디지털미디어랩 머신러닝 여름캠프 3주차

(1) 기본적인 머신러닝의 용어와 개념

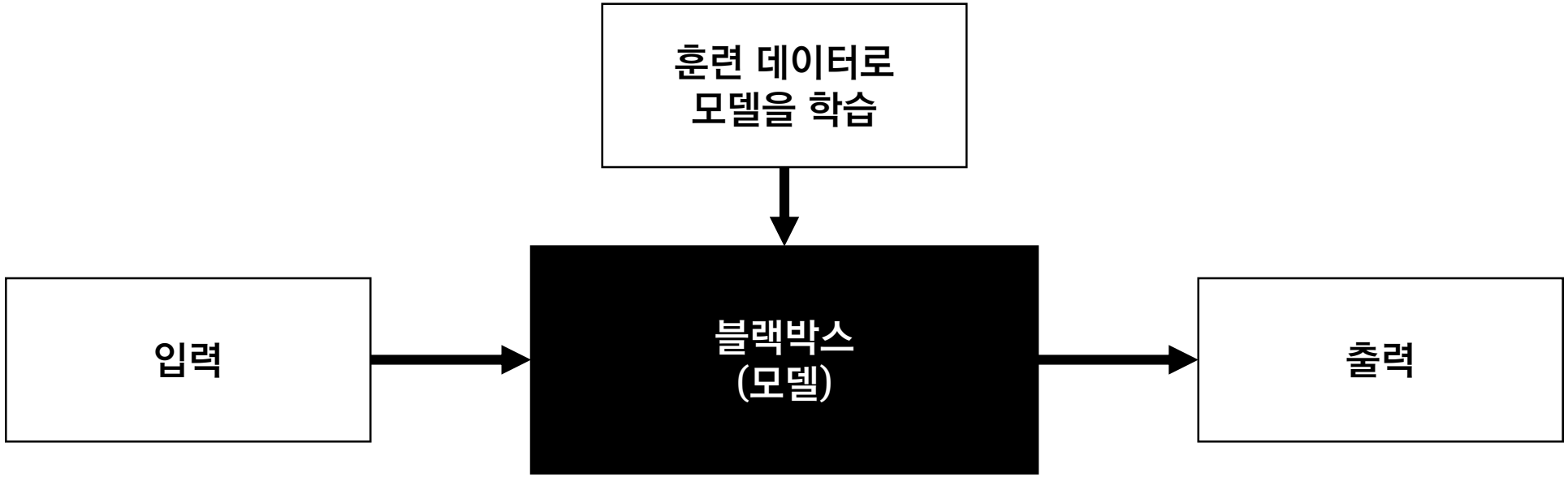
YONSEI
UNIVERSITY
DIGITAL
MEDIA LAB.

