동양미래대학교

동양미래대학교 인공지능소프트웨어학과

4차산업혁명 시대의 인공지능
"누구나 이해할 수 있는 인공지능"

Dongyang Mirae University
Dept. Of Artificial Intelligence

DONGYANG MIRAE UNIVERSITY
Dept. of Artificial Intelligence

누구나 이해하는 인공지능

머신러닝 복습

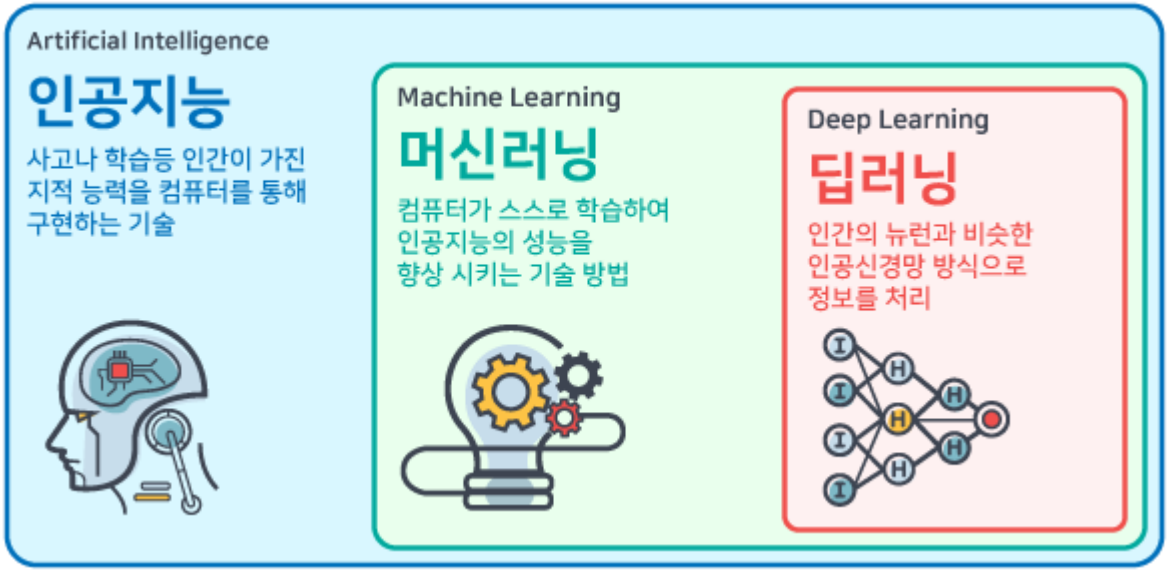
강환수 교수

AI

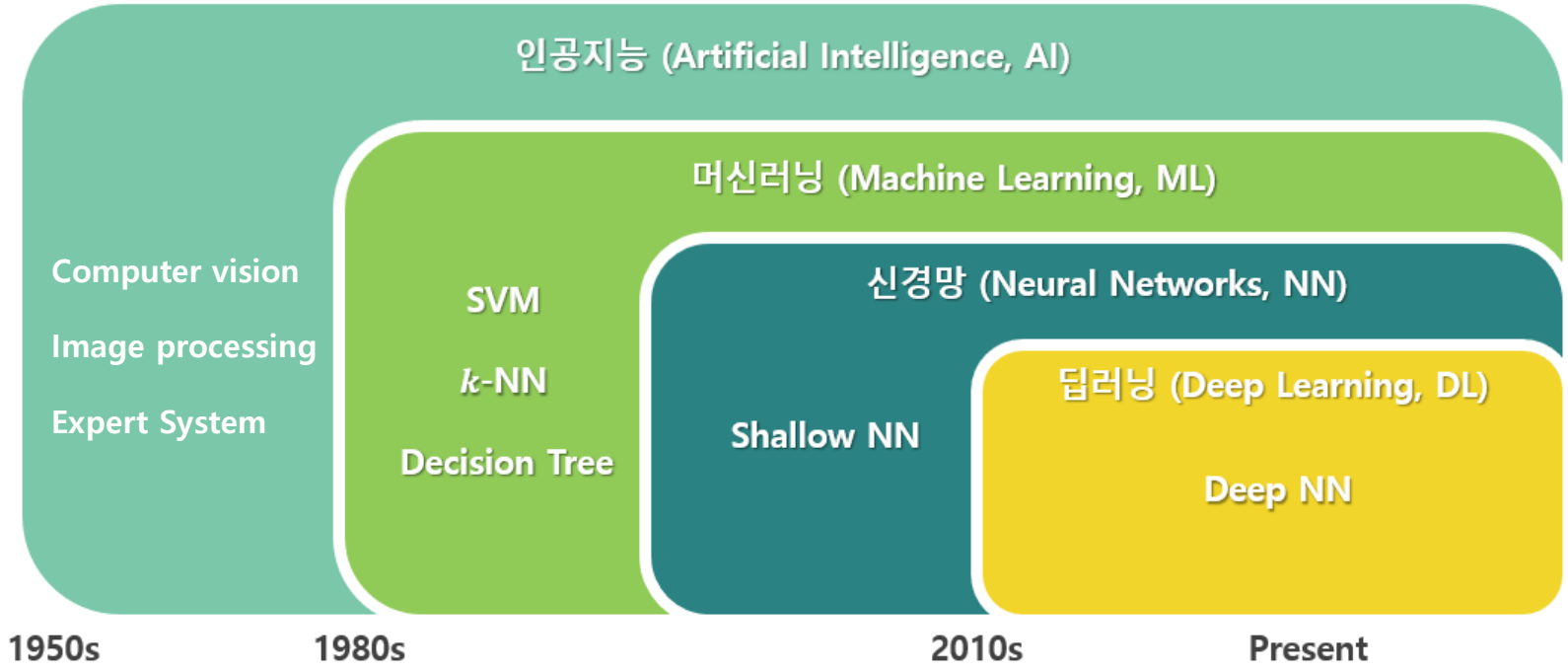
DMUAi
동양미래대학교
인공지능소프트웨어학과

머신러닝

- **회귀와 분류**
- **회귀**
  - 선형 회귀
- **분류**
  - KNN
  - SVM



머신러닝의 다양성


인공지능 (Artificial Intelligence, AI)
머신러닝 (Machine Learning, ML)
신경망 (Neural Networks, NN)
딥러닝 (Deep Learning, DL)
Computer vision
Image processing
Expert System
SVM
k-NN
Decision Tree
Shallow NN
Deep NN
1950s
1980s
2010s
Present

머신러닝 용어의 등장

- **머신러닝 (Machine Learning, ML)**
