데이터마이닝팀

4팀

장이준
이선민
김영호
김현우
박시언

CONTENTS

1

데이터 마이닝

데이터 마이닝

데이터 마이닝이란?

Definition of Data Mining

DATA(데이터) + MINING(채굴)

"A Process of extracting useful information and patterns from large amount of data"

대량의 데이터로부터

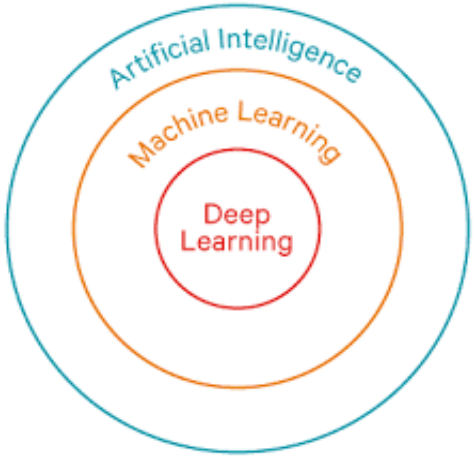
유용한 정보와 패턴을 추출해내는 과정

데이터 마이닝

인공 지능, 머신 러닝, 딥 러닝?

인공 지능(AI, Artificial Intelligence)

컴퓨팅을 이용한 학습 과정을 모두 포함하는
포괄적인 개념
머신 러닝과 딥 러닝을 모두 포함



기계 학습(Machine Learning)

사람의 개입이 최소화된 학습 수행 방법
분류/예측에 따른 적절한 모델을 선정하면
컴퓨터가 스스로 데이터를 학습 후 결과 도출

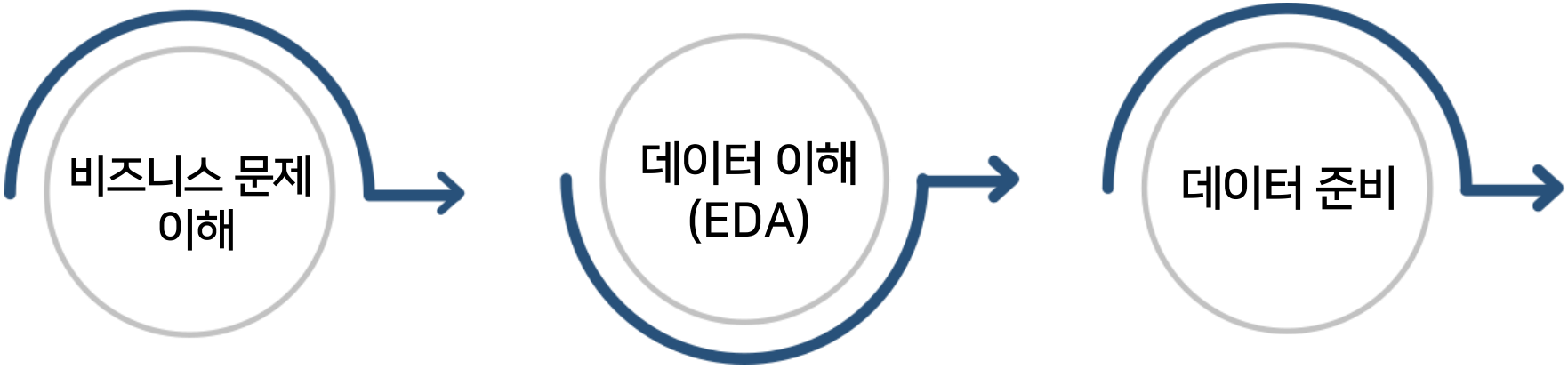
딥 러닝(Deep Learning)

사람의 신경망과 유사한 학습 체계를 구축해
목적 달성을 위한 과정을 수행
도출된 결과 해석이 어려워 '블랙박스 모델' 이라고 불림

데이터 마이닝

방법론: CRISP_DM

CRoss-Industry Standard Process for Data Mining



- 과제의 목적과 요구사항 이해
- 도메인 지식을 활용하여 초기 프로젝트 계획 수립

- 해당 데이터에 대한 '탐색과 이해'
- 변수 분포, 추이, 상관관계 시각화, 통계량 확인 등
- 이상치(outlier)와 결측치 확인이 중요

- 데이터 전처리 과정
- 모델의 성능에 상당한 영향을 미치는 중요한 단계

데이터 마이닝

방법론: CRISP_DM

CRoss-Industry Standard Process for Data Mining

- 모델링 과정 수행 및 파라미터 최적화 단계
- 모델링 기법 선택, 모델 테스트 계획 설계, 모델 작성과 평가

- 모델링의 성과 평가
- 분류모델 (ex. misclassification rate)
- 회귀모델 (ex. RMSE, MAE)

- 실제 서비스 런칭 등의 유의미한 결론 도출

2

모델링

2 모델링

Train data, Test data란?

