

The Bias-Variance Decomposition

When your learner outputs a classifier that is 100% accurate on the training data but only 50% accurate on test data, when in fact it could have output one that is 75% accurate on both, it has overfit.

One way to understand overfitting is by decomposing generalization error into *bias* and *variance*. Bias is a learner's tendency to consistently learn the same wrong thing. variance is the tendency to learn random things irrespective of the real signal.

It is possible to show that the expected test MSE, for a given x_0 , can always be decomposed into the sum of three fundamental quantities: the variance of $\hat{f}(x_0)$, the squared bias of $\hat{f}(x_0)$ and the variance of the error terms ϵ . That is,

$$E(y_0 - \hat{f}(x_0))^2 = \text{Var}(\hat{f}(x_0)) + [\text{Bias}(\hat{f}(x_0))]^2 + \text{Var}(\epsilon)$$

Here the notation $E(y_0 - \hat{f}(x_0))^2$ defines the *expected* test MSE, and refers to the average test MSE that we would obtain if we repeatedly estimated f using a large number of training sets, and tested each at x_0 . The overall expected test MSE can be computed by averaging $E(y_0 - \hat{f}(x_0))^2$ over all possible values of x_0 in the test set.

The Equation tells us that in order to minimize the expected test error, we need to select a statistical learning method that simultaneously achieves low variance and low bias. Variance refers to the amount by which \hat{f} would change if we estimated it using a different training data set. Since the training data are used to fit the statistical learning method, different training data sets will result in a different \hat{f} . But ideally the estimate for f should not vary too much between training sets. However, if a method has high variance then small changes in the training data can result in large changes in \hat{f} . In general, more flexible statistical methods have higher variance.

On the other hand, *bias* refers to the error that is introduced by approximating a real-life problem, which may be extremely complicated, by a much simpler model. For example, linear regression assumes that there is a linear relationship between Y and X_1, X_2, \dots, X_p . It is unlikely that any real-life problem truly has such a simple linear relationship, and so performing a linear regression will undoubtedly result in some bias in the estimate of f . As a general rule, as we use more flexible methods, the variance will increase and the bias will decrease. The relative rate of change of these two quantities determines whether increases or decreases. As we increase the flexibility of a class of methods the bias tends to initially decrease faster than the variance increases. Consequently, the expected test MSE declines. However, at some point increasing flexibility has little impact on the bias but starts

to significantly increase the variance. When this happens the test MSE increases while the training MSE keeps decreasing. (i.e. overfitting)