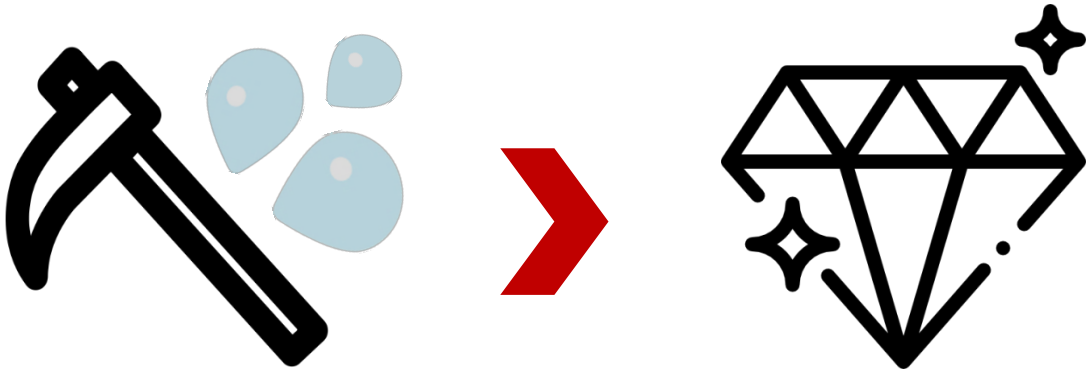
클린업 1주차

4팀 데이터마이닝 

박정현
전규리
김지민
노정화
염예빈

어원

Data Mining

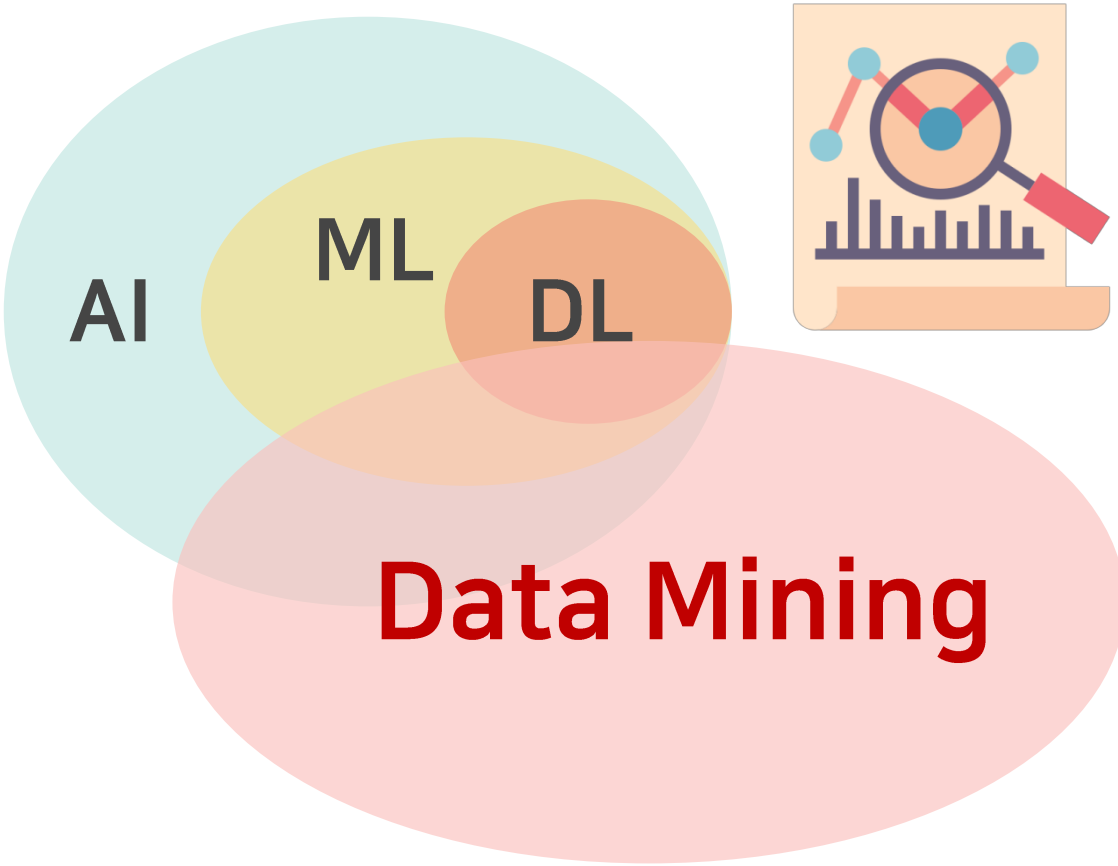


