

# 데이터 분석의 기초와 실습

Wan Ju Kang

- 기계학습에 대해서 high-level 개념 이해 및 용어 소개
- 기계학습의 주요 알고리즘 소개 및 이해
- 주요 딥러닝 알고리즘 소개 및 이해
- 학습된 내용을 실제 python으로 구현 및 실습

- Weeks 1~2) 데이터 분석의 기초와 실습
  - Part 1) 기계학습(machine learning) 및 python 소개, 선형 회귀(linear regression) 알고리즘
  - Part 2) 비선형 회귀(non-linear regression) 및 분류(classification) 알고리즘
- Weeks 3~4) 머신러닝 응용 알고리즘과 실습
  - Part 3) 비지도 학습(unsupervised learning) 개념 및 알고리즘 소개
  - Part 4) 앙상블 방법 및 의사 결정 나무(decision tree) 알고리즘 소개
- Weeks 5~6) 딥러닝 (deep learning) 알고리즘과 실습
  - Part 5) 퍼셉트론(perceptron) 및 역전파(backpropagation) 소개
  - Part 6) 합성곱 신경망(convolutional neural network) 및 회귀 신경망(recurrent neural network) 응용

# Part 1) 기계학습 및 python 소개, 선형 회귀 알고리즘

# Contents

---

## 1. 기계학습의 소개

- 1) 기계학습이 실생활에서 쓰이는 곳
- 2) 기계학습이란 무엇인가?
- 3) 기계학습의 종류

## 2. 회귀

- 1) 선형 회귀 분석 알고리즘 소개
- 2) 알고리즘 평가(evaluation) 방법 소개
- 3) 다중 선형 회귀 분석(multiple linear regression) 알고리즘 소개

# 이미지의 동물을 맞히는 알고리즘 어떻게?



개? 고양이? 구분하는 하는 법?

# 그림을 그리는 알고리즘

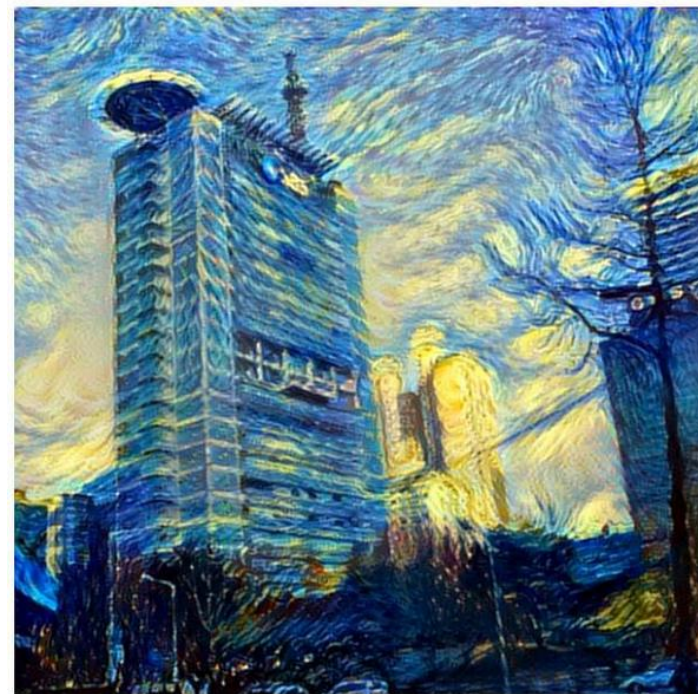
- 생성적 적대 신경망
  - Generative Adversarial Network (GAN)

입력 사진



+

고흐 화풍

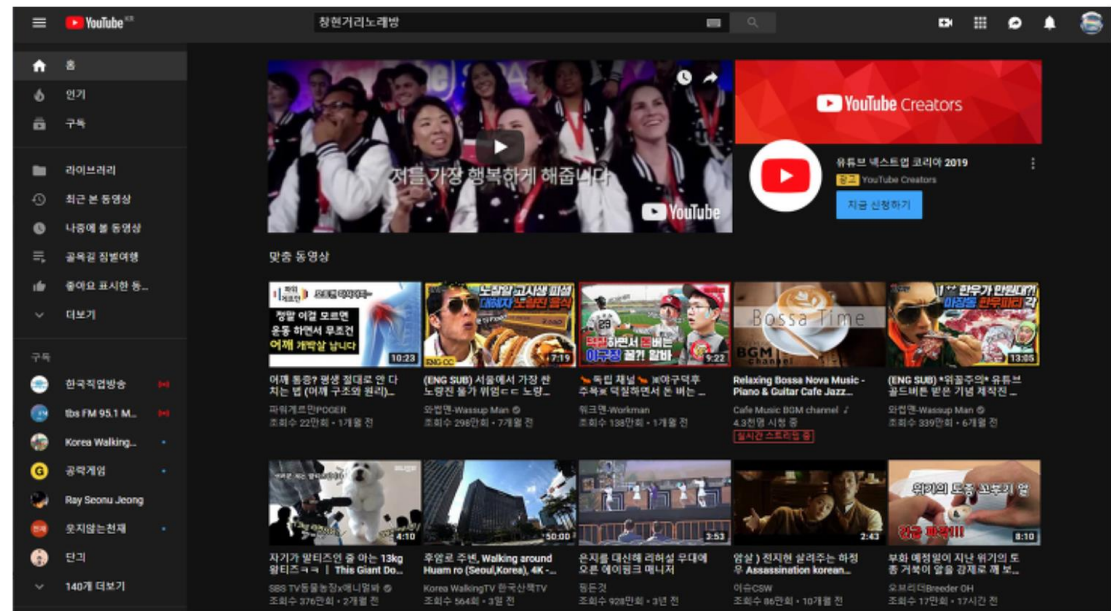










입력된 사진을  
고흐 화풍으로 그려본 그림



# 추천 알고리즘

- Youtube, Netflix, 쿠팡 등등



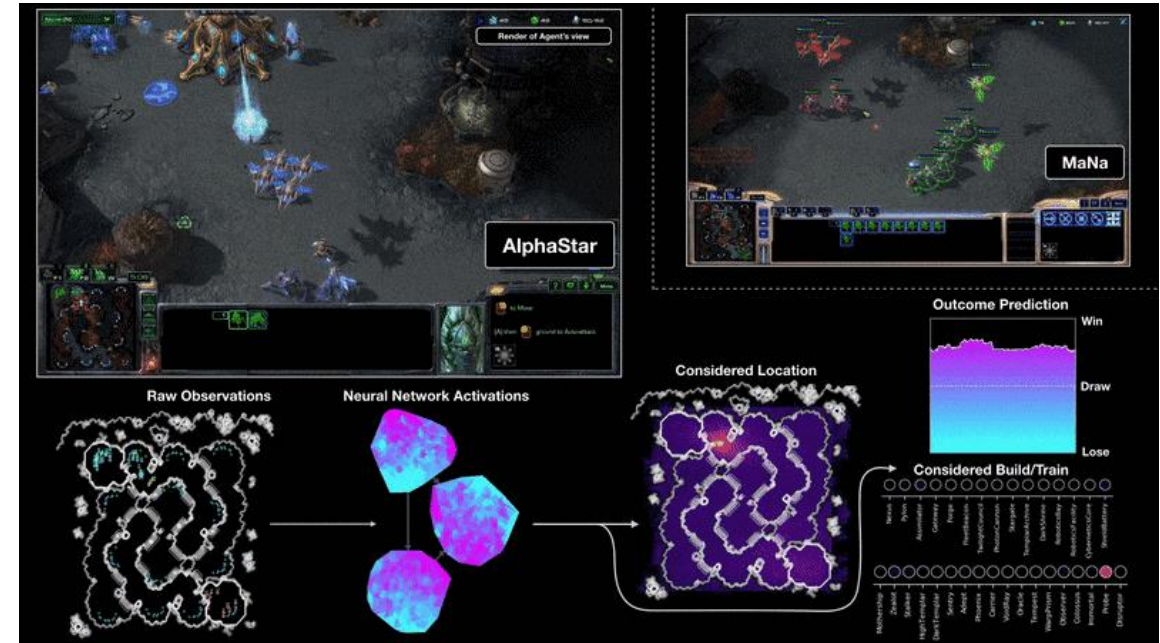
 <p>메일로우 오리지널 우유, 200mL, 24팩</p> <p>23% 18,000원 <b>13,740원</b> 로켓배송 내일(목) 12/16 도착 보장</p> <p>★★★★★ (116980) 최대 687원 적립</p>	 <p>탐사수, 500mL, 40개</p> <p>29% 11,900원 <b>8,390원</b> 로켓배송 내일(목) 12/16 도착 보장</p> <p>★★★★★ (893196) 최대 420원 적립</p>	 <p>제주상다수, 2L, 12개</p> <p>29% 11,760원 <b>11,760원</b> 로켓배송 내일(목) 12/17 도착 예정 (일반 택배)</p> <p>★★★★★ (435112) 최대 588원 적립</p>	 <p>무향주원 추가할인 쿠폰 규당 당도선별 감귤 (로열과, 3kg(S-M), 1박스)</p> <p>와우할인가 53% 29,800원 <b>13,900원</b> 로켓배송 내일(목) 새벽 도착 보장</p> <p>★★★★★ (14502) 최대 695원 적립</p>
 <p>추가할인 쿠폰 공공 GAP고양도 제주감귤, 2.5kg, 1박스</p> <p>와우할인가 47% 24,800원 <b>12,990원</b> 로켓배송 내일(목) 새벽 도착 보장</p> <p>★★★★★ (15786) 최대 650원 적립</p>	 <p>펩시 콜라 제로 슈거 라병형, 210mL, 30개</p> <p>37% 23,250원 <b>14,910원</b> 로켓배송 내일(목) 12/16 도착 보장</p> <p>세 상종, 박스 묶음 (28) 최저14,460원 ★★★★★ (26693) 최대 745원 적립</p>	 <p>제주상다수, 2L, 24개</p> <p>29% 23,520원 <b>23,520원</b> 로켓배송 내일(목) 12/17 도착 예정 (일반 택배)</p> <p>★★★★★ (435111) 최대 1,176원 적립</p>	 <p>메일 소화가 잘되는 우유, 190mL, 24개</p> <p>36% 24,000원 <b>15,180원</b> 로켓배송 내일(목) 12/16 도착 보장</p> <p>★★★★★ (45717) 최대 759원 적립</p>



# 바둑, 스타크래프트를 하는 알고리즘



AlphaGO



AlphaStar

이러한 알고리즘들은 모두 **기계 학습**을 기반으로 사람의 지능이 있어야만 가능하다고 여겨진 일들을 할 수 있음!

# Contents

---

## 1. 기계학습의 소개

- 1) 기계학습이 실생활에서 쓰이는 곳
- 2) 기계학습이란 무엇인가?
  1. 기계학습 = 패턴 찾기
  2. 기계학습 vs 인공지능, 기계학습 vs 통계
- 3) 기계학습의 종류

## 2. 회귀

- 1) 선형 회귀 분석 알고리즘 소개
- 2) 알고리즘 평가(evaluation) 방법 소개
- 3) 다중 선형 회귀 분석(multiple linear regression) 알고리즘 소개

# 기계 학습(machine learning)이란 무엇인가?

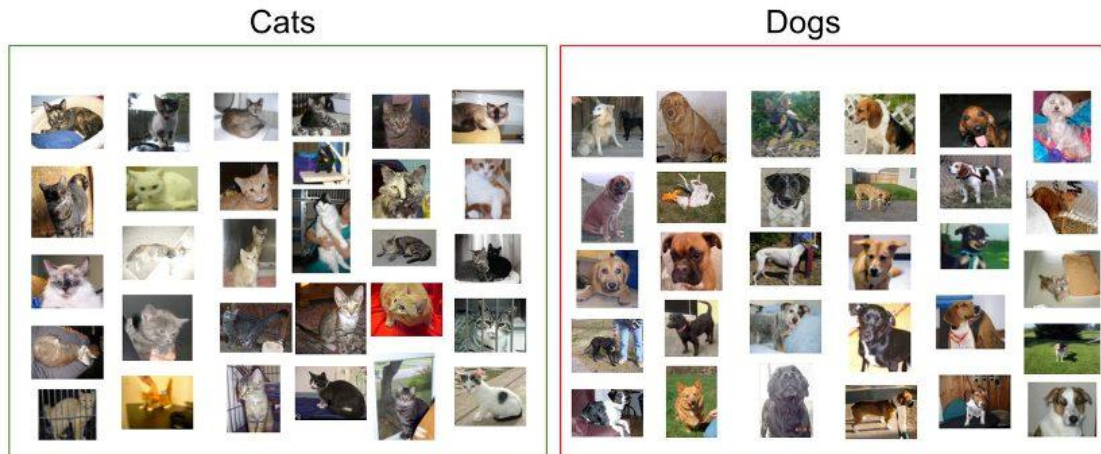
- 기계학습의 정의

- 수집된 데이터 셋(data set)에 존재하는 특정 패턴을 학습하여
- 유용한 알고리즘을 개발하는 컴퓨터 공학의 한 갈래

데이터

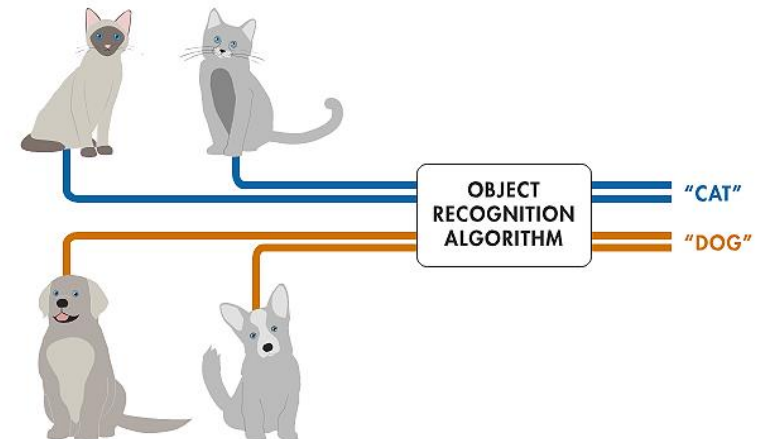
machine learning

알고리즘



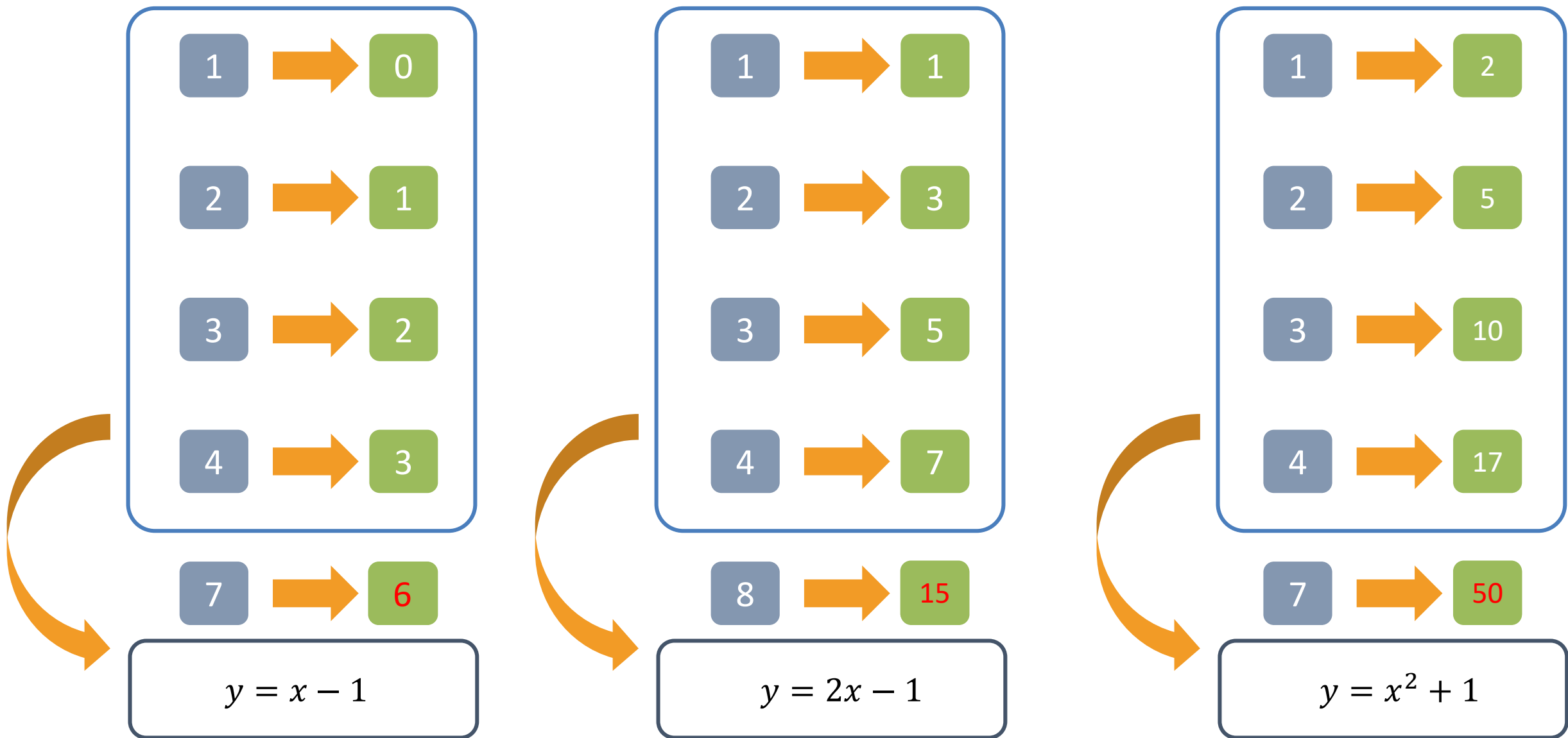
학습용 데이터 (개, 고양이)

machine learning



개와 고양이를 구분하는 분류 알고리즘

# 기계학습 = 패턴 찾기



# 단지 패턴이 복잡할 뿐...

1 → 0

2 → 1

3 → 2

4 → 3

7 → 6

$$y = x - 1$$



0



1



0



0

고양이 = 0, 개 = 1

$$y = F(x)$$

F는 상당히 복잡한 형태의 함수이기에 찾기 어렵다!

# Contents

---

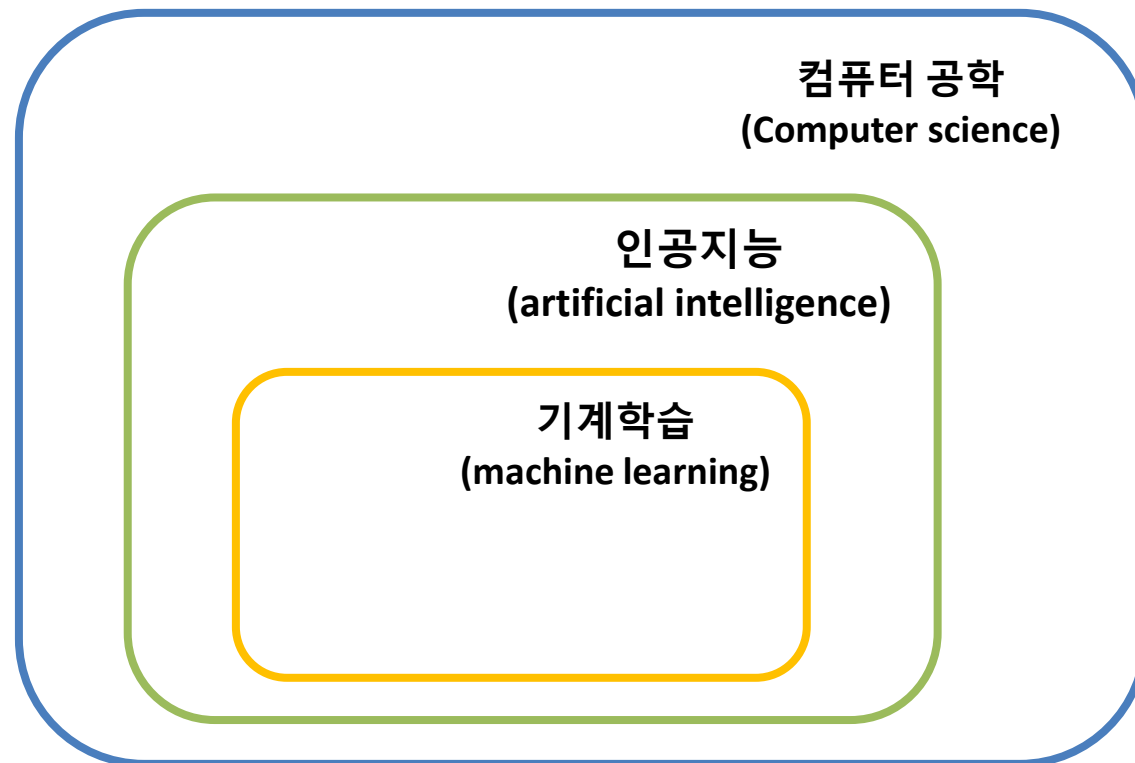
## 1. 기계학습의 소개

- 1) 기계학습이 실생활에서 쓰이는 곳
- 2) 기계학습이란 무엇인가?
  1. 기계학습 = 패턴 찾기
  2. 기계학습 vs 인공지능, 기계학습 vs 통계
- 3) 기계학습의 종류

## 2. 선형 회귀 분석

- 1) 선형 회귀 분석 알고리즘 소개
- 2) 알고리즘 평가(evaluation) 방법 소개
- 3) 다중 선형 회귀 분석(multiple linear regression) 알고리즘 소개

- 인공지능
  - 기계 혹은 시스템에 의해 만들어진 지능
  - 사람처럼 주변 환경과 상호작용하는 다양한 방법을 총칭
- 머신러닝
  - 기계(컴퓨터)가 직접 기본적인 규칙을 가지고 입력받은 **데이터**를 분석
  - **데이터** 안의 패턴 및 규칙성을 학습하여 유용한 상호작용 방법
- 머신러닝은 인공지능의 한 방법!



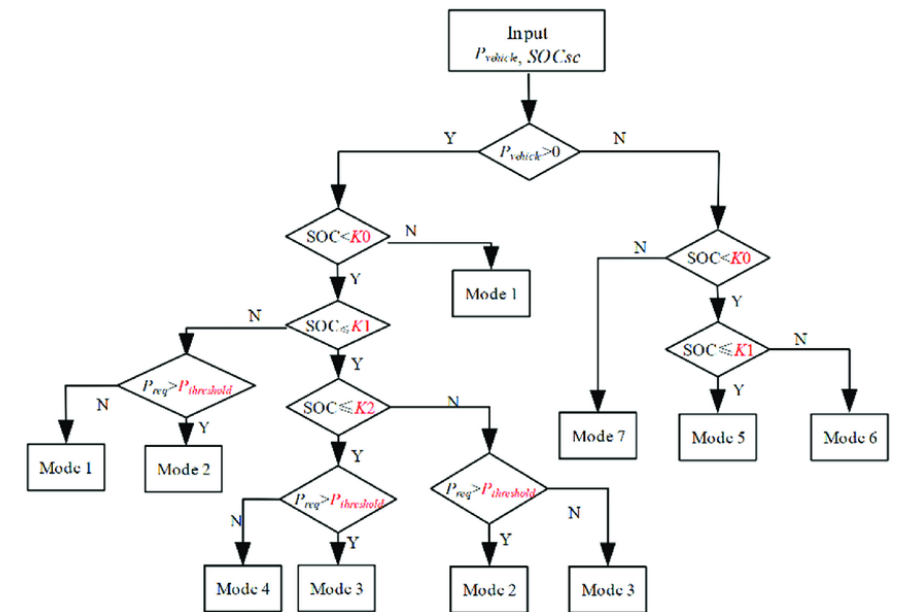


# 규칙 기반 인공지능(Rule-based AI)

- 1999년에 이미 컴퓨터 대전 모드 존재
- 특징
  - 사람이 미리 규칙을 정하여서 그 규칙대로 자신의 전략을 진행 (**rule-based algorithm**)
    - 일꾼 7마리 → 첫번째 건물 건설 → 8마리로 공격
  - 데이터를 바탕으로 학습한 것이 아니라 사람의 직관을 바탕으로 정함
  - 특정 상황에서는 상당히 강력함
  - 사람이 미리 상정하지 않은 상황에서는 성능이 매우 저하됨
- 인공지능이지만, 머신러닝은 아님!



