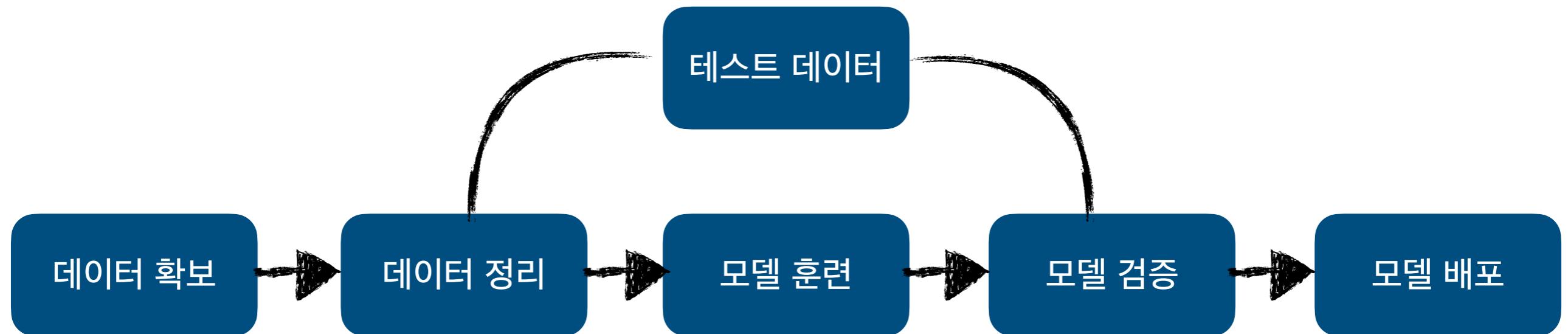


BIG DATA ANALYTICS

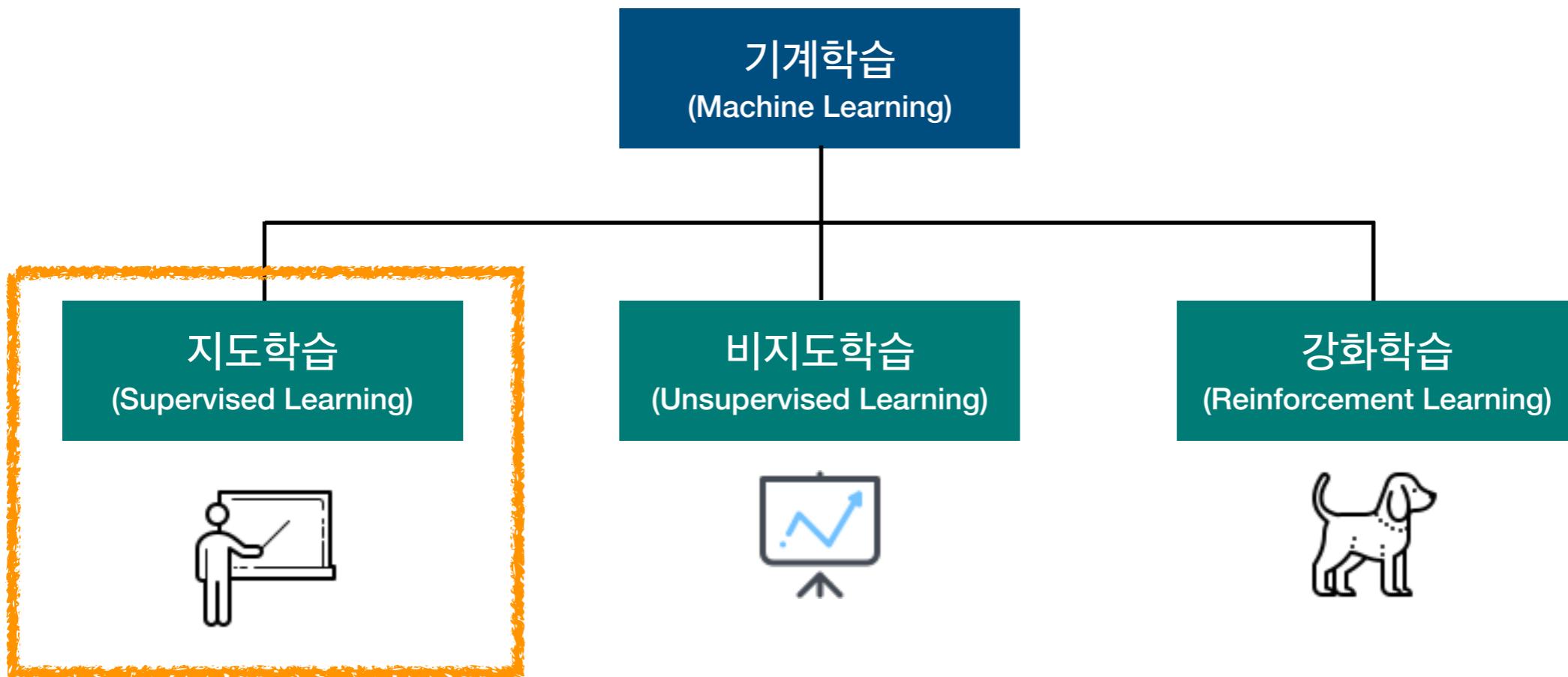
WEEK-07 | Supervised Learning - Classification

**Yonsei University
Jungwon Seo**

Machine Learning Process



기계학습의 종류



기계학습의 종류

	지도학습	비지도학습	강화학습
Training	Training Data Testing Data	No Training	Reward를 최대화
불연속 데이터	Classification	Clustering	Simulation 기반의 최적화
연속 데이터	Regression	Dimension Reduction	Autonomous Devices