데이터에서 **유용한 정보와**

패턴을 추출하는 과정

1 정의

머신러닝, 딥러닝, statistical learning과의 차이

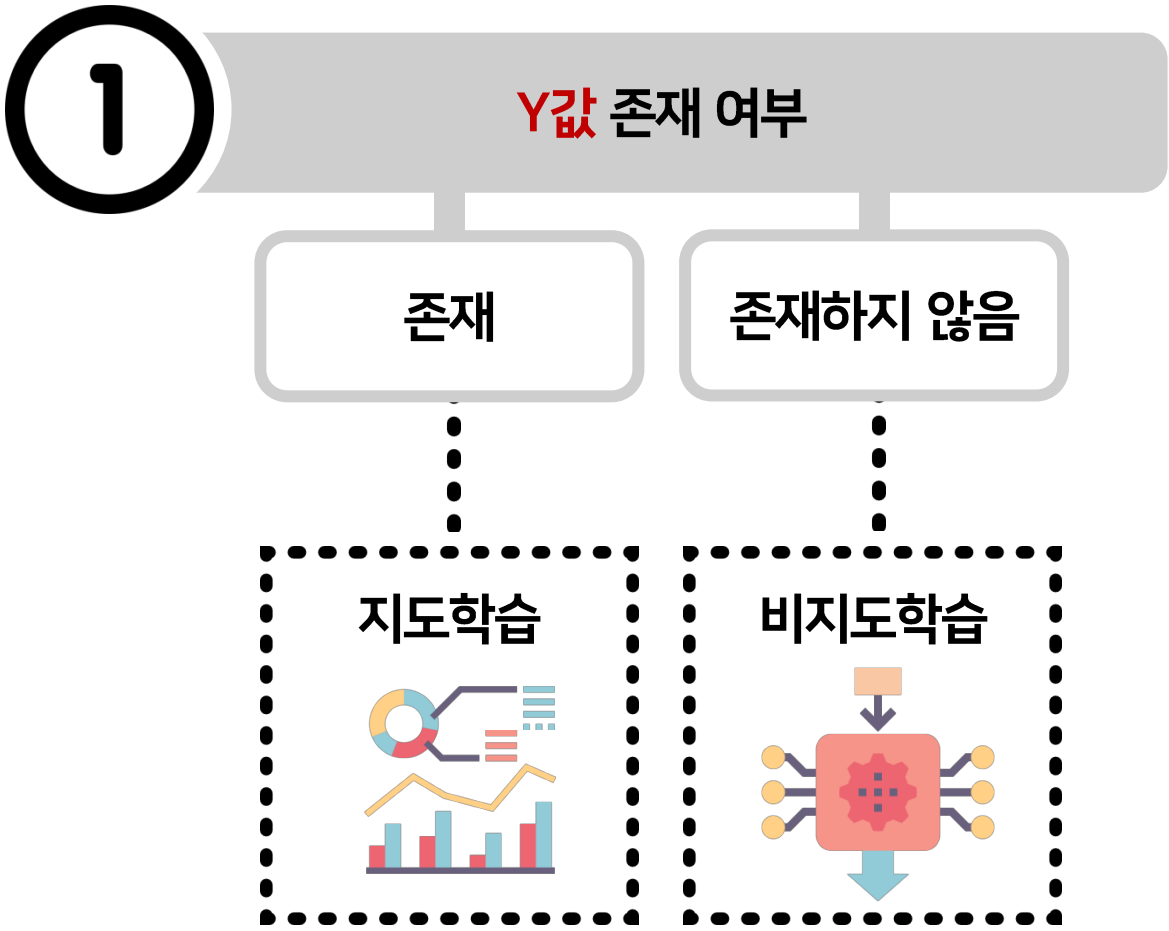


1 과정

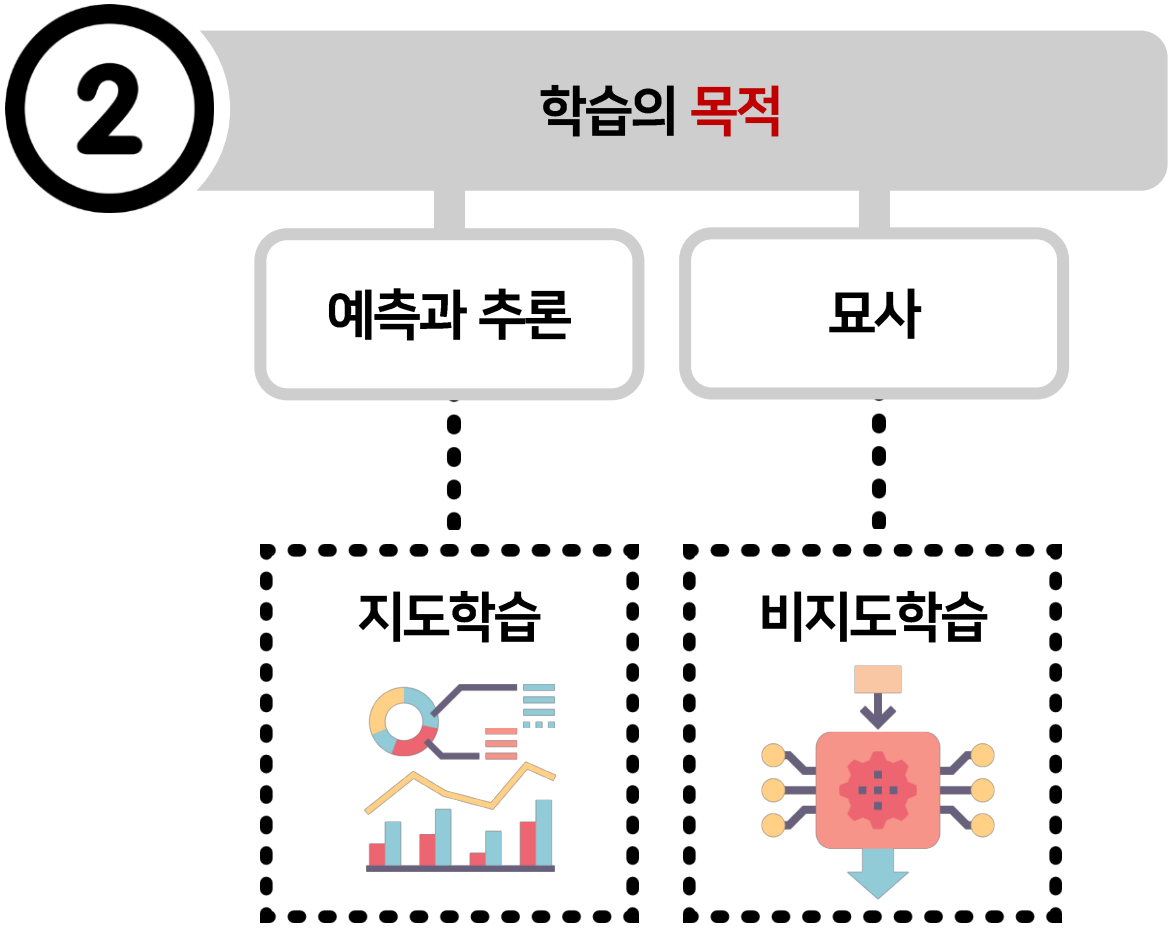
CRISP-DM으로 보는 데이터분석 과정

2 정의

지도학습과 비지도학습의 비교



지도학습과 비지도학습의 비교



2 편향-분산 트레이드 오프

Reducible error, irreducible error

MSE(Mean Squared Error)

