PRESENTATION

4팀

김동영
강수경
김재희
유경민
최윤혜

INDEX

1 데이터 마이닝

Data Mining

데이터

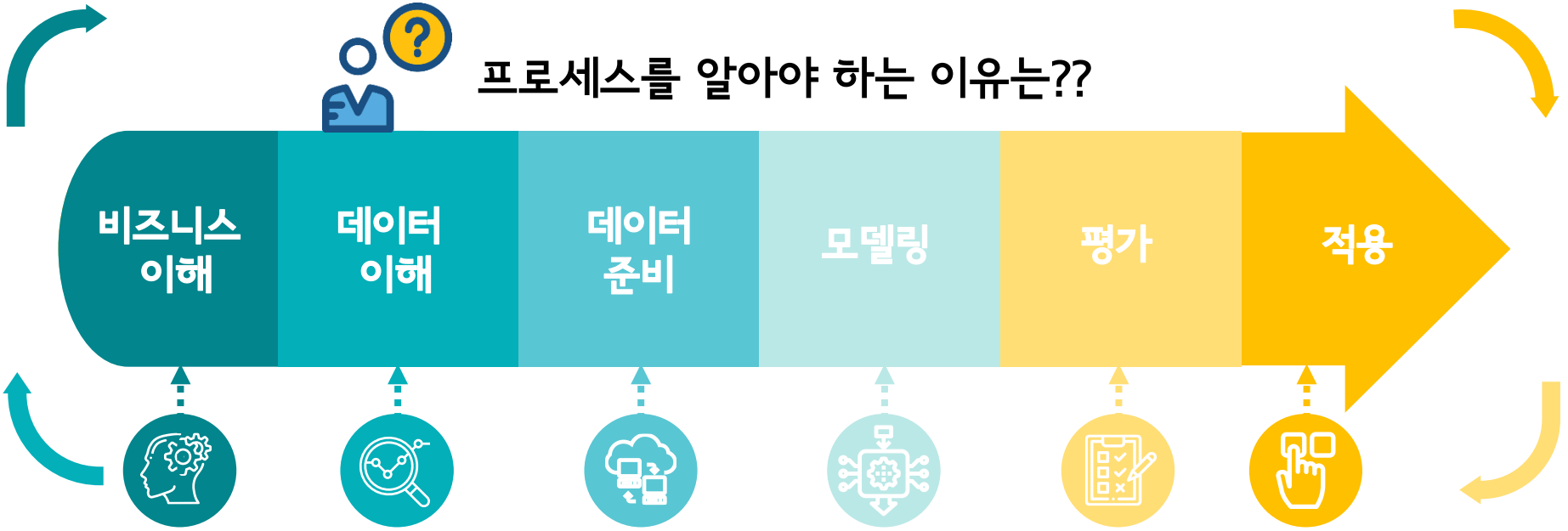
Data mining

유의미한 정보

대용량 데이터를 조사함으로써 의미 있는 정보를 발굴해내는 과정
다양한 학문과의 결합이 요구됨

CRISP-DM

프로세스를 알아야 하는 이유는??