지도학습(Supervised Learning)

- 정답이 있는 데이터를 이용해 학습하는 방식
 - 100,000개의 스팸메일 여부가 표시(Labeling)된 데이터셋
 - 10년치의 주가가 표시된 데이터셋
 - 강아지, 고양이와 같은 동물이 이름이 표시된 이미지 데이터셋
- 분류 문제 (Classification Task)
- 회귀 문제 (Regression Task)
 - 回: 돌아올 회, 歸: 돌아갈 귀

Classification Task

- Categorical 값을 예측하는 문제

- True or False
- A, B, C, D

$[x_1, x_2, x_3, \dots, x_{n-1}]$

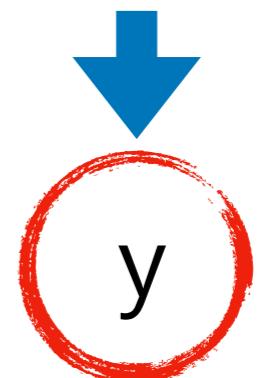
- 훈련 데이터셋 구조

- $X' = [x_1, x_2, x_3, \dots, x_n]$
- $y = x_n$ (클래스)
- $X = [x_1, x_2, x_3, \dots, x_{n-1}]$



- 테스트 데이터셋 구조

- $X' = [x_1, x_2, x_3, \dots, x_{n-1}]$



- x_n 즉, y 를 예측하는 것이 목적

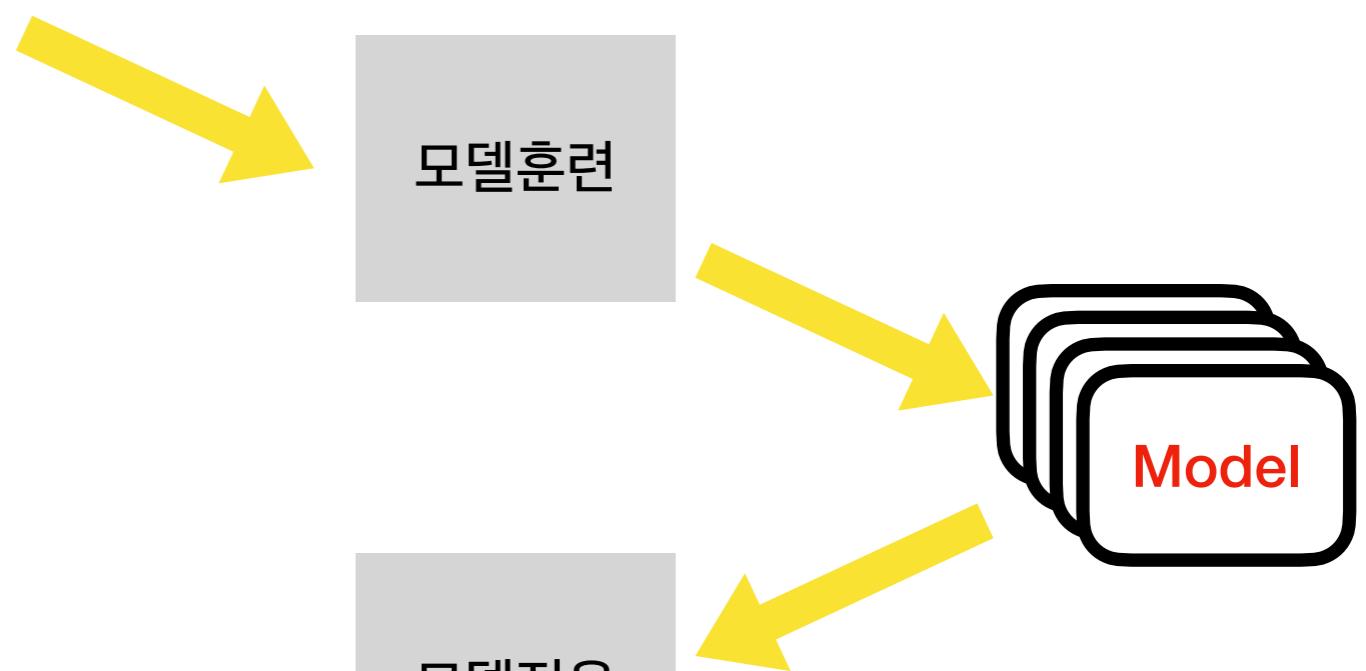
Classification Task

ID	Attr1	Attr2	Attr3	Class
1	Yes	170	66	Yes
2	Yes	180	67	Yes
3	No	160	57	No
4	No	184	75	No
5	No	192	86	Yes
6	Yes	193	99	No
7	No	175	83	Yes
8	No	165	61	Yes
9	Yes	156	50	No

훈련 데이터

ID	Attr1	Attr2	Attr3	Class
10	No	179	76	Yes
11	No	170	67	Yes
12	Yes	169	54	No
13	Yes	180	86	No
14	Yes	182	88	No

검증 데이터



Classification Techniques

- Statistical Methods
- Rule-based Methods
- Decision Tree based Methods
- Memory based reasoning
- Neural Networks
- Naïve Bayes and Bayesian Belief Networks
- Support Vector Machines

