CS584 HOMEWORK 3

Name: Sreeram Bangaru

Miner username: sreerambangaru

G#: G01406291

Given problem statement is to implement Adaboost algorithm using gini index as the error measure for learning decision stumps and to compare the results of train and test error with respect to number of rounds of boosting. Adaboosting typically uses weak learners as models which perform slightly better than random guessing. These weak learners are then combined to form a stronger model that generalizes much better on new data. This is done because individual weak learners are simple and have high bias hence are prone very less to over fitting when compared to other models. Adaboost uses a hyperparameter alpha which controls the contribution of each weak learner to the final ensemble model. It is basically proportional to the accuracy on the training data and inversely proportional to the updated weights in the ensemble model.

In our case, we have used decision stump as the weak classifier and have run Adaboost algorithm for maximum of 200 rounds on both training data and then on test data.

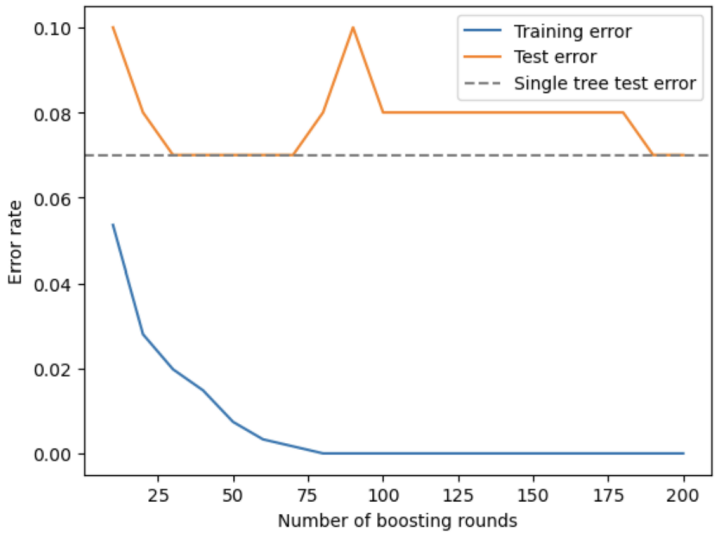
Number of Iterations	**Training Error**	**Test Error**
10	0.053542	0.099999
20	0.028006	0.079999
30	0.019769	0.069999
40	0.014827	0.069999
50	0.007413	0.069999
60	0.003294	0.069999
70	0.001647	0.069999
80	0.0	0.079999
90	0.0	0.079999
100	0.0	0.099999
110	0.0	0.079999
120	0.0	0.079999
130	0.0	0.079999
140	0.0	0.079999
150	0.0	0.079999
160	0.0	0.079999
170	0.0	0.079999
180	0.00	0.079999
190	0.0	0.069999
200	0.0	0.069999

During the iterative process of the Adaboost algorithm we can see that on training data the algorithm is able to learn quickly from the data and a convergence is achieved at 80 iterations and hundred percent accuracy at this point. From this point onwards till two hundred iterations the error rate is constant at zero and did not change hence shows that the model has generalized well on the training data.

But when it comes to generalizing on the test data, we can see that the error initially decreases as the number of rounds of boosting increases, indicating that the algorithm is learning better on test data. However, after a certain number of rounds(around 90 iterations) the test error starts to increase again from which we can infer that as the number of rounds increased, the algorithm lost some of its generalization ability to predict unseen data, causing the accuracy to decrease. After some further rounds, the algorithm may have corrected this situation and stabilized the accuracy at the correct level.

The comparison of the above results with the test error of single decision tree is informative. The error of the single decision tree is constant and hence is a reasonable baseline where its error is 0.07, but we can see in the graph that at the beginning the error of a single decision tree is less than the weak learner at small number of rounds, but eventually Adaboost achieved better performance.

To summarize the concern of overfitting for Adaboost basically depends on the amount of data and the number of iterations we are using to generate weak models.



Citation:

sklearn.ensemble.AdaBoostClassifier — scikit-learn 1.2.2 documentation

https://www.youtube.com/watch?v=wF5t4Mmv5us