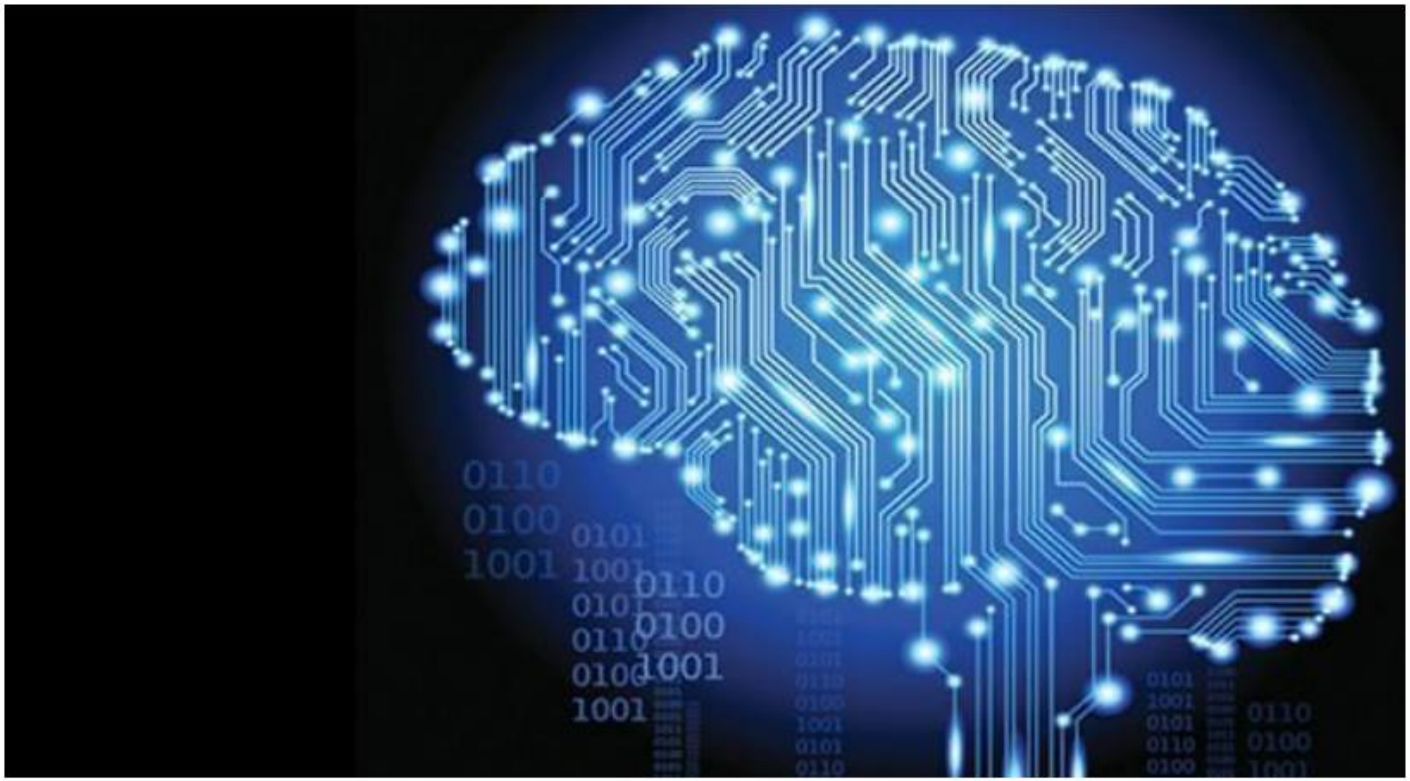
머신 러닝 개요

인공지능이란?