K-Nearest Neighbor

- 간단한 Classification 알고리즘
- 이름에서 알 수 있듯이, 분류를 원하는 데이터를 기준으로 **가장 가까운 K개의 주변** 데이터 포인트들을 확인
- K개의 데이터 포인터들의 Class를 기준으로 새 데이터의 Class를 판별
- “친구를 보면 그 사람을 알 수 있다”

K-Nearest Neighbor

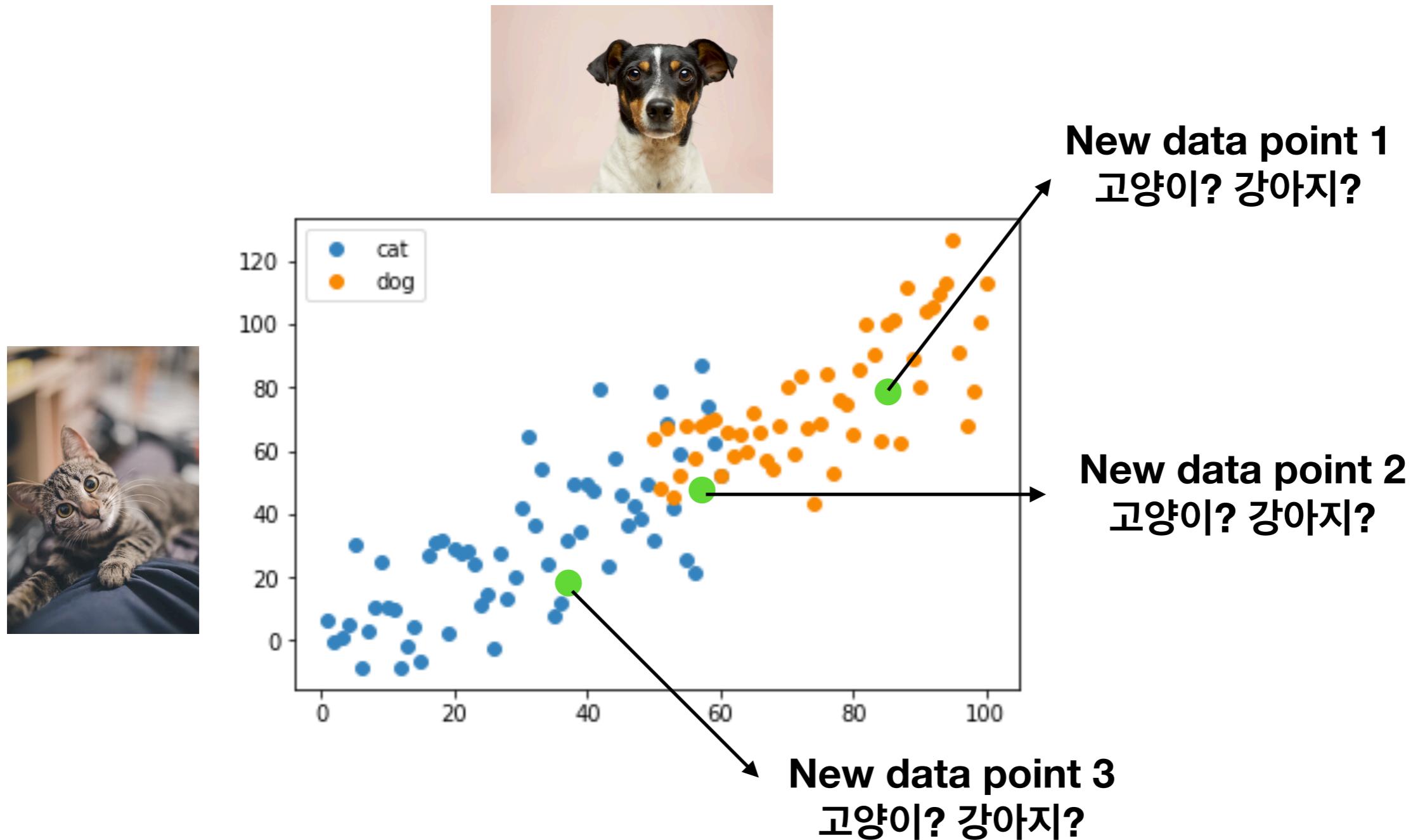


Photo by [Victor Grabarczyk](#) on [Unsplash](#)

