ISL
5.1.4~5.2

AI명예학회

SKHU

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

Cross Validation과 LOOCV

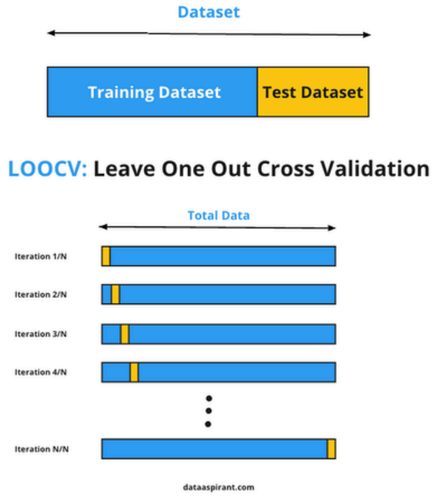
- LOOCV는 K-fold Cross Validation의 k = n인 특수한 경우라고 볼 수 있다.
- n이 늘어날 경우, LOOCV는 연산량이 늘어나는 문제가 있다.
- 연산량의 문제는 제하고, test error를 더 잘 추정할 수 있는 방법론은 무엇일까?

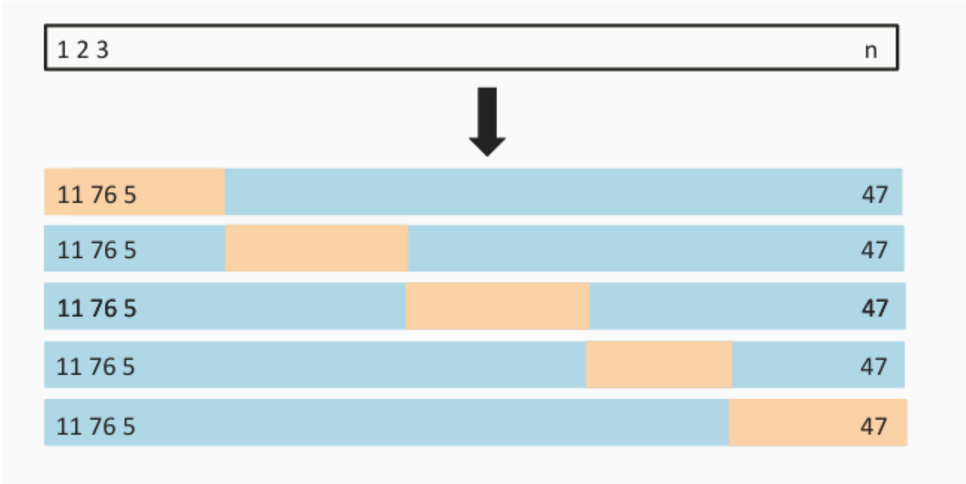
AI명예학회
SKHU

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

bias-variance trade off

- 전체 데이터 집합의 관측 수와 거의 동일한 n-1 개의 관측이 포함되어 있으므로 LOOCV는 테스트 오류에 대해 거의 편향되지 않은 추정치를 제공.
- Cross Validation은 어느 정도 편향이 발생



