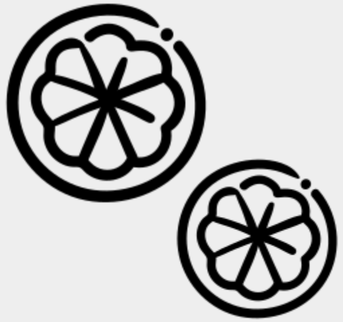


데이터 다루기

훈련 세트와 테스트 세트

9기 교육부 김진영, 김은혜

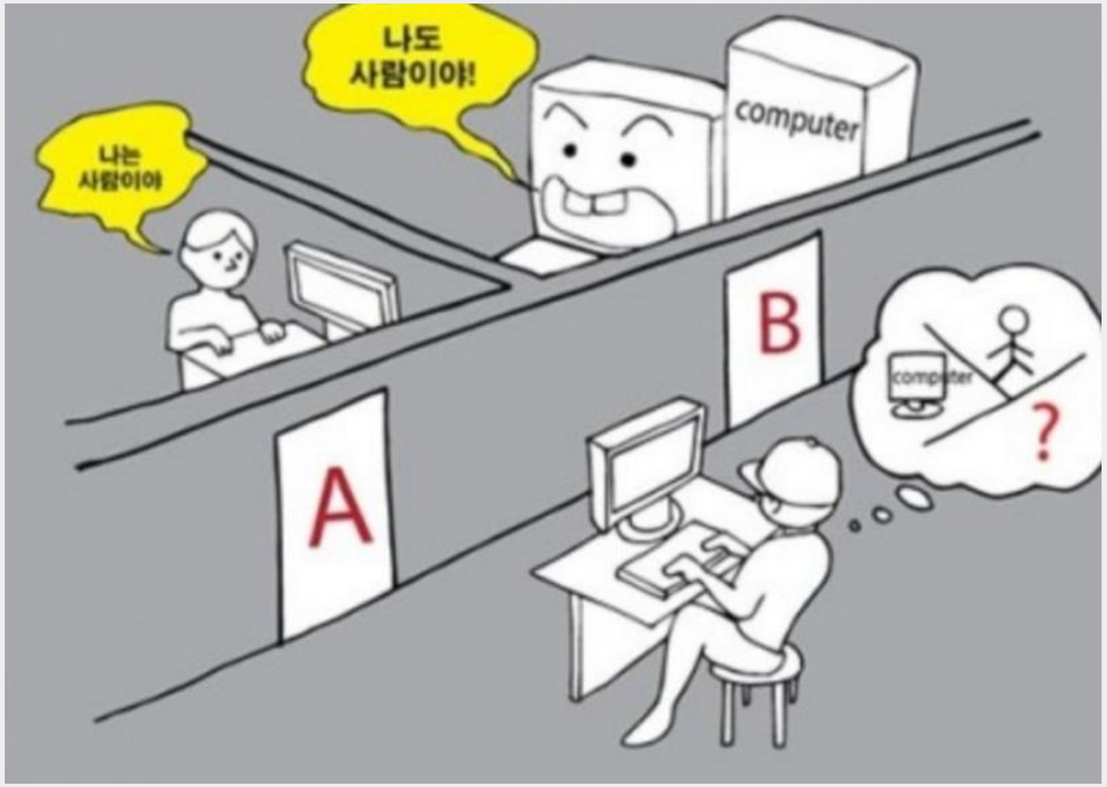
0. 내용 복습 

01 인공지능, 머신러닝, 딥러닝

02 데이터 분석을 위한 3종 패키지

03 복습 과제 풀이 (ex.)

기계도 생각할 수 있는가? _ Alan Turing (앨런 튜링)



