

Assume $Y = f(x) + \varepsilon$, with $E(\varepsilon) = 0$

We predict Y with $\hat{Y} = \hat{f}(x)$

We have:

$$\begin{aligned} E((Y - \hat{Y})^2) &= E\left((f(x) + \varepsilon - \hat{f}(x))^2\right) \\ &= E\left((f(x) - \hat{f}(x))^2 + \varepsilon^2 + 2\varepsilon \cdot (f(x) - \hat{f}(x))\right) \\ &= E\left((f(x) - \hat{f}(x))^2\right) + E(\varepsilon^2) + 2E(\varepsilon \cdot (f(x) - \hat{f}(x))) \end{aligned}$$

We have

$$\text{Var}(\varepsilon) = E(\varepsilon^2) - \underbrace{(E(\varepsilon))^2}_{=0} = E(\varepsilon^2)$$

$$\text{Cov}(\varepsilon, f(x) - \hat{f}(x)) = E(\varepsilon \cdot (f(x) - \hat{f}(x))) - E(\varepsilon) \cdot E(f(x) - \hat{f}(x))$$

We know that $\hat{f}(x)$ depends on $Y_1, \dots, Y_N$, which are independent of Y , which is a new data point, so $\hat{f}(x)$ is independent of $Y = f(x) + \varepsilon$, hence, $\hat{f}(x)$ is independent of ε and

$$\text{Cov}(\varepsilon, \hat{f}(x)) = 0$$

$$\rightarrow \text{Cov}(\varepsilon, f(x) - \hat{f}(x)) = 0$$

$$\rightarrow E(\varepsilon \cdot (f(x) - \hat{f}(x))) = \underbrace{E(\varepsilon)}_{=0} \cdot E(f(x) - \hat{f}(x)) = 0$$

$$\rightarrow E((Y - \hat{Y})^2) = \underbrace{E((f(x) - \hat{f}(x))^2)}_{\text{Reducible}} + \underbrace{\text{Var}(\varepsilon)}_{\text{Non-reducible}}$$

Recall that

$$E(Y|X=x) = \int y p(y|x) dy = \int y \frac{p(y,x)}{p(x)} dy$$

We have :

$$\begin{aligned} E((f(x) - \hat{f}(x))^2) &= E\left((f(x) - E(\hat{f}(x)) + E(\hat{f}(x)) - \hat{f}(x))^2\right) \\ &= E\left((f(x) - E(\hat{f}(x)))^2 + (E(\hat{f}(x)) - \hat{f}(x))^2\right. \\ &\quad \left.+ 2(f(x) - E(\hat{f}(x))) \cdot (E(\hat{f}(x)) - \hat{f}(x))\right) \\ &= E\left(\underbrace{(f(x) - E(\hat{f}(x)))^2}_{= \text{Bias}}\right) + E\left(\underbrace{(\hat{f}(x) - E(\hat{f}(x)))^2}_{= \text{Var}(\hat{f}(x))}\right) \\ &\quad + 2(f(x) - E(\hat{f}(x))) \cdot \underbrace{E(E(\hat{f}(x)) - \hat{f}(x))}_{= E(\hat{f}(x)) - E(\hat{f}(x)) = 0} \\ &= (f(x) - E(\hat{f}(x)))^2 + \text{Var}(\hat{f}(x)) \\ &= \text{Bias}^2 + \text{Variance} \end{aligned}$$