5. Resampling Methods

Resampling Method란 training set에서 표본을 추출해 model fitting을 하는 과정을 반복하여 결과물이 불어나도록 하는 것이다. Training set을 전부 사용하여 단 한번 모델 피팅을 하는 것에 비해 훨씬 더 많은 정보를 얻을 수 있다. 물론 모델 피팅을 여러 번 하는 것이기 때문에 computational한 비용이 생긴다.

5장에서는 resampling method 중 두 가지 방법인 Cross-Validation과 Bootstrap을 소개한다.

5.1 Cross-Validation

Cross-Validation은 모델의 성능을 평가(model assessment)하기 위해, 혹은 적합한 유연성을 채택(model selection)하기 위해 사용할 수 있다.

Training error rate과 test error rate은 다르다. Training error rate은 정확한 수치를 알아낼 수 있는 반면에 test error rate은 실제 test set이 없으므로 알 수 없다. 이 문제를 해결하기 위해 training error rate을 이용하여 test error rate을 추정하는 방법이 있지만 이는 6장에서 다룬다. 5장에서는 training observation을 가지고 test error rate을 추정하는 방법을 논의한다.

5.1.1 The Validation Set Approach

가지고 있는 training observation을 둘로 나누어 하나는 training set, 다른 하나는 validation set(hold-out set)으로 지정한다. 후자가 pseudo test set의 역할을 한다고 볼 수 있다. Training set을 가지고 모델 학습을 거친 후 validation set에 대한 예측치를 구하여 validation set error rate을 정확히 구할 수 있다. 이를 바탕으로 모델의 성능을 평가한다.

그림5.2를 보면 이 경우 일차항만 있는 모델보다 더 높은 차수의 항을 포함하는 모델들의 MSE가 대체로 더 낮은 것을 쉽게 알 수 있다.

한 세트의 observation을 반으로 나누기를 반복한다면 매번 다른 test MSE 추정치가 나올 것이다. 이 과정을 10번 반복했다고 가정했을 때, 우측 그림에서 각 선은 서로 다른 수행을 의미한다. 높은 차수를 포함하는 모델의 MSE가 언제나 낮은 것을 다시 한번 확인할 수 있다. 그러나 각 수행의 변동이 너무 크다. 따라서 여기서 내릴 수 있는 결론은 이 데이터에는 선형회귀식이 적합하지 않다는 것뿐이다.

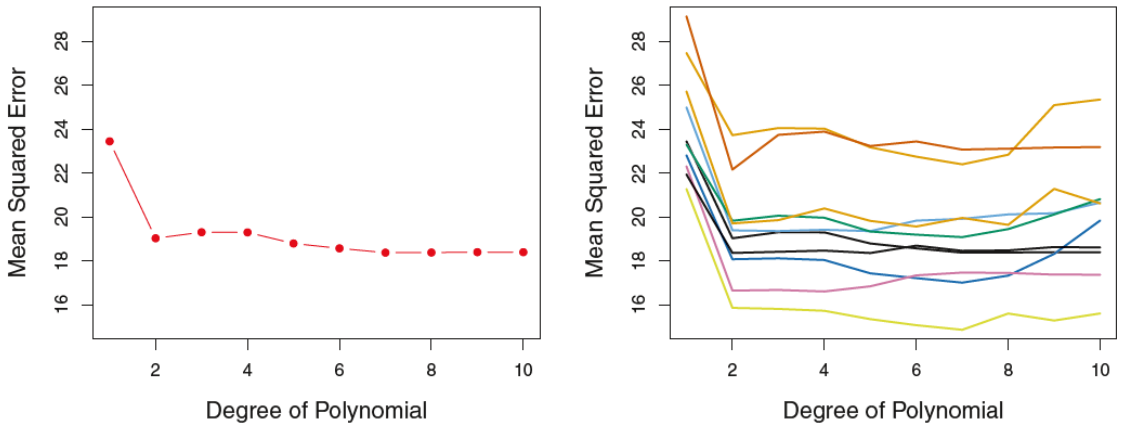

FIGURE 5.2. *The validation set approach was used on the* `Auto` *data set in order to estimate the test error that results from predicting* `mpg` *using polynomial functions of* `horsepower`. Left: *Validation error estimates for a single split into training and validation data sets.* Right: *The validation method was repeated ten times, each time using a different random split of the observations into a training set and a validation set. This illustrates the variability in the estimated test MSE that results from this approach.*

Validation set approach의 단점은 다음과 같다. 1. 기존의 observation set을 어떻게 나누었는가에 따라, 즉 validation set에 어떤 observation들이 포함되었는지에 따라 test error rate 추정치의 변동이 심하다. 2. 데이터를 둘로 나누었으므로 training에 활용할 수 있는 데이터가 줄어든다. 데이터가 적을수록 모델의 성능도 떨어지므로 실제 test error rate보다 높게 추정할 수 있다.