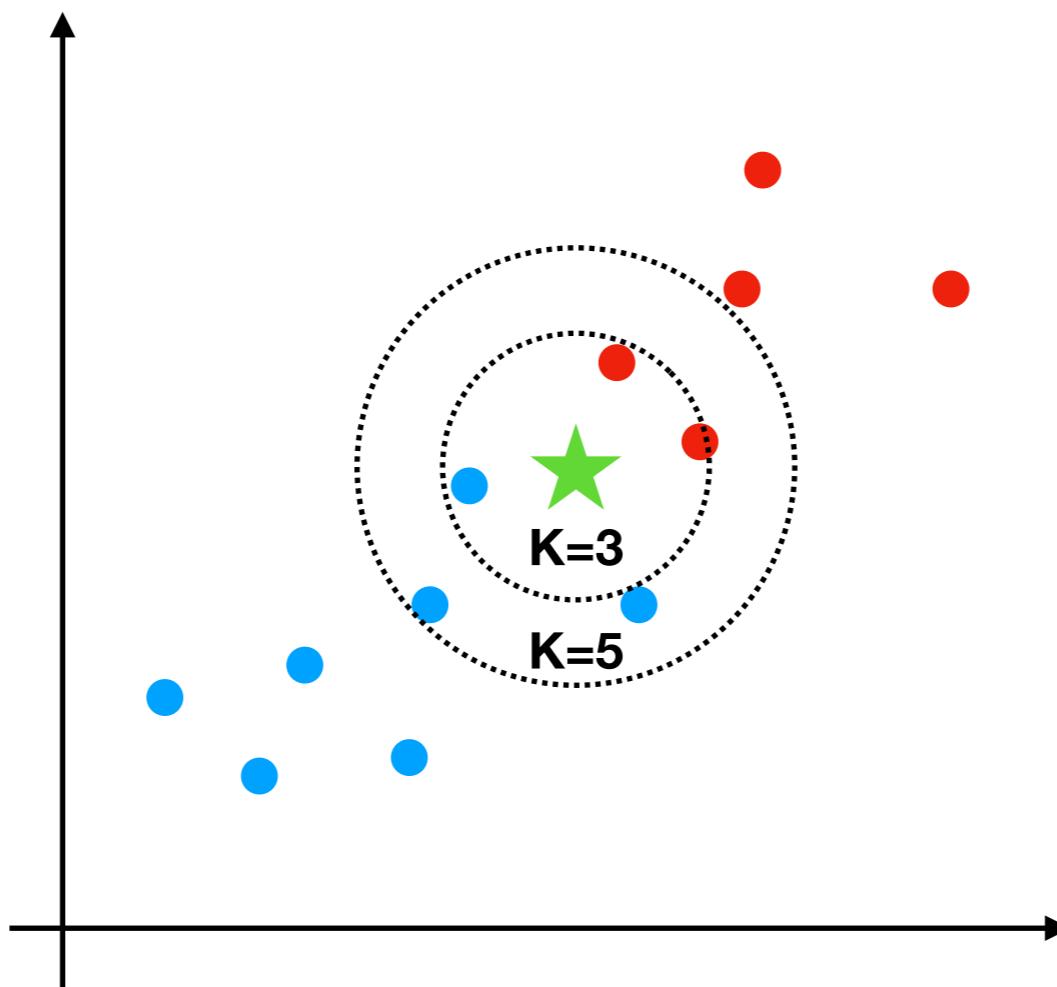
Photo by [Ramiz Dedaković](#) on [Unsplash](#)

K-Nearest Neighbor

- 학습 알고리즘
 - 모든 데이터 포인트를 저장
 - $\text{data} = [\text{data1}, \text{data2}, \dots]$
 - $\text{data1} = [x_1, x_2, \dots, x_{n-1}], \text{class}$
 - $\text{data2} = [x_1, x_2, \dots, x_{n-1}], \text{class}$
 - $\text{data3} = [x_1, x_2, \dots, x_{n-1}], \text{class}$
- 예측 알고리즘
 - 새 데이터 입력: $\text{new_data} = [x_1, \dots, x_{n-1}]$
 - new_data 를 기준으로 기존의 데이터들과의 거리계산
 - 데이터가 100,000개면 100,000번의 거리 계산
 - 거리 순으로 데이터들을 정렬 (오름차순)
 - 상위 K개의 데이터의 Class를 다수결로 적용하여 new_data 의 Class 결정
 - If $K = 3$
 - ($\text{data1} = \text{"dog"}, \text{data2} = \text{"dog"}, \text{data3} = \text{"cat"}$) $\Rightarrow \text{new_data} = \text{"dog"}$

K-Nearest Neighbor

- K를 어떻게 결정하느냐에 따라 결과가 달라질 수 있음



- K가 짹수면? K가 너무 크면? K가 너무 작으면?

