Machine Learning for Physical Scientists 2021

A Careful Derivation of Bias-Variance Decomposition

Thiparat Chotibut

Chula Intelligent and Complex Systems,

Department of Physics, Chulalongkorn University, Bangkok, Thailand

We've seen subtle issues with the definition of expected *out-of-sample* error. In class we follow the derivation of Mehta's Physics Report 2019, and define the expected *out of sample* error as sample averages over training data set $D = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ and noise realization ϵ as

$$\mathbb{E}_{D,\epsilon}\left[\frac{1}{n}\sum_{i=1}^{n}\ell(y_i, h(\mathbf{x}_i; \boldsymbol{\theta}_D))\right].$$

However, as some of you have already noticed in the discussion session, this definition is still somewhat like an *in-sample-error* because the expectation value is taken with respect to the training data set. Although the expectation value over noise is taken, we still have not evaluated the ability of the empirically optimal hypothesis $h_{\boldsymbol{\theta}_D}$ to *predict* the god-given function value for inputs $\mathbf{x}$ *not in the training set* D.

Now we'll be the most careful with the definition of the expected *out-of-sample* error which will account for *unseen inputs.* For notational brevity, we'll define the concepts you've learned in a more compact notation. First, recall the main assumption behind statistical learning theory, that is the training set $D = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ are drawn i.i.d. from some god-given distribution $P(X, Y)$. Without loss of generality, we may assume a regression setting, where $y \in \mathbb{R}$; the generalization to a discrete label setting (classification) is straightforward. Given the training set D, our machine learning algorithm $\mathcal{A}$ spits out the empirically optimal hypothesis as $h_D = \mathcal{A}(D)$. Then, we should define the *generalization error* of h_D as

Expected Test (out-of-sample) Error

$$\mathbb{E}_{(\mathbf{x},y)\sim P}\left[(h_D(\mathbf{x}) - y)^2\right] = \iint (h_D(\mathbf{x}) - y)^2 P(\mathbf{x}, y) d\mathbf{x} dy. \tag{1}$$

Note that one can use other loss functions. We use squared loss because it has nice mathematical properties, and it is also the most common loss function. The previous statement is true for a given training set D; However, remember that D is drawn from P^n, and therefore is a random variable. Furthermore, h_D is a function of D, and thus is also a random variable (function). Thus we can compute the expectation value of the empirically optimal hypothesis. Therefore, given a learning algorithm $\mathcal{A}$, we can calculate

Expected Predictor (Classifier)

$$\bar{h} = \mathbb{E}_{D\sim P^n}\left[h_D\right] = \int h_D P(D) dD, \tag{2}$$

where $P(D)$ is the probability of sampling D from P^n. Here, $\bar{h}$ is a weighted average over functions. We can also use the fact that h_D is a random variable to compute the expected out-of-sample error only given a learning algorithm $\mathcal{A}$, by taking the expectation over the data sample D as

Expected Test Error, given a learning algorithm $\mathcal{A}$,

$$\mathbb{E}_{\substack{(\mathbf{x},y)\sim P \\ D\sim P^n}}\left[(h_D(\mathbf{x})-y)^2\right] = \int\int\int (h_D(\mathbf{x})-y)^2\,\mathrm{P}(\mathbf{x},y)\mathrm{P}(D)d\mathbf{x}dydD. \tag{3}$$

Note that in this notation, D is our **training set** (a set of measure zero) whereas all the other $(\mathbf{x}, y)$ pairs are the **test set**. Therefore, this expression precisely quantifies the *generalization error* or expected *out-of-sample error* of a machine learning algorithm $\mathcal{A}$, with respect to sampling the training set D generated from the joint data distribution $P(X, Y)$. Having defined all necessary ingredients as above, now we may proceed to bias-variance (and noise) decomposition as usual.

The expected *out-of-sample error* in (3) can be decomposed as follow.

$$\begin{aligned}
\mathbb{E}_{\mathbf{x},y,D}\left[[h_D(\mathbf{x})-y]^2\right] &= \mathbb{E}_{\mathbf{x},y,D}\left[\left[\left(h_D(\mathbf{x})-\bar{h}(\mathbf{x})\right)+(\bar{h}(\mathbf{x})-y)\right]^2\right] \\
&= \mathbb{E}_{\mathbf{x},D}\left[\left(\bar{h}_D(\mathbf{x})-\bar{h}(\mathbf{x})\right)^2\right] + 2\mathbb{E}_{\mathbf{x},y,D}\left[\left(h_D(\mathbf{x})-\bar{h}(\mathbf{x})\right)(\bar{h}(\mathbf{x})-y)\right] \\
&\quad + \mathbb{E}_{\mathbf{x},y}\left[(\bar{h}(\mathbf{x})-y)^2\right].
\end{aligned}$$

