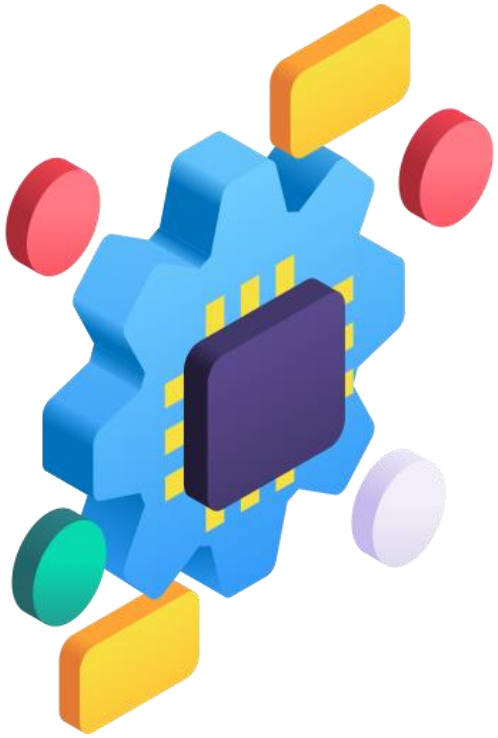
머신 러닝 Machine Learning

Classification

Clustering

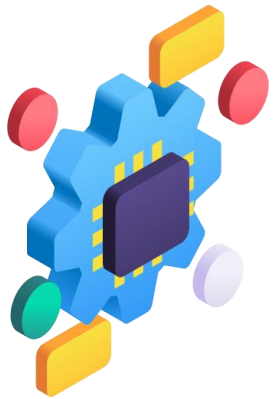


ensemble

Assessment

Regression

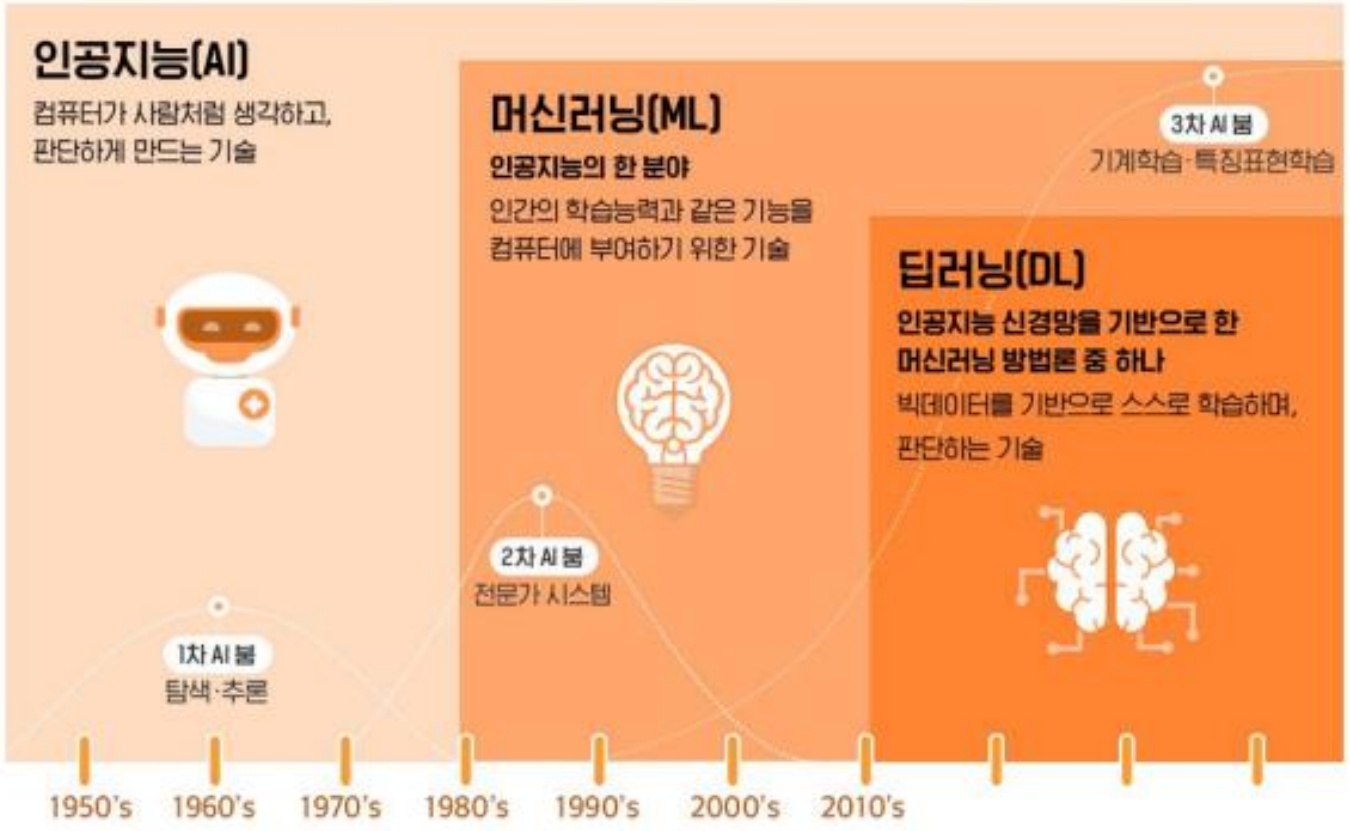
INDEX



머신 러닝이란?