데이터분석 **전체의 흐름**에 대한 이해가 있어야

길을 잃지 않고 좋은 분석을 할 수 있다

모델링

모델링의 종류

머신 러닝

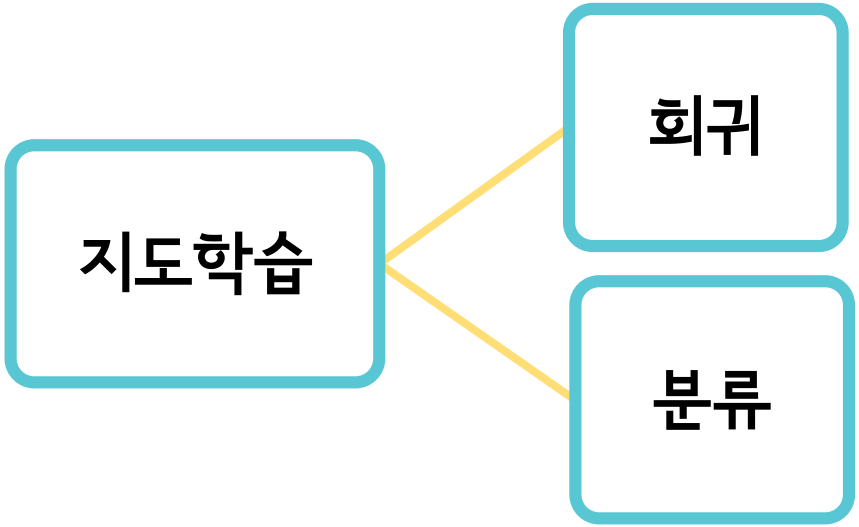
- 지도학습
  - 회귀
  - 분류
- 비지도학습
- 강화학습

모델링이란?

지도학습

모델의 종류에 따라 처리할 수 있는 문제가 다르다!

➜ 학습 데이터의 반응변수(Y값)에 따라 결정됨

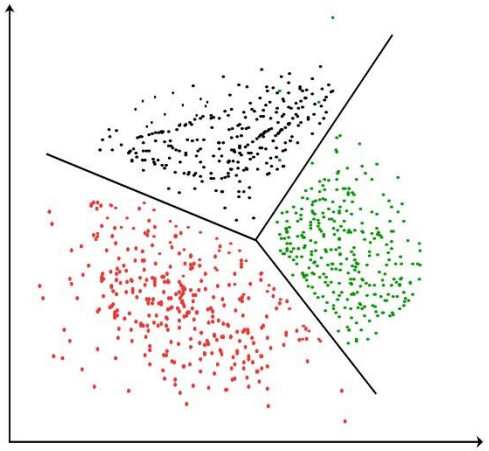


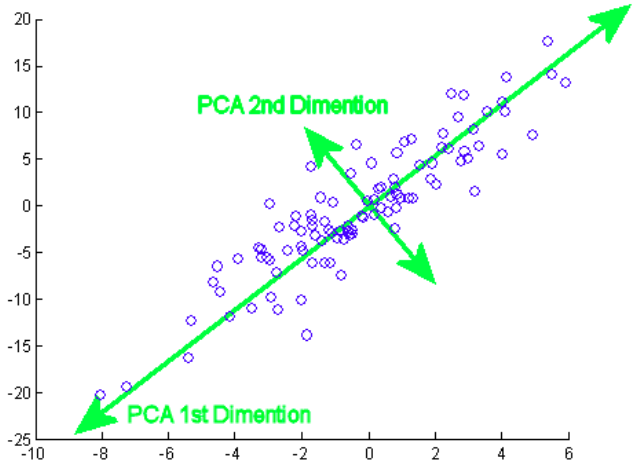
반응변수가 연속형
ex) 주가 예측, 확률 예측

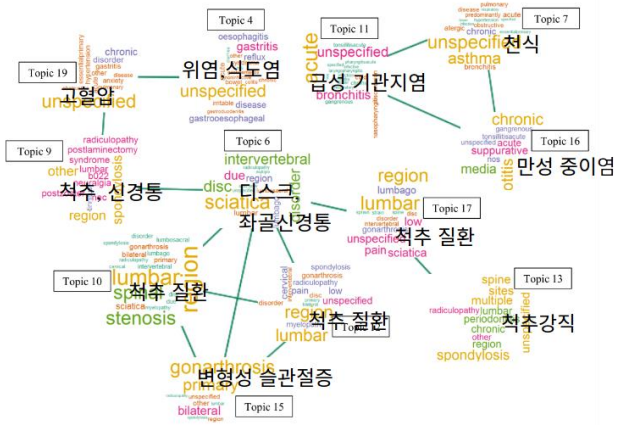
반응변수가 범주형
ex) 고양이/개 판단 문제

모델링이란?

