데이터마이닝팀

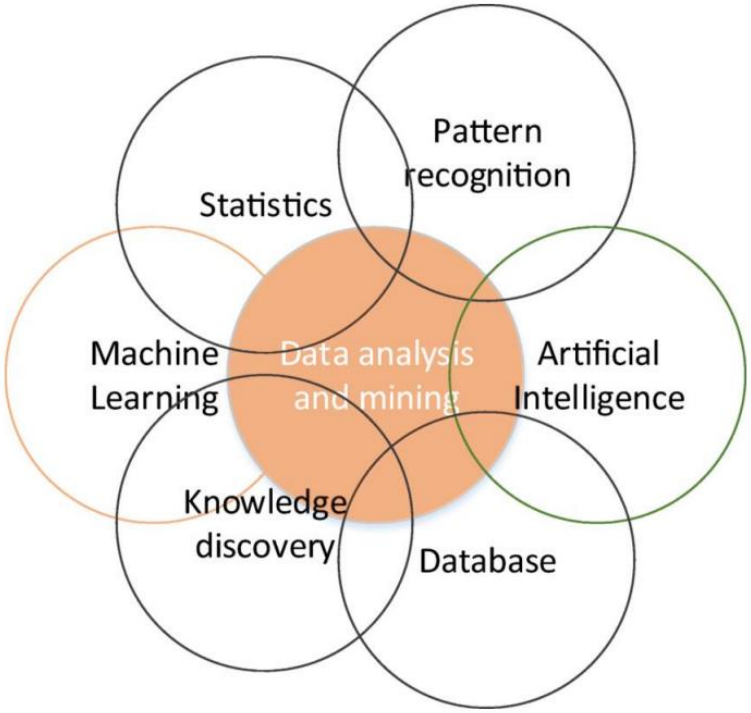
4팀

이진모
이은서
임주은
박지민
장이준

CONTENTS

데이터 마이닝의 범학문적 특성

Definition of Data Mining

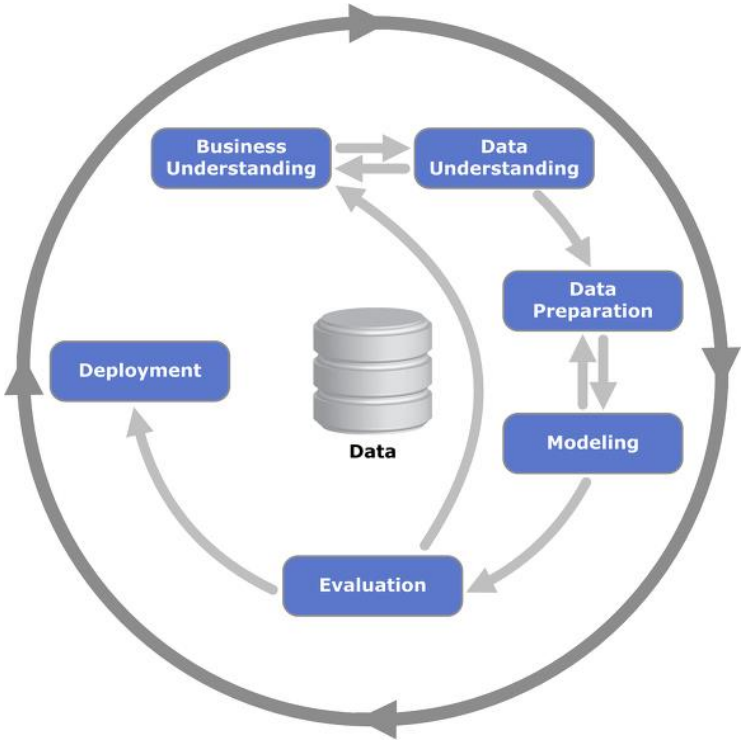


데이터 마이닝은 머신 러닝, 통계학과 같은 여러 학문들의 교집합에 위치

