Foundations of AI **Fall 2024**
Professors Greenwald and Ewing **Machine Learning**

Statistics for Machine Learning, Continued

1 Bias-Variance Decomposition Theorem

Imagine we wish to estimate a function $f : X \to Y$ from data, where the available dataset $\mathcal{D}$ is drawn from a noisy data-generating function as follows: given a data point $x \in X$, $y \in Y$ is generated as $y \sim f(x) + \epsilon$, where ϵ is a random variable centered at 0. That is, y is a noisy observation of $f(x)$, while $f(x)$ itself is the true value of the function f at x, also called the **ground truth**.

Let $\hat{f}_{\mathcal{D}}(x)$ denote the estimated, or predicted, value of $f(x)$ at x. We can compute the expected value of the squared error between our estimate/prediction $\hat{f}_{\mathcal{D}}(x)$ and the observed value y as follows:

$$\mathbb{E}_{y,\mathcal{D}}\left[\left(\hat{f}_{\mathcal{D}}(x) - y\right)^2\right] = \mathbb{E}_{y,\mathcal{D}}\left[\left(\hat{f}_{\mathcal{D}}(x) - f(x) + f(x) - y\right)^2\right] \tag{1}$$

$$= \mathbb{E}_{\mathcal{D}}\left[\left(\hat{f}_{\mathcal{D}}(x) - f(x)\right)^2\right] + \mathbb{E}_{y}\left[\left(f(x) - y\right)^2\right] + 2\mathbb{E}_{y,\mathcal{D}}\left[\left(\hat{f}_{\mathcal{D}}(x) - f(x)\right)\left(f(x) - y\right)\right] \tag{2}$$

$$= \mathbb{E}_{\mathcal{D}}\left[\left(\hat{f}_{\mathcal{D}}(x) - f(x)\right)^2\right] + \underbrace{\mathbb{E}_{y}\left[\left(f(x) - y\right)^2\right]}_{\text{irreducible error}} \tag{3}$$

Equation 3 follows from the fact that $\mathbb{E}_{y,\mathcal{D}}\left[\left(\hat{f}_{\mathcal{D}}(x) - f(x)\right)\left(f(x) - y\right)\right] = 0$, as $\mathbb{E}_{y \sim f(x)+\epsilon}\left[y\right] = f(x)$, since ϵ is centered at 0.

The right-hand term in Equation 3 is sometimes called **irreducible error**, as it represents error that arises from the fact that y is generated via a noisy data-generating function. The other term, however, is potentially **reducible error**, as it may vary with the choice of estimator $\hat{f}_{\mathcal{D}}$.

Define $\bar{f}(x) = \mathbb{E}_{\mathcal{D}}\left[\hat{f}_{\mathcal{D}}(x)\right]$. Note that this term is *not* the expected value of the estimator across datasets. It is the expected value of the estimated values (i.e., the predictions) across datasets.

Now let's take a closer look at the reducible error term, which is the expected value of the squared error of our estimate/prediction $\hat{f}_{\mathcal{D}}(x)$ and the *ground truth* $f(x)$:

$$\mathbb{E}_{\mathcal{D}}\left[\left(\hat{f}_{\mathcal{D}}(x) - f(x)\right)^2\right] = \mathbb{E}_{\mathcal{D}}\left[\left(\hat{f}_{\mathcal{D}}(x) - \bar{f}(x) + \bar{f}(x) - f(x)\right)^2\right] \tag{4}$$

$$= \mathbb{E}_{\mathcal{D}}\left[\left(\hat{f}_{\mathcal{D}}(x) - \bar{f}(x)\right)^2\right] + \left(\bar{f}(x) - f(x)\right)^2 + 2\mathbb{E}_{\mathcal{D}}\left[\left(\hat{f}_{\mathcal{D}}(x) - \bar{f}(x)\right)\left(\bar{f}(x) - f(x)\right)\right] \tag{5}$$

$$= \underbrace{\mathbb{E}_{\mathcal{D}}\left[\left(\hat{f}_{\mathcal{D}}(x) - \bar{f}(x)\right)^2\right]}_{\text{variance}} + \underbrace{\left(\mathbb{E}_{\mathcal{D}}\left[\hat{f}_{\mathcal{D}}(x)\right] - f(x)\right)^2}_{\text{bias}^2} \tag{6}$$

Note that $\mathbb{E}_{\mathcal{D}}\left[\left(\hat{f}_{\mathcal{D}}(x) - \bar{f}(x)\right)\left(\bar{f}(x) - f(x)\right)\right] = 0$, since $\mathbb{E}_{\mathcal{D}}\left[\hat{f}_{\mathcal{D}}(x)\right] = \bar{f}(x)$.

In summary, the expected value of the squared error between $\hat{f}_{\mathcal{D}}(x)$ and y is a combination of three terms:

$$\mathbb{E}_{y,\mathcal{D}}\left[\left(\hat{f}_{\mathcal{D}}(x)-y\right)^2\right]=\underbrace{\mathbb{E}_{\mathcal{D}}\left[\left(\hat{f}_{\mathcal{D}}(x)-\bar{f}(x)\right)^2\right]}_{\text{variance}}+\underbrace{\left(\mathbb{E}_{\mathcal{D}}\left[\hat{f}_{\mathcal{D}}(x)\right]-f(x)\right)^2}_{\text{bias}^2}+\underbrace{\mathbb{E}_{y}\left[(f(x)-y)^2\right]}_{\text{irreducible error}} \tag{7}$$

More specifically, the reducible error is the sum of the variance and the bias squared. The bias-variance tradeoff is precisely the fact that reducible error can be "allocated" across bias and variance. This allocation decision is the choice of a high or a low bias model, which impacts the variance accordingly.