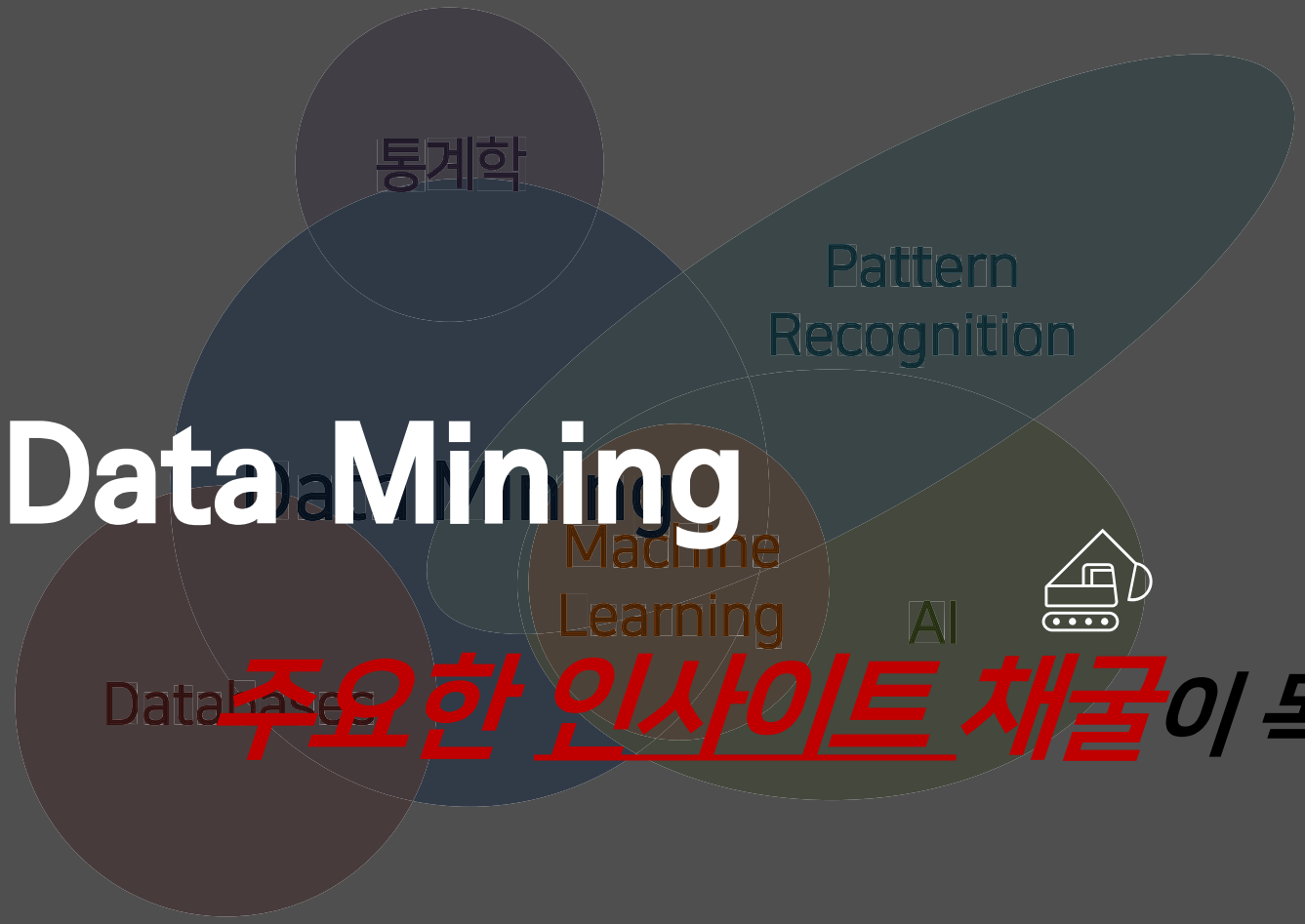
CONTENTS

1 What is Data Mining?

정의 및 접근법

여러 학문과 밀접히 맞닿아있어!



Data Mining

주요한 인사이트 채굴이 목표!

1 What is Data Mining?

프로세스: CRISP-DM

1. 비즈니스 문제 이해

- 비즈니스 상황의 배경지식 쌓기
- 데이터 마이닝 과정의 성공 여부 기준 세우기

2. 데이터 이해 (EDA)

- 시각화를 통해 데이터 직관적 이해 달성
- 변수의미, 변수 간의 관계 파악
- 이상치, 결측치 유무 파악

3. 데이터 준비

- 데이터 전처리 과정
- 모델의 성능 개선에 주요한 역할

프로세스: CRISP-DM

4. 데이터 분석과 모델링

- 머신러닝 / 딥러닝 기법 적용
- 추천, 예측, 해석 등

5. 분석 모델의 평가

- '모델링이 잘 되었는지' 평가
- 범주형 데이터 (misclassification rate)
- 연속형 데이터 (RMSE)

6. 분석 결과의 적용

- 실제 비즈니스 상황에 적용

What is Modeling?

지도학습 (supervised learning)

독립변수 종속변수

Bedrooms	Sq. feet	Neighborhood	Sale price
3	2000	Normaltown	$250,000
2	800	Hipsterton	$300,000
2	850	Normaltown	$150,000
1	550	Normaltown	$78,000
4	2000	Skid Row	$150,000

[training data]

Bedrooms	Sq. feet	Neighborhood	Sale price
3	2000	Hipsterton	???