방법론: CRISP_DM

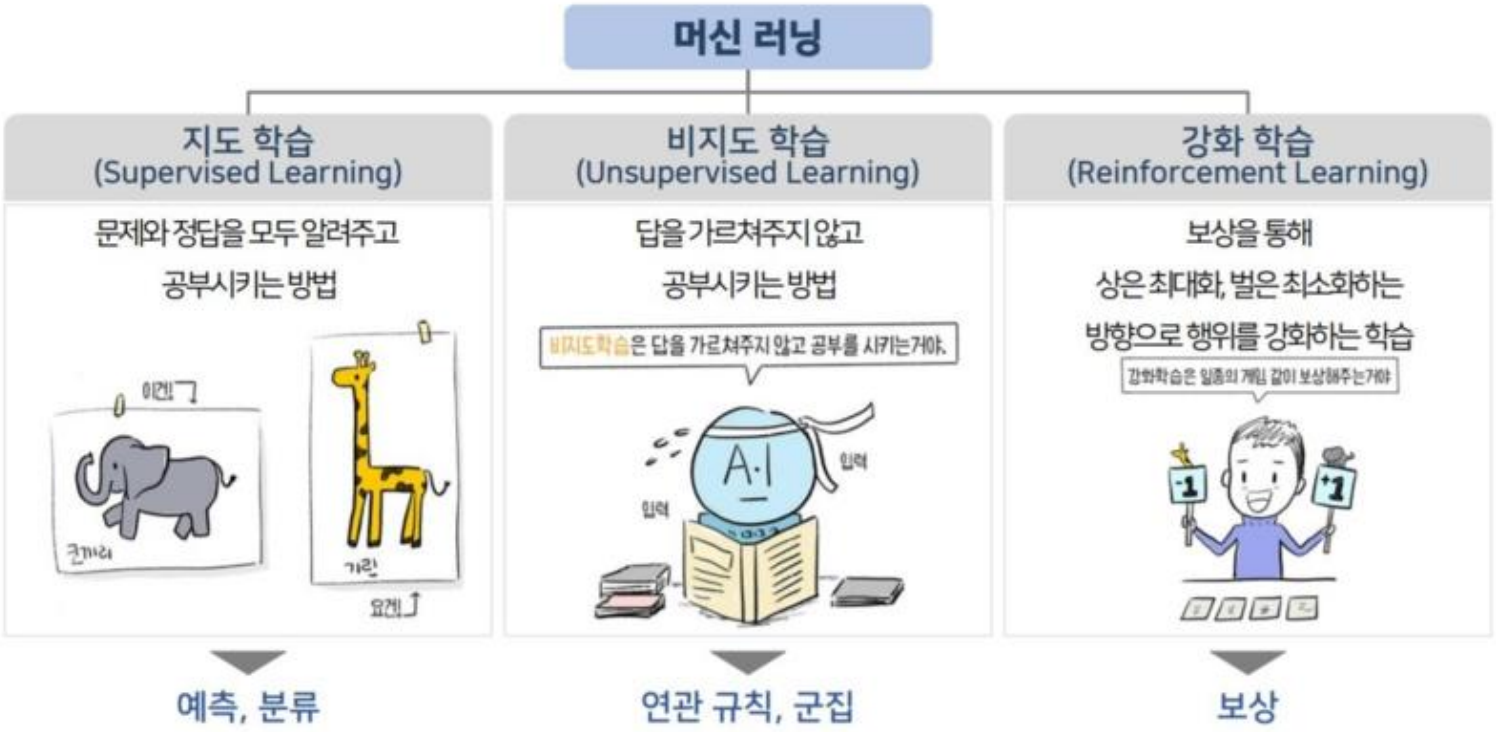
Cross-Industry Standard Process for Data Mining

1) Business Understanding
2) Data Understanding
3) Data Preparation
4) Analysis & Modeling
5) Evaluation
6) Deployment