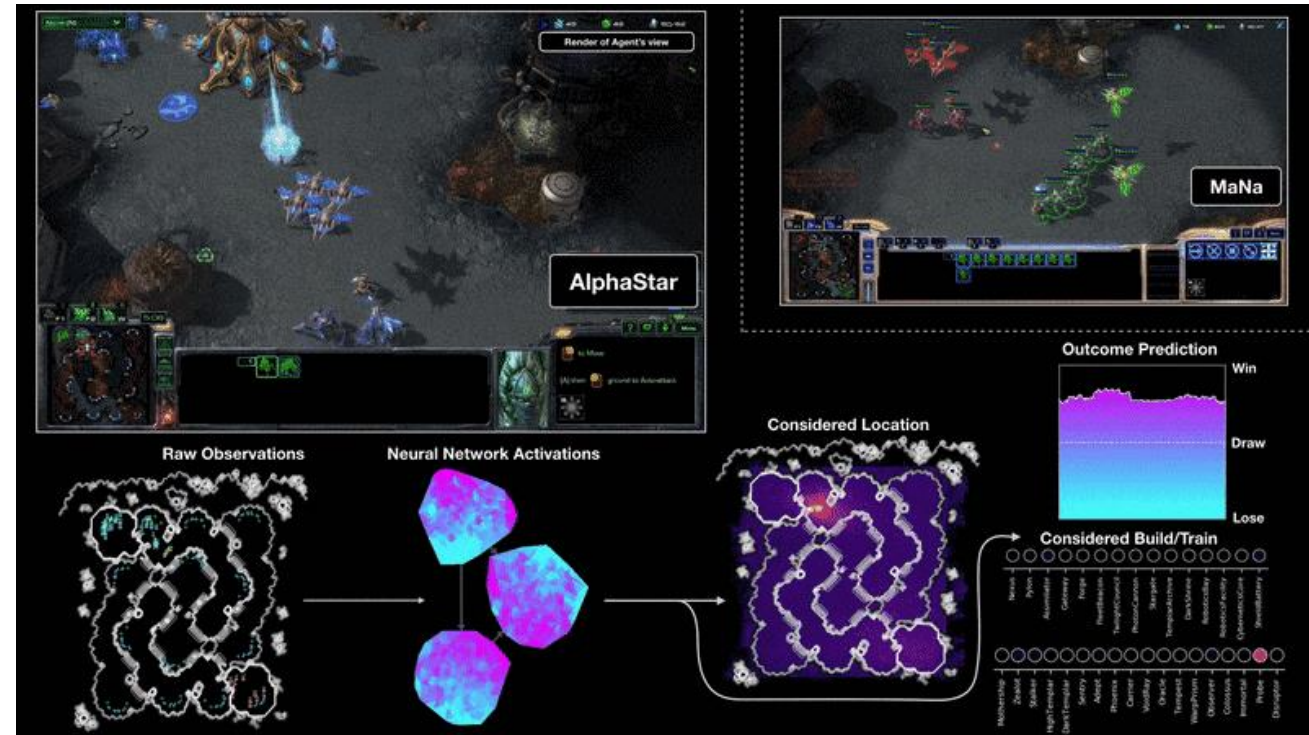
VS computer



Rule-based algorithm

- Alphastar

- 머신러닝 기법을 활용하여서,
- **사람의 실제 경기 데이터를** 기반으로 실제 경기와 비슷하게 플레이하는 알고리즘 개발
- **컴퓨터 알고리즘들끼리 플레이한 경기 데이터를** 기반으로 알고리즘들이 스스로 자신들의 전략을 개선
  - 이 과정에서, 컴퓨터 알고리즘들은 스스로 '패스트 다크템플러' (극단적인 도박수)을 배움
  - 나중에는 이 전략을 카운터 치는 전략도 스스로 학습
- 최종적으로 프로게이머 수준으로 성능이 올라감!



## • 규칙 기반 알고리즘

### • 장점

- 데이터가 없거나 적은 경우에도 개발 가능
- 사람이 직접 만들기에, 알고리즘의 버그 등이 발견되었을 때 개선이 쉬움
- 최소한의 성능이 보장됨

### • 단점

- 모든 상황을 다 고려해서 알고리즘을 만들기가 어려움
- 잠재적인 성능이 머신러닝에 비해서 떨어짐

## • 머신러닝

### • 장점

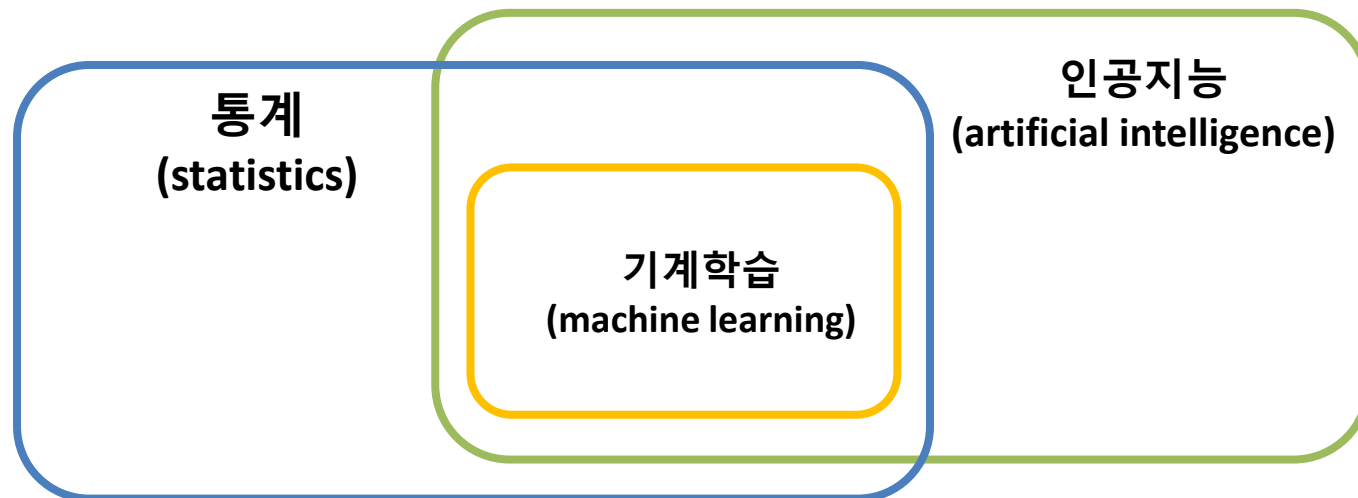
- 데이터가 충분히 많다면, 최종적인 잠재력이 높음
- 사람이 직접 모든 경우의 수를 고려하지 않아서 편리함

### • 단점

- **충분한 양의 정확한 데이터가 필요함**
- 잘못된 방식으로 학습할 경우 성능이 불안정한 경우가 종종 나옴
- 학습을 하였는데, 성능이 잘 나오지 않을 경우 개선하기가 어려움

**실제 상황에서는 두가지 방법의 장점을 균형 있게 섞어서 사용하는 것이 중요!**

- 두 가지 방법 둘다 데이터를 기반으로 통계적 패턴을 분석하는 공통점이 있다.
- 다만, 통계는 주어진 데이터의 분석에 더 초점을 가지고 있고,
  - ex) 통계를 내보니, 사람들이 상품 A의 만족도가 90%, B는 80%이다.
- 기계학습의 경우에는, 분석을 바탕으로 자동적으로 변화하는 알고리즘까지 개발한다는 특징이 있다.
  - ex) 상품 A의 만족도가  $x$ , B의 만족도가  $y$ 일 경우 상품 A를 추천 비율은  $\frac{x}{x+y}$ 로 하자.



# Contents

---

## 1. 기계학습의 소개

- 1) 기계학습이 실생활에서 쓰이는 곳
- 2) 기계학습이란 무엇인가?
  1. 기계학습 = 패턴 찾기
  2. 기계학습 vs 인공지능, 기계학습 vs 통계
- 3) 기계학습의 종류
  1. 지도학습, 비지도학습, 강화학습
  2. 지도학습의 종류: 회귀, 분류

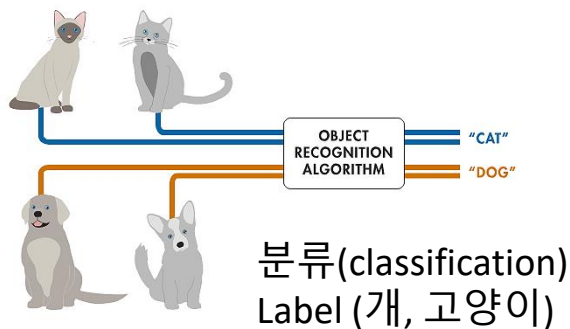
## 2. 선형 회귀 분석

- 1) 선형 회귀 분석 알고리즘 소개
- 2) 알고리즘 평가(evaluation) 방법 소개
- 3) 다중 선형 회귀 분석(multiple linear regression) 알고리즘 소개

## 기계학습(Machine Learning)

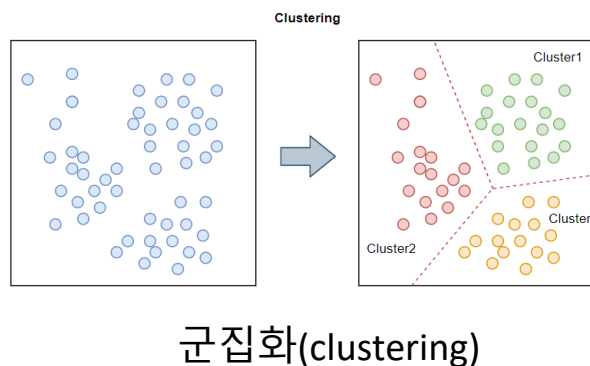
### 지도학습 (Supervised Learning)

정답(Label)이 **있는** data  
를 받고, 새로운 문제에  
대해서 정답을 맞추는 기  
계학습



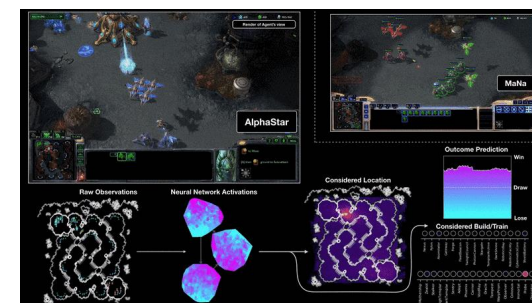
### 비지도학습 (Unsupervised Learning)

정답 (Label)이 **없는** data  
를 받고, data의 구조를  
분석하는 기계학습



### 강화학습 (Reinforcement Learning)

스스로 **자신에게 가장 필  
요한 data**를 선택하면서  
학습을 하는 기계학습



전략 시뮬레이션 학습

# Contents

---

## 1. 기계학습의 소개

- 1) 기계학습이 실생활에서 쓰이는 곳
- 2) 기계학습이란 무엇인가?
  1. 기계학습 = 패턴 찾기
  2. 기계학습 vs 인공지능, 기계학습 vs 통계
- 3) **기계학습의 종류**
  1. 지도학습, 비지도학습, 강화학습
  2. **지도학습의 종류: 회귀, 분류**

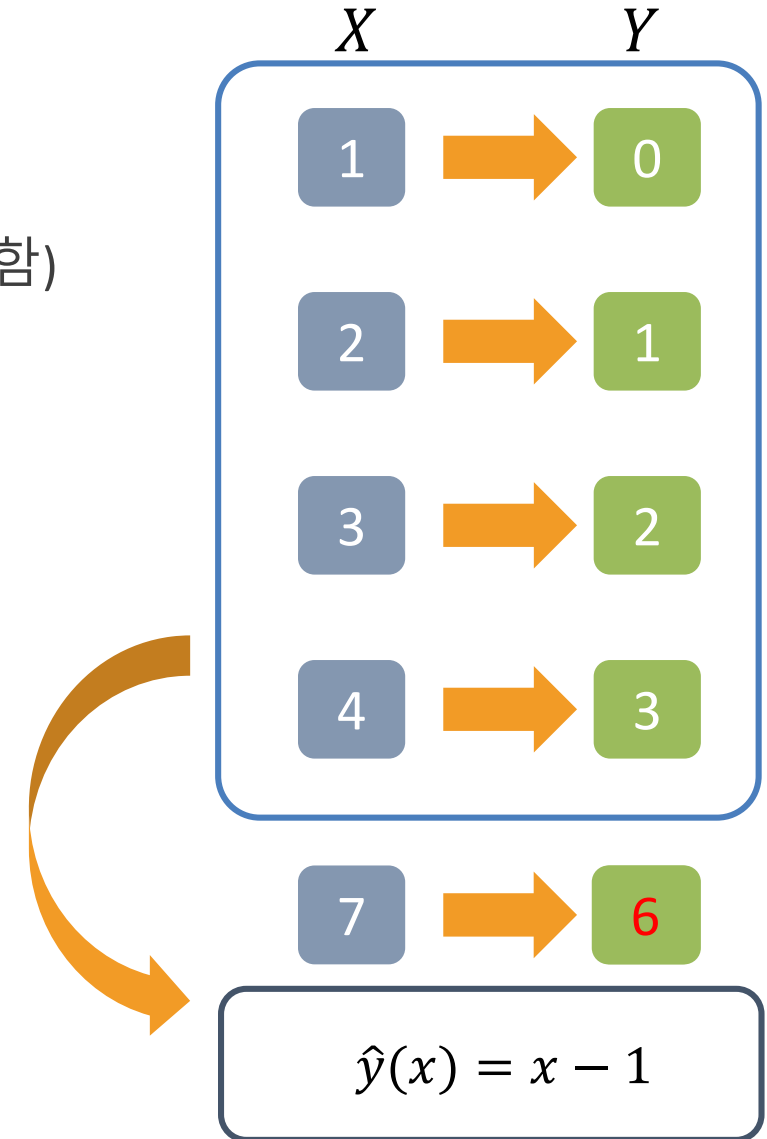
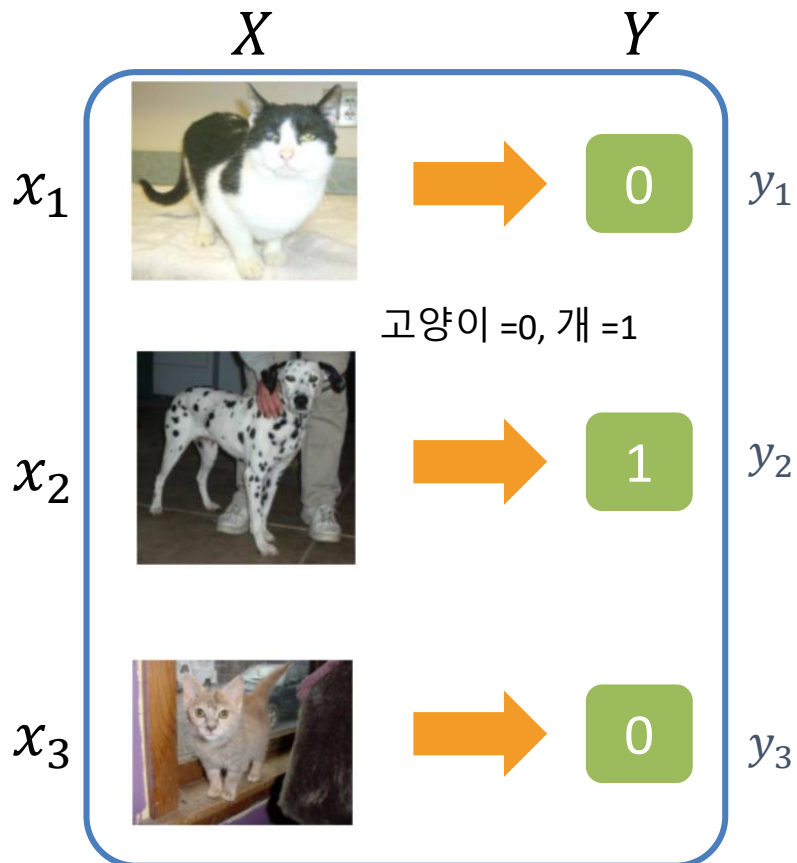
## 2. 회귀

- 1) 선형 회귀 분석 알고리즘 소개
- 2) 알고리즘 평가(evaluation) 방법 소개
- 3) 다중 선형 회귀 분석(multiple linear regression) 알고리즘 소개



# 지도 학습(Supervised Learning)

- 주어진 데이터  $(x_i, y_i)_{1 \leq i \leq n}$ 들의 관계를 보고
  - $\hat{y}(x_i) \approx y_i$ 가 되도록 추정함수  $\hat{y}(\cdot)$ 를 학습하는 것
  - $X$ : 독립변수 (Independent variable)
  - $Y$ : 종속변수 (Dependent variable) (정답 혹은 Label이라고도 함)



# 지도학습의 2가지 종류

- 회귀 분석(Regression) (Part 1)
  - 목표로 하는  $Y$ 가 연속적일 때 (continuous)



## Regression

What is the temperature going to be tomorrow?

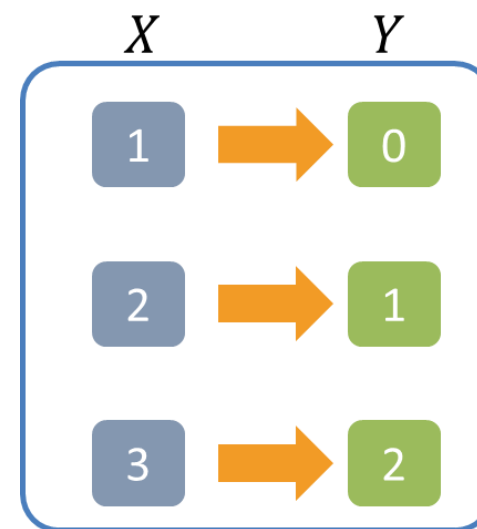
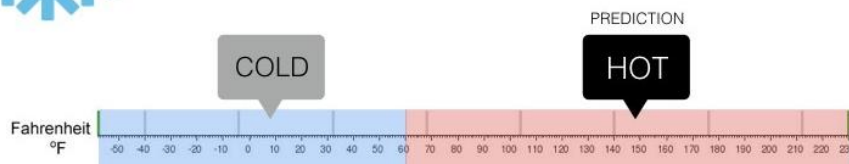


- 분류(Classification) (Part 2)
  - 목표로 하는  $Y$ 가 이산적일 때 (discrete, categorical)



## Classification

Will it be Cold or Hot tomorrow?



$Y$ 는 임의의 실수가 될 수 있다.



$Y$ 는 0 또는 1만 가능

# Contents

---

## 1. 기계학습의 소개

- 1) 기계학습이 실생활에서 쓰이는 곳
- 2) 기계학습이란 무엇인가?
  1. 기계학습 = 패턴 찾기
  2. 기계학습 vs 인공지능, 기계학습 vs 통계
- 3) 기계학습의 종류
  1. 지도학습, 비지도학습, 강화학습
  2. 지도학습의 종류: 회귀, 분류

## 2. 회귀 분석

- 1) **회귀 문제 소개: 회귀란, 회귀 문제의 종류, 회귀 분석에 사용되는 알고리즘**
- 2) 선형 회귀 분석 알고리즘 소개
- 3) 알고리즘 평가(evaluation) 방법 소개
- 4) 다중 선형 회귀 분석(multiple linear regression) 알고리즘 소개

- $X$ : 독립 변수 (independent variable)
  - 지능지수, 나이, 성별
- $Y$ : 종속 변수 (dependent variable), 정답 (label)
  - 연소득
  - continuous하다
- 목표
  - 주어진 (지능지수, 나이, 성별)에 대해서
  - 예상 연소득 값을 예측하자

IQ	나이	성별	연봉
107	22	여	6305
95	23	여	5730
114	57	여	8735
83	55	여	6735
101	21	여	6170
119	26	남	7805
92	22	남	6205
108	59	남	8830
129	44	남	9075
104	45	남	7935
94	20	남	6415
112	26	남	??

# 회귀 분석 문제의 종류

- 단순 회귀 분석 (Simple Regression)
  - $X$ 가 1차원인 경우
  - ex) 지능지수  $\rightarrow$  연소득 or 나이  $\rightarrow$  연소득 예측
- 다중 회귀 분석 (Multiple Regression)
  - $X$ 가 2차원 이상인 경우
  - ex) (지능지수, 나이)  $\rightarrow$  연소득

IQ	나이	성별	연봉
107	22	여	6305
95	23	여	5730
114	57	여	8735
83	55	여	6735
101	21	여	6170
119	26	남	7805
92	22	남	6205
108	59	남	8830
129	44	남	9075
104	45	남	7935
94	20	남	6415
112	26	남	??

- 선형 회귀(Linear regression)



**Part 1**

- 인공신경망 회귀(Neural network regression)
  - 딥러닝 (Deep Learning)의 한 종류



**Part 5**

- 다항 함수 회귀(Polynomial regression), 베이시안 선형 회귀(Bayesian linear regression), 푸아송 회귀(Poisson regression)

# Contents

---

## 1. 기계학습의 소개

- 1) 기계학습이 실생활에서 쓰이는 곳
- 2) 기계학습이란 무엇인가?
  1. 기계학습 = 패턴 찾기
  2. 기계학습 vs 인공지능, 기계학습 vs 통계
- 3) 기계학습의 종류
  1. 지도학습, 비지도학습, 강화학습
  2. 지도학습의 종류: 회귀, 분류

## 2. 회귀 분석

- 1) 회귀 문제 소개
- 2) **선형 회귀 분석 알고리즘 소개**
- 3) 알고리즘 평가(evaluation) 방법 소개
- 4) 다중 선형 회귀 분석(multiple linear regression) 알고리즘 소개



# 단순 선형 회귀 분석(Simple Linear Regression)

- 지능지수( $X$ )와 연소득( $Y$ )에 선형 관계가 있다고 가정

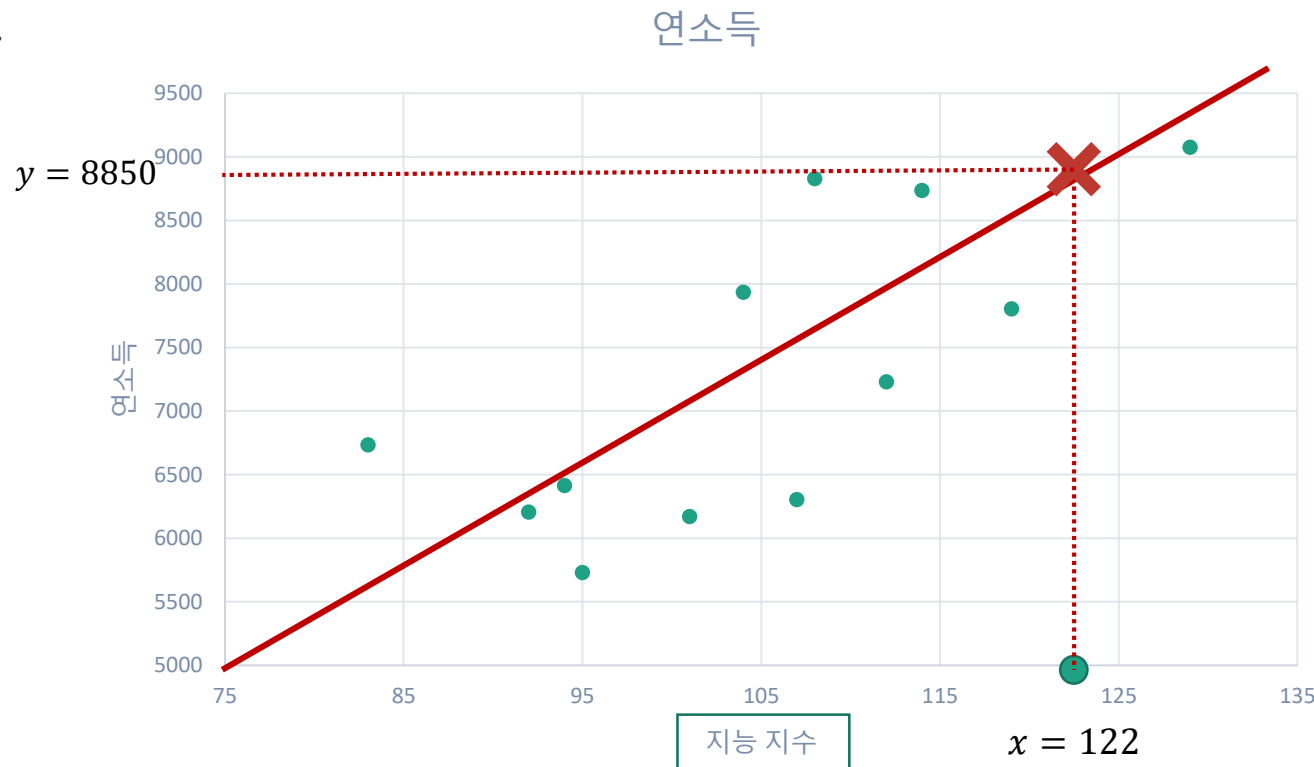
- $Y \approx \theta_0 + \theta_1 X$

- $F(X) = \theta_0 + \theta_1 X$ 에서

- 적절한  $\theta_0$  (절편),  $\theta_1$  (기울기)를 찾은 후,

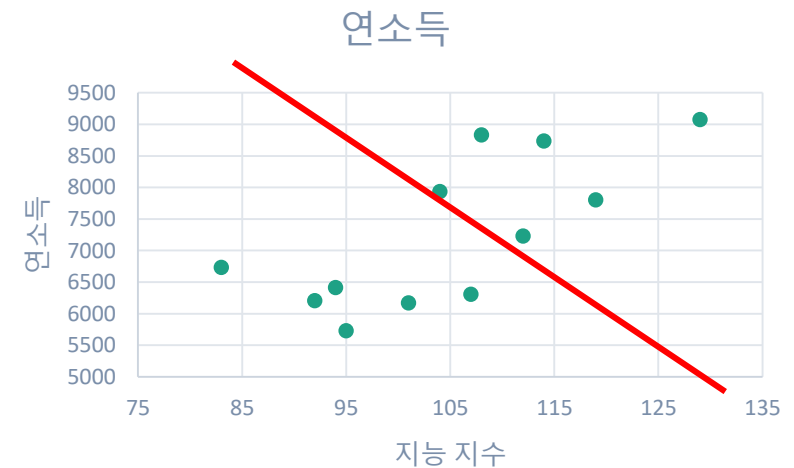
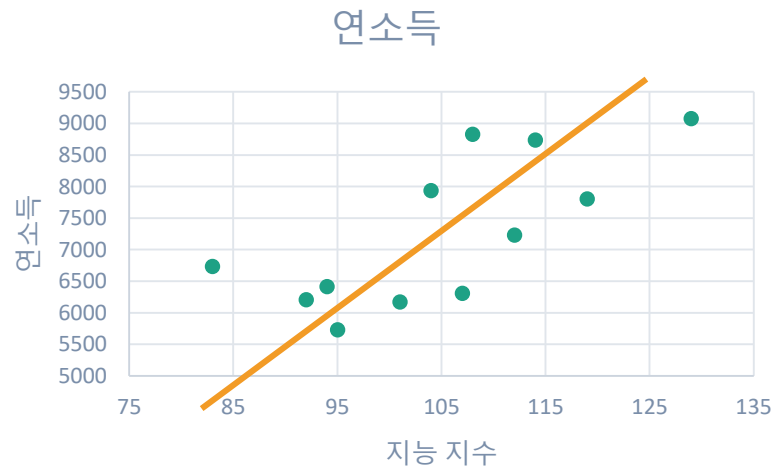
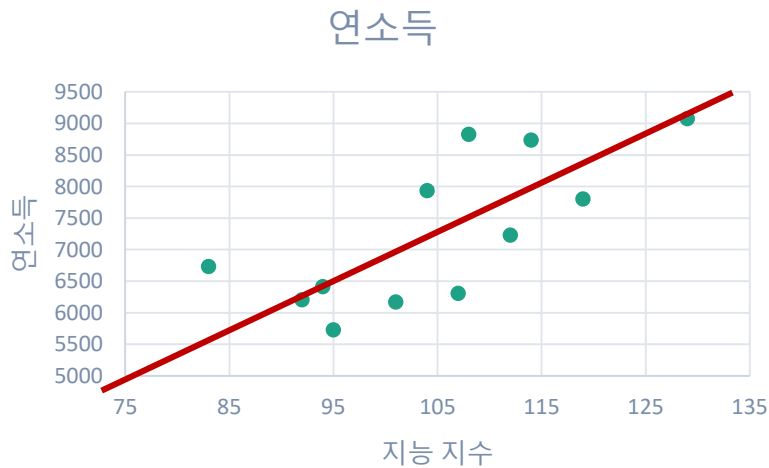
- 지능지수가 122 일 때 예상되는 연소득 값?

