Notes: bias-variance tradeoff

CSCI145/MATH166, Mike Izbicki

We generate data according to the process

$$t \sim f(\mathbf{x}) + \epsilon \tag{1}$$

where the $\sim$ symbol should be read as "sampled from" or "has distribution." The f function is unknown, and our goal is to estimate it from a sample $D = \{(t_1, \mathbf{x}_1), ..., (t_n, \mathbf{x}_n)\}$ using a parametric family of functions y. This family is often defined to be linear

$$y(\mathbf{x}; \mathbf{w}) = \mathbf{x}^T \mathbf{w}, \tag{2}$$

but it does not have to be. Our only assumption is that there exists some value of $\mathbf{w}$ such that $f(\cdot) = y(\cdot; \mathbf{w})$. (This is sometimes called the **realizability assumption**.)

We measure the quality of our estimator y using the squared error, which is defined as

$$\text{squared error} = (t - y(\mathbf{x}; D))^2 \tag{3}$$

where $y(\mathbf{x}; D) = y(\mathbf{x}; \hat{\mathbf{w}})$ and $\hat{\mathbf{w}}$ is the parameter estimate of $\mathbf{w}$ on dataset D. The squared error of an estimator can be decomposed into three terms:

$$\text{squared error} = \text{bias}^2 + \text{variance} + \text{noise}, \tag{4}$$

where each term is defined to be

$$\text{bias} = \left| \mathop{\mathbb{E}}_D y(\mathbf{x}; D) - f(x) \right| \tag{5}$$

$$\text{variance} = \mathop{\mathbb{E}}_D \left(y(\mathbf{x}; D) - \mathop{\mathbb{E}}_D y(\mathbf{x}; D) \right)^2 = \mathop{\mathbb{E}}_D y(\mathbf{x}; D)^2 - (\mathop{\mathbb{E}}_D y(\mathbf{x}; D))^2 \tag{6}$$

$$\text{noise} = (t - f(x))^2 = \epsilon^2 \tag{7}$$

The definitions above are (slightly) different than the definitions in Bishop (Eq. 3.41-3.44). Bishop takes the expectation of everything with respect to $\mathbf{x}$ and t (which I haven't done above) to get the *mean* squared error.

Theorem 1 (informal)**.** For any "reasonable" maximum likelihood problem,

$$\text{bias} = O(n^{-1}) \qquad \text{and} \qquad \text{variance} = \Theta(n^{-1}). \tag{8}$$

The bias may decay at a faster rate, and in particular may be zero. But for the variance, this rate is tight.

Problem 1. If n is large, which is larger: bias or variance?

Problem 2. In the limit as $n \to \infty$, what does the squared error equal?

Visualizing bias and variance

	high variance	low variance
high bias	. $\mathbf{w}^*$	. $\mathbf{w}^*$
low bias	. $\mathbf{w}^*$	. $\mathbf{w}^*$

How to control the squared error

Action	bias	variance	noise
adding more data			
more complex model			
stronger regularization/prior			
adding more features			