딥러닝과 머신러닝의 관계

1959년, 아서 사무엘은 기계 학습을 "기계가 일일이 코드로 명시하지 않은 동작을 데이터로부터 학습하여 실행할 수 있도록 하는 알고리즘을 개발하는 연구 분야"라고 정의하였다. 기계 학습의 핵심은 표현(representation)과 일반화(generalization)에 있다. 표현이란 데이터의 평가이며, 일반화란 아직 알 수 없는 데이터에 대한 처리이다.



머신 러닝이란?

머신 러닝의 종류

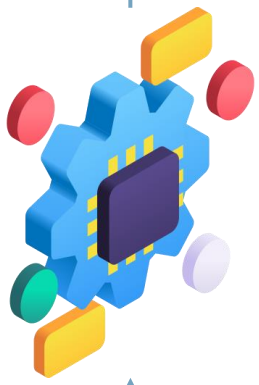
	지도학습 (Supervised Learning)	**비지도학습** (Unsupervised Learning)	**강화학습** (Reinforcement Learning)
훈련 방식	정답이 있는 데이터로 훈련	정답이 없는 데이터로 훈련	자신의 행동에 대한 보상을 받으며 목표를 달성하는 방향으로 학습
주요 알고리즘	분류, 회귀	군집화, 차원축소	로보틱스, 시뮬레이션
예시	• 강아지와 고양이 사진 분류하기(분류) • 집 값 예측하기(회귀)	• 라벨이 없는 데이터를 n개의 집단으로 구분(군집화) • 변수의 여러가지 특징을 보다 작은 수로 축소(차원 축소)	• 자율주행 차 • 알파고 등

머신 러닝의 핵심 개념

독립변수와 종속변수

라벨(Label), 종속변수

집값(Y)
?



라벨의 유무

지도학습
(Supervised Learning)

정답이 있는 데이터로 훈련

분류, 회귀

비지도학습
(Unsupervised Learning)

