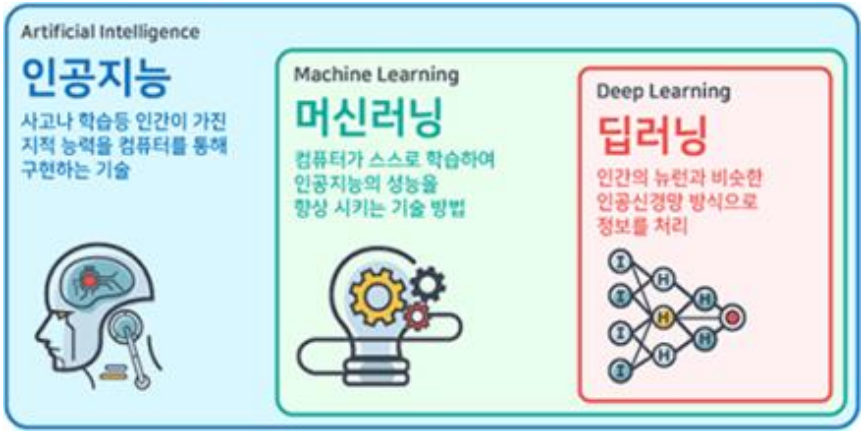
머신러닝 소개

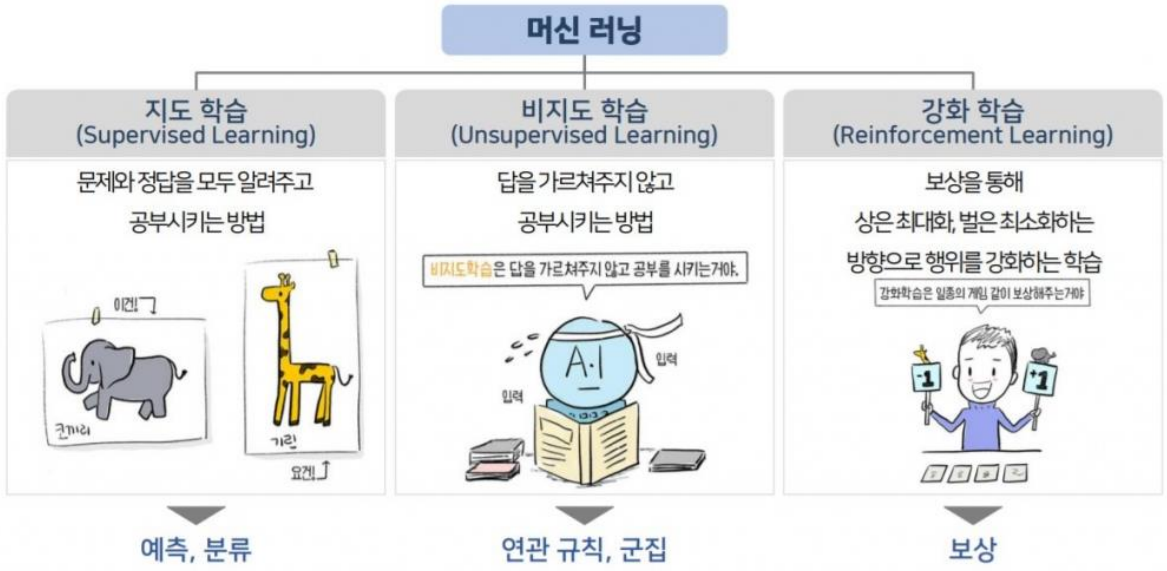
강사 이성구

2018년 10월 v1.0

- 인공지능(Artificial Intelligence)
  - 인간이 지닌 지적 능력의 일부 또는 전체, 혹은 그렇게 생각되는 능력을 인공적으로 구현하는 기술
- 머신러닝(Machine Learning)
  - 명시적인 프로그래밍 없이도 컴퓨터가 스스로 학습할 수 있도록 하는 기술
  - 기계(컴퓨터 알고리즘) 스스로 데이터를 학습하여 서로 다른 변수 간의 관계를 찾아 나가는 과정
  - 학습 = 표현(Representation) + 평가(Evaluation) + 최적화(Optimization)



- 머신러닝의 목적
  - 주어진 알고리즘을 이용해 입력 데이터를 처리한 후 최적의 결과를 도출하기 위한 모델을 구축하고, 이 모델을 이용하여 새로운 데이터에 대한 예측을 수행
- 학습 유형
  - 지도 학습
    Supervised Learning
  - 비지도 학습
    Unsupervised Learning
  - 강화 학습
    Reinforcement Learning