5.1.2 Leave-One-Out Cross-Validation

Validation set approach의 단점을 보완한다. 여기서도 마찬가지로 observation을 나누지만, 얼추 비슷한 몸집으로 두 set을 나누었던 validation set approach와 달리 LOOCV는 단 한 개의 observation만을 validation set으로 사용한다는 것이 특징이다. n개의 observation 중 n-1개를 가지고 모델 피팅을 한 후, 사용하지 않은 한 개의 observation에 대한 예측치를 구하여 실제 값과 비교하는 것이다. 여기서 구한 MSE는 단 하나의 값에 대한 것이므로 unbiased하지만 variance가 크다.

모든 observation이 한 번씩 validation set 역할을 하도록 위 과정을 n번 반복하면 각각에 해당하는 MSE 값이 나온다. 이를 평균내면 test MSE의 추정치를 구할 수 있다(수식5.1).

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{MSE}_i. \tag{5.1}$$

LOOCV의 장점은 다음과 같다. 1. training에 n-1개의 데이터를 모두 사용할 수 있다. 따라서 test error rate을 지나치게 높게 추정할 위험도 적어진다(lower bias). 2. Validation set을 임의로 뽑는 것이 아니라 단 한 개만 뽑아내므로 LOOCV를 여러 번 반복하더라도 결과의 변동성이 작다. 3. 다양한 모델링에 사용할 수 있다.

한편 데이터의 수만큼 모델 피팅을 n번 해야 하므로 몸집이 큰 데이터의 경우 시간이 굉장히 오래 걸릴 수 있다. 이 때 최소제곱법 회귀나 다항회귀인 경우에만 빠르게 계산할 수 있는 방법이 있다(수식5.2).

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - h_i}\right)^2, \tag{5.2}$$

5.1.3 k-Fold Cross-Validation

이전까지 데이터를 두개의 그룹으로 나누었다면 k-fold CV는 데이터를 k개의 비슷한 몸집의 그룹(fold)으로 나눈다. 이 중 하나의 fold가 validation set의 역할을 하고, 나머지 (k-1)개의 fold가 training set의 역할을 한다. 전자에 대하여 MSE를 구한 후 이 과정을 k번 반복해 k개의 MSE를 구하여 평균을 내면 test MSE의 추정치를 구할 수 있다(수식5.3).

$$\mathrm{CV}_{(k)} = \frac{1}{k}\sum_{i=1}^{k}\mathrm{MSE}_i. \tag{5.3}$$

여기서 k=n으로 지정한다면 LOOCV와 동일한 것이다. 그러나 일반적으로 k=5 이거나 k=10으로 지정한다(경험적으로 합의된 수치인 것 같다. High bias와 high variance로부터 비교적 자유롭다고 한다).

k-Fold CV는 다음과 같은 장점을 가진다. 1. LOOCV에 비해 computational한 비용이 훨씬 적게 들어간다. 2. MSE가 최소가 되게 하는 지점을 대강 알 수 있다(그림5.6). 실제 test MSE와 완전히 겹치지는 않지만 전체적인 형태나 골짜기를 이루는 지점이 엇비슷하게 나오기 때문이다.