모델링(머신 러닝) 종류

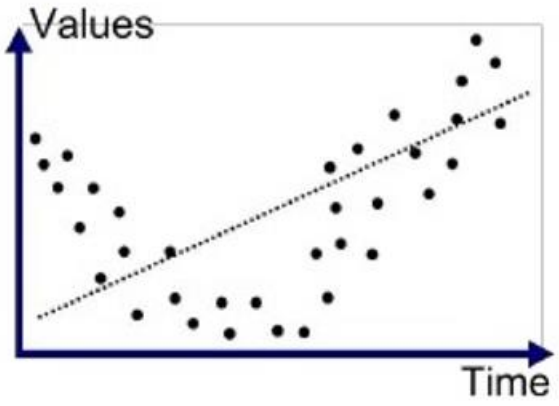
type of modeling(machine learning)



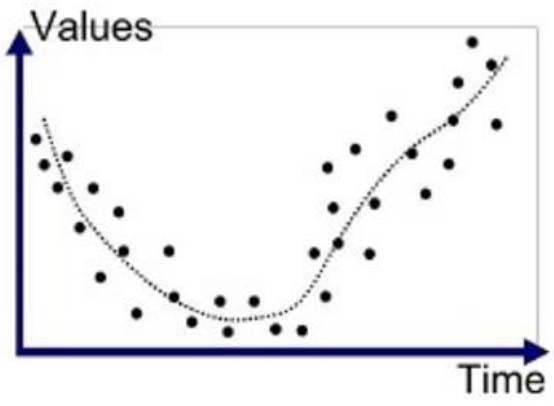
지도학습

지도학습 방법 - 수학적 모델링

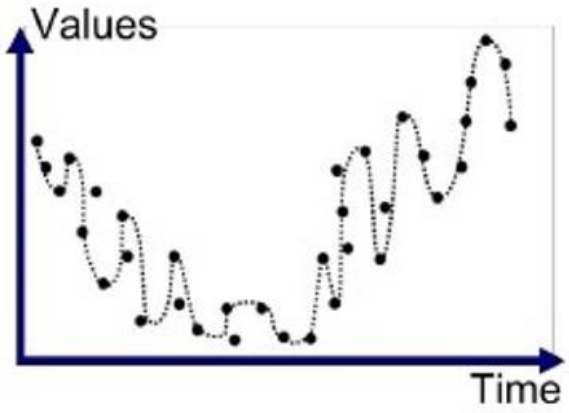
"하이퍼 파라미터"에 따라 모델 변화!



