simpli learn

AI & Machine Learning

Video Tutorials   Articles   Ebooks   Free Practice Tests   On-demand Webinars

Everything You Need To Know About Bias And Variance

Lesson 25 of 33   By Simpl

Last updated on Sep 6, 2022

Previous   Next

Tutorial Playlist

Table of Contents

Errors in Machine Learning

What is Bias?

What is Variance?

Bias-Variance Tradeoff

Plotting Bias and Variance Using Python

View More

While discussing model accuracy, we need to keep in mind the prediction errors, ie: Bias and Variance, that will always be associated with any machine learning model. There will always be a slight difference in what our model predicts and the actual predictions. These differences are called errors. The goal of an analyst is not to eliminate errors but to reduce them. There is always a tradeoff between how low you can get errors to be. In this article titled 'Everything you need to know about Bias and Variance', we will discuss what these errors are.

Errors in Machine Learning

We can describe an error as an action which is inaccurate or wrong. In Machine Learning, error used to see how accurately our model can predict on data it uses to learn; as well as new, un data. Based on our error, we choose the machine learning model which performs best for a particular dataset.

FREE Machine Learning Course

Learn In-demand Machine Learning Skills and Tools

START NOW

There are two main types of errors present in any machine learning model. They are Reducible Errors and Irreducible Errors.

- Irreducible errors are errors which will always be present in a machine learning model, because of unknown variables, and whose values cannot be reduced.

- Reducible errors are those errors whose values can be further reduced to improve a model. They are caused because our model's output function does not match the desired output function and can be optimized.

We can further divide reducible errors into two: Bias and Variance.
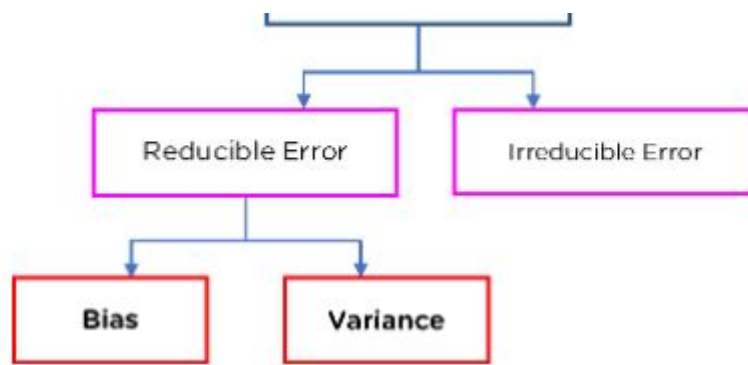
Machine Learning
Errors

Figure 1: Errors in Machine Learning

What is Bias?

To make predictions, our model will analyze our data and find patterns in it. Using these patterns we can make generalizations about certain instances in our data. Our model after training learns these patterns and applies them to the test set to predict them.

Bias is the difference between our actual and predicted values. Bias is the simple assumptions that our model makes about our data to be able to predict new data.
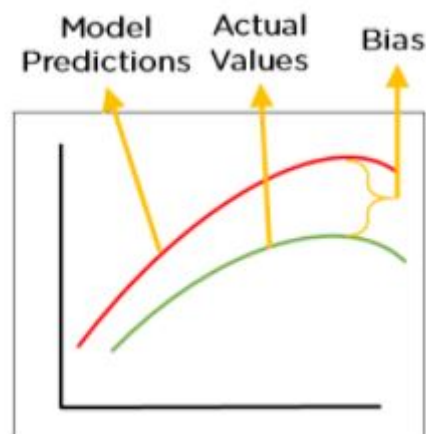

Figure 2: Bias

When the Bias is high, assumptions made by our model are too basic, the model can't capture the important features of our data. This means that our model hasn't captured patterns in the training data and hence cannot perform well on the testing data too. If this is the case, our model cannot perform on new data and cannot be sent into production.

This instance, where the model cannot find patterns in our training set and hence fails for both seen and unseen data, is called Underfitting.

