Bias and Variance Tradeoff

Zhe Li

November 4, 2014

1 Bias and Variance tradeoff

First of all, Let me say this truth that I have been thinking and struggling this topic for a long time. For this topic, the basic assumption is that data is generated from an underlying distribution \((x, y) \sim p(x, y)\). We consider the square loss function \(\ell(f; x, y) = (f(x) - y)^2\), So the expected loss:

\[
E[\ell(f)] = E_{x,y}[(f(x) - y)^2] = \int \int (f(x) - y)^2 p(x, y) dx dy \tag{1}
\]

Intuitively, we want the expected loss \(E[\ell(f)]\) to be minimum. In order to get the minimum of \(E[\ell(f)]\), we compute the gradient of \(E[\ell(f)]\) w.r.t \(f(x)\). It is better to write Eq. (1) in the following way:

\[
E[\ell(f)] = E_{x,y}[(f(x) - y)^2] = \int (f(x_1) - y)^2 p(x_1, y) dy + \int (f(x_2) - y)^2 p(x_2, y) dy + \cdots + \int (f(x_n) - y)^2 p(x_n, y) dy
\]

Then computing gradient of \(E[\ell(f)]\) w.r.t \(f(x)\) and setting it to zero:

\[
\frac{\partial E[\ell(f)]}{\partial f(x)} = 2 \int (f(x) - y)p(x, y) dy = 0 \tag{2}
\]

Rearranging the above equation, it gives

\[
f^*(x) = \frac{\int yp(x, y)dy}{p(x)} = E_{y|x}[y] \tag{3}
\]

Here, it is very important to notice that \(E_{y|x}[y]\) is independent to \(y\), and it is the function of \(x\). The above gives you the mathametic derivation of optimal \(f^*(x)\), which is hard to understand. However, \(E_{y|x}[y]\) is nothing more than the “average” of target \(y\) given the specific \(x\). For example, for given \(x_1\), maybe there are several data points corresponding
to $x_1$ such as $\{(x_1, y_{1,1}), (x_1, y_{1,2}), \ldots, (x_1, y_{1,m})\}$, assume that those points have same probability drawing from $p(x, y)$, then the optimal $f^*(x_1)$ is
\[
 f^*(x_1) = \frac{1}{m}(y_{1,1} + y_{1,2} + \cdots + y_{1,m}) \quad (4)
\]
Decompose the expected loss using optimal solution $f^*(x)$,
\[
 (f(x) - y)^2 = (f(x) - E_{y|x}[y] + E_{y|x}[y] - y)^2
\]
\[
 = (f(x) - E_{y|x}[y])^2 + (E_{y|x}[y] - y)^2 + 2(f(x) - E_{y|x}[y])(E_{y|x}[y] - y)
\]
Keep in mind that $E_{y|x}[y]$ does not dependent on $y$. For example, for this specific point $(\hat{x}, \hat{y})$, we have
\[
 (f(\hat{x}) - \hat{y})^2 = (f(\hat{x}) - E_{y|\hat{x}}[y] + E_{y|\hat{x}}[y] - y)^2
\]
\[
 = (f(\hat{x}) - E_{y|\hat{x}}[y])^2 + (E_{y|\hat{x}}[y] - y)^2 + 2(f(\hat{x}) - E_{y|\hat{x}}[y])(E_{y|\hat{x}}[y] - y)
\]
Since $(\hat{x}, \hat{y})$ are drawn from $p(x, y)$, take the expectation to both side of the equation, specifically for the last term,
\[
 E_{x,y}\{(f(\hat{x}) - E_{y|\hat{x}}[y])(E_{y|\hat{x}}[y] - y)\} \quad (5)
\]
Notice that
\[
 E_{x,y}\{E_{y|\hat{x}}[y]\} = E_{x,y}[y] \quad (6)
\]
That gives us that
\[
 E_{x,y}\{(f(\hat{x}) - E_{y|\hat{x}}[y])(E_{y|\hat{x}}[y] - y)\} = 0
\]
So the expected loss,
\[
 E[(f(x) - y)^2] = E[(f(x) - E_{y|x}[y])^2] + E[(E_{y|x}[y] - y)^2] \quad (7)
\]
In Eq. (7), the last term does not involve with $f(x)$, that is to say, no matter what the predictive function $f(x)$ is, the last term is always there, even you have the best predictive function $f^*(x)$. We called this term is \textit{noise} and that noise is from data itself or depends on $p(x, y)$, so we have no any power to control this noise term.

In the above, we only concern one function $f(x)$ from an underlying distribution $p(x, y)$. The fact is in reality what we have is data $\mathcal{D}$. Different data $\mathcal{D}$ will lead to different functions $f(x)$, which brings the uncertainty in $f(x)$. Here you can consider that there are lots of functions $f(x)$. Based on the above, we can take expected loss over $f(x)$. Here I did not agree the notation in the book, for which they used the notation $E_{\mathcal{D}}\{E[\ell(f)]\}$. I think it might be more clear to use the notation $E_f\{E[\ell(f)]\}$. But on the other hand, the notation
$E_D\{E[\ell(f)]\}$ also makes sense, since different data $D$ will lead to different function $f(x)$, as said before. Put that formally,

$$E_D\{E[\ell(f)]\} = E_D\left\{E[(f(x) - E_y|x[y])^2]\right\} + noise \tag{8}$$

If we decompose expected loss with $E_D[f(x; D)]$,

$$(f(x; D) - E_{x,y}[y])^2 = (f(x; D) - E_D[f(x; D)])^2 + (E_D[f(x; D)] - E_y|x[y])^2 + 2(f(x; D) - E_D[f(x; D)])(E_D[f(x; D)] - E_y|x[y])$$

Taking the expectation on both sides of the above equation on $D$, and first consider the last term of RHS,

$$E_D\left\{(f(x; D) - E_D[f(x; D)])\right\} = 0 \tag{9}$$

If one does not understand the above equation, think of $E[p - E[p]] = 0$, which is similar to the above equation. Note that

$$E_D\left\{(E_D[f(x; D)] - E_y|x[y])^2\right\} = (bias)^2 \tag{10}$$

and

$$E_D\left\{(f(x; D) - E_D[f(x; D)])^2\right\} = Variance \tag{11}$$

finally, we reach the end,

$$Expected \ loss = (Bias)^2 + Variance + Noise \tag{12}$$

In summary, for understanding Bias, you can consider $E_{y|x}[y]$ is the "underlying standard", Bias measures how far the overall predictive function $f(x; D)$ is away from this "underlying standard". For understanding Variance, variance measures how much the predictive function $f(x; D)$ varies from the "average" prediction function $E_D[f(x; D)]$. For noise, just as said before, it has nothing with predictive function $f(x; D)$, only depends on the data itself.