How Much Does Training Error Hide?

Due: Wednesday, October 7 at 11:59pm

Instructions: Please submit your *handwritten* solutions as a single PDF on Gradescope.

Several sensors measure the same constant quantity μ, but their readings contain noise. We can give each sensor its own prediction or pool their readings into a shared prediction.

There are $n \geq 2$ sensors. Sensor i contributes a reading Y_i to the training set and a reading Y_i' to the test set:

$$Y_i = \mu + \varepsilon_i, \qquad Y_i' = \mu + \varepsilon_i'.$$

All $2n$ errors are independent, with mean zero and variance $\sigma^2 > 0$. Before observing any readings, partition the sensors into k equal groups, where k divides n. Write $m = n/k$ and let $G(i)$ be sensor i's group. Predict its test reading using its group's initial average:

$$\widehat{Y}_i = \frac{1}{m} \sum_{\substack{j=1 \\ j \in G(i)}}^{n} Y_j.$$

Define the training-set and test-set mean squared errors by

$$\mathcal{L}_{\text{train}} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2, \qquad \mathcal{L}_{\text{test}} = \frac{1}{n} \sum_{i=1}^{n} (Y_i' - \widehat{Y}_i)^2.$$

1. The training error uses the readings used to fit the predictions. What is its expectation as a function of n, k, and σ^2?

2. The test error uses readings with fresh noise. What is its expectation as a function of n, k, and σ^2?

3. We can choose how many groups to use. Does choosing k by training error necessarily minimize expected test error?

Takeaway: *More separate predictions lower training error but raise expected test error in this example. Choosing how many predictions to fit using training error alone rewards fitting the noise.*