동양미래대학교

인공지능

사람과 유사한 지능을 가지도록 인간의 학습능력, 추론능력, 지각능력,
자연어 이해 능력 등을 컴퓨터 프로그램으로 실현하는 기술



인공지능, 머신러닝, 딥러닝

동양미래대학교

인공지능
Artificial Intelligence

인간의 지적 능력을 컴퓨터를 통해 구현하는 기술의 총칭

머신 러닝
Machine Learning

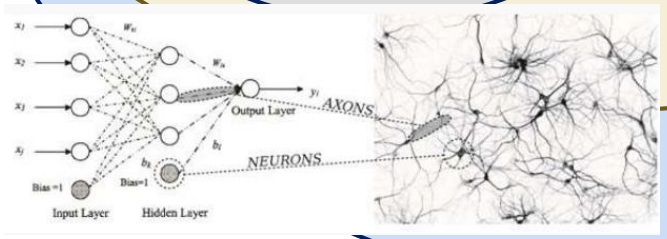
입력된 데이터를 기반으로 컴퓨터가 스스로 학습하여 인공지능의 성능을 향상시키는 기술 방법

e.g. 회귀분석, SVM, 결정트리, 군집화 등

딥 러닝
Deep Learning

인간의 뉴런과 비슷한 인공신경망 방식으로 컴퓨터를 스스로 학습하고 문제를 해결하는 방법

e.g. CNN, RNN, Neural Netrowk

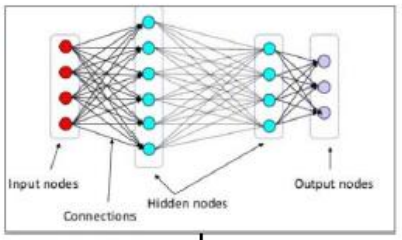


인공지능의 역사

Alan Turing (1912~1954)

Turing Test ('1950) :
기계(컴퓨터)가 인공지능을
갖추었는지를 판별하는 실험

AI Resurgence (부할)
Deep Learning



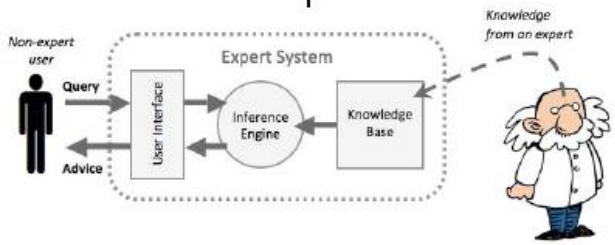
1st AI Winter
60년대 말 ~ 70년대

2nd AI Winter
80년대 후반 ~ 90년대 초

1950 1960 1970 1980 1990 2000 2010

1956 Dartmouth AI Project

('1956)
인공지능 용어 등장



AI Boom (80년대 중반)
전문가시스템
전문가와 동일한 또는 그 이상의
문제 해결 능력을 가질 수 있도록
만들어진 시스템

2020 2040 2060

Artificial Narrow Intelligence

- 약한 인공지능
- 정해진 목적에 특화된 작업 수행
- 자율 주행차, 번역기

Artificial General Intelligence

- 강한 인공지능
- 인간의 수준의 지능을 보유하여 전반적인 문제 해결 가능

Artificial Super Intelligence

- 초 인공지능
- 모든 영역에서 가장 유능한 사람보다 뛰어난 능력을 보유

현재 인공지능 붐 원인

머신러닝이 잘 동작할 수 있는 조건은?

멋진 알고리즘

기존의
학습 알고리즘들

요즘 뜨는
Deep Learning

많은 데이터

사람들의 생활이
디지털로 기록됨

클라우드에
집약된 데이터들

좋은 컴퓨터

Public 또는 private
클라우드에 있는
수백대의 컴퓨터들

때는 지금이다!

머신러닝의 정의

머신 러닝은 명시적인 프로그래밍 없이 컴퓨터가 학습하는 능력을 갖추게 하는 연구 분야이다.

아서 사무엘(Arthur Samuel, 1959)

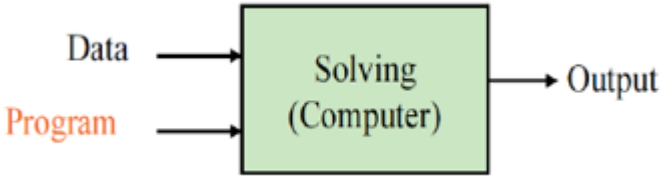
어떤 작업 T에 대한 컴퓨터 프로그램의 성능을 P로 측정했을 때 경험 E로 인해 성능이 향상됐다면, 이 컴퓨터 프로그램은 작업 T와 성능 측정 P에 대해 경험 E로 학습한 것이다.