The below figure shows an example of Underfitting. As we can see, the model has found no patterns in our data and the line of best fit is a straight line that does not pass through any of the data points. The model has failed to train properly on the data given and cannot predict new data either.
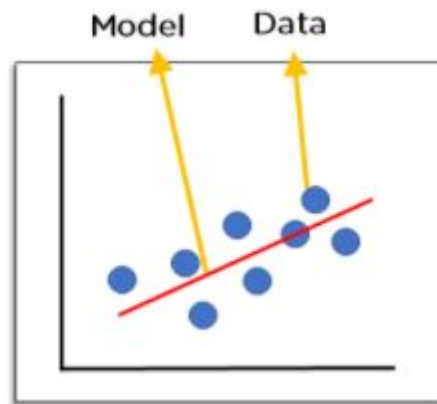

Figure 3: Underfitting

What is Variance?

Variance is the very opposite of Bias. During training, it allows our model to 'see' the data a certain number of times to find patterns in it. If it does not work on the data for long enough, it will not find patterns and bias occurs. On the other hand, if our model is allowed to view the data too many times, it will learn very well for only that data. It will capture most patterns in the data, but it will also learn from the unnecessary data present, or from the noise.

We can define variance as the model's sensitivity to fluctuations in the data. Our model may learn from noise. This will cause our model to consider trivial features as important.
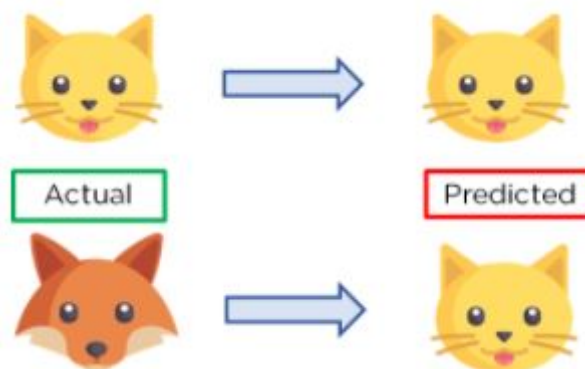

Figure 4: Example of Variance

In the above figure, we can see that our model has learned extremely well for our training data, which has taught it to identify cats. But when given new data, such as the picture of a fox, our model predicts it as a cat, as that is what it has learned. This happens when the Variance is high, our model will capture all the features of the data given to it, including the noise, will tune itself to the data, and predict it very well but when given new data, it cannot predict on it as it is too specific to training data.

Hence, our model will perform really well on testing data and get high accuracy but will fail to perform on new, unseen data. New data may not have the exact same features and the model won't be able to predict it very well. This is called Overfitting.
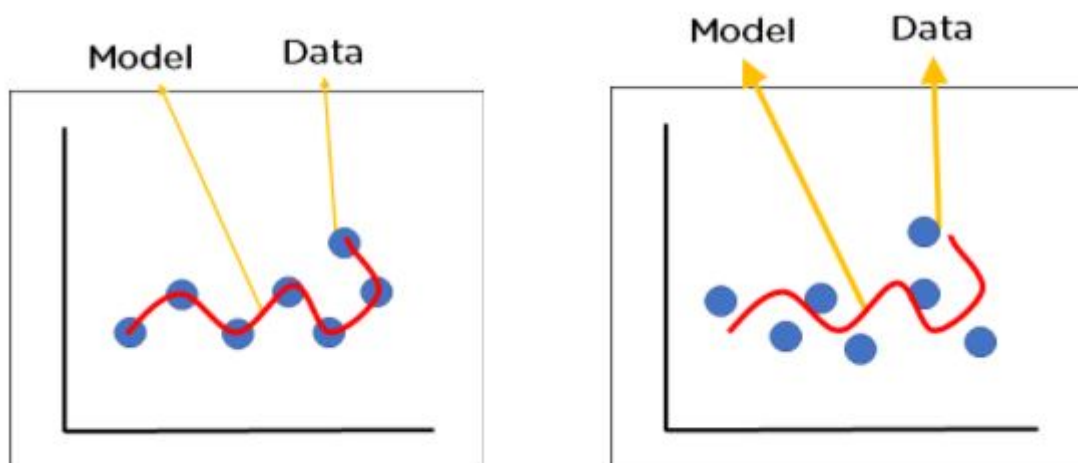