- $\hat{Y} = \theta_0 + \theta_1 X = 8850$ 으로 예측



# 좋은 $\theta_0, \theta_1$ 이란 어떤 의미일까?

- 선형 회귀 분석에서는 적절한  $\theta_0, \theta_1$ 을 찾는 것이 중요!
- 여러가지 후보들이 존재
- 적절함의 기준이 필요하다
  - 빨강 → 경향과 맞지 않는다, 주황 or 갈색 중에서는 누가 더 좋을까?



# 오차(Error)와 평균 제곱 오차(Mean Squared Error) LANADA

32

- 오차 =  $y - \hat{y}(x)$

- $\hat{y}$ : 예측 값
- $y$ : 실제 참 값 ( $\hat{y} := \theta_0 + \theta_1 x$ )

- $x = 101$ 일 때

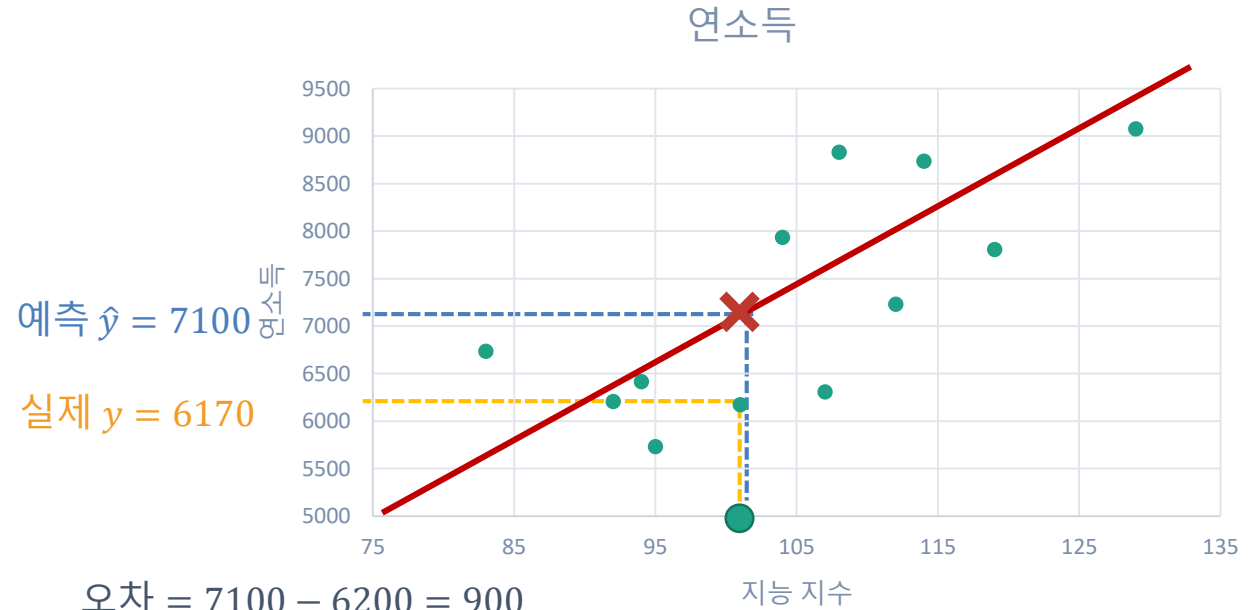
- $\hat{y}$ : 예측 값 = 7100
- $y$ : 실제 참 값 = 6170
- 오차 = -930

- (적절함의 기준) MSE를 줄이는  $\theta_0, \theta_1$ 을 찾자!

- $MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$
- $\hat{y}_i = \hat{y}(x_i)$

- (심화) 하필 왜 MSE가 적절한 기준이 되는가?

- $MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$  를 사용할 수도 있을텐데..?



MAE도 기준으로 사용할 수 있다.  
다만, 일반적인 가정에서 MSE가 **가장 좋은 기준**  
이라는 것이 이론적으로 증명됨!

# $\theta_0, \theta_1$ 의 계산

- [목표]  $MSE = \frac{1}{n} \sum_{i=1}^n (\theta_0 + \theta_1 x_i - y_i)^2$ 를 가장 작게 하는  $\theta_0, \theta_1$ 을 찾자!

- [답] (통계 기법)

- $\theta_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = 69.08$ 
  - $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{11} (107 + 95 + \dots + 94) = 104.2$
  - $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i = \frac{1}{11} (4020 + 3660 + \dots + 4750) = 7267.3$
- $\theta_0 = \bar{y} - \theta_1 \cdot \bar{x}$
- $\theta_0 = 7267.3 - 69.08 \cdot 104.2 = 69.2$

$$\hat{y}(x) = 69.2 + 69.08x$$

IQ	나이	성별	연봉
$x_1 =$ 107	22	여	$y_1 =$ 6305
$x_2 =$ 95	23	여	$y_2 =$ 5730
$x_3 =$ 114	57	여	$y_3 =$ 8735
$x_4 =$ 83	55	여	$y_4 =$ 6735
101	21	여	6170
...	26	남	...
...	26	남	7805
92	22	남	6205
108	59	남	8830
129	44	남	9075
104	45	남	7935
$x_n =$ 94	20	남	$y_n =$ 6415

# 선형 회귀 분석을 통한 실제 예측

- 지능지수가 112인 사람의 예측 연 소득

- $\hat{y}(x) = \theta_0 + \theta_1 x$

- $\hat{y}(x) = 69.2 + 69.08 \cdot x$

- $\hat{y}(112) = 69.2 + 69.08 \cdot 112 = 7806$

IQ	나이	성별	연봉
107	22	여	6305
95	23	여	5730
114	57	여	8735
83	55	여	6735
101	21	여	6170
119	26	남	7805
92	22	남	6205
108	59	남	8830
129	44	남	9075
104	45	남	7935
94	20	남	6415
112	26	남	??

# 선형 회귀 분석 ( $Y = \theta_0 + \theta_1 X$ )의 장단점

- 장점:

- 단순한 만큼 쉽게 파라미터인  $\theta_0, \theta_1$ 이 계산 가능

- 생각보다 많은 종류의 문제에서 무난하게 동작

$$\begin{aligned} \bullet \theta_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ \bullet \bar{x} &= \frac{1}{n} \sum_{i=1}^n x_i, \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \\ \bullet \theta_0 &= \bar{y} - \theta_1 \cdot \bar{x} \end{aligned}$$

파라미터 계산 공식

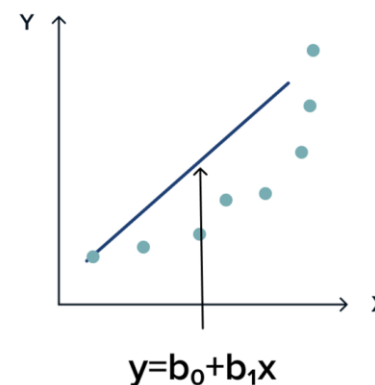
- 단점

- 기본적으로 선형만 고려하기에, 주어진 데이터의 분포가 선형이 아닐 경우 어떤  $\theta_0, \theta_1$ 로도 잘 fit하지 않게 된다.

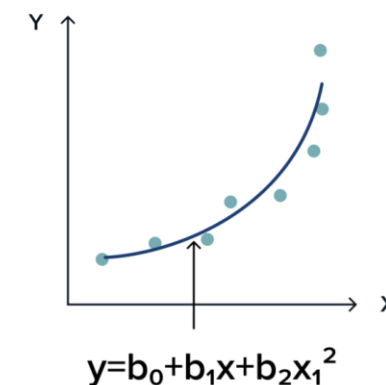
- 이 경우 비선형 회귀 분석 (Non-linear regression)을 진행

- ex) 다항 회귀 분석(Polynomial regression)

Simple linear model



Polynomial model



선형 회귀모델이 잘 동작 안 하는 경우

# (목차) 모델 평가(Model Evaluation)

- 모델 평가(model evaluation)
- 학습 데이터 셋(train data set)과 검증 데이터 셋(test data set)
- K-fold 교차검증(K-fold cross validation) 기법
- 여러가지 평가지표들(evaluation metrics)



# Contents

---

## 1. 기계학습의 소개

- 1) 기계학습이 실생활에서 쓰이는 곳
- 2) 기계학습이란 무엇인가?
  1. 기계학습 = 패턴 찾기
  2. 기계학습 vs 인공지능, 기계학습 vs 통계
- 3) 기계학습의 종류
  1. 지도학습, 비지도학습, 강화학습
  2. 지도학습의 종류: 회귀, 분류

## 2. 회귀 분석

- 1) 회귀 문제 소개
- 2) 선형 회귀 분석 알고리즘 소개
- 3) **알고리즘 평가(evaluation) 방법 소개**
- 4) 다중 선형 회귀 분석(multiple linear regression) 알고리즘 소개

# 모델 평가(Model Evaluation)

- 학습된 모델은 **얼마만큼 정확한가?**
- 학습이 완료된 모델의 예측을 **얼마만큼 신뢰할 수 있는가?**
- 다른 말로, 학습된 모델의 **전반적인 오차의 크기**는 얼마인가?

지능지수	연봉
107	6305
95	5730
114	8735
83	6735
101	6170
119	7805
92	6205
108	8830
129	9075
104	7935
94	6415

데이터 셋

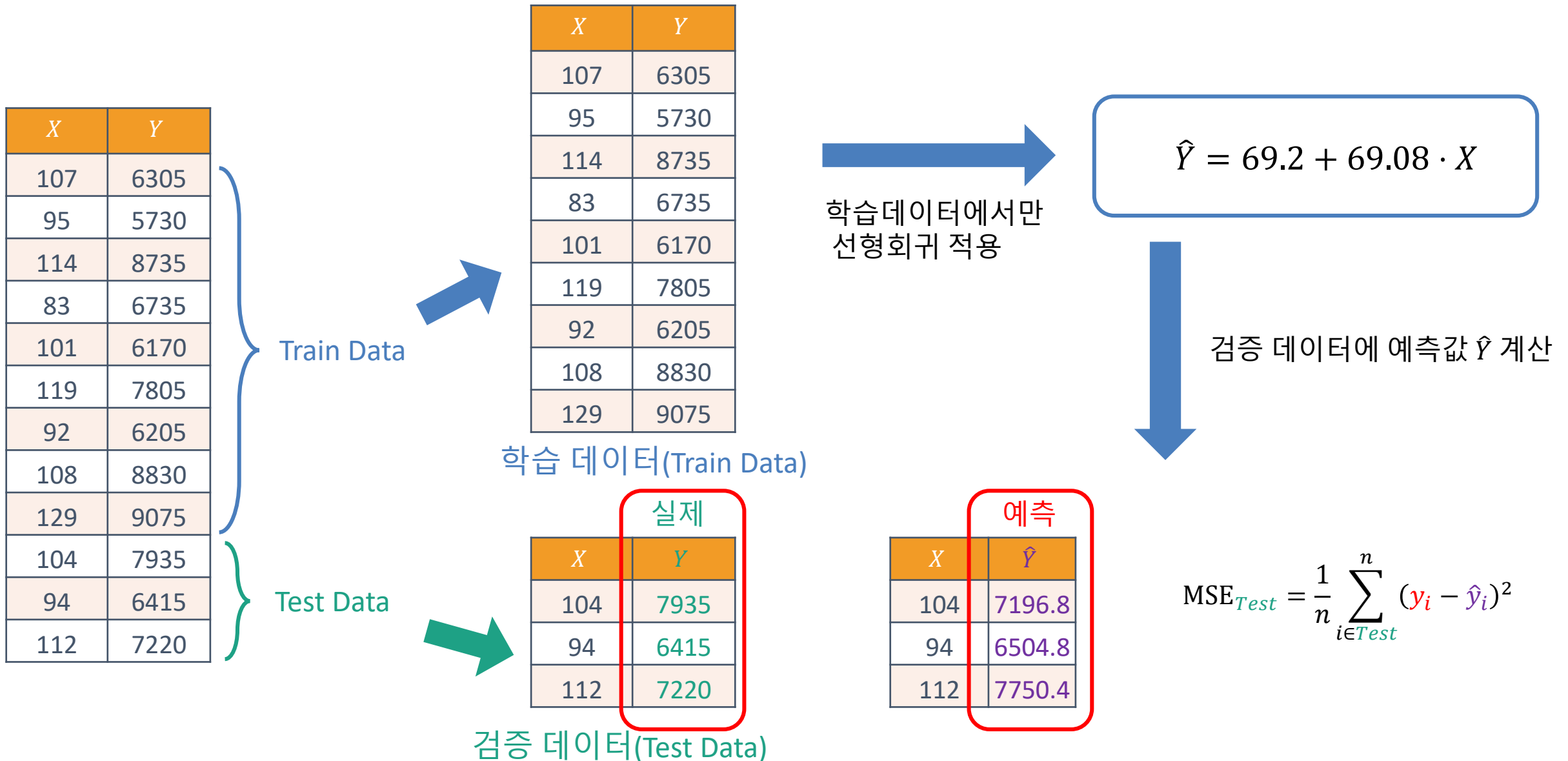
선형 회귀 학습

$$\hat{Y} = 69.2 + 69.08 \cdot X$$

예측이 100% 정확하리라 믿을 수는 없다...  
그렇다면 얼마나 정확할까?

# 학습 데이터(Train Data) vs 검증 데이터(Test Data) LANADA

39



# 학습 데이터(Train Data) vs 검증 데이터(Test Data)

40

- 전체 데이터를 두 종류로 나눔
  - 학습 데이터(Train Data)
    - 모델을 학습하는 데 사용되는 데이터
  - 검증 데이터(Test Data)
    - 학습이 완료된 모델을 검증하는 데 사용되는 데이터
- 학습 정확도(Training accuracy)
  - $MSE_{Train} = \frac{1}{n} \sum_{i \in Train} (\hat{y}_i - y_i)^2$
- 검증 정확도(Test accuracy)
  - $MSE_{Test} = \frac{1}{n} \sum_{i \in Test} (\hat{y}_i - y_i)^2$

Train Data	IQ	나이	성별	연봉
	107	22	여	6305
	95	23	여	5730
	114	57	여	8735
	83	55	여	6735
	101	21	여	6170
	119	26	남	7805
	92	22	남	6205
	108	59	남	8830
Test Data	129	44	남	9075
	104	45	남	7935
	94	20	남	6415
	112	26	남	7220

검증 정확도가 높다는 것은 처음 보는 (학습에 사용x) 데이터의 예측이 정확하다는 것  
**검증 정확도를 전체 모델의 정확도로 활용한다!**

# 검증 데이터를 정하는 기준

- 일반적으로 전체 데이터 중 랜덤하게 다음 비율로 선택한다
  - 80%는 학습용
  - 20%는 검증용
- 통계와 마찬가지로 기계학습에서도 학습데이터가 많을 수록 정확해짐
- 반대로 검증용 데이터가 너무 적으면, 검증 정확도의 variance가 지나치게 커질 수 있음
- 상황에 따라 달라질 수 있다...

Train Data

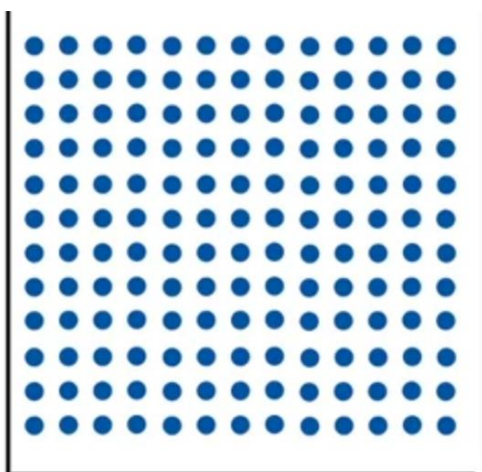
Test Data

IQ	나이	성별	연봉
107	22	여	6305
95	23	여	5730
114	57	여	8735
83	55	여	6735
101	21	여	6170
119	26	남	7805
92	22	남	6205
108	59	남	8830
129	44	남	9075
104	45	남	7935
94	20	남	6415
112	26	남	7220

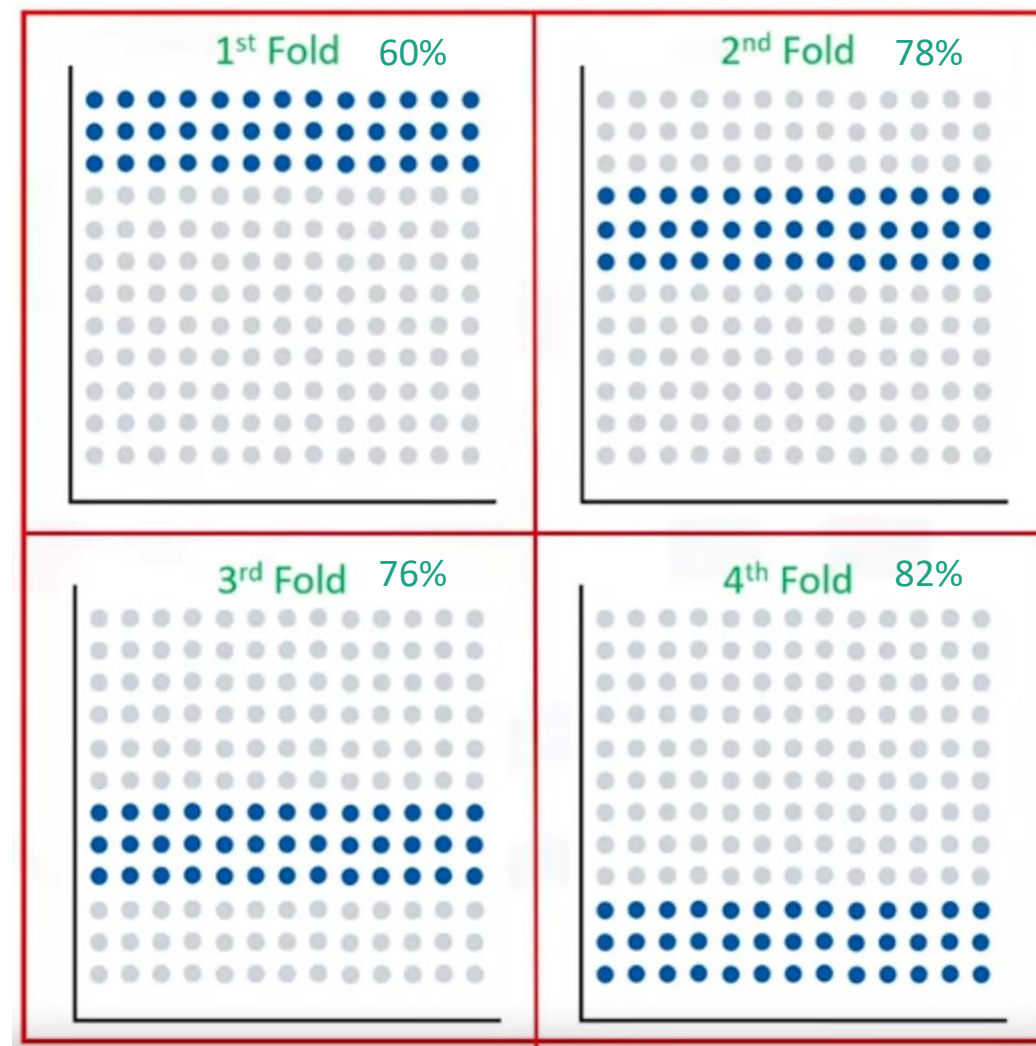
# K-fold 교차검증(K-fold cross validation) 기법

- 데이터를 K등분하고,
  - 그중 하나를 검증데이터로 사용한다
  - 나머지 K-1/K의 데이터는 학습데이터로
  - 이 과정을 K번 반복

- Ex) K=4인 경우



- :검증 데이터 셋
- :학습 데이터셋



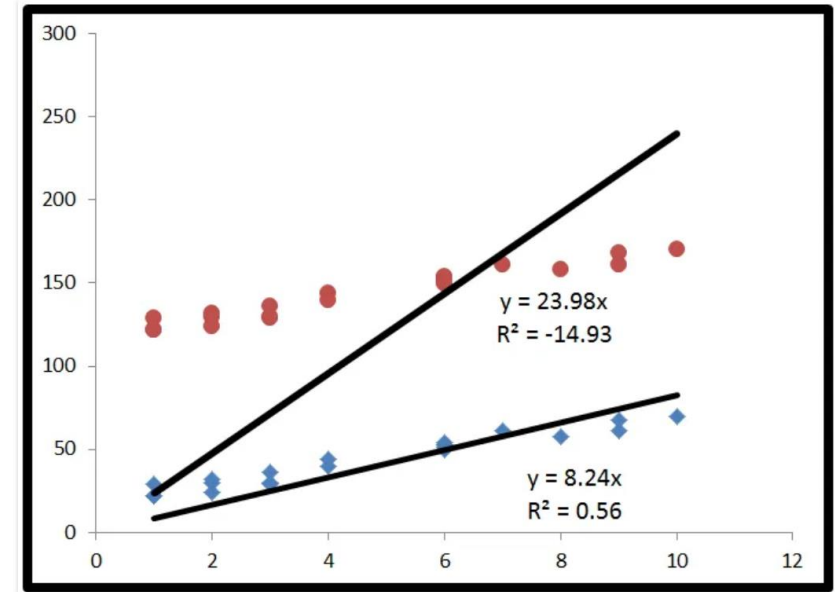
$$\text{평균 test accuracy} = \frac{1}{4} (60 + 78 + 76 + 82) = 74\%$$

# MSE를 제외한 다른 종류의 정확도 측정

- 검증 정확도(Test accuracy)는 여러가지 다른 척도(metric)이 사용될 수 있다

- $MSE_{Test} = \frac{1}{n} \sum_{i \in Test} (\hat{y}_i - y_i)^2$  (Mean squared error) (L2 loss)
- $MAE_{Test} = \frac{1}{n} \sum_{i \in Test} |\hat{y}_i - y_i|$  (Mean absolute error) (L1 loss)
- $RMSE_{Test} = \sqrt{\frac{1}{n} \sum_{i \in Test} (\hat{y}_i - y_i)^2}$  (Root mean absolute error)
  - 일반적으로 가장 많이 쓰이는 방식
  - $RMSE = \sqrt{MSE}$ 이므로 본질적으로 MSE를 쓰는 것과 같다
    - 통계에서 분산(variance) 대신에 표준편차(deviation)를 사용하는 경우와 유사

- $RSE_{Test} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$  (Relative squared error)
- $R^2_{Test} = 1 - RSE_{Test}$  (R squared error)
  - 값이 1에 가까울 수록 정확함



$R^2$  예시: 파란색의  $R^2$ 가 훨씬 1에 가깝다

목표에 따라서 적절한 metric 사용 필요!

# Contents

---

## 1. 기계학습의 소개

- 1) 기계학습이 실생활에서 쓰이는 곳
- 2) 기계학습이란 무엇인가?
  1. 기계학습 = 패턴 찾기
  2. 기계학습 vs 인공지능, 기계학습 vs 통계
- 3) 기계학습의 종류
  1. 지도학습, 비지도학습, 강화학습
  2. 지도학습의 종류: 회귀, 분류

## 2. 회귀 분석

- 1) 회귀 문제 소개
- 2) 선형 회귀 분석 알고리즘 소개
- 3) 알고리즘 평가(evaluation) 방법 소개
- 4) 다중 선형 회귀 분석(multiple linear regression) 알고리즘 소개



# 다중 선형 회귀 분석(Multiple Linear Regression)

- 단순 선형 회귀 분석 (여태까지 배운 것)
  - 하나의 독립변수  $x$ 로 종속 변수  $y$ 를 예측한다
    - $F: IQ \rightarrow \text{연봉}$
    - $F: \text{나이} \rightarrow \text{연봉}$
- 다중 선형 회귀 분석
  - 동시에 여러 개의 독립변수  $X_1, X_2, \dots, X_d$ 를 활용하여  $Y$  예측
    - $F: (IQ, \text{나이}, \text{성별}) \rightarrow \text{연봉}$
- 일반적으로 다중 선형 회귀 분석이 더 정확!
  - 더 많은 종류의 정보를 활용해서 예측하기에

IQ	나이	성별	연봉
107	22	여	6305
95	23	여	5730
114	57	여	8735
83	55	여	6735
101	21	여	6170
119	26	남	7805
92	22	남	6205
108	59	남	8830
129	44	남	9075
104	45	남	7935
94	20	남	6415
112	26	남	??