톰 미첼(Tom Mitchell, 1997)

프로그래밍 방식의 차이점

일반적인 프로그램

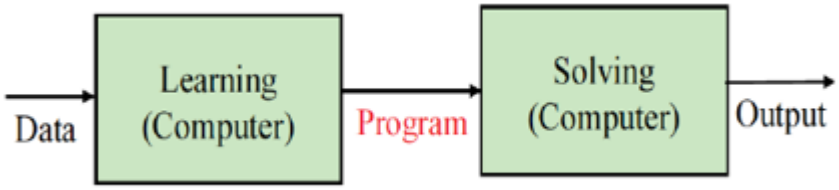
- 사람이 주어진 문제를 풀기 위한 알고리즘을 설계하고 구현



머신러닝 프로그램

- 컴퓨터에게 머신러닝 알고리즘과 데이터를 입력

- 컴퓨터가 입력데이터에 맞는 정답을 찾아내는 알고리즘(모델)을 생성



- 머신러닝이 필요한 문제
  - 명시적 문제해결 지식의 부재 (알고리듬 부재)
  - 프로그래밍이 어려운 문제 (예: 음성인식)
  - 지속적으로 변화하는 문제 (예: 자율이동로봇)
- 머신러닝 더욱 중요해지는 이유
  - 빅데이터의 존재 (학습에 필요)
  - 컴퓨팅 성능의 향상 (고난도 학습이 가능)
  - 서비스와 직접 연결 (비지니스적 효과)
  - 비즈니스 가치 창출 (회사 가치 향상)

머신 러닝의 종류

학습 데이터에 레이블(label)이 있는 경우와 그렇지 않은 경우에 따라 지도학습과 비지도학습으로 구분하고, 강화학습은 지도학습 중 하나로 분류되거나 또는 독립적인 세 번째 머신러닝 모델로 분류하기도 한다.

Types	Tasks	Algorithms
지도학습 (Supervised Learning)	분류 (Classification)	▪ KNN : k Nearest Neighbor ▪ SVM : Support Vector Machine ▪ Decision Tree (의사결정 나무) ▪ Logistic Regression
	예측 (Prediction)	▪ Linear Regression (선형 회귀)
비지도학습 (Unsupervised Learning)	군집 (Clustering)	▪ K-Means Clustering ▪ DBSCAN Clustering ▪ Hierarchical Clustering (계층형 군집)
강화학습 (Reinforcement Learning)		▪ MDP : Markov Decision Process

? 기계 학습 시 고려해야 할 사항은?

- 그 방법으로 문제를 풀 수 있는가?
- 그 방법을 적용할 경우 성능에 문제가 없는가?
- 그 방법을 적용하기 위해 데이터를 충분히 준비할 수 있는가?

머신 러닝 알고리즘들

동양미래대학교

지도 학습

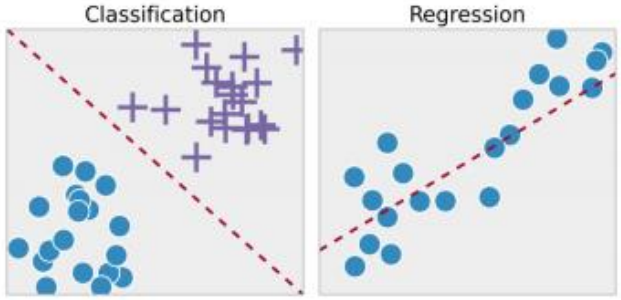
동양미래대학교

Label이 있는 학습 데이터(Training Set)를 이용해서 학습.

【 지도학습의 Training Set의 예】

- **분류와 회귀의 비교**

	분류 (Classification)	회귀 (Regression)
결과	학습데이터의 레이블 중 하나를 예측 (discrete)	연속된 값을 예측 (Continuous)
예제	학습데이터가 A, B, C 인 경우 결과는 A, B, C 중 하나다. 예) 스팸메일 필터	결과 값이 어떠한 값도 나올 수 있다. 예) 주가 분석 예측