비지도학습







✓ 클러스터링

✓ PCA(주성분 분석)

✓ 토픽 모델링

클러스터링, PCA, 토픽 모델링 등등

분산과 편향

Bias-Variance Trade-off

예측을 목적으로 하는 지도 학습에서는 오차가 필연적

$$E\left[\left(y-\hat{f}\right)^2\right] = \sigma^2 + Bias[\hat{f}]^2 + Var[\hat{f}]$$

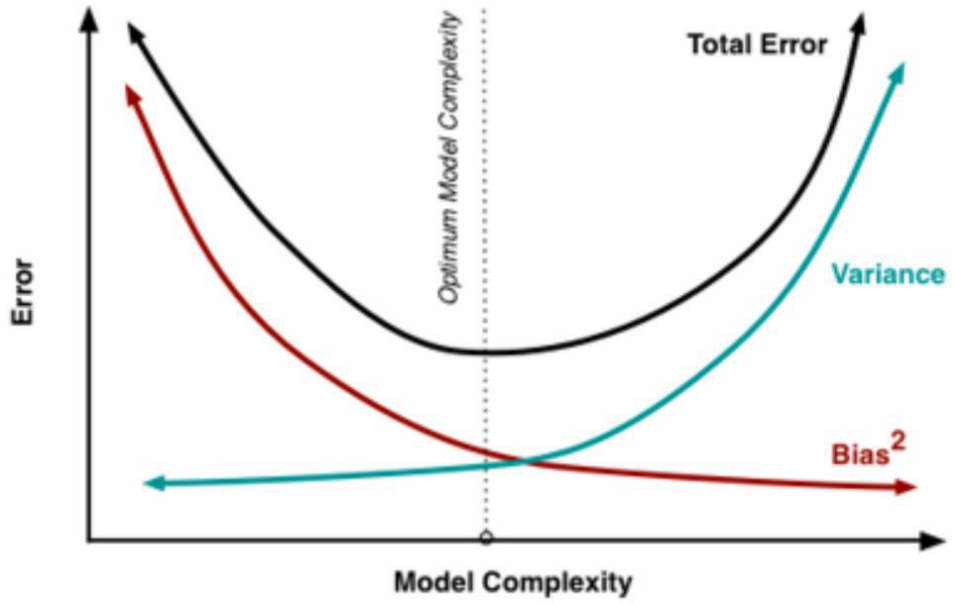
오차 = 데이터 자체의 오차 + 편향 + 분산

Irreducible error

Reducible error

분산과 편향

Bias-Variance Trade-off