# 다중 선형 회귀 분석(Multiple Linear Regression)

- 전반적으로 단순 선형 분석과 매우 유사함

- 연봉 =  $\theta_0 + \theta_1 \cdot IQ + \theta_2 \cdot \text{나이} + \theta_3 \cdot \text{성별}$

- $\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3$

- $\Leftrightarrow \hat{y} = W^T X$

$$\begin{aligned} & \bullet W^T = [\theta_0, \theta_1, \dots, \theta_3] \quad X = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_3 \end{bmatrix} \end{aligned}$$

- 적절한 매개변수  $W^T = [\theta_0, \theta_1, \dots, \theta_3]$ 를 찾는 것이 목표



IQ	나이	성별	연봉
107	22	여	6305
95	23	여	5730
114	57	여	8735
83	55	여	6735
101	21	여	6170
119	26	남	7805
92	22	남	6205
108	59	남	8830
129	44	남	9075
104	45	남	7935
94	20	남	6415
112	26	남	??

# 다중 선형 회귀 모델에서의 MSE

- 오차(error)를 단순 선형 회귀 모델과 같은 식으로 정의

- $error = y - \hat{y}$

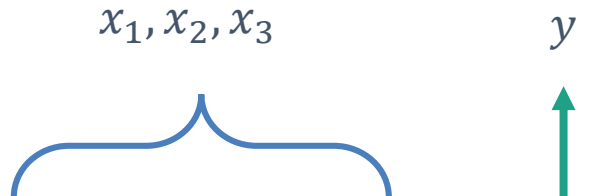
- 예시 (4번째 사람)

- $\hat{y}_4 = W^T X_4 = 8760$
  - $y_4 = 8735$

- $error = y_4 - \hat{y}_4 = 8735 - 8760 = -25$

- $MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$ 
  - 선형 모델과 마찬가지로 정의

- MSE를 가장 낮추는  $W^T = (\theta_0, \theta_1, \dots, \theta_3)$ 를 찾자



	IQ	나이	성별	연봉	
	107	22	여	6305	
	95	23	여	5730	
	83	55	여	6735	
$X_4$	114	57	여	8735	$y_4$
	101	21	여	6170	
	119	26	남	7805	
	92	22	남	6205	
	108	59	남	8830	
	129	44	남	9075	
	104	45	남	7935	
	94	20	남	6415	

# 최소 자승법(Ordinary Least Squares): $\Theta$ 계산하기

•  $MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \frac{1}{n} \|Y - MW\|^2$

1	107	22	0
1	95	23	0
1	114	57	0
1	83	55	0
1	101	21	0
1	119	26	1
1	92	22	1
1	108	59	1
1	129	44	1
1	104	45	1
1	94	20	1

$M: n \times (d + 1)$  행렬

$\theta_0$
$\theta_1$
$\theta_2$
$\theta_3$

$W: (d + 1) \times 1$  행렬

6305
5730
8735
6735
6170
7805
6205
8830
9075
7935
6415

$Y: n \times 1$  행렬

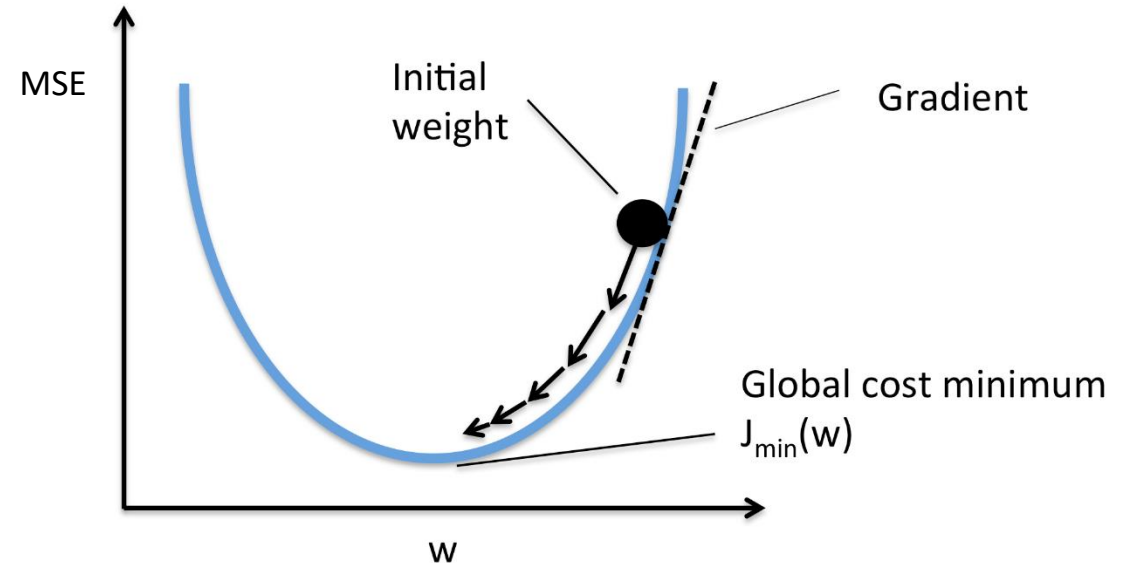
IQ	나이	성별	연봉
107	22	여	6305
95	23	여	5730
114	57	여	8735
83	55	여	6735
101	21	여	6170
119	26	남	7805
92	22	남	6205
108	59	남	8830
129	44	남	9075
104	45	남	7935
94	20	남	6415

$M, Y$ 로 쪼개서 표현

최소 자승법(Ordinary Least Squares)에 의해서  $MSE = \frac{1}{n} \|Y - MW\|^2$ 는  $W = (M^T M)^{-1} M^T Y$  일 때 최소가 된다.

# 경사 하강법(Gradient Descent)

- 최소 자승법의 장점
  - 100% 정확하다
- 최소 자승법의 단점
  - $W = (M^T M)^{-1} M^T Y$ 의 역행렬계산 과정에서 계산 비용 큼
- 경사 하강법(Gradient Descent)
  - 각 point에서 MSE를 낮추는 방향으로 조금씩 이동
  - 다소 부정확할 수 있지만, 효율적임
  - 깊은 신경망 학습에서도 사용됨
  - 자세한 설명은 Part 4에서 진행!



# 다중 선형 회귀 분석에서 예측 예시

- $W^T = (100, 50, 60, 70)$ 인 경우 실제 예측 계산
- $\hat{y} = W^T \cdot X$
- $\hat{y} = 100 \cdot 1 + 50x_1 + 60x_2 + 70x_3$
- $\hat{y} = 100 + 50 \cdot IQ + 60 \cdot \text{나이} + 70 \cdot (\text{남자일경우})$
- 4번째 사람의 예측값
  - $\hat{y}_4 = 100 \cdot 1 + 50 \cdot 114 + 60 \cdot 57 + 70 \cdot 0 = 9220$
- 오차 =  $y - \hat{y} = 8735 - 9220 = -85$

100
50
60
70

W:  $(d + 1) \times 1$  행렬

	IQ	나이	성별	연봉	
	107	22	여	6305	
	95	23	여	5730	
	83	55	여	6735	
$X_4$	114	57	여	8735	$y_4$
	101	21	여	6170	
	119	26	남	7805	
	92	22	남	6205	
	108	59	남	8830	
	129	44	남	9075	
	104	45	남	7935	
	94	20	남	6415	

---

# Thank you!

Any Questions?

---

- 1. 분류(Classification) 소개
  - 무엇을 하는 것인지? 회귀분석과의 차이점? 분류의 예시? 왜 중요한지? 분류 알고리즘 소개
- 2. K-NN 알고리즘(K-Nearest Neighbors)
  - 알고리즘 목표, 작동 원리, 좋은 k 찾기, 장단점
- 3. 서포트 벡터 머신(Support Vector Machine, SVM)
  - 알고리즘 목표, 작동 원리, 문제점, Kernel 함수, Fine-tuning, 장단점
- 4. 로지스틱 회귀 분석(Logistic regression)
  - 알고리즘 목표, 작동 원리, Sigmoid 함수, Log-loss (Cross entropy loss), 경사하강법(Gradient descent)
- 5. 분류 알고리즘에서의 성능 평가
  - Test accuracy, Precision and recall, F1-score, Cross entropy loss



## 2-1. 분류(Classification) 소개

# Contents

---

## 1. Classification (분류) 소개

- 1) 지도학습(Supervised learning) 복습
- 2) Classification의 문제 예시
- 3) Classification 주요 알고리즘 소개
- 4) 여러 개의 Label이 존재하는 경우

## 2. K-NN 알고리즘(K-nearest neighbor)

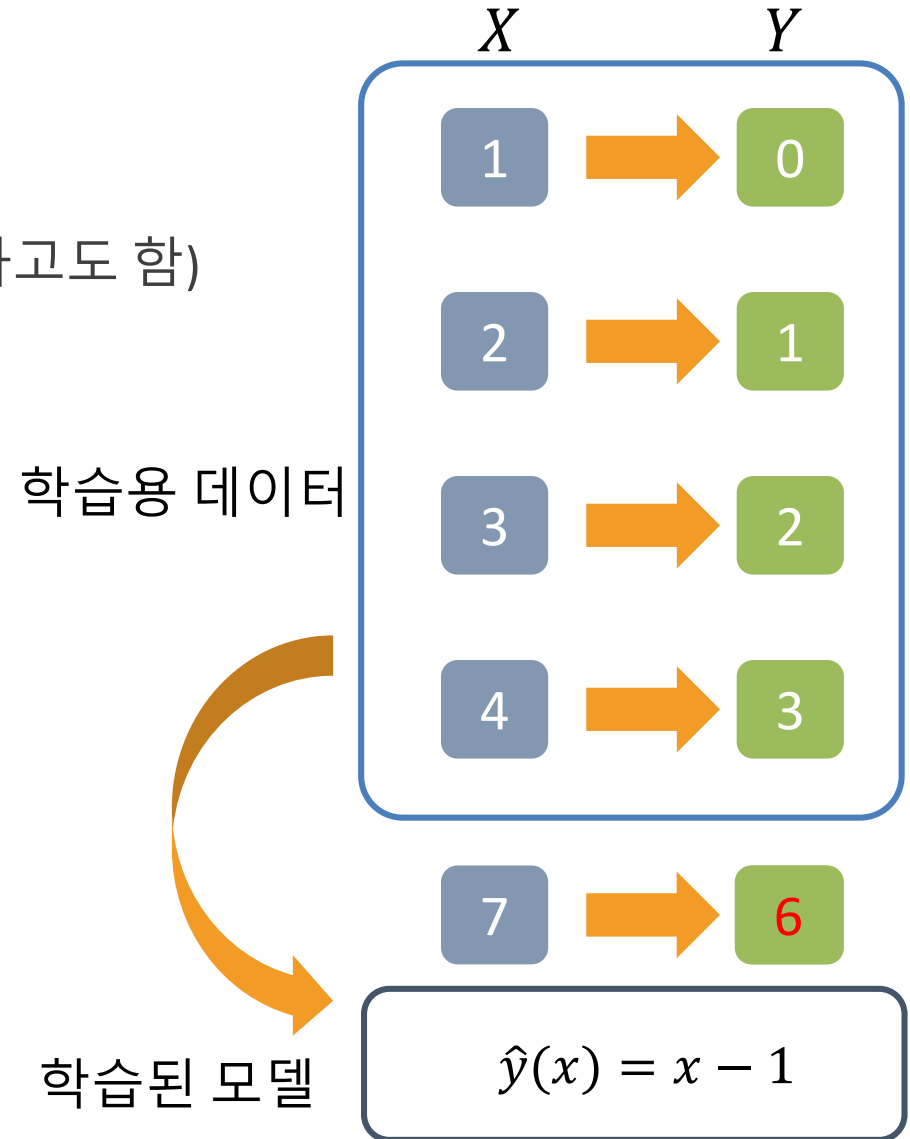
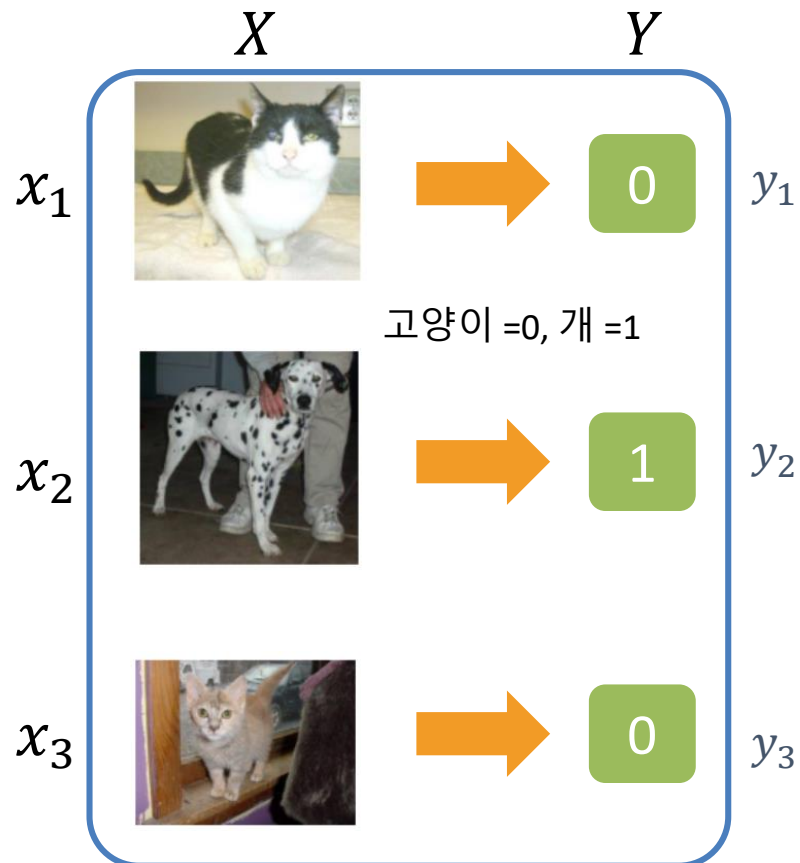
## 3. 서포트 벡터 머신(Support vector machine, SVM)

## 4. 로지스틱 회귀 분석(Logistic regression)

## 5. 알고리즘 성능 평가(Model Evaluation)

# 지도 학습(Supervised Learning)

- 주어진 데이터  $(x_i, y_i)_{1 \leq i \leq n}$ 들의 관계를 보고
  - $\hat{y}(x_i) \approx y_i$ 가 되도록 추정함수  $\hat{y}(\cdot)$ 를 학습하는 것
  - $X$ : 독립변수 (Independent variable)
  - $Y$ : 종속변수 (Dependent variable) (정답 혹은 Label이라고도 함)



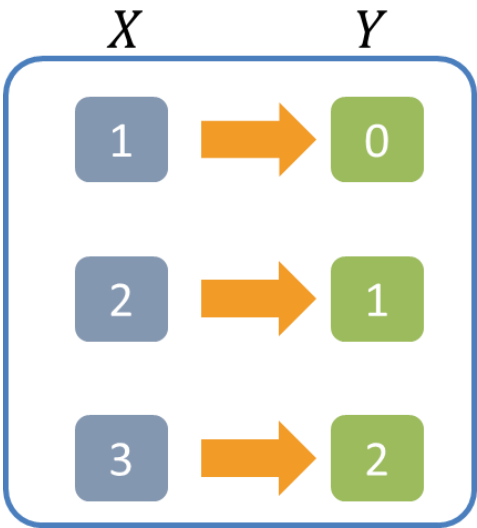
# 지도 학습의 2가지 종류

- 회귀 분석(Regression) (Part 1)
  - 목표로 하는  $Y$ 가 연속적일 때 (continuous)



## Regression

What is the temperature going to be tomorrow?



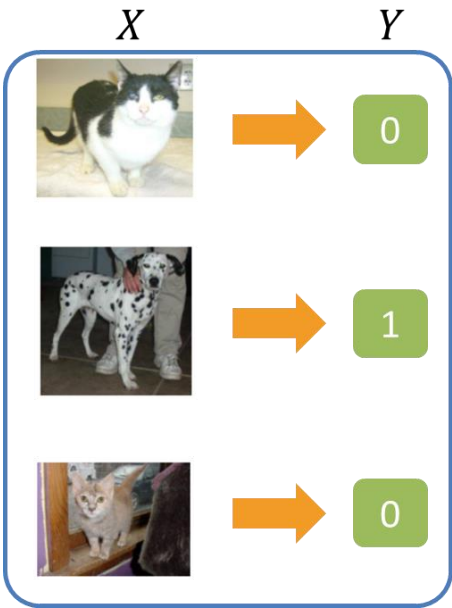
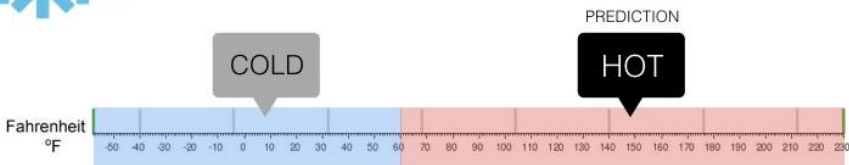
$Y$ 는 임의의 실수가 될 수 있다.

- 분류(Classification) (Part 2)
  - 목표로 하는  $Y$ 가 이산적일 때 (discrete, categorical)



## Classification

Will it be Cold or Hot tomorrow?



$Y$ 는 0 또는 1만 가능

# Contents

---

## 1. Classification (분류) 소개

- 1) 지도학습(Supervised learning) 복습
- 2) **Classification의 문제 예시**
- 3) Classification 주요 알고리즘 소개

## 2. K-NN 알고리즘(K-nearest neighbor)

## 3. 서포트 벡터 머신(Support vector machine, SVM)

## 4. 로지스틱 회귀 분석(Logistic regression)

## 5. 알고리즘 성능 평가(Model Evaluation)

# 분류(Classification) 예시

- $X$ : 독립 변수 (independent variable)
  - 지능지수, 연봉
- $Y$ : 종속 변수 (dependent variable), 정답 (label)
  - 성별 or 나이
    - 성별: 2가지 경우만 있다 → **이산 분류(binary classification)**
    - 나이: 여러가지 후보 있음 → **다중 레이블 분류 (multi-label classification)**
  - 이산적이다(discrete)
- 목표
  - 주어진 (지능지수, 연봉, 성별)에 대해서
  - 예상 성별 or 나이를 예측하자

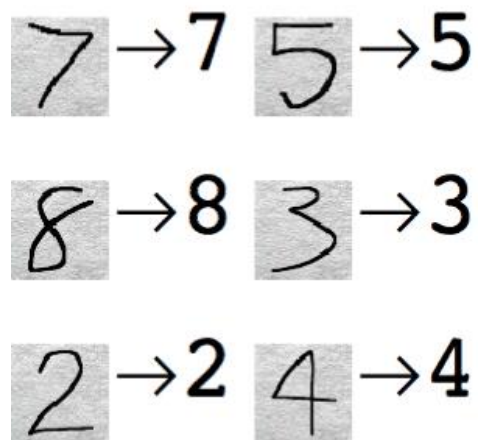
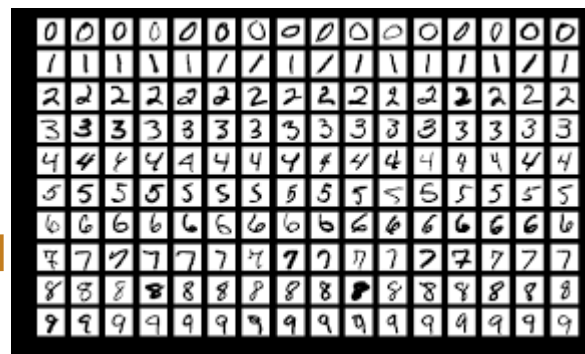
$X$ : 독립		$Y$ : 종속	
IQ	연봉	나이	성별
107	6305	22	여
95	5730	23	여
114	8735	57	여
83	6735	55	여
101	6170	21	여
119	7805	26	남
92	6205	22	남
108	8830	59	남
129	9075	44	남
104	7935	45	남
94	6415	20	남
112	7800	??	??

학습데이터 (known data)

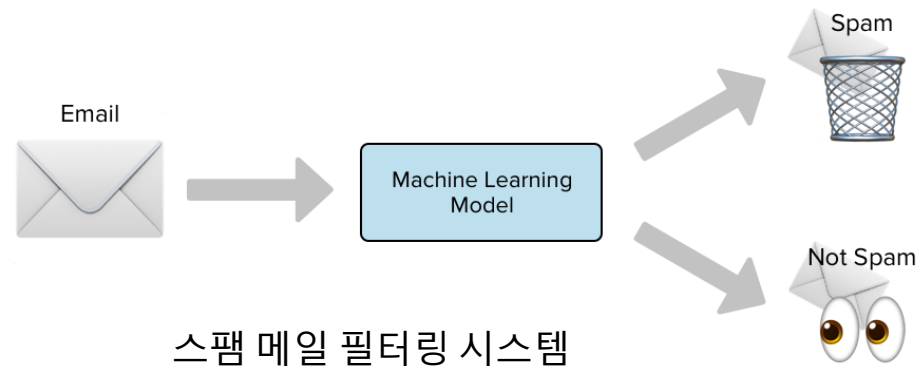
목표 데이터 (unknown data)

본 프레젠테이션에선 이산분류만 다루나, 모든 알고리즘은 다중 레이블에서도 확장 가능함!

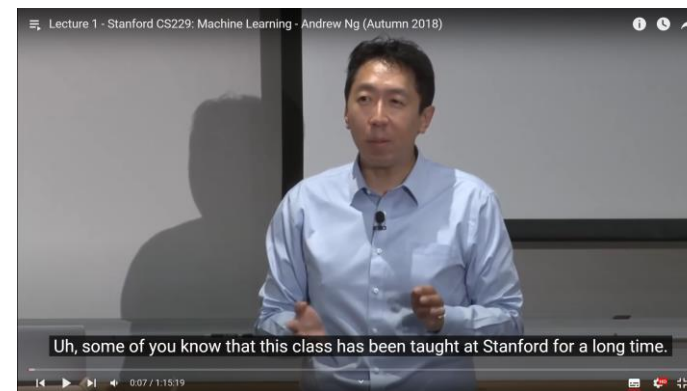
# 분류 알고리즘 실제 사례



이미지 분류 (MNIST)  
 $X$ : image,  $Y$ : 0,1,...,9



스팸 메일 필터링 시스템  
 $X$ : messages,  $Y$ : spam (1) or not spam (0)



음성 인식 후 자막 생성 알고리즘  
 $X$ : 목소리 음성 파일,  $Y$ : 각각의 영어단어

# Contents

---

## 1. Classification (분류) 소개

- 1) 지도학습(Supervised learning) 복습
- 2) Classification의 문제 예시
- 3) **Classification** 주요 알고리즘 소개

## 2. K-NN 알고리즘(K-nearest neighbor)

## 3. 서포트 벡터 머신(Support vector machine, SVM)

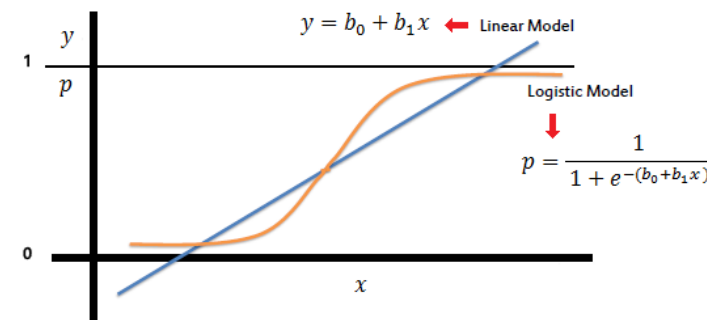
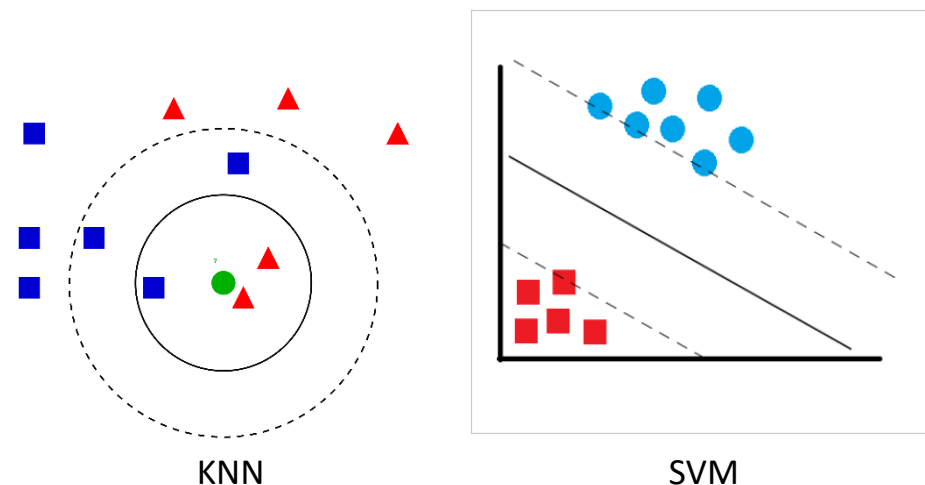
## 4. 로지스틱 회귀 분석(Logistic regression)

## 5. 알고리즘 성능 평가(Model Evaluation)



# Classification 주요 알고리즘 소개

- KNN 알고리즘(K-Nearest Neighbors)
  - Section 2
- 서포트 벡터 머신(Support Vector Machines, SVM)
  - Section 3
- 로지스틱 회귀 분석(Logistic Regression)
  - Section 4
- 의사결정 나무(Decision Tree)
  - Part 4
- 인공 신경망(Neural Network)
  - Part 5