$$= \mathrm{E}[(Y - \hat{Y})^2]$$

$$= bias(\hat{f})^2 + Var(\hat{f}) + \sigma^2$$

- σ^2 : **irreducible error**

샘플링 방법을 다르게 하거나 측정 방법을 달리하여 줄일 수 있지만 일반적으로 데이터분석 시 학습할 데이터셋이 주어져 있는 경우

줄이지 못하는 오차(irreducible error)이다.

2 편향-분산 트레이드 오프

Reducible error, irreducible error

MSE

$$= \mathrm{E}[(Y - \hat{Y})^2]$$

$$= bias(\hat{f})^2 + Var(\hat{f}) + \sigma^2$$

- $bias(\hat{f})$ **(편향)**$^2 + Var(\hat{f})$**(분산)**:**reducible error**

더 좋은 모델을 선택함으로써 줄일 수 있기 때문에

줄일 수 있는 오차 (reducible error) 이다.

2 편향-분산 트레이드 오프

Reducible error, irreducible error

MSE

$= E\left[(Y - \hat{Y})^2\right]$

$= \{bias(\hat{f})\}^2 + Var(\hat{f}) + \sigma^2$

따라서 우리는 적절한 모델을 선택함으로써

reducible error를 최소화하고자 한다.

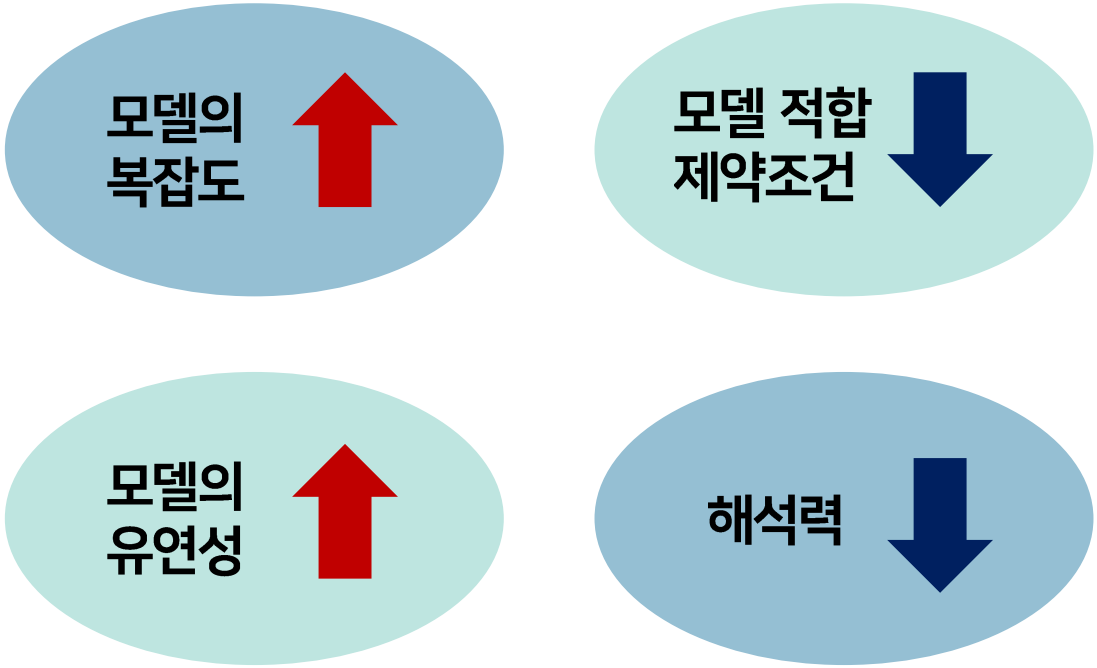
- Bias(편향), $Var(\hat{f})$(분산) : (reducible error)

더 좋은 모델을 선택함으로써 줄일 수 있기 때문에

줄일 수 있는 오차 (reducible error) 이다.

