Bias-Variance Tradeoff

Jacob Fine

Oct 9th, 2024

In this exercise, we will define the loss function, the risk function, and use it to illustrate the bias-variance tradeoff. In essence, the bias-variance trade states that all else being equal, when we make our model more complex to reduce bias (i.e., by fitting the training data), we increase the chances of overfitting, making there be higher variance and worse generalization of our model. Given some parameter θ and an estimator $\hat{\theta}$, suppose we are interested in the difference between our estimate and the true value. This is captured by the loss function

$$L(\theta, \hat{\theta})$$

There are different ways to define the loss function, for instance, the squared error loss

$$L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$$

or the absolute loss

$$L(\theta, \hat{\theta}) = |\theta - \hat{\theta}|$$

We could also use information theoretic loss functions, for instance, based on the Kullback-Leibler divergence or cross-entropy loss, which allows us to capture more information about the distribution of our estimator and the true parameter.

When we take the expected value of the loss function, this is defined as the 'risk' of an estimator $\hat{\theta}$, $R(\theta, \hat{\theta})$. Therefore,

$$R(\theta, \hat{\theta}) = E[L(\theta, \hat{\theta})]$$

Based on the definition of the expected value and using the law of the unconscious statistician, we may write this as

$$R(\theta, \hat{\theta}) = \int L(\theta, \hat{\theta}) f(x|\theta) dx$$

We will now use a squared error loss function to illustrate how bias and variance are related. We note that bias and variance are defined as, respectively

$$\text{bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$$

$$\text{var}(\hat{\theta}) = E[\hat{\theta} - E(\hat{\theta})]^2$$

We can start by using the square loss function

$$E[L(\theta, \hat{\theta})] = E[(\theta - \hat{\theta})^2]$$

Now we add in terms that cancel out to expand the square

$$= E[(\theta - E(\hat{\theta}) + E(\hat{\theta}) - \hat{\theta})^2]$$

$$= E[(\theta - E(\hat{\theta}))^2 + 2(\theta - E(\hat{\theta}))(E(\hat{\theta}) - \hat{\theta}) + (E(\hat{\theta}) - \hat{\theta})^2]$$

We can apply the expectation operation to each term, and notice that the middle term disappears since in the expectation of the product, the first factor is constant (the parameter and the expected value of the estimate are fixed), so

$$E(2(\theta - E(\hat{\theta}))(E(\hat{\theta}) - \hat{\theta})) = 2(\theta - E(\hat{\theta})) \cdot E[(E(\hat{\theta}) - \hat{\theta})]$$

We notice that in the second term, $E(E(\hat{\theta})) = E(\hat{\theta})$ since the argument is a constant, which leads to the cancellation $E(\hat{\theta}) - E(\hat{\theta}) = 0$, making the entire product zero. The overall expression can therefore be written as

$$= [E(\hat{\theta}) - \theta]^2 + E[\hat{\theta} - E(\hat{\theta})]^2$$

Considering the definitions above of bias and variance, we can clearly see that this is equivalent to

$$\text{mean squared error} = \text{bias}(\hat{\theta})^2 + \text{var}(\hat{\theta})$$

It follows from this that for a given MSE, there is a tradeoff between bias and variance. We may explore the results of decomposing other risk functions, but the results may be less straightforward with respect to bias and variance.