01_머신러닝 기초

학습: 데이터 입력과정
예측: 학습 후 예측

< 학습내용>
1. 학습전 : 학습방법 결정/ 데이터 전처리
2. 학습후 : 성능향상(재학습, 반복학습 등)
3. 예측후 : 테스트/ 배포

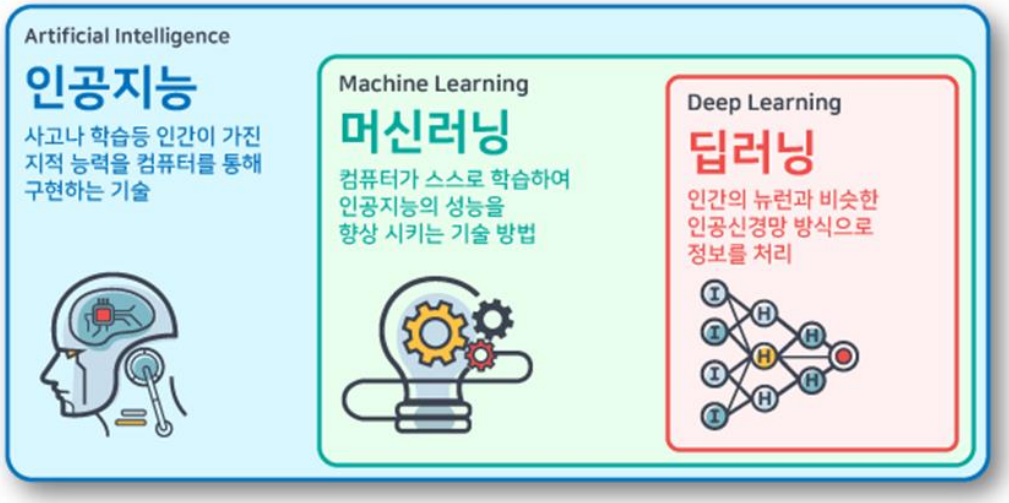
머신러닝 기초

- AI

> **인공지능**
  - 인공지능은 인간의 지능이 갖고 있는 기능을 갖춘 컴퓨터 시스템이며, 인간의 지능을 기계 등에 인공적으로 구현한 가장 큰 범주이다.

> **머신러닝** 통계알고리즘에 기반함
  - 기계가 학습할 수 있도록 하는 알고리즘과 기술을 개발하는 분야

> **딥러닝**
  - 인간의 뉴런구조와 비슷한 인공신경망 알고리즘을 사용하여 학습하는 분야

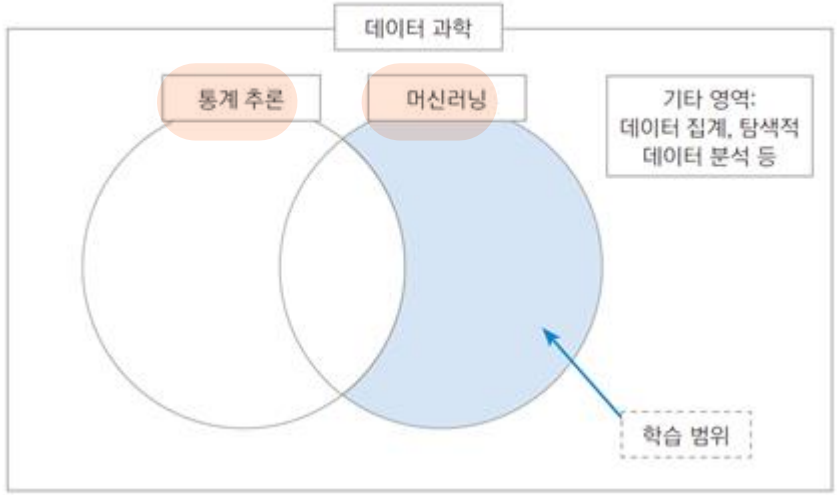


• 머신러닝이란?

> 인과 관계(causal relation) 혹은 상관관계(correlation)를 모델링하는 기법을 사용하여 입력값(독립변수)과 출력값(종속변수)이 주어졌을 때의 알고리즘을 추정 통계 기반

> 해결하려는 문제에 따라 예측(prediction), 분류(classification), 군집(clustering) 알고리즘 등으로 나눠진다.