2 편향-분산 트레이드 오프 II

모델 복잡도(complexity)와 해석력(interpretability)

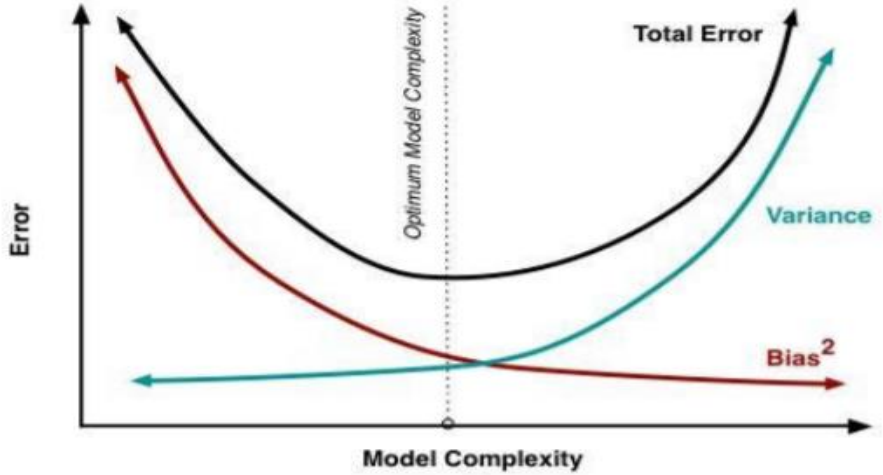


2 편향-분산 트레이드 오프 II

편향과 분산의 균형

모델의 복잡도를 올릴수록 모델의 편향은 줄어들고, 분산은 커진다.

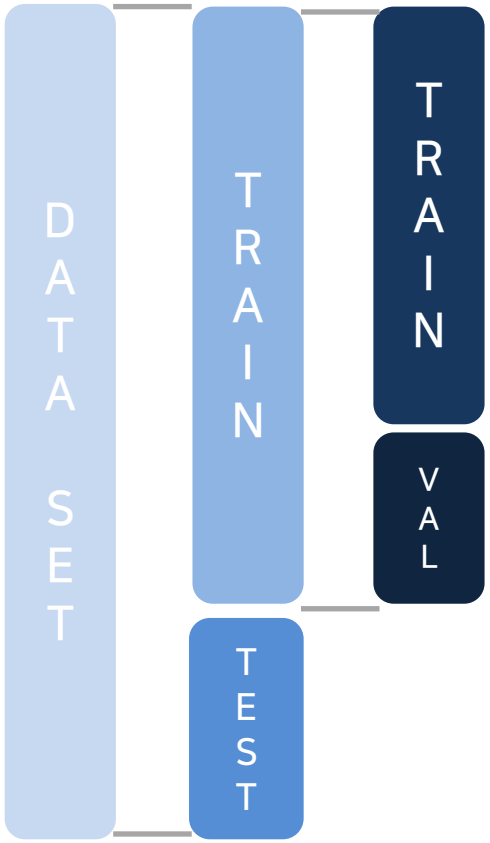
∴ 모델의 성능에 대한 평가 지표인 MSE를 구성하는 편향과 분산은 **트레이드오프(Trade-off)관계** 이다.



모델의 복잡도는 Test MSE가 가장 **최소**가 되는 지점에서 결정!