Figure 5: Over-fitted model where we see model performance on, a) training data b) new data

Bias-Variance Tradeoff

For any model, we have to find the perfect balance between Bias and Variance. This just ensures that we capture the essential patterns in our model while ignoring the noise present it in. This is

called Bias-Variance Tradeoff. It helps optimize the error in our model and keeps it as low as possible.

An optimized model will be sensitive to the patterns in our data, but at the same time will be able to generalize to new data. In this, both the bias and variance should be low so as to prevent overfitting and underfitting.
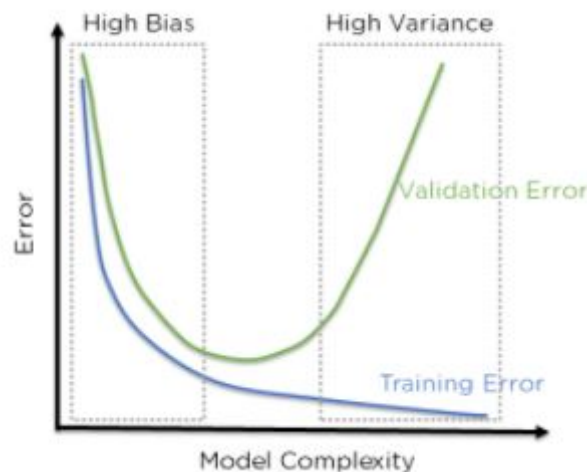

Figure 6: Error in Training and Testing with high Bias and Variance

In the above figure, we can see that when bias is high, the error in both testing and training set is also high.If we have a high variance, the model performs well on the testing set, we can see that the error is low, but gives high error on the training set. We can see that there is a region in the middle, where the error in both training and testing set is low and the bias and variance is in perfect balance.
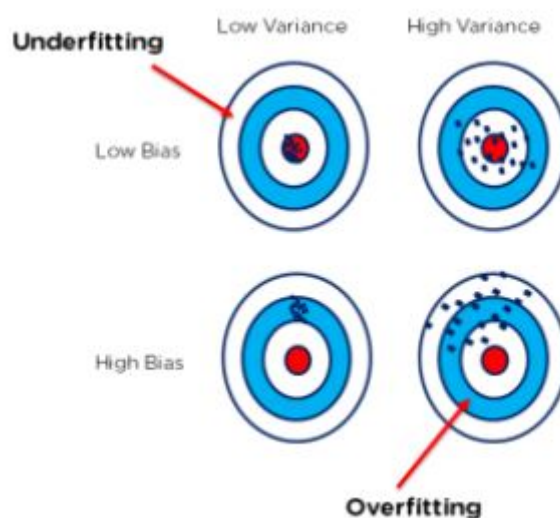

Figure 7: Bull's Eye Graph for Bias and Variance

The above bull's eye graph helps explain bias and variance tradeoff better. The best fit is when

the data is concentrated in the center, ie: at the bull's eye. We can see that as we get farther and farther away from the center, the error increases in our model. The best model is one where bias and variance are both low.

Plotting Bias and Variance Using Python

Let's find out the bias and variance in our weather prediction model. For this we use the daily forecast data as shown below:

Figure 8: Weather forecast data

We start off by importing the necessary modules and loading in our data.

Figure 9: Importing modules

In the data, we can see that the date and month are in military time and are in one column. The day of the month will not have much effect on the weather, but monthly seasonal variations are important to predict the weather. So, let's make a new column which has only the month.

Figure 10: Creating new month column

The dataset now looks as shown below.

Figure 11: New dataset

Dropping unnecessary columns.

Figure 12: Dropping columns