튜링 테스트 : 기계가 "생각할 수 있는지"를 확인하는 검사

"기계가 생각한다" 의 조건

대화하는 상대가 사람인지 기계인지 구분할 수 없을 때

사람이 전혀 이상한 점을 느낄 수 없을 때

↓

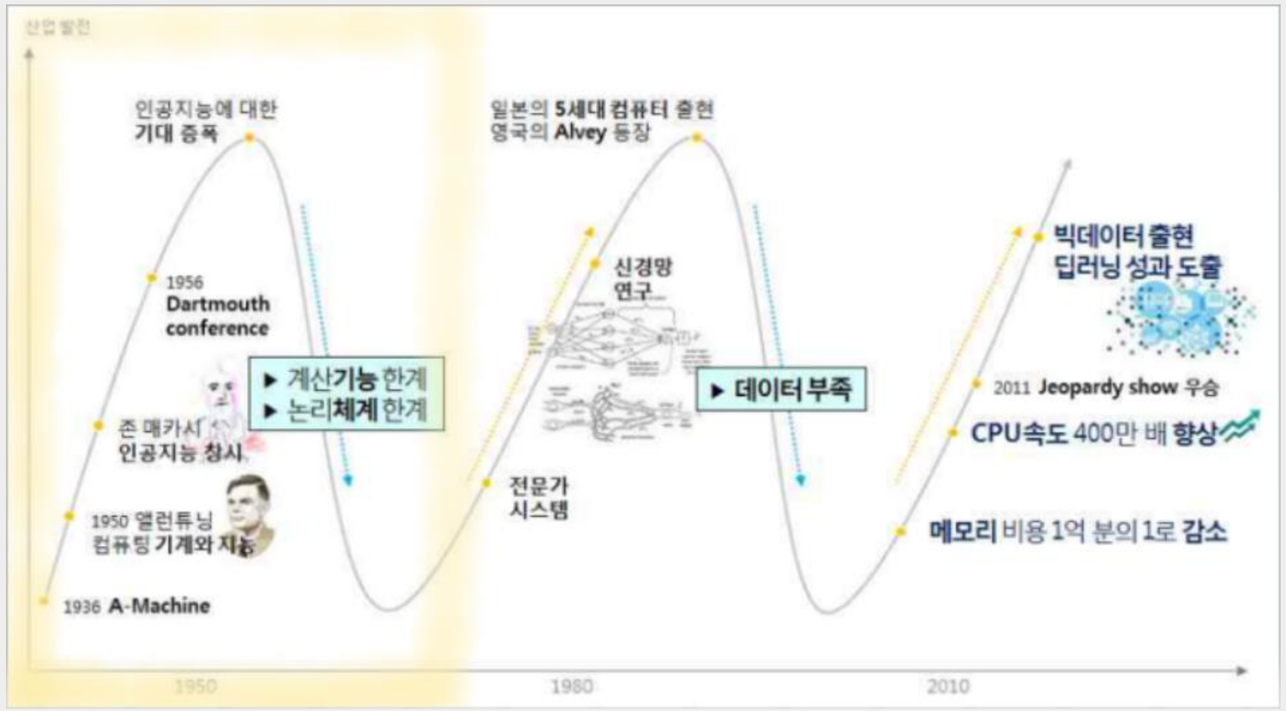
강한 인공지능

사람과 동격으로 자유로운 사고를 할 수 있는 인공지능

01 인공지능, 머신러닝, 딥러닝

인공지능 연구자들의 1차 목표 : '강한 인공지능'