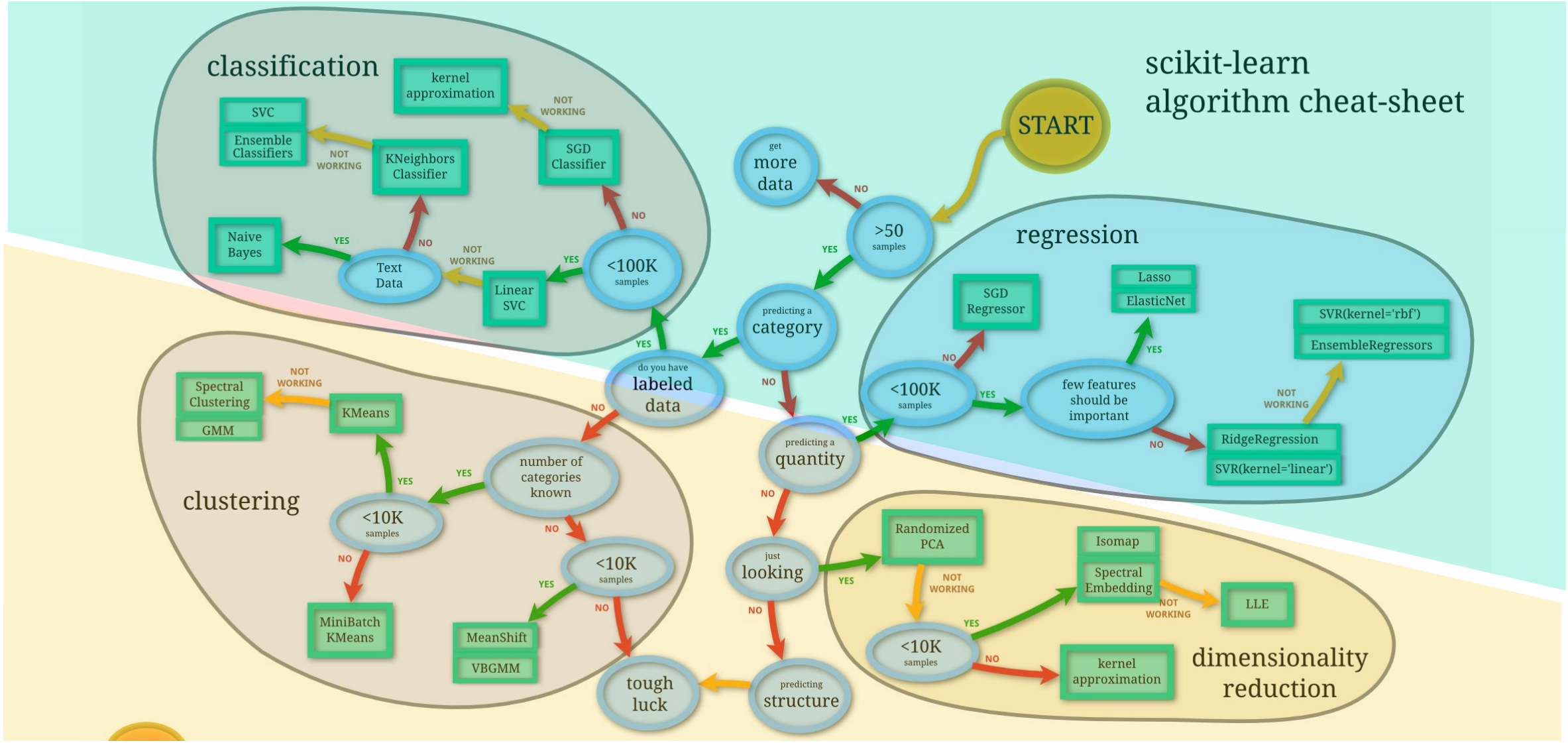
정답이 없는 데이터로 훈련

군집화, 차원축소

특징(feature), 독립변수

위치(X1)	크기(X2)
강남	32평

지도학습과 비지도학습



https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

지도학습의 특징

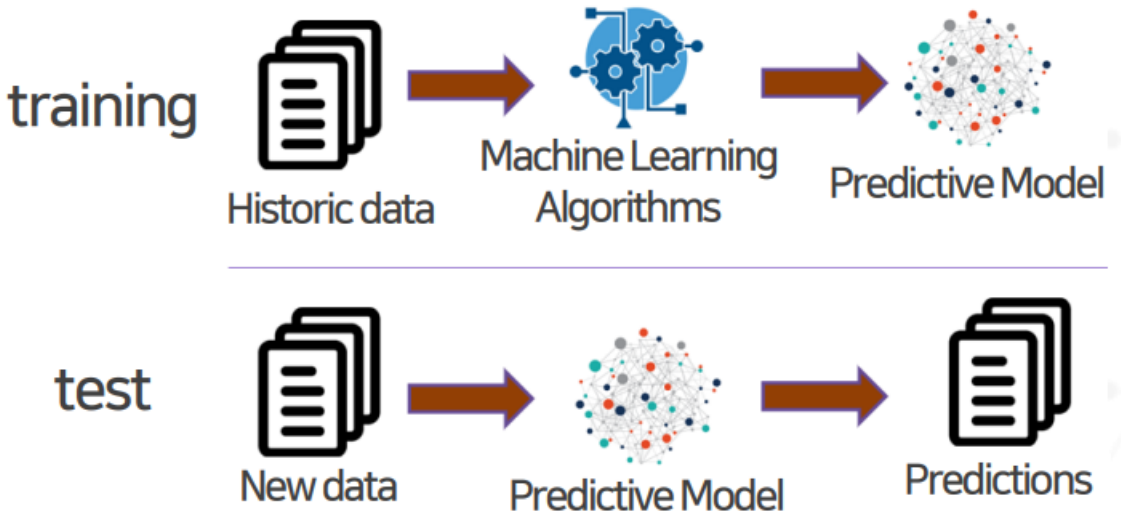
- Y(output)가 존재
  - 종속변수, 타겟(target), 라벨(label)
- X(input)가 존재
  - 독립변수, 특징(feature)
- Classification(분류) 문제에서 Y는 범주형(이산형) 값
  - 붓꽃의 종류, spam메일인지 아닌지(Yes/No) 여부 등
- N개의 Training Data로 학습

$$(x_1, y_2), \ldots, (x_N, y_N)$$

지도학습의 특징

N개의 training Data를 기반으로

- 본 적이 없는(학습데이터에 없었던) test data의 output을 예측(prediction)
- 어떤 input이 output에 어떻게 영향을 미쳤는지 이해하고 분석(inference)
- 모델을 평가하고, 다시 훈련하는 반복과정을 거쳐 성능을 향상시킴



https://singaporebusinessintelligence.blogspot.com/2018/10/what-is-automated-machine-learning.html

지도학습 - 종류

번호	알고리즘 명	분류	회귀
1	선형 회귀 (Linear Regression)	X	○
2	정규화 (Regularization)	X	○
3	로지스틱 회귀 (Logistic Regression)	○	X
4	서포트 벡터 머신 (Support Vector Machine)	○	○
5	나이브 베이즈 분류 (Naïve Bayes Classification)	○	X
6	랜덤 포레스트	○	○
7	K-최근접 이웃 (K-neareast neighborhood)	○	○
8	신경망**	○	○

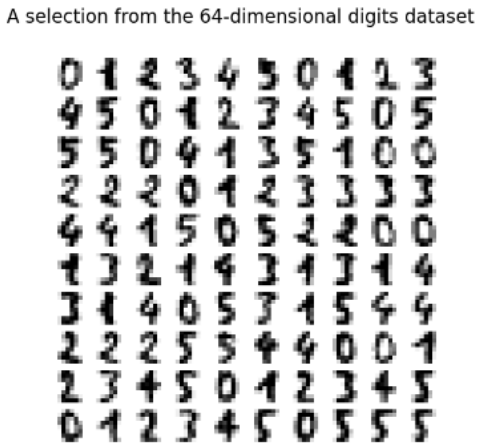
비지도학습의 특징

- Y(output)가 존재하지 않음
- X(input)만 존재
- 머신러닝의 목표가 지도학습에 비해 불명확함 - 차원 축소, 군집화
- 학습의 결과에 대해 평가하기 어려움
- 지도학습의 전처리(preprocessing) 과정으로 유용함

