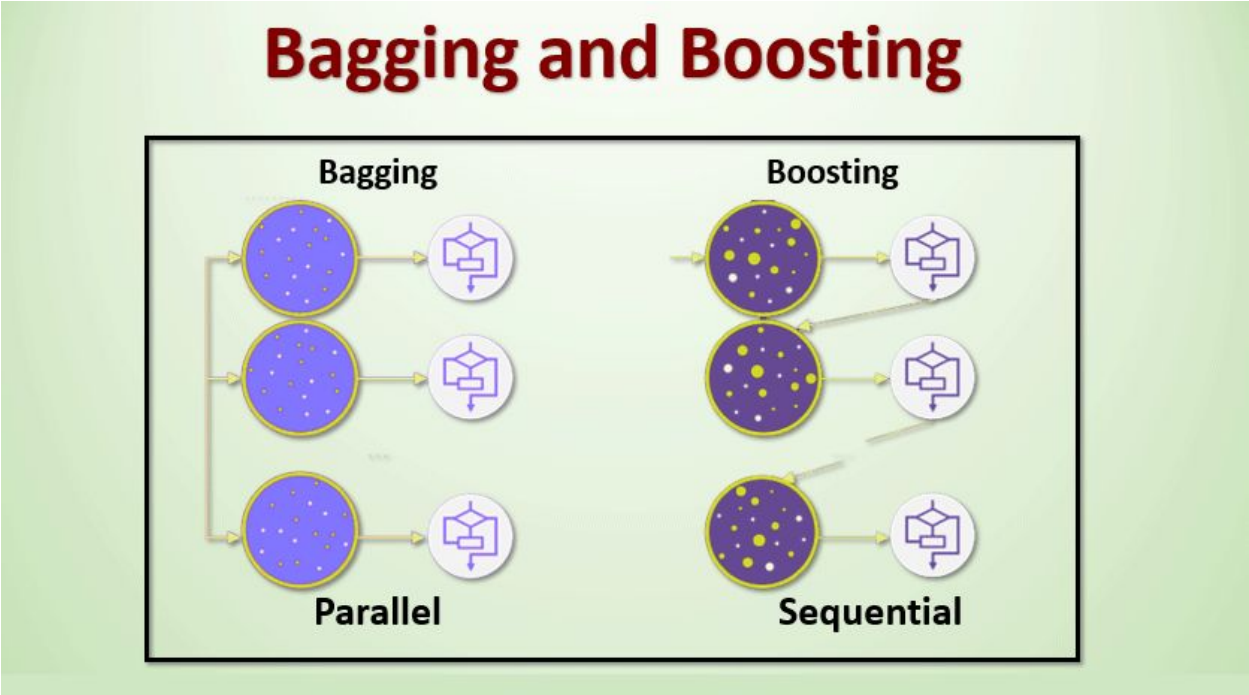
"The future is ours to shape. I feel we are in a race that we need to win. It's a race between the growing power of the technology and the growing wisdom we need to manage it." **~Max Tegmark**



Assignment 3

Bagging and Boosting

Nitesh Yadav(35), Rohit Rana(40), Rohit Shakya(41)
M.Sc. Computer Science
Department of Computer Science,
University of Delhi

Overview

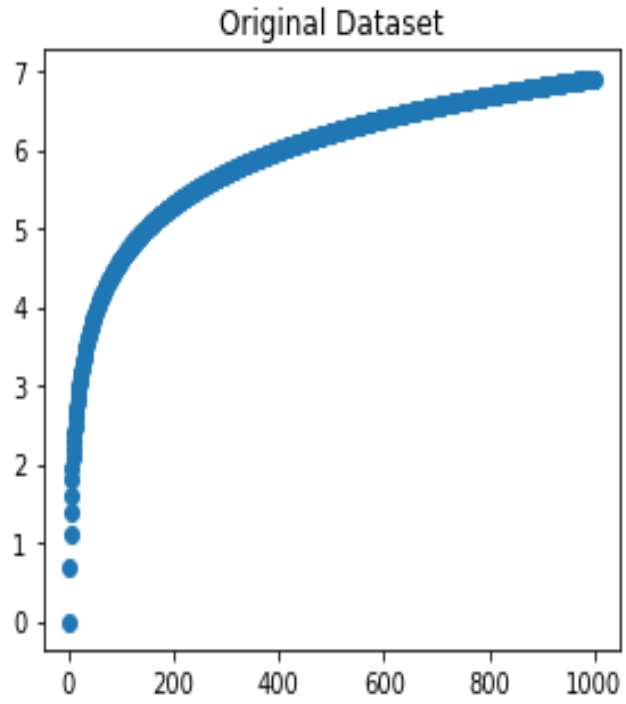
Choose a nonlinear function y=f(x) and generate 1000 data points. Add noise(~N(0,σ^2)) to create a noisy dataset χ. Generate 100 bootstrap samples χ1 to χ100. For each χi fit a linear regression model Mi.

