Notes on derivation of bias-variance decomposition in linear regression

Greg Shakhnarovich

October 2, 2011

Derivation of bias-variance decomposition for regression

Reminder: we are considering a model

$$y = F(\mathbf{x}) + \nu$$

where ν is additive *white* noise with variance σ_ν^2 (note: noise does not have to be Gaussian, but does have to be white). This means, in particular, that for any $\mathbf{x}_0$,

$$F(\mathbf{x}_0) = E_{y|\mathbf{x}}\left[y_0|\mathbf{x}_0\right]. \tag{1}$$

Let us start with writing down, and manipulating a bit, the expected loss with a predictor $\hat{f}$:

$$\begin{aligned}
E_{\mathbf{x},y}\left[\left(y_0 - \hat{f}(\mathbf{x}_0)\right)^2\right] &= E_{\mathbf{x},y}\left[\left(y_0 - F(\mathbf{x}_0) + F(\mathbf{x}_0) - \hat{f}(\mathbf{x}_0)\right)^2\right] && (2)\\
&= E_{\mathbf{x},y}\left[(y_0 - F(\mathbf{x}_0))^2\right] && (3)\\
&\quad + E_{\mathbf{x},y}\left[\left(F(\mathbf{x}_0) - \hat{f}(\mathbf{x}_0)\right)^2\right] && (4)\\
&\quad + 2E_{\mathbf{x},y}\left[(y_0 - F(\mathbf{x}_0))\left(F(\mathbf{x}_0) - \hat{f}(\mathbf{x}_0)\right)\right] && (5)
\end{aligned}$$

Focusing on (5), we have

$$\begin{aligned}
&E_{\mathbf{x},y}\left[(y_0 - F(\mathbf{x}_0))\left(F(\mathbf{x}_0) - \hat{f}(\mathbf{x}_0)\right)\right] && (6)\\
&\qquad = \int\int (y_0 - F(\mathbf{x}_0))\left(F(\mathbf{x}_0) - \hat{f}(\mathbf{x}_0)\right) p(y_0|\mathbf{x}_0)p(\mathbf{x}_0)dy_0 d\mathbf{x}_0 && (7)\\
&\qquad = \int \left\{E_{y|\mathbf{x}}\left[(y_0 - F(\mathbf{x}_0))\right]\right\}\left(F(\mathbf{x}_0) - \hat{f}(\mathbf{x}_0)\right) p(\mathbf{x}_0)d\mathbf{x}_0 && (8)\\
&\qquad = 0, && (9)
\end{aligned}$$

the last step due to (1). So, the term in (5) vanishes. Next we consider (3): this is the expected squared deviation of y_0 from $F(\mathbf{x}_0)$, which by definition is the variance of the noise, σ_ν^2.

Now on to (4). We will repeat a trick similar to that used in (2), subtracting and adding $\bar{f}(\mathbf{x}_0) = E_X\left[\hat{f}(\mathbf{x}_0)\right]$. The expectation here is taken w.r.t. the random training data set X

which produces the fit. Note that we should really use the notation $\hat{f}(\mathbf{x}_0, X)$ to indicate the dependence of $\hat{f}$ on the data, but we will sometimes write $\hat{f}(\mathbf{x}_0)$ for brevity. Let us work with the term inside the expectation in (4).

$$
\begin{aligned}
\left(F(\mathbf{x}_0) - \hat{f}(\mathbf{x}_0)\right)^2 &= \left(F(\mathbf{x}_0) - \bar{f}(\mathbf{x}_0) + \bar{f}(\mathbf{x}_0) - \hat{f}(\mathbf{x}_0)\right)^2 && (10)\\
&= \left(F(\mathbf{x}_0) - \bar{f}(\mathbf{x}_0)\right)^2 && (11)\\
&+ \left(\bar{f}(\mathbf{x}_0) - \hat{f}(\mathbf{x}_0)\right)^2 && (12)\\
&+ 2\left(F(\mathbf{x}_0) - \bar{f}(\mathbf{x}_0)\right)\left(\bar{f}(\mathbf{x}_0) - \hat{f}(\mathbf{x}_0)\right) && (13)
\end{aligned}
$$

Now take the expectation of the terms above w.r.t. X. For the last term in (13), we have

$$
\begin{aligned}
E_X\left[2\left(F(\mathbf{x}_0) - \bar{f}(\mathbf{x}_0)\right)\left(\bar{f}(\mathbf{x}_0) - \hat{f}(\mathbf{x}_0)\right)\right] &= 2\left(F(\mathbf{x}_0) - \bar{f}(\mathbf{x}_0)\right) E_X\left[\left(\bar{f}(\mathbf{x}_0) - \hat{f}(\mathbf{x}_0)\right)\right] && (14)\\
&= 0, && (15)
\end{aligned}
$$

the last step by definition of $\bar{f}$. Furthermore, note that the term in (11) does not depend on X. So, we have

$$
\begin{aligned}
E_X\left[\left(F(\mathbf{x}_0) - \hat{f}(\mathbf{x}_0)\right)^2\right] &= \underbrace{\left(F(\mathbf{x}_0) - \bar{f}(\mathbf{x}_0)\right)^2}_{\text{bias}^2} && (16)\\
&+ \underbrace{E_X\left[\left(\bar{f}(\mathbf{x}_0) - \hat{f}(\mathbf{x}_0)\right)^2\right]}_{\text{variance}}. && (17)
\end{aligned}
$$

Putting it all together with (2)-(5), we get the following decomposition of the expected error:

$$
\begin{aligned}
E_{X,\mathbf{x}_0,y_0}\left[\left(y_0 - \hat{f}(\mathbf{x}_0, X)\right)^2\right] &= \sigma_\nu^2 \qquad \text{(noise variance)} && (18)\\
&+ \int \left(F(\mathbf{x}_0) - \bar{f}(\mathbf{x}_0)\right)^2 p(\mathbf{x}_0) d\mathbf{x}_0 \qquad \text{(expected squared bias)} && (19)\\
&+ \int E_X\left[\left(\bar{f}(\mathbf{x}_0) - \hat{f}(\mathbf{x}_0)\right)^2\right] p(\mathbf{x}_0) d\mathbf{x}_0 \qquad \text{expected variance.} && (20)
\end{aligned}
$$