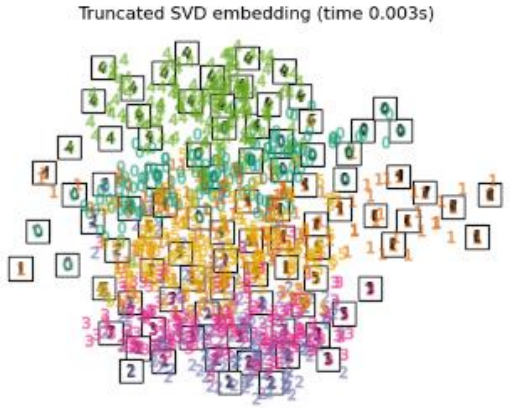
비지도학습의 특징

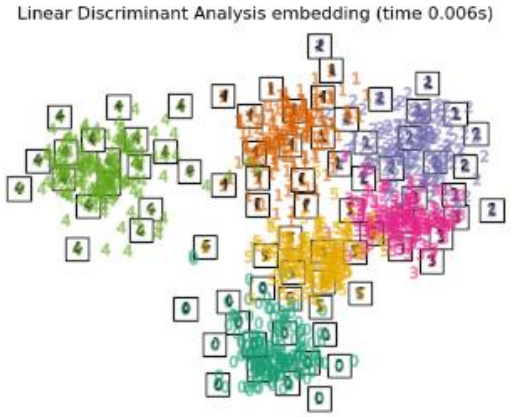
N개의 training Data를 기반으로

- 데이터의 특징을 활용하여 군집화/차원축소
- 분석가는 군집화/차원축소의 결과물을 활용하여 데이터를 분석하거나, 지도학습을 수행
- 목표와 일치하지 않는 군집화/차원축소인 경우, 다른 모델을 쓰거나 재군집화/차원축소



비지도
학습





비지도학습의 종류

번호	알고리즘 명	차원 축소	군집화
1	주성분 분석 (PCA)	○	X
2	잠재 의미 분석 (LSA)	○	X
3	음수 미포함 행렬 분해 (NMF)	○	X
4	잠재 디리클레 할당 (LDA)	○	X
5	K-평균 알고리즘 (K-means)	X	○
6	가우시안 혼합 모델	X	○
7	국소 선형 임베딩	○	X
8	T-분포 확률적 임베딩	○	X

https://github.com/jpub/MLBook

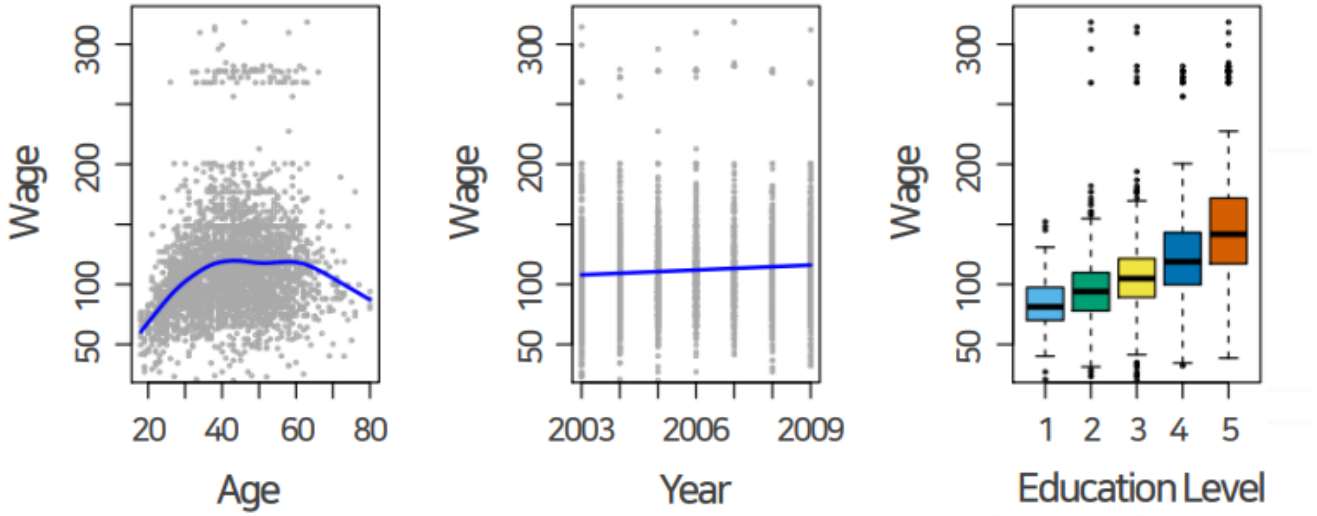
어떤 학습일까?

임금 예측

Data 미국 Central Atlantic 지역 남성의 임금 데이터

Input 나이, 연도, 교육수준

Label 임금



Q. 어떤 학습일까?

지도학습

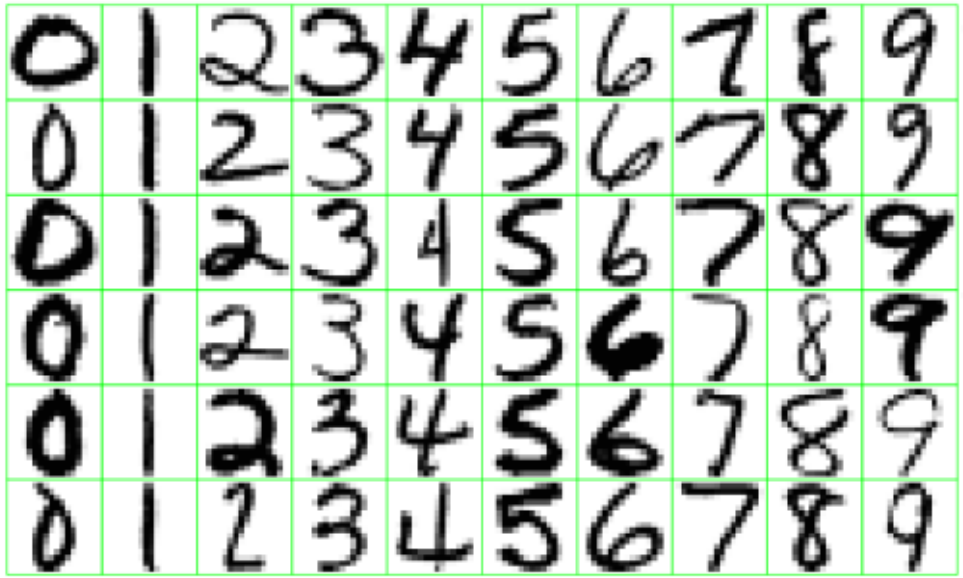
어떤 학습일까?

