Cross-Validation :

- more like model selection

Bootstrap :

- more like model assessment

5.1 Cross-Validation

- training set , training error rate
- test set , test error rate

5.1.1 The Validation Set Approach

- Suppose we would like to estimate the test error associated with fitting a particular statistical learning method on a set of observations. The validation set approach involves randomly dividing the available set of observations into two parts, training set and validation set.
- But randomly dividing data into two parts is risky in that MSE varies greatly for each trial, and we cannot fully make use of the available data.

drawbacks

5.1.2 Leave-One-Out Cross-Validation

- Using only one observation as a test set and using the rest of the data as the training set. Repeat this for all individual data.

  => $n-1$ training observation, 1 test observation

$MSE_i = (Y_i - \hat{Y}_i)^2$, where $\hat{Y}_i$ is the fitted model without the i^{th} observation and Y_i is the actual value of the i^{th} observation.

there are n MSE_i's

=> $CV_{(n)} = \frac{1}{n} \sum_{i=1}^{n} MSE_i$

Pros of LOOCV :

- First, LOOCV has far less bias, consequently tending not to overestimate the test error rate. Second, there is no randomness, or variation, in MSE.

But!

- It is very time consuming and computationally loading.

5.1.3 K-Fold Cross-Validation

- randomly dividing the set of observations into k groups, and using each fold as a test set for each trial.

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} MSE_i$$

- requires significantly less computational support compared to LOOCV.
- Bias-Variance Trade-off

  ※ $MSE = Variance + Bias^2$
- MSE is stable for each trial.

5.1.4 Bias-Variance Trade-Off for k-Fold Cross-Validation

- In terms of bias reduction, it is obvious to choose LOOCV over k-fold for cross-validation since LOOCV uses almost full data.
- It turns out that LOOCV has higher variance than does k-fold. When we perform LOOCV, we are in effect averaging the outputs of n fitted models, each of which is trained on an almost identical set of observations; therefore, these outputs are highly, positively correlated. Since the mean of many highly correlated quantities has higher variance than does the mean of many quantities that are not as highly correlated, the test error estimate resulting from LOOCV tends to have higher variance than does the test error estimate resulting from k-fold CV.

→ MSE of LOOCV :

$CV = \frac{1}{n}\sum_{i=1}^{n} MSE_i = \overline{MSE}$, 여기서 X_i들끼리 "highly correlated" 하다는건, Covariance가 높다는 것. 이 상태에서 $\overline{MSE}$의 분산을 구하게 된다면

$$\Rightarrow Var(\overline{MSE}) = Var\left(\frac{1}{n}(MSE_1 + MSE_2 + \cdots + MSE_n)\right)$$
$$= \frac{1}{n^2}\cdot Var\left((MSE_1 + MSE_2 + \cdots + MSE_n)\right)$$
$$= \frac{1}{n^2}\left\{Var(MSE_1) + Var(MSE_2) + \cdots + Var(MSE_n) + 2Cov(MSE_1, MSE_2) + \cdots\right.$$

- 여기서 Covariance를 더하는 과정에서 MSE_i 들은 높은 상관관계를 가졌기 때문에 $Var(\overline{MSE})$가 매우 커진다.