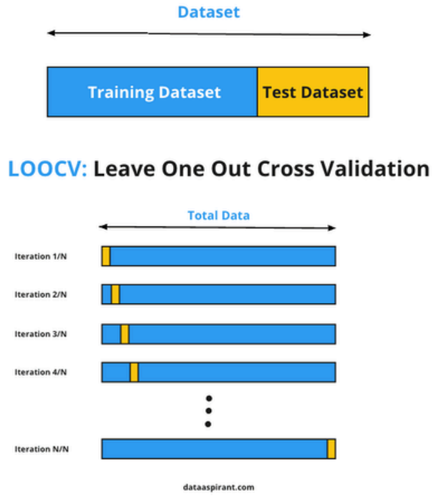


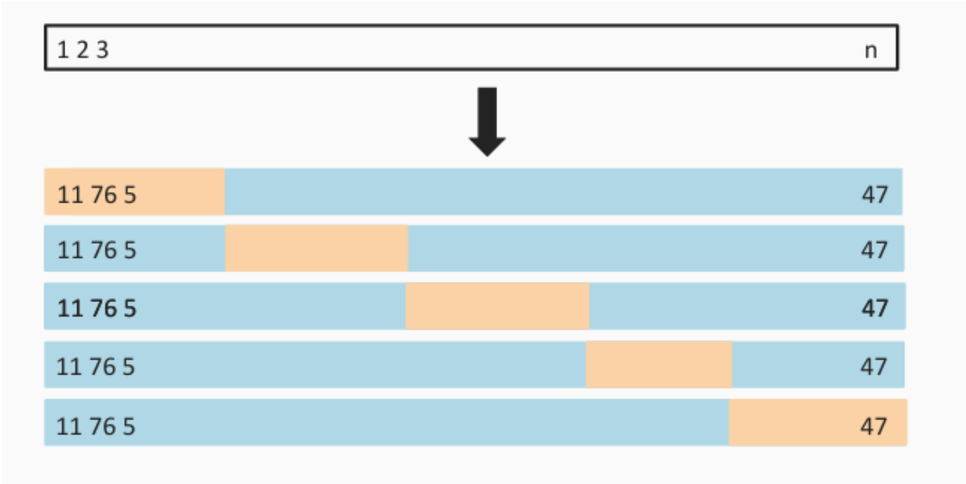
AI명예학회
SKHU

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

bias-variance trade off

- 거의 동일한 세트를 사용한 모델들의 평균을 사용
  -> 결과 간의 높은 상관관계
- Cross Validation은 서로 상관관계가 적은 k개의 모델 사용
- variance는 Cross Validation이 더 적다.





AI명예학회
SKHU

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

bias-variance trade off

- 거의 같은 데이터 세트를 사용하기 때문에, LOOCV는 훈련 데이터에 대한 과대적합 위험이 높다.

- K-fold Cross Validation는 k가 커질수록 bias는 작아지지만 variance는 커질 것이고, 적절한 k 값을 찾는 것이 중요하다.

- 일반적으로 5나 10을 사용한다.

