# Bias & Variance In Machine-learning

CCTS 40500
(Thanks to Peter McHale)

February 18, 2019

## 1 True model

Let a target variable $Y$ and a feature variable $X$ are related via

$$Y = f(X) + \epsilon \tag{1}$$

where $X$ and $\epsilon$ are independent random variables with zero average error, *i.e.*, $E[\epsilon] = 0$.

Let $\mathcal{D}$ be a data set that obeys this relationship. Because data sets are always of finite size, we may think of $\mathcal{D}$ as a random variable. Consider a set of values $d$, where

$$d = \{(x_1, y_1), \ldots, (x_m, y_m)\} \tag{2}$$

where $x_i$ and $y_i$ are realizations of random variables $X$ and $Y$.

## 2 Estimated model

Machine learning uses a particular realization $d$ of $\mathcal{D}$ to train an estimate of the function $f(x)$, called the hypothesis $h_d(x)$. The subscript $d$ reminds us that the hypothesis is a random function that varies over training data sets.

## 3 Test error of the estimated model

Having learned an hypothesis for a particular training set $d$, we next evaluate the error made in predicting the value of $y$ on an unseen test value $x$. In linear regression, that test error is quantified by taking a test data set (also drawn from the distribution of $\mathcal{D}$) and computing the average of $(Y - h_d)^2$ over the data set. If the size of the test data set is large enough, this average is approximated by:

$$E_{X,\epsilon}[(Y(X,\epsilon) - h_d(X))^2] \tag{3}$$

As the training data set $d$ varies, so does the test error; in other words, test error is a random variable, the average of which over all training sets is given by

$$\text{expected test error} = E_{\mathcal{D}}\left[E_{X,\epsilon}\left[(Y(X,\epsilon) - h_{\mathcal{D}}(X))^2\right]\right].$$

In the following sections, I will show how this error can be decomposed into three parts: a *bias* that quantifies how much (the average of) the hypothesis deviates from $f$; a *variance* term that quantifies how much the hypothesis varies among training data sets; and an *irreducible error* that describes the fact that one's ability to predict is always limited by the noise $\epsilon$.

## 4 Establishing a useful order of integration

To compute the expected test error analytically, we rewrite the expectation operators in two steps. The first step is to recognize that $E_{X,\epsilon}[\ldots] = E_X\left[E_\epsilon[\ldots]\right]$, since $X$ and $E$ are independent. The second step is to use Fubini's theorem to reverse the order in which $X$ and $D$ are integrated out. The final result is that the expected test error is given by

$$\text{expected test error} = E_X\left[E_{\mathcal{D}}\left[E_\epsilon\left[(Y-h)^2\right]\right]\right] \tag{4}$$

,

where I have dropped the dependence of $Y$ and $h$ on $X$, $\epsilon$ and $\mathcal{D}$ in the interests of clarity.

## 5 Reducible and irreducible error

We fix values of $X$ and $\mathcal{D}$ (and therefore $f$ and $h$) and compute the inner-most integral in the expected test error:

$$\begin{aligned} E_\epsilon\left[(Y-h)^2\right] &= E_\epsilon\left[(f+\epsilon-h)^2\right] \\ &= E_\epsilon\left[(f-h)^2+\epsilon^2+2\epsilon(f-h)\right] \\ &= (f-h)^2 + E_\epsilon\left[\epsilon^2\right] + 0 \\ &= (f-h)^2 + Var_\epsilon\left[\epsilon\right]. \end{aligned}$$

The last term remains unaltered by subsequent averaging over $X$ and $D$. It represents the irreducible error contribution to the expected test error.

The average of the first term, $E_X\left[E_{\mathcal{D}}\left[(f-h)^2\right]\right]$, is sometimes called the reducible error.

## 6 Decomposing the reducible error into 'bias' and 'variance'

We relax our constraint that $\mathcal{D}$ is fixed (but keep the constraint that $X$ is fixed) and compute the innermost integral in the reducible error:

$$\begin{aligned} E_{\mathcal{D}}\left[(f-h)^2\right] &= E_{\mathcal{D}}\left[f^2+h^2-2fh\right] \\ &= f^2 + E_{\mathcal{D}}\left[h^2\right] - 2fE_{\mathcal{D}}\left[h\right] \end{aligned}$$

Adding and subtracting $E_{\mathcal{D}}\left[h\right]^2$, and rearranging terms, we may write the right-hand side above as

$$\left(f - E_{\mathcal{D}}\left[h\right]\right)^2 + Var_{\mathcal{D}}\left[h\right].$$

Averaging over $X$, and restoring the irreducible error, yields finally:

$$\boxed{\text{expected test error} = E_X\left[\left(f - E_{\mathcal{D}}\left[h\right]\right)^2\right] + E_X\left[Var_{\mathcal{D}}\left[h\right]\right] + Var_\epsilon\left[\epsilon\right].}$$

The first term is called the bias and the second term is called the variance.

The variance component of the expected test error is a consequence of the finite size of the training data sets. In the limit that training sets contain an infinite number of data points, there are no fluctuations in $h$ among the training sets and the variance term vanishes. Put another way, when the size of the training set is large, the expected test error is expected to be solely due to bias (assuming the irreducible error is negligible).

## 7 More info

An excellent exposition of these concepts and more can be found here.