Meng Deng
Dr. Jennifer Neville
CS 573, hw 4
April 16, 2017

Ensembles Analysis

1. Assess whether ensembles improve performance.

   a) Learning Curves of Decision Tree (DT), Bagged Tree (BT), Random Forest (RF), and SVM. The graph shows that in general, as the training sample size increases, all of the three models increase their performance (loss drops). Among them, SVM seems to have highest performance of all time. The performance of RF seems to increase drastically when the sample size become bigger. The standard deviation of the three ensembles seems to be higher than SVM.

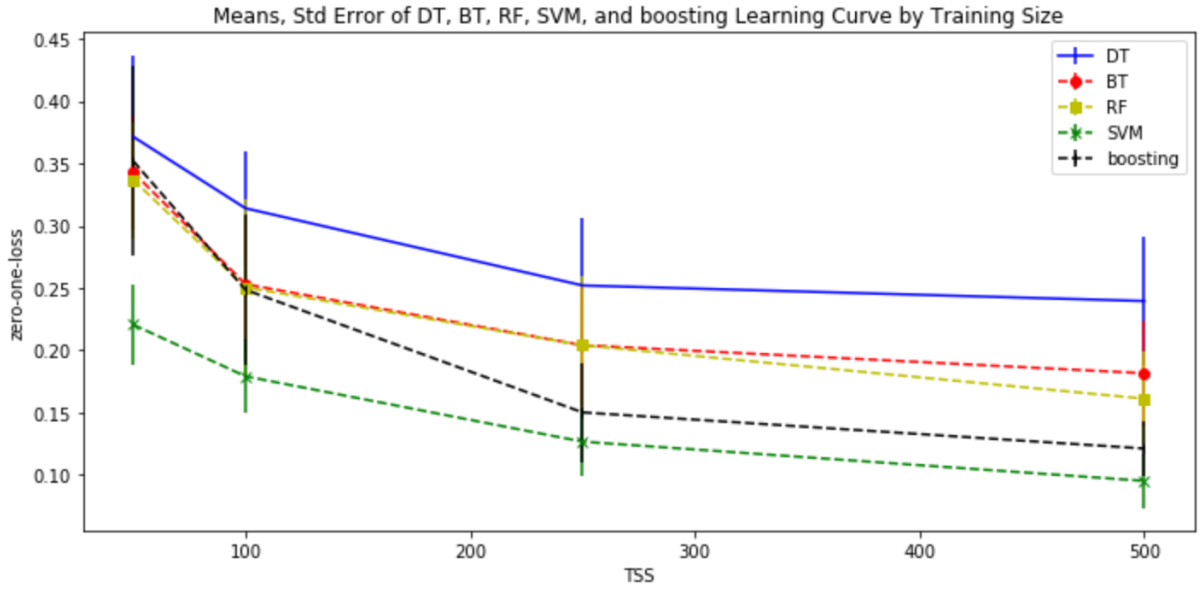

Figure 1: Means, Std Error of Decision Tree, Bagged Tree, Random Forest, and SVM by Training Set Size.

b) H_0: $L_{SVM} = L_{DT}$

H-alternative: $L_{SVM} < L_{DT}$.

From the observed mean and error of the learning curve of SVM and DT, I conclude that SVM performs significantly better than Decision Tree.

2. Assessing whether the number of features affects performance.

a) I don't see the general loss drop trend in Figure 2. SVM seems to drop all the time, while the other three does not seem to change following certain pattern. The std. error of RF seems to grow bigger as the feature number increases.