Underfitted



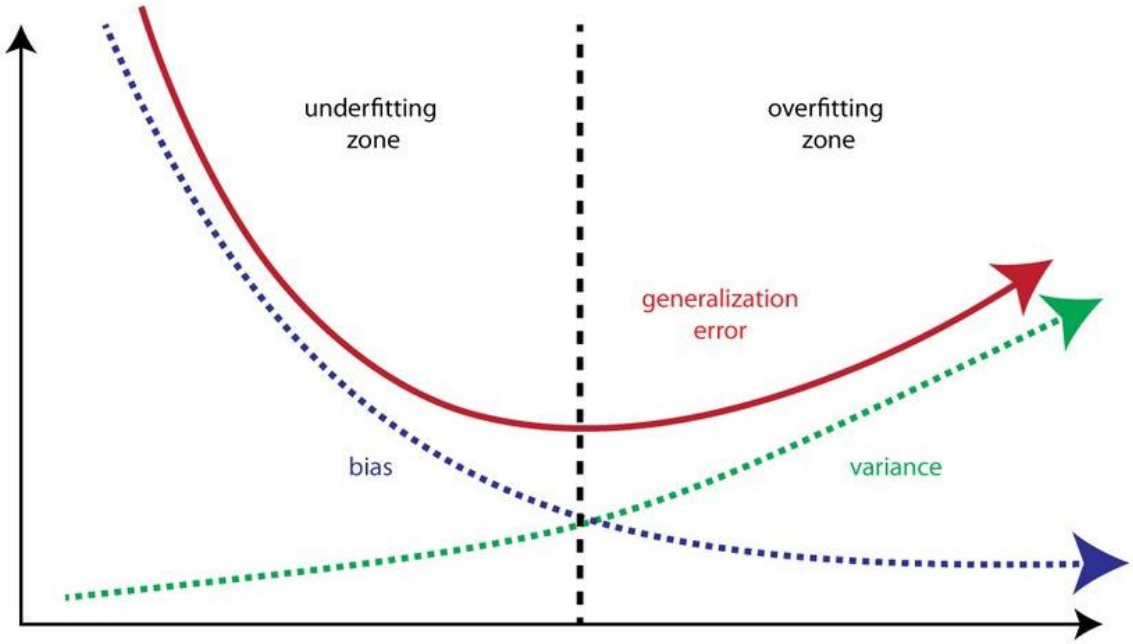
Good Fit/Robust



Overfitted

지도학습

Variance - Bias Tradeoff

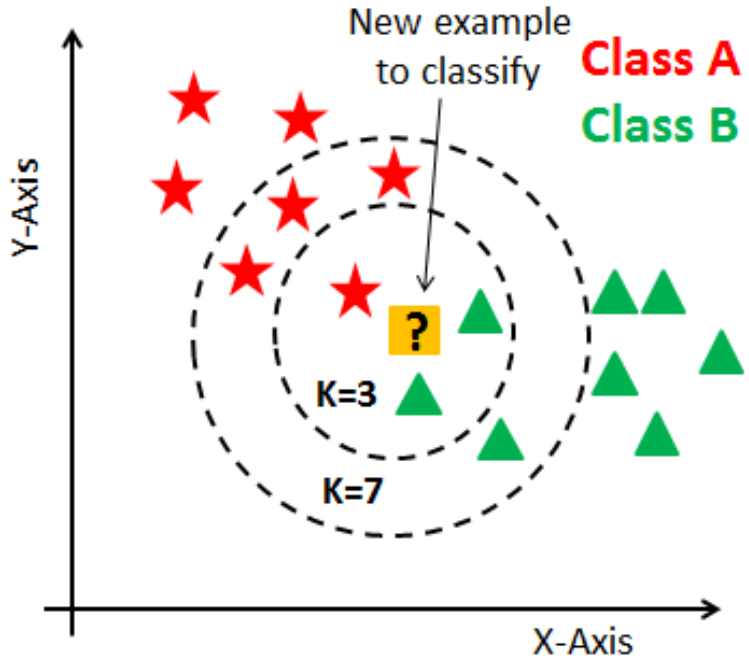


MSE를 줄이기 위해서는 모델의 편향과 분산을 줄여야 함

그러나! 모델의 편향과 분산을 **동시에** 원하는 만큼 줄이는 것은 **불가능**

지도학습

KNN Model (K-Nearest Neighbors)



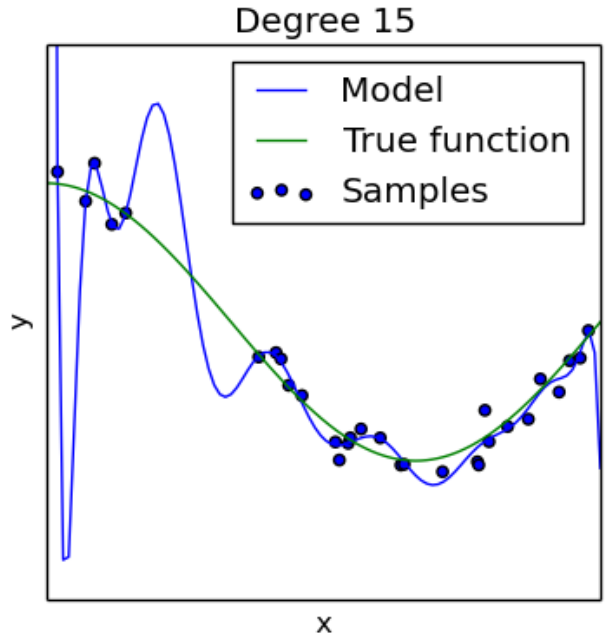
분류 과정에서의 KNN model

파라미터 K값에 따라서 다른 결과를 도출

(K=3 : 초록 세모, K=7 : 빨간 별)

Overfitting?