우편번호 인식

Data 우편물로부터 수집한 숫자 데이터

Input 손글씨 이미지

Label 0, 1, 2, 3, 4, 5, 6, 7, 8, 9



Q. 어떤 학습일까?

지도학습

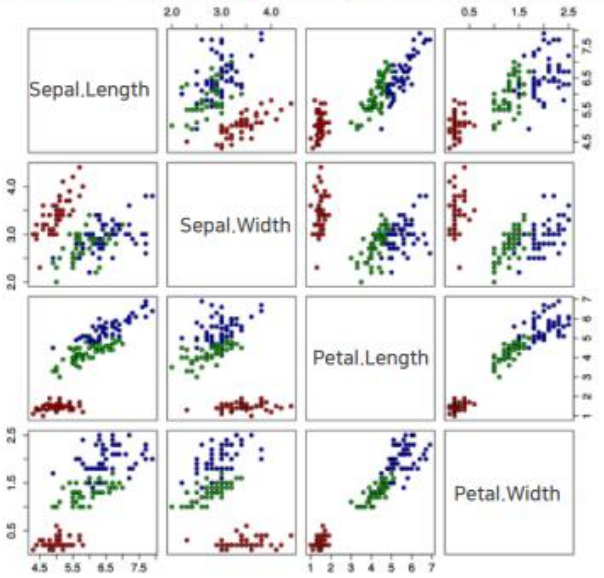
어떤 학습일까?

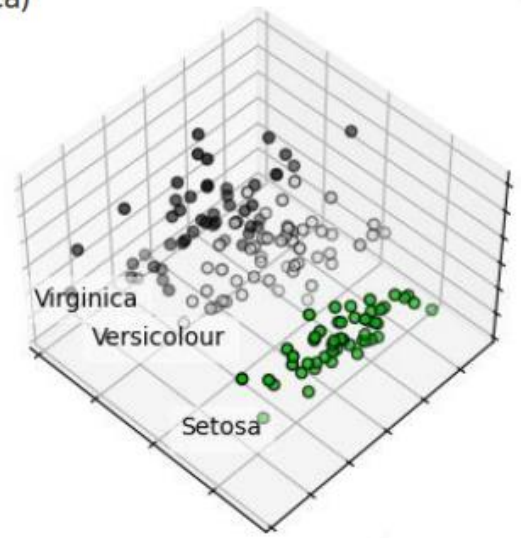
붓꽃 데이터의 차원 축소(dimensionality reduction)

Data 붓꽃의 꽃받침 길이와 너비, 꽃잎의 길이와 너비

PCA(Principal Component Analysis)를 적용하여 4 ➡ 3차원으로 축소

Iris Data(red=setosa, green=versicolor, blue=virginica)





