


머신러닝에서 주로 마주하는 문제들-훈련데이터/편향/데이터품질/EDA/과대적합/과소적합

3강. 머신러닝에서 주로 마주하는 문제들

- 클래스 불균형
- 과대적합, 과소적합
- 탐색적 데이터 분석(EDA)

■ 머신러닝에서 주로 마주하는 문제들

- 
- 훈련 데이터의 부족
 - 대표성 없는 훈련 데이터
 - 낮은 품질의 데이터
 - 관련 없는 특성
 - 과대적합(Overfitting)
 - 과소적합(Underfitting)

머신러닝에서 주로 마주하는 문제들

· 훈련 데이터의 부족

· 대표성 없는 훈련 데이터

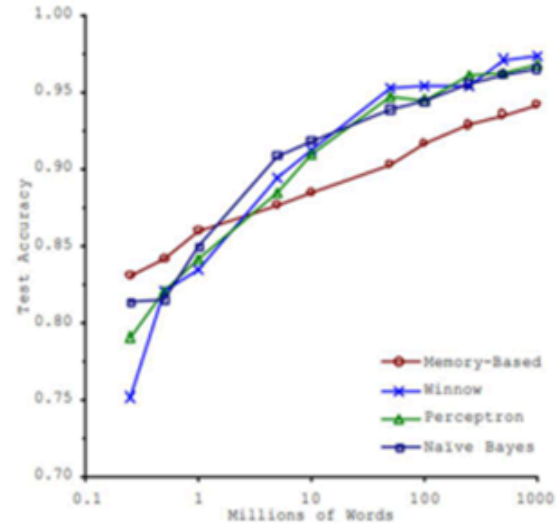
· 낮은 품질의 데이터

· 관련 없는 특성

· 과대적합(Overfitting)

· 과소적합(Underfitting)

대부분의 머신러닝 알고리즘이 잘 작동하려면 데이터가 많아야 함



※ 출처 : Michele Banko, Eric Brill의 'Scaling to Very Very Larger Corpora for Natural Language Disambiguation'
 Link - [Scaling to Very Very Large Corpora for Natural Language Disambiguation - ACL Anthology](#)

머신러닝에서 주로 마주하는 문제들

· 훈련 데이터의 부족

· 대표성 없는 훈련 데이터

· 낮은 품질의 데이터

· 관련 없는 특성

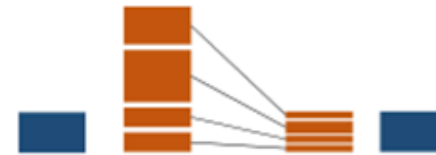
· 과대적합(Overfitting)

· 과소적합(Underfitting)

클래스 불균형 문제

수집된 데이터의 각 클래스가 갖고 있는 데이터 양의 차이가 큰 경우

Under Sampling



· 랜덤 추출법

랜덤으로 데이터를 추출

· 계통 추출법

데이터에 순번을 정해 등간격으로 추출

· 집락 추출법

데이터의 무리 안에서 랜덤 추출

· 층화 추출법

특징이 이질적인 층을 나눈 다음에 각 층에서 추출

머신러닝에서 주로 마주하는 문제들

· 훈련 데이터의 부족

· 대표성 없는 훈련 데이터

· 낮은 품질의 데이터

· 관련 없는 특성

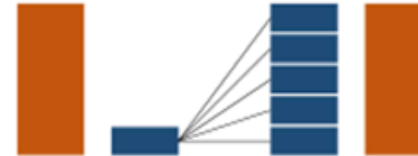
· 과대적합(Overfitting)

· 과소적합(Underfitting)

클래스 불균형 문제

수집된 데이터의 각 클래스가 갖고 있는 데이터 양의 차이가 큰 경우

Over Sampling



· SMOTE (Synthetic minority oversampling technique)
거리가 가까운 유사한 데이터들을 합성하여 새로운 데이터를 생성

· ADSYN (Adaptive Synthetic Sampling Approach)
SMOTE에서 분포를 고려하여 체계적으로 조절

· ROS (Random Over Sampling)
무작위로 선택하여 반복추출

머신러닝에서 주로 마주하는 문제들

· 훈련 데이터의 부족

· 대표성 없는 훈련 데이터 (샘플링 편향)

