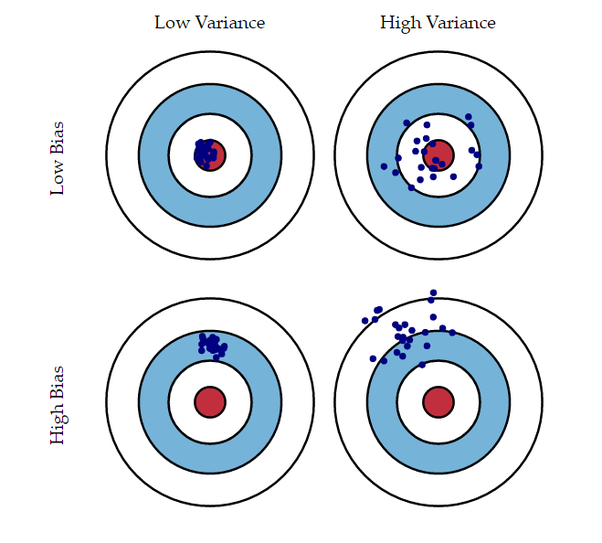
**Error due to Bias**

Error due to bias is the amount by which the expected model prediction differs from the true value of the training data. It is introduced by approximating the complicated model by much simpler model. High bias algorithms are easier to learn but less flexible, due to this they have lower predictive performance on complex problems. Linear algorithms and oversimplified model lead to high bias in the model.

**Error due to variance**

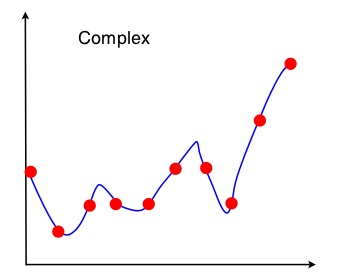
Error due to variance is the amount by which the prediction, over one training set, differs from the expected value over all the training sets. In machine learning, diﬀerent training data sets will result in a diﬀerent estimation. But ideally it should not vary too much between training sets. However, if a method has high variance then small changes in the training data can result in large changes in results.

**Bias and Variance Trade-off**

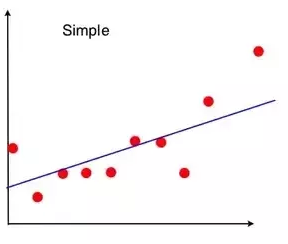


Suppose the red circle is the target which perfectly predicts the correct values. If we move away from that, our prediction gets worse. The above figure displays four different cases representing combinations of both high and low bias and variance. High bias is when all dots are far from bulls eye and high variance is when all dots are scattered.

There is always a trade-off between bias and variance because it is easy to achieve low bias but high variance (complex model)



and low variance and high bias (simple model)



The challenge lies in ﬁnding a method/model which exhibits both low bias and variance.

**What is R-Squared?**

R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression.

The definition of R-squared is fairly straight-forward; it is the percentage of the response variable variation that is explained by a linear model. Or:

R-squared = Explained variation / Total variation

R-squared is always between 0 and 100%:

* 0% indicates that the model explains none of the variability of the response data around its mean.
* 100% indicates that the model explains all the variability of the response data around its mean.

In general, the higher the R-squared, the better the model fits your data. However, there are important conditions for this guideline that I’ll talk about both in this post and my next post.

**Hypothesis Testing**

Hypothesis: A premise or claim that we want to test.

Null Hypothesis: Currently accepted Value. Represented by Ho

Alternate Hypothesis: Also called Research Hypothesis. Involves the claim to be tested. Represented by Ha.

**Note:** Null Hypothesis and Alternate Hypothesis are mathematically opposite

Possible Outcomes of this test:

* Reject Null Hypothesis
* Failed to Reject Null Hypothesis

Eg: Newton has invented gravitational force when apple falls. This is currently everyone believing. So this is null hypothesis. If any researcher will come and tells null hypothesis is false. He has to test to claim null hypothesis is false.

Problem:

In Manufacturing company of Chocolate bars. One employee claims chocolate bars are no more than 5 gm. Here what us H0 and Ha?

H0 : Mean = 5

Ha: Mean <> 5

(Note: To prove this he must perform some test)

Level of confidence(C): How confident are we in our decision. Eg: 95%

Level of significance(α): α= 1-C Eg: α = 1-0.95 = 0.05