Bias-Variance-Noise Decomposition

By: Behzad Asadi

Generalization Error Decomposition

Training: Consider the training dataset

$$D = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\},$$

where samples are drawn i.i.d. according to the distribution $p(\mathbf{x}, y)$. We fit a learner to this dataset which estimates the dependent variable, y_i, using explanatory variables, $\mathbf{x}_i$.

Expected Test Error: Assuming a regression problem with squared error loss where $h_D(\mathbf{x})$ is the regression function, the expected test error is defined as

$$E_{\substack{(\mathbf{x}, y) \sim p(\mathbf{x}, y) \\ D \sim p^n(\mathbf{x}, y)}} \left[(h_D(\mathbf{x}) - y)^2 \right] = \int_D \int_{\mathbf{x}} \int_y (h_D(\mathbf{x}) - y)^2 p(\mathbf{x}, y) p(D) d_{\mathbf{x}} d_y d_D$$

where training points, D, and test points, $(\mathbf{x}, y)$, are independent from each other. The expected test error can be divided into three terms as

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

where $\bar{h}(\mathbf{x})$ is the expected model

$$\bar{h}(\mathbf{x}) = E_D \left[h_D(\mathbf{x}) \right] = \int_D h_D(\mathbf{x}) p(D) d_D.$$

Generalization Error Decomposition

The term $E_{\mathbf{x},y,D}\left[\left(h_D(\mathbf{x})-\bar{h}(\mathbf{x})\right)\left(\bar{h}(\mathbf{x})-y\right)\right]$ is equal to zero

$$\begin{aligned}
&E_{\mathbf{x},y,D}\left[\left(h_D(\mathbf{x})-\bar{h}(\mathbf{x})\right)\left(\bar{h}(\mathbf{x})-y\right)\right]\\
&\quad=E_{\mathbf{x},y}\left[E_D\left[h_D(\mathbf{x})-\bar{h}(\mathbf{x})\right]\left(\bar{h}(\mathbf{x})-y\right)\right]\\
&\quad=E_{\mathbf{x},y}\left[\left(\bar{h}(\mathbf{x})-\bar{h}(\mathbf{x})\right)\left(\bar{h}(\mathbf{x})-y\right)\right]\\
&\quad=0.
\end{aligned}$$

The term $E_{\mathbf{x},y}\left[\left(\bar{h}(\mathbf{x})-y\right)^2\right]$ can be similarly divided into three terms as

$$\begin{aligned}
E_{\mathbf{x},y}\left[\left(\bar{h}(\mathbf{x})-y\right)^2\right]&=E_{\mathbf{x},y}\left[\left(\bar{h}(\mathbf{x})-\bar{y}(x)+\bar{y}(x)-y\right)^2\right]\\
&=E_{\mathbf{x}}\left[\left(\bar{h}(\mathbf{x})-\bar{y}(x)\right)^2\right]+E_{\mathbf{x},y}\left[\left(\bar{y}(x)-y\right)^2\right]\\
&\qquad+2E_{\mathbf{x},y}\left[\left(\bar{h}(\mathbf{x})-\bar{y}(x)\right)\left(\bar{y}(x)-y\right)\right],
\end{aligned}$$

where $\bar{y}(\mathbf{x})$ is the expected label given $\mathbf{x}$, calculated as $\bar{y}(\mathbf{x})=\int_y yp(y\mid\mathbf{x})d_y$, and the last term is equal to zero

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

Generalization Error Decomposition

Therefore, we can write the expected test error as

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

Variance: This part shows changes in the classifier as the result of changes in the training dataset. It is high when we have overfitting, and the learner is over-specialized to the training dataset.

Bias: This part shows the error even when we have infinite training data. It is high when we have underfitting.

Noise: This is due to the noise in the data.

References

1. Kilian Weinberger, Lecture Notes, https://www.cs.cornell.edu/courses/cs4780/2018fa/lectures/lecturenote12.html.
2. Kevin P. Murphy, Machine Learning: A Probabilistic Perspective.
3. Trevor Hastie, Robert Tibshirani, and Jerome Friedman, The Elements of Statistical Learning.