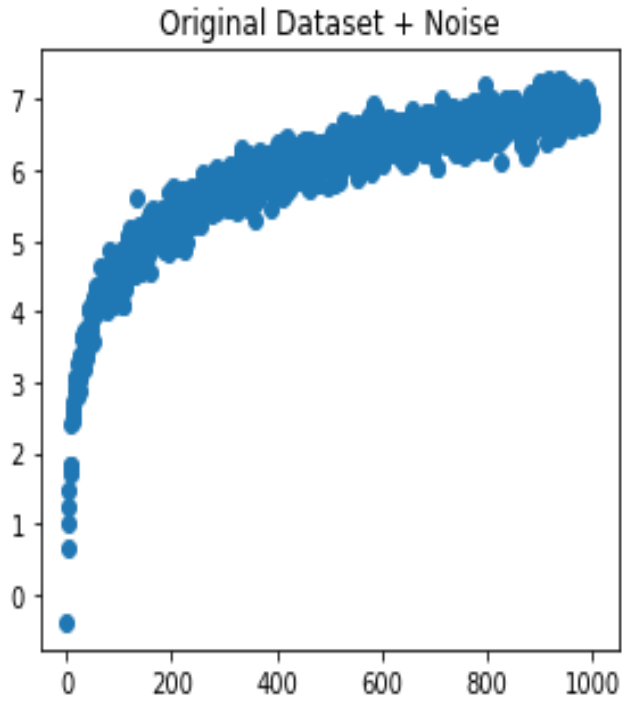
1. For each Mi, find an out-of-bag error and plot prob density curve for the error.
2. Take all unique out-of-bag samples and create the test dataset.
3. Create bagging ensembles of sizes L=10,20,...100 and find bias and variance.
4. Plot B^2 and variance against ensemble size.

About Data

Created dataset using log function (non-linear).

Generated noise using random function along with normal distribution having zero mean and 0.2 variances.





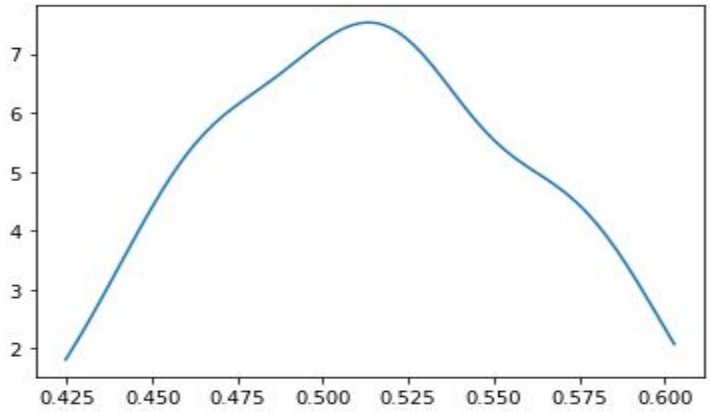
Ensemble Method

Bagging is a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression. It also reduces variance and helps to avoid overfitting.

Observations

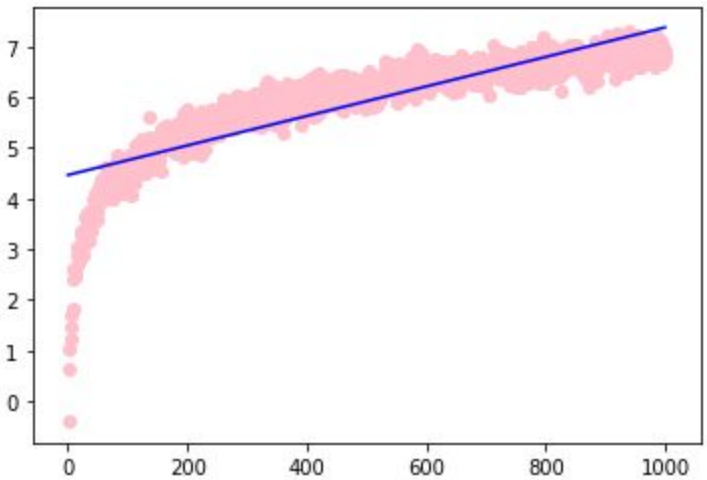
I. Probability Density Curve For The Error

In probability theory, a probability density function, or density, is a function whose value at any given sample in the sample space can be interpreted as providing a relative likelihood that the value of the random variable would equal that sample.



This distribution is a little bit similar to the Gaussian distribution.

