Risk and Bias-Variance Tradeoff

Ed Scerbo

Suppose pairs (x, y) with $y \in \mathbb{R}$ are drawn from a population with probability distribution D. We wish to estimate y in terms of x. We choose a hypothesis class $\mathcal{H}$ and a nonnegative pointwise loss function $l = l(y, \hat{y})$; i.e., l computes the error between an observed value y and our prediction $\hat{y}$. Given some **training set** $S = ((x_i, y_i))_{i=1}^n$ constructed by making independent random draws from D, we generally try to minimize the average pointwise loss made by a hypothesis $h \in \mathcal{H}$:

$$L(h) = \frac{1}{n} \sum_{i=1}^{n} l(y_i, h(x_i)).$$

That is, we try to find $\hat{h} \in \mathcal{H}$ that minimizes L.

The first purpose of this note is to explain why we usually use this definition of L, as opposed to, say, maximum pointwise loss.

Note: 1) If we wish to emphasize the role of the training set S, we shall write L_S, $\hat{h}_S$ in place of L, $\hat{h}$.

2) If $\mathcal{H}$ is a parametric hypothesis class, say $\mathcal{H} = \{f_\theta : \theta \in \Theta\}$, it is common to write $L(\theta)$ instead of $L(h)$.

Definition. The **true risk**, or simply **risk**, of a particular hypothesis $h \in \mathcal{H}$ using the pointwise loss function l is

$$R(h) = \mathbb{E}_{(x,y)}[l(y, h(x))],$$

where the expectation is over all points (x, y) in the population.

That is, $R(h)$ is the average error that h makes over the entire population. We would like to minimize $R(h)$, i.e., find a hypothesis h^* that yields, on average, the least error. In practice, of course, we have no hope of finding h^*,

as computing $R(h)$ requires we have access to the entire population, which we ordinarily will not.

Instead, select a sample $S = ((x_i, y_i))_{i=1}^n$ as above by making independent random draws from the population. If n is large, the distribution of S will likely approximate the underlying distribution D.

Definition. Given S, the **empirical risk** is the true risk where we treat S as a population in its own right, i.e.,

$$\hat{R}(h) = \hat{R}_S(h) = \mathbb{E}[l(y, h(x))],$$

where the expectation is over all points in S.

Each point (x_i, y_i) in S has equal probability $\frac{1}{n}$ of being selected from S. Thus we compute

$$\hat{R}(h) = \mathbb{E}[l(y, h(x))] = \frac{1}{n}\sum_{i=1}^{n} l(y_i, h(x_i)) = L(h)$$

and see that empirical risk equals average pointwise loss, which is why we use that definition for L above. Recall our original goal was to find a hypothesis h^* minimizing the true risk R. Since this is impossible and since S yields an approximation of D, we instead estimate h^* by finding a **tuned hypothesis** $\hat{h}_S$ that minimizes $\hat{R}_S = L_S$. Since there may be several optimal hypotheses for a given S, we assume that we have some (deterministic) function associating to each S a unique $\hat{h}_S$. We emphasize that $\hat{h}_S$ is a random quantity depending on the particular training set S drawn.

Question: How do the estimates $\hat{h}_S$ vary with different training sets S?

In order to say anything about this, we specialize to the case where our pointwise loss function is squared error, i.e., $l(y, \hat{y}) = (y - \hat{y})^2$. We also make the assumption that the distribution D has a density, which allows us to consider conditional expectations. With these assumptions, it is clear that for a given x, the best prediction we can hope for is the expected value of y over all points (x, y) with that particular value of x. Set $f(x) = \mathbb{E}_{y|x}[y|x]$, and write $y = f(x) + \epsilon_x$, where ϵ_x is defined by this equality. Note that ϵ_x is a random variable, being a function of y. Clearly for each x, $\mathbb{E}_{y|x}[\epsilon_x] = 0$. Also set $\sigma_x^2 = \mathrm{Var}_{y|x}(\epsilon_x)$.

Definition. The **bias** of our hypothesis class $\mathcal{H}$ at a point x is

$$\text{Bias}(x) = \mathbb{E}_S[f(x) - \hat{h}_S(x)]^2,$$

and the **variance** at x is

$$\text{Var}(x) = \text{Var}_S(\hat{h}_S(x)) = \mathbb{E}_S[(\hat{h}_S(x) - \mathbb{E}_S[\hat{h}_S(x)])^2].$$

As denoted, the expectations are over all possible training sets S. Global analogs of these may be defined as: $\text{Bias} = \mathbb{E}_x[\text{Bias}(x)]$, $\text{Var} = \mathbb{E}_x[\text{Var}(x)]$.

That is, for a given x, $\text{Bias}(x)$ measures how far on average the predictions $\hat{h}_S(x)$ deviate from the value $f(x)$ we are trying to approximate, and $\text{Var}(x)$ measures the variance of the values $\hat{h}_S(x)$ for different training sets S. Note that what we have called $\text{Bias}(x)$ is called $\text{Bias}(x)^2$ by some authors.

Theorem 1 (Bias-Variance Tradeoff)**.** *For each x, the mean squared error of the predictions $\hat{h}_S(x)$ made by the various tuned hypotheses is*

$$\mathbb{E}_{S,y|x}[(y - \hat{h}_S(x))^2] = \text{Bias}(x) + \text{Var}(x) + \sigma_x^2,$$

where, as indicated, the expectation is over all training sets S and the various values of y for the given x. Thus, the average true risk for our tuned hypotheses $\hat{h}_S$ over all training sets S is

$$\mathbb{E}_S[R(\hat{h}_S)] = \text{Bias} + \text{Var} + \text{Noise},$$

where $\text{Noise} = \mathbb{E}_x[\sigma_x^2]$.

Proof.

$$\begin{aligned}
\mathbb{E}_{S,y|x}[(y - \hat{h}_S(x))^2] &= \mathbb{E}_{S,y|x}[((f(x) - \hat{h}_S(x)) + \epsilon_x)^2] \\
&= \mathbb{E}_S[(f(x) - \hat{h}_S(x))^2] + 2\mathbb{E}_{y|x}[\epsilon_x]\mathbb{E}_S[f(x) - \hat{h}_S(x)] + \mathbb{E}_{y|x}[\epsilon_x^2] \\
&= \mathbb{E}_S[f(x) - \hat{h}_S(x)]^2 + \text{Var}_S(f(x) - \hat{h}_S(x)) + \sigma_x^2 \\
&= \text{Bias}(x) + \text{Var}(x) + \sigma_x^2,
\end{aligned}$$

where the third equality holds because $\mathbb{E}_{y|x}[\epsilon_x] = 0$ and because for any random variable X, $\mathbb{E}[X^2] = \mathbb{E}[X]^2 + \text{Var}(X)$.

To get the second result, simply note that

$$\mathbb{E}_S[R(\hat{h}_S)] = \mathbb{E}_S[\mathbb{E}_{(x,y)}[(y - \hat{h}_S(x))^2]] = \mathbb{E}_x[\mathbb{E}_{S,y|x}[(y - \hat{h}_S(x))^2]]$$

and apply the above. □