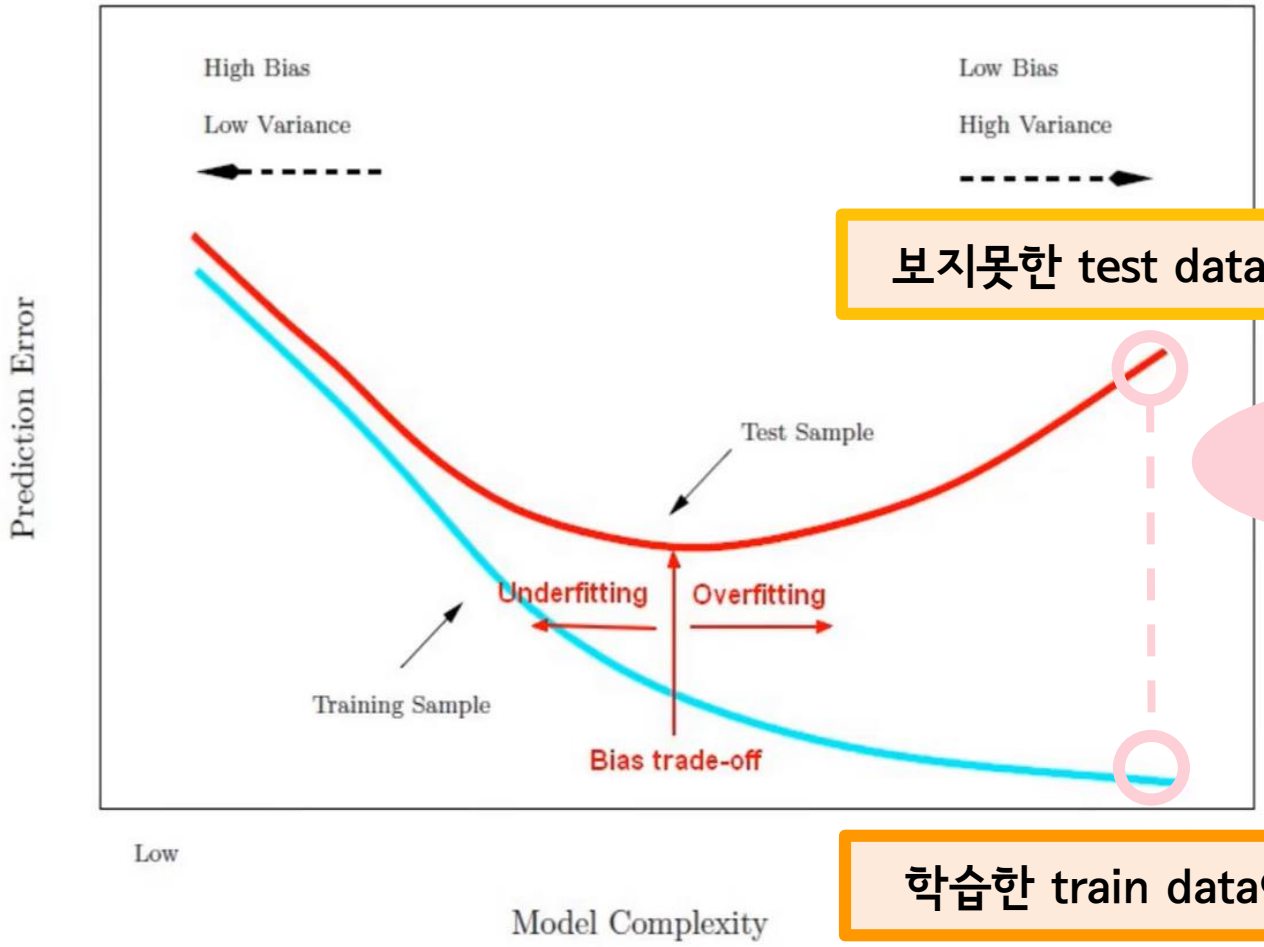


편향과 분산은 Trade-off 관계

데이터의 복잡도에 따라 모델을 선택하여 오차 줄일 수 있음

분산과 편향

왜 Test Data를 만들어?



보지못한 test data에게는 나쁜 성능!

!Overfitting!

학습한 train data에게는 결과 좋아!

분산과 편향

여전히 문제는 있다!

Dataset이 작다면!
train/test data에 어떤 관측치가 있는가에 따른 **변동이 크다**!

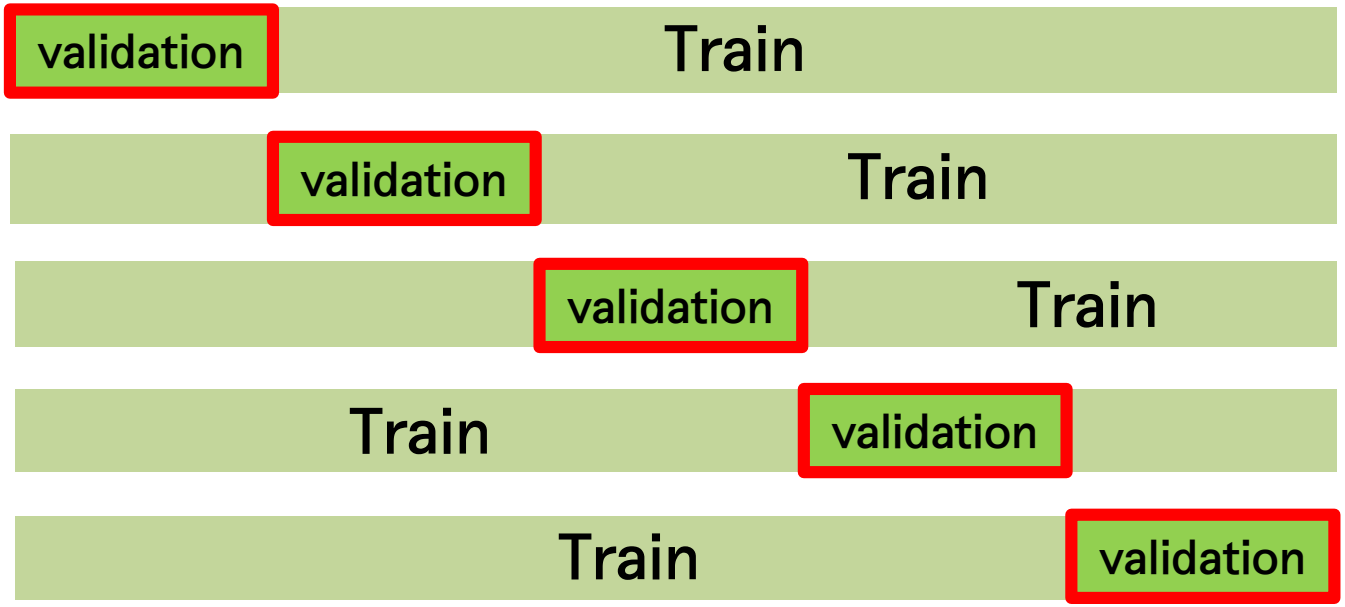
Test 성능 높이려고 반복한다면!
Test 데이터에 또 다시 **과적합**!