목차

머신러닝의 등장 배경

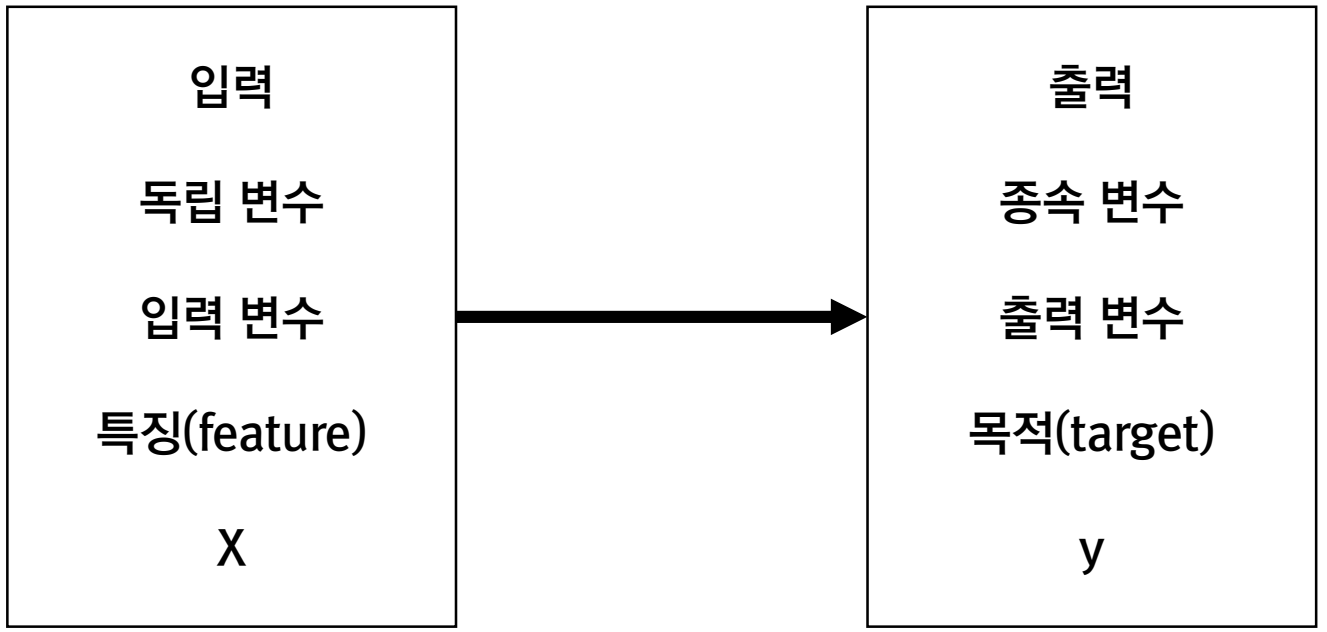
- 기존의 명시적 프로그래밍 방식의 한계
- 컴퓨터가 어떤 규칙을 이해하기 위해서는 프로그래머가 일일이 규칙을 명시적으로 정해줘야 했다.
- 그러나 스팸 메일을 필터링하거나, 자동차가 자율주행하는 등의 문제는 규칙이 너무나 복잡하기 때문에 프로그래머가 모든 규칙을 알려줄 수 없었다.
- 이에 1959년에 Arthur Samuel은 “명시적으로 프로그래밍하지 않고, 컴퓨터가 학습 능력을 갖게 하도록 연구하는 분야”를 머신러닝(ML)이라고 부르기 시작한다.

머신러닝의 기본적인 구조



- 훈련 데이터로 모델을 학습 시킨다.
- 학습된 모델에 입력 데이터를 넣으면 출력 데이터를 만들어 준다.

머신러닝의 기본적인 구조



- 간단하게 X로 y를 알아내는 문제로 이해할 수 있다.