[출처] https://en.wikipedia.org/wiki/Iris_flo

Q. 어떤 학습일까?

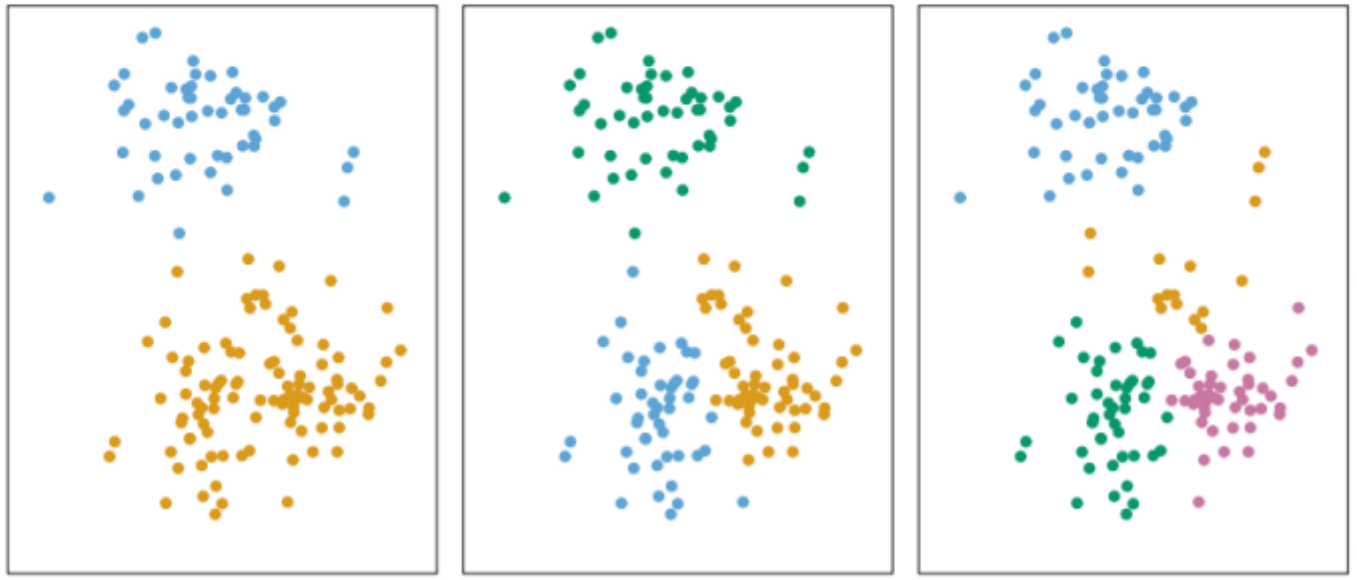
비지도학습

어떤 학습일까?

농구선수 군집화(clustering)

Data 게임당 득점, 리바운드, 도움, 가로채기

Clustering을 적용하여 2~4개의 군집(cluster)으로 구분



Q. 어떤 학습일까?

비지도학습

머신 러닝의 단계

1. 데이터 불러오기
2. 데이터 확인하기(통계적 특징, 데이터의 크기 등)
3. 데이터 전처리(결측치 및 이상값 정리, 스케일링 등)
4. train_test_split 및 x, y 데이터의 정의
5. 머신러닝 모델 정의 및 훈련, 검증

지도학습 - 분류

분류 문제는 Y가 이산적인 값

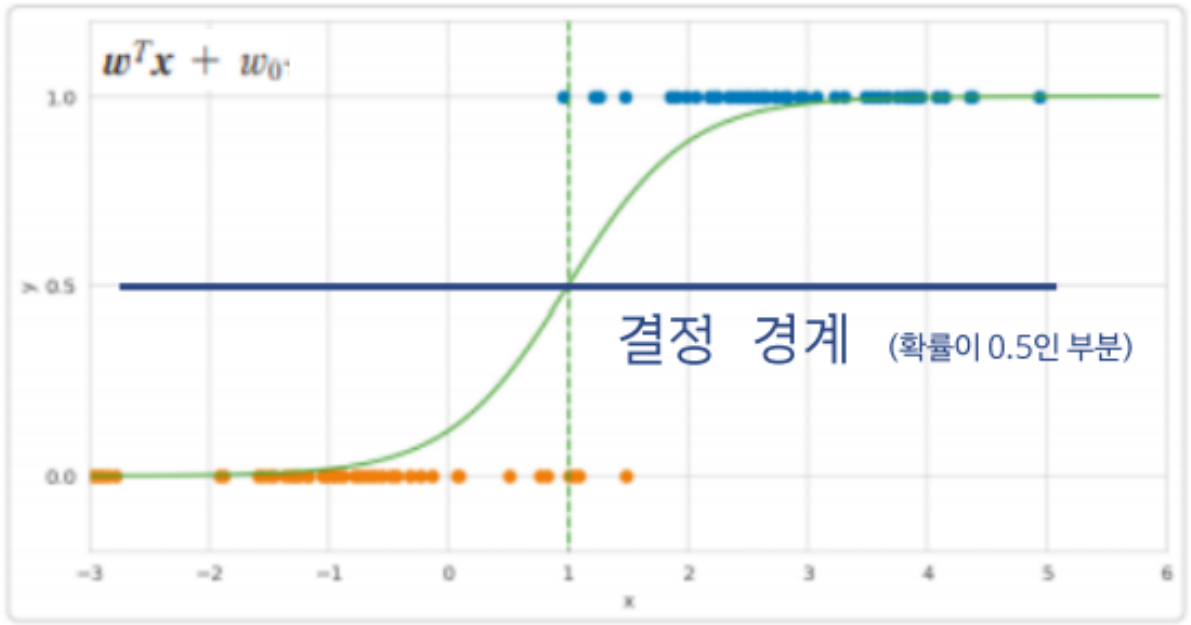
- Classifier(분류기)는 Training에 사용되지 않았던 새로운 데이터 X를 클래스 중 하나로 분류
- X가 각각의 클래스로 분류될 확률을 평가
- 각각의 input X=(X1, X2, …Xp)의 역할을 이해
- 즉, X와 Y의 관계성을 파악

