**Homework 3**

**COSC 6342: Machine Learning**

**Submitted by**

**S M Salah Uddin Kadir (1800503)**

**Rubayat Jinnah (1891217)**

In this experiment, we are using “Ionosphere” dataset. The dataset has 351 examples, where each example has 34 attributes. All attribute values are continuous numbers.

* Number of examples: 351
* Number of attributes: 34

The target of this dataset are free electrons in the ionosphere. If the returns from the radar, the attributes, shows an evidence of some type of structure in the ionosphere, then we classify that example as a “Good” return otherwise “Bad” return. So, it is a classification problem, and the target has two categorical classes.

* Good
* Bad

We used three estimators to classify the dataset with Bagging and Boosting approach.

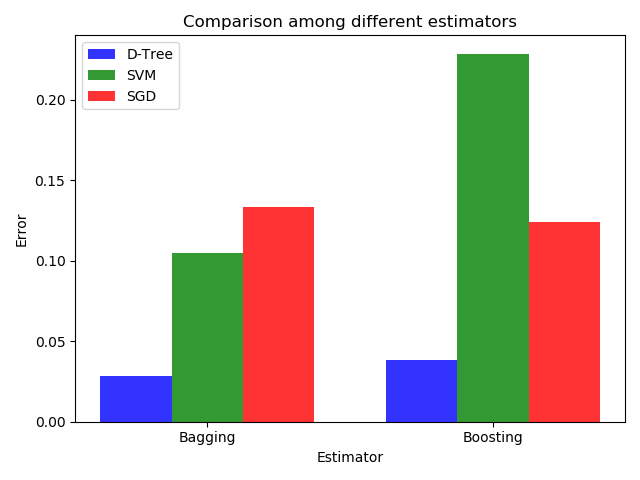
* Support Vector Machine (SVM)
* Decision Tree (d-tree)
* Stochastic gradient descent (SGD)

We used 100 trees for bagging and boosting.

(c)

***General approach (without k-fold):***

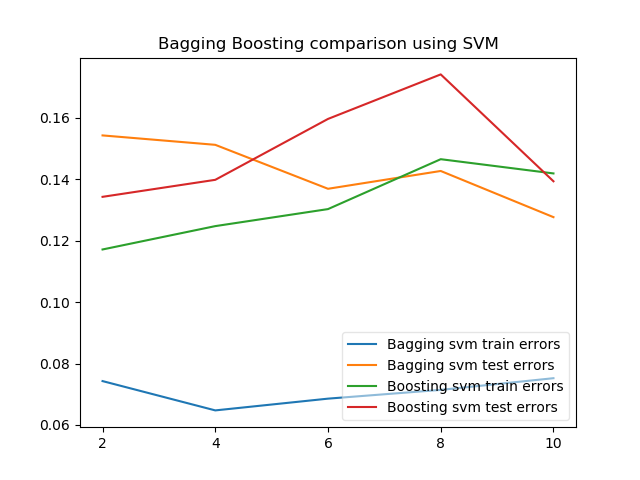
We used 70% of the dataset to train the model and remaining 30% to test the model.



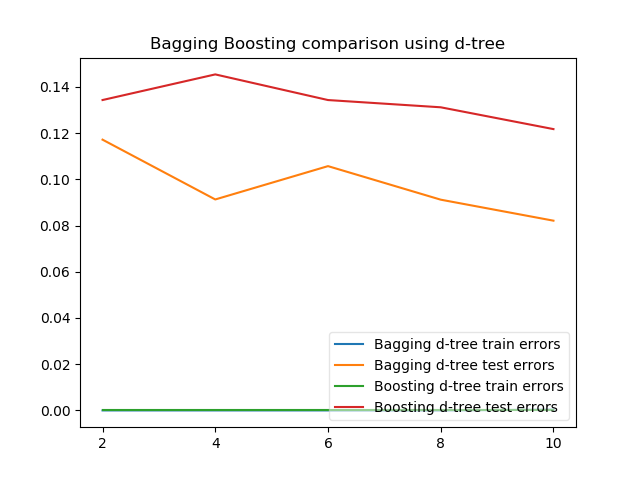
We applied Bagging and Boosting to classify the “Ionoshere” dataset. We used 3 estimators: Decision tree, Support vector machine (SVM) and the Stochastic gradient descent (SGD) algorithms. From the plot, we can see that the error rate of decision tree is lower than the other estimators for both Bagging and Boosting.

***Comparison with k-fold cross-validation:***

We had an experiment with k-fold cross-validation using the same estimators and dataset for the Bagging and Boosting approach. We calculated the test and train error rates with 2, 4, 6, 8, and 10 folds.

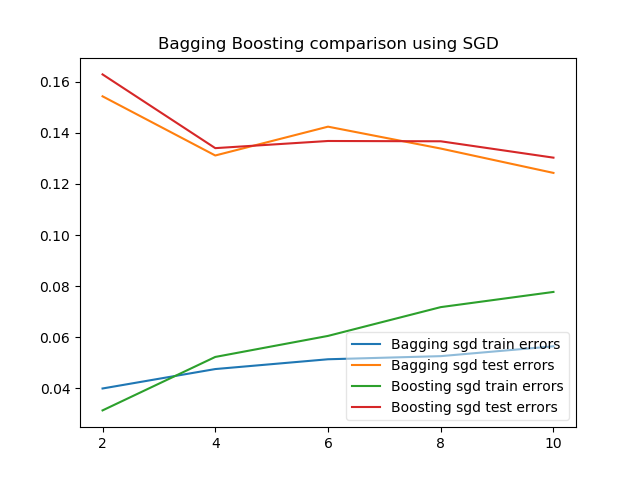
***Comparison using SVM classifier:***

Using SVM classifier, we see that the train error is very low for bagging than the boosting. Although, the test error is similar for both.