5.1.5 on Classification

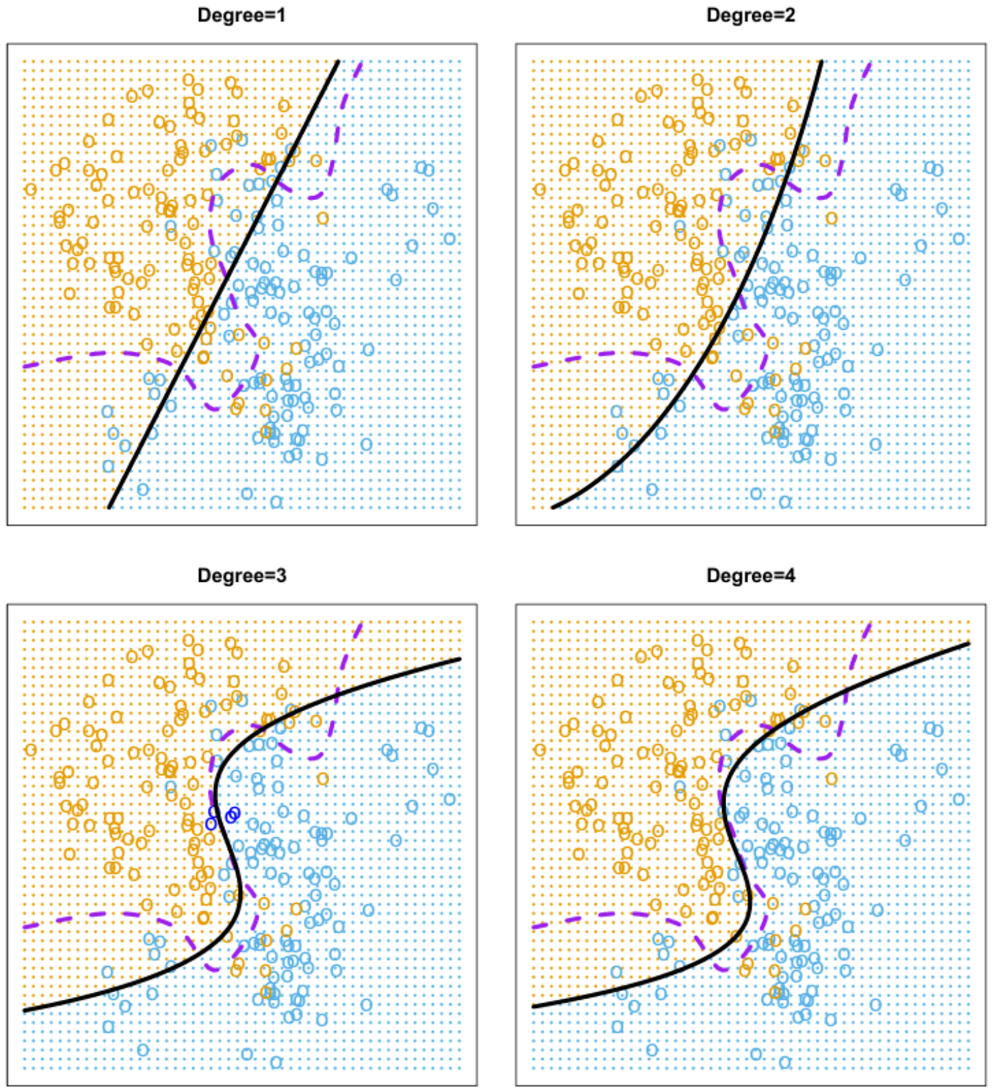
- Cross Validation은 분류 문제일 때도 효과적이다.
- 동일하게 적용되지만, test error를 측정할 때 MSE 대신 잘못 분류된 라벨의 수를 사용한다.

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

AI명예학회
SKHU

5.1.5 on Classification

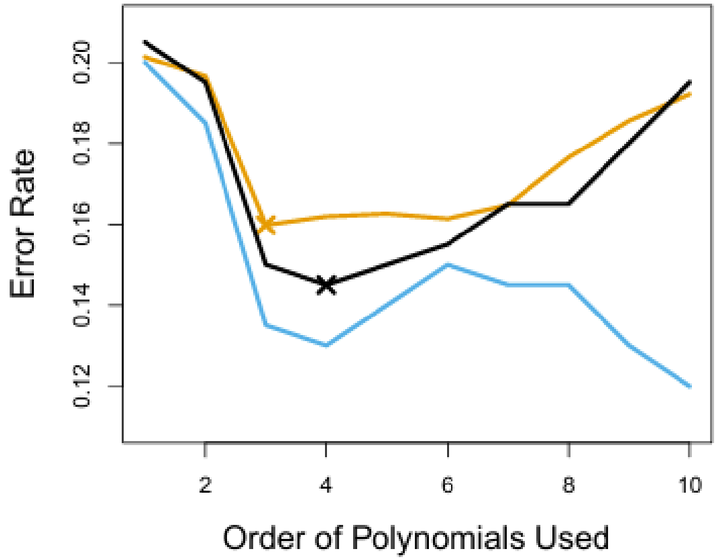
- Logistic Regression의 degree 조정
- 최대한 베이지안 결정경계에 가깝게 추정해야 한다.
- degree를 너무 높이면, 훈련데이터에 과대적합되는 문제가 있음.
- 즉, cross validation을 통해 test error를 추정하는 작업이 중요하다.