> 데이터 과학이라는 큰 영역에서 머신러닝은 통계 추론과 함께 큰 두 줄기를 구성함



빅데이터분석은 머신런닝

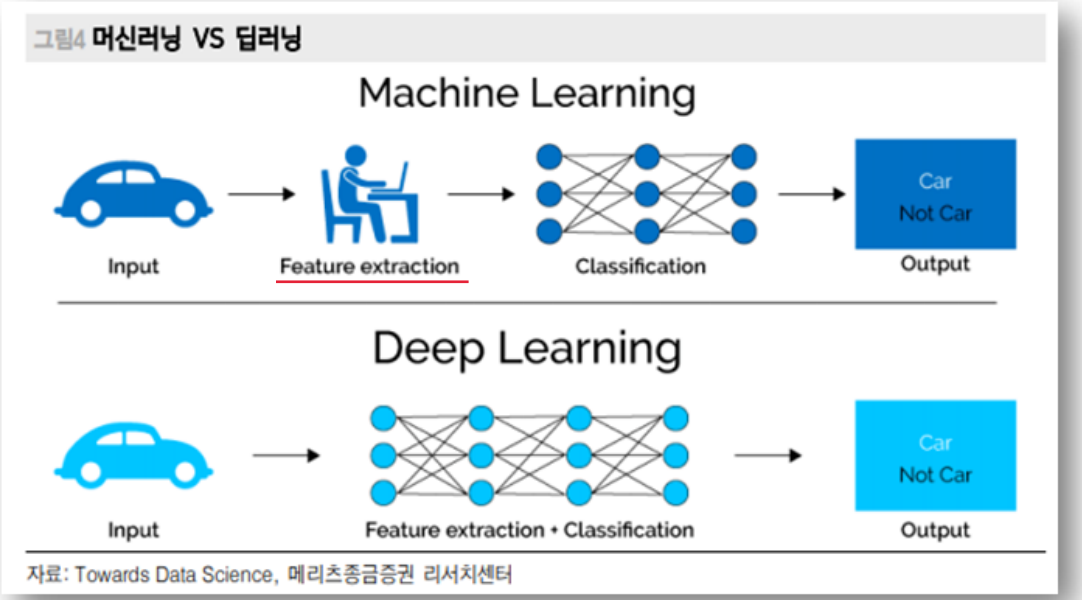
• 머신러닝 vs 딥러닝

> **머신러닝**
  - 데이터를 분석하고, 데이터로부터 학습한 다음, 학습한 것을 적용해 정보에 입각한 결정을 내리는 알고리즘

==> 결과 예측

> **딥러닝**
  - 머신러닝의 하위 분야로 인공 신경망을 계층으로 구성하여 자체적으로 학습하여 결정을 내리는 알고리즘

그림4 **머신러닝 VS 딥러닝**



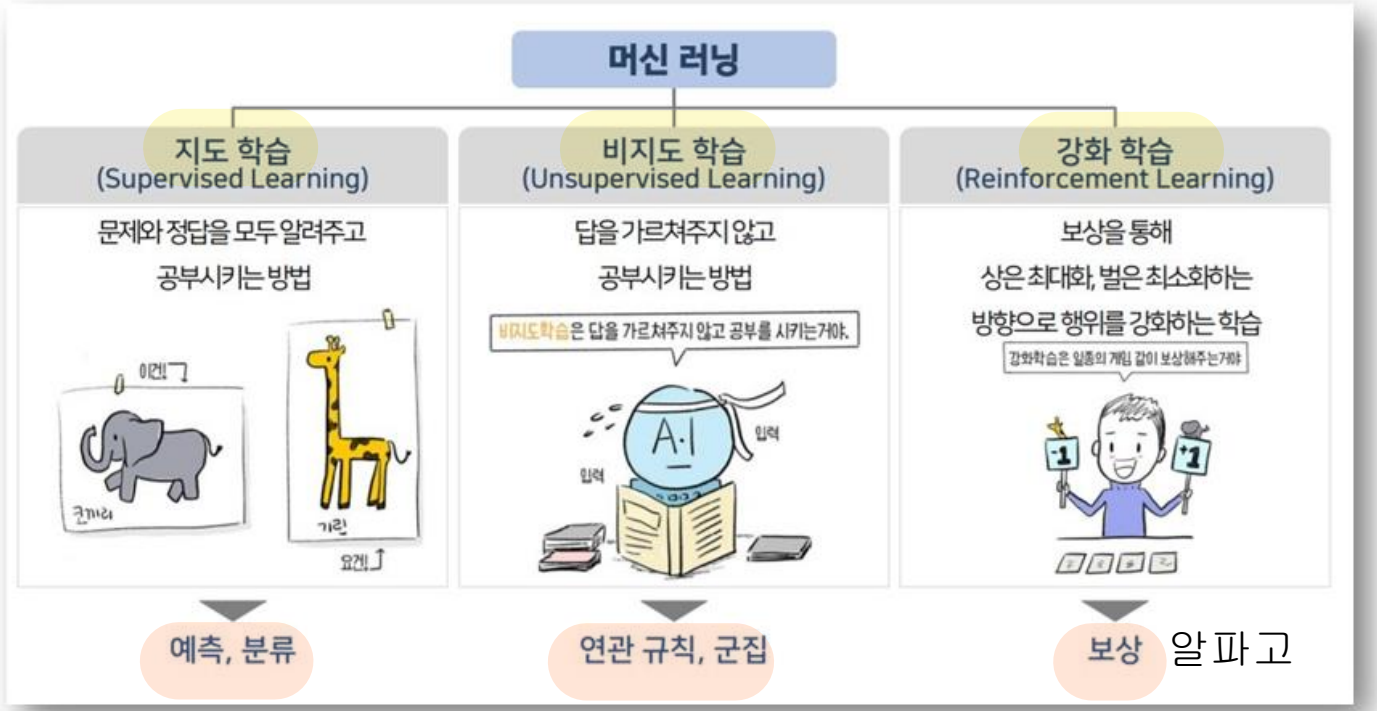
자료: Towards Data Science, 메리츠종금증권 리서치센터

통계

수학

• 머신러닝의 종류

> 머신러닝은 크게 3가지로 나누어진다.



• 머신러닝의 종류

> 지도 학습

종속변수와 독립변수 모두 학습

- 사례들을 기반으로 예측을 수행한다.
- 입력 피처(feature)와 입력 피처에 해당하는 목표 변수의 쌍이 주어졌을 때 이 관계를 모델링하여 입력으로 들어온 새로운 피처로 목표 변수를 예측함.
- 분류, 회귀 등

독립변수

종속변수

종속변수 – 수치형: 회귀(예측)
–범주형: 분류

> 비지도 학습

독립변수만 학습

- 데이터만을 학습하여 패턴을 찾아낸다.
- 피처 공간만 주어진 상태에서 이를 학습하여 피처 공간의 분포를 모델링하고 인사이트를 도출함.
- 군집화, 차원 축소, 이상값 탐지 등 지도학습전 데이터 군집화, 차원축소 등 _> 지도학습 도움

