1 Bias-Variance Trade-Off

When we are talking the concept of bias-variance trade off, try to image that one could repeatedly building a model multiple times based on different data (i.e. by resampling or gathering new data). Based on it, the following definitions are given.

Definition 1.1. Bias is the difference between the average prediction among our repeatedly built models and the actual value which we are trying to predict. Therefore,

$$\text{Bias}\Big(\hat{f}(x_0)\Big) = \text{E}[\hat{f}(x_0)] - f(x_0) \tag{1}$$

where x_0 is the sample that is trying to be predicted and $f(x_0)$ is the true value without considering the effect of noise ϵ (i.e. the true value corresponds to x_0 is y_0 and $y_0 = f(x_0) + \epsilon$).

Definition 1.2. Variance is defining how the model prediction varies for a given data point when repeatedly building models based on different sample data. Therefore,

$$\text{Var}\Big(\hat{f}(x_0)\Big) = \text{E}[\hat{f}(x_0) - \text{E}(\hat{f}(x_0))]^2 \tag{2}$$

where x_0 is the sample that is trying to be predicted and $f(x_0)$ is the true value without considering the effect of noise ϵ (i.e. the true value corresponds to x_0 is y_0 and $y_0 = f(x_0) + \epsilon$).

Based on definition 1.1 and 1.2, we know that model with high bias pays very little attention to the training data and oversimplifies the model, and model with high variance pays a lot of attention to training data and does not generalize on the data which it has not seen before.

Now, let's focusing on proving

$$\text{E}\left(y_0 - \hat{f}(x_0)\right)^2 = \text{Var}\left(\hat{f}(x_0)\right) + \left[\text{Bias}\left(\hat{f}(x_0)\right)\right]^2 + \text{Var}(\epsilon) \tag{3}$$

Here, we also assume that $\epsilon \sim \mathcal{N}(0, \sigma^2)$, which means we assume that the noise is the gaussian white noise (i.e. mean = zero). For simplicity, the following derivative process uses y for representing y_0, f for representing $f(x_0)$, and $\hat{f}$ for representing $\hat{f}(x_0)$

$$\begin{aligned}
\text{E}\Big((y - \hat{f})\Big)^2 &= \text{E}\left[(f + \epsilon - \hat{f})^2\right] \\
&= \text{E}\left[\Big(f + \epsilon - \hat{f} + \text{E}(\hat{f}) - \text{E}(\hat{f})\Big)^2\right] \\
&= \text{E}\left[\Big((f - \text{E}(\hat{f})) - (\hat{f} - \text{E}(\hat{f})) + \epsilon\Big)^2\right] \\
&= \text{E}\Big[(f - \text{E}(\hat{f}))^2 + (\hat{f} - \text{E}(\hat{f}))^2 + \epsilon^2 - 2(f - \text{E}(\hat{f}))(\hat{f} - \text{E}(\hat{f})) \\
&\quad + 2\epsilon(f - \text{E}(\hat{f})) - 2\epsilon(\hat{f} - \text{E}(\hat{f}))\Big] \quad {}^{1} \\
&= \text{E}\Big[(f - \text{E}(\hat{f}))^2\Big] + \text{E}(\hat{f} - \text{E}(\hat{f}))^2 + \text{E}(\epsilon^2) \\
&\quad - 2\text{E}\Big[(f - \text{E}(\hat{f}))(\hat{f} - \text{E}(\hat{f}))\Big] + 2\text{E}\Big[\epsilon(f - \text{E}(\hat{f}))\Big] - 2\text{E}\Big[\epsilon(\hat{f} - \text{E}(\hat{f}))\Big]
\end{aligned} \tag{4}$$

[1] $\forall a, b, c \in R, (a + b + c)^2 = a^2 + b^2 + c^2 + 2ab + 2ac + 2bc.$

Since f is deterministic, $f - E(\hat{f})$ is a constant. Thus,

$$\begin{aligned}\mathrm{E}\Big((y-\hat{f})\Big)^2 &= \mathrm{E}\Big[(f-\mathrm{E}(\hat{f}))^2\Big] + \mathrm{E}(\hat{f}-\mathrm{E}(\hat{f}))^2 + \mathrm{E}(\epsilon^2) \\ &\quad - 2\mathrm{E}\Big[(f-\mathrm{E}(\hat{f}))(\hat{f}-\mathrm{E}(\hat{f}))\Big] + 2\mathrm{E}\Big[\epsilon(f-\mathrm{E}(\hat{f}))\Big] - 2\mathrm{E}\Big[\epsilon(\hat{f}-\mathrm{E}(\hat{f}))\Big] \\ &= (f-E(\hat{f}))^2 + \mathrm{E}(\hat{f}-E(\hat{f}))^2 + \mathrm{E}(\epsilon^2) \\ &\quad - 2(f-E(\hat{f}))\mathrm{E}(\hat{f}-E(\hat{f})) + 2(f-E(\hat{f}))\mathrm{E}(\epsilon) - 2\mathrm{E}(\epsilon)\mathrm{E}(\hat{f}-E(\hat{f}))\end{aligned} \quad (5)$$

[2]

Based on definition 1.1 and 1.2, we know that the term $(f-\mathrm{E}(\hat{f}))^2$ in (5) is $\Big[\mathrm{Bias}(\hat{f})\Big]^2$, and the term $\mathrm{E}(\hat{f}-\mathrm{E}(\hat{f}))^2$ is $\mathrm{Var}\Big(\hat{f}\Big)$. Also, based on our assumption that $\mathrm{E}(\epsilon)=0$, we know that $\mathrm{Var}(\epsilon)=\mathrm{E}(\epsilon^2)$. Therefore,

$$\mathrm{E}\Big((y-\hat{f})\Big)^2 = \mathrm{Var}\Big(\hat{f}\Big) + \Big[\mathrm{Bias}(\hat{f})\Big]^2 + \mathrm{Var}(\epsilon) - 2(f-E(\hat{f}))\mathrm{E}(\hat{f}-E(\hat{f})) \quad (6)$$

Since $\mathrm{E}(\hat{f}-E(\hat{f})) = \mathrm{E}(\hat{f}) - E(\hat{f}) = 0$, the final conclusion would be

$$\mathrm{E}\Big((y-\hat{f})\Big)^2 = \mathrm{Var}\Big(\hat{f}\Big) + \Big[\mathrm{Bias}(\hat{f})\Big]^2 + \mathrm{Var}(\epsilon) \quad (7)$$

2 References

[1] S, Fortmann-Roe, Understanding the Bias-Variance Tradeoff, June. 2012. Accessed on: http://scott.fortmann-roe.com/docs/BiasVariance.html

[2] Some background knowledge that should be known: 1. $\forall a,\, b \in R$, $\mathrm{E}(a+bX) = a + b\mathrm{E}(X)$; 2. If X, Y are two independent variables, then $\mathrm{E}(XY) = \mathrm{E}X \cdot \mathrm{E}Y$. That is why $\mathrm{E}\big[\epsilon(\hat{f}-\mathrm{E}(\hat{f}))\big] = \mathrm{E}(\epsilon)\mathrm{E}(\hat{f}-E(\hat{f}))$.