계산주의 _ 사람이 가진 지식을 컴퓨터로 표현하고 이를 활용하여 현상 분석/ 문제 해결



자료: IDC&EMC, 디지털 유니버스 스터디 2011, 한국정보화진흥원

계산주의에서의 지능

계산능력을 통한 논리적 추론의 조합

의의

논리체계를 프로그래밍으로 다루는 응용분야에 많은 성과를 남김

한계

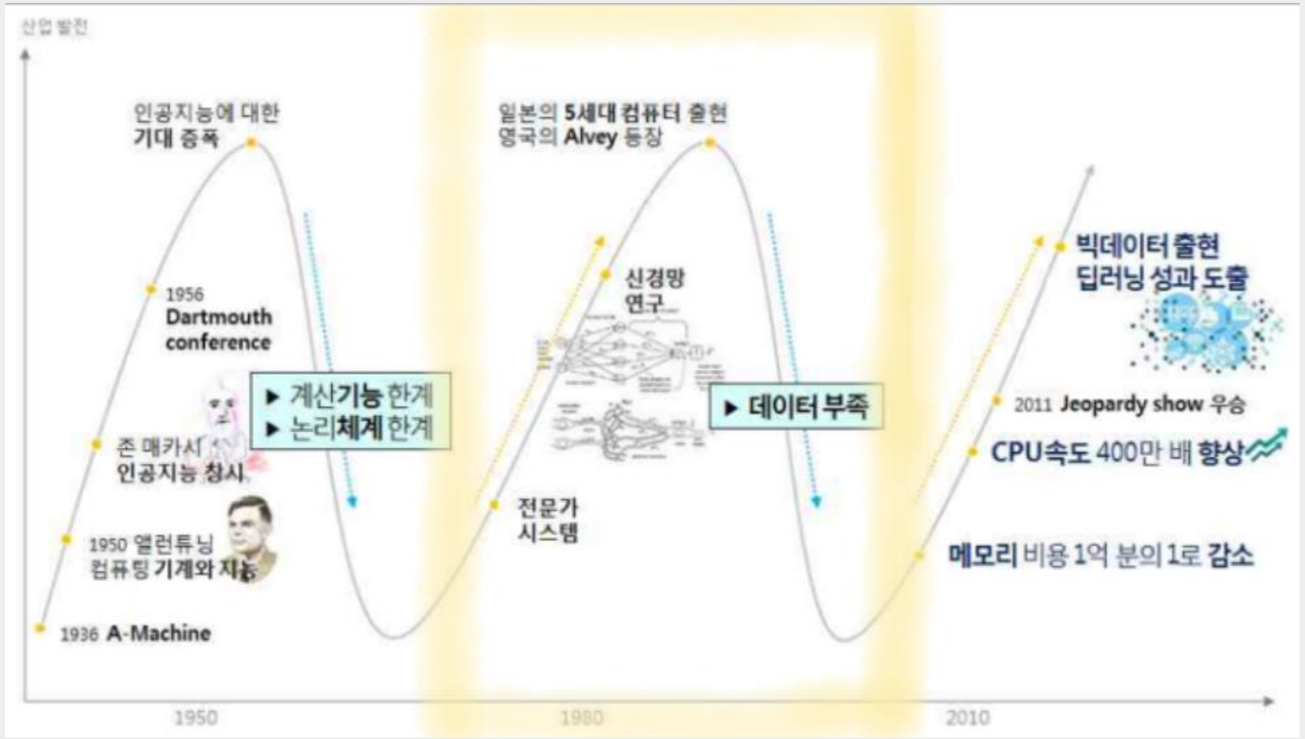
기계의 메모리(기억장치) 에 모두 기호로 구성 불가능

=> 강한 인공지능 만드는 데 실패

01 인공지능, 머신러닝, 딥러닝

인공지능 연구자들의 1차 목표 : '강한 인공지능'