  - 컴퓨터가 명시적으로 프로그램 되지 않고도 학습할 수 있도록 하는 연구 분야를 말함
  - 용어는 1959년에 아서 사무엘이 학술지 <IBM Journal of Research and Development>에 기고한 논문에서 처음 사용함

[사진출처] https://en.wikipedia.org/wiki/Arthur_Samuel

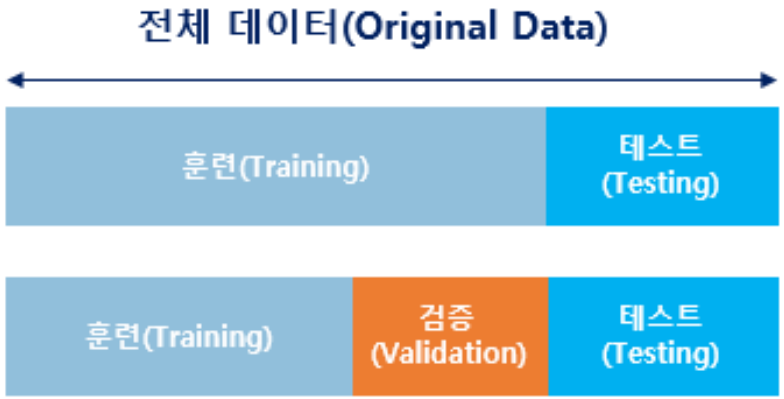
아서 사무엘의 머신러닝 정의

"머신러닝은 컴퓨터가 명시적으로 프로그램되지 않고도 학습할 수 있도록 하는 연구 분야를 말합니다."

(Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed.)

머신러닝 개념(1/3)

- **일반적인 프로그래밍 작업**
  - 입력값에 따라 원하는 결과값이 출력되도록 사람이 내부 동작을 작성
- **머신러닝**
  - 사람이 컴퓨터에게 입력값과 결과값만 충분히 전달해 주면
    컴퓨터가 스스로 입력값과 결과값의 관계를 만족시키는 내부 동작을 찾아냄
    - **학습 데이터 (훈련 데이터, Training Data)**
      - **내부 동작을 만들 때 사용한 데이터 (입력값과 결과값)**
    - **시험 데이터 (Test Data)**
      - **만들어진 내부 동작의 성능을 평가할 때 사용하는 데이터**