K-Nearest Neighbor

장점

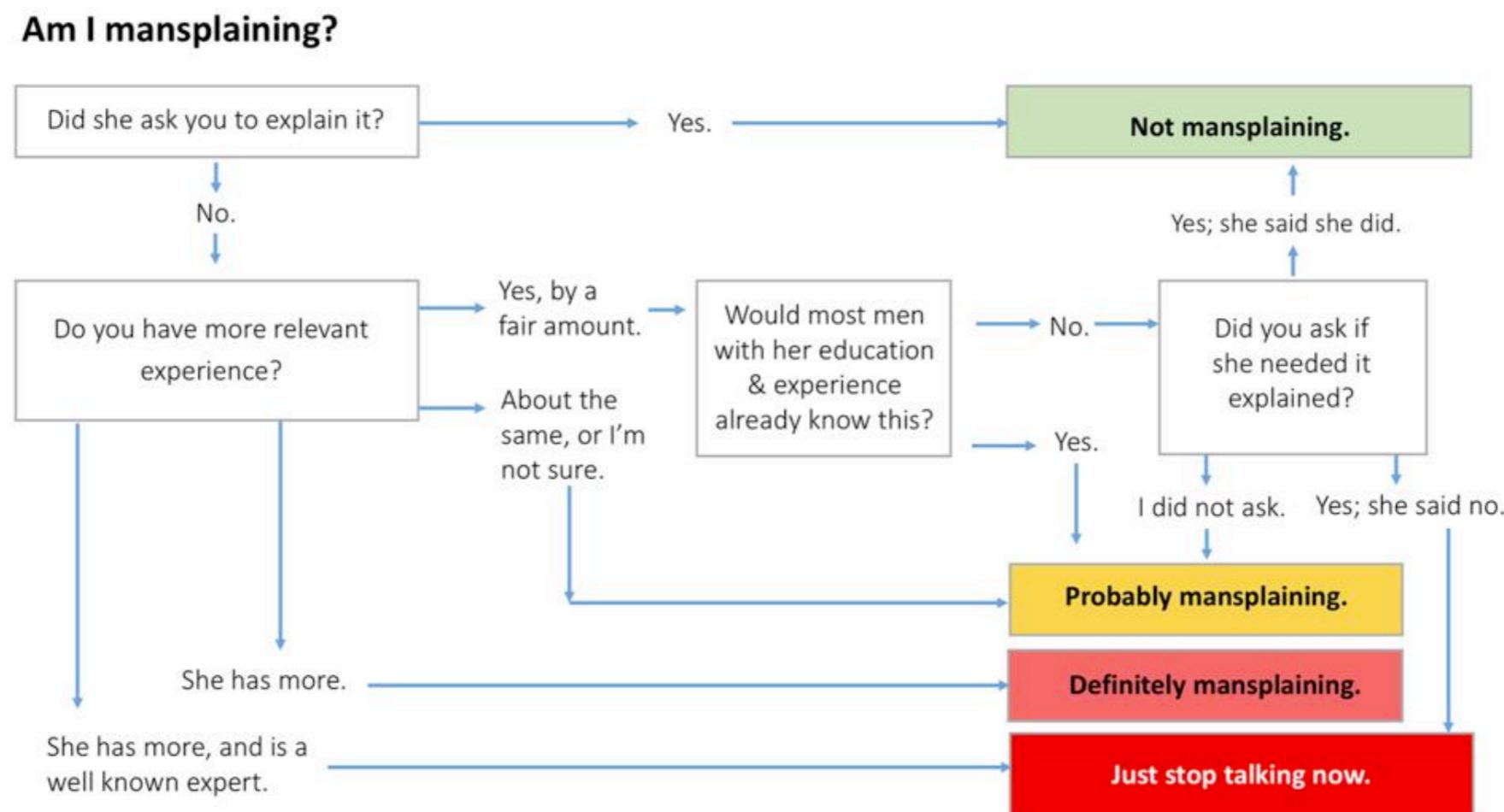
- 간단함
- 학습이 별로 안 중요함
- 어떤 수의 Class에도 적용가능
- 새로운 데이터 추가의 용이함
- 적은 파라미터: K, Distance

단점

- ▶ 예측을 위한 계산 비용이 큼
- ▶ 고차원 데이터에 부적합
- ▶ 범주형 feature에 부적합

Decision Tree

- 알고리즘을 시각화 하는것과 비슷



Tree?