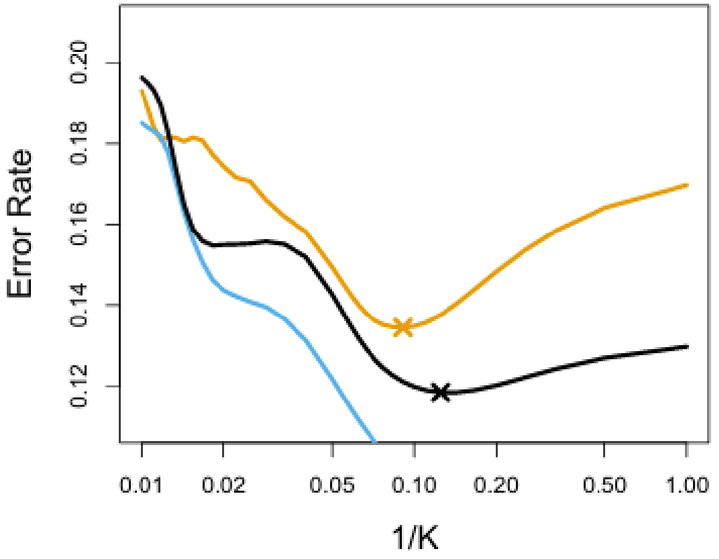


AI명예학회
SKHU

5.1.5 on Classification

- train error는 degree가 올라갈수록 작아지는 경향성
- test error는 일정 수준에서 다시 올라감
- Cross Validation은 이와 비슷한 U자 형태를 띔
- degree가 높을수록, cross validation을 통해 모델을 추정하면 오차가 커질 수밖에 없다.





AI명예학회
SKHU

5.2 Bootstrap

- 통계적 학습 방법에서 불확실성을 측정하는데 유용
- 통계 소프트웨어에서 제공하지 않는 방법들도 포함
- 문제 가정: 최적의 투자 배분을 결정하고자 하는 문제

- X와 Y의 수익률을 내는 두 금융 자산
- X와 Y가 무작위인 수량
- 수익률에는 변동성이 있기 때문에 분산을 최소화하는 결정을 내고자 한다.
- 즉, 가장 안전하고 안정적인 선택을 찾는 문제

AI명예학회
SKHU

5.2 Bootstrap

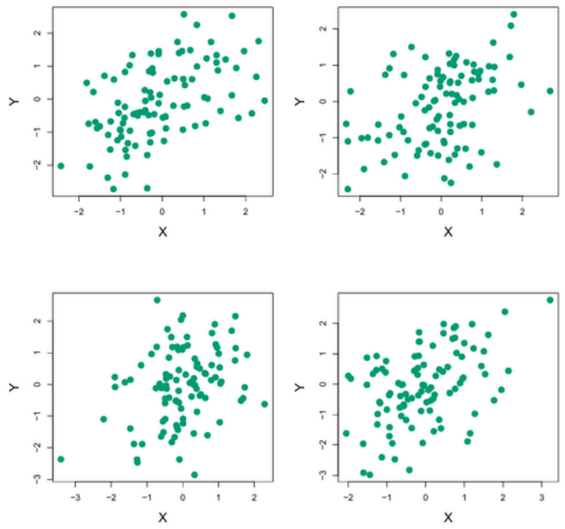
- X와 Y의 수익률을 내는 두 금융 자산
- X와 Y가 무작위인 수량
- 수익률에는 변동성이 있기 때문에 분산을 최소화하는 결정을 내고자 한다.
- 즉, 가장 안전하고 안정적인 선택을 찾는 문제
- X를 a만큼 선택, Y를 1-a만큼 선택
- Var(aX+(1-a)Y)

$$\alpha = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}}, \qquad \hat{\alpha} = \frac{\hat{\sigma}_Y^2 - \hat{\sigma}_{XY}}{\hat{\sigma}_X^2 + \hat{\sigma}_Y^2 - 2\hat{\sigma}_{XY}}.$$

AI명예학회
SKHU

5.2 Bootstrap

- 100쌍의 수익률을 1000번 시뮬레이션
- 1000개의 a가 나올 것이고, 계산한 결과를 통해 SE를 추정한다.
- 모집단으로부터 추출한 샘플이 평균적으로 n과 0.08정도 차이가 날 수 있다는 것을 보여준다.