연결주의 _ 지식과 정보가 포함된 데이터를 제공하고 컴퓨터가 스스로 필요한 정보를 학습



지능을 담는 두뇌의 물리적 구조에 주목

↓

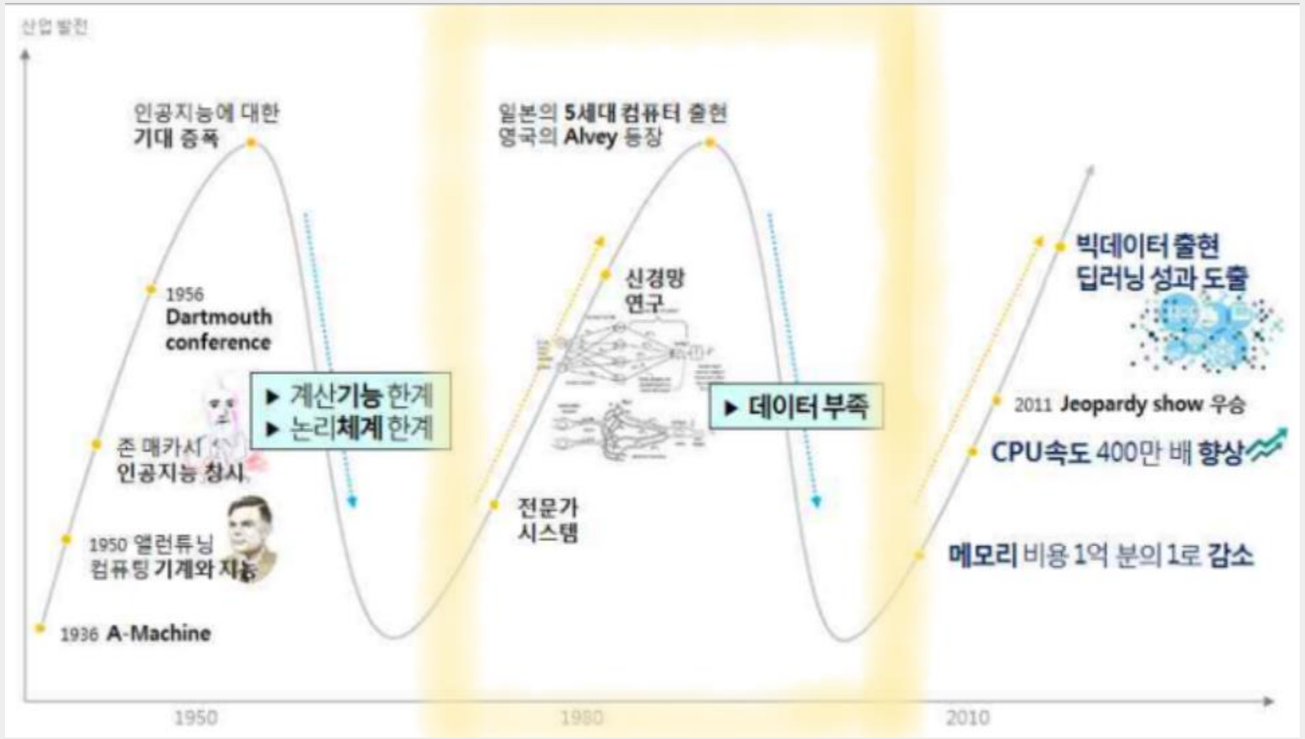
인공 신경망의 탄생

수많은 뉴런이 시냅스를 통해 연결된 신경망인 인간의 두뇌 구조를 컴퓨터 프로그램을 통해 복제하여 탄생

01 인공지능, 머신러닝, 딥러닝

인공지능 연구자들의 1차 목표 : '강한 인공지능'

연결주의 _ 지식과 정보가 포함된 데이터를 제공하고 컴퓨터가 스스로 필요한 정보를 학습



의의

"학습"의 역할을 부각

적절한 학습 병행시, 매우 복잡한 문제 해결 가능

한계

컴퓨터 성능의 부족

학습에 필요한 데이터의 부족

=> 강한 인공지능 만드는 데 실패

계산주의, 연결주의의 강인공지능 연구 실패...

인공지능 연구의 노력은 헛된 것인가?