[test data]

독립변수
Feature

특정 주택 내 방의 개수,
면적, 주택이 속한 동네

종속변수
Target

주택 가격

편향-분산 트레이드 오프(Bias-Variance Tradeoff)

구한 예측치가 실제값과의 **차이가 작을수록** 좋은 모델!

* **회귀분석의 관점)** MSE(Mean Squared Error)를 최소화

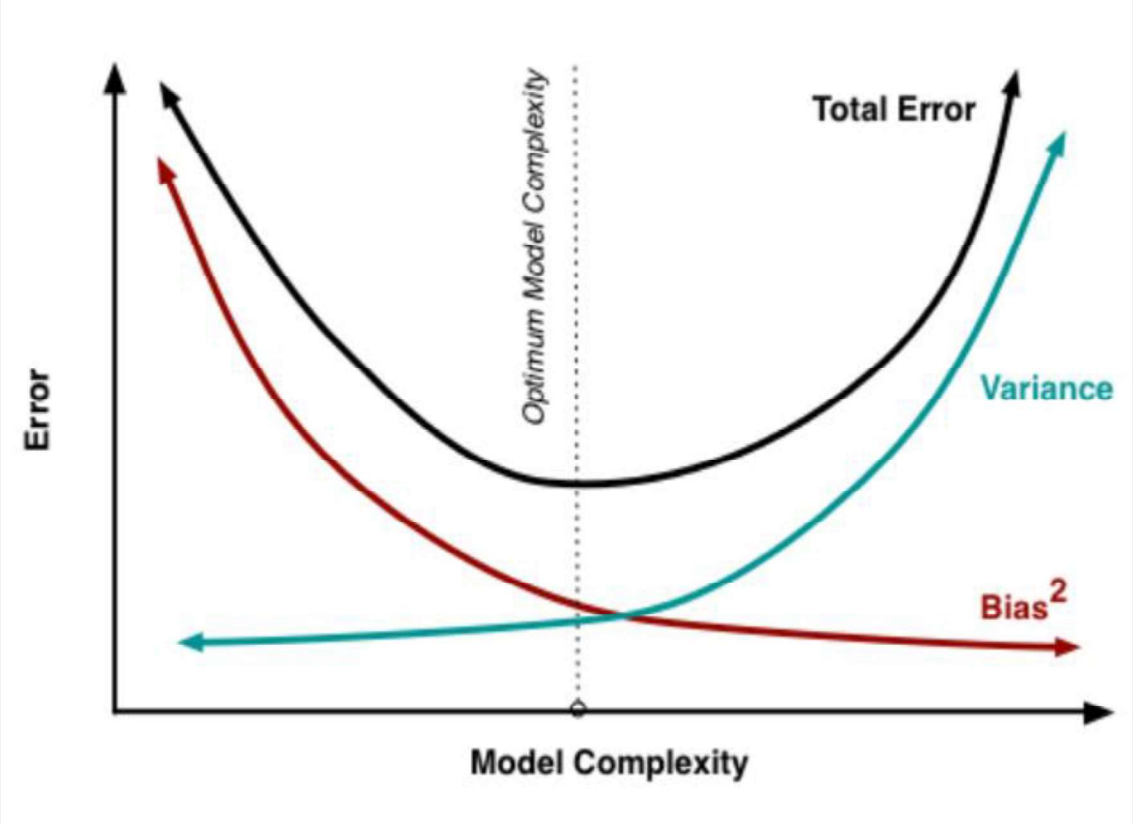
MSE

$$
\begin{aligned}
\mathrm{E}[(y-\hat{f})^2] &= \mathrm{E}[y^2+\hat{f}^2-2y\hat{f}] \\
&= \mathrm{E}[y^2]+\mathrm{E}[\hat{f}^2]-\mathrm{E}[2y\hat{f}] \\
&= \mathrm{Var}[y]+\mathrm{E}[y]^2+\mathrm{Var}[\hat{f}]+\mathrm{E}[\hat{f}]^2-2f\mathrm{E}[\hat{f}] \\
&= \mathrm{Var}[y]+\mathrm{Var}[\hat{f}]+(f-\mathrm{E}[\hat{f}])^2 \\
&= \sigma^2+\mathrm{Var}[\hat{f}]+\mathrm{Bias}[\hat{f}]^2
\end{aligned}
$$

Irreducible Error

Reducible Error

편향-분산 트레이드 오프(Bias-Variance Tradeoff)



하지만,

Bias와 Variance를
동시에 원하는 수준으로
줄이기 어려움.

3 How to avoid overfitting

Overfitting [과적합 문제]

따라서,

우리가 설계한 **모델의 성능**을 평가할 때

조금 더 객관적일 필요가 있다!

완벽히 설명하는 모델 설계

새로운 데이터가 들어온다면?

새로운 데이터 (검증 데이터)에 대해서

어떻게 반응할지 궁금하다!

모델의 '재사용성'은 이렇게 날라가고…!

3 How to Avoid Overfitting

K-fold CV [K-fold 교차검증]

주의
우리나라에서 만들어서 앞에 K 붙은 거 아님

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)으로 나눈 후

한 개의 데이터셋을 검증 데이터셋으로,

나머지 k-1개의 데이터셋을 학습 데이터셋으로 사용

3 How to Avoid Overfitting

자세한 내용은 회귀/선대/딥팀 교안 참고하시고~!

차원의 저주 [Curse of Dimensionality]

따라서, 너무 적지도 많지도 않은 적절한 변수 개수를 설정해야 하는데…

몇 가지 방법 소개해드림…!

1. Feature Selection

EX) Forward Selection,
Stepwise Selection

2. Feature Extraction

EX) Principal Component
Analysis(PCA, 주성분 분석)