CSCI 145 Problem 8 Solution: How Much Does Training Error Hide?

Solution

1. Expected Training and Test Errors

For a fixed group G of size $m = n/k$, write

$$\overline{\varepsilon}_G = \frac{1}{m} \sum_{j \in G} \varepsilon_j.$$

Every sensor $i \in G$ receives prediction $\widehat{Y}_i = \mu + \overline{\varepsilon}_G$. Independence and centering give

$$\mathbb{E}[\overline{\varepsilon}_G] = 0, \qquad \mathbb{E}[\overline{\varepsilon}_G^2] = \frac{1}{m^2} \sum_{j \in G} \mathbb{E}[(\varepsilon_j)^2] = \frac{\sigma^2}{m}.$$

Thus every prediction is unbiased for μ, but its variance grows when we use smaller groups.

Training Error and Test Error

The training residual for sensor $i \in G$ is $\varepsilon_i - \overline{\varepsilon}_G$. These two terms are dependent: the group average includes ε_i. Their expected product is

$$\begin{aligned} \mathbb{E}[\varepsilon_i \overline{\varepsilon}_G] &= \frac{1}{m} \left(\mathbb{E}[(\varepsilon_i)^2] + \sum_{\substack{j \in G \\ j \neq i}} \mathbb{E}[\varepsilon_i \varepsilon_j] \right) \\ &= \frac{\sigma^2}{m}. \end{aligned}$$

Only products involving distinct, independent errors have expectation zero. Expanding the squared residual therefore gives

$$\begin{aligned} \mathbb{E}[(Y_i - \widehat{Y}_i)^2] &= \mathbb{E}[(\varepsilon_i)^2] - 2\mathbb{E}[\varepsilon_i \overline{\varepsilon}_G] + \mathbb{E}[\overline{\varepsilon}_G^2] \\ &= \sigma^2 - \frac{2\sigma^2}{m} + \frac{\sigma^2}{m} = \sigma^2 \left(1 - \frac{1}{m}\right). \end{aligned}$$

Averaging over the n sensors yields

$$\mathbb{E}[\mathcal{L}_{\text{train}}] = \sigma^2 \left(1 - \frac{k}{n}\right).$$

The test residual is $\varepsilon_i' - \bar{\varepsilon}_G$. The test error is independent of all initial errors, so

$$\mathbb{E}[\varepsilon_i' \bar{\varepsilon}_G] = \mathbb{E}[\varepsilon_i'] \mathbb{E}[\bar{\varepsilon}_G] = 0.$$

Consequently,

$$\begin{aligned} \mathbb{E}[(Y_i' - \widehat{Y}_i)^2] &= \mathbb{E}[(\varepsilon_i')^2] + \mathbb{E}[\bar{\varepsilon}_G^2] \\ &= \sigma^2 + \frac{\sigma^2}{m}, \end{aligned} \qquad \mathbb{E}[\mathcal{L}_{\text{test}}] = \sigma^2 \left(1 + \frac{k}{n}\right).$$

The first term is the new reading's noise. The second is noise retained in the fitted prediction. Subtracting the expected training error gives

$$\mathbb{E}[\mathcal{L}_{\text{test}} - \mathcal{L}_{\text{train}}] = \frac{2k\sigma^2}{n}.$$

2. Choosing How Much to Pool

Among these fixed grouping rules, expected training error decreases with k, while expected test error increases with k. At the two extremes,

	$\mathbb{E}[\mathcal{L}_{\text{train}}]$	$\mathbb{E}[\mathcal{L}_{\text{test}}]$
$k = 1$	$\sigma^2(1 - 1/n)$	$\sigma^2(1 + 1/n)$
$k = n$	0	$2\sigma^2$

When $k = n$, each prediction is exactly its sensor's initial reading. This always achieves zero training error, the smallest possible value, by memorizing the initial noise. Other choices can tie on particular samples. Here, the zero-error training fit has the largest expected test error among the grouping rules.

Pooling all sensors, $k = 1$, has the smallest expected test error among the grouping rules. Because every sensor has the same true mean, separate predictions fit differences due to noise and increase prediction variance. Choosing k by training error favors these noisier fits.