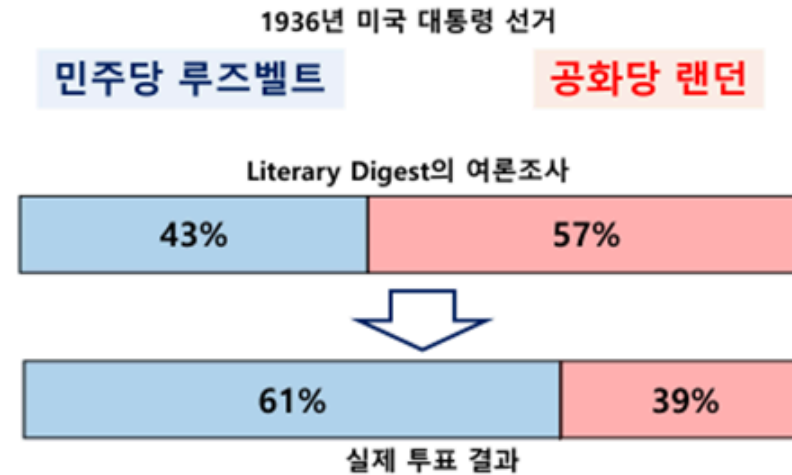
일반화하기 위하여 훈련 데이터가
대부분의 사례를 대표하는 것이 중요

· 낮은 품질의 데이터

· 관련 없는 특성

· 과대적합(Overfitting)

· 과소적합(Underfitting)



■ 머신러닝에서 주로 마주하는 문제들

· 훈련 데이터의 부족

· 대표성 없는 훈련 데이터

· 낮은 품질의 데이터

· 관련 없는 특성

· 과대적합(Overfitting)

· 과소적합(Underfitting)

나이	성별	키	몸무게
27	남	169	84
*12	-	1.2	-
34	-	99999	-
-	여	151	42
19	여	-	51
-	남	-	103
34	-	190	0.12
70	-	0.123	20
23	-	221Q	71
11	qewd	-	32

에러, 이상치, 잡음 등이 가득한 데이터

이상치 처리, 결손값 처리, 표준화 등
데이터 전처리가 중요

머신러닝에서 주로 마주하는 문제들

· 훈련 데이터의 부족

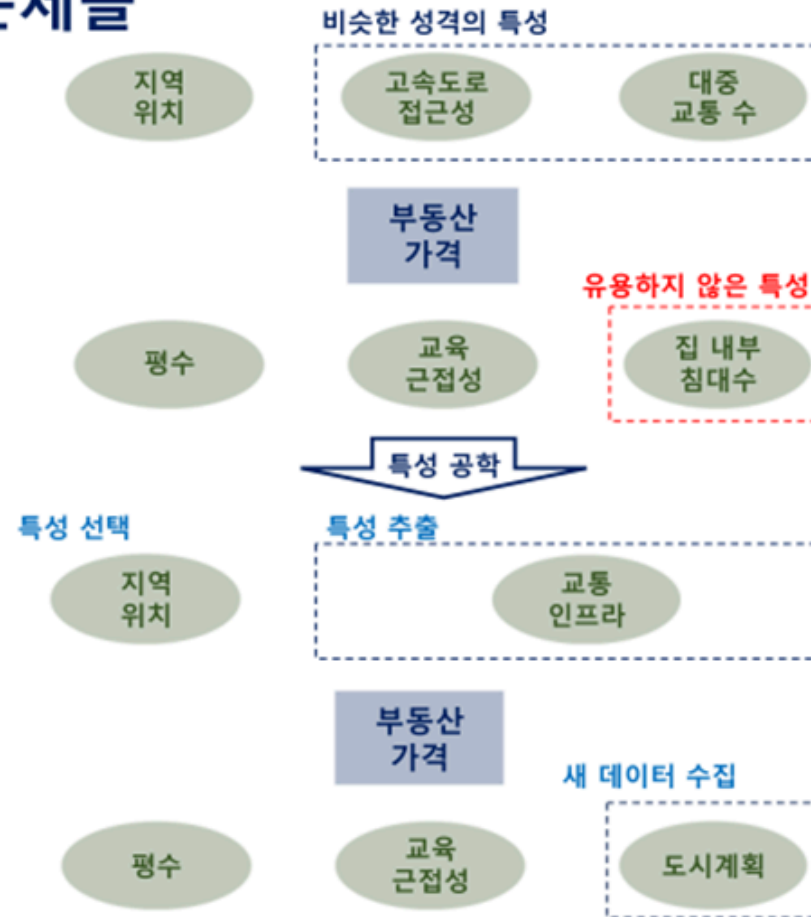
· 대표성 없는 훈련 데이터

· 낮은 품질의 데이터

· 관련 없는 특성

· 과대적합(Overfitting)

· 과소적합(Underfitting)



머신러닝에서 주로 마주하는 문제들

· 훈련 데이터의 부족

· 대표성 없는 훈련 데이터

· 낮은 품질의 데이터

· 관련 없는 특성

· 과대적합(Overfitting)

학습 데이터와 일치하지만 일반성이 떨어지는 경우

· 과소적합(Underfitting)

규칙을 제대로 찾지 못해 학습 성능이 낮은 경우



머신러닝에서 주로 마주하는 문제들

· 훈련 데이터의 부족

· 대표성 없는 훈련 데이터

· 낮은 품질의 데이터

· 관련 없는 특성

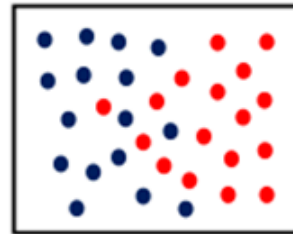
· 과대적합(Overfitting)

