김은영

제5장 회귀모형의 선택

5.4 모형의 확인 (Model Validation[Checking])

5.4.1 교차확인 (Cross Validation)

- 예측오차를 최소로 하는 모형이 좋은 모형
- 예측오차에 대한 추정치로 교차확인(cross validation)을 이용

- 교차확인의 기본 아이디어

• 주어진 n개의 자료를 훈련자료(training data)와 시험자료(test data)의 두 부분으로 임의로 나누어서 사용한 다음 이를 바탕으로 예측오차를 추정한다.

(1) k-접힘 교차확인(k-fold cross validation)

- 집합 $\{1, \cdots, n\}$을 임의로 K개로 분할: $I_j \ (j=1, \cdots, K)$
- $\kappa : \{1, \cdots, n\} \to \{1, \cdots, K\}$
  '$\kappa(i)=j$'와 'i가 I_j에 속한다'가 같은 의미를 나타내도록 정의된 함수

- 원래의 자료 $\{(X_1, Y_1), \cdots, (X_n, Y_n)\}$에서 j번째 분할에 속하는 자료를 제외한 나머지 자료들을 이용하여 적합시킨 추정치를 $\hat{f}_{-j}$라고 하면, 예측오차에 대한 교차확인 추정치는 다음과 같이 정의된다.

$$\widehat{PE}_{CV} = \frac{1}{n}\sum_{i=1}^{n} L\left(Y_i, \ \hat{f}_{-\kappa(i)}(X_i)\right)$$

- 한점소거교차확인(Leave-One-Out CV; LOOCV): $K=n$인 특별한 경우

n개의 자료에서 i-번째$(i=1, \cdots, n)$ 자료를 제외한 나머지 $(n-1)$개의자료를 훈련자료로 사용하고 i-번째 자료를 시험자료로 사용한다.

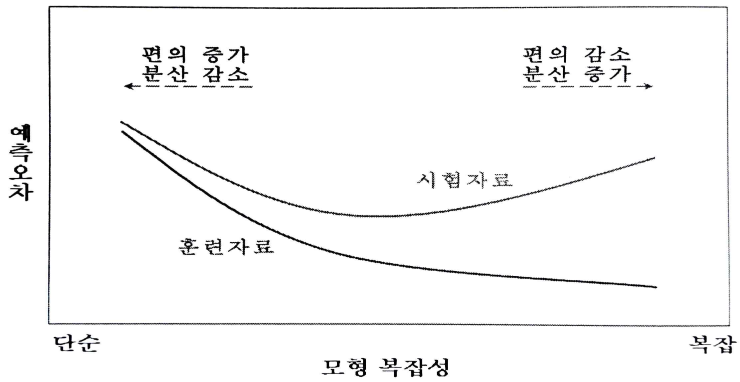
- 일반적으로

$$L_q - LOOCV = \sum_{i=1}^{n} |y_i - \hat{y}_{i(i)}|^q \qquad \text{여기서, } q > 0$$

• $LOOCV$는 $q=2$인 경우

	1	2	3	4	5	6	7	8	9	10
1	훈련자료									시험자료
2										
3										
4										
5										
6										
7										
8										
9										
10										

5.4.2 훈련자료와 시험자료, 그리고 예측오차



5.6 모형선택의 기타 논제

5.6.1 젠센 부등식과 칼백-라이블러 거리

[정리 5.1]

ϕ가 열린 구간(open interval) I에서 아래로 볼록함수(convex function)이고, 확률변수 X가 $P(X \in I) = 1$와 $E(X) < \infty$이면

$$\phi[E(X)] \le E[\phi(X)]$$

이 항상 성립되며 이를 **젠센 부등식(Jensen's Inequality)**이라 부른다.

[정의 5.1]

두 확률밀도함수 f와 g가 주어져 있을 때

g와 f 간의 칼백-라이블러 거리 [f에 대한 f와 g 간의 엔트로피 거리]는

$$E_f\left[\log\left\{\frac{f(X)}{g(X)}\right\}\right] = \int \log\left\{\frac{f(x)}{g(x)}\right\} f(x)\,dx$$

[예]

로그함수: 위로 볼록함수 → 칼백-라이블러 거리에 젠센 부등식을 이용하면

$$
\begin{aligned}
E_f\left[\log\left\{\frac{f(X)}{g(X)}\right\}\right] = E_f\left[-\log\left\{\frac{g(X)}{f(X)}\right\}\right] \\
&\geq -\log E_f\left\{\frac{g(X)}{f(X)}\right\} \\
&= -\log \int \frac{g(x)}{f(x)} f(x)dx \\
&= -\log 1 = 0
\end{aligned}
$$

※ $KL-distance \geq 0$