Bias and Variance in Machine Learning

Introduction

A machine learning model's performance is considered good based on it prediction and how well it generalizes on an independent test dataset. Models are assessed based on the prediction error on a new test dataset.

Whenever we discuss model prediction, it's important to understand prediction errors. Gaining a proper understanding of these errors would help us not only to build accurate models but also to avoid the mistake of overfitting and underfitting.

Recall that, *Error is when the predicted value is different from the actual value.*

Bias:

Bias is how far are the predicted values from the actual values. In other words, Bias is the difference between the average prediction of our model and the correct value which we are trying to predict.

Model with high bias pays very little attention to the training data and oversimplifies the model which causes model to miss relevant relationship between input and output variable. This leads to high error on training and test data and called Underfitting as the model is too simple and does not capture the complexity of data

Variance:

Variance occurs when the model performs good on the training dataset but does not do well (has high error rates) on a dataset that it is not trained on, like a test dataset or validation dataset.

This is because 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.

In simple terms, model with high variance pays a lot of attention to training data and does not generalize on the data which it hasn't seen before. This is called overfitting. As a result, such models perform very well on training data but has high error rates on test data.

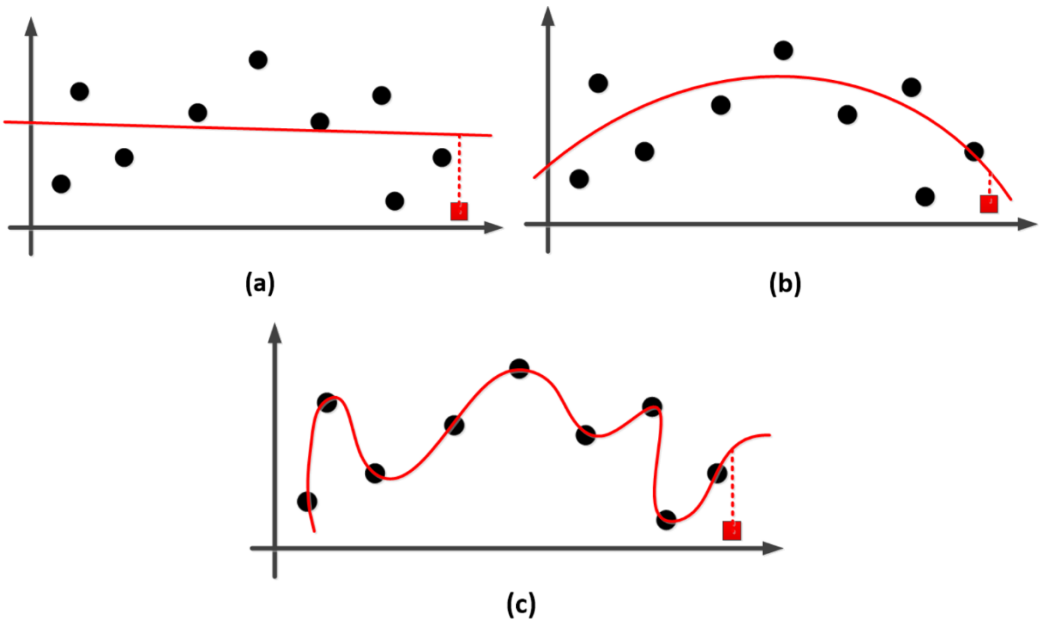
High variance causes overfitting. that implies that the algorithm models random noise present in the training data.

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.