***Comparison Using Decision Tree classifier:***

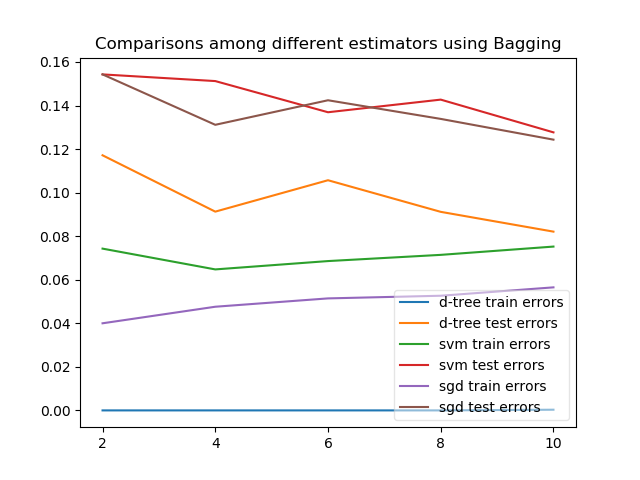
For decision tree, the train error is 0 for both Bagging and Boosting. The test error is relatively high for Boosting.

***Comparison Using SGD classifier:***



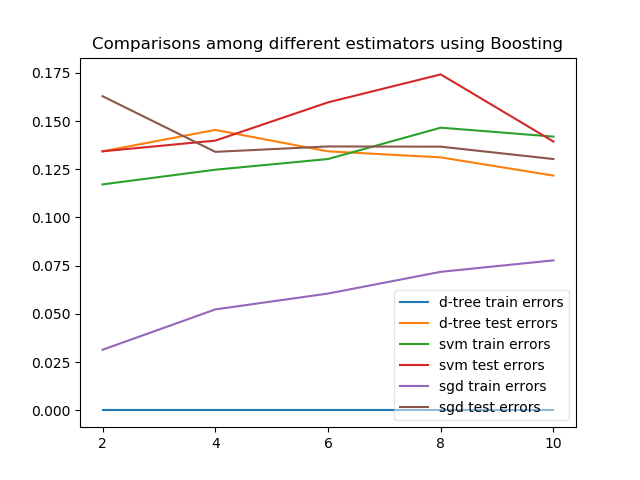
For SGD classifier, we see that the performance of Bagging and Boosting is very close for both training and testing.

***Comparison among all classifiers using only Bagging:***

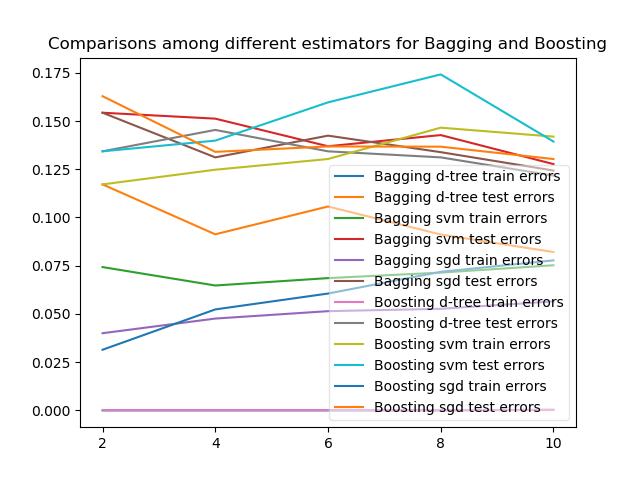


Now, we want to compare among different classifiers using only Bagging. From the plot, we see that the training error is higher for SVM among these three classifiers. We can also notice that ***the d-tree has the best performance for both training and testing among these three classifiers***.

***Comparison among all classifiers using only Boosting:***



There is ***no training error for decision tree.*** Although, the performance on the testing dataset ***does not differ very much*** with each other.

***Comparison among all classifier for both Bagging and Boosting:***

The plot is the combination of all classifiers for both Bagging and the Boosting so that we can compare with each other. We can conclude that ***the performance of Decision tree is relatively better than other classifiers***.

***Complexity Graph with fixed K (6 fold):***

(d)

Bias:

Vairance:

Noise:

Overafitting:

Underfitting:

(e)

(f)

When to choose Boosting over Bagging:

(g)

**Random forest:**

Each individual tree in the random forest splits out a class prediction and the class with the most votes becomes our model’s prediction

* The random forests algorithm applies the general technique of bootstrap aggregating like bagging to tree learners.
* Additionally use “feature bagging”
  + Random subset of features. Eliminate strong features bias. Ensures the correlation of the trees in an ordinary bootstrap sample.

So, random forest ends up with trees that are not only trained on different sets of data like bagging, but also use different features to make decisions. This forces even more variation amongst the trees in the model and ultimately results in lower correlation across trees and more diversification.