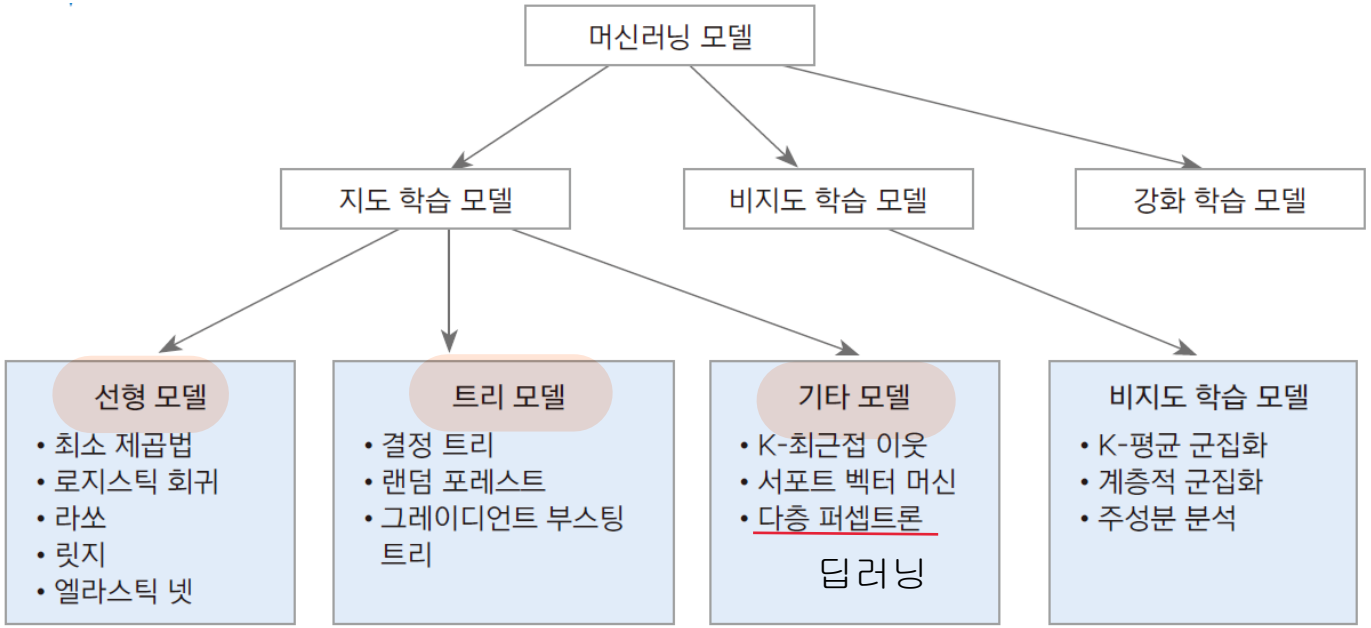
> 강화학습

- 피드백을 기반으로 행동을 분석하고 최적화한다.
- 시행 착오와 지연 보상을 통해 최적의 방법을 스스로 찾아간다.

• 머신러닝의 알고리즘 종류

학습: 데이터 입력과정
예측: 학습 후 예측

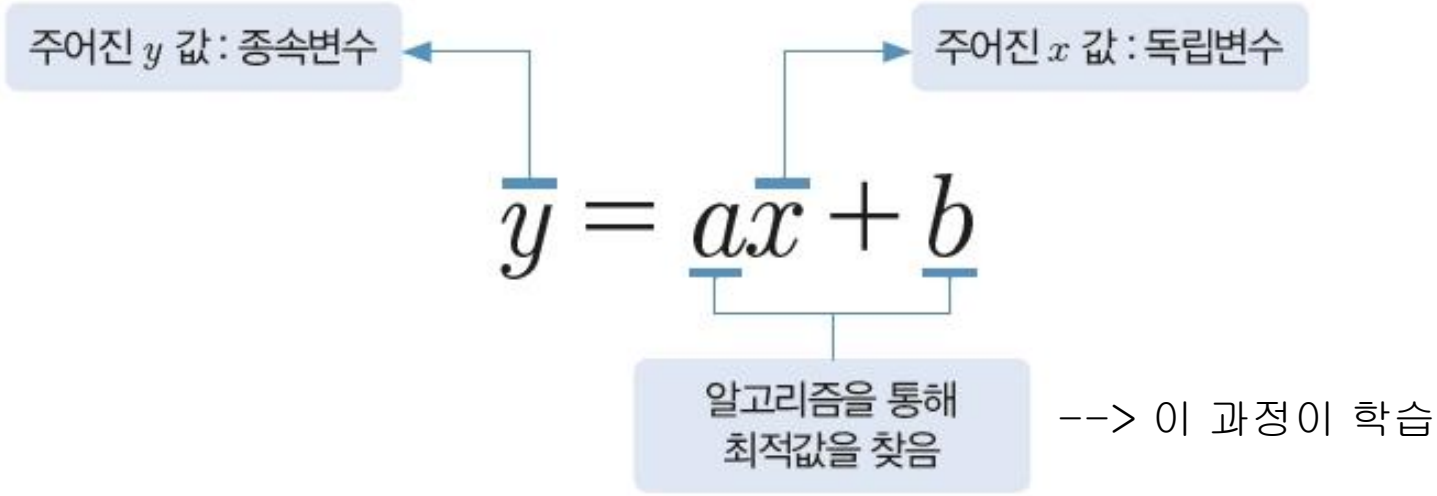
1. 학습전 : 학습방법 결정/ 데이터 전처리
2. 학습후 : 성능향상(재학습, 반복학습 등)
3. 예측후 : 테스트/ 배포