VS



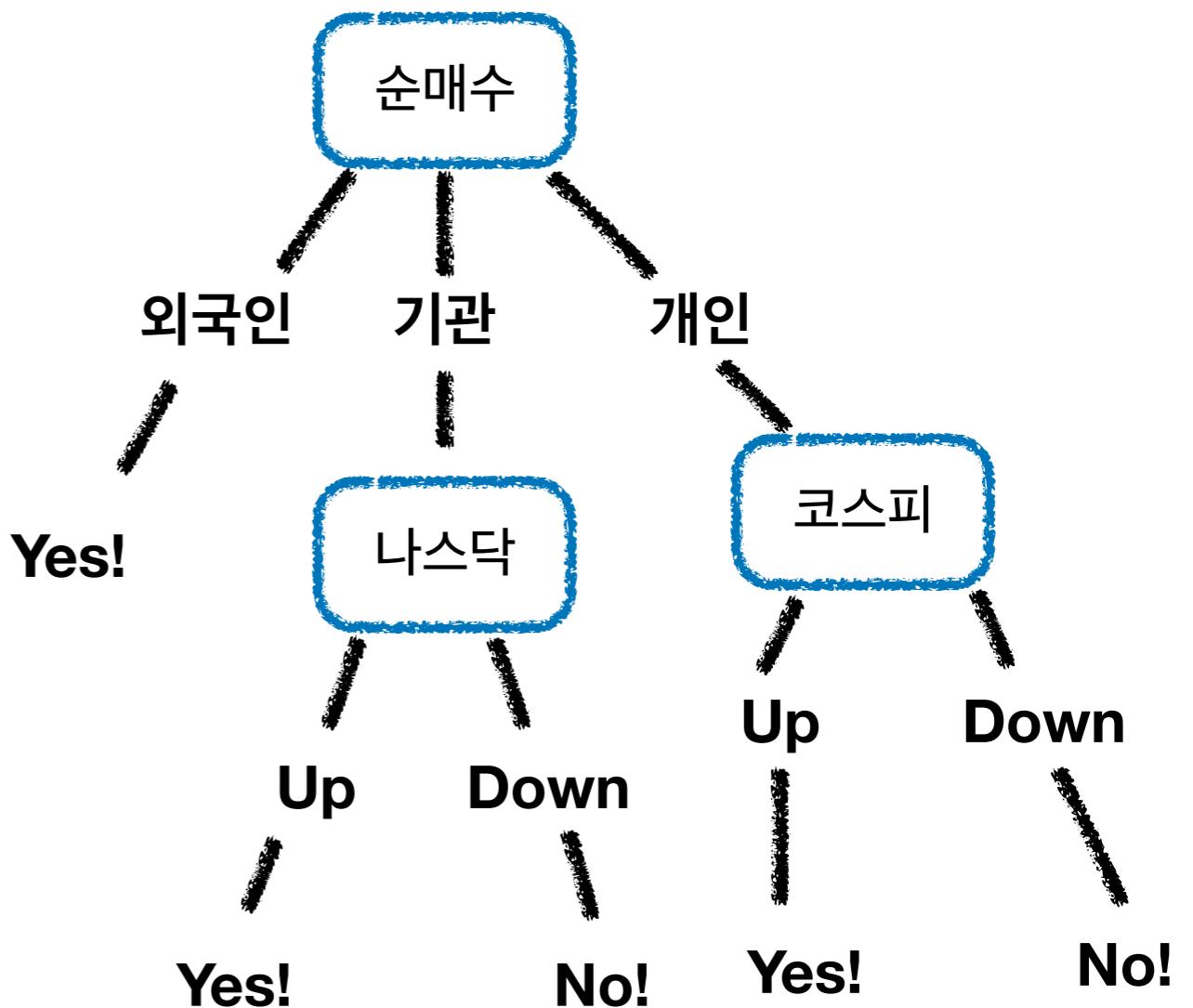
Computer science

The rest of the world

Photo by [Gilly Stewart](#) on [Unsplash](#)

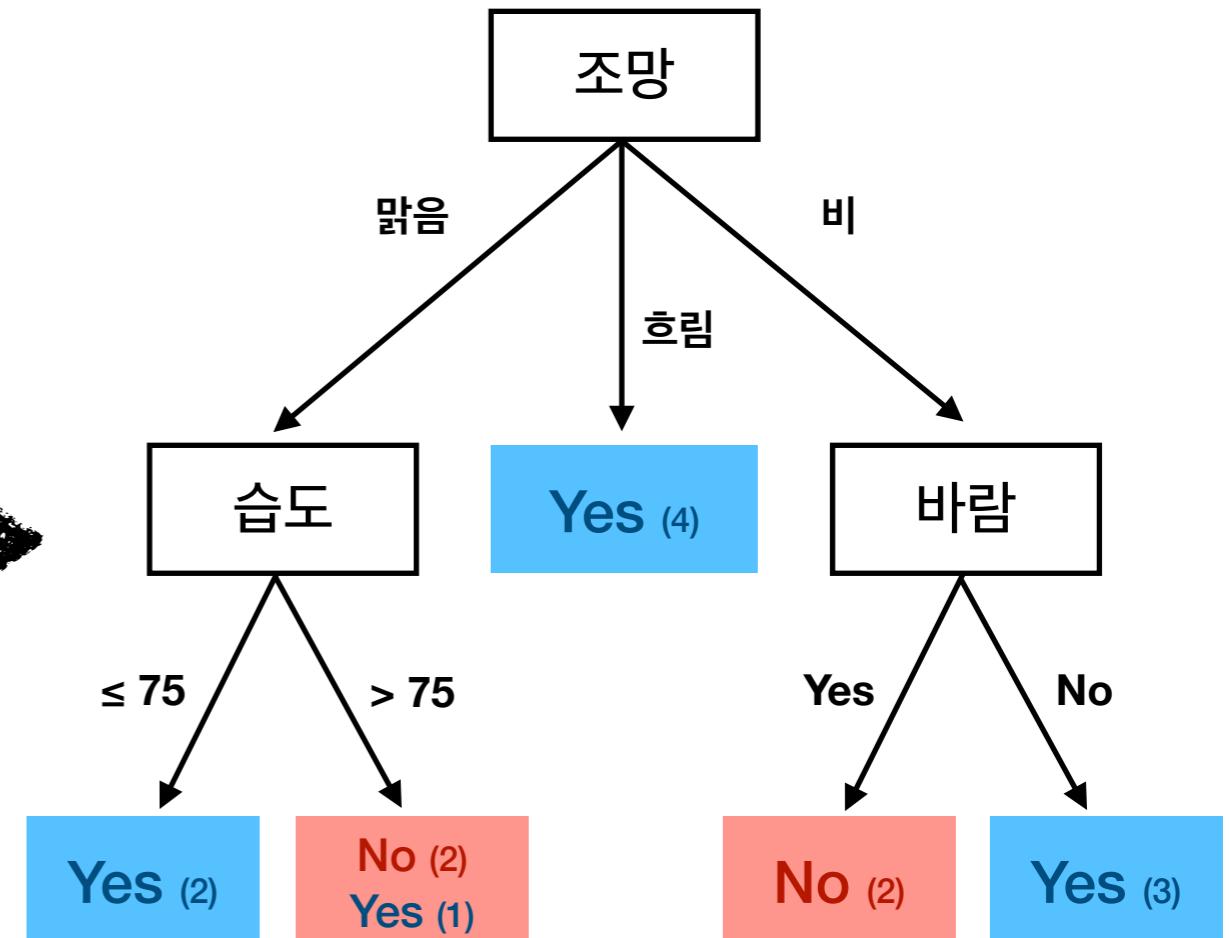
의사 결정 나무

- Node (○) : 변수 / Feature
- Edge (―) : 값 / 기준값



의사 결정 나무

온도	조망	습도	바람	테니스
보통	맑음	80	No	Yes
더움	맑음	75	Yes	No
더움	흐림	77	No	Yes
시원함	비	70	No	Yes
시원함	흐림	72	Yes	Yes
보통	맑음	77	No	No
시원함	맑음	70	No	Yes
보통	비	69	No	Yes
보통	맑음	65	Yes	Yes
보통	흐림	77	Yes	Yes
더움	흐림	74	No	Yes
보통	비	77	Yes	No
시원함	비	73	Yes	No