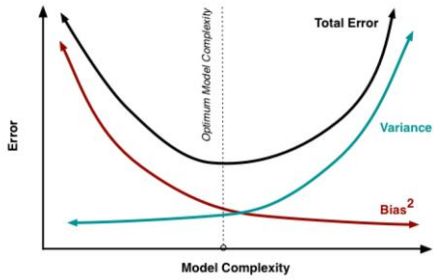
Train Data	Test Data
학습을 위한 데이터 종속변수, 독립변수 모두 존재	모델을 테스트하기 위한 데이터 종속변수 존재하지 않음

Variance-bias Trade-off

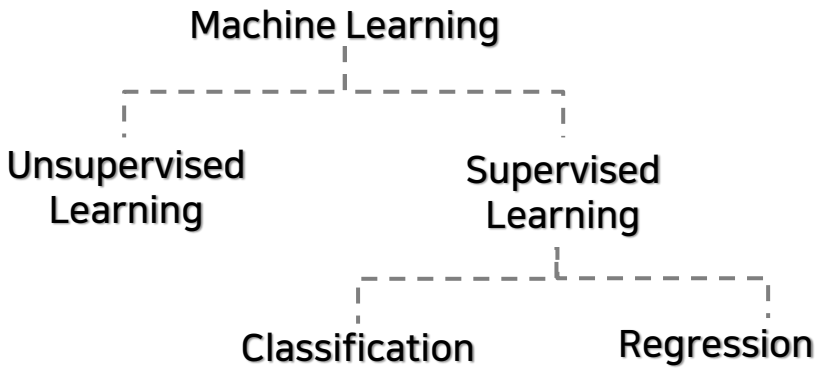
Bias 추정한 모델이 실제 모델을 얼마나 잘 설명하는지와 관련된 지표 $(f - E[\hat{f}])^2$

Variance 추정한 모델과 다른 데이터셋을 적합했을 때 모델이 달라지는 정도 $E\left[(E[\hat{f}] - \hat{f})^2\right]$

" Variance-Bias Trade-off "