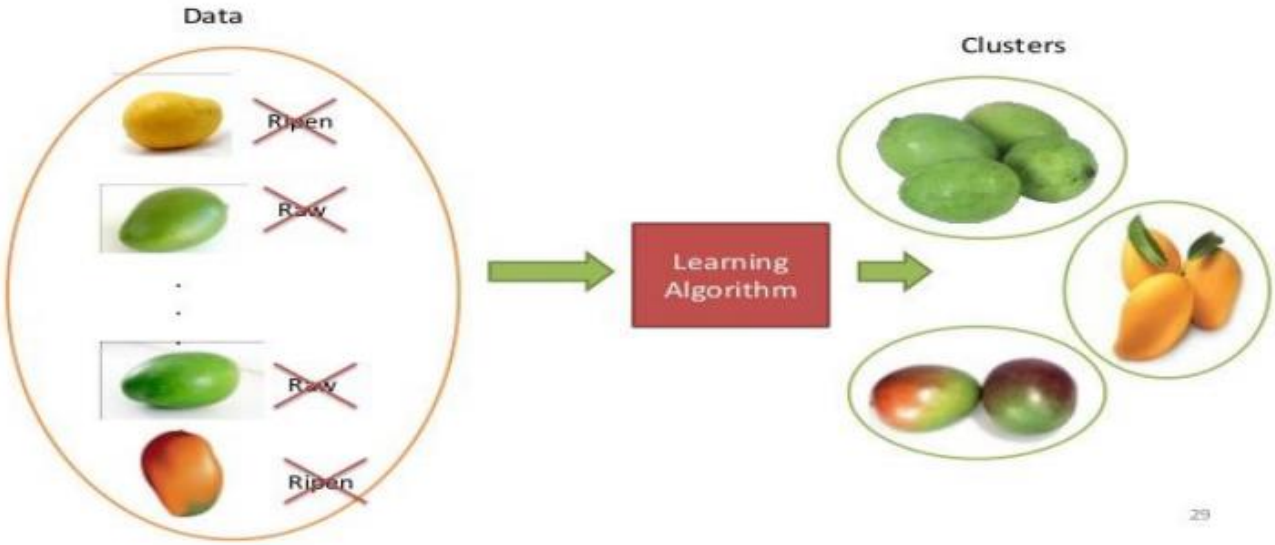


비지도 학습

동양미래대학교

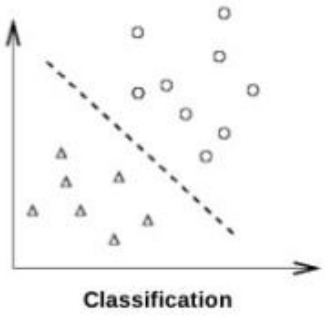
Label이 없는 학습 데이터(Training Set)를 이용해서 학습.

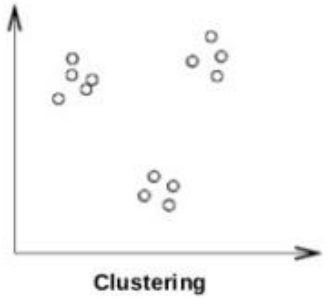
【 비지도학습의 Training Set의 예】



▪ **분류와 군집의 비교**

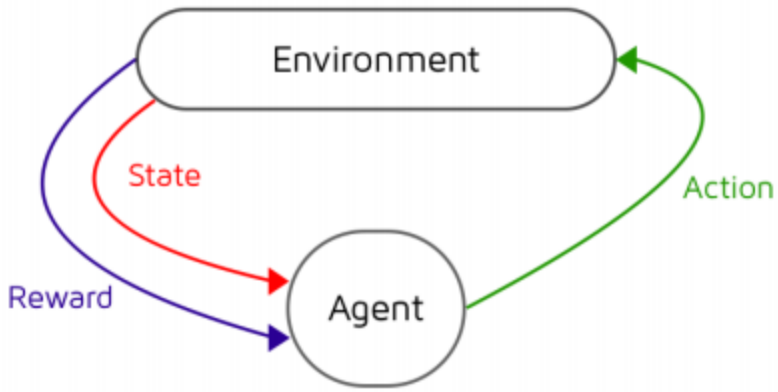
	분류 (Classification)	군집 (Clustering)
공통점	입력된 데이터들이 어떤 형태로 그룹을 형성하는지가 관심사	
차이점	레이블이 있다.	레이블이 없다. 예) 의학 임상실험 환자군 구별 예) 구매자 유형 분류