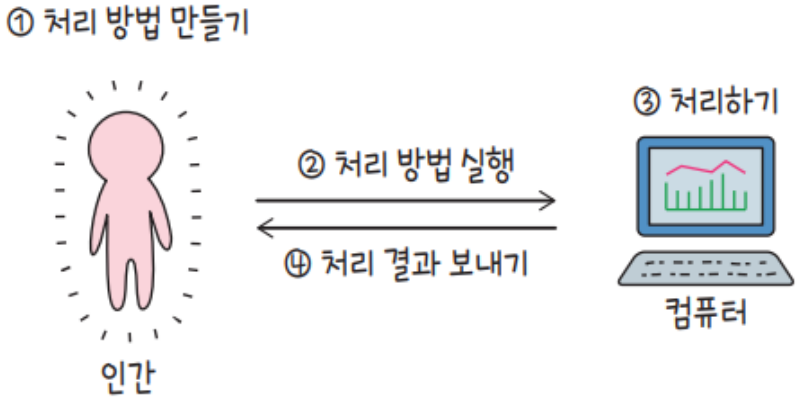
머신러닝 개념(2/3)

- **머신러닝**
  - 여러 개의 입력값과 결과값을 컴퓨터에 제공하기만 하면 이 데이터를 바탕으로 컴퓨터가 스스로 내부 동작을 만들어 냄
    - **이를 위해서 사람은 양질의 많은 학습 데이터를 공급해야 함**

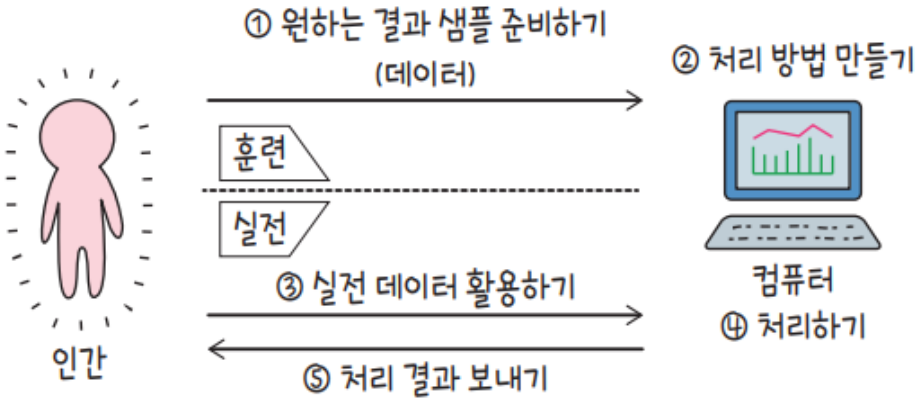
그림 2-2 일반적인 프로그래밍 방식과 머신러닝 방식 비교

머신러닝 개념(3/3)

- **사람의 학습 방식을 흉내 내려는 것이 기계 학습**
  - 주어진 데이터로부터 원하는 결과를 더 효율적으로 정확하게 찾기 위한 학습 방법을 프로그래밍하는 것
    - **학습 방법이 프로그래밍된 기계에 샘플 데이터로 훈련**
      - **훈련을 통해 원하는 결과를 도출할 수 있는 처리 방법을 학습한 후**
    - **새롭게 입력 받은 실전 데이터를 앞서 학습한 방식으로 처리**