II. The best-Fit line for the dataset



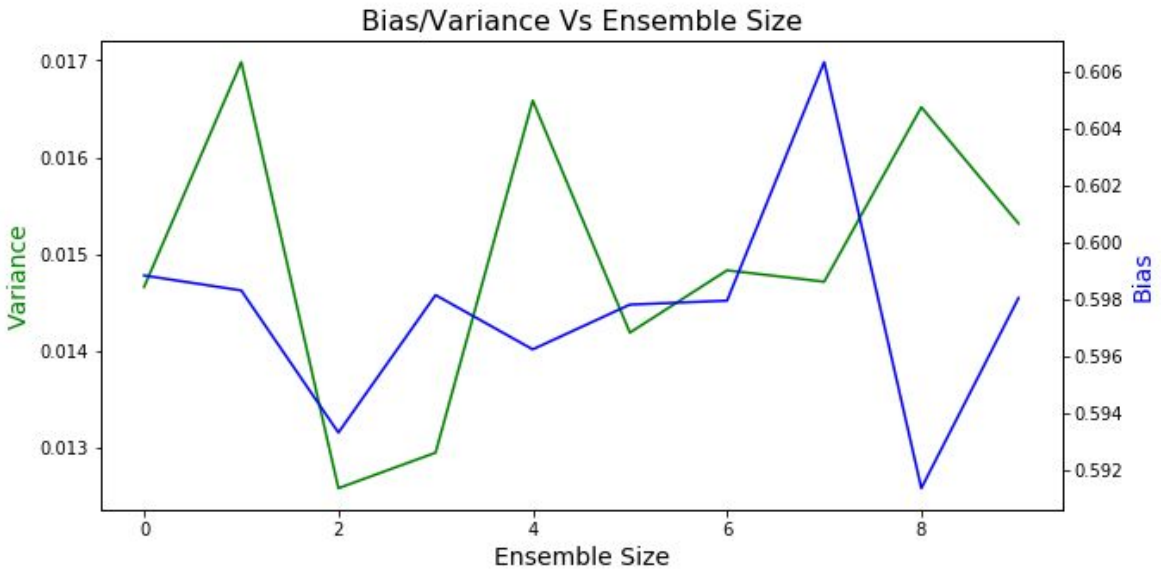
III. Bagging, variance, and error for ensembles of diff sizes

	Bias	Variance	Total_error
0	0.598851	0.014660	0.613511
1	0.598325	0.016982	0.615308
2	0.593329	0.012582	0.605911
3	0.598163	0.012949	0.611111
4	0.596253	0.016585	0.612838
5	0.597827	0.014188	0.612015
6	0.597966	0.014832	0.612798
7	0.606352	0.014714	0.621067
8	0.591374	0.016517	0.607890
9	0.598059	0.015316	0.613374

According to bias-variance decomposition, total error = bias + variance.

In our bagging models, we can clearly observe that it has high bias and low variance which means our model is under fit.

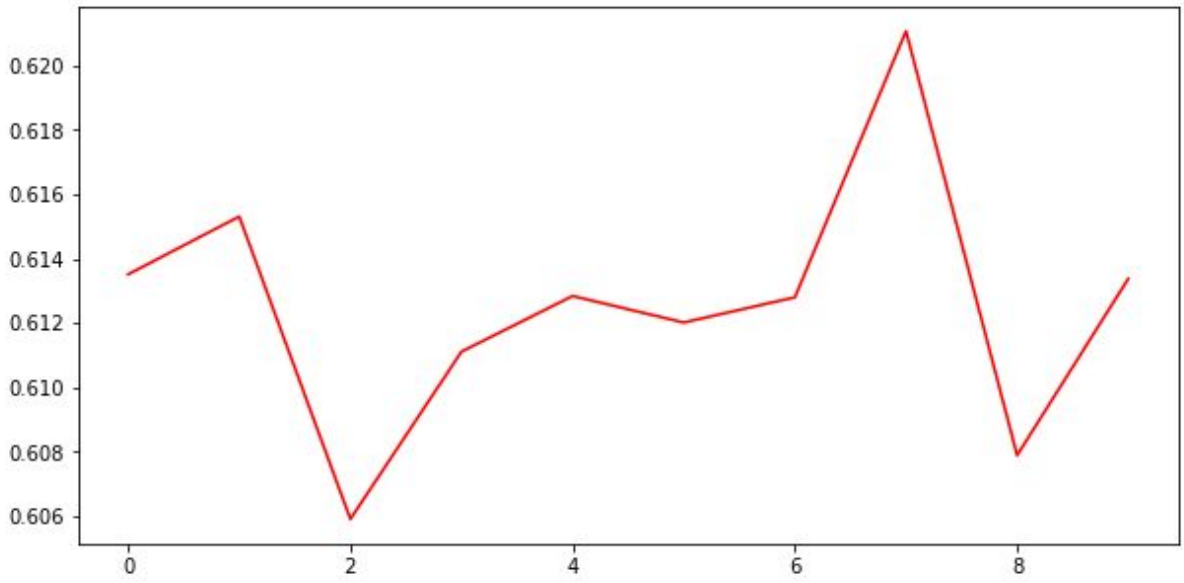
IV. B^2 and variance against ensemble size



We can notice the trend of variance is increasing as the ensemble size is increased.

That means increasing the ensemble size makes our model good (because of high bias and variance).

V. Error against ensemble size



According to bias-variance decomposition, the total error is equal to bias + variance.

And in our model due to lower variance, the total error is quite similar to the biases of the model.

Conclusion

The bag of size 30 gives the best result with the least error.

Due to low variance and high biases our linear regression model is underfit.