Derivation

March 29, 2022

1 Derivation of the Bias-Variance Decomposition

1.0.1 Goal

Decompose the model risk into recognizable components.

1.0.2 Step 1

$$\begin{aligned}
\text{model risk} &= \mathbb{E}\big((Y - \hat{Y}(x))^2\big) \\
&= \mathbb{E}\big((g(x) + \epsilon - \hat{Y}(x))^2\big) \\
&= \mathbb{E}\big((\epsilon + (g(x) - \hat{Y}(x)))^2\big) \\
&= \mathbb{E}(\epsilon^2) + 2\mathbb{E}(\epsilon(g(x) - \hat{Y}(x))) + \mathbb{E}\big((g(x) - \hat{Y}(x))^2\big)
\end{aligned}$$

On the right hand side:

- The first term is the observation variance σ^2.
- The cross product term is 0 because ϵ is independent of $g(x) - \hat{Y}(x)$ and $\mathbb{E}(\epsilon) = 0$
- The last term is the mean squared difference between our predicted value and the value of the true function at x

1.0.3 Step 2

At this stage we have

$$\text{model risk} = \text{observation variance} + \mathbb{E}\big((g(x) - \hat{Y}(x))^2\big)$$

We don't yet have a good understanding of $g(x) - \hat{Y}(x)$. But we do understand the deviation $D_{\hat{Y}(x)} = \hat{Y}(x) - \mathbb{E}(\hat{Y}(x))$. We know that

- $\mathbb{E}(D_{\hat{Y}(x)}) = 0$
- $\mathbb{E}(D^2_{\hat{Y}(x)}) = \text{model variance}$

So let's add and subtract $\mathbb{E}(\hat{Y}(x))$ and see if that helps.

$$g(x) - \hat{Y}(x) \;=\; (g(x) - \mathbb{E}(\hat{Y}(x)) + (\mathbb{E}(\hat{Y}(x) - \hat{Y}(x))$$

The first term on the right hand side is the model bias at x. The second term is $-D_{\hat{Y}(x)}$. So

$$g(x) - \hat{Y}(x) \;=\; \text{model bias} - D_{\hat{Y}(x)}$$

1.0.4 Step 3

Remember that the model bias at x is a constant, not a random variable. Think of it as your favorite number, say 10. Then

$$\begin{aligned}
\mathbb{E}\big((g(x) - \hat{Y}(x))^2\big) &= \text{model bias}^2 - 2(\text{model bias})\mathbb{E}(D_{\hat{Y}(x)}) + \mathbb{E}(D_{\hat{Y}(x)}^2) \\
&= \text{model bias}^2 - 0 + \text{model variance} \\
&= \text{model bias}^2 + \text{model variance}
\end{aligned}$$

1.0.5 Step 4: Bias-Variance Decomposition

In Step 2 we had

$$\text{model risk} \;=\; \text{observation variance} + \mathbb{E}\big((g(x) - \hat{Y}(x))^2\big)$$

Step 3 showed

$$\mathbb{E}\big((g(x) - \hat{Y}(x))^2\big) \;=\; \text{model bias}^2 + \text{model variance}$$

Thus we have shown the bias-variance decomposition

$$\text{model risk} \;=\; \text{observation variance} + \text{model bias}^2 + \text{model variance}$$

That is,

$$\mathbb{E}\big((Y - \hat{Y}(x))^2\big) \;=\; \sigma^2 + \mathbb{E}\big((g(x) - \mathbb{E}(\hat{Y}(x))^2\big) + \mathbb{E}\big((\hat{Y}(x) - \mathbb{E}(\hat{Y}(x))^2\big)$$

1.0.6 Special Case $\hat{Y}(x) = f_{\hat{\theta}}(x)$

In the case where we are making our predictions by fitting some function f that involves parameters θ, our estimate $\hat{Y}$ is $f_{\hat{\theta}}$ where $\hat{\theta}$ has been estimated from the data and hence is random.

In the bias-variance decomposition

$$\mathbb{E}\big((Y - \hat{Y}(x))^2\big) \;=\; \sigma^2 + \mathbb{E}\big((g(x) - \mathbb{E}(\hat{Y}(x))^2\big) + \mathbb{E}\big((\hat{Y}(x) - \mathbb{E}(\hat{Y}(x))^2\big)$$

just plug in the particular prediction $f_{\hat{\theta}}$ in place of the general prediction $\hat{Y}$:

$$\mathbb{E}\big((Y - f_{\hat{\theta}}(x))^2\big) \;=\; \sigma^2 + \mathbb{E}\big((g(x) - \mathbb{E}(f_{\hat{\theta}}(x))^2\big) + \mathbb{E}\big((f_{\hat{\theta}}(x) - \mathbb{E}(f_{\hat{\theta}}(x))^2\big)$$