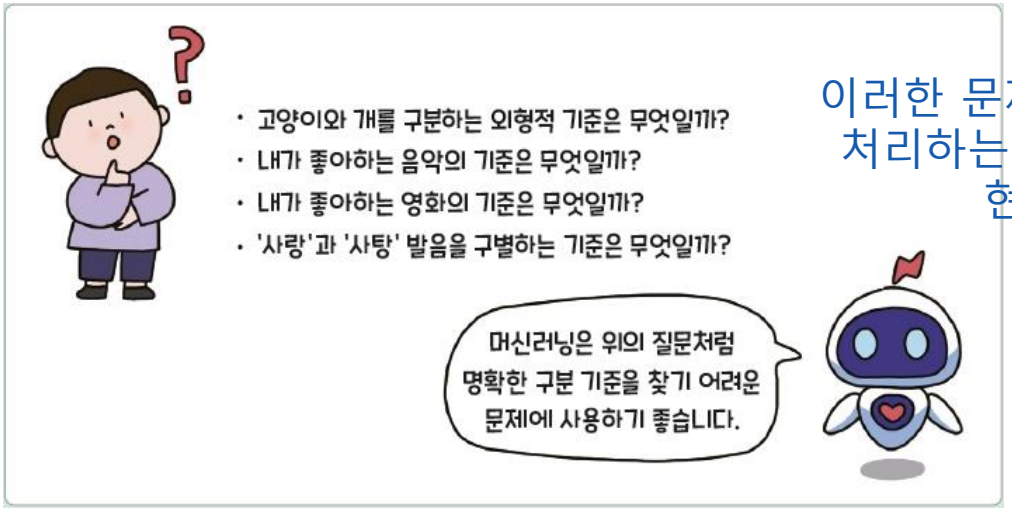
▲ 일반적인 프로그래밍



▲ 기계 학습 프로그래밍

머신러닝은 언제 주로 사용될까? (1/2)

- **명시적으로 알고리즘을 설계하고 프로그래밍 하는 것이 어렵거나 불가능 한 경우에 주로 사용됨**
  - ① 문제에서 나타날 수 있는 경우의 수가 너무 많은 경우
  - ② 규칙 기반 프로그램으로 답을 내기가 어려운 경우
- **예를 들어**
  - ① 바둑 경기에서 모든 경우의 수를 찾아서 if-else와 같은 문장으로 처리하는 것은 거의 불가능함
  - ② 스팸 메일을 자동으로 걸러내는 작업에도 많은 경우가 있기 때문에 프로그래밍하는 것은 거의 불가능함



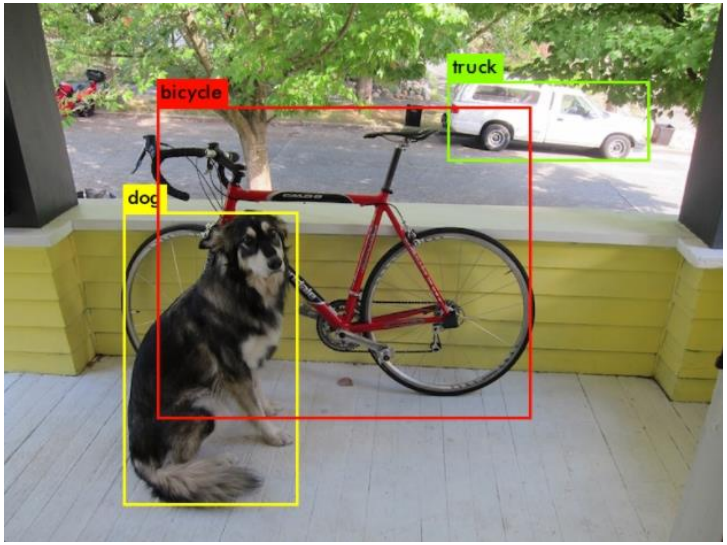
이러한 문제를 if-else 구조를 사용하여 처리하는 프로그램을 만든다는 것은 현실적으로 불가능함

머신 러닝 분류


기계 학습
지도 학습
(Supervised Learning)
원숭이
하마
문제와 정답을 모두 알려 주고
공부시키는 방법
비지도 학습
(Unsupervised Learning)
답을 가르쳐 주지 않고 공부시키는 방법
강화 학습
(Reinforcement Learning)
보상을 통해 상은 최대화, 벌은 최소화하는
방향으로 행위를 강화하는 학습

머신러닝은 언제 주로 사용될까? (2/2)

- **머신러닝의 응용 분야**
  - 주로, 복잡한 데이터들이 있고 이 데이터들에 기반하여 결정을 내려야 하는 분야
  - 머신러닝 모델을 학습(Learning) 시키려면 많은 데이터가 필요하기 때문
    - **머신러닝은 빅데이터(Big Data)와 아주 밀접한 관계가 있음**
  - 예를 들어
    - **객체 검출 (Object Detection)**
    - **음성 인식 (Voice Recognition)**
    - **글자 인식 (Character Recognition)**



머신러닝 모델의 생성 과정

- **머신러닝 모델을 생성**
  - 머신러닝으로 문제를 해결하기 위해서는 문제 적합한 머신러닝 모델을 생성해야 함
    - **문제 속의 데이터를 해결에 잘 설명할 수 있는 머신러닝 모델 선정이 중요**
- **학습과 적용**
  - 모델로부터 학습 데이터에 최적화된 구체적인 함수를 찾아내는 과정
    - **과정을 학습(Learning)이라고 함**
  - 학습된 모델(함수)를 실제 문제에 적용