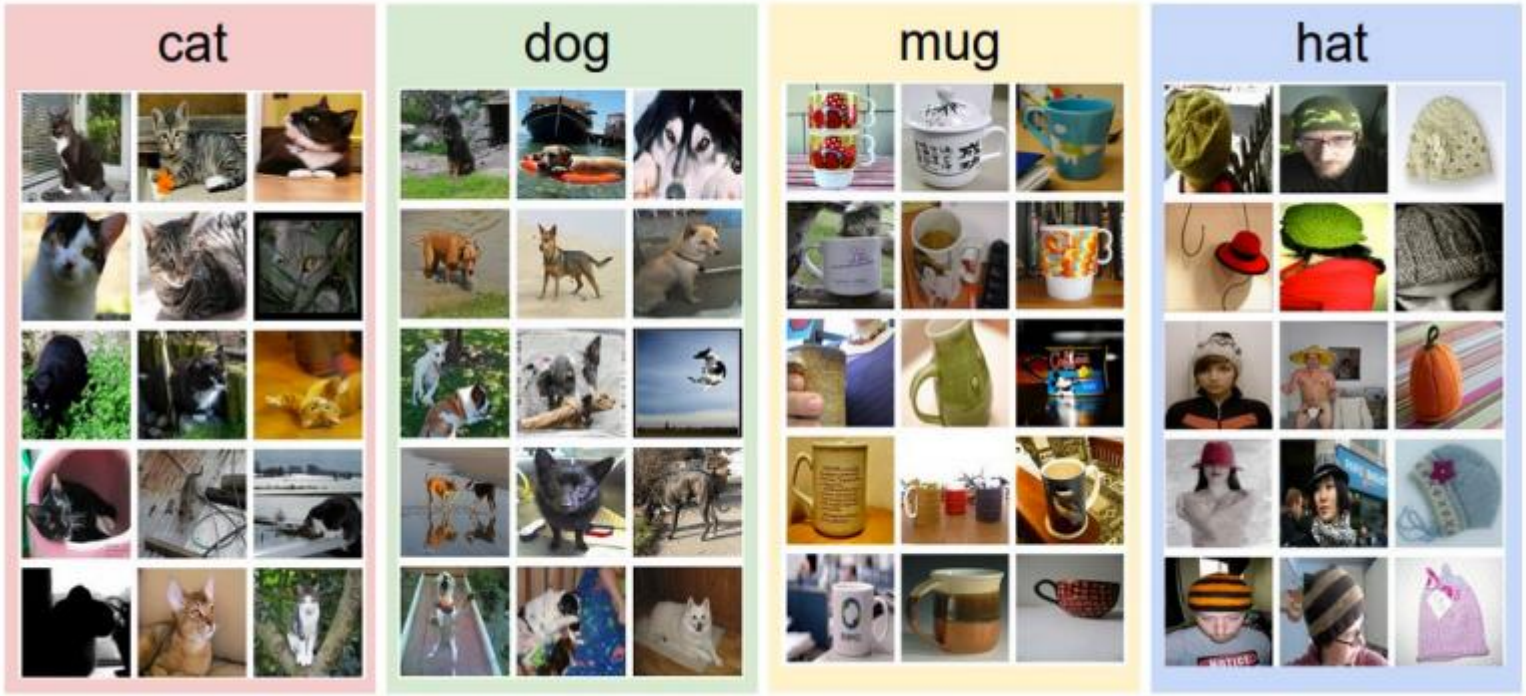
지도학습/비지도학습

- 머신러닝은 크게 두 종류의 학습이 있다.

- 지도 학습 (Supervised Learning)
  : 레이블링된 데이터(Labeled data)를 가지고 학습하는 것

- 비지도 학습 (Unsupervised Learning)
  : 레이블링되지 않은 데이터로 학습을 하는 것

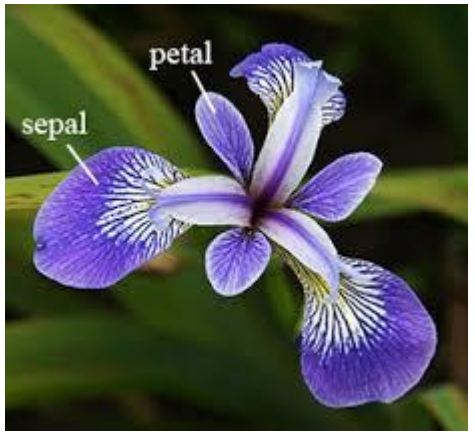
지도학습



- 각 이미지 데이터가 무엇을 의미하는지 알려준다.
- X : 이미지 데이터
- y : 이미지 종류 (고양이, 강아지 등)