강한 인공지능의 "지능" 범위 축소

강한 인공지능에서의 지능 : 인간처럼 생각하기

↓

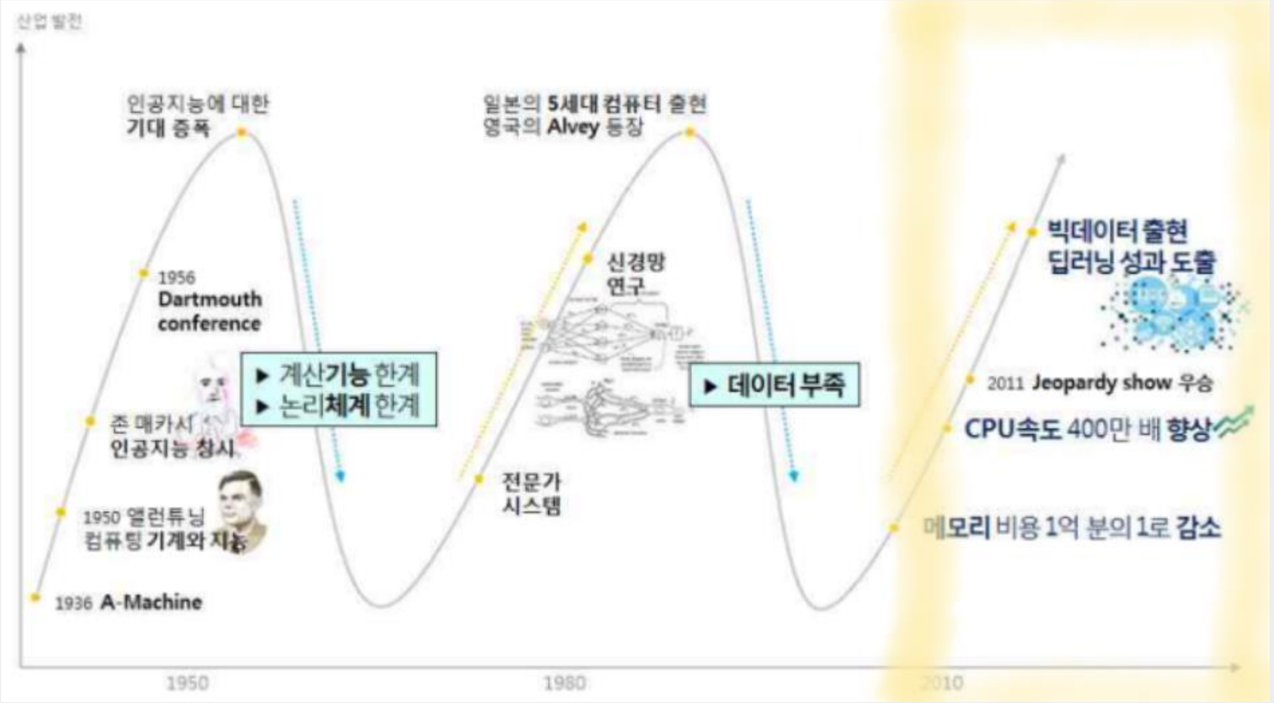
약한 인공지능

약한 인공지능에서의 지능 : 특정한 문제를 인간처럼 풀기

약한 인공지능의 범주에 속하는 각종 기술은
인간을 넘어서는 능력을 보여주고 있음

2번의 AI 겨울을 겪은

인공지능 분야의 급부상



컴퓨팅 파워의 급증

GPU를 활용한 분산처리 기술의 발전은 이전에 제기되었던 방대한 양의 계산 과제 대부분 해결

딥러닝에서의 새로운 해결방안 제시

빅데이터의 등장

클라우드 컴퓨팅과 IoT와 IoB로 발현된 엄청난 데이터는 인공지능의 결정적인 학습 도구를 제공

02 데이터분석을 위한 3종 패키지

Numpy

파이썬에서 배열을 사용하기 위한 표준 패키지

수치 해석용 파이썬 패키지로, 벡터/행렬 사용하는 선형대수 계산에 주로 사용

Pandas

고수준의 자료 구조와 빠르고 쉬운 데이터 분석 도구를 제공하는 파이썬 라이브러리

pandas의 자료구조에는 series, dataframe이 있음.

matplotlib

파이썬에서 자료를 차트나 플롯으로 시각화하는 패키지

03 복습 과제 풀이(ex.)

복습과제가 문제(퀴즈)였던 경우

문제 [3-3]

- mpg1이라는 새로운 데이터 프레임을 생성 후, mpg1 데이터를 출력하세요.
- 이 데이터 프레임은 fl, price_fl이라는 두 열을 가지고 있습니다.

- fl값이 c이면, price_fl 값은 2.66
- fl값이 d이면, price_fl 값은 2.23
- fl값이 e이면, price_fl 값은 2.11
- fl값이 p이면, price_fl 값은 2.89
- fl값이 r이면, price_fl 값은 2.49

```
mpg1 = pd.DataFrame({'fl': ['c', 'd', 'e','p','r'],
                     'price_fl': [2.66,2.23,2.11,2.89,2.49]})
mpg1
```

key : dataframe의 생성

step1. 하나의 열 데이터를 일차원 배열(리스트)로 준비

fl 열에 해당하는 데이터 : ['c', 'd', 'e','p','r'] 로 준비

price_fl 열에 해당하는 데이터 : [2.66,2.23,2.11,2.89,2.49]로 준비

step2. 각 열에 대한 이름을 키로 가지는 딕셔너리 생성

{'fl': ['c', 'd', 'e','p','r'], 'price_fl': [2.66,2.23,2.11,2.89,2.49]}

step3. dataframe 생성자에 데이터 넣기

pd.DataFrame({'fl': ['c', 'd', 'e','p','r'],'price_fl': [2.66,2.23,2.11,2.89,2.49]})