- **시험 공부 시간으로 시험 성적을 예상하는 문제**
  - 머신러닝으로 해결
    - **머신러닝 모델 생성 과정을 개념적으로 표현하면 [그림 2-4]와 같음**

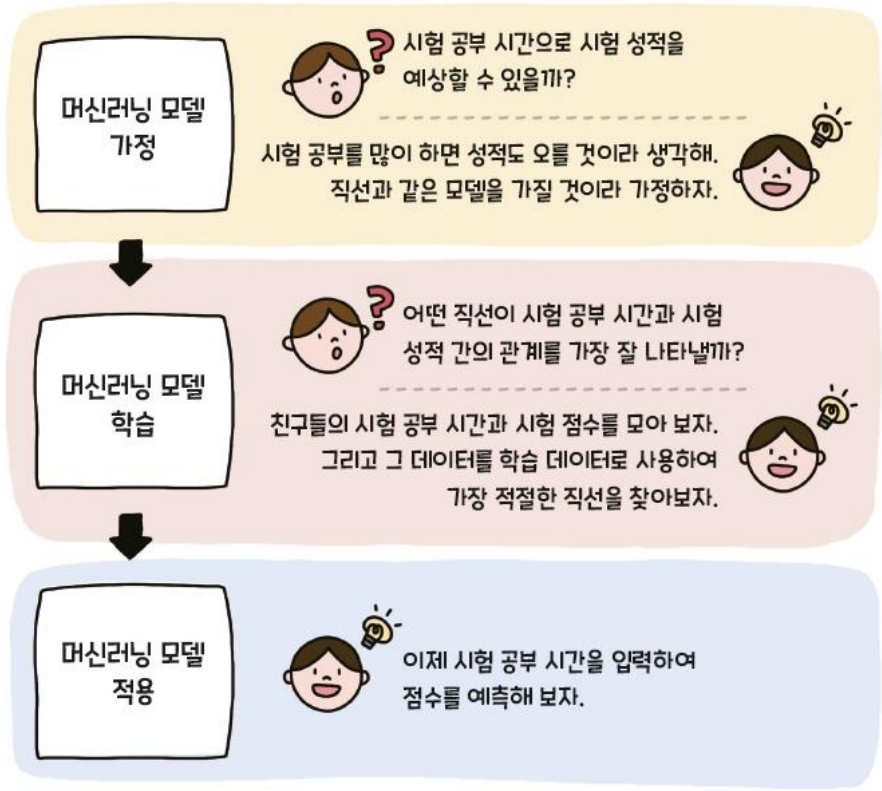


그림 2-4 머신러닝 모델 생성 과정의 예

머신러닝 구현 과정 예제 (1/4)

- **사람의 키를 입력했을 때 몸무게를 추측하는 작업을 머신러닝으로 구현하면?**

① 일반적으로 키가 커지면 몸무게도 늘어날 것이라 가정해,
[몸무게 = a × 키 + b]와 같은 직선의 방정식을 만들어 머신러닝 모델로 가정

간단하지만 실제 사용하는 모델

$$몸무게 = a \times 키 + b$$
$$y = a \times x + b$$

그림 2-5 머신러닝 구현 과정 예제의 학습 데이터 © Hanbit Academy Inc.

		x	y
No.	이름	키 (cm)	몸무게 (kg)
1	김민성	100	30
2	박다인	120	40
3	윤이안	130	45
4	최서연	160	60
5	문진승	190	75

머신러닝 구현 과정 예제 (2/4)

- **사람의 키를 입력했을 때 몸무게를 추측하는 작업을 머신러닝으로 구현하면?**
  - ② 머신러닝 모델로 가정한 직선의 방정식은 아직 기울기 a와 y절편 b의 값이 결정되지 않은 상태
    - **학습 데이터를 확보하여 최적화된 직선을 구함**
  - 학습 데이터와 최적화된 직선을 구한다는 것
    - **학습 데이터와 오차가 가장 적은 직선의 기울기 (= a)와 y절편 (= b)을 구한다는 의미**

		x	y
No.	이름	키 (cm)	몸무게 (kg)
1	김민성	100	30
2	박다인	120	40
3	윤이안	130	45
4	최서연	160	60
5	문진승	190	75