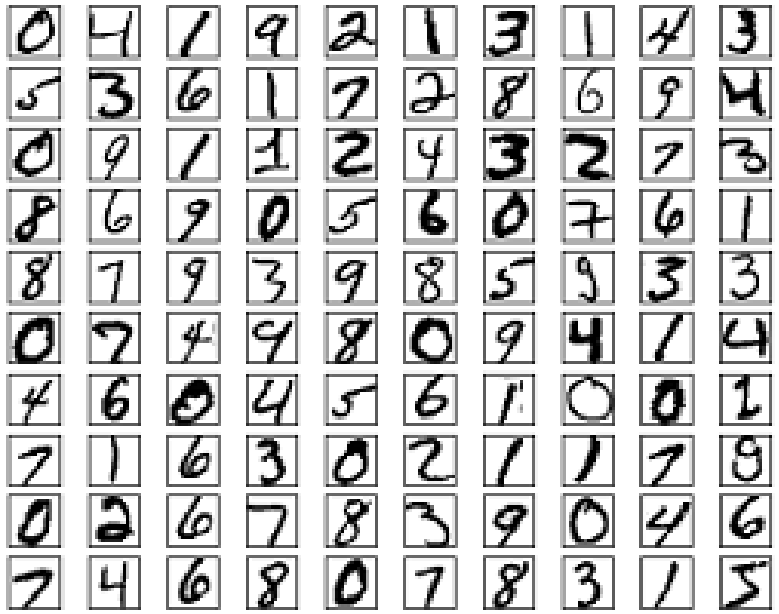
지도학습 (Iris data 분류)

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa
11	5.4	3.7	1.5	0.2	setosa
12	4.8	3.4	1.6	0.2	setosa
13	4.8	3.0	1.4	0.1	setosa
14	4.3	3.0	1.1	0.1	setosa
15	5.8	4.0	1.2	0.2	setosa



- 각 데이터가 어떤 꽃의 종류인지 알려준다.
- X : sepal length, sepal width, petal length, petal width
- y : 꽃의 종류 (Setosa, Virginica, Versicolor)

지도학습 (MNIST 데이터 분류)



- 각 이미지 데이터가 어떤 숫자를 의미하는 지 알려준다.
- X : 이미지 데이터
- y : 숫자

지도학습 (회귀)

광고비	매출
2800	4000
3000	3500
500	1000
1000	1200
1500	1750
650	650
7000	11000

- 매출도 레이블된 데이터로 볼 수 있음
- X : 광고비
- y : 매출

지도학습 (회귀)