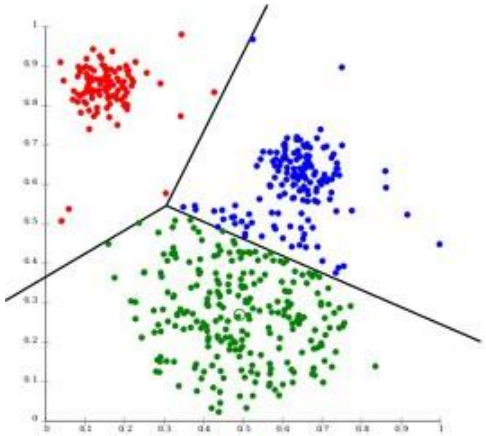
● 지도 학습(Supervised Learning)

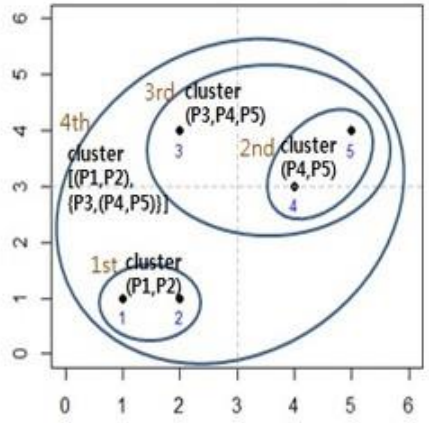
- 특징(features)이 이미 정해진 데이터를 사용하여 학습하는 방법. 이 때 각 데이터에 정해진 특징은 레이블(label)이라고도 표현하며, 레이블이 있는 데이터들의 집합을 트레이닝 세트(Training Set)라고 한다.

X_1	X_2	X_3	Y
3	6	9	3
2	5	7	2
2	3	5	1

- 주어진 트레이닝 세트를 학습하면 데이터를 기반으로 하는 **모델**이 생성되고, 이 모델을 사용하여 어떠한 특징을 갖는 데이터가 어떤 레이블에 속할지 **예측**할 수 있다.

- 비지도 학습(Unsupervised Learning)
  - 학습에 사용하는 데이터에 특징(레이블)이 부여되어 있지 않다.
  - 지도 학습이 기존에 있는 데이터를 기반으로 새로운 데이터에 대한 특징을 추론하는 것을 목표로 한다면, 비지도 학습은 주어진 데이터들이 어떻게 구성되어있는지를 분석하는 군집분석을 목표로 한다.
  - 비지도 학습의 대표적인 사례
    - ✓ 구글 뉴스 서비스: 비슷한 주제의 뉴스끼리 묶어서 보여주는 기능 제공
    - ✓ 단어 클러스터링: 유사한 단어끼리의 묶음 구성





- 학습 유형별 모델 구분
  - 회귀(Regression): 주가 환율 등 경제지표 예측
  - 분류(Classification): 은행에서 고객을 분류하여 대출 승인•거부 결정
  - 군집(Clustering): 비슷한 소비패턴의 고객을 분류하여 군집 구성

	지도학습	비지도학습
분석 모형	● 회귀 분석 ● 분류	● 군집 분석
특징	✓ 제공된 정답을 이용한 학습 ✓ 다양한 모델 평가 방법	✓ 정답 없이 비슷한 데이터를 찾아 그룹화 ✓ 제한적 모델 평가 방법

● 회귀(Regression)

- 회귀 분석은 어떠한 변수에 영향을 받는 결과가 연속적인 경우에 사용.
  - ✓ 시험 공부에 투자한 시간(변수)에 따라 예상되는 기말고사 점수(0~100 사이의 연속적인 값) 추측
  - ✓ 시험 공부에 투자한 시간과 실제로 획득한 성적을 담고 있는 트레이닝 세트
  - ✓ 회귀분석 모델 기반, 시험 공부에 7시간을 투자한 학생의 예상 점수를 대략 75점 정도로 예측

X (TIME SPENT FOR EXAM)	Y (SCORE)
10	90
9	80
3	50
2	30

- 분류(Classification)
  - 어떠한 변수에 영향을 받는 결과를 연속적이지 않은 값들로 나눌 때 사용
  - 시험 공부에 투자한 시간(변수)에 따라 예상되는 합격 여부(Pass or Fail) 혹은 학점(A, B, C, D, F) 추측
  - 데이터를 합격(P) 혹은 불합격(F) 두 가지로 나뉘고 이러한 데이터 구분은 Binary Classification 이라 한다.

X (TIME SPENT FOR EXAM)	Y (PASS/FAIL)
10	P
9	P
3	F
2	F

  - 주어진 데이터를 두 개 이상으로 분류하였으므로 이는 Multi-label classification이라 할 수 있습니다.