강화 학습

상과 벌이라는 보상을 통해 현재의 행위의 그 방향 혹은 반대 방향으로 행위를 강화화는 학습 방향



- **강화학습**

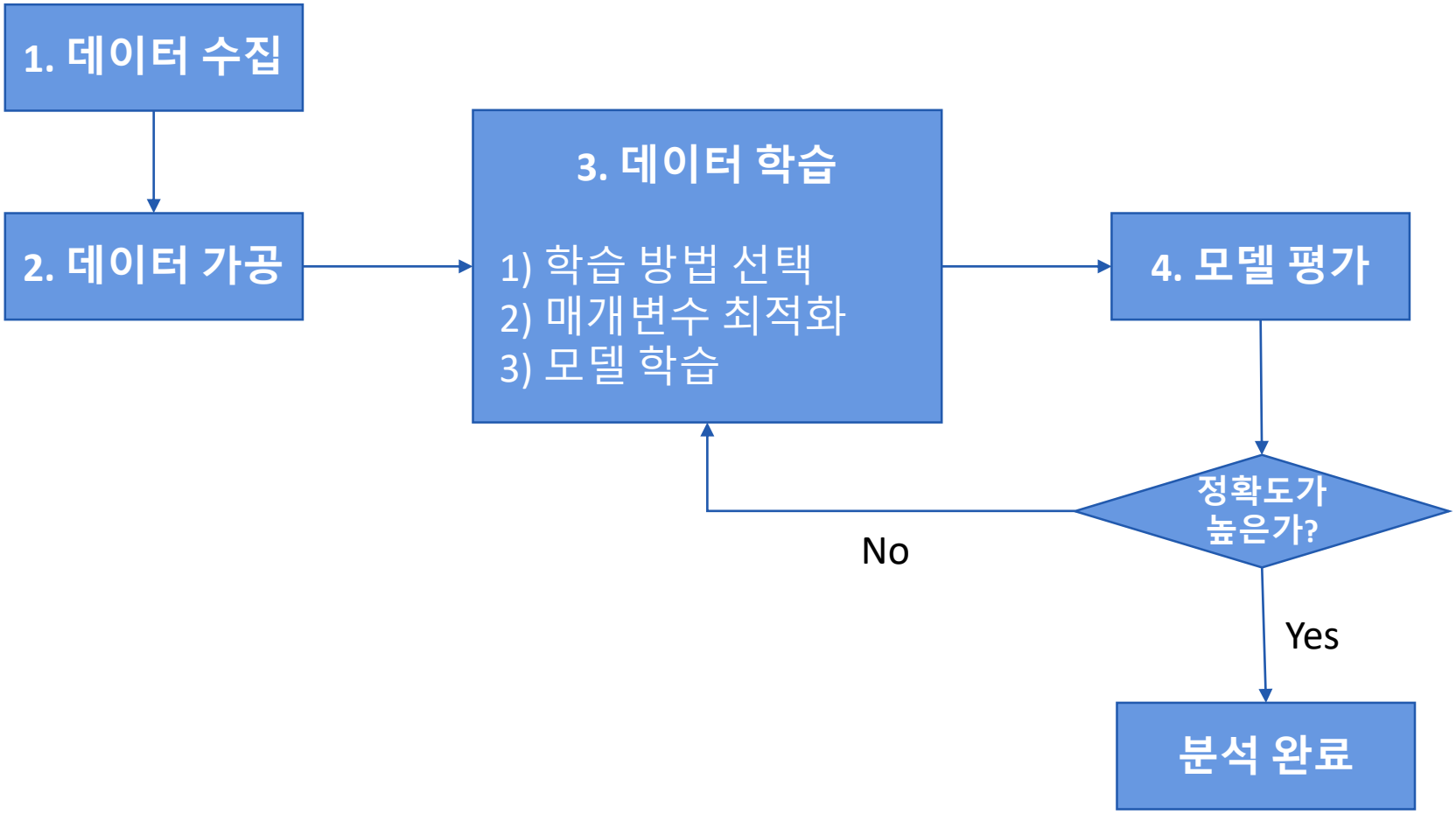
✓ 시행착오 과정을 거쳐 학습하기 때문에 사람의 학습방식과 유사

✓ Agent는 환경으로부터 상태를 관측하고 이에 따른 적절한 행동을 하면 이 행동을 기준으로 환경으로부터 보상을 받는다.