- 1, 2, …k까지 K개의 클래스가 존재한다면, X가 클래스 K 일 확률을 다음과 같이 정의

$$p_k(x) = \Pr(Y = k|X = x), k = 1, 2, \ldots, K$$

지도학습 – 분류-로지스틱 회귀

번호	알고리즘 명	분류	회귀
3	로지스틱 회귀 (Logistic Regression)	○	X

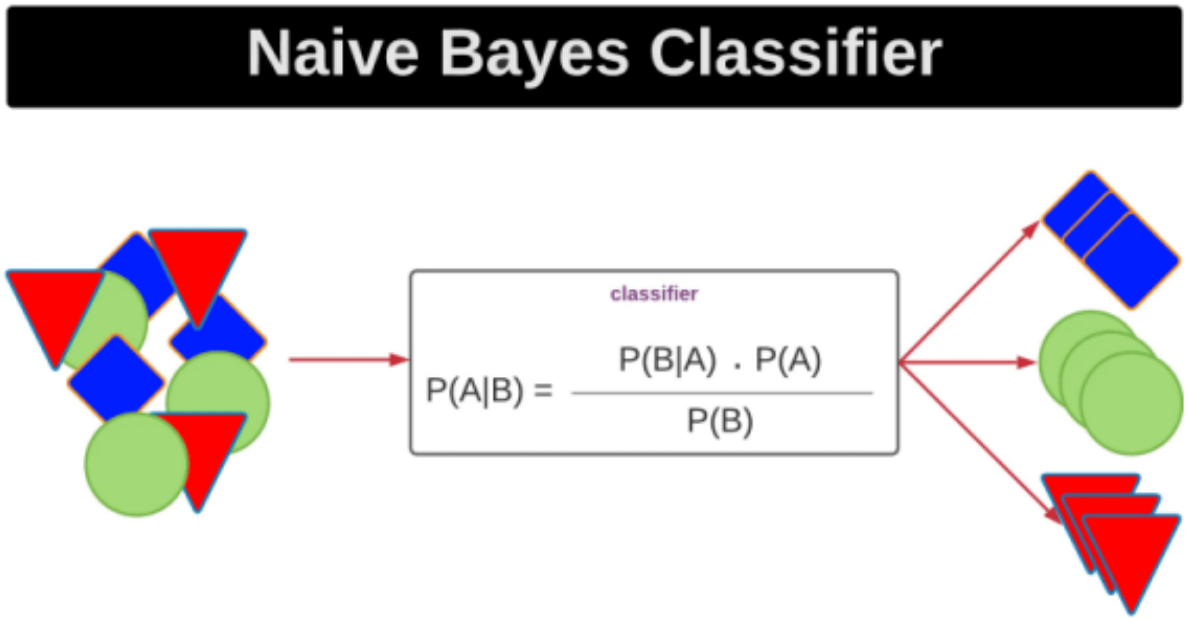


- 이름은 회귀지만… 분류입니다.
- 데이터가 집단에 속할 확률을 계산
- 사건 발생 vs 사건 미 발생 이진 분류
- 0에서 1사이의 확률로 결과를 나타냄
- 손실함수 : 로그손실
- 경사 하강법을 통해 최솟값을 찾음