The middle term in the last equality vanishes because

$$\begin{aligned}
\mathbb{E}_{\mathbf{x},y,D}\left[\left(h_D(\mathbf{x})-\bar{h}(\mathbf{x})\right)(\bar{h}(\mathbf{x})-y)\right] &= \mathbb{E}_{\mathbf{x},y}\left[\mathbb{E}_D\left[h_D(\mathbf{x})-\bar{h}(\mathbf{x})\right](\bar{h}(\mathbf{x})-y)\right] \\
&= \mathbb{E}_{\mathbf{x},y}\left[\left(\mathbb{E}_D\left[h_D(\mathbf{x})\right]-\bar{h}(\mathbf{x})\right)(\bar{h}(\mathbf{x})-y)\right] \\
&= \mathbb{E}_{\mathbf{x},y}[(\bar{h}(\mathbf{x})-\bar{h}(\mathbf{x}))(\bar{h}(\mathbf{x})-y)] \\
&= \mathbb{E}_{\mathbf{x},y}[0] \\
&= 0
\end{aligned}$$

Therefore, the earlier expression reduces to the sum between the variance and the other term as

$$\mathbb{E}_{\mathbf{x},y,D}\left[(h_D(\mathbf{x})-y)^2\right] = \underbrace{\mathbb{E}_{\mathbf{x},D}\left[\left(h_D(\mathbf{x})-\bar{h}(\mathbf{x})\right)^2\right]}_{\text{Variance}} + \mathbb{E}_{\mathbf{x},y}\left[(\bar{h}(\mathbf{x})-y)^2\right],$$

whereas the last term can be decomposed as

$$\begin{aligned}\mathbb{E}_{\mathbf{x},y}\left[(\bar{h}(\mathbf{x})-y)^2\right] &= \mathbb{E}_{\mathbf{x},y}\left[(\bar{h}(\mathbf{x})-\bar{y}(\mathbf{x}))+(\bar{y}(\mathbf{x})-y)^2\right] \\ &= \underbrace{\mathbb{E}_{\mathbf{x},y}\left[(\bar{y}(\mathbf{x})-y)^2\right]}_{\text{Noise}} + \underbrace{\mathbb{E}_{\mathbf{x}}\left[(\bar{h}(\mathbf{x})-\bar{y}(\mathbf{x}))^2\right]}_{\text{Bias }^2} + 2\mathbb{E}_{\mathbf{x},y}[(\bar{h}(\mathbf{x})-\bar{y}(\mathbf{x}))(\bar{y}(\mathbf{x})-y)],\end{aligned}$$

where we have defined

Expected Label, given $\mathbf{x} \in \mathbb{R}^d$, as

$$\bar{y}(\mathbf{x}) = \mathbb{E}_{y|\mathbf{x}}[Y] = \int yP(y \mid \mathbf{x})dy. \tag{4}$$

Now we show that the last term in the equality right above the definition of expected label vanishes:

$$\begin{aligned}\mathbb{E}_{\mathbf{x},y}[(\bar{h}(\mathbf{x})-\bar{y}(\mathbf{x}))(\bar{y}(\mathbf{x})-y)] &= \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{y|\mathbf{x}}[\bar{y}(\mathbf{x})-y](\bar{h}(\mathbf{x})-\bar{y}(\mathbf{x}))\right] \\ &= \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{y|\mathbf{x}}[\bar{y}(\mathbf{x})-y](\bar{h}(\mathbf{x})-\bar{y}(\mathbf{x}))\right] \\ &= \mathbb{E}_{\mathbf{x}}\left[\left(\bar{y}(\mathbf{x})-\mathbb{E}_{y|\mathbf{x}}[y]\right)(\bar{h}(\mathbf{x})-\bar{y}(\mathbf{x}))\right] \\ &= \mathbb{E}_{\mathbf{x}}[(\bar{y}(\mathbf{x})-\bar{y}(\mathbf{x}))(\bar{h}(\mathbf{x})-\bar{y}(\mathbf{x}))] \\ &= \mathbb{E}_{\mathbf{x}}[0] \\ &= 0.\end{aligned}$$

Finally, writing all the term together, we have the bias-variance (and noise) decomposition as required

$$\underbrace{\mathbb{E}_{\mathbf{x},y,D}\left[(h_D(\mathbf{x})-y)^2\right]}_{\text{Expected Test Error}} = \underbrace{\mathbb{E}_{\mathbf{x},D}\left[\left(h_D(\mathbf{x})-\bar{h}(\mathbf{x})\right)^2\right]}_{\text{Variance}} + \underbrace{\mathbb{E}_{\mathbf{x},y}\left[(\bar{y}(\mathbf{x})-y)^2\right]}_{\text{Noise}} + \underbrace{\mathbb{E}_{\mathbf{x}}\left[(\bar{h}(\mathbf{x})-\bar{y}(\mathbf{x}))^2\right]}_{\text{Bias }^2}. \tag{5}$$