Logistic Regression

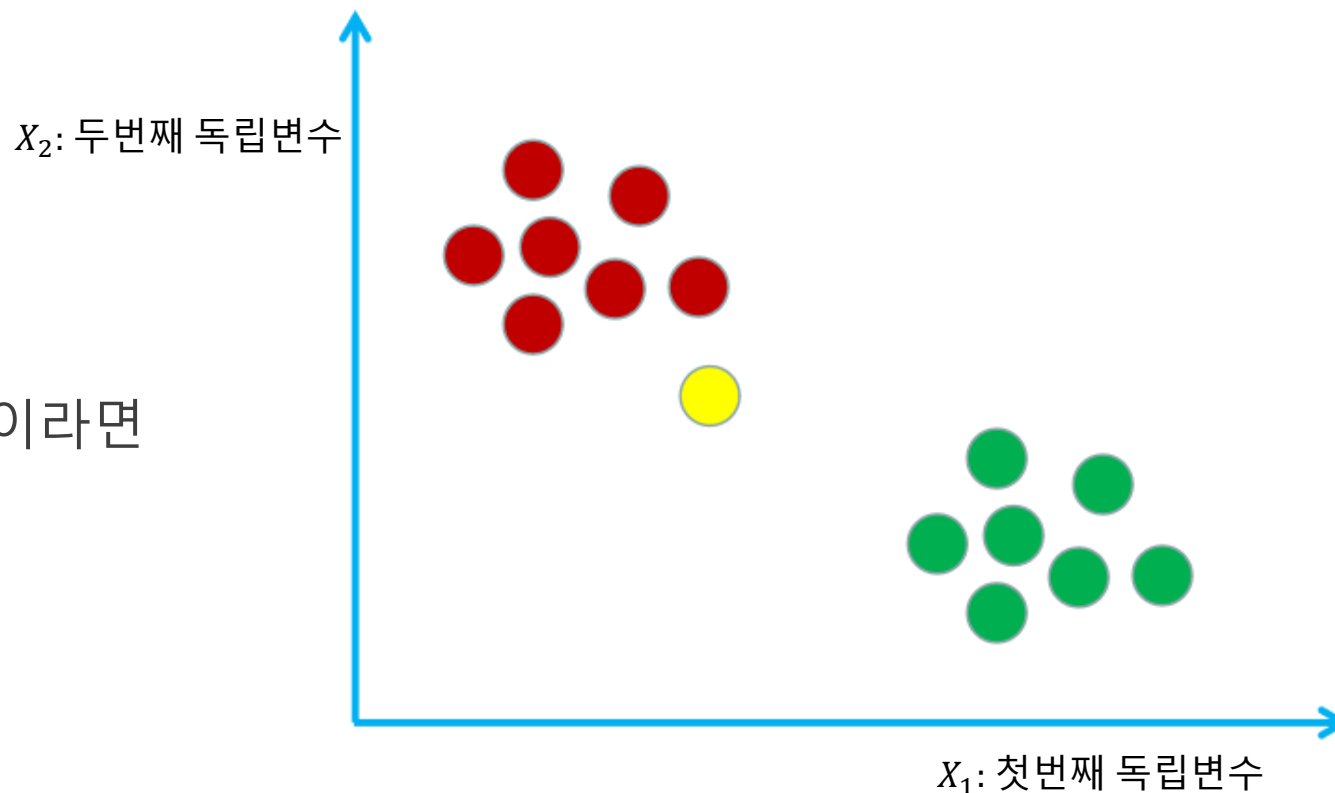
# Contents

---

1. Classification (분류) 소개
- 2. K-NN 알고리즘(K-nearest neighbors)**
  - 1) 알고리즘 아이디어 소개
  - 2) K 정하기
  - 3) KNN의 장단점
3. 서포트 벡터 머신(Support vector machine, SVM)
4. 로지스틱 회귀 분석(Logistic regression)
5. 알고리즘 성능 평가(Model Evaluation)

# Quiz: 노란색 원은 어느 클래스에 속할까?

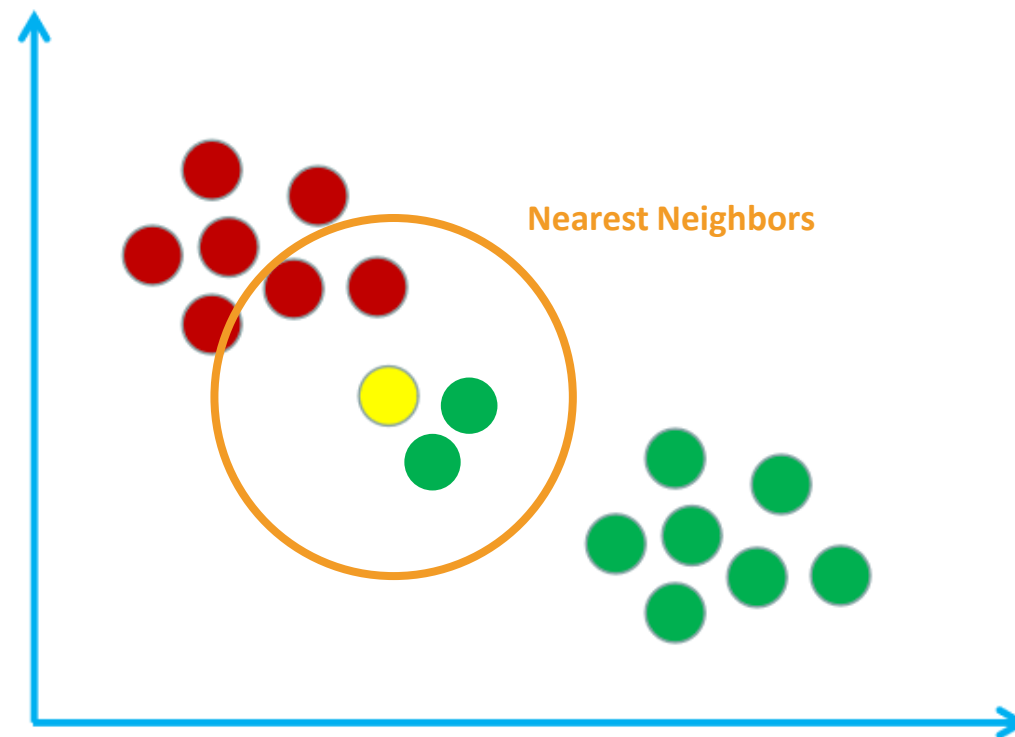
- 학습 데이터
  - 빨간색 원: class 1
  - 초록색 원: class 2
- 예측하고자 하는 원: 노란색
  - class 1 or class 2 둘 중에 하나의 색이라면
  - 무슨 색일까?



**직관적인 대답: 노란색원 → 빨강원일 가능성이 가능성이 높다!**

# 왜 그렇게 생각했을까?

- 노란색 원과 비슷한 성질을 가지고 있는
  - (근처에 있는) or (주황색 원안에 있는)
- 원들이 주로 빨강색이었기 때문에!



## KNN 알고리즘 작동 원리

1. 주어진 unknown 원을 중심으로 가까운 k개의 이웃 원들을 찾자 (K=3)
2. 이때 속하는 원들의 색이 노란색 원의 색이 될 것이다. (3개다 빨강)
3. 만약에 색이 같된다면? → 다수결 (빨강 1, 초록 2) → **초록**

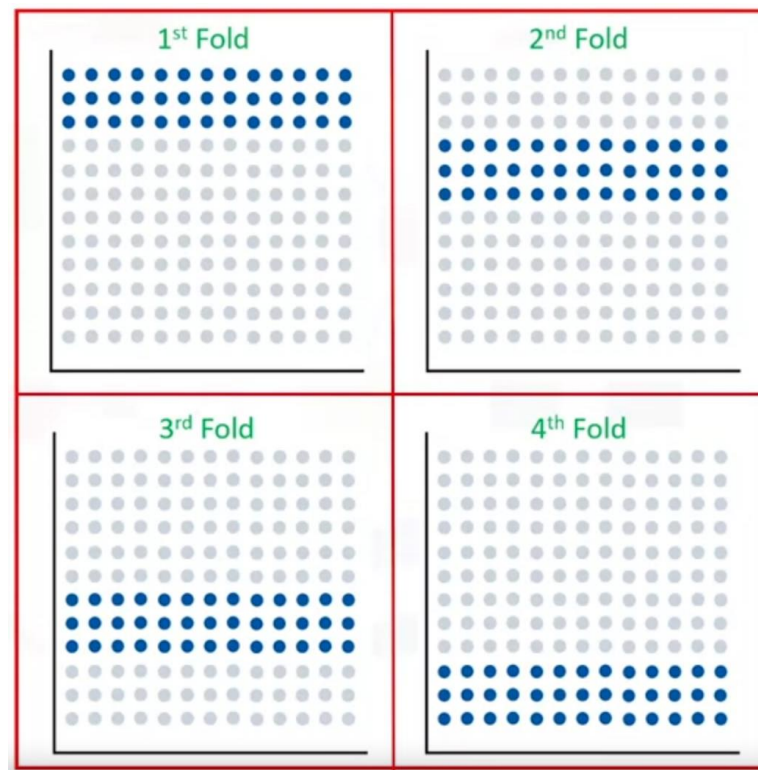
# Contents

---

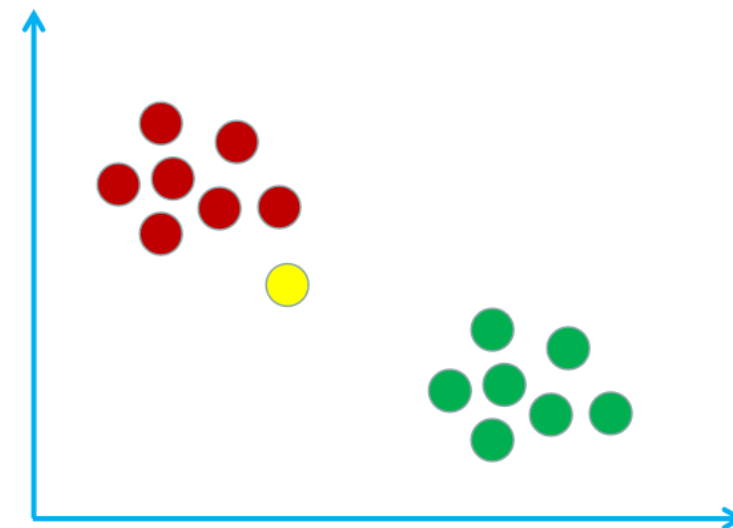
1. Classification (분류) 소개
- 2. K-NN 알고리즘(K-nearest neighbor)**
  - 1) 알고리즘 아이디어 소개
  - 2) k 정하기**
  - 3) KNN의 장단점
3. 서포트 벡터 머신(Support vector machine, SVM)
4. 로지스틱 회귀 분석(Logistic regression)
5. 알고리즘 성능 평가(Model Evaluation)

# K를 적절하게 정하는 방법

- 정답이 없다
- 각각의 K에 대한 직접 성능 평가 후 결정
  - $K=1 \rightarrow 70\%$
  - $K=2 \rightarrow 72\%$
  - $K=3 \rightarrow 73\%$
  - $K=4 \rightarrow 75\%$
  - ...
  - $K=10 \rightarrow 50\%$
- $K=4!$
- 머신러닝의 단점
- 다만, 일반적으로 추천하는  $K = 5 \sim 15$  or  $K = \sqrt{n}$ 
  - ( $n$ : 전체 sample 수)



K-fold 교차검증



$n = 14, K \approx 4$

# Contents

---

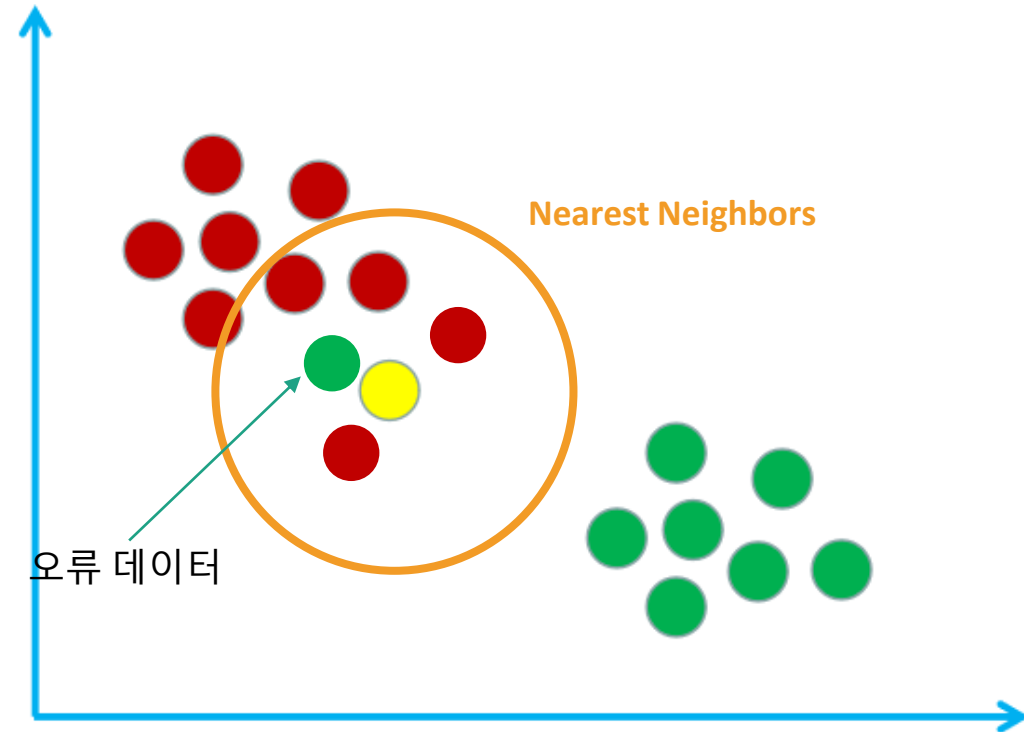
1. Classification (분류) 소개
- 2. K-NN 알고리즘(K-nearest neighbor)**
  - 1) 알고리즘 아이디어 소개
  - 2) K 정하기
  - 3) KNN의 장단점**
3. 서포트 벡터 머신(Support vector machine, SVM)
4. 로지스틱 회귀 분석(Logistic regression)
5. 알고리즘 성능 평가(Model Evaluation)

- 장점

- 단순하고 적용하기 쉽다
- 성능이 안정적이다(robust)
  - 데이터 하나가 오류가 있어도 다수결로 안정적
- 데이터 수가 적어도 ( $n$ 이 작을 때) 잘 동작

- 단점

- 전체 데이터의 수가 많을 때 ( $n$ 이 클때)
  - 새로운 unknown data (노란색 데이터포인트)가 생길때마다 일일이 나머지 모든 데이터들과의 거리 계산이 필요
    - 가장 가까운  $k$ 개의 원을 골라야하니까!
  - 다른 분류 알고리즘 보다 느리게 동작하는 편





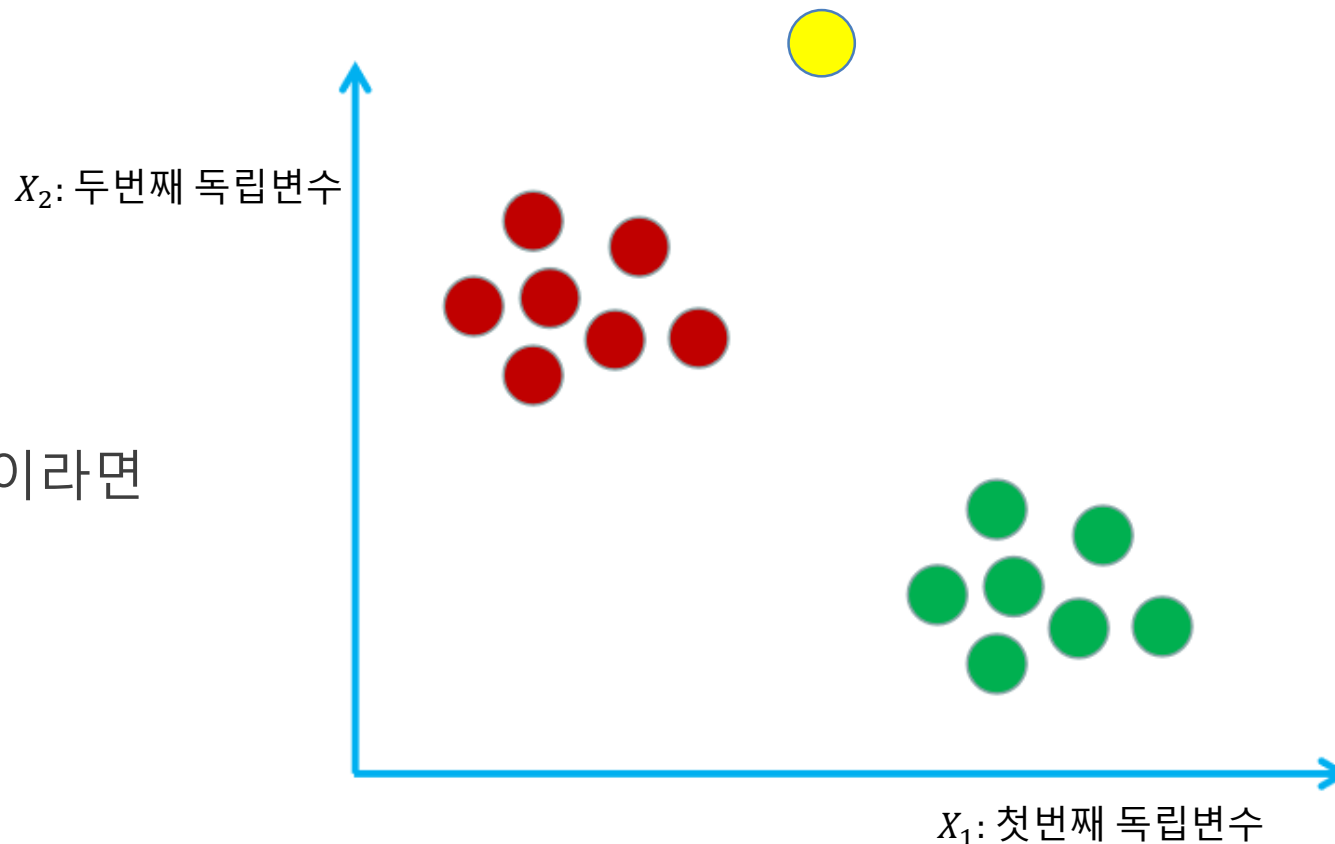
# Contents

---

1. Classification (분류) 소개
2. K-NN 알고리즘(K-nearest neighbor)
- 3. 서포트 벡터 머신(Support vector machine, SVM)**
  - 1) 알고리즘 작동 원리
  - 2) 문제점
  - 3) 해결책: C, Kernel 함수
  - 4) Fine tuning: C랑 Kernel을 찾자
  - 5) 장단점
4. 로지스틱 회귀 분석(Logistic regression)
5. 알고리즘 성능 평가(Model Evaluation)

# Quiz: 노란색 원은 어느 클래스에 속할까?

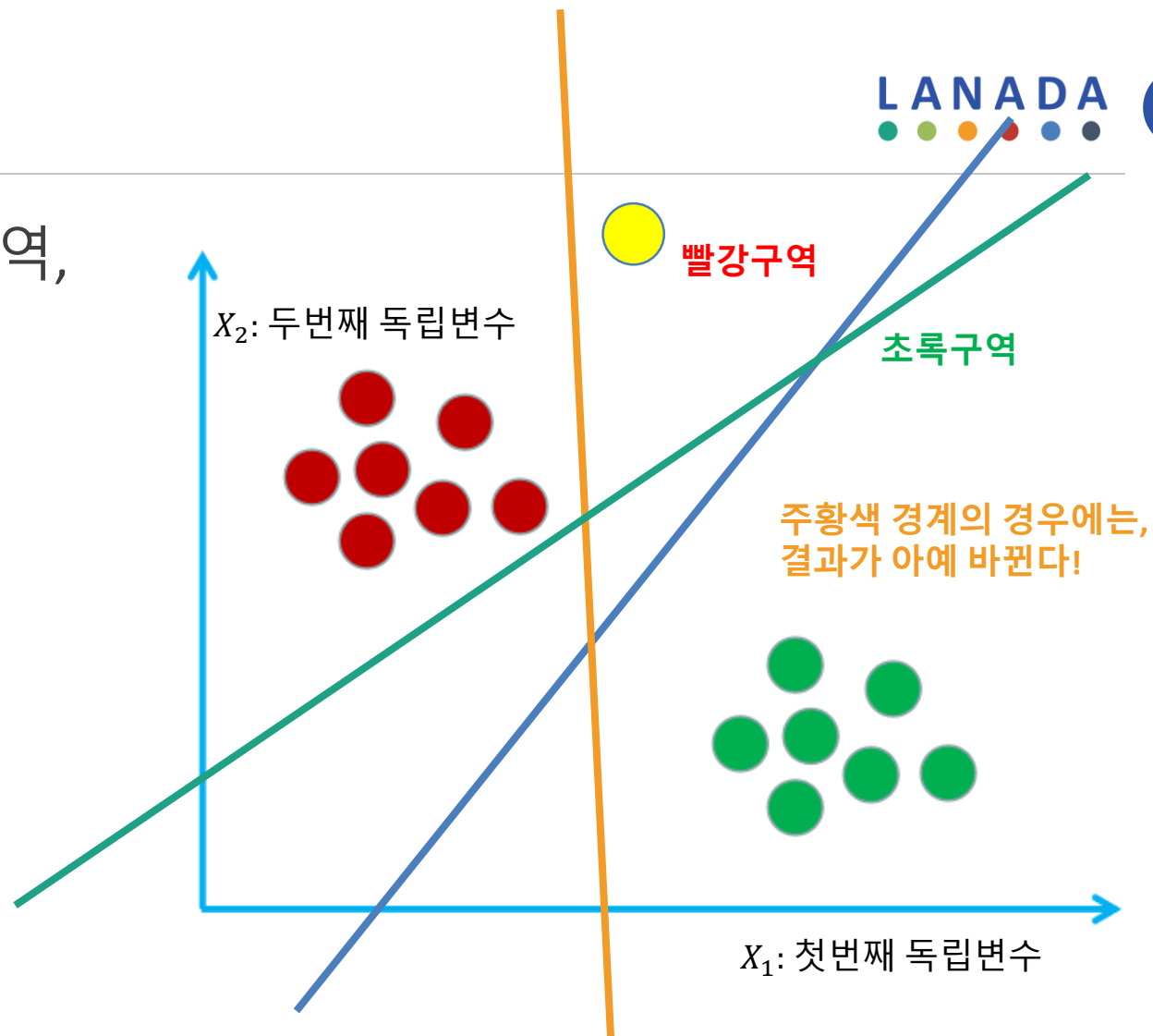
- 학습 데이터
  - 빨간색 원: class 1
  - 초록색 원: class 2
- 예측하고자 하는 원: 노란색
  - class 1 or class 2 둘 중에 하나의 색이라면
  - 무슨 색일까?



**직관적인 대답: 노란색원 → 빨강원일 가능성이 높다!**

# 왜 그렇게 생각했을까?

- 왠지 파란색의 경계를 나누면 빨강구역, 초록 구역으로 나뉘질 거 같다
- 빨강 구역의 원은 빨강색일거같다



71

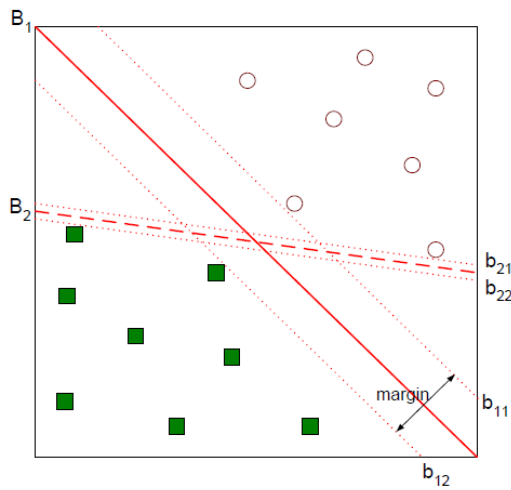
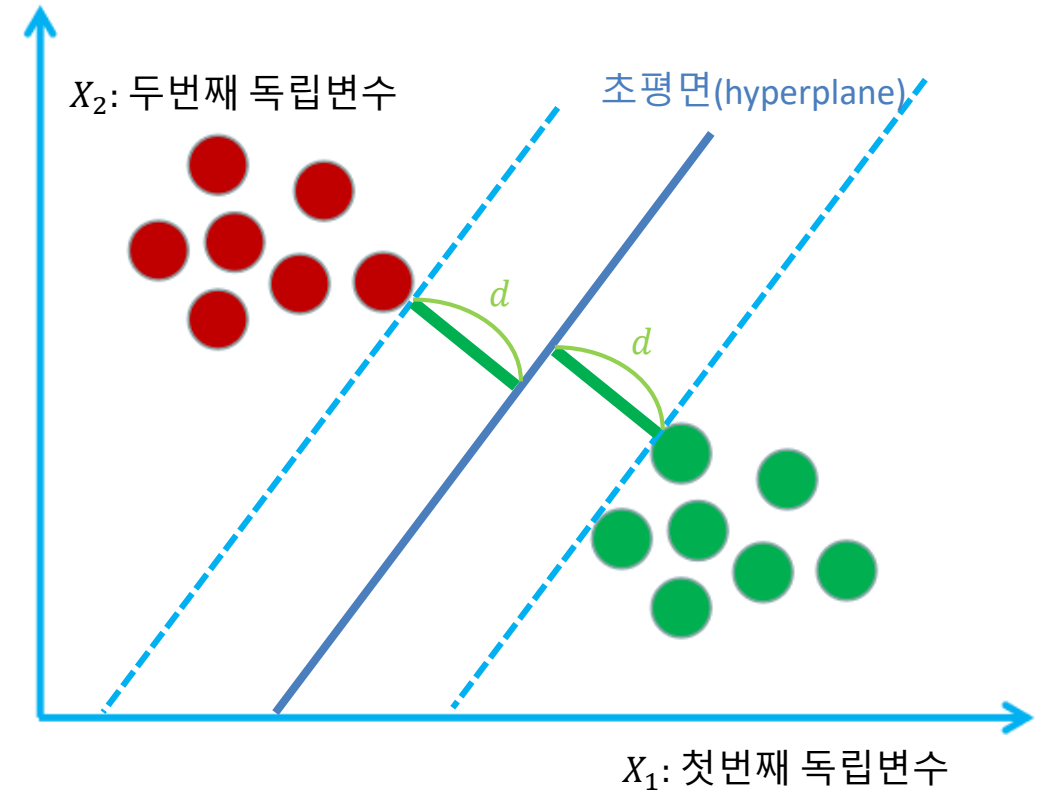
SVM 알고리즘 작동 원리

1. 주어진 원들을 기준으로 **적절한 경계(hyperplane)**를 그린다
2. 이때 노란색원이 경계를 기준으로 빨강색에 속하면 → **빨강**

주어진 학습용 데이터에서 **적절한 경계(hyperplane)**란 무엇이 될까?