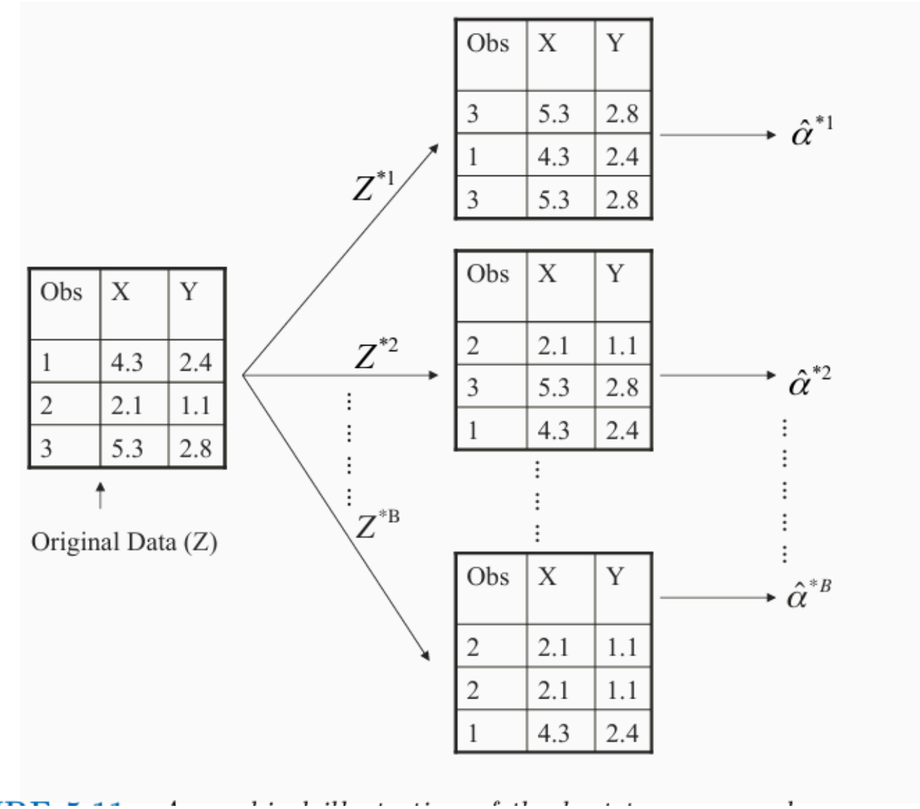
$$\sqrt{\frac{1}{1000-1}\sum_{r=1}^{1000}(\hat{\alpha}_r - \bar{\alpha})^2} = 0.083.$$

AI명예학회
SKHU

5.2 Bootstrap

- 하지만 실제 데이터에서는 모집단에서 계속 새로운 샘플을 뽑아낼 수 없다.
- 부트스트랩은 컴퓨터를 사용해 새로운 표본집단을 애뮬레이션해서 추가 표본을 생성하지 않고도 이 작업을 수행할 수 있도록 해준다.
- 관측치를 반복적으로 샘플링해서 별개의 데이터 집합을 얻는다.