2 모델 복잡도 결정

Hold-out method



Validation set:

- 라벨 값(y) 있음
- Test error를 간접적으로 측정해 모델의 예측력을 높임

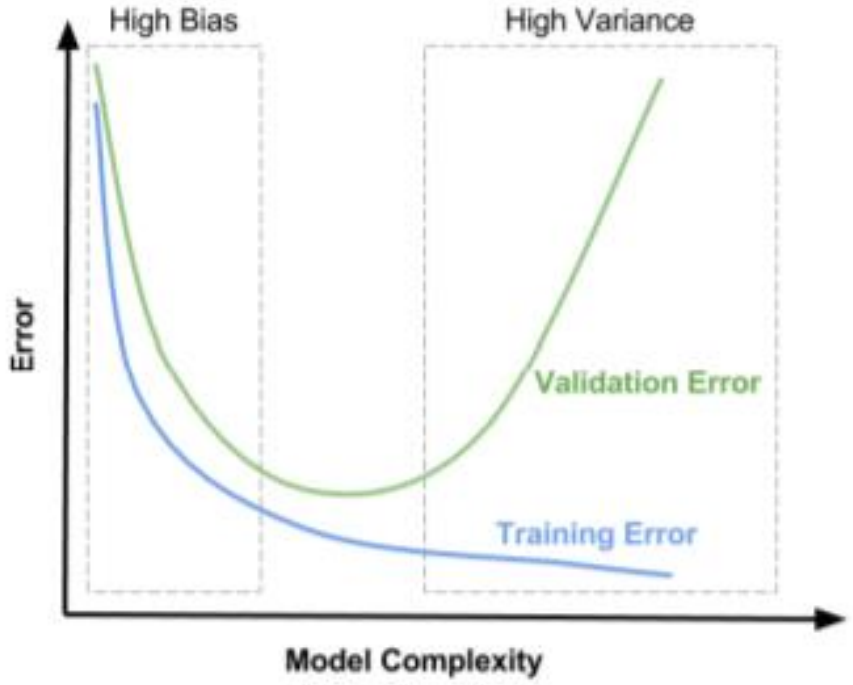
Test set :

- 라벨 값(y) 없음

2 모델 복잡도 결정

Hold-out method

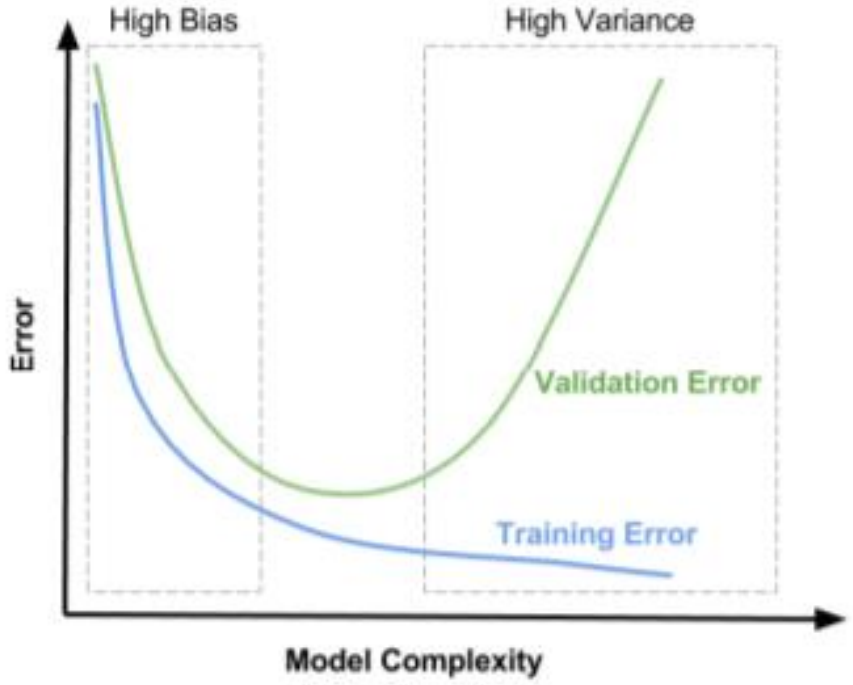
training error는 모델의 복잡도가 증가할수록 감소하는 반면 test set 혹은 validation set에 대한 error는 어느 지점을 기준으로 다시 상승하게 된다.