Why Avoid Overfitting



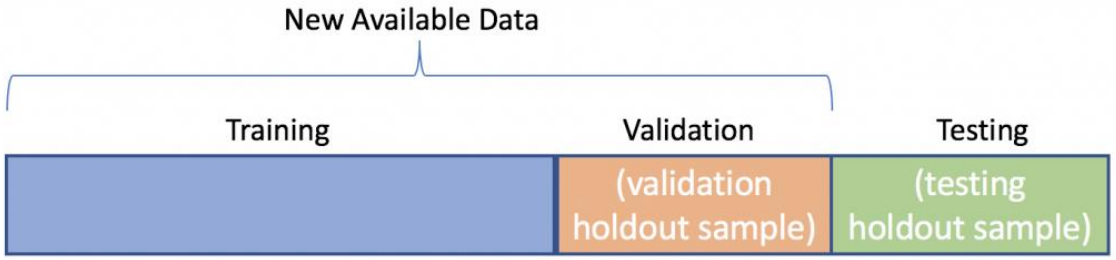
추정된 모델이 데이터의

관측 값을 하나하나 따르는 형태

모델의 분산이 커진다!

Train-Test Split

Hold-out Method

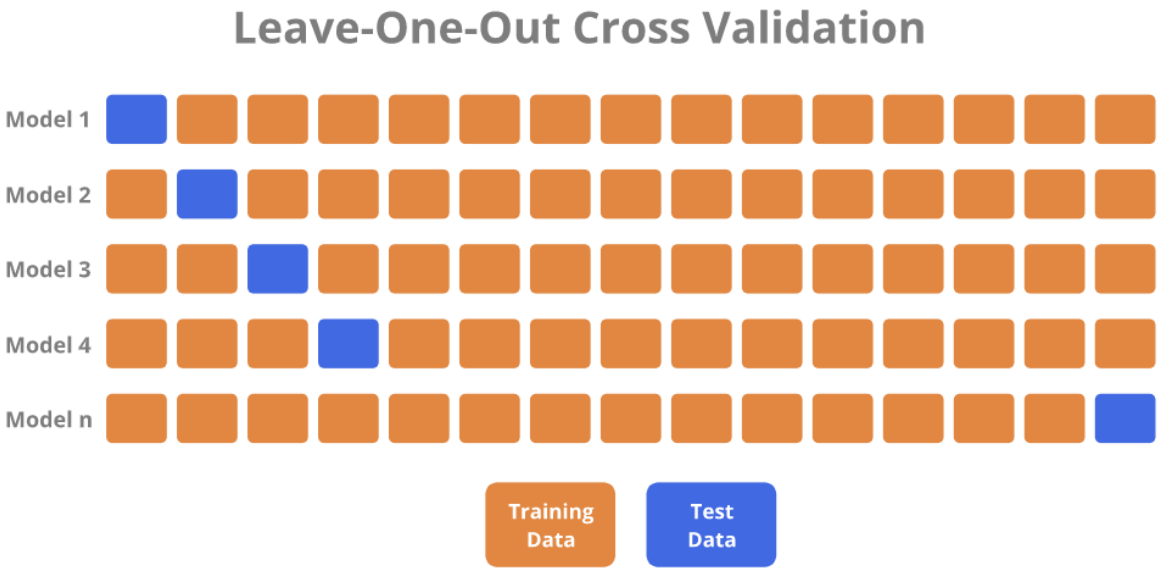


학습 데이터의 일부를 검증 데이터로 삼아

테스트 데이터가 적합 되었을 때 모델 성능을 평가하자!