Underfitting vs Overfitting

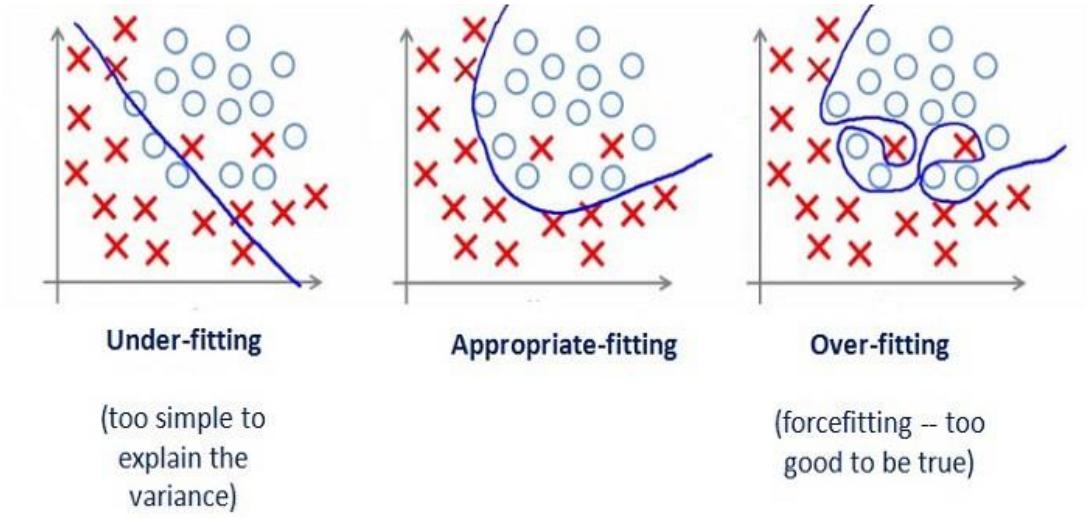
In supervised learning, underfitting happens when a model unable to capture the underlying pattern of the data. These models usually have high bias and low variance. It happens when we have very less amount of data to build an accurate model or when we try to build a linear model with a nonlinear data. 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.



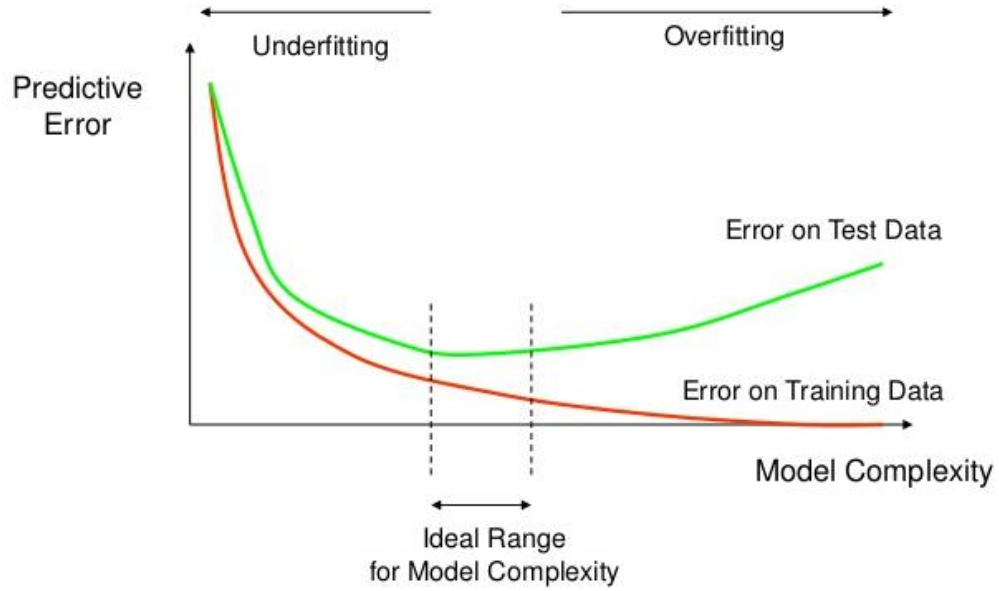
An example for (a) underfitting, high bias (b) good fit, and (c) overfitting, high variance. The black circles and red square are training and test instances, respectively. The red curve is the fitted curve.

