**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.

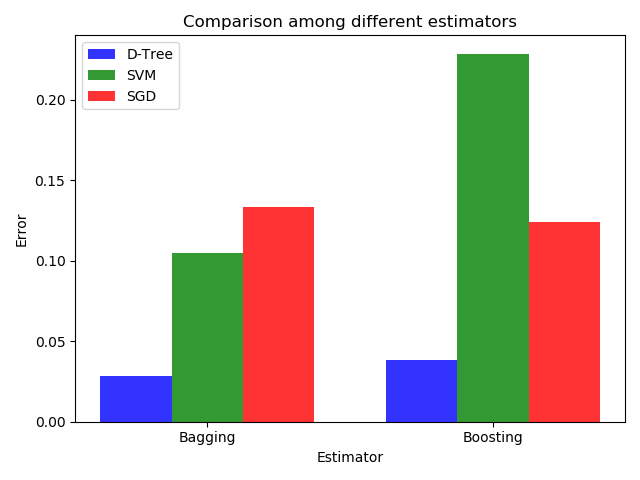
* Number of examples: 351
* Number of attributes: 34

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 60% of the dataset to train the model and remaining 30% to test the model.

(c)



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.

(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.