# 적절한 초평면(Hyperplane)의 기준

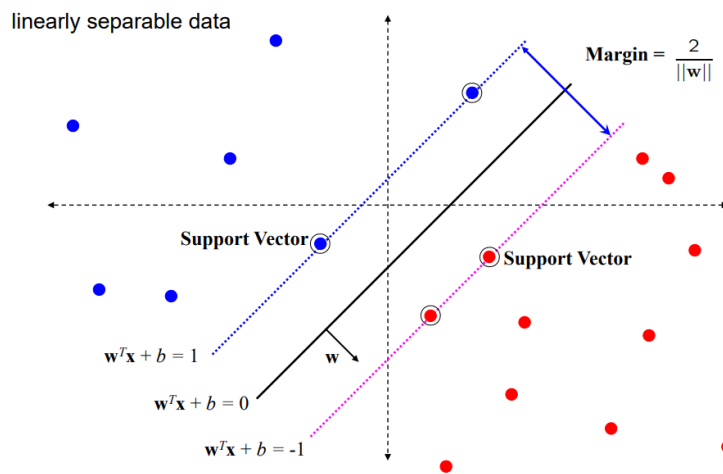
- 가장 적절한 초평면(hyperplane)의 기준
  - 두 종류의 데이터들의 수직이등분선
- $d$ 는 (margin) 경계선을 기준으로 각각 가장 가까운 빨강원과 초록색원의 거리
  - 가장 가까운 두 원을 **서포트 벡터(support vector)** 라고 부름
- 이 margin이 가장 커지는 경계선을 그리는 것이 가장 공평한 경계라고 할 수 있다



굵은 실선의 폭과, 가는 점선을 기준으로 분리했을 때 폭이 다르다

- 경사하강법(Gradient Descent) 사용해서 계산 가능
  - Section 4) 로지스틱 회귀(Logistic Regression)에서 설명 예정
- 

$$\min_{\mathbf{w}} \|\mathbf{w}\|^2 \quad \text{subject to } y_i (\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 \text{ for } i = 1 \dots N$$

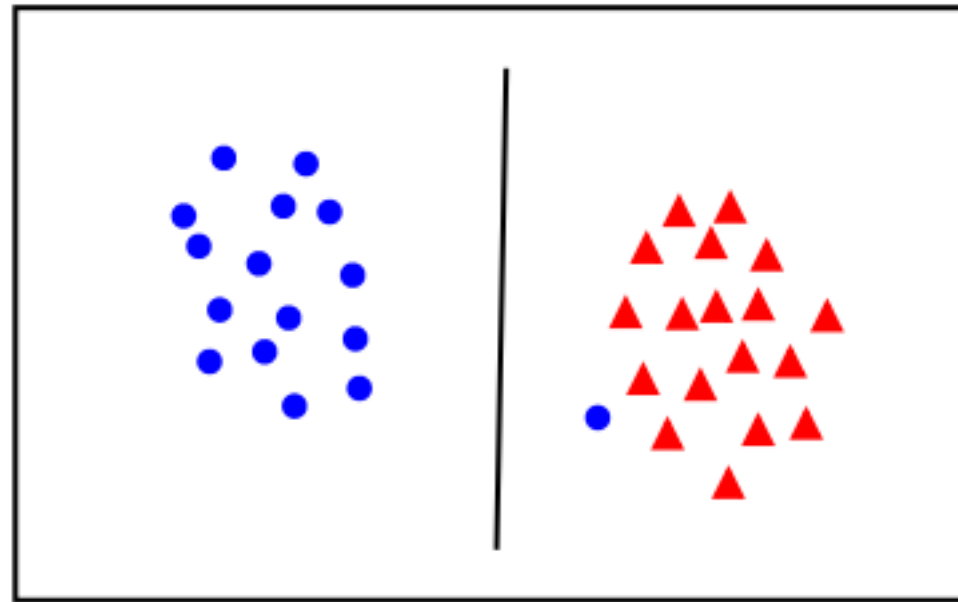
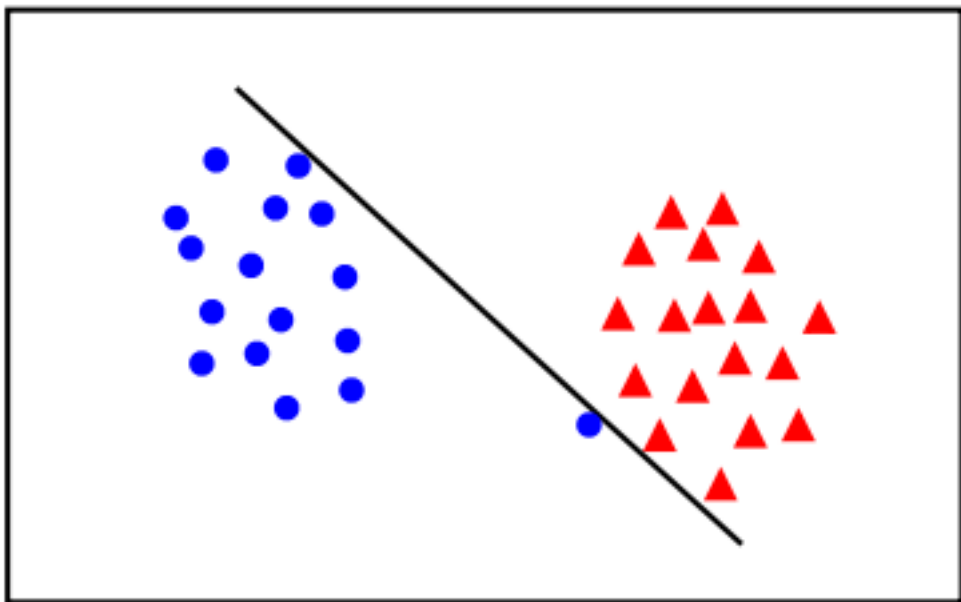


# Contents

---

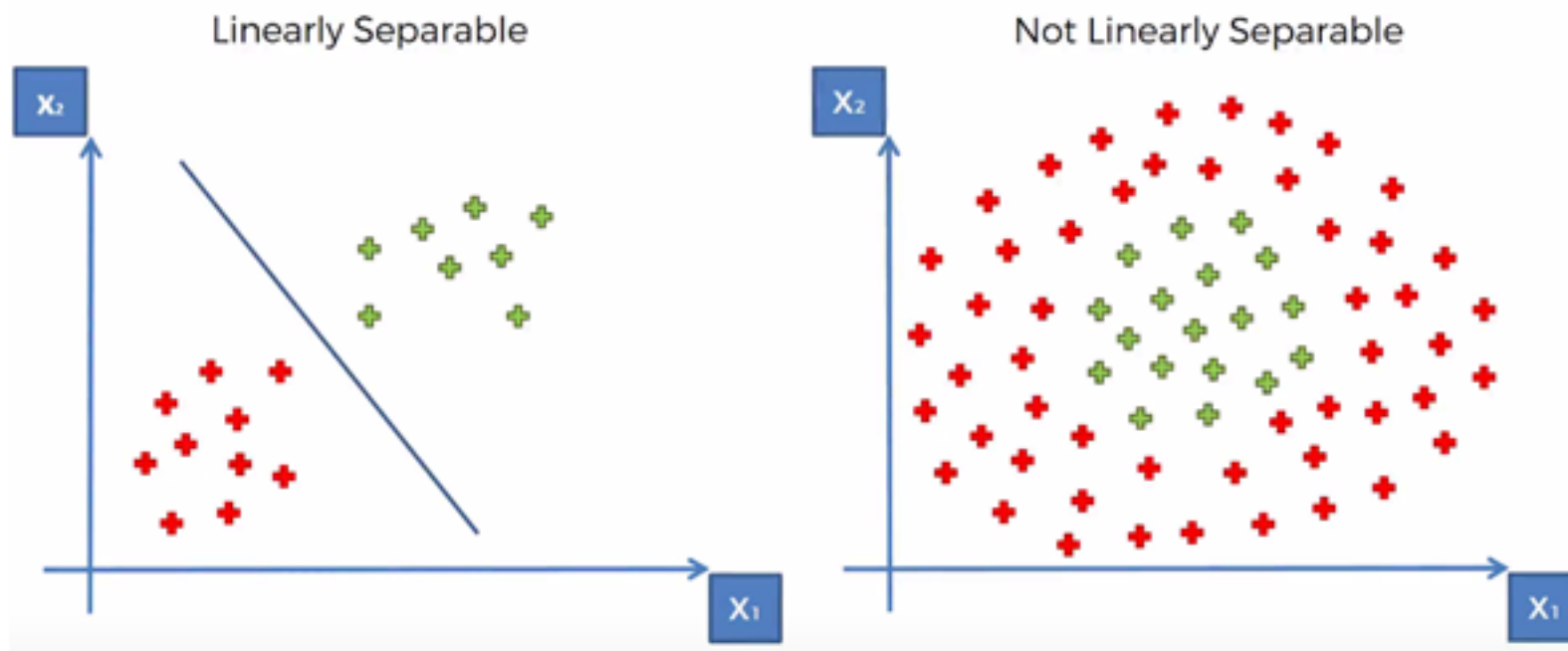
1. Classification (분류) 소개
2. K-NN 알고리즘(K-nearest neighbor)
3. **서포트 벡터 머신(Support vector machine, SVM)**
  - 1) 알고리즘 작동 원리
  - 2) **문제점**
  - 3) 해결책: C, Kernel 함수
  - 4) Fine tuning: C랑 Kernel을 찾자
  - 5) 장단점
4. 로지스틱 회귀 분석(Logistic regression)
5. 알고리즘 성능 평가(Model Evaluation)

- 왼쪽 그림처럼 데이터를 정확히 분류 가능한 경우에도,
- 오른쪽 그림처럼 전체적인 폭을 넓히고 정확한 분류를 포기하는 것이 더 좋을 수 있다
  - 일부 특이점(outlier)를 적절히 무시할 필요가 생김



# 서포트 벡터 머신 문제점 2

- 사람은 쉽게 2개의 클래스를 구분 가능하지만, 선형 함수로는 구분이 불가능한 경우 존재
- 직선이 아닌 곡선의 초평면 필요



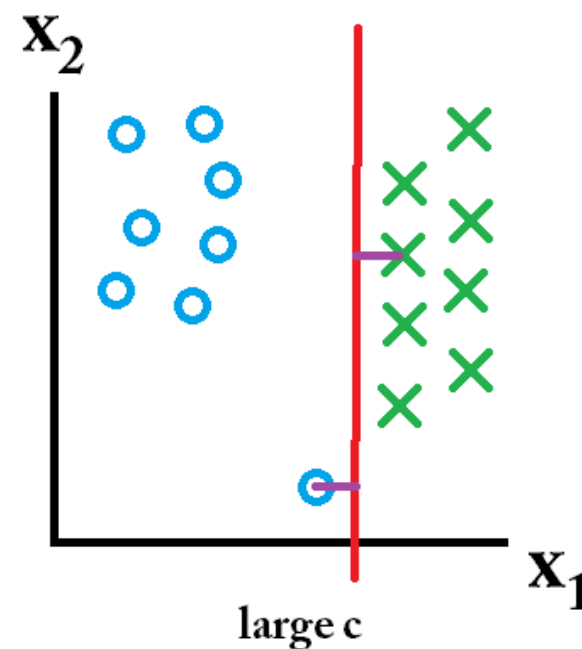
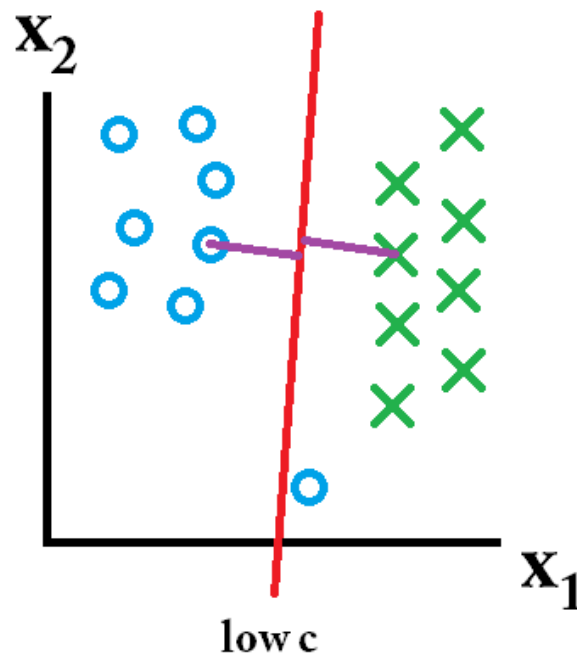
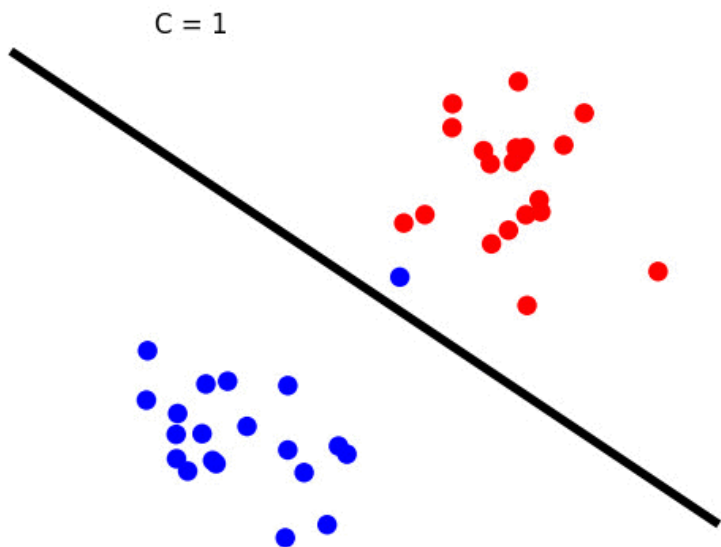


# Contents

---

1. Classification (분류) 소개
2. K-NN 알고리즘(K-nearest neighbor)
- 3. 서포트 벡터 머신(Support vector machine, SVM)**
  - 1) 알고리즘 작동 원리
  - 2) 문제점
  - 3) 해결책: C, Kernel 함수**
  - 4) Fine tuning: C랑 Kernel을 찾자
  - 5) 장단점
4. 로지스틱 회귀 분석(Logistic regression)
5. 알고리즘 성능 평가(Model Evaluation)

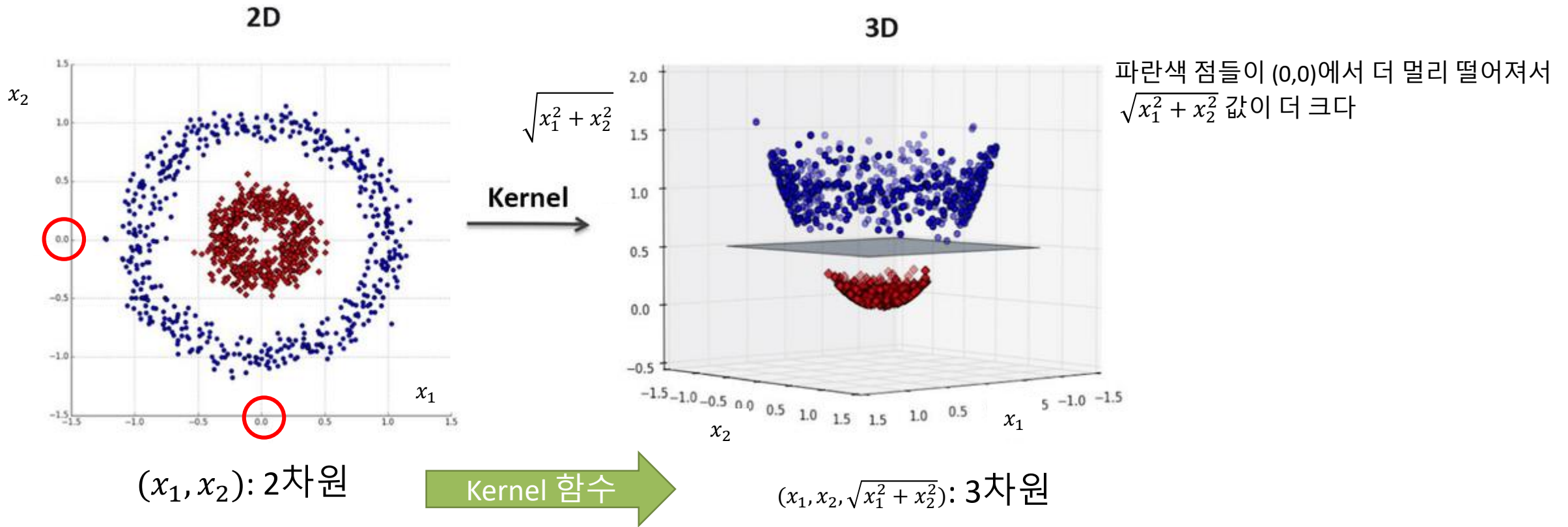
- 새로운 매개 변수  $C$ 를 도입
  - $C$ 가 크면 분류를 정확히 하는걸 우선
  - $C$ 가 작을수록 몇몇 데이터를 무시하고 나머지 데이터들끼리의 폭을 최대화 하는 것을 목표
    - 기존 SVM은  $C = \infty$ 인 경우



적절한  $C$ 는 어떻게 고르는가?

# Kernel 함수의 도입

- 차원을 더 고차원으로 변형해서, hyperplane이 존재하도록 유도



적절한 Kernel은 어떻게 고르는가?

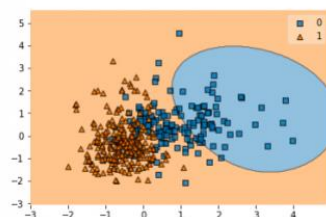
# Contents

---

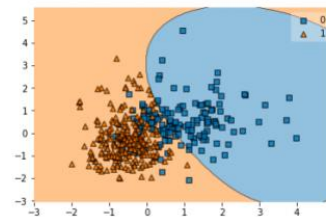
1. Classification (분류) 소개
2. K-NN 알고리즘(K-nearest neighbor)
3. **서포트 벡터 머신(Support vector machine, SVM)**
  - 1) 알고리즘 작동 원리
  - 2) 문제점
  - 3) 해결책: C, Kernel 함수
  - 4) **Fine tuning: C랑 Kernel을 찾자**
  - 5) 장단점
4. 로지스틱 회귀 분석(Logistic regression)
5. 알고리즘 성능 평가(Model Evaluation)

# C와 Kernel에 따라 경계가 여러가지로 바뀐다

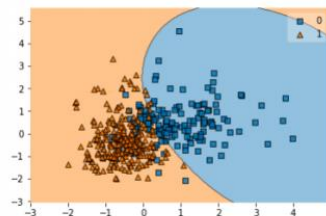
- 같은 데이터에서도 전혀 다른 boundary가 나온다



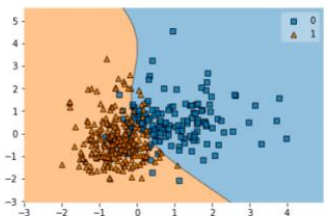
C = 0.02  
Accuracy: 81.3%



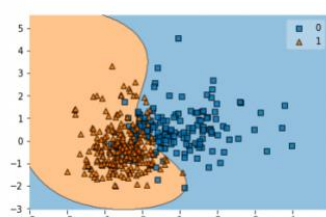
C = 0.03  
Accuracy: 88.3%



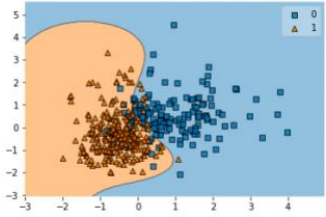
C = 0.08  
Accuracy: 90.6%



C = 1.0  
Accuracy: 90.6%

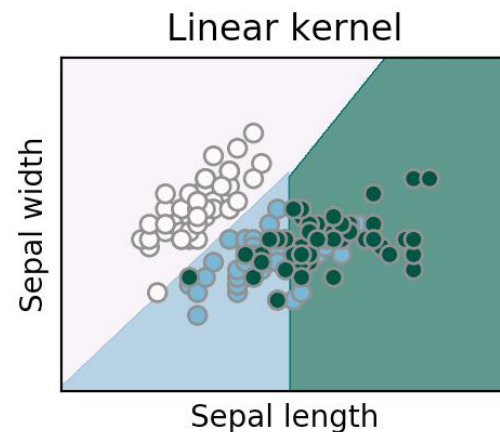


C = 10.0  
Accuracy: 90.1%

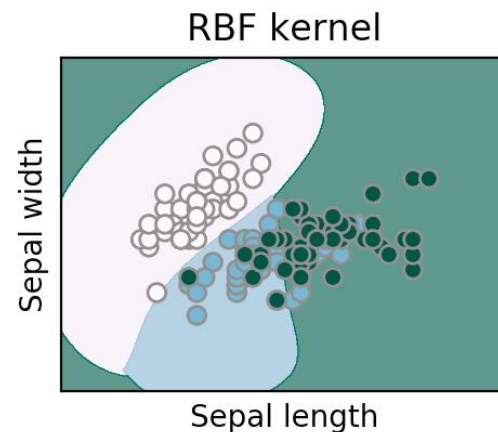


C = 100.0  
Accuracy: 90.1%

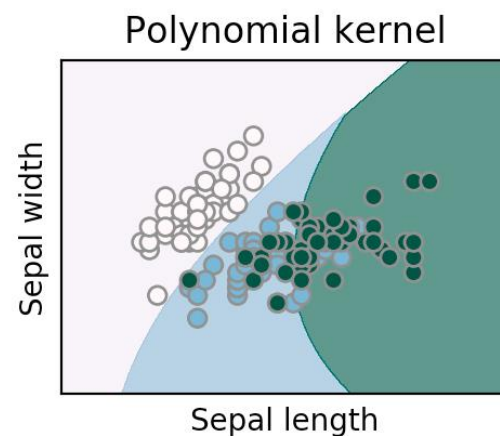
C에 따라 바뀌는 형태



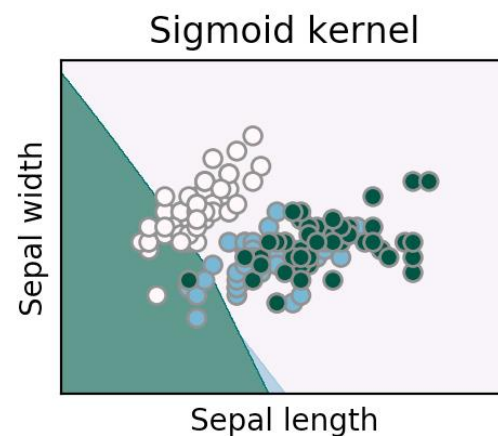
Linear kernel



RBF kernel



Polynomial kernel

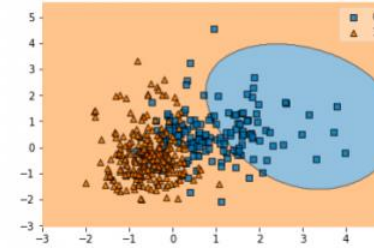


Sigmoid kernel

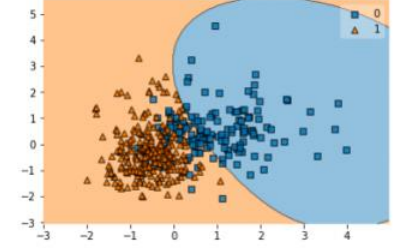
4가지 종류의 Kernel 함수에 따라 바뀌는 형태  
Linear Kernel, Polynomial Kernel, RBF Kernel, Sigmoid Kernel

# Fine Tuning

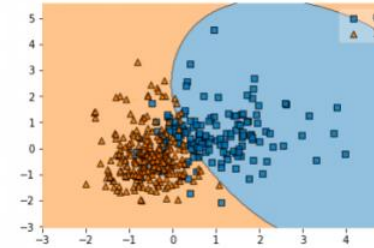
- Part 1의 선형회귀에서는 최적의  $\theta_0, \theta_1$ 를 계산가능
  - **선형회귀의 큰 장점!**
- 일반적인 기계학습에서는 매개변수를 최적으로 정하는 공식은 없다
  - Ex)
    - KNN에서 K 정하기
    - SVM에서 C or Kernel 정하기
- 따라서, 여러 매개변수를 테스트해보면서 성능이 잘 찾아야함
  - 이 과정에서 데이터의 구조등을 고민하고 경험이 생기면 더 효율적으로 진행이 됨