2 모델 복잡도 결정

Hold-out method

그 지점을 모델 복잡도로 선정하고 **최적의 모델**로 전체 train set을 **다시 학습**시킨다

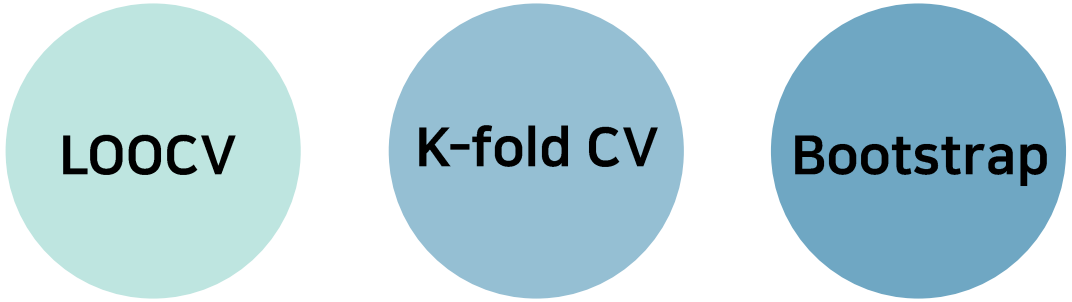


3 Sample Reuse Methods

재표본 방법이란?

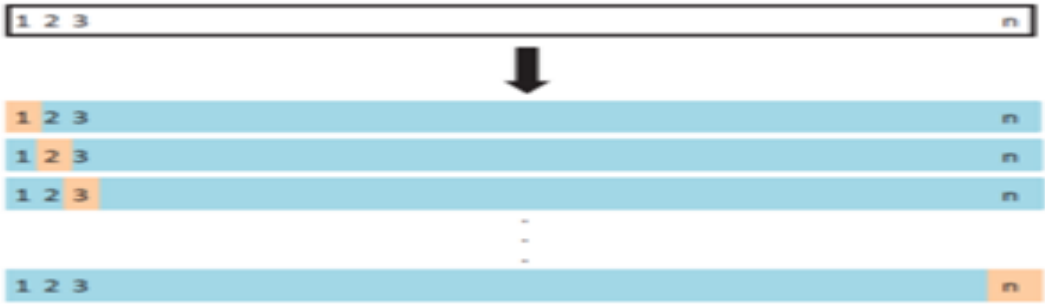
통계학에서 무작위 변동성을 알아보기 위해
관찰한 데이터 값에서 **표본을 반복적으로 추출**하는 것

오늘 다뤄볼 귀여운 재표본 칭구들?>-<



3 Cross Validation

LOOCV(Leave-One-Out CV)