학습 데이터셋

의사결정트리

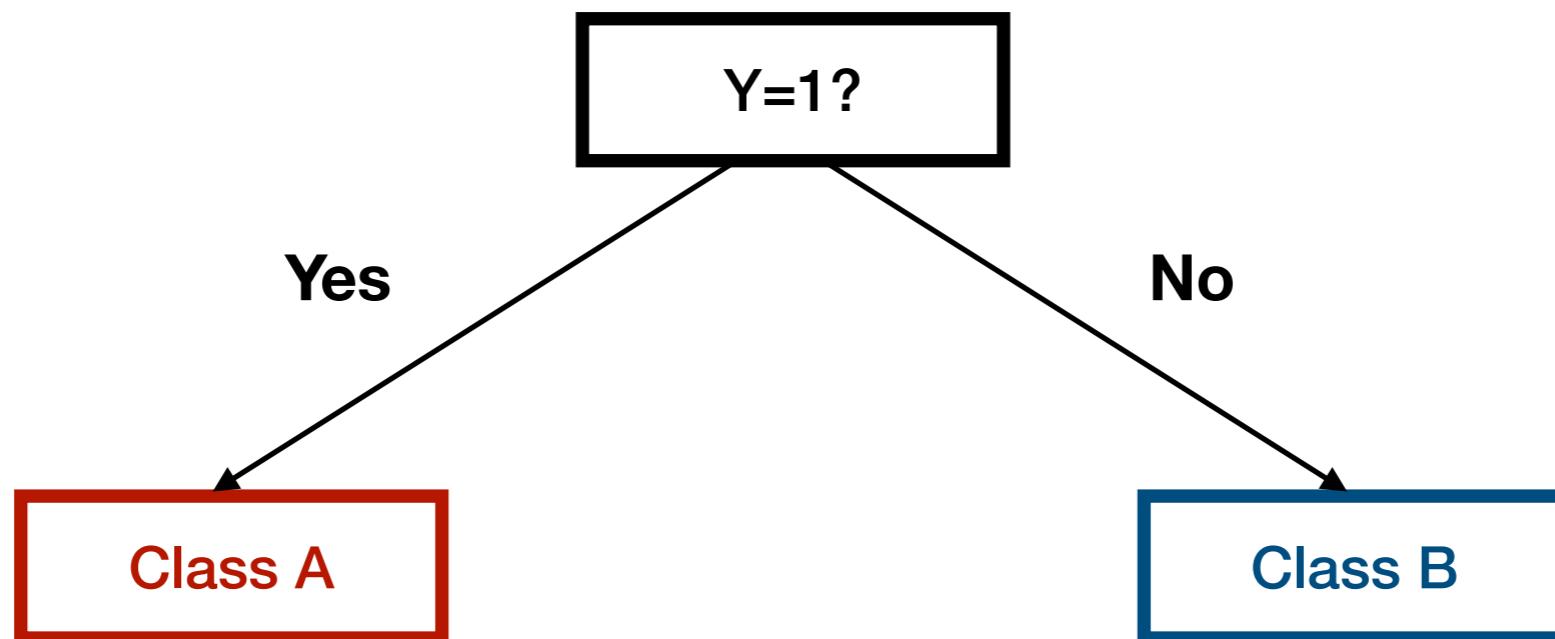
가지 나누기(split) 원리

- 3개의 feature와 2개의 class로 이루어진 dataset

X	Y	Z	Class
1	1	1	A
1	1	0	A
0	0	1	B
1	0	0	B

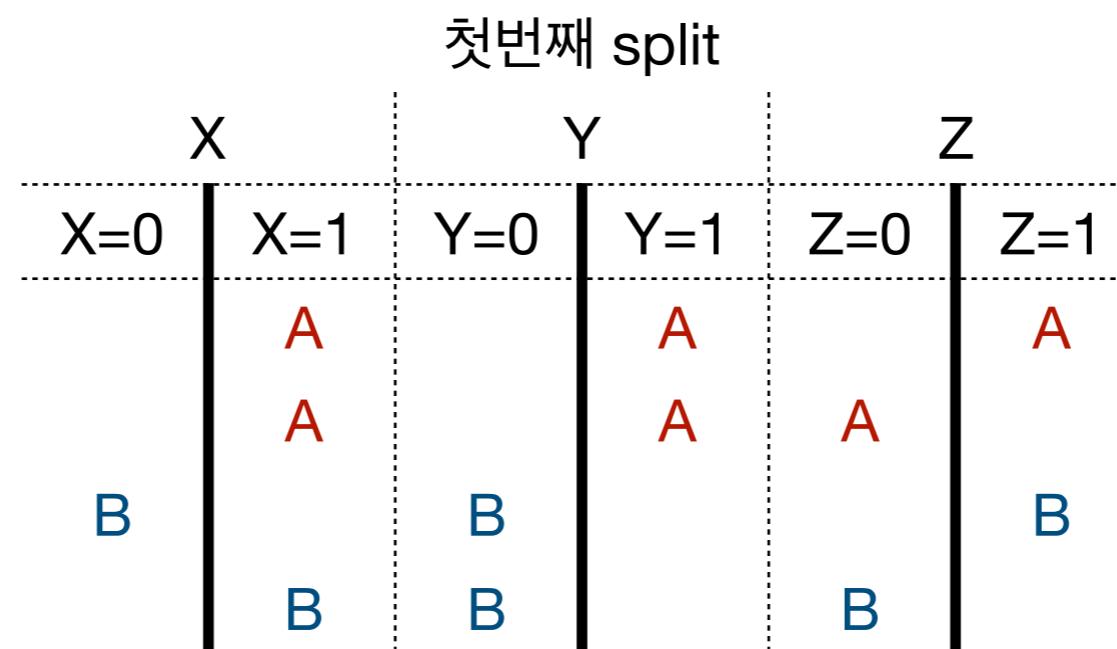
가지 나누기(split) 원리

- Feature Y로 나눌 시 두 클래스를 분리 가능



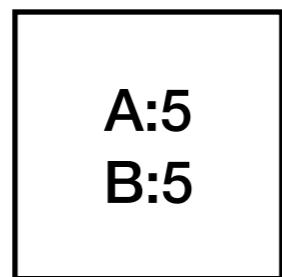
가지 나누기(split) 원리

- 다른 feature로 split 할 경우

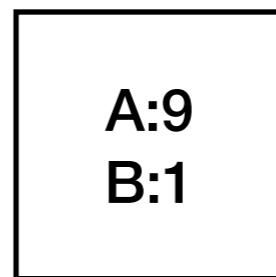


가지 나누기(split) 원리