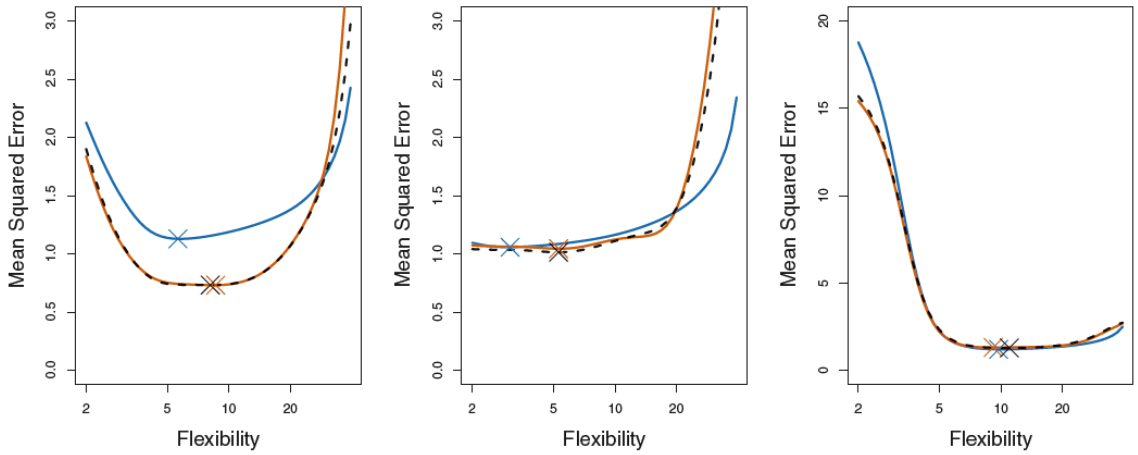

FIGURE 5.6. *True and estimated test MSE for the simulated data sets in Figures 2.9 (*left*), 2.10 (*center*), and 2.11 (*right*). The true test MSE is shown in blue, the LOOCV estimate is shown as a black dashed line, and the 10-fold CV estimate is shown in orange. The crosses indicate the minimum of each of the MSE curves.*

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

앞서 살펴본 validation set approach나 LOOCV는 bias가 지나치게 높거나 아주 낮았다. 그에 비해 k-fold CV는 (k=5 정도일 때) 전자보다는 많은, 그러나 후자보다는 적은 수의 표본을 사용하므로 중간 수준의 bias를 보일 것이다. 따라서 bias를 최소화하고자 한다면 LOOCV를 사용할 수 있다.

그러나 LOOCV는 거의 똑같은 training set에 대하여 반복적으로 모델 피팅을 수행하므로 test MSE 추정치들 사이의 정적 상관관계가 아주 높을 것이다. 상관관계가 높은 값들의 평균치는 높은 variance를 가진다. 이에 반해 k-fold CV는 매 수행마다 사용되는 데이터 간의 중복되는 정도가 덜하므로 상대적으로 낮은 상관관계와 variance를 갖게 된다. 즉 k의 값이 지나치게 커질(LOOCV) 경우 variance가 높아진다고 할 수 있다.

5.1.5 Cross-Validation on Classification Problems

실제 상황에서는 Bayes decision boundary와 test error rate을 알 수 없으므로 마찬가지로 cross validation을 한다. 이 때 MSE 대신 잘못 분류된 observation의 숫자를 n으로 나눈 값을 사용한다 (수식5.4).

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{Err}_i, \tag{5.4}$$

그림5.8과 같이 모델의 차수를 늘려가며, 혹은 k의 값에 변화를 주며 그래프를 그려보면 가장 낮은 test error rate이 어디쯤 에서 나올지 가늠할 수 있다. 여기서 10-fold CV가 실제 test error rate과 유사한 트렌드를 가지고 있음을 알 수 있다.

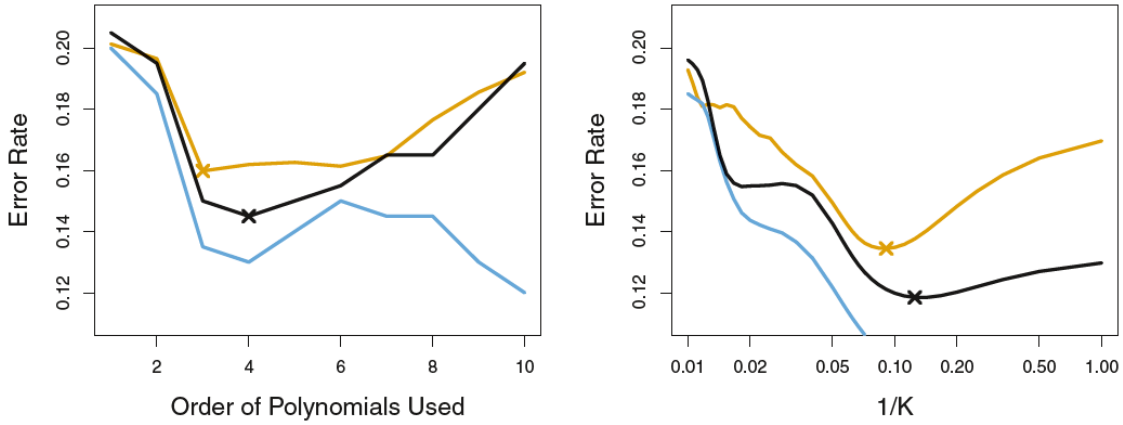

FIGURE 5.8. *Test error (brown), training error (blue), and 10-fold CV error (black) on the two-dimensional classification data displayed in Figure 5.7.* Left: *Logistic regression using polynomial functions of the predictors. The order of the polynomials used is displayed on the x-axis.* Right: *The KNN classifier with different values of K, the number of neighbors used in the KNN classifier.*

5.2 The Bootstrap

Bootstrap은 learning method나 parameter 추정의 정확도를 파악하기 위해 사용할 수 있다.