1. 훈련 세트와 테스트 세트

01 지도학습, 비지도학습, 강화학습

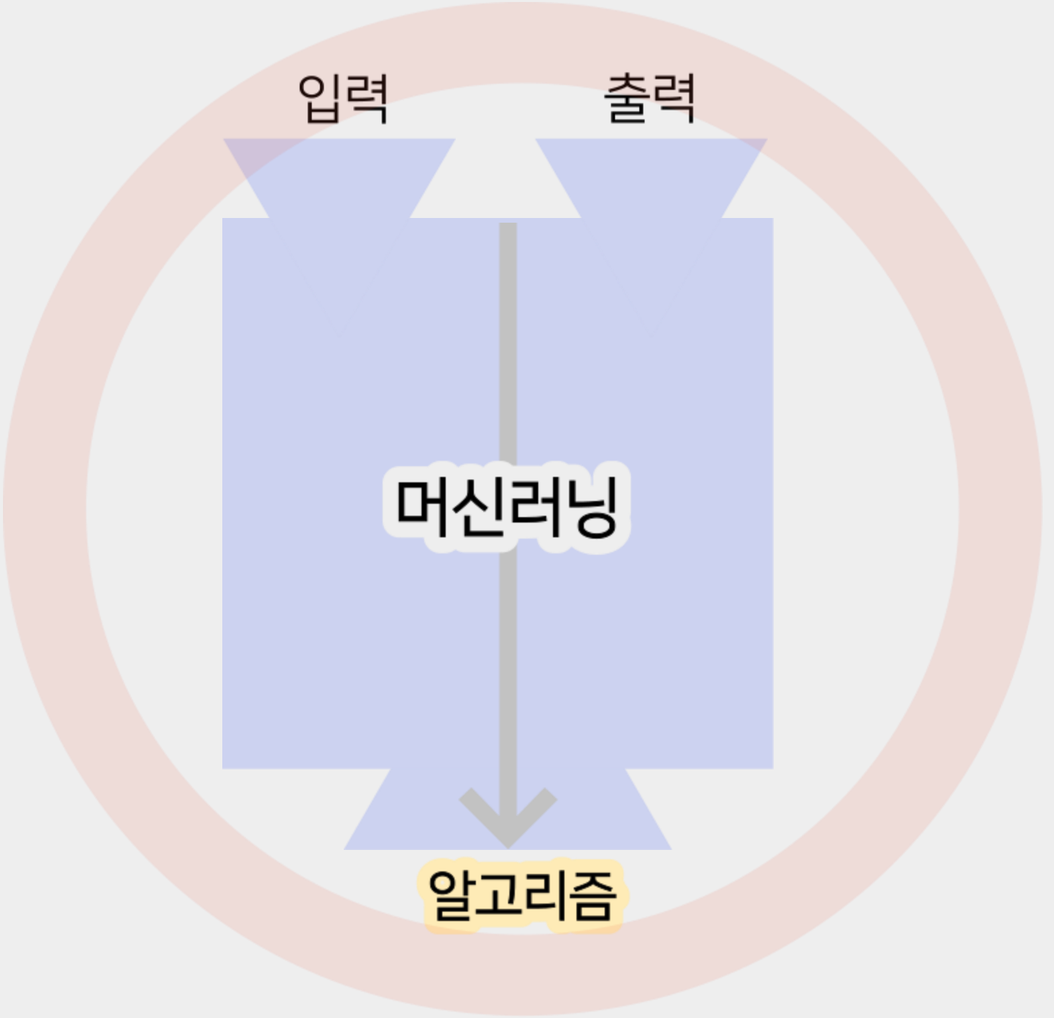
02 지도학습에서의 훈련세트, 테스트세트

01 지도학습, 비지도학습, 강화학습

머신러닝

인공지능의 한 분야

컴퓨터가 데이터를 이용하여 학습하는 알고리즘 기술



머신러닝으로 해결할 수 있는 문제

분류
Classification

예측
Forecast

지도(감독) 학습

이상값 감지
Anomaly Detection

그룹화
Clustering

비지도(무감독) 학습

강화학습
Reinforcement Learning

강화 학습

머신러닝으로 해결할 수 있는 문제

분류
Classification

예측
Forecast

지도(감독) 학습

분류? 어떤 여러개의 카테고리가 있을 때,
내가 가지고 있는 데이터는 어떤 카테고리에 속하는지

알고리즘 KNN, SVM, Decision Tree, Logistic Regression

분류문제의 ex) 손글씨 숫자 이미지 분류

머신러닝으로 해결할 수 있는 문제

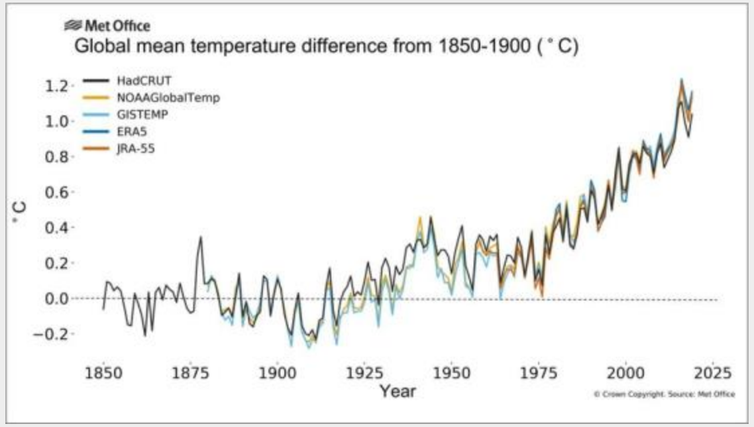
분류
Classification

예측
Forecast

지도(감독) 학습

예측 어떤 상황/특징이 주어졌을 때,
내가 가지고 있는 데이터의 (연속된) 값을 예측

알고리즘 Linear Regression (선형 회귀)