X (TIME SPENT FOR EXAM)	Y (GRADE)
10	A
9	B
3	D
2	F

● 머신러닝 프로세스

데이터 정리

↳ 데이터 분리(훈련•검증)

↳ 알고리즘 준비

↳ 모델 학습(훈련 데이터)

↳ 예측(검증 데이터)

↳ 모델 평가

↳ 모델 활용

1) 데이터 정리

- 머신러닝 데이터 분석을 시작하기 전에 알고리즘이 이해할 수 있는 형태로 데이터 변환 작업이 선행되어야 한다.

2) 데이터 분리(훈련 데이터 & 검증 데이터)

- 분석 대상에 관해 수집한 관측 값(observation)을 속성 값(feature 또는 variable)을 기준으로 정리
- numpy 또는 pandas 등의 도구를 활용하여 구조적 데이터 형식으로 변환
- 배열 또는 데이터프레임의 열은 속성을 나타내는 변수들이 위치하고 행은 하나의 관측 값을 나타내며 관측 값의 개수만큼 행의 수는 증가한다.

3) 모형 학습(훈련 데이터 이용)

- 알고리즘이 이해할 수 있도록 데이터프레임으로 변환한 다음에는 여러 속성(변수) 간의 관계를 분석하여 결과를 예측하는 모형을 학습을 통해 찾는다.
- 모형 학습에 사용하는 데이터를 훈련 데이터(train data)라고 한다.

4) 예측(검증 데이터)

- 학습을 마친 모형의 예측 능력을 평가하기 위한 데이터를 검증 데이터(test data)라고 한다.

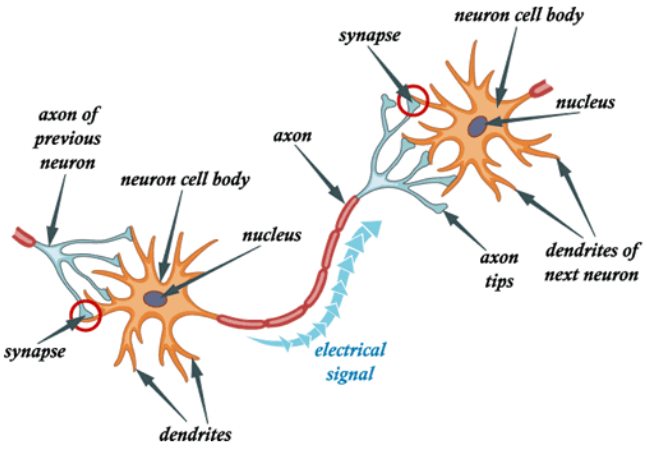
5) 모형 평가

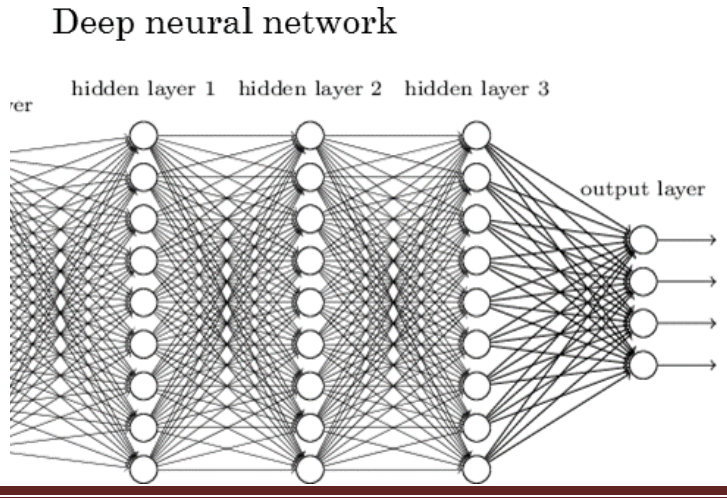
- 검증 과정을 통해 학습을 마친 모형의 예측 능력을 평가한다.

6) 모형 활용

- 평가 결과를 바탕으로 최종 모형을 확정하여 문제 해결에 적용한다.

- 딥러닝(Deep Learning)
  - 머신러닝 알고리즘 중 하나인 인공 신경망(Neural Network)을 기반으로 하는 머신러닝의 한 부류
  - 인공 신경망은 인간의 외부 자극에 대한 뉴런과 뉴런 사이의 연결 구조에서 영감을 받아 구현
  - 인간의 뇌가 동작하는 방식을 모방하였으나 실제 동작 방식은 인간의 뇌와 많은 차이가 있으며 범용적인 분야가 아닌 특정 분야에 한정하여 사용





- 딥러닝은 기존의 머신러닝이 처리하기 어려운 데이터를 더 잘 처리
  - ✓ 머신러닝이 잘 처리하는 데이터: 데이터베이스, CSV, 엑셀 등에 저장된 정형 데이터
  - ✓ 딥러닝이 잘 처리하는 데이터: 이미지•영상, 음성•소리, 텍스트•번역 등의 비정형 데이터