Approximation error versus variability

Daniel Hsu (COMS 4771)

Theorem. Assume $(\vec{X}_1, Y_1), \ldots, (\vec{X}_n, Y_n)$ (training data) and $(\vec{X}, Y)$ are IID random examples, and also assume the predictor F only depends on the training data. For any feature vector $\vec{x}$,

$$\begin{aligned}
\mathbb{E}[(F(\vec{x}) - Y)^2 \mid \vec{X} = \vec{x}] &= (\mathbb{E}[F(\vec{x})] - \eta(\vec{x}))^2 && (\text{"approximation error"}) \\
&\quad + \operatorname{var}(F(\vec{x})) && (\text{"variability"}) \\
&\quad + \mathbb{E}[(Y - \eta(\vec{x}))^2 \mid \vec{X} = \vec{x}] && (\text{"error of the optimal predictor"})
\end{aligned}$$

where $\eta(\vec{x}) := \mathbb{E}[Y \mid \vec{X} = \vec{x}]$.

Proof. By assumption, $F(\vec{x})$ is independent of $(\vec{X}, Y)$. We shall use this fact a couple of times below.

Let us use the notation $\mathbb{E}_{\vec{x}}[\,\cdot\,]$ for conditional expectation given $\vec{X} = \vec{x}$, i.e., $\mathbb{E}[\,\cdot \mid \vec{X} = \vec{x}]$. By the tower property of conditional expectation,

$$\mathbb{E}_{\vec{x}}[(F(\vec{x}) - Y)^2] = \mathbb{E}_{\vec{x}}\Big[\underbrace{\mathbb{E}_{\vec{x}}[(F(\vec{x}) - Y)^2 \mid F(\vec{x})]}_{(*)}\Big] \tag{1}$$

Let us first consider the "inner" conditional expectation $(*)$. By the bias-variance decomposition,

$$\begin{aligned}
\mathbb{E}_{\vec{x}}[(F(\vec{x}) - Y)^2 \mid F(\vec{x})] &= (F(\vec{x}) - \mathbb{E}_{\vec{x}}[Y \mid F(\vec{x})])^2 && (\text{"squared bias"}) \\
&\quad + \mathbb{E}_{\vec{x}}[(Y - \mathbb{E}_{\vec{x}}[Y \mid F(\vec{x})])^2 \mid F(\vec{x})] && (\text{"variance"}) \\
&= (F(\vec{x}) - \mathbb{E}_{\vec{x}}[Y])^2 \\
&\quad + \mathbb{E}_{\vec{x}}[(Y - \mathbb{E}_{\vec{x}}[Y])^2] \\
&= (F(\vec{x}) - \eta(\vec{x}))^2 + \mathbb{E}_{\vec{x}}[(Y - \eta(\vec{x}))^2]
\end{aligned}$$

where the second step uses the independence of $F(\vec{x})$ and $(\vec{X}, Y)$. Plugging this back into Equation (1), we obtain

$$\begin{aligned}
\mathbb{E}_{\vec{x}}[(F(\vec{x}) - Y)^2] &= \mathbb{E}_{\vec{x}}[(F(\vec{x}) - \eta(\vec{x}))^2] + \mathbb{E}_{\vec{x}}[(Y - \eta(\vec{x}))^2] \\
&= \underbrace{\mathbb{E}[(F(\vec{x}) - \eta(\vec{x}))^2]}_{(**)} + \mathbb{E}_{\vec{x}}[(Y - \eta(\vec{x}))^2] \qquad (2)
\end{aligned}$$

where, again, the second step uses the independence of $F(\vec{x})$ and $(\vec{X}, Y)$. This first term on the right-hand side $(**)$ in Equation (2) can be written as

$$\begin{aligned}
\mathbb{E}[(F(\vec{x}) - \eta(\vec{x}))^2] &= (\mathbb{E}[F(\vec{x})] - \eta(\vec{x}))^2 && (\text{"squared bias"}) \\
&\quad + \operatorname{var}(F(\vec{x})) && (\text{"variance"})
\end{aligned}$$

by the bias-variance decomposition. Plugging this back into Equation (2) finally gives

$$\begin{aligned}
\mathbb{E}_{\vec{x}}[(F(\vec{x}) - Y)^2] &= (\mathbb{E}[F(\vec{x})] - \eta(\vec{x}))^2 \\
&\quad + \operatorname{var}(F(\vec{x})) \\
&\quad + \mathbb{E}_{\vec{x}}[(Y - \eta(\vec{x}))^2].
\end{aligned}$$

□