In case of a classification problem,

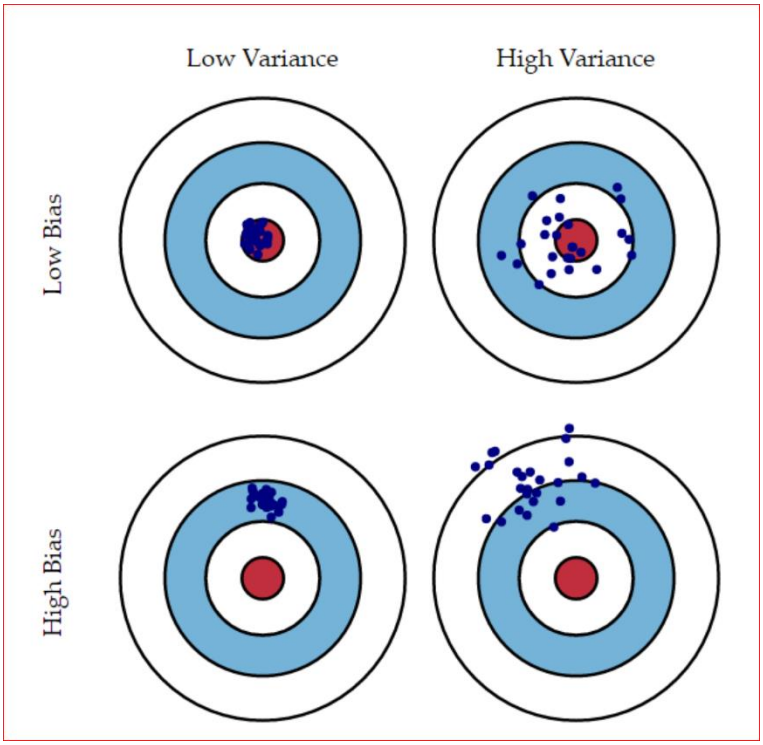


If we look at the diagrams above, we see that a model with high bias looks very simple. A model with high variance tries to fit most of the data points making the model complex and difficult to model.

This can also be visible from the plot below between test and training prediction error as a function of model complexity.



We would like to have a model complexity that trades bias off with variance so that we minimize the test error and would make our model perform better. A Bulls eye is illustrated to explain this trade off below:



In the above diagram, center of the target is a model that perfectly predicts correct values i.e., the perfect model that we can have considering all the combinations of the data that we can get. Each blue dot (·) marks a model that we have learned on the basis of combinations of the dataset and features that we get. 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.

- Models with low bias, low variance, represented by the top left image, will learn a good general structure of underlying data patterns and relationships that will be close to the hypothetical model and predictions will be consistent and hit the bull's eye!

- Models with low bias, high variance, represented by the top right image, are models that generalize to some extent (learn proper relationships\patterns) and perform decently on average due to low bias but are sensitive to the data it is trained on leading to high variance and hence predictions keep fluctuating.

- Models with high bias, low variance will tend to make consistent predictions irrespective of datasets on which the models are built leading to low variance but due to high bias, it will not learn the necessary patterns\relationships in the data that are required for correct predictions and hence misses the mark due to the high bias error on average, as depicted in the bottom-left image.

- Models with high bias, high variance are the worst sort of models possible, as they will not learn necessary data attribute relationships that are essential to correlation with output responses. Also, they will be extremely sensitive to data and outliers and noise leading to highly fluctuating predictions which result in high variance, as depicted in the bottom-right image

Extreme Cases of Bias-Variance

In real-world modelling, we will always have a trade-off between decreasing bias and variance simultaneously. To understand why we have this trade-off, we must first consider the two possible extreme cases of bias and variance.

Underfitting

Consider a linear model that is lazy and always predicts a constant value. This model will have extremely low variance (in fact it will be a zero variance model) as the model is not dependent at all on which subset of data it gets. It will always predict a constant and hence have stable performance. But on the other hand it will have extremely high bias as it has not learned anything from the data and made a very rigid and erroneous assumption about the data. This is the case of model underfitting, in which we fail to learn anything about the data, its underlying patterns, and relationships.

Overfitting

Consider the opposite case in which we have model that attempts to fit every data point it encounters (the closest example would be fitting an nth order polynomial curve for an n-observation dataset so that the curve passes through each point). In this case, we will get a model which will have low bias as no assumption to structure of data was made (even when there was some structure) but the variance will be very high as we have tightly fit the model to one of the possible subsets of data (focusing too much on the training data). Any subset different from the training set will lead to a lot of error. This is the case of overfitting, where we have built our model so specific to the data at hand that it fails to do any generalization over other subsets of data.