로지스틱 회귀: 회귀분석이 아니라 분류를 함

• 데이터의 이해

> **피쳐의 개념**
  - 피쳐(feature) : 특성이나 특징이라는 의미
  - 모델을 구성하는 데 데이터가 가장 큰 영향을 줌

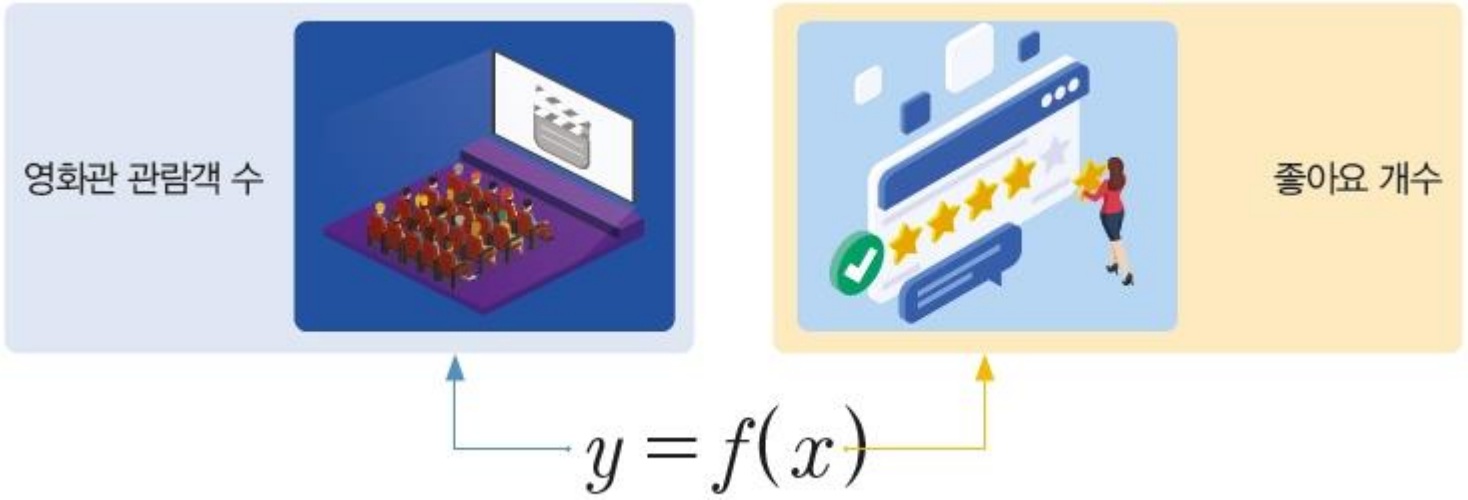


<기본적인 선형회귀식>

• 데이터의 이해

> 피쳐의 개념

- 영화관 관람객 수(y)를 인터넷 웹사이트의 좋아요 개수(x)로 예측할 수 있다고 가정하면, 아래와 같은 간단한 식으로 표현할 수 있다.



• 데이터의 이해

> 보스턴 집값 예측(Boston House Price) 데이터 셋

x 변수 13개	[01] CRIM	자치시(town)별 1인당 범죄율
	[02] ZN	25,000 평방피트를 초과하는 거주지역의 비율
	[03] INDUS	비소매상업지역이 점유하고 있는 토지의 비율
	[04] CHAS	찰스강에 대한 더미변수(강의 경계에 위치한 경우는 1, 아니면 0)
	[05] NOX	10ppm 당 농축 일산화질소
	[06] RM	주택 1가구당 평균 방의 개수
	[07] AGE	1940년 이전에 건축된 소유주택의 비율
	[08] DIS	5개의 보스턴 직업센터까지의 접근성 지수
	[09] RAD	방사형 도로까지의 접근성 지수
	[10] TAX	10,000달러 당 재산세율
	[11] PTRATIO	자치시(town)별 학생/교사 비율
	[12] B	1000(Bk−0.63)^2, 여기서 Bk는 자치시별 흑인의 비율을 말함
	[13] LSTAT	모집단의 하위계층의 비율(%)
y 변수	[14] MEDV	본인 소유의 주택가격(중앙값) (단위 : $1,000)

• 데이터의 이해

> **보스턴 집값 예측(Boston House Price) 데이터 셋**

- 범죄율, 방의 개수, 재산세율 등의 값(독립변수)들이 어떻게 집값(종속변수)에 영향을 주는지에 대한 모델
- 서로 다른 독립변수 13개를 x로 하고 가중치를 선형 결합하여 나타냄

딥러닝의 기본식: $y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \cdots + \beta_{13} x_{13} + \beta_0 \times 1$

잔차 1개

> **머신러닝에서 독립변수 x를 피쳐라고 한다.**

- 종속변수에 영향을 주는 특성, 데이터가 가지고 있는 특성