Cross Validation (CV)

Leave-One-Out CV (LOOCV)



전체 N개의 데이터 중 한 개만 검증 데이터로,

나머지 N-1개는 학습데이터로 사용해 총 N번의 검증을 반복

Cross Validation (CV)

K-Fold CV

Estimation 1	Test	Train	Train	Train	Train
Estimation 2	Train	Test	Train	Train	Train
Estimation 3	Train	Train	Test	Train	Train
Estimation 4	Train	Train	Train	Test	Train
Estimation 5	Train	Train	Train	Train	Test

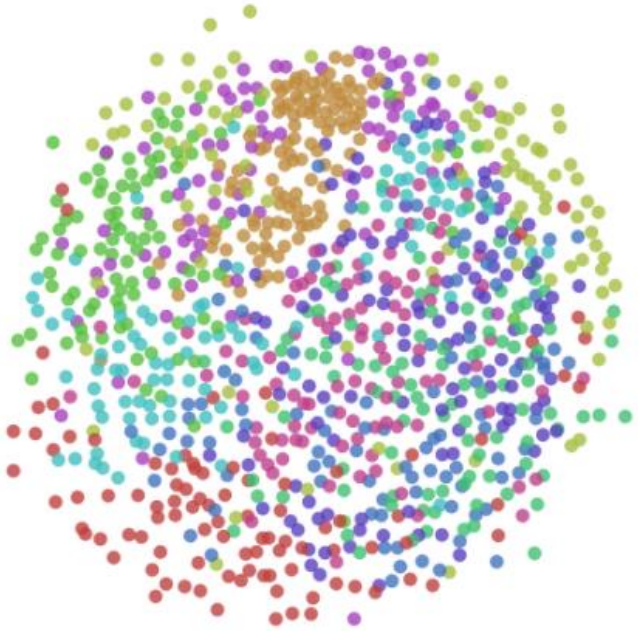
