bias}\,[g(x)]^2$$

Since, $y = f(x) + \varepsilon$, where ε is the noise, then Var[y] signifies the variance of the data which contains noise.

Thus, in eq(4), we have decomposed the expected square loss into bias, variance and noise terms;

where,

σ^2 is the noise terms,

Var[g(x)] is the variance, and

$E[f(x) - g(x)]^2$ is the bias term.