Validation을 만들자!

분산과 편향

교차검증은 무엇인가요?

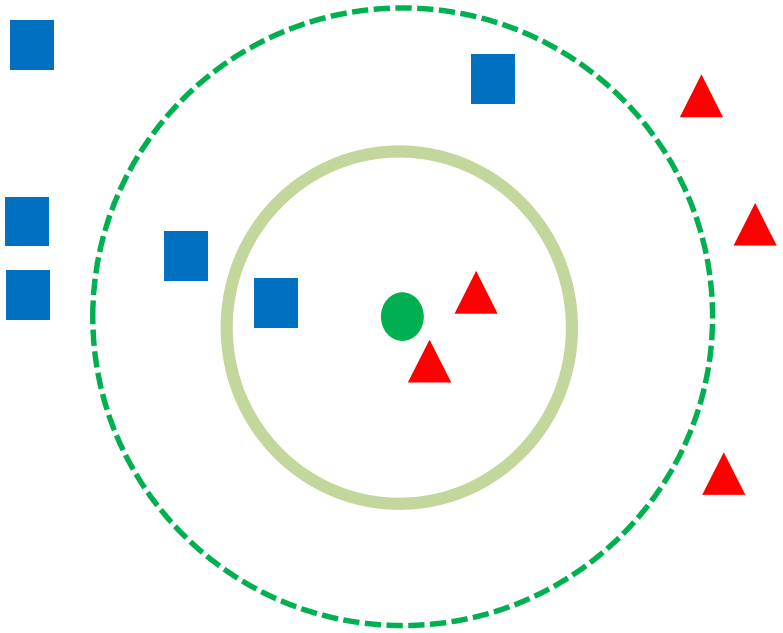
"Train / Validation 나누는 것을 반복하기!"



k-NN이란?

새롭게 들어온 점에 대해 가장 가까운 점 k개를 찾아

다수결의 원칙으로 점 분류!



k = 1 → 빨강

k = 2 → 빨강

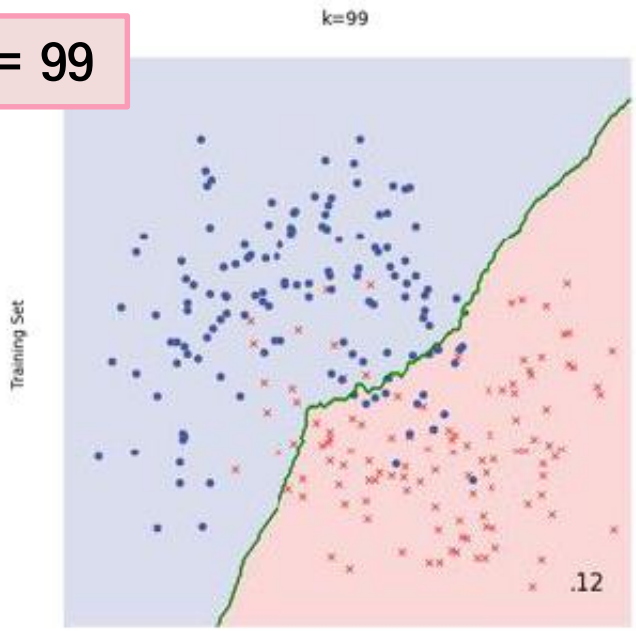
k = 3 → 빨강

k= 5 → 파랑

KNN

편향, 분산을 k-NN에 적용!

