Derivation of the Bias-Variance Decomposition

Goal

Decompose the model risk into recognizable components.

Step 1

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

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 $\mathrm{E}(\epsilon) = 0$
- The last term is the mean squared difference between our predicted value and the value of the true function at x

Step 2

At this stage we have

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

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

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

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

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

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

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

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

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

Step 4: Bias-Variance Decomposition

In Step 2 we had

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

Step 3 showed

$$\mathrm{E}\left((g(x) - \hat{Y}(x))^2\right) = \text{model bias}^2 + \text{model variance}$$

Thus we have shown the bias-variance decomposition

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

That is,

$$\mathrm{E}\left((Y - \hat{Y}(x))^2\right) = \sigma^2 + \mathrm{E}\left((g(x) - \mathrm{E}(\hat{Y}(x))^2\right) + \mathrm{E}\left((\hat{Y}(x) - \mathrm{E}(\hat{Y}(x))^2\right)$$

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

$$\mathrm{E}\left((Y - \hat{Y}(x))^2\right) = \sigma^2 + \mathrm{E}\left((g(x) - \mathrm{E}(\hat{Y}(x))^2\right) + \mathrm{E}\left((\hat{Y}(x) - \mathrm{E}(\hat{Y}(x))^2\right)$$

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

$$\mathrm{E}\left((Y - f_{\hat{\theta}}(x))^2\right) = \sigma^2 + \mathrm{E}\left((g(x) - \mathrm{E}(f_{\hat{\theta}}(x))^2\right) + \mathrm{E}\left((f_{\hat{\theta}}(x) - \mathrm{E}(f_{\hat{\theta}}(x))^2\right)$$