Notes on bias-variance tradeoff

BY XING CHAO

1 Bias–variance decomposition of squared error

Suppose

$$y(x) = f(x) + \varepsilon$$

where the noise ε has zero mean and variance σ^2. $\hat{f}(x)$ is to approximate the true $f(x)$ by using information from $y(x)$. For notational convenience, abbreviate $f = f(x)$, $\hat{f} = \hat{f}(x)$, $y = y(x)$.

$$\begin{aligned}
E(f - \hat{f})^2 &= E(f - E\hat{f} + E\hat{f} - \hat{f})^2 \\
&= E(f - E\hat{f})^2 + E(E\hat{f} - \hat{f})^2 + 2E(f - E\hat{f})E(E\hat{f} - \hat{f}) \\
&= E(f - E\hat{f})^2 + E(E\hat{f} - \hat{f})^2 \\
&= (\mathrm{Bias}(\hat{f}))^2 + \mathrm{Var}(\hat{f})
\end{aligned}$$

2 Least square estimation

Linear regression problem can be represented as

$$Y_{n\times 1} = X_{n\times n}\theta_{n\times 1} + \varepsilon_{n\times 1}$$

where ε is noise with n row and 1 column. $E(\varepsilon\varepsilon^T) = \sigma^2 I$.

The ordinary least square estimation of θ is

$$\begin{aligned}
\hat{\theta} &= \arg\min_{\theta} (Y - X\theta)^T (Y - X\theta) \\
&= (X^T X)^{-1} X^T Y \\
&= (X^T X)^{-1} X^T (X\theta + \varepsilon) \\
&= \theta + (X^T X)^{-1} X^T \varepsilon \\
E\hat{\theta} &= \theta \\
\mathrm{Var}\hat{\theta} &= (X^T X)^{-1} X^T E(\varepsilon\varepsilon^T) X (X^T X)^{-1} \\
&= \sigma^2 (X^T X)^{-1}
\end{aligned}$$

The regularized least square estimation of θ is

$$\begin{aligned}
\hat{\theta}_r &= \arg\min_{\theta} (Y - X\theta)^T (Y - X\theta) + \lambda\theta^T\theta \\
&= (X^T X + \lambda I)^{-1} X^T Y \\
&= (X^T X + \lambda I)^{-1} X^T (X\theta + \varepsilon) \\
&= (X^T X + \lambda I)^{-1} X^T X\theta + (X^T X + \lambda I)^{-1} X^T \varepsilon \\
E\hat{\theta}_r &= (X^T X + \lambda I)^{-1} X^T X\theta \\
\mathrm{Var}\hat{\theta}_r &= (X^T X + \lambda I)^{-1} X^T E(\varepsilon\varepsilon^T) X (X^T X + \lambda I)^{-1} \\
&= \sigma^2 (X^T X + \lambda I)^{-1} X^T X (X^T X + \lambda I)^{-1}
\end{aligned}$$

When $\lambda > 0$, there are $E\hat{\theta}_r > E\hat{\theta}, \mathrm{Var}\hat{\theta}_r < \mathrm{Var}\hat{\theta}$. It is a better choice to use regularized least square estimation in some cases to decrease squared estimating error $(\theta - \hat{\theta})^2$.

3 Mean filtering

Suppose there is a signal with random noise as

$$y(n) = f(n) + \varepsilon(n)$$

$\varepsilon(n)$ is identical independence random variable with $E\varepsilon(n) = 0, \mathrm{Var}(\varepsilon(n)) = \sigma^2$

The value of signal $f(n)$ can be estimated by using mean filtering on $y(n)$

$$\begin{aligned}
\hat{f}_3(n) &= \frac{1}{3}\sum_{i=-1}^{1} y(n+i) \\
E\hat{f}_3(n) &= \frac{1}{3}\sum_{i=-1}^{1} f(n+i) \\
\mathrm{Var}\,\hat{f}_3(n) &= \frac{\mathrm{Var}(\varepsilon)}{3} \\
&= \frac{\sigma^2}{3}
\end{aligned}$$

When using $\hat{f}_1(n) = y(n)$ to estimate $f(n)$ directly,

$$\begin{aligned}
E\hat{f}_1(n) &= f(n) \\
\mathrm{Var}\big(\hat{f}_1(n)\big) &= \sigma^2
\end{aligned}$$

Bias is increased but variance is decreased in mean filtering.