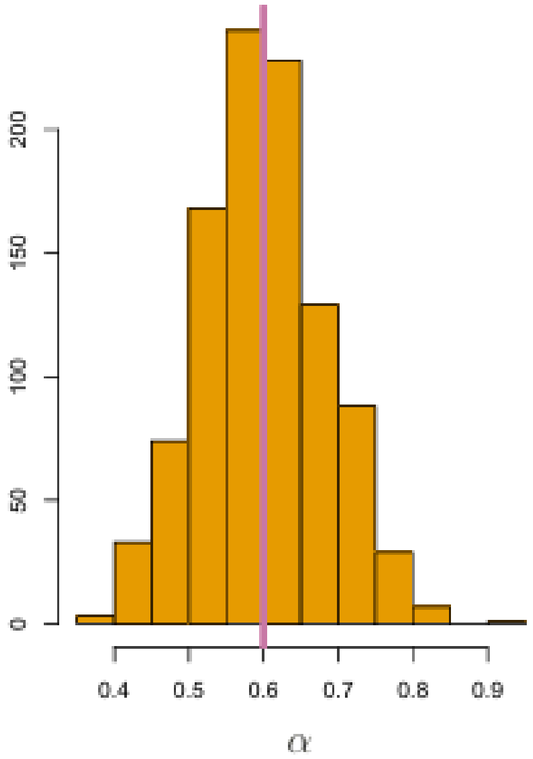
5.2 Bootstrap

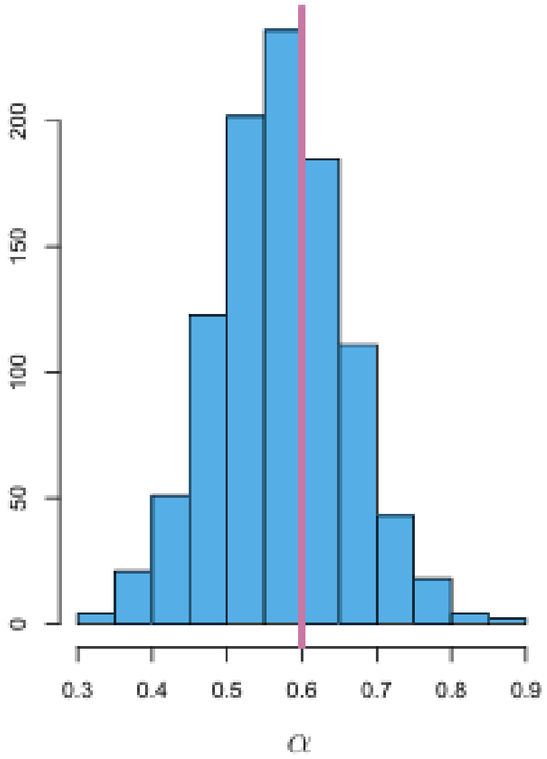


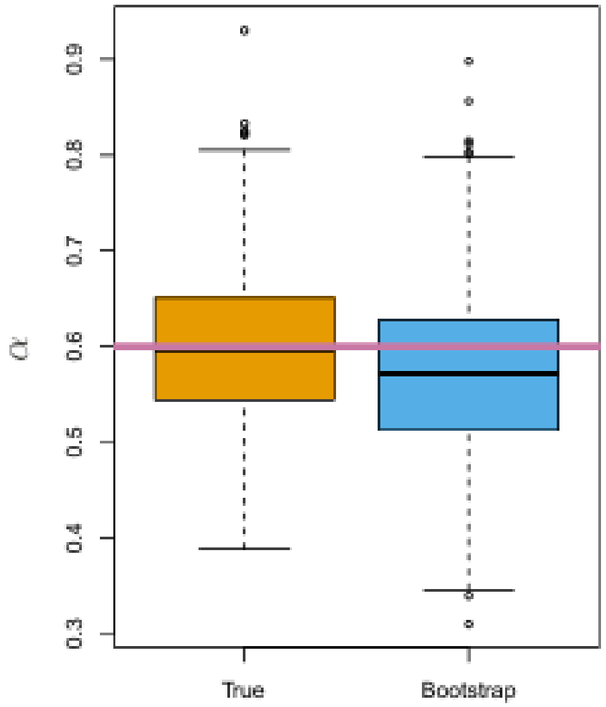
$$\mathrm{SE}_B(\hat{\alpha}) = \sqrt{\frac{1}{B-1}\sum_{r=1}^{B}\left(\hat{\alpha}^{*r} - \frac{1}{B}\sum_{r'=1}^{B}\hat{\alpha}^{*r'}\right)^2}.$$

AI명예학회
SKHU

5.2 Bootstrap







AI명예학회
SKHU