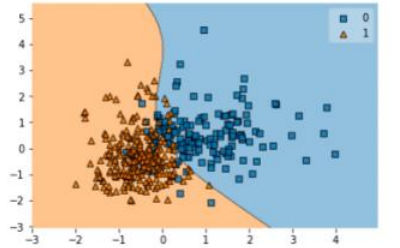
C = 0.02  
Accuracy: 81.3%



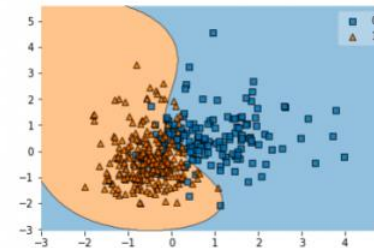
C = 0.03  
Accuracy: 88.3%



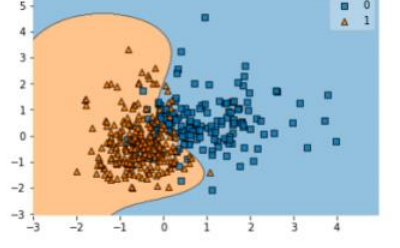
C = 0.08  
Accuracy: 90.6%



C = 1.0  
Accuracy: 90.6%



C = 10.0  
Accuracy: 90.1%



C = 100.0  
Accuracy: 90.1%

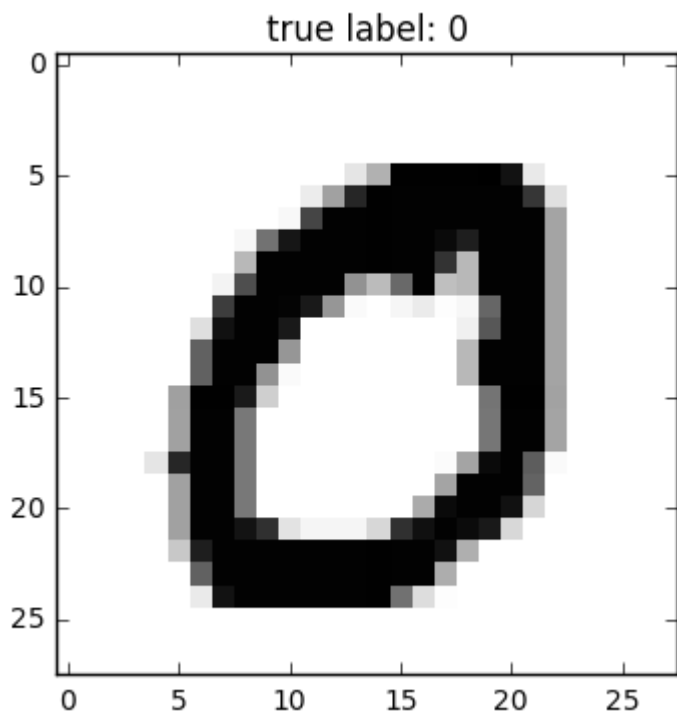
# Contents

---

1. Classification (분류) 소개
2. K-NN 알고리즘(K-nearest neighbor)
- 3. 서포트 벡터 머신(Support vector machine, SVM)**
  - 1) 알고리즘 작동 원리
  - 2) 문제점
  - 3) 해결책: C, Kernel 함수
  - 4) Fine tuning: C랑 Kernel을 찾자
  - 5) 장단점**
4. 로지스틱 회귀 분석(Logistic regression)
5. 알고리즘 성능 평가(Model Evaluation)

- 장점

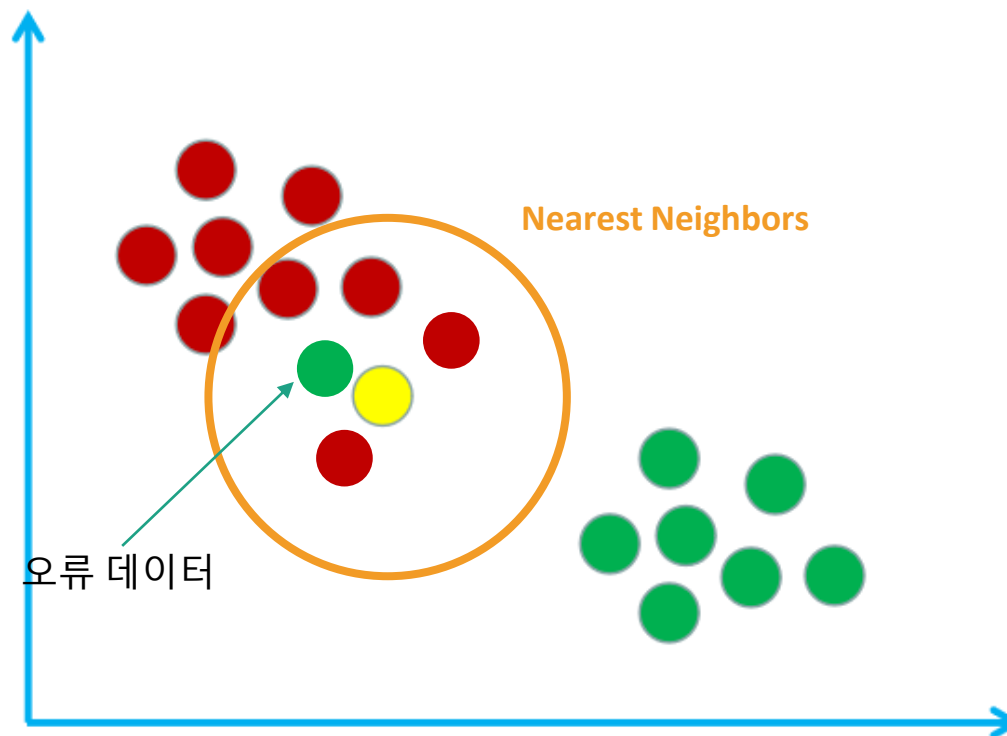
- 고차원적인 data에서도 잘 동작함



MNIST 이미지:  $28 \times 28 = 784$  차원

- 단점

- KNN의 경우에 오류 데이터가 있어도 다수결로 robust 하지만
- SVM은 복잡한 경계를 그리려는 성질이 있다





# Contents

---

1. Classification (분류) 소개
2. K-NN 알고리즘(K-nearest neighbor)
3. 서포트 벡터 머신(Support vector machine, SVM)
- 4. 로지스틱 회귀(Logistic regression)**
  - 1) 일반적인 분류 문제와의 차이점
  - 2) 로지스틱 회귀를 사용하는 경우
  - 3) Sigmoid 함수
  - 4) Log-loss (Cross entropy loss)
5. 알고리즘 성능 평가(Model Evaluation)

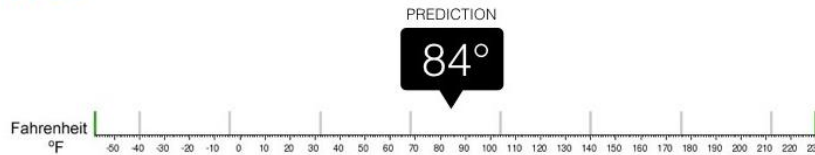
# Remind) 로지스틱 회귀(Logistic Regression)

- 회귀(Regression) (Part 1)
  - 목표로 하는  $Y$ 가 **연속적**일 때 (continuous)



## Regression

What is the temperature going to be tomorrow?



회귀와 분류는 다른 문제

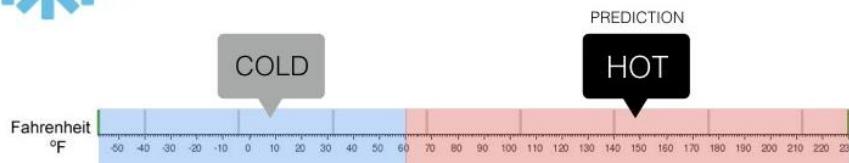
그런데 왜 분류 문제를 푸는 **로지스틱 회귀**는 이름이 **로지스틱 분류**가 아니지?

- 분류(Classification) (Part 2)
  - 목표로 하는  $Y$ 가 **이산적**일 때 (discrete, categorical)



## Classification

Will it be Cold or Hot tomorrow?



# 로지스틱 회귀는 연속한 값(확률)을 목표로 한다

- 일반적인 분류 문제

- 학습데이터를 바탕으로 패턴을 익혀서
- **??**의 부분의 답을 맞힌다

X: 독립      Y: 종속

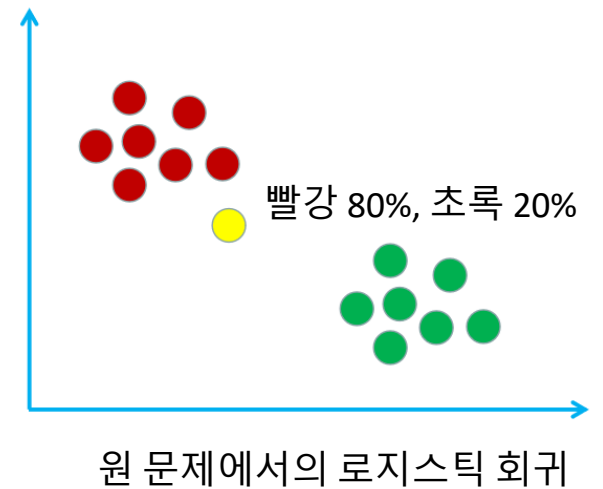
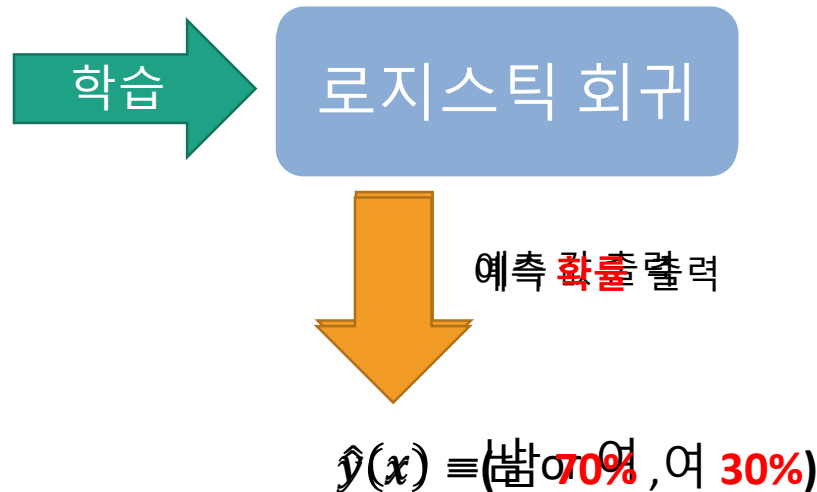
IQ	연봉	성별
107	6305	여
95	5730	여
114	8735	여
83	6735	여
101	6170	여
119	7805	남
92	6205	남
108	8830	남
129	9075	남
104	7935	남
94	6415	남
112	7800	<b>??</b>

학습데이터 (known data)

목표 데이터 (unknown data)

- 로지스틱 회귀

- 학습데이터를 바탕으로 패턴을 익혀서
- **??**에서 각 클래스가 나올 **확률**을 예측한다
  - 여기서 확률은 **연속하다(continuous)**
  - **따라서 회귀 알고리즘**



# Contents

---

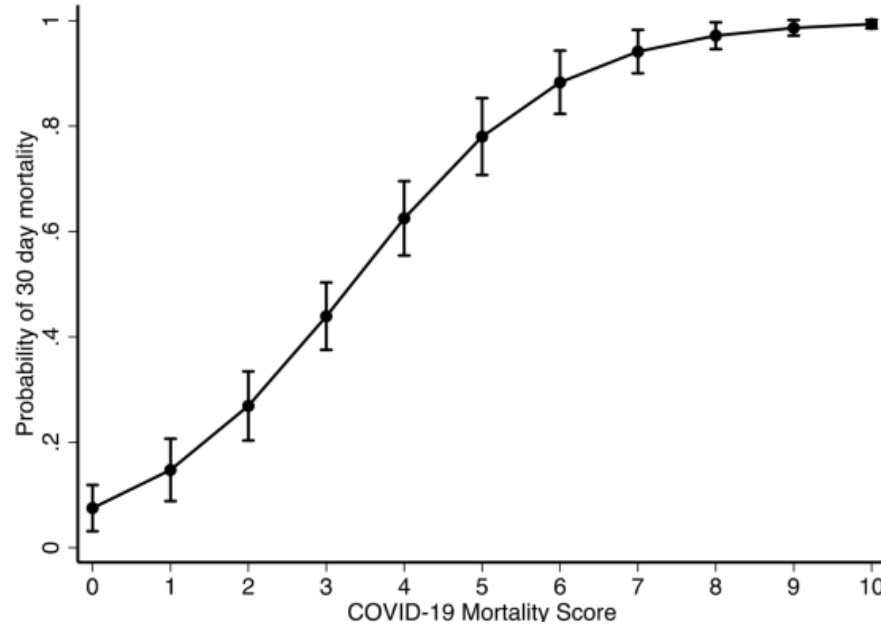
1. Classification (분류) 소개
2. K-NN 알고리즘(K-nearest neighbor)
3. 서포트 벡터 머신(Support vector machine, SVM)
- 4. 로지스틱 회귀(Logistic regression)**
  - 1) 일반적인 분류 문제와의 차이점
  - 2) 로지스틱 회귀를 사용하는 경우
  - 3) Sigmoid 함수
  - 4) Log-loss (Cross entropy loss)
5. 알고리즘 성능 평가(Model Evaluation)

# 로지스틱 회귀를 사용하는 예1: 확률 예측

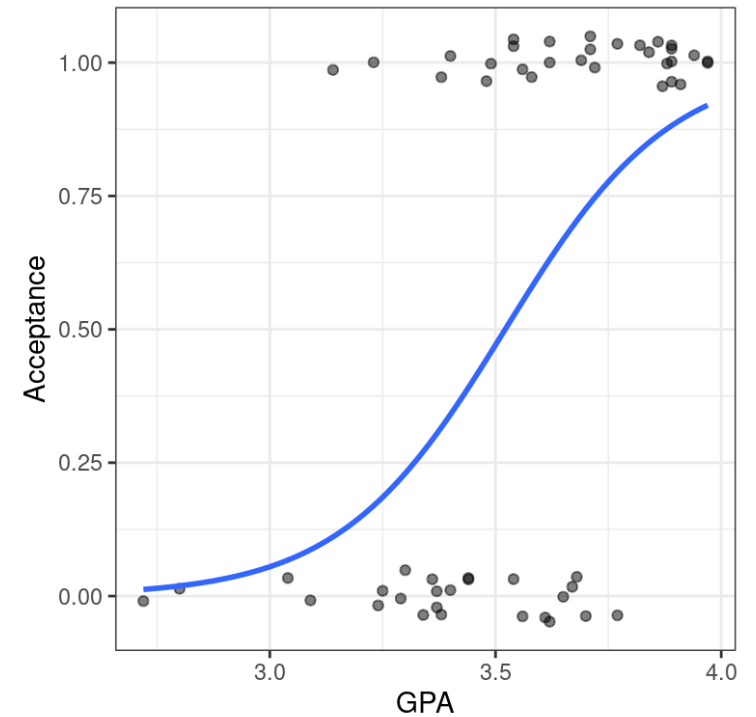
- 어떤 사건이 일어날 확률을 알고 싶을 때

내일 (23일, 토)		최고/최저기온 : 26°C / 18°C				
시간	날씨	기온	바람	습도	강수 확률	강수량
24		19°C	동 1m/s	85%	30%	
3	흐림	19°C	동 1m/s	85%	20%	-
6	구름 많음	18°C	동 1m/s	90%	10%	
9	구름 조금	21°C	남동 1m/s	75%	0%	-
12	맑음	25°C	남서 1m/s	55%	0%	
15	맑음	26°C	서 2m/s	45%	0%	-
18	맑음	24°C	서 2m/s	55%	0%	

일기예보: 강수 확률  
X: 현재 기상 정보, Y: 강수 확률



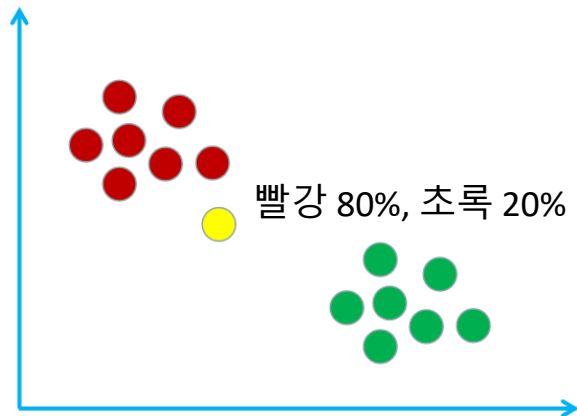
코로나19 사망위험 예측모델 개발  
X: 환자 중증도, Y: 30일 이내 사망 확률



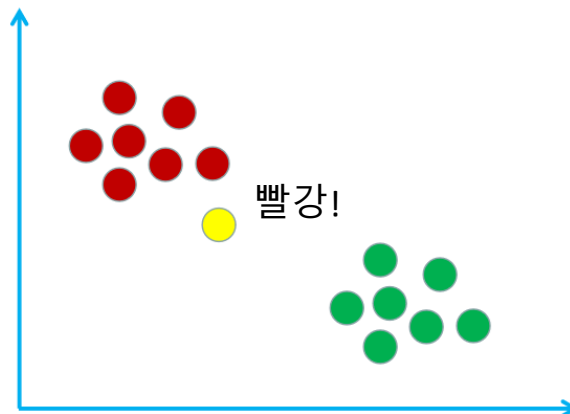
학점 - 합격 확률 예측  
X: 학점, Y: 합격률

# 로지스틱 회귀를 사용하는 예1: 일반 분류 문제

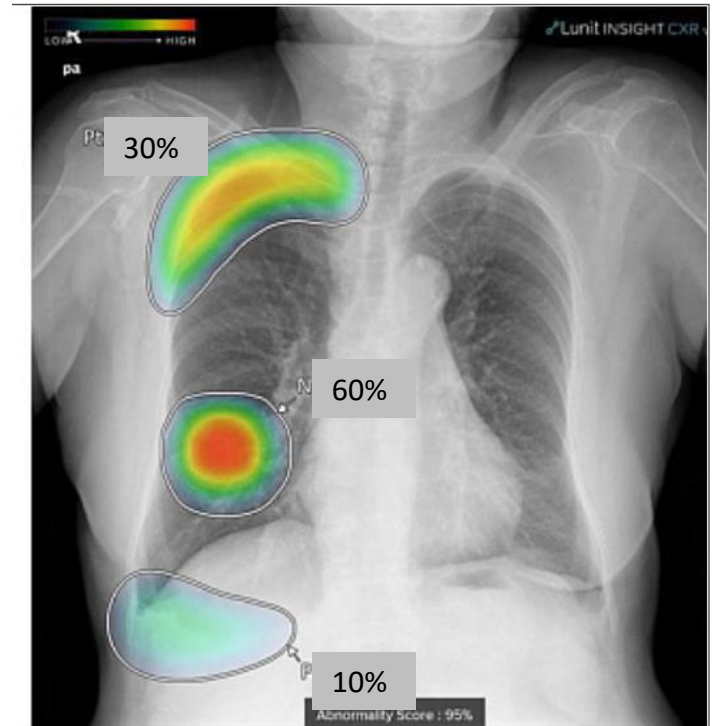
- 로지스틱 회귀에서 나온 확률을 보고 50%이상인 클래스를 고른다
  - 일반적인 분류 알고리즘: KNN, SVM과 같은 방식
- 암환자 같이 조금이라도 의심스러울 경우 무조건 예측 해야 할 시
  - 기준치를 50%보다 낮춰서 예측 가능 (ex, 10% 이상이면 예측)
    - 미리 boundary를 그리는 SVM은 불가능한 방식



원 문제에서의 로지스틱 회귀



로지스틱 회귀를 이용한 분류문제



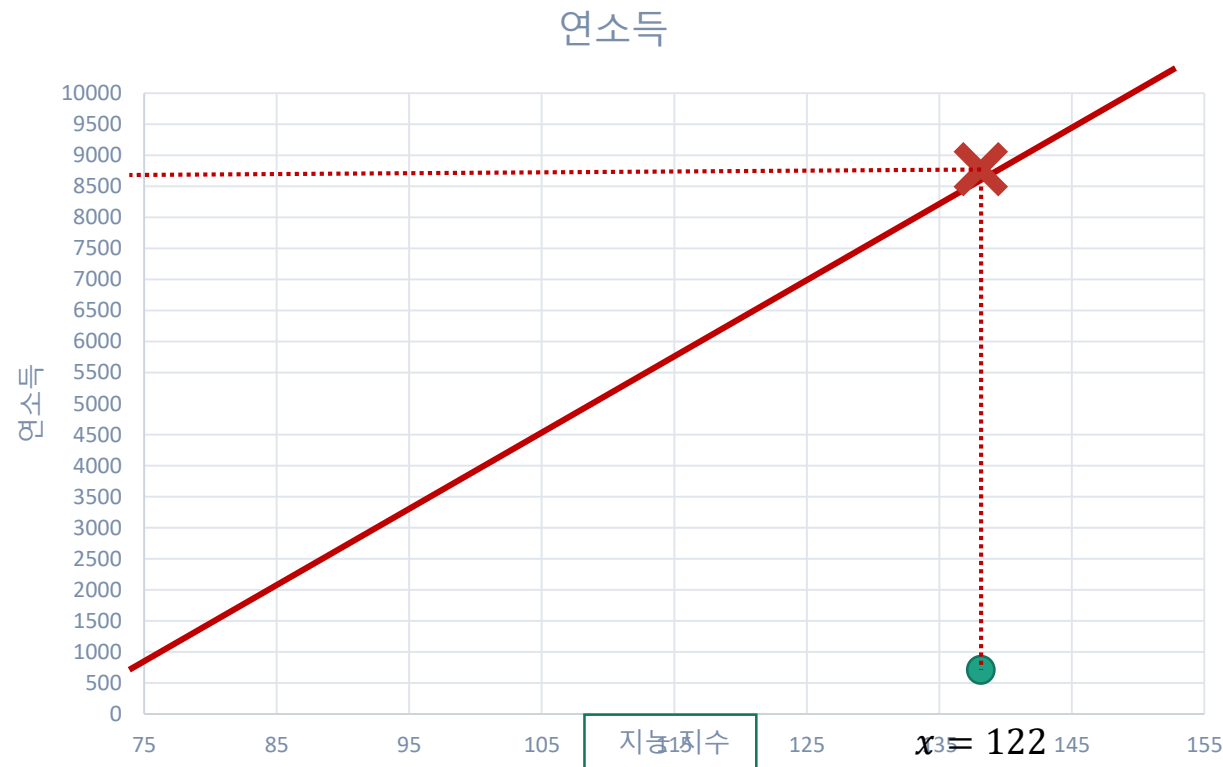
암환자 예측에서의 로지스틱 회귀  
안전할 확률이 더 높더라도 '위험'으로 분류 필요

- 선형 회귀

- 타겟으로 하는  $\hat{y}(x) = \theta_0 + \theta_1 x$ 의 범위가  $(-\infty, \infty)$ 로 넓다
- 자유로운 범위의 예측

- 로지스틱 회귀

- $\hat{y}(x)$ 의 범위가  $[0,1]$ 로 제한됨
- 주로 확률에 관련된 예측에 집중



# Contents

---

1. Classification (분류) 소개
2. K-NN 알고리즘(K-nearest neighbor)
3. 서포트 벡터 머신(Support vector machine, SVM)
- 4. 로지스틱 회귀(Logistic regression)**
  - 1) 일반적인 분류 문제와의 차이점
  - 2) 로지스틱 회귀를 사용하는 경우
  - 3) Sigmoid 함수**
  - 4) Log-loss (Cross entropy loss)
5. 알고리즘 성능 평가(Model Evaluation)



# Sigmoid 함수

- $\hat{y}(x) = \frac{1}{1+e^{-(\theta_0+\theta_1 x)}} = \frac{1}{1+e^{-W^T \cdot X}} \quad W^T = (\theta_0, \theta_1), X = (1, x)$

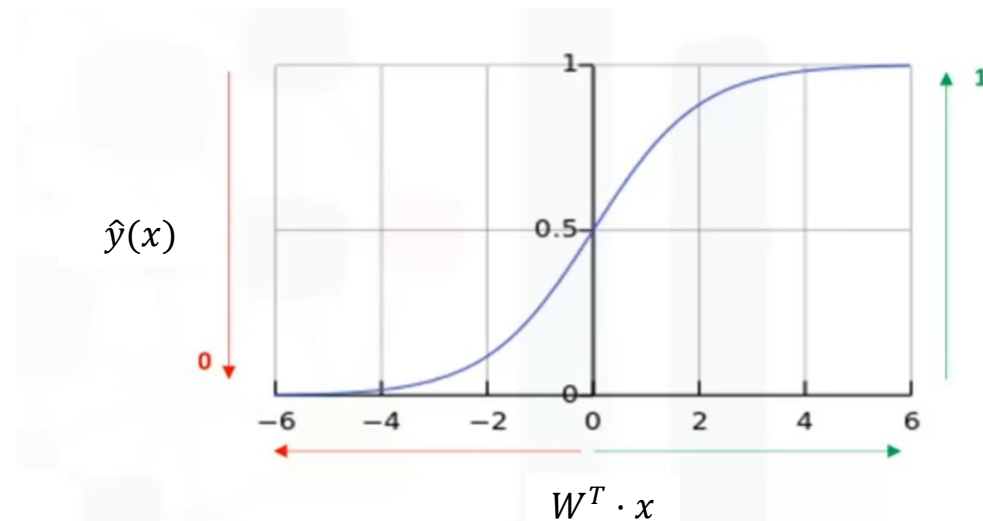
- $W^T \cdot x$  값이  $+\infty$ 로 커질 수록
  - $\hat{y}(x) \rightarrow 1$

- $W^T \cdot x$ 가  $-\infty$ 로 작아질 수록
  - $\hat{y}(x) \rightarrow 0$

- $W^T \cdot x = 0$  일 때,
  - $\hat{y}(x) = \frac{1}{2} \rightarrow$  반반의 확률(전혀 모르겠다는 의미)

- 단 어떤  $x$ 에 대해서도 100% or 0% 존재하지 않음
  - 100% 완벽한 예측은 없다는 것

- $\theta_0 + \theta_1 x$  or  $W^T \cdot x \rightarrow$  현재 확률에 대한 믿음을 나타냄



0: 여자,  
1: 남자 일때,

$\hat{y}(x_i) = p(y = 1|x_i) = 0.2 \rightarrow$  (남자: 20%, 여자: 80% 로 예측)

$\hat{y}(x_i) = p(y = 1|x_i) = 0.99 \rightarrow$  (남자: 99%, 여자: 1% 로 예측)

$\hat{y}(x_i) = p(y = 1|x_i) = 0.5 \rightarrow$  (남자: 50%, 여자: 50% 로 예측)

# Contents

---

1. Classification (분류) 소개
2. K-NN 알고리즘(K-nearest neighbor)
3. 서포트 벡터 머신(Support vector machine, SVM)
- 4. 로지스틱 회귀(Logistic regression)**
  - 1) 일반적인 분류 문제와의 차이점
  - 2) 로지스틱 회귀를 사용하는 경우
  - 3) Sigmoid 함수
  - 4) Log-loss (Cross entropy loss)**
5. 알고리즘 성능 평가(Model Evaluation)

## 오차

- $Error(y_i, \hat{y}(x_i))$

- ex)  $Error(0, 0.2)$

- $y$ 의 0 → 실제 성별이 여자 였다는 의미
    - $\hat{y}(x_i) = 0.2$  → (남자: 20%, 여자: 80%)

- ex)  $Error(1, 0.4)$

- $y$ 의 1 → 실제 성별이 남자 였다는 의미
    - $\hat{y}(x_i) = 0.4$  → (남자: 40%, 여자: 60%)

## $Error(0, 0.2)$ vs $Error(1, 0.4)$ 무엇이 더 큰 오차?

- $Error(0, 0.2)$ : 남자 확률 20%라고 예상했는데, 실제로 여자 였던 경우
- $Error(1, 0.4)$ : 남자 확률 40%라고 예상했는데, 실제로 남자 였던 경우

$$Error(1, 0.4) \gg Error(0, 0.2)$$

X: 독립      Y: 종속

IQ	연봉	성별
107	6305	0
95	5730	0
114	8735	0
83	6735	0
101	6170	0
119	7805	1
92	6205	1
108	8830	1
129	9075	1
104	7935	1
94	6415	1
112	7800	??

