

An Observation About Subsection 6.6: On Average Training MSE Underestimates Validation MSE

The validating mean squared error is often used to evaluate the performance of a predictor. The expected value of the validating mean squared error taken over the distribution of training data y and validating data y_v , conditional on known covariates, is shown in equation (6.65) on page 285:

$$E_{y,y_v}(\text{MSE}_v) = \sigma^2 + \frac{1}{N} \sum_i \text{Var}(\hat{y}_{v,i}) + \text{bias}^2, \quad (6.65)$$

where for the i th term in the sum, $\text{bias}^2(i) = (E(y_{v,i}) - E(\hat{y}_{v,i}))^2$.

On the other hand, the expected value of the training mean squared error, taken over the distribution of training data y , conditional on known covariates, is shown in equation (6.66):

$$E_y(\text{MSE}_t) = \sigma^2 + \frac{1}{N} \sum_i \text{Var}(\hat{y}_i) + \text{bias}^2 - \frac{2}{N} \sum_i \text{Cov}(y_i, \hat{y}_i), \quad (6.66)$$

where for the i th term in the sum, $\text{bias}^2(i) = (E(y_i) - E(\hat{y}_i))^2$, indicating that the training mean squared error

$$\text{MSE}_t = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2$$

is a poor estimate of (6.65).

The predicted values in training and validating data are $\hat{y}_i = \hat{f}(x_i)$ and $\hat{y}_{v,i} = \hat{f}(x_{v,i})$, a function of known covariates x_i and $x_{v,i}$, respectively. When these covariates do not take the same values in training and validating data, the difference between (6.65) and (6.66) is a function of not only $\frac{2}{N} \sum_i \text{Cov}(y_i, \hat{y}_i)$ but also of the second and third terms of (6.65) and (6.66). When predictions are evaluated at the same values of the covariates in the training and validating data, $E(y_i) = E(y_{v,i})$, $\hat{y}_{v,i} = \hat{y}_i$ and $\frac{1}{N} \sum_i \text{Var}(\hat{y}_{v,i}) = \frac{1}{N} \sum_i \text{Var}(\hat{y}_i)$. The last two equalities hold when $\hat{f}(x_i) = x_i' \hat{b}$. This is not made explicit in the book. Then it follows that, exactly,

$$E_y(\text{MSE}_t) = E_{y,y_v}(\text{MSE}_v) - \frac{2}{N} \sum_i \text{Cov}(y_i, \hat{y}_i). \quad (6.67)$$

Evaluation of training and validating MSE at the same value of the covariates is less restrictive than may seem at first glance. When the objective is to obtain a measure of the validating MSE committing all the records as training data and the choice of value of the covariates is arbitrary, it is reasonable to perform the calculations using the covariates available of the training data.

In the least squares linear regression setting, when predictions $\hat{y}_i = x_i' \hat{b}$ are evaluated at the same values of the covariates, $X_t = X_v = X$, the second terms in (6.65) and (6.66)

take the form

$$\begin{aligned}\frac{1}{N} \sum_i \text{Var}(\hat{y}_i) &= \frac{1}{N} \text{tr} \left[\text{Var} \left(X\hat{b} | y, \sigma^2 \right) \right] \\ &= \frac{1}{N} \text{tr} \left[(X'X)^{-1} X'X \right] \sigma^2 \\ &= \frac{p}{N} \sigma^2\end{aligned}$$

where p is the number of columns of the $(N \times p)$ full column rank matrix X . In this setting, with bias = 0 and $\sum_i \text{Cov}(y_i, \hat{y}_i) = p\sigma^2$, the training mean squared error is equal to the maximum likelihood estimator of the residual variance in the normal linear regression model. Its expectation takes the form

$$\begin{aligned}\text{E}_y(\text{MSE}_t) &= \sigma^2 + \frac{p}{N} \sigma^2 - \frac{2p}{N} \sigma^2 \\ &= \frac{N-p}{N} \sigma^2.\end{aligned}$$

See page 267.

References