Lecture 19

Bias and variance trade-off

Irreducible Error

Errors that cannot be reduced no matter what algorithm you apply is called an irreducible error. It is usually caused by unknown variables that may be having an influence on the output variable.

Reducible Error has two components — **bias and variance.**

Presence of bias or variance causes overfitting or underfitting of data.

Bias

Bias is how far are the predicted values from the actual values. If the **average predicted values are far off from the actual values then the bias is high**.

High bias causes algorithm to miss relevant relationship between input and output variable. When a model has a high bias then it implies that the model is too simple and does not capture the complexity of data thus **underfitting the data**.

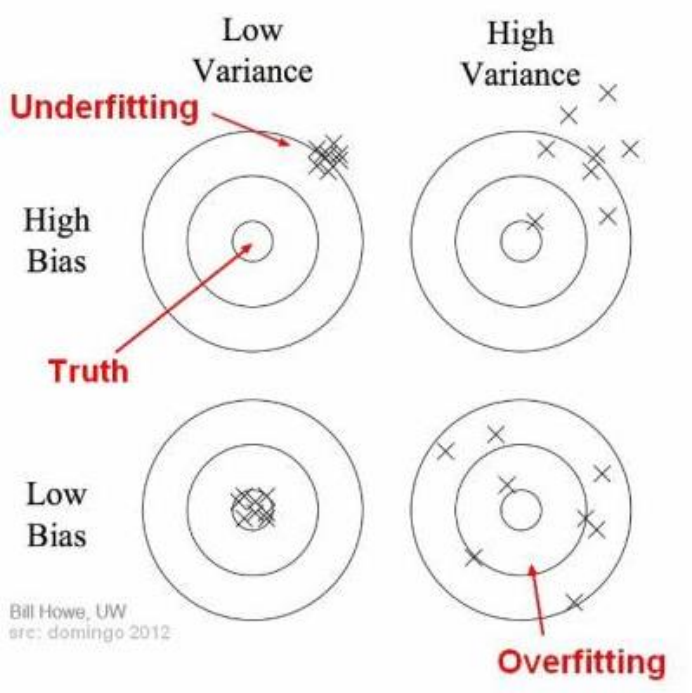
Variance

Variance occurs when the model performs good on the trained dataset but does not do well on a dataset that it is not trained on, like a test dataset or validation dataset. **Variance tells us how scattered are the predicted value from the actual value**.

High variance causes overfitting that implies that the algorithm models random noise present in the training data.

when a model has a high variance then the model becomes very flexible and tune itself to the data points of the training set. when a high variance model encounters a different data point that it has not learnt then it cannot make right prediction.

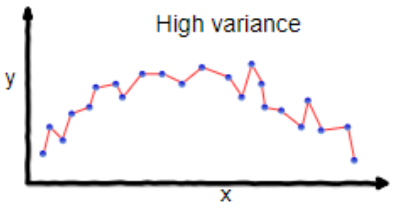
Bias and variance using bulls-eye diagram



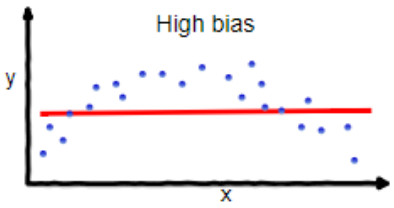
In the above diagram, center of the target is a model that perfectly predicts correct values. As we move away from the bulls-eye our predictions become get worse and worse. We can repeat our process of model building to get separate hits on the target.

In supervised learning, **underfitting** happens when a model unable to capture the underlying pattern of the data. These models usually have high bias and low variance. It happens when we have very less amount of data to build an accurate model or when we try to build a linear model with a nonlinear data. Also, these kind of models are very simple to capture the complex patterns in data like Linear and logistic regression.

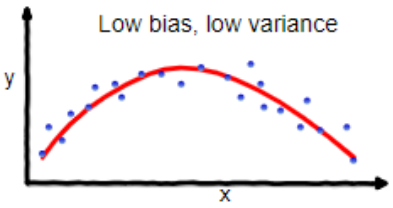
In supervised learning, **overfitting** happens when our model captures the noise along with the underlying pattern in data. It happens when we train our model a lot over noisy dataset. These models have low bias and high variance. These models are very complex like Decision trees which are prone to overfitting.



overfitting



underfitting



Good balance

Why is Bias Variance Tradeoff?

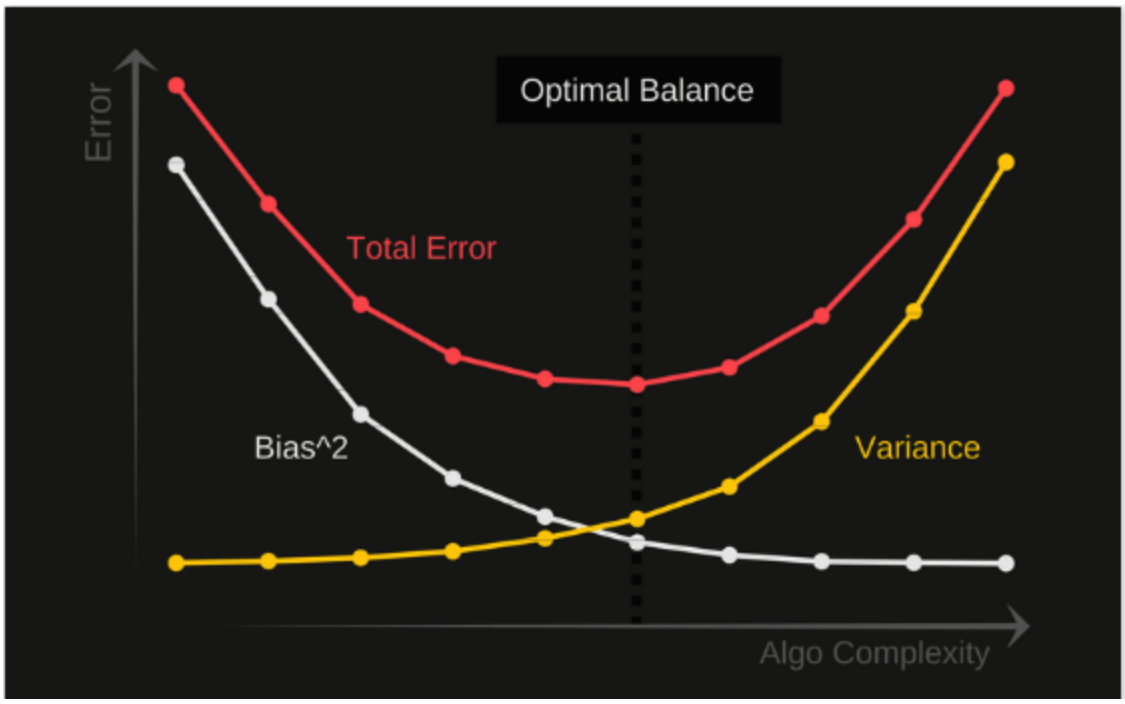
If our model is too simple and has very few parameters then it may have high bias and low variance. On the other hand if our model has large number of parameters then it's going to have high variance and low bias. So we need to find the right/good balance without overfitting and underfitting the data.

This tradeoff in complexity is why there is a tradeoff between bias and variance. An algorithm can't be more complex and less complex at the same time.

Total Error

To build a good model, we need to find a good balance between bias and variance such that it minimizes the total error.

Total Error = Bias^2 + Variance + Irreducible Error



An optimal balance of bias and variance would never overfit or underfit the model.

Therefore understanding bias and variance is critical for understanding the behavior of prediction models.