0: 여자, 1: 남자

학습데이터 (known data)

목표 데이터 (unknown data)

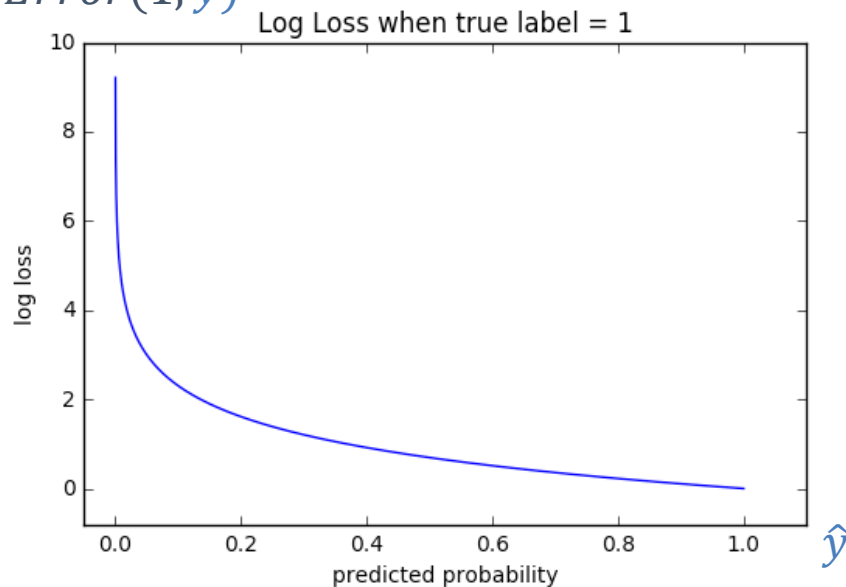
선형 회귀에서 사용했던 목표 오차

$$MSE(y, \hat{y}) = \text{mean}(y - \hat{y})^2$$

# Log loss (Cross-Entropy Loss)

- $Error(y, \hat{y}) := Log\_loss(y, \hat{y}) = -(y \cdot \log \hat{y} + (1 - y) \cdot \log(1 - \hat{y}))$

$Error(1, \hat{y})$



$y = 1$  일 때  $Error(1, \hat{y})$  그래프

$y = 1$  일 확률이 100% 라고 예측한 경우

$\hat{y} = 1$  일 때  $Error(1, \hat{y}) = 0$

100% 확신을 가졌다가 맞으면,  
0 error 발생

$\hat{y} = 0$  일 때  $Error(1, \hat{y}) = \infty$

100% 확신을 가졌다가 틀리면,  
 $\infty$  error 발생

$y = 1$  일 확률이 0% 라고 예측한 경우

	오차 함수	학습방법
선형 회귀	$MSE(y, \hat{y}) = (y - \hat{y})^2$	직접 계산
로지스틱 회귀	$Log_{loss}(y, \hat{y}) = -(y \cdot \log \hat{y} + (1 - y) \cdot \log(1 - \hat{y}))$	경사하강법

$$\theta_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\theta_0 = \bar{y} - \theta_1 \cdot \bar{x}$$

(Remind) 선형회귀에서 학습 방법

# Contents

---

1. 분류(Classification) 소개
2. K-NN 알고리즘(K-nearest neighbor)
3. 서포트 벡터 머신(Support vector machine, SVM)
4. 로지스틱 회귀(Logistic regression)
5. **분류 알고리즘 성능 평가(Model Evaluation)**
  - 1) 선형 회귀와의 비교: 검증 데이터 셋 분리
  - 2) 검증 정확도(test accuracy)
  - 3) Precision and recall
  - 4) F1-score

# 분류 알고리즘 성능 평가(Model Evaluation)

- 분류 학습된 모델은 얼마만큼 정확한가?
  - KNN, SVM, 로지스틱 회귀
- 일반 회귀 학습의 모델 평가와 공통점과 차이점?
  - 검증 데이터 분리하는 할 것인가? MSE나 MAE같은 오차함수를 또 사용할 것인가?

지능지수	성별
107	0
95	0
114	0
83	0
101	0
119	0
92	1
108	1
129	1
104	1
94	1

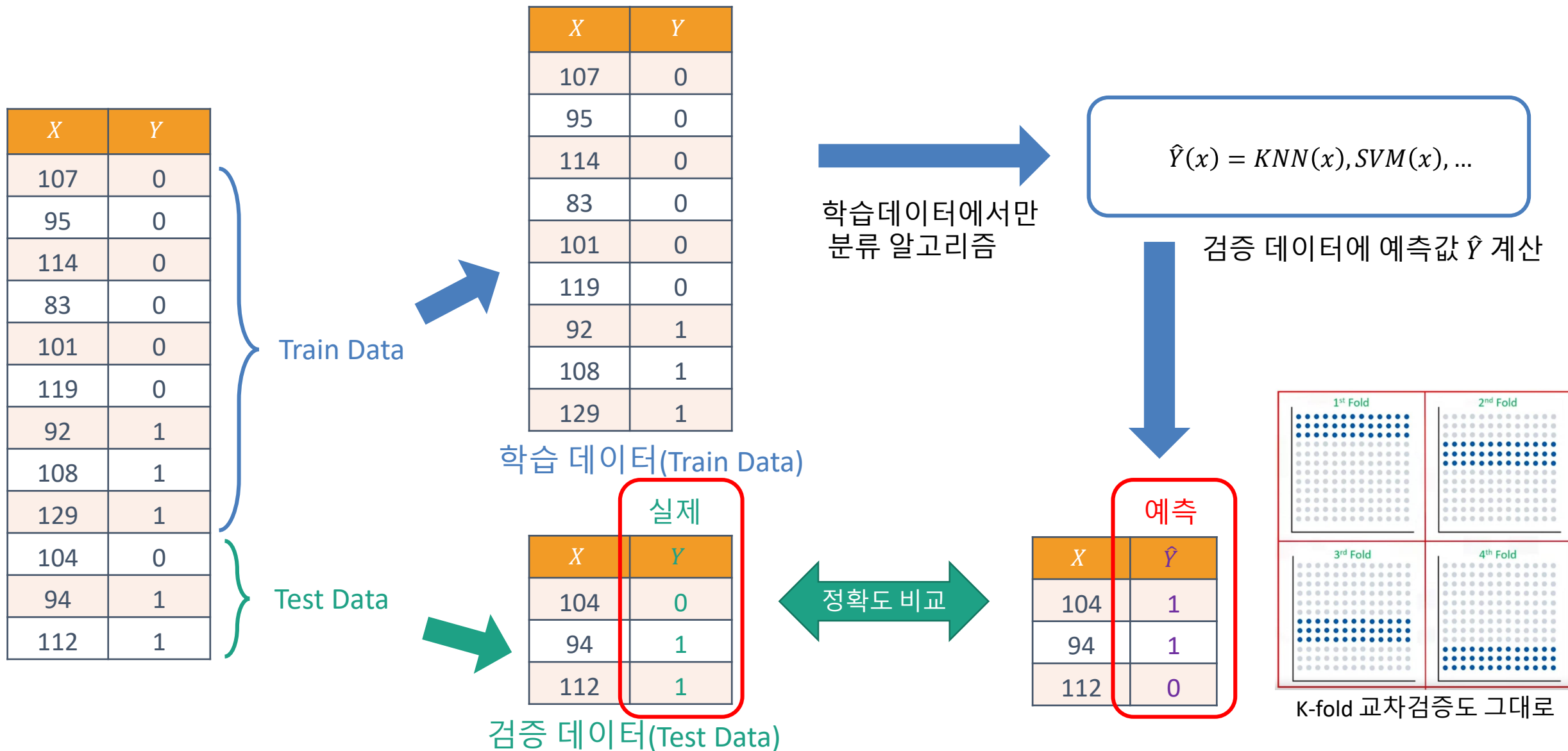
데이터 셋

분류 알고리즘

$$\hat{y}(x) = 0 \text{ or } 1$$

예측이 100% 정확하리라 믿을 수는 없다...  
그렇다면 얼마나 정확할까?

## 99



# 차이점) 다른 종류의 정확도 함수

- (Remind) 선형 회귀에서는  $\hat{y}$ 가 연속한 값을 가짐
  - $MSE_{Test} = \frac{1}{n} \sum_{i \in Test} (\hat{y}_i - y_i)^2$
  - $MAE_{Test} = \frac{1}{n} \sum_{i \in Test} |\hat{y}_i - y_i|$
  - $RSE_{Test} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}, R_{Test}^2 = 1 - RSE_{Test}$
- 분류 문제에서는 (로지스틱 회귀를 제외하고) +-1 만 사용됨
  - 일반 분류 문제 정확도 평가
    - Test accuracy, precision and recall, F1-score
  - 로지스틱 회귀 정확도 평가
    - Log\_loss: 로지스틱 학습에 쓰였던 그대로
      - $Error(y, \hat{y}) := Log\_loss(y, \hat{y}) = -(y \cdot \log \hat{y} + (1 - y) \cdot \log(1 - \hat{y}))$



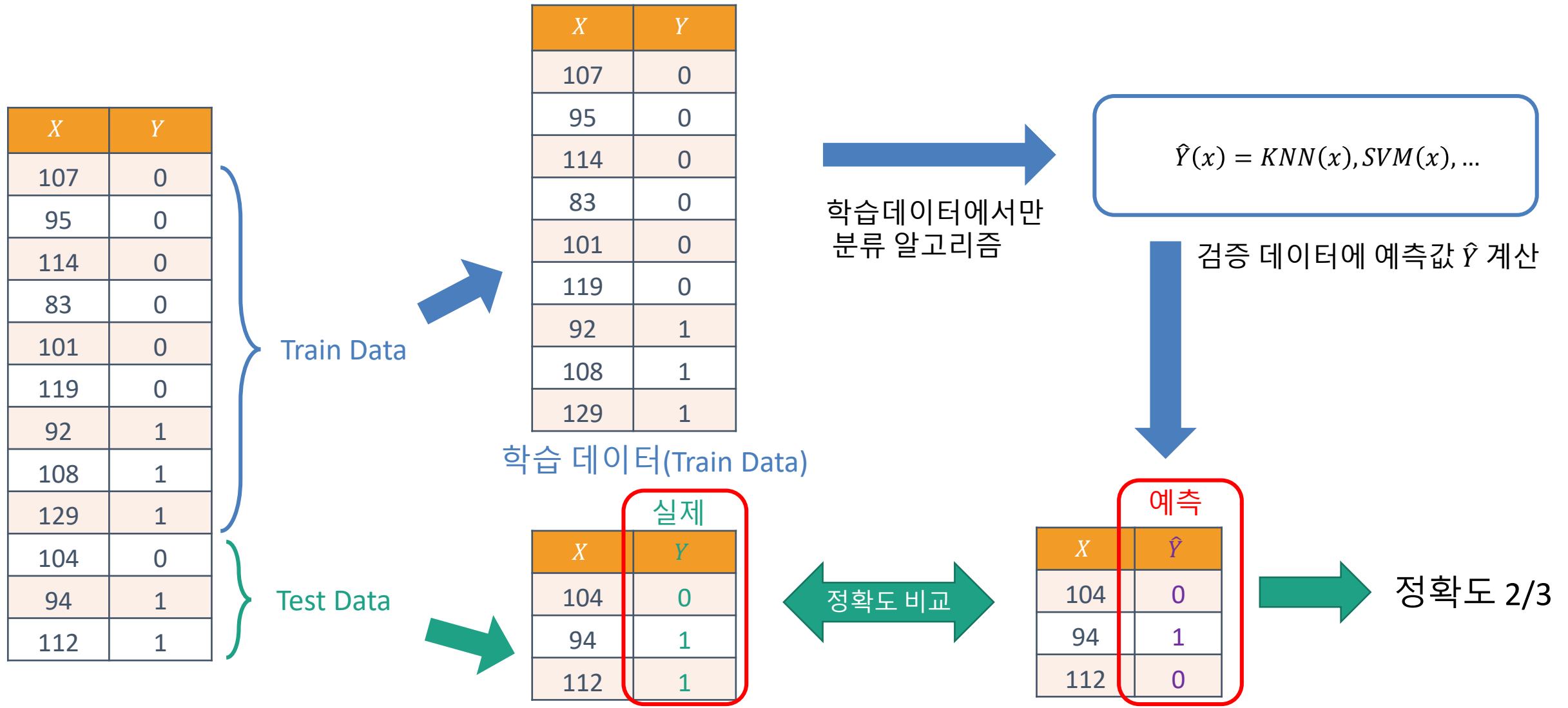
# Contents

---

1. 분류(Classification) 소개
2. K-NN 알고리즘(K-nearest neighbor)
3. 서포트 벡터 머신(Support vector machine, SVM)
4. 로지스틱 회귀(Logistic regression)
5. **분류 알고리즘 성능 평가(Model Evaluation)**
  - 1) 선형 회귀와의 비교: 검증 데이터 셋 분리
  - 2) **분류 정확도(Classification accuracy)**
  - 3) Precision and recall
  - 4) F1-score

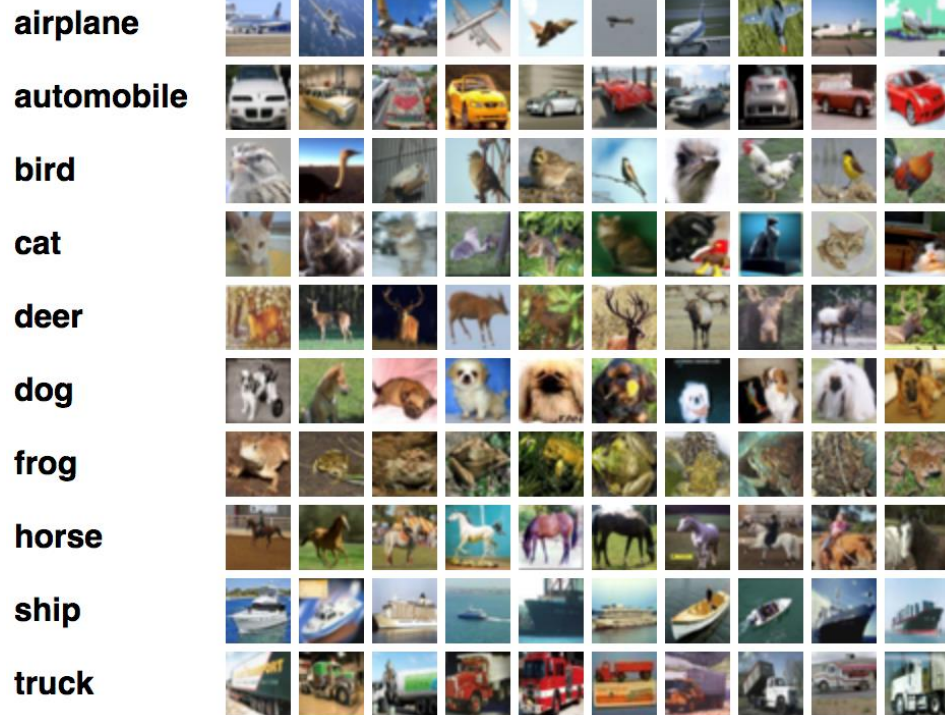
# 분류 정확도(Classification Accuracy)

- 이지 선다 시험문제 20문제 중에서 16문제를 맞혔다 → 80점 → 0.8 accuracy



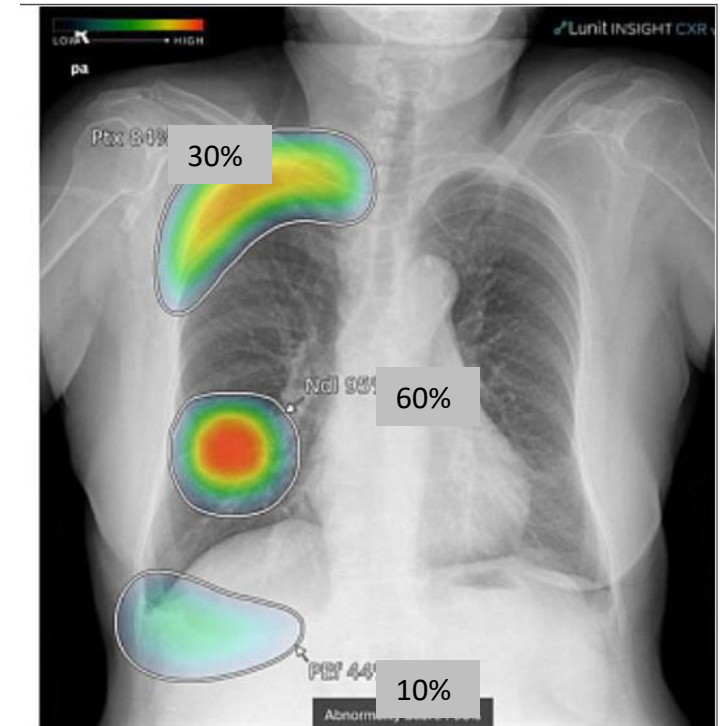
# 분류 정확도가 사용되는 예시

- 가장 무난한 평가기준
- 일반적인 분류 문제의 평가에서 가장 많이 쓰임
  - 각각의 label이 서로 동등한 역할일 때 자주 쓰임 → 두루두루 잘 맞혀야 하는 상황



이미지 분류 문제: 모든 클래스를 두루두루 잘 푸는게 목표

이 경우를 위한  
다른 정확도 기준 존재



암 검진의 경우는 암(1) or 이상없음(0) 에서  
암인 경우를 특히 잘 맞추는게 중요하다

- 예시)
  - 전체 1000명이 검진
  - 10명이 암환자, 990명은 정상
- 알고리즘 1
  - 모든 사람을 정상으로 분류하는 (바보같은) 알고리즘
  - 분류 정확도 =  $\frac{990}{1000} = 99\%$
- 알고리즘 2
  - 50명의 사람을 의심환자, 950명의 사람을 정상으로 분류
    - 하지만, 10명의 암환자를 모두 검진 성공, 정상인 40명은 잘못 검진
  - 분류 정확도 =  $\frac{1000-40}{1000} = 96\%$

# Contents

---

1. 분류(Classification) 소개
2. K-NN 알고리즘(K-nearest neighbor)
3. 서포트 벡터 머신(Support vector machine, SVM)
4. 로지스틱 회귀(Logistic regression)
5. **분류 알고리즘 성능 평가(Model Evaluation)**
  - 1) 선형 회귀와의 비교: 검증 데이터 셋 분리
  - 2) 분류 정확도(Classification accuracy)
  - 3) **Precision and recall**
  - 4) F1-score

# 혼동 행렬(Confusion Matrix)

- $Y$ : 실제 암환자 여부 (1: 암, 0: 정상)
- $\hat{Y}$ : 분류 알고리즘 예측 이상 여부 (1: 암 의심, 0: 정상)
- **Positive or Negative**: 검진에서 이상소견이 나왔는지 아닌지
- **True or False**: 검진 결과가 실제랑 맞는지 아닌지

	$\hat{Y} = 1$	$\hat{Y} = 0$
$Y = 1$	(TP)	(FN)
$Y = 0$	(FP)	(TN)

혼동 행렬

- $(Y = 1, \hat{Y} = 1)$ : 암환자가 실제로 이상 소견을 받은 경우의 수
  - True Positive
- $(Y = 1, \hat{Y} = 0)$ : 암환자가 검진에서 이상 없음을 받은 경우의 수
  - False Negative
- $(Y = 0, \hat{Y} = 1)$ : 정상인이 이상소견을 받은 경우의 수
  - False Positive
- $(Y = 0, \hat{Y} = 0)$ : 정상인이 이상 없음을 받은 경우의 수
  - True Negative

	$\hat{Y} = 1$	$\hat{Y} = 0$
$Y = 1$	3	4
$Y = 0$	5	6

예시) 혼동 행렬

Quiz: 다음 예시에서 분류 정확도는?

$$\text{분류 정확도} = \frac{3+6}{(3+4+5+6)} = \frac{TP+TN}{TP+FP+TN+FN}$$

# Precision and Recall

- Precision: 검출 정확도
  - 실제로 분류 알고리즘이 이상소건이 있을때 진짜 암환자일 확률
  - $P(Y = 1 | \hat{Y} = 1) = \frac{TP}{TP + FP}$
- Recall: 민감도
  - 실제 암환자 중에서 검출로 이상이 나올 비율
  - $P(\hat{Y} = 1 | Y = 1) = \frac{TP}{TP + FN}$
  - 암의 검진에서는 Recall >> Precision
- 둘 다 높은 게 가장 좋다
- 일반적인 경우는? 분류 알고리즘의 목표에 따라 상황마다 다르다

	$\hat{Y} = 1$	$\hat{Y} = 0$
$Y = 1$	(TP)	(FN)
$Y = 0$	(FP)	(TN)

	$\hat{Y} = 1$	$\hat{Y} = 0$
$Y = 1$	3	4
$Y = 0$	5	6

# Contents

---

1. 분류(Classification) 소개
2. K-NN 알고리즘(K-nearest neighbor)
3. 서포트 벡터 머신(Support vector machine, SVM)
4. 로지스틱 회귀(Logistic regression)
5. **분류 알고리즘 성능 평가(Model Evaluation)**
  - 1) 선형 회귀와의 비교: 검증 데이터 셋 분리
  - 2) 분류 정확도(Classification accuracy)
  - 3) Precision and recall
  - 4) **F1-score**



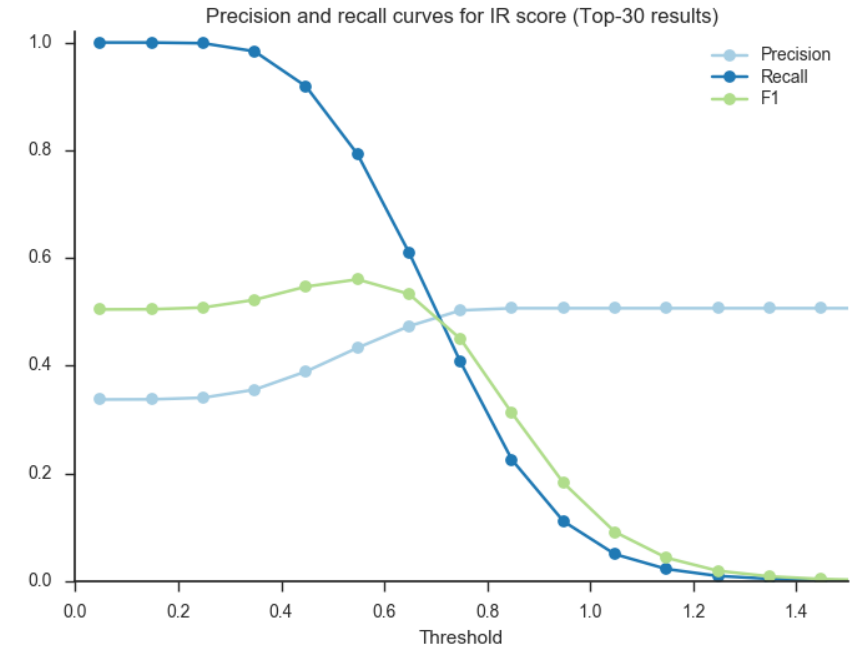
# F1-Score

- Precision 과 Recall를 동시에 고려하는 정확도 기준

- $$\text{F1-score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{(\text{Precision} + \text{Recall})} \text{ (조화 평균)}$$

- 분류 정확도와 비교

- $P(Y = 1) \approx P(Y = 0)$  인 경우 (Balanced label) → 분류 정확도
- $P(Y = 1) \ll P(Y = 0)$  or  $P(Y = 1) \gg P(Y = 0)$  인 경우 (Unbalanced label) → F1-score



---

# Thank you!

Any Questions?

---