✓ 관측 – 행동 – 보상의 상호작용을 반복하면서 환경으로부터 얻는 보상을 최대화하는 태스크를 수행하기 위한 일련의 과정.

✓ 관측 – 행동 – 보상의 과정을 경험(Experience)이라고도 한다.

머신 러닝의 과정

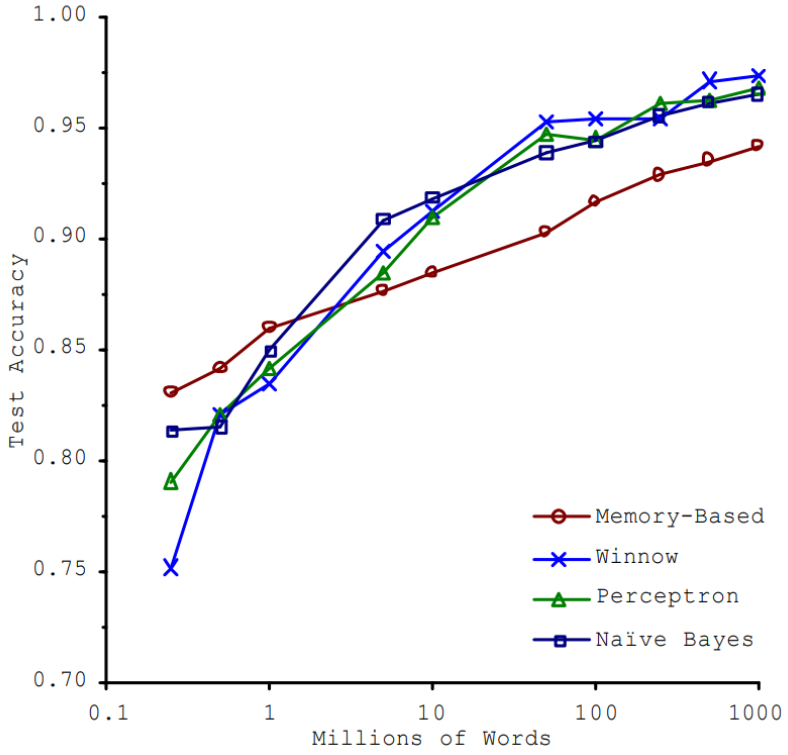

1. 데이터 수집
2. 데이터 가공
3. 데이터 학습
1) 학습 방법 선택
2) 매개변수 최적화
3) 모델 학습
4. 모델 평가
정확도가 높은가?
No
Yes
분석 완료

머신 러닝의 주요 도전 과제 - 데이터

동양미래대학교

1. 충분하지 않은 양의 훈련 데이터

충분한 데이터가 주어지면, 복잡한 자연어 중의성 해소 가능하다는 것을 증명
(Michele Banko&Eric Brill, 2001)



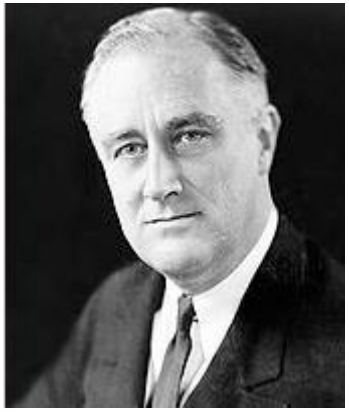
머신 러닝의 주요 도전 과제 - 데이터

2. 대표성 없는 훈련 데이터

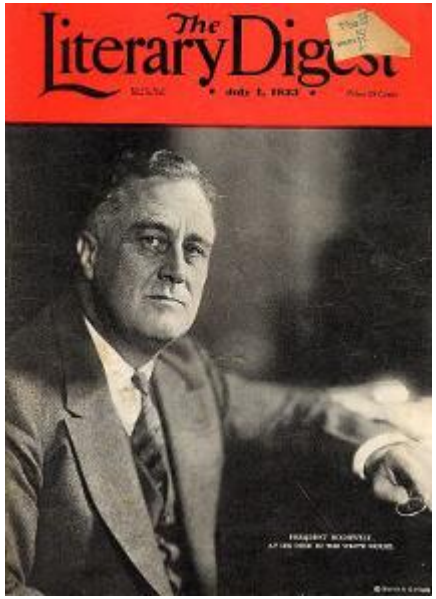
1936 미국 대선

Randon
(공화당)