예를 들어 X와 Y라는 결과값을 주는 두가지 종목에 투자하여 이익을 얻고자 한다고 하자. 총 금액의 α만큼을 X에, 나머지 1- α만큼을 Y에 투자하여 가장 변동성이 낮은, 즉 안정적인 투자를 가능하게 하는 α값을 구하고 싶다. 다시 말해 $(\alpha X + (1-\alpha)Y)$의 분산이 최소가 되게 하는 α를 찾는 것이다. 이를 구하기 위해서는 X와 Y의 공분산과 각각의 분산을 추정해야 한다(수식5.7).

$$\hat{\alpha} = \frac{\hat{\sigma}_Y^2 - \hat{\sigma}_{XY}}{\hat{\sigma}_X^2 + \hat{\sigma}_Y^2 - 2\hat{\sigma}_{XY}}. \tag{5.7}$$

이 때 이 추정치들로 추정한 α의 정확도를 평가하기 위해서는 α를 추정하는 과정을 여러 번 반복하여 평균을 내고, 이를 활용하여 표준오차를 구해야 한다.

이를 위해 같은 모집단에서 표본을 무한정 추출할 수 있으면 좋겠지만, 실제 상황에서의 표본의 수는 한정되어 있다. 여기서 '모집단에서 새로운 표본을 추출하는 효과'를 한정된 표본만을 가지고 모방할 수 있게끔 하는 것이 bootstrap이다. 우리가 가지고 있는 observation에서 필요한 만큼 복원추출 하는 것이다. 이 방법을 통해 α의 추정치를 여러 개 얻어낼 수 있고 α의 추정치의 표준

오차 또한 추정할 수 있다.

모집단에서 표본을 1000번 추출한 경우와(좌측) 가지고 있는 데이터에서 bootstrap으로 1000번 복원추출 한 경우(가운데)의 그래프들을 비교해보면 아주 유사하다(그림5.10). Bootstrap이 매우 효과적이라는 것을 알 수 있다.

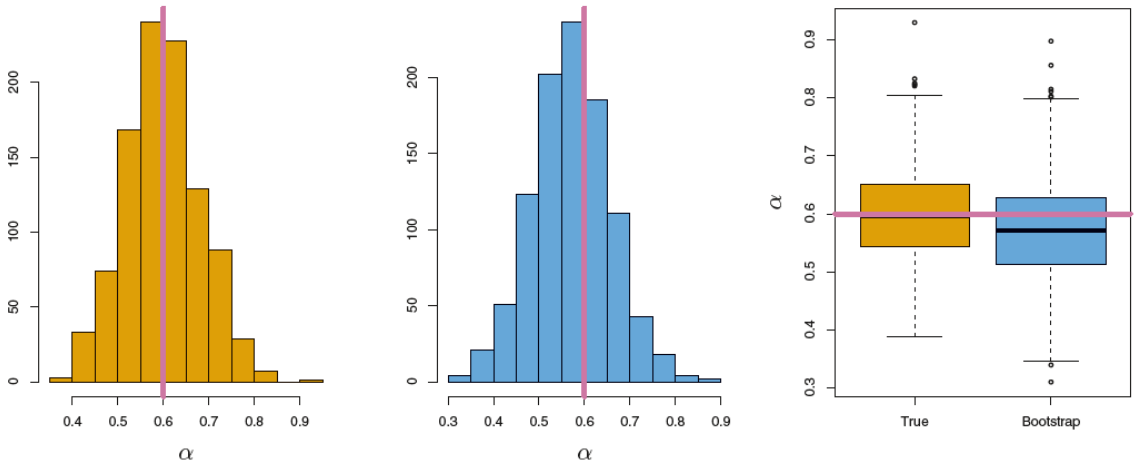

FIGURE 5.10. Left: *A histogram of the estimates of* α *obtained by generating 1,000 simulated data sets from the true population.* Center: *A histogram of the estimates of* α *obtained from 1,000 bootstrap samples from a single data set.* Right: *The estimates of* α *displayed in the left and center panels are shown as boxplots. In each panel, the pink line indicates the true value of* α.

참고문헌

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). New York: springer.