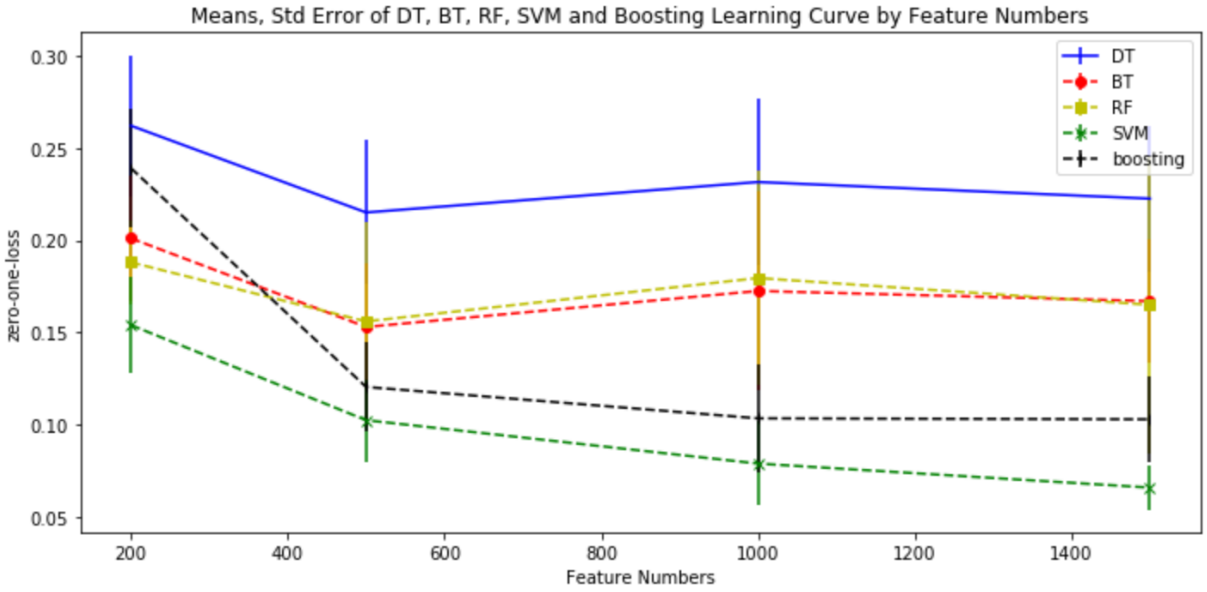

Figure 2: Means, Std Error of Decision Tree, Bagged Tree, Random Forest, and SVM by Feature Numbers.

b) H_0: $L_{SVM} = L_{BT}$

H-alternative: $L_{SVM} < L_{BT}$.

The mean and standard error of the learning curve shows that SVM performs significantly better than BT.

3. Assess the depth of the tree's and the performance.

a). Learning Curves of Decision Tree (DT), Bagged Tree (BT), and Random Forest (RF).

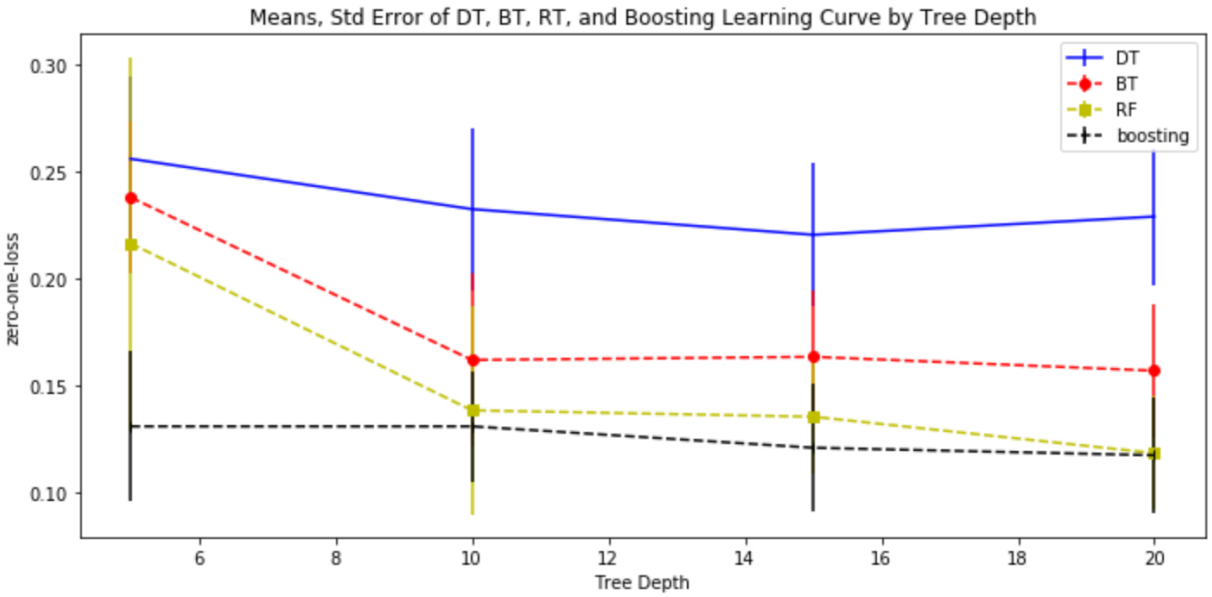

Figure 3: Means, Std Error of Decision Tree, Bagged Tree, and Random Forest by Tree Depth.

b). H0: $L_{BT} = L_{DT}$

H-alternative: $L_{BT} < L_{DT}$.