- 어떤 feature로 나누는 것이 최선일까?
- 수치적인 근거는 없을까?
- Greedy Approach
 - 같은(homogenous) 클래스끼리 묶어 주는 Split
 - Impurity(불순도)를 측정할 방법이 필요



Non-homogenous
High degree of impurity

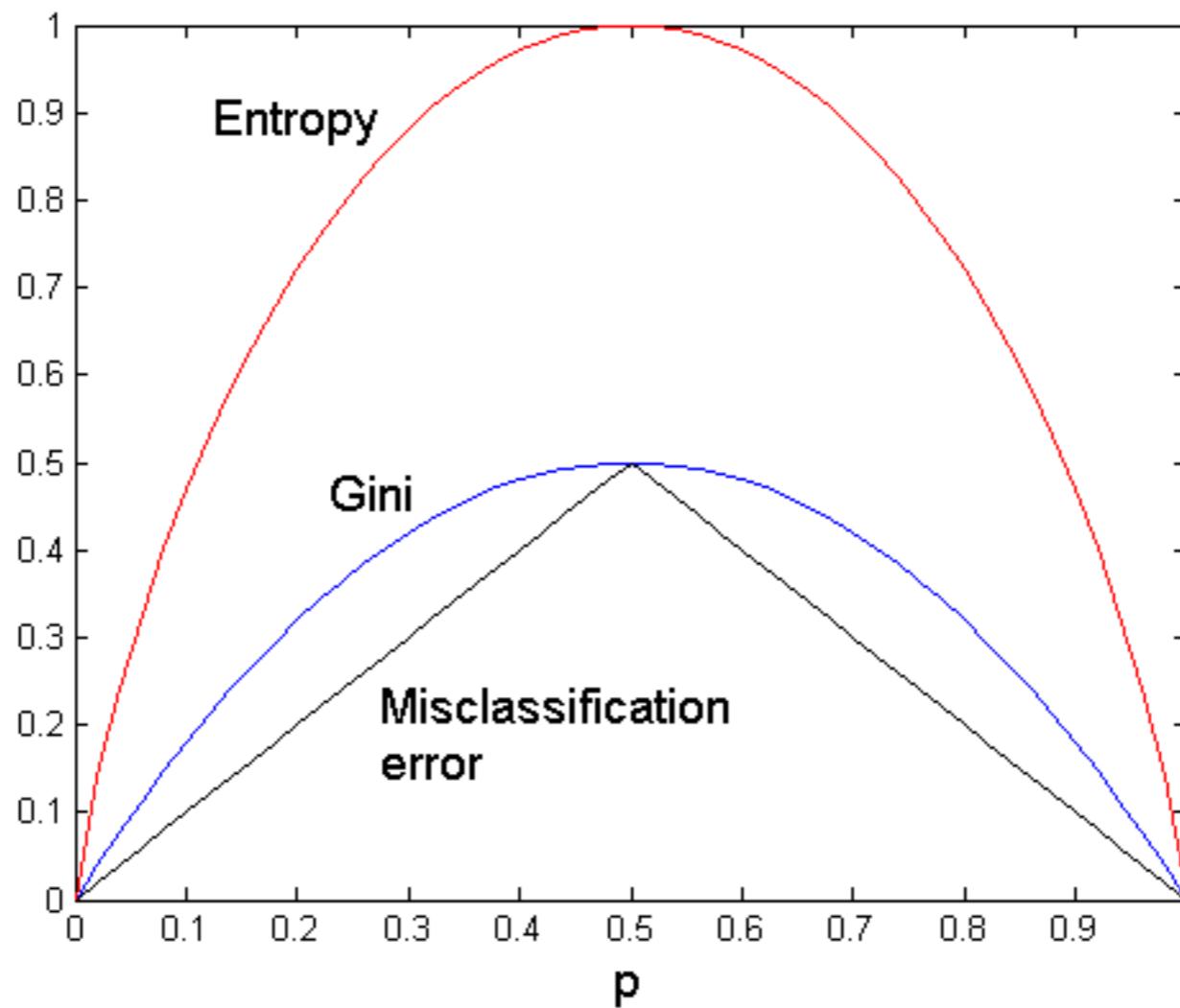


Homogenous
Low degree of impurity

노드 불순도 측정

- Gini index = $1 - \sum_j p_j^2$
- Entropy = $-\sum_j p_j \log_2 p_j$
- Misclassification error = $1 - \max p_j$

가지 나누기(split) 원리



E.O.D