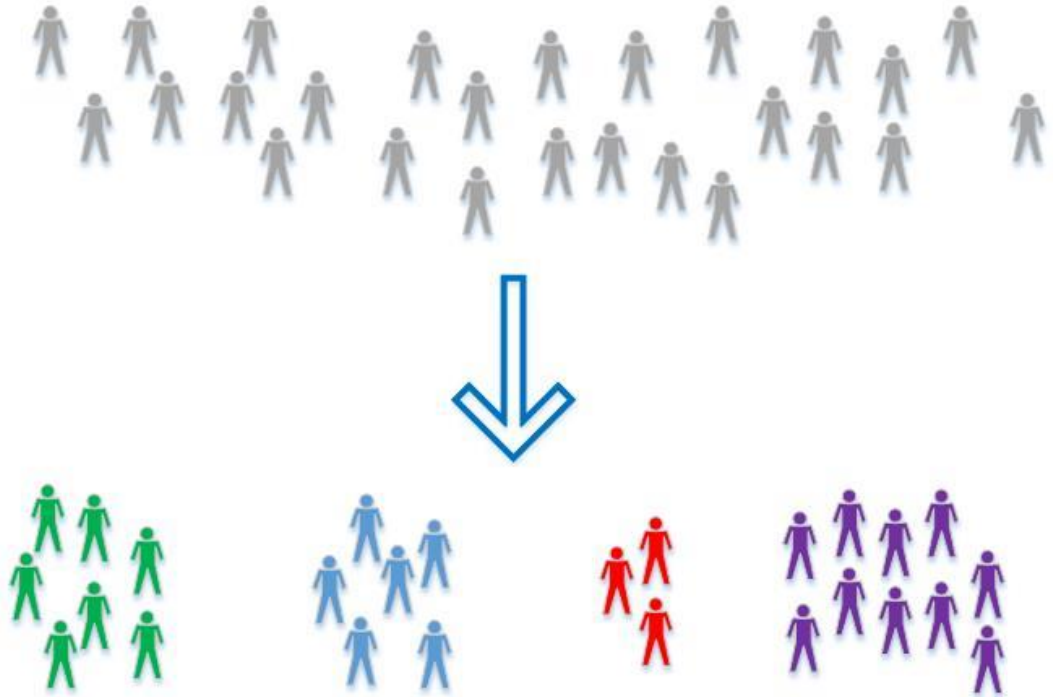
datetime	season	holiday	workingda	weather	temp	atemp	humidity	windspeed	casual	registered	count
2011-01-01 0:00	1	0	0	1	9.84	14.395	81	0	3	13	16
2011-01-01 1:00	1	0	0	1	9.02	13.635	80	0	8	32	40
2011-01-01 2:00	1	0	0	1	9.02	13.635	80	0	5	27	32
2011-01-01 3:00	1	0	0	1	9.84	14.395	75	0	3	10	13
2011-01-01 4:00	1	0	0	1	9.84	14.395	75	0	0	1	1
2011-01-01 5:00	1	0	0	2	9.84	12.88	75	6.0032	0	1	1
2011-01-01 6:00	1	0	0	1	9.02	13.635	80	0	2	0	2
2011-01-01 7:00	1	0	0	1	8.2	12.88	86	0	1	2	3
2011-01-01 8:00	1	0	0	1	9.84	14.395	75	0	1	7	8
2011-01-01 9:00	1	0	0	1	13.12	17.425	76	0	8	6	14
2011-01-01 10:00	1	0	0	1	15.58	19.695	76	16.9979	12	24	36
2011-01-01 11:00	1	0	0	1	14.76	16.665	81	19.0012	26	30	56
2011-01-01 12:00	1	0	0	1	17.22	21.21	77	19.0012	29	55	84
2011-01-01 13:00	1	0	0	2	18.86	22.725	72	19.9995	47	47	94
2011-01-01 14:00	1	0	0	2	18.86	22.725	72	19.0012	35	71	106
2011-01-01 15:00	1	0	0	2	18.04	21.97	77	19.9995	40	70	110
2011-01-01 16:00	1	0	0	2	17.22	21.21	82	19.9995	41	52	93
2011-01-01 17:00	1	0	0	2	18.04	21.97	82	19.0012	15	52	67
2011-01-01 18:00	1	0	0	3	17.22	21.21	88	16.9979	9	26	35
2011-01-01 19:00	1	0	0	3	17.22	21.21	88	16.9979	6	31	37
2011-01-01 20:00	1	0	0	2	16.4	20.455	87	16.9979	11	25	36
2011-01-01 21:00	1	0	0	2	16.4	20.455	87	12.998	3	31	34
2011-01-01 22:00	1	0	0	2	16.4	20.455	94	15.0013	11	17	28
2011-01-01 23:00	1	0	0	2	18.86	22.725	88	19.9995	15	24	39

- 자전거 대여 수요예측
- X : 날짜, 계절, 휴일 여부, 날씨, 온도, 습도 등
- y : 대여 수

지도학습

즉, 학습과정에서 데이터가 무엇을 나타내는지 알려주는 것

비지도학습



- y가 무엇인지는 모르겠지만 어떤 기준으로 나눈다.
- 데이터를 이해하기 위해 활용

지도학습의 종류

- 우리가 공부할 내용은 대부분 지도학습에 관한 것이다.

- 지도 학습은 크게 2가지로 나눌 수 있다.

- 회귀 (Regression)
  : y가 연속형 변수인 것

- 분류 (Classification)
  : y가 비연속형(명목) 변수인 것

Regression

X (hours)	y (scores)
1	30
2	50
9	80
10	90

- 공부한 시간을 기반으로 시험 점수를 예측
- 5시간 공부한 사람은 몇 점일까?

Binary Classification

X (hours)	y (P/NP)
1	NP
2	NP
9	P
10	P

- 공부한 시간을 기반으로 패스여부를 예측
- 7시간 공부한 사람은 패스일까 논패스일까?

Multi-Labeled Classification

X (hours)	y (grades)
1	D
5	C
7	B
10	A

- 공부한 시간을 기반으로 패스여부를 예측
- 8시간 어떤 학점을 받았을까?

참고 자료

모두를 위한 머신러닝/딥러닝
http://hunkim.github.io/ml/