전체 데이터를 K개의 그룹(Fold)로 나눈 후

한 개의 그룹을 검증 데이터셋으로,

나머지 K-1 개의 그룹을 학습 데이터셋으로 설정해 K번의 검증을 반복

고차원 모델

변수가 많아 복잡한 모델



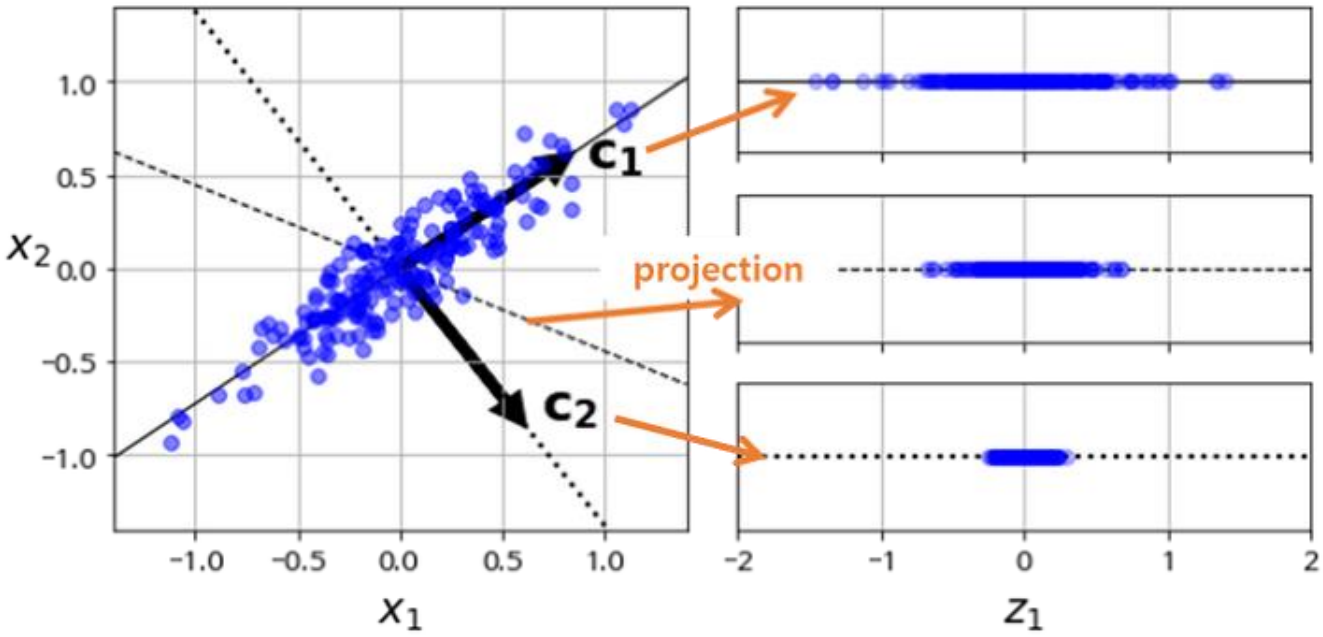
많은 변수 = 복잡한 모델

→ 과적합 초래

변수선택 필요

차원 축소

PCA

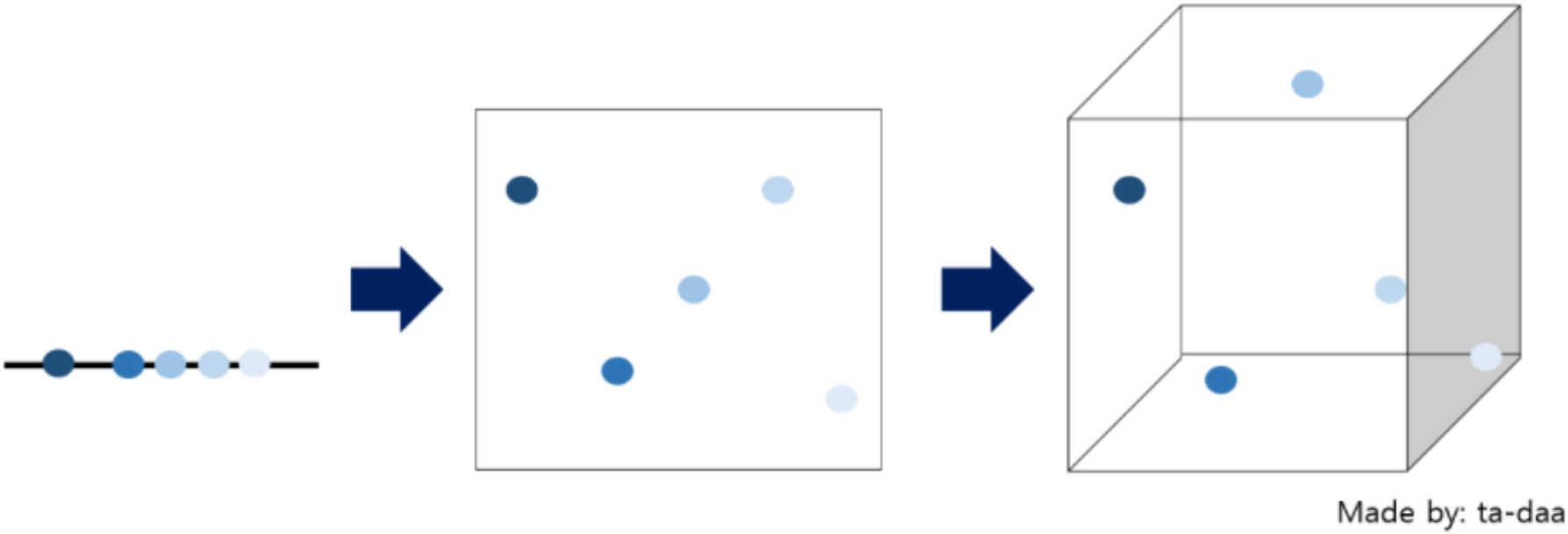


데이터의 분산(variance)을 최대한 보존하면서

서로 직교하는 새 기저를 찾아 차원 축소

차원의 저주

Curse of Dimensionality



Made by: ta-daa

차원이 증가(변수가 증가)함에 따라 모델 성능 저하

데이터 사이에 빈 공간 생김 (관측값 없음)