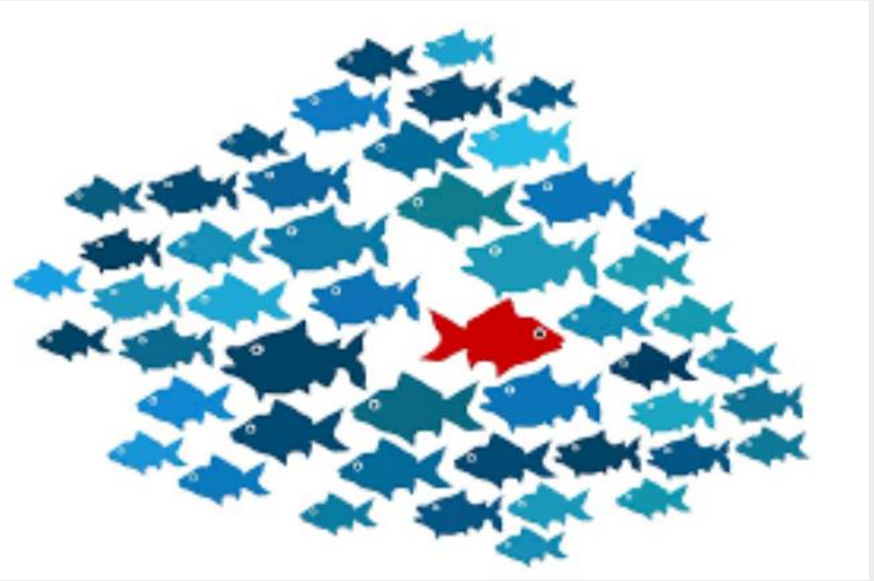
예측문제의 ex) n년 후, 기온 상승 예측

머신러닝으로 해결할 수 있는 문제

이상값 감지
Anomaly Detection

그룹화
Clustering

비지도(무감독) 학습



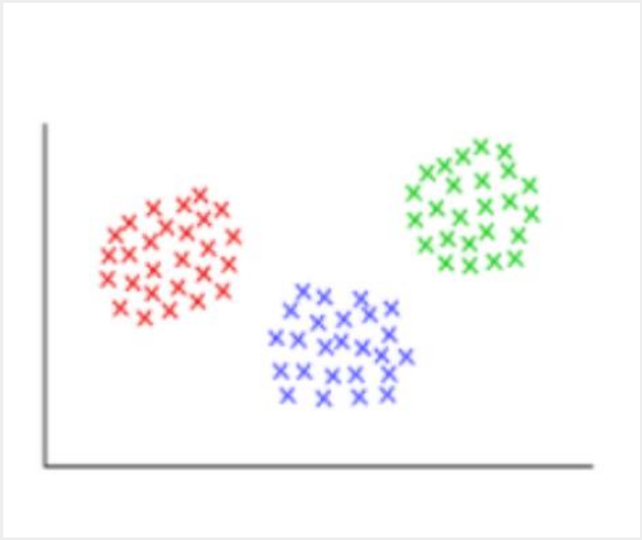
이상값 감지 평소와는 다른 패턴/값을 파악

머신러닝으로 해결할 수 있는 문제

이상값 감지
Anomaly Detection

그룹화
Clustering

비지도(무감독) 학습



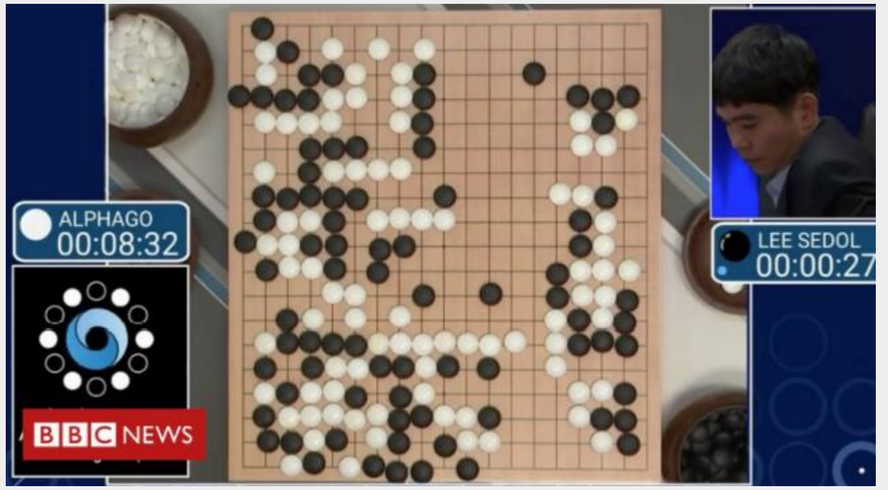
그룹화 어떤 특징을 기준으로,
내가 가지고 있는 데이터를 그룹화

알고리즘 K-Means Clustering, DBSCAN Clustering

머신러닝으로 해결할 수 있는 문제

강화학습
Reinforcement Learning

강화 학습



강화학습 ex) AlphaGo

강화학습

시행착오를 통해 학습

실수와 보상을 통해 학습하여 목표를 향해 감

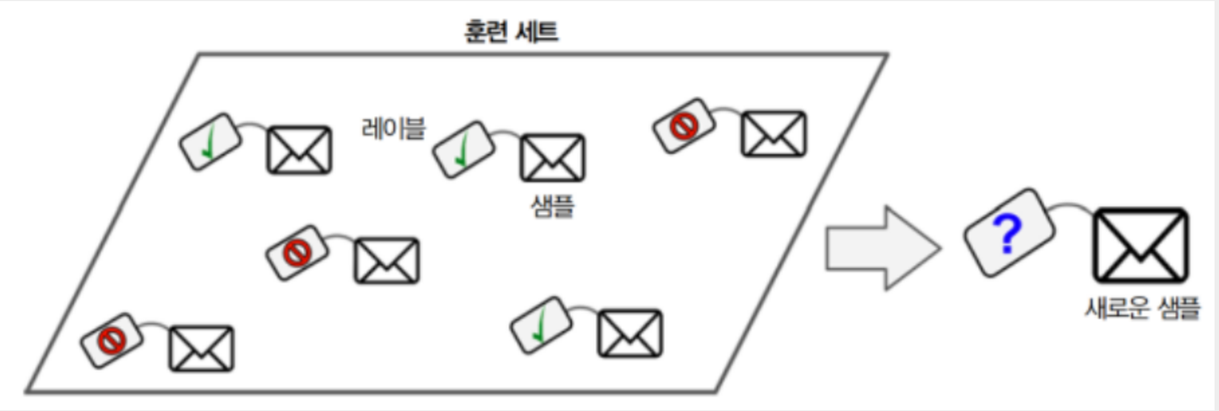
지도(감독)학습 _ Supervised Learning

지도학습으로 해결할 수 있는 문제

- **분류** Classification
- **예측** Forecast

지도(감독) 학습