그림 출처 : 머신러닝 도감, 아키바 신야 외, 2020 , 제이펍

지도학습 – 분류-나이브 베이즈 분류

번호	알고리즘 명	분류	회귀
5	나이브 베이즈 분류 (Naïve Bayes Classification)	○	X

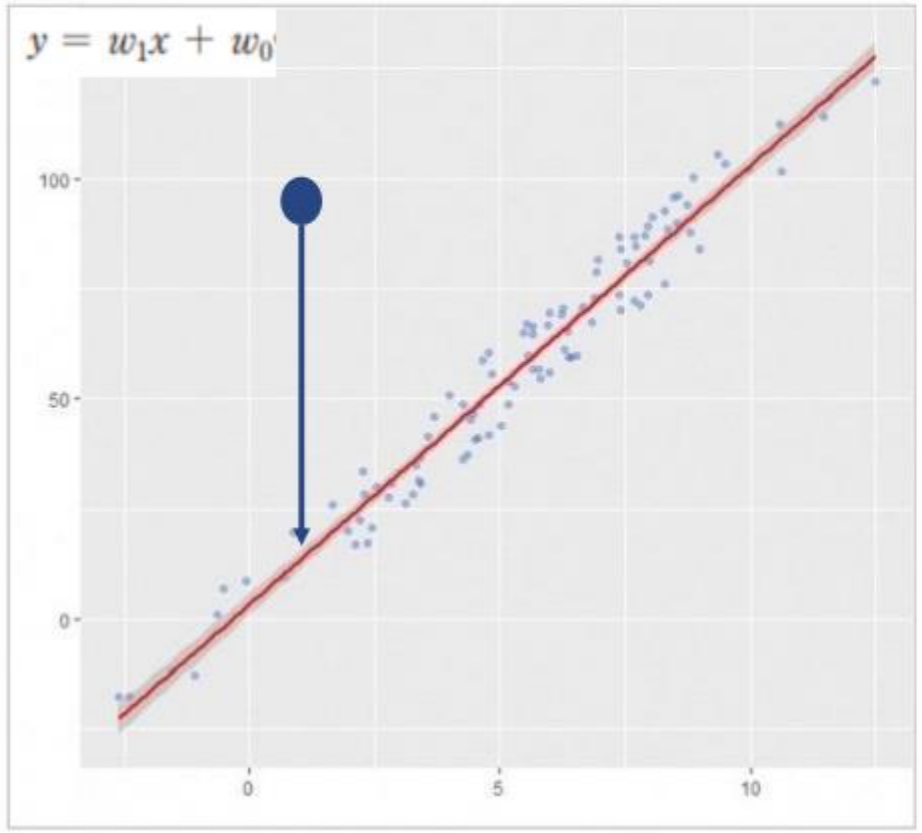


- 확률에 따른 결과 예측
- 주로 자연어 분류에 활용
- 데이터가 어떤 라벨에 속할지 확률 계산

그림 출처 : 머신러닝 도감, 아키바 신야 외, 2020 , 제이펍

지도학습 – 회귀 - 선형회귀

번호	알고리즘 명	분류	회귀
1	선형 회귀 (Linear Regression)	X	○



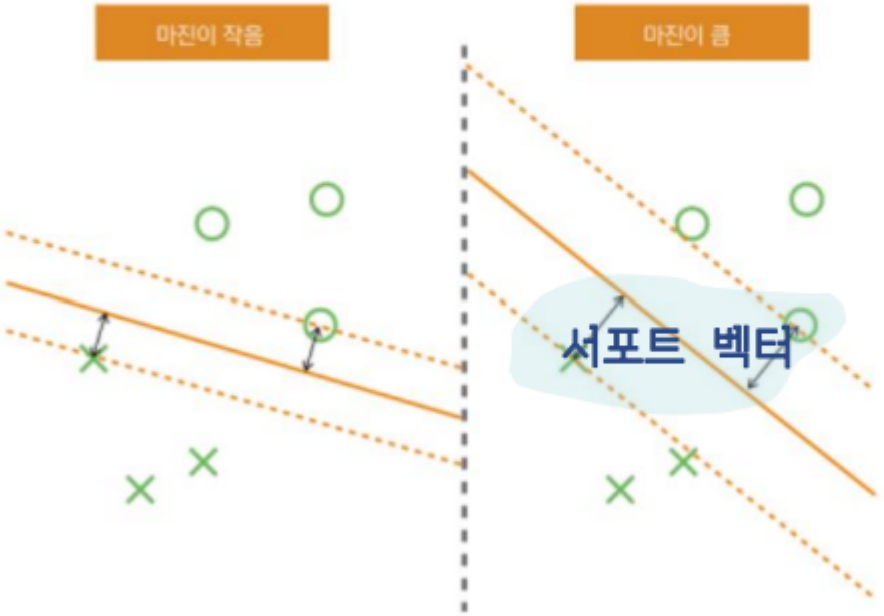
- 하나 이상의 독립변수를 사용하여 답을 찾음
- 하나일 때는 단순선형회귀 (simple LR)
- 여러 개는 다중선형회귀 (multiple LR)
- 독립변수 (x) 는 하나지만 n제곱 형태일 때 다항회귀
- loss : 평균제곱오차 => ((오차) 제곱) 평균

$$\frac{\sum_{i=1}^{n}\{y_i-(w_0+w_1x_i)\}^2}{n}$$

그림 출처 : 머신러닝 도감, 아키바 신야 외, 2020 , 제이펍

지도학습 – SVM(서포트 벡터 머신)

번호	알고리즘 명	분류	회귀
4	서포트 벡터 머신 (Support Vector Machine)	○	○

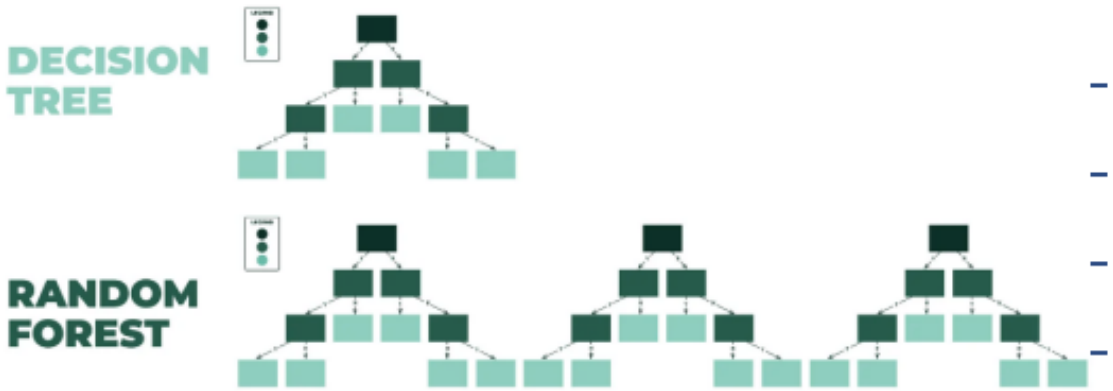


- 데이터에서 되도록 먼 결정 경계를 학습
- 집단 사이의 마진을 최대화 (더 명확한 분류)
- 소프트 마진 : 일부 데이터가 마진 안에 포함 되는 것을 허용
  > 그리드 탐색, 랜덤 탐색
- 서포트 벡터 : 마진 데이터, 마진 안 데이터

그림 출처 : 머신러닝 도감, 아키바 신야 외, 2020 , 제이펍

지도학습 – 랜덤 포레스트

번호	알고리즘 명	분류	회귀
6	랜덤 포레스트	○	○



- 여러 모델을 합해 높은 성능을 이끌어 냄
- 결정 트리 (1개) 의 의견을 다수결로 종합함
- 결정 트리는 조건에 따라 학습 데이터를 나눔
- 조건: 불순도가 작아지도록 함
- 지니 계수를 사용함

그림 출처 : 머신러닝 도감, 아키바 신야 외, 2020 , 제이펍

지도학습 – K-최근접 이웃

번호	알고리즘 명	분류	회귀
7	K-최근접 이웃 (K-neareast neighborhood)	○	○

그림 2-45 다수결로 알 수 없는 데이터 분류

- 입력 데이터의 주변 k개를 통해 입력 데이터를 분류
- 데이터 개수가 적거나 차원이 낮을 때 적절함
- 데이터 개수가 많으면 -> 시간/공간의 소요가 늘어남
- 차원이 많은 경우 -> 근접을 찾기 어려움

그림 출처 : 머신러닝 도감, 아키바 신야 외, 2020 , 제이펍