Causes of Error

Now that we have covered some basic metrics for measuring model performance, let us turn our attention to reasons why models exhibit errors in the first place.

In model prediction there are two main sources of errors that a model can suffer from. *Bias* due to a model being unable to represent the complexity of the underlying data or *variance* due to a model that is overly sensitive to the limited data it has been trained on. We will go over both in a bit more detail.

Error due to Bias - Accuracy and Underfitting

As mentioned before, bias occurs when a model has enough data but is not complex enough to capture the underlying relationships. As a result, the model consistently and systematically misrepresents the data, leading to low accuracy in prediction. This is known as *underfitting*.

Simply put, bias occurs when we have an inadequate model. An example might be when we have objects that are classified by color and shape, but our model can only partition and classify objects just by color (it is an overly simplified model) and therefore consistently mislabels future objects.

Or perhaps we have continuous data that is polynomial in nature but our model can only represent linear relationships. In this case it does not matter how much data we feed the model because it just cannot represent the underlying relationship that we need a more complex model for.

Error due to Variance - Precision and Overfitting

When we train a model, we typically use a limited number of samples from a larger population (the training set). If we repeatedly train a model with randomly selected subsets of data, we would expect its predictons to be different based on the specific examples given to it. Here *variance* is a measure of how much the predictions vary for any given test sample.

Some variance is normal, but too much variance indicates that the model is unable to generalize its predictions to the larger population from which training samples were drawn. High sensitivity to the training set is also known as *overfitting*, and generally occurs when either the model is too complex and/or we do not have enough data to support it.

We can typically reduce the variability of a model's predictions and increase precision by training on more data.

Improving the Validity of a Model

As we can see there is a trade-off when a model is too simple or complex given with a fixed set of data. If too simple, our model cannot learn about the data and misrepresents the data. However if we make a really complex model we need more data to learn the underlying relationship otherwise it is very common for a model to infer relationships that might not actually exist in the data.

The key is to find the sweet spot that minimizes bias and variance by finding the right model complexity. And of course the more data the better any model can improve over time.

In order to learn more about bias and variance, we recommend this essay by Scott Fortmann-Roe. In addition to the subset of data chosen for training, what features you use from a given dataset can also greatly affect the bias and variance of your model.