학습 데이터와 일치하지만 일반성이 떨어지는 경우

· 과소적합(Underfitting)

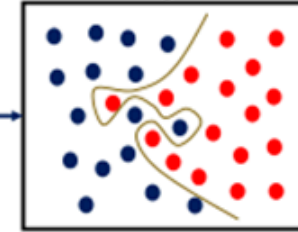
규칙을 제대로 찾지 못해 학습 성능이 낮은 경우

Training Data

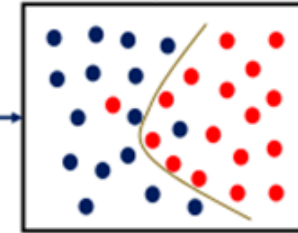


학습

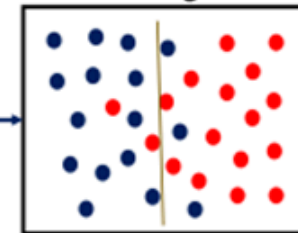
Overfitting



Justfitting



Underfitting



머신러닝에서 주로 마주하는 문제들

· 훈련 데이터의 부족

· 대표성 없는 훈련 데이터

· 낮은 품질의 데이터

· 관련 없는 특성

· 과대적합(Overfitting)

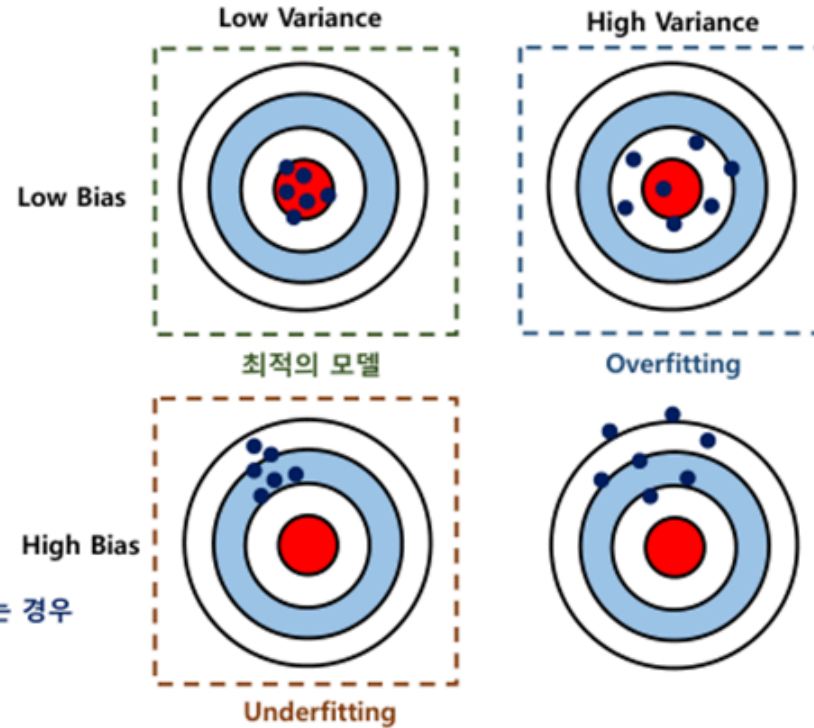
학습 데이터와 일치하지만 일반성이 떨어지는 경우

· 과소적합(Underfitting)

규칙을 제대로 찾지 못해 학습 성능이 낮은 경우

· 분산(Variance) : 데이터가 얼마나 퍼져 있는지 척도

· 편향(Bias) : 정답에서 멀리 떨어져 있는지 척도



머신러닝에서 주로 마주하는 문제들

· 훈련 데이터의 부족

· 대표성 없는 훈련 데이터

· 낮은 품질의 데이터

· 관련 없는 특성

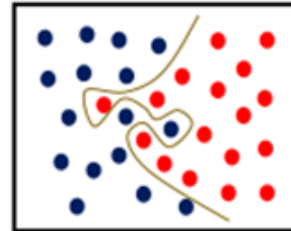
· 과대적합(Overfitting)

학습 데이터와 일치하지만 일반성이 떨어지는 경우

· 과소적합(Underfitting)

규칙을 제대로 찾지 못해 학습 성능이 낮은 경우

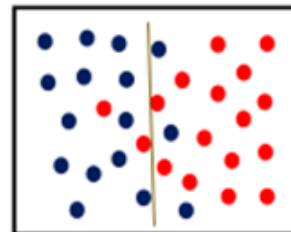
과대적합



높은 분산(High Variance)

- 파라미터 수가 적은 모델 선택
- 많은 훈련 데이터 수집
- 데이터의 잡음 제거
- 모델에 규제를 가함
 - * 하이퍼파라미터 조절

과소적합



높은 편향(High Bias)

- 파라미터가 더 강력한 모델 선택
- 특성 공학을 통한 좋은 특성 제공
- 모델의 규제를 줄임

