Bias-variance tradeoff

UATX Statistical Learning

Bias–variance decomposition

Let $\hat{f}(x)$ be a noisy estimate of some function $f(x)$, evaluated at some point x. Define the mean-squared error of the estimate as

$$\text{MSE}(\hat{f}, f) = \text{E}\{[f(x) - \hat{f}(x)]^2\}\,.$$

Prove that $\text{MSE}(f, \hat{f}) = B^2 + v$, where

$$B = \text{E}\{\hat{f}(x)\} - f(x) \quad \text{and} \quad V = \text{var}\{\hat{f}(x)\}\,.$$

A simple example

Some people refer to the above decomposition as the *bias–variance trade-off.* Why a tradeoff? Here's a simple example to convey the intuition.

Suppose we observe $x_1, \dots, x_n$ from some distribution F, and want to estimate $f(0)$, the value of the probability density function at 0. Let h be a small positive number, called the *bandwidth*, and define the quantity

$$\pi_h = P\left(-\frac{h}{2} < X < \frac{h}{2}\right) = \int_{-h/2}^{h/2} f(x)dx\,.$$

Clearly $\pi_h \approx hf(0)$.

(A) Let Y be the number of observations in a sample of size n that fall within the interval $(-h/2, h/2)$. What is the distribution of Y? What are its mean and variance in terms of n and π_h? Propose a simple estimator $\hat{f}(0)$ involving Y.

(B) Suppose we expand $f(x)$ in a second-order Taylor series about 0:

$$f(x) \approx f(0) + xf'(0) + \frac{x^2}{2}f''(0)\,.$$

Use this in the above expression for π_h, together with the bias–variance decomposition, to show that

$$\text{MSE}\{\hat{f}(0), f(0)\} \approx Ah^4 + \frac{B}{nh}$$

for constants A and B that you should (approximately) specify. What happens to the bias and variance when you make h small? When you make h big?

(C) Use this result to derive an expression for the bandwidth that minimizes mean-squared error, as a function of n. You can approximate any constants that appear, but make sure you get the right functional dependence on the sample size.