VS



Roosevelt
(민주당)



https://ko.wikipedia.org/wiki/1936%EB%85%84_%EB%AF%B8%EA%B5%AD_%EB%8C%80%ED%86%B5%EB%A0%B9_%EC%84%A0%EA%B1%B0

머신 러닝의 주요 도전 과제 - 데이터

3. 낮은 품질의 데이터

에러 (error)

이상치(outlier)

잡음(noise)

데이터 정제 (전처리)

데이터 품질 향상

머신 러닝의 주요 도전 과제 - 데이터

4. 관련 없는 특성

Garbage In Garbage Out

특성 공학(Feature Engineering)

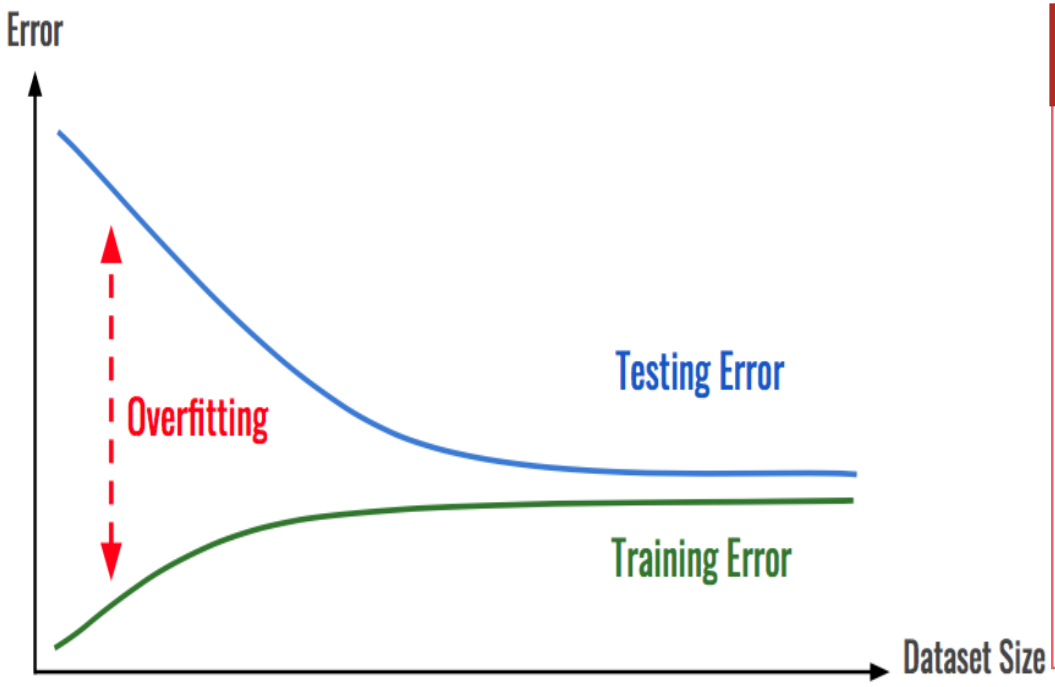
특성 선택feature selection : 가장 유용한 특성 선택

특성 추출feature extraction : 특성을 결합하여 더 유용한 새로운 특성을 만듬.(e.g. 차원 축소)

머신 러닝의 주요 도전 과제

5. 과대 적합 (overfitting)

훈련 데이터에 잡음이 많거나, 데이터셋의 크기가 작은 경우 자주 발생



해결책

1) 더 많은 훈련 데이터 확보

2) 훈련 데이터의 잡음 제거 (데이터 수정, 이상치 제거, 정규화 등)

3) 모델의 단순화 (알고리즘 변경, 규제(regulation) 등)

머신 러닝의 주요 도전 과제

6. 과소적합(underfitting)