I observe from the learning curve of figure 3 that BT has significant lower loss than DT.

4. Assess the number of trees and the performance.

a). Figure 4 shows the loss changes of the three models by tree numbers.

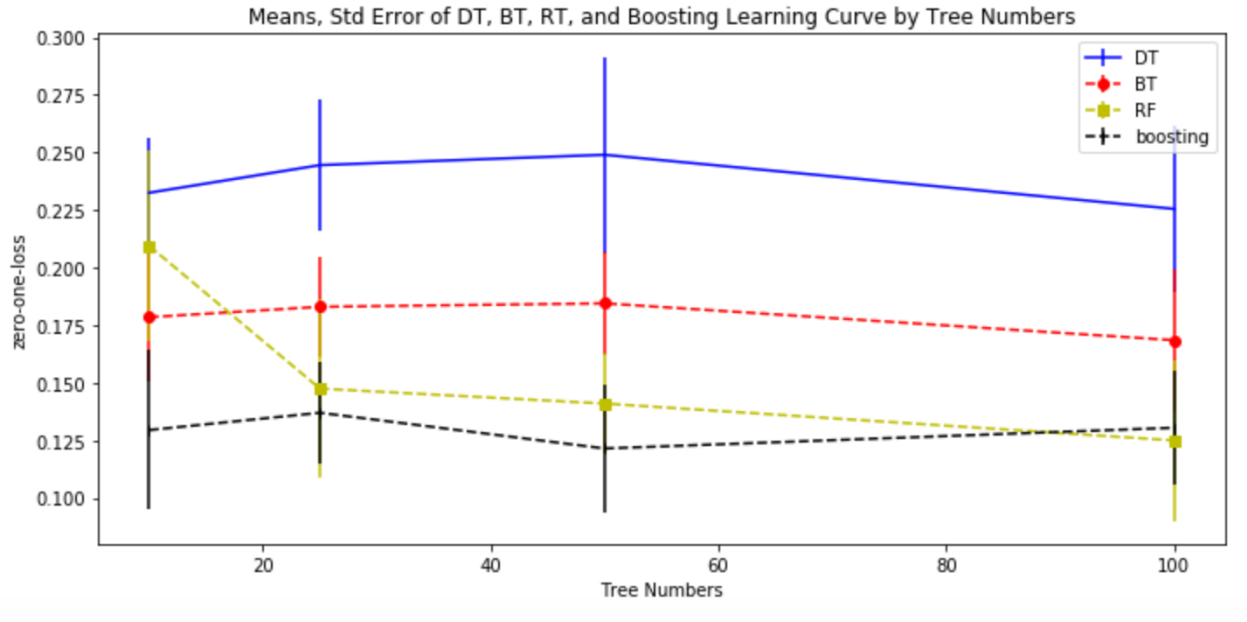

Figure 4: Means, Std Error of Decision Tree, Bagged Tree, and Random Forest by Tree Numbers.

b). H0: $L_{BT} = L_{DT}$

H-alternative: $L_{BT} < L_{DT}$.

I observe from the learning curve of figure 4 that BT has significant lower loss than DT as tree numbers change.

5.

$$E[L_{square}(x, y)] = E[(y - E(y) + E(y) - E(x) + E(x) - x)^2]$$
$$= E[(y - E(y))^2] + 2E[y - E(y)]E[E(y) - E(x)] + E[(E(y) - E(x))^2]$$
$$+ 2E[y - E(y)]E[E(x) - x] + E[(x - E(x))^2] + 2E[E(x) - x]E[E(y) - E(x)] =$$
$$(E(y) - E(x))^2 + E[(y - E(y))^2] + E[(x - E(x))^2]$$

where, $E[y-E(y)] = E[E(x)-x] = 0$. Therefore, we have bias as $(E(y)-E(x))^2$, variance as $E[(y-E(y))^2]$ and noise as $E[(x-E(x))^2]$.