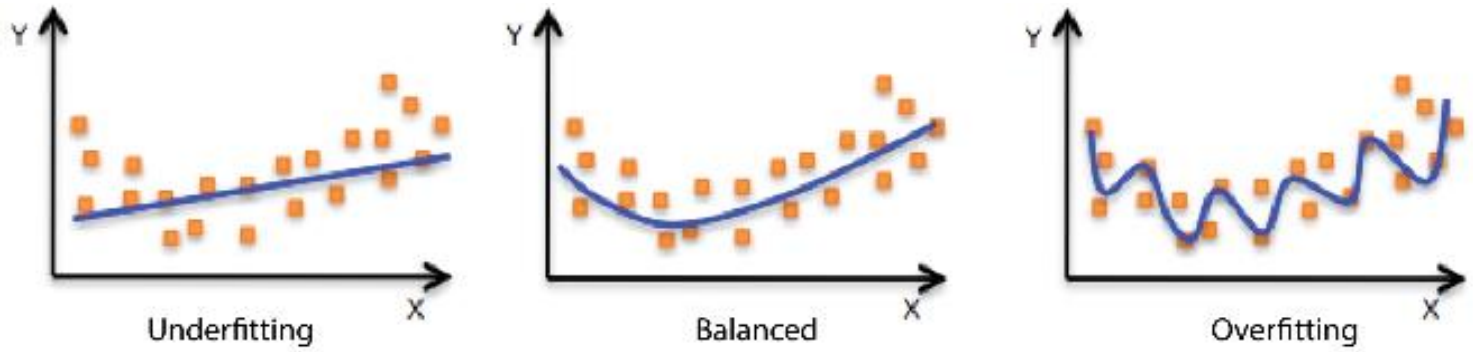


모델링(머신 러닝)의 종류



Variance-Bias Trade-off

Complex model

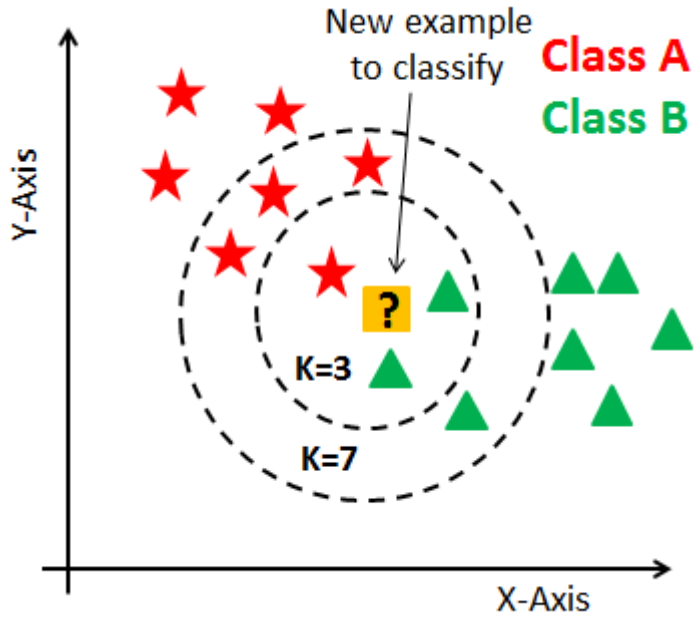


Bias와 Var가 적당히 작아 MSE가 최소가 되는

model을 찾아내는 것이 관건!

KNN(K-Nearest Neighbor)

비모수적 지도학습의 대표적 모델



Hyperparameter

“ k의 개수에 따라 달라지는
model의 decision boundary “

3

과적합 방지법

3 과적합을 방지하는 방법

과적합이란?

과적합

Train data에 대한 설명력은 높아도
Test data에 대해서는 설명을 잘 못하게 되는 현상

Cross Validation (교차 검증)

분석과정에서 주어진 train data를 다시 train data와 test data로 나누어 모델의
적절성을 평가하는 방법

Hold-out

Train-test split을 통해 단일한 검증
데이터셋을 생성해내는 방법

LOOCV(Leave-one-out CV)

n개의 전체 데이터에서 한 개의 데이터를
검증 데이터로, 나머지 n-1개의 데이터를
학습데이터로 사용하여 n번의 검증

K-Fold 교차검증(K-Fold CV)

전체 데이터를 K 개의 집합으로 나눈 후
하나의 집합을 검증 데이터셋으로, 나머지
K-1개의 집합을 학습 데이터로
사용하여 총 K번의 검증을 시행

3 과적합을 방지하는 방법

차원의 저주

모델에서 고려하는 변수가 많은 경우, 즉 독립 변수가 많아
데이터의 차원이 높은 경우에 과적합이 발생

차원 축소

고차원의 데이터를 저차원으로 축소하는 것으로
차원의 저주를 방지할 수 있음

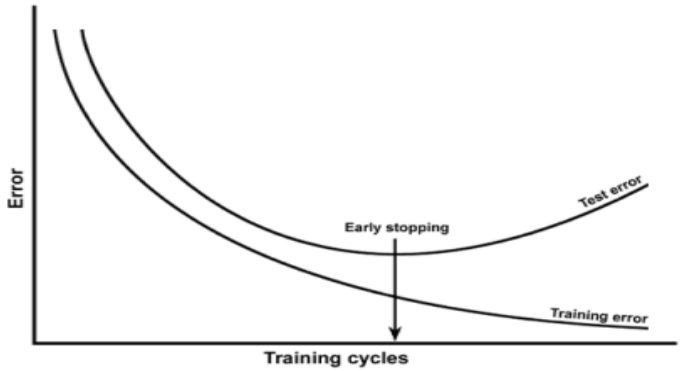
Feature Selection

데이터의 특성을
가장 잘 설명하는 변수를
추가하거나 제거해가며
모델을 적합시킴

Feature Extraction

데이터의 차원을
고차원에서 저차원으로
변환함으로써
모델을 적합시킴

Early Stopping(학습 조기 종료)



학습을 진행할 때 소요되는 시간에 제한,
혹은 모델의 성능이 일정 수준에 도달하게 되면
학습을 조기에 종료