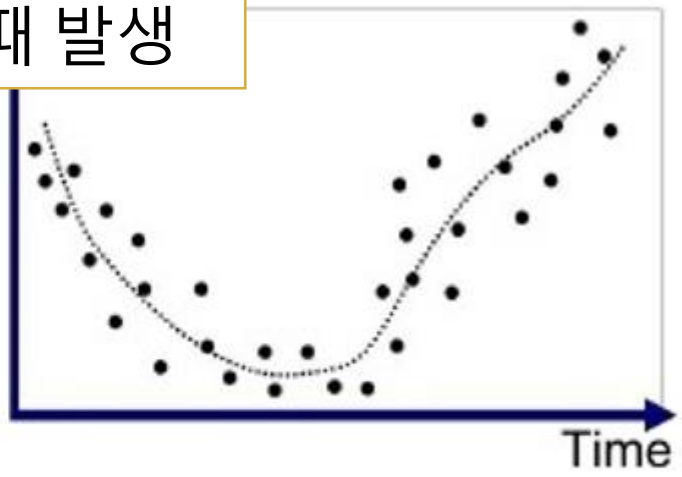
모델이 너무 단순해서 데이터의 내재된 구조를 학습하지 못할 때 발생

해결책

1) 특성 공학을 통한 데이터 정제

2) 더 강력한 모델을 선택

3) 모델의 규제(regulation)을 줄임.



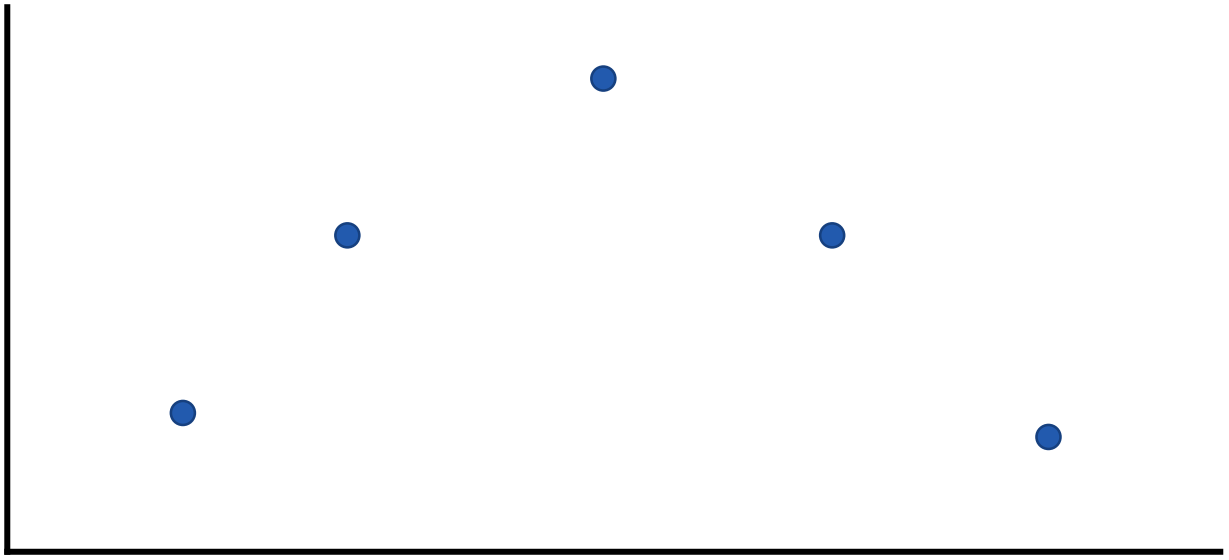
Underfitted

Good Fit/Robust

많은 머신 러닝 방법 중 어떤 알고리즘을 선택하는 게 가장 좋을까?

No Free Lunch 이론

어떤 알고리즘도 모든 문제에 대해서 다른 알고리즘보다 항상 좋을 수 없다.



1) 논문 원문「The Lack of A Priori Distinctions Between Learning Algorithms」, D. Wolperts (1996), https://goo.gl/uJxSvo
2) 간단 설명 : https://ml-dnn.tistory.com/1

배치 학습과 온라인 학습

동양미래대학교

▪ 배치 학습

- 가용한 데이터를 모두 사용하여 훈련
- 시간과 자원을 많이 소모하므로 보통 오프라인에서 수행
- 기훈련된 학습 결과(모델)을 시스템에 적용만 함. (업데이트 X) → 오프라인 학습

▪ 온라인 학습

- 데이터를 한개씩 또는 미니배치 단위로 입력받아 훈련
- 매 학습 단계가 빠르고 비용이 적게 듬.