Bias Vs. Variance Tradeoff

What do you mean it?

It means that when bias decreases, variance increases and vice versa. Ideally, we need both of them to be as low as possible for a model to generalize well on unseen data. But that’s not, mostly, always possible, so we try to moderate them to have acceptable values for a model to generalize well.

Variance would cause a model to be inconsistent and bias (bias is the error in predicting actual values) would lead to prediction error. We ought to avoid both the situations, so we care about the tradeoff between them.

The major errors in building a machine learning model are due to bias and variance. But these two are not the only source as there is some irreducible error (due to randomness) that cannot be controlled or prevented from happening.

Bias can be thought simply as the error we make through our assumptions. For example: we might assume a linear relationship between target variable and features, but, in reality, the linear assumption may not capture the complexity that exists between the target and features. In that situation, the error between the predicted and actual values will be high. The model is said to underfit the data in this situation.

Variance means that your estimator (or learning algorithm) varies a lot depending on the data that you give it. A model is said to overfit the data when the variance is high. For example: Let’s assume the real estate startup, Opendoor is trying to predict the selling price of houses for 10 major neighborhoods in San Francisco for the year 2020. It has deployed a model to predict the selling price. Let’s say the model gets new data as input for every three weeks and the process started implemented in the year 2019.

Example:

The model predicts the selling price at the upper level of the confidence interval in SOMA to be 1.75M in Feb 2019 (for the year 2020), but in April 2019 it predicts the value to be 4.5M. The model clearly changes its predictions; for the same neighborhood it predicts different values with different datasets. It seems to overfit the data in this situation.

If your algorithm is able to fit your data extremely well every single time and even a single data point perturbation changes the algorithm a lot, then the algorithm is having high variance. This type of high variance is called overfitting. Thus, usually overfitting is related to high variance. This is bad because it means your algorithm is probably not robust to noise for example. Notice that overfitting implies high variance but not the other way around.

In real problems, you’ll generally see “high variance, low bias” and “overfitting” occur when models have more degrees of freedom, more parameters to fit, less regularization, more flexibility.