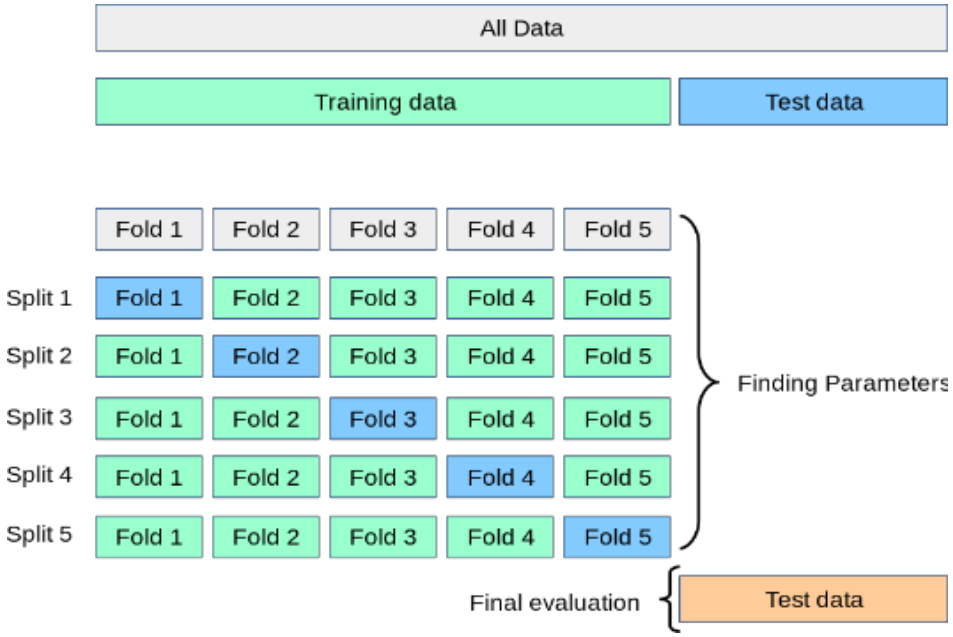
관측치가 n개일 때, 각각의 관측치를 **하나의 fold** 취급

하나의 관측치만을 validation으로 두고
n-1개의 fold에 대해 학습

각각의 validation에 대해 MSE를 계산하고, 평균을 낸다.

3 Cross Validation

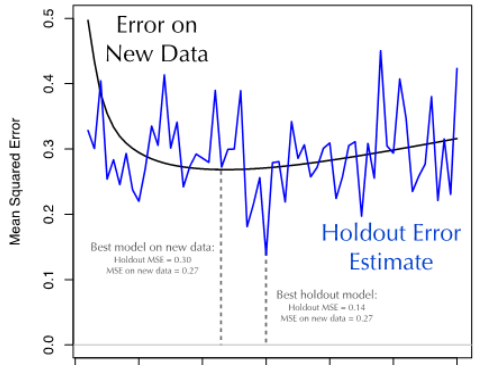
k-fold CV의 원리



- 데이터를 K개의 fold로 나누고 남은 k-1개 fold에 대해 학습
- 이 과정을 k번 반복하여 평균 내어 CV Error로 삼는다

3 Cross Validation

Holdout Error VS k-fold CV Error



Holdout Error

- Validation set만을 이용
- Test Error를 overestimate하는 경향

K-fold CV Error

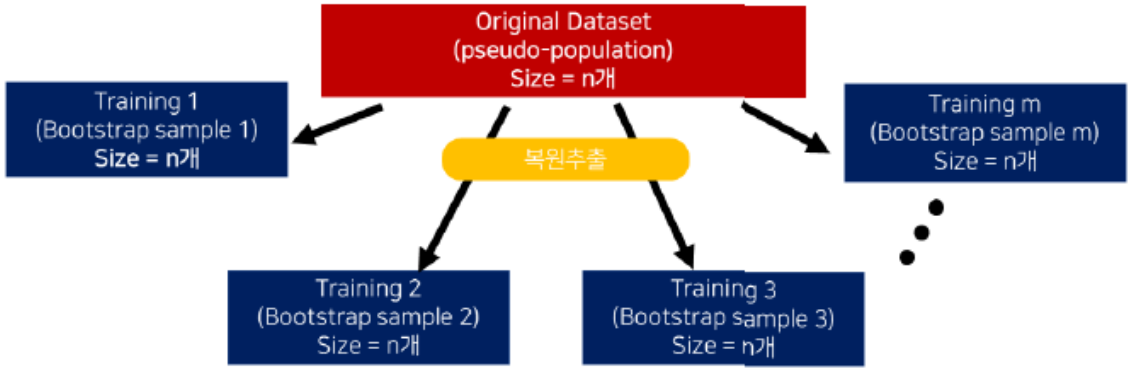
- Test Error를 안정적이고 근접하게 따라감

3 Bootstrap

부트스트랩이란?

- 파라미터 튜닝보다는 **표준편차 추정**이나 가설 검정 기법의 유효성 검증을 위한 **시뮬레이션**에서 사용

- 트리기반 모델들의 기반이 되는 재표본 방법



부트스트랩의 한계

- 기존 데이터셋을 validation set으로 사용하여 만든 샘플을 사용하기 때문에 기존 데이터와 **상당수의 관측치가 겹침**
- 데이터크기가 **작을** 경우, Test Error를 잘 **대변하는 추정치라고 보기 어려움**

모델 성능 측정 시 사용빈도