정답값(label, target)이 있는 데이터셋을 통해 학습하는 것



지도(감독)학습에서의 훈련 세트, 테스트 세트

훈련 세트 모델을 훈련할 때 사용하는 데이터 (입력 + 정답(target, label))
훈련 세트가 클수록 좋다

테스트 세트 훈련 세트로 학습시킨 모델을 평가하는데 사용하는 데이터

※ 머신러닝 알고리즘의 성능을 제대로 평가하려면, 훈련 데이터와 평가에 사용하는 데이터가 달라야 한다.

훈련 세트, 테스트 세트로 나눌 때 유의할 점

1. 훈련 세트의 데이터가 테스트 세트보다 더 많아야 한다.

2. 훈련 세트와 테스트 세트에 샘플이 골고루 섞여 있어야 한다.

특정 종류의 샘플이 과도하게 많은 샘플링 편향을 가지고 있다면,
제대로 된 지도 학습 모델을 만들 수 없다.

※ 샘플링 편향 : 훈련 세트와 테스트 세트에 샘플이 골고루 섞여 있지 않아, 한쪽으로 치우친 상태

훈련 세트, 테스트 세트로 나누는 방법

1. numpy 배열의 인덱스 섞기

2. scikit-learn 패키지의 train_test_split() 함수 이용하기

train_test_split() : 훈련 데이터를 훈련 세트와 테스트 세트로 나누는 함수

test_size 매개변수 지정 : 기본값은 0.25

stratify 매개변수 : target 데이터를 매개변수로 지정시, 클래스 비율 맞게 훈련세트와 테스트 세트로 나눔

실습

using google Colab