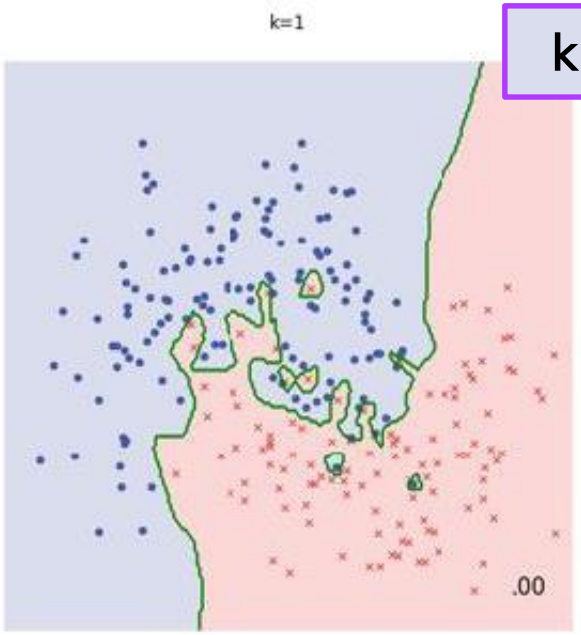
k = 99



편향 ↑ 분산 ↓

"Underfitting"

k = 1

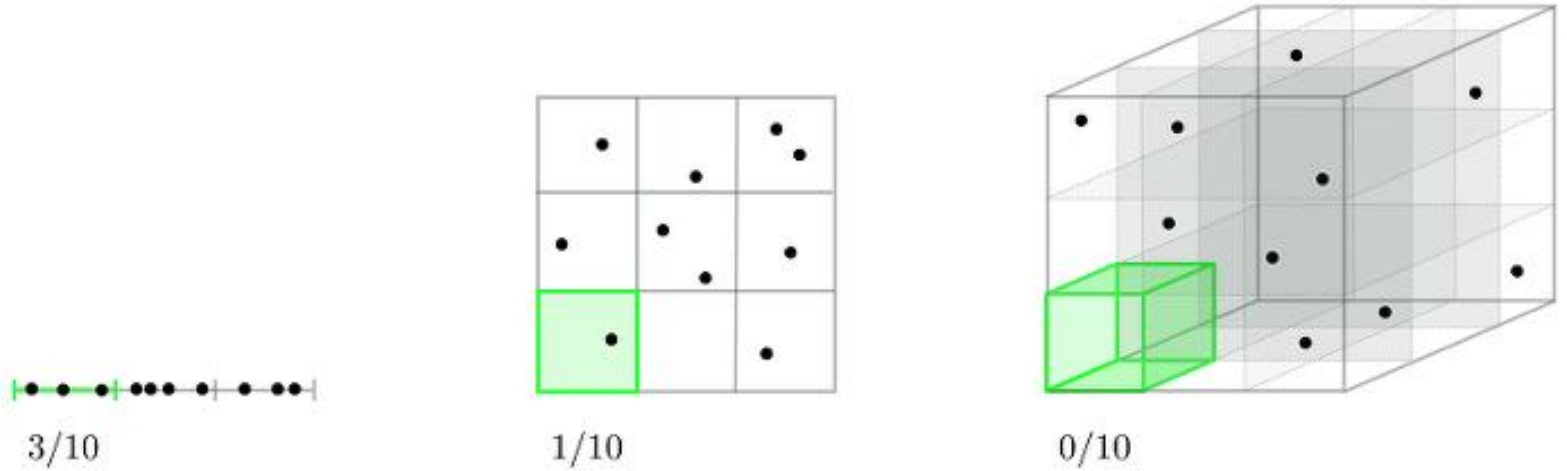


편향 ↓ 분산 ↑

"Overfitting"

KNN

차원의 저주



차원↑ --> 데이터 밀도↓ --> 경향학습 어려워져!

"성능 저하"

이웃들이 멀어지며
주변에 제대로 존재x