High Bias Low Variance: Models are consistent but inaccurate on average

High Bias High Variance: Models are inaccurate and also inconsistent on average

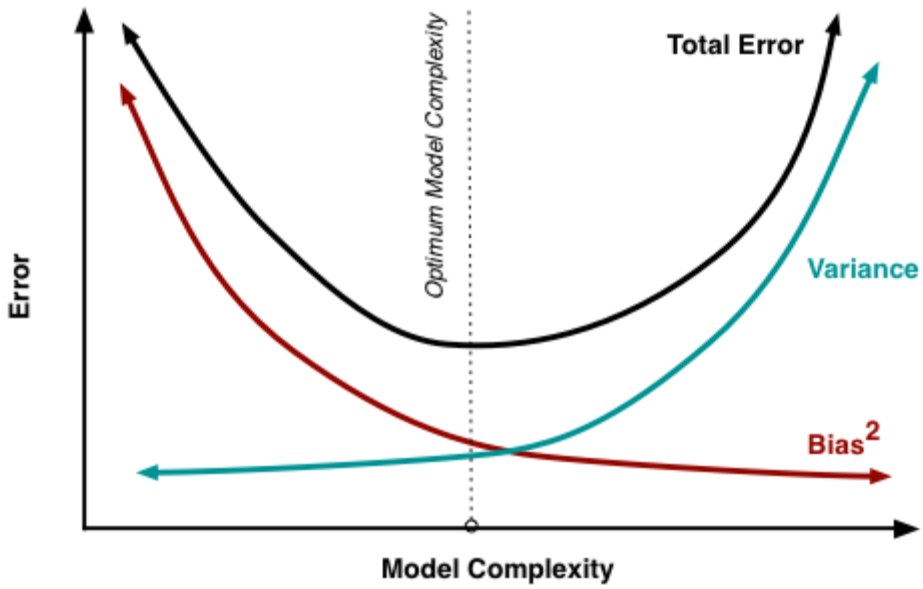
Low Bias Low Variance: Models are accurate and consistent on averages. We strive for this in our model

Low Bias High variance: Models are somewhat accurate but inconsistent on averages. A small change in the data can cause a large error.

Why is Bias Variance Trade-off?

The total generalization error of any model will be a sum of its bias error, variance error, and irreducible error, as depicted in the following equation.

Generalization Error = Bias Error + Variance Error + Irreducible Error



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

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

Is there a way to find when we have a high bias or a high variance?

High Bias can be identified when we have

- High training error
- Validation error or test error is same as training error

High Variance can be identified when

- Low training error
- High validation error on high test error

How do we fix high bias or high variance in the data set?

High bias is due to a simple model and we also see a high training error. To fix that we can do following things

- Add more input features
- Add more complexity by introducing polynomial features
- Decrease Regularization term

High variance is due to a model that tries to fit most of the training dataset points and hence gets more complex. To resolve high variance issue we need to work on

- Getting more training data
- Reduce input features
- Increase Regularization term

Regularization is a technique where we penalize the loss function for a complex model which is very flexible. This helps with overfitting. It does this by penalizing the different parameters or weights to reduce the noise of the training data and generalize well on the test data Regularization significantly reduces the variance without substantially increasing bias.

Extra Reading

When we have an input x and we apply a function f on the input x to predict an output y. Difference between the actual output and predicted output is the *error*. Our goal with machine learning algorithm is to generate a model which minimizes the error of the test dataset. Models are assessed based on the prediction error on a new test dataset.

Mathematically,

Let the variable we are trying to predict as Y and other covariates as X. We assume there is a relationship between the two such that

$$Y = f(X) + e$$

Where e is the error term and it's *normally distributed* with a mean of 0.

We will make an estimated model $\hat{f}(X)$ of $f(X)$ using linear regression or any other modelling technique. So the expected squared error at a point x is:

$$Error(x) = E\left[\left(Y - \hat{f}(x)\right)^2\right]$$

This $Error(x)$ can be further decomposed as :

$$Error(x) = E\left(E\left[\hat{f}(x)\right] - f(x)\right)^2 + E\left[\left(\hat{f}(x) - E[\hat{f}(x)\right)^2\right] + \sigma_e^2$$

which is simply,

$$Error(x) = Bias^2 + Variance + Irreducible\ Error$$

Irreducible error is the error that can't be reduced by creating good models. It is a measure of the amount of noise in our data. Here it is important to understand that no matter how good we make our model, our data will have certain amount of noise or irreducible error that cannot be removed.

Reducible Error has two components — bias and variance. Presence of bias or variance causes overfitting or underfitting of data.

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 behaviour of prediction models.