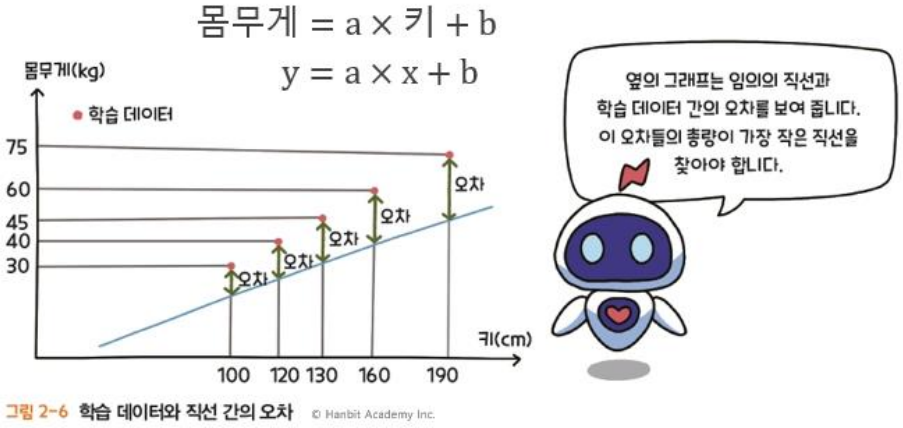


그림 2-6 학습 데이터와 직선 간의 오차 © Hanbit Academy Inc.

교수님,

난 몰라요!

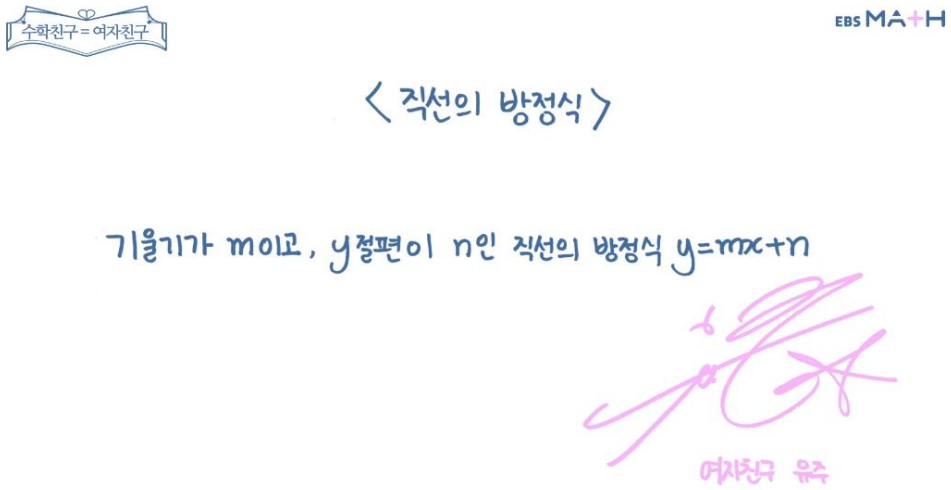
직선의 기울기 (= a)와 y절편 (= b)

걱정하지 말아요!

마법, 정보의 보고 인터넷이 있어요.

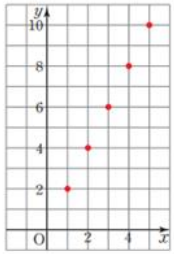
직성의 방정식, ebs 수학

https://www.ebsmath.co.kr/resource/rscView?cate=11003&cate2=11028&cate3=11088&rscTpDscd=RTP10&grdCd=HGRD01&sno=28911&historyYn=study



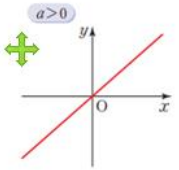
❸ 일차함수와 그 그래프

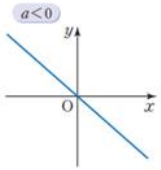
1. 함수 $y=f(x)$에서
$y=ax+b$ (a, b는 상수, $a\neq0$)
와 같이 y가 x에 대한 일차식으로 나타내어질 때, 이 함수 $y=f(x)$를 x에 대한 **일차함수**라고 한다.

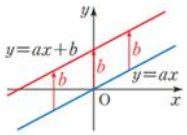


2. 함수 $y=2x$에 대하여 x의 값과 y의 값을 순서쌍 (x, y)로 나타내면 $(1, 2)$, $(2, 4)$, $(3, 6)$, $(4, 8)$, $(5, 10)$이고, 이를 좌표로 하는 점을 좌표평면 위에 나타내면 오른쪽 그림과 같다.
이와 같이 함수 $y=f(x)$에서 x와 그 함숫값 $f(x)$로 이루어진 순서쌍 $(x, f(x))$를 좌표로 하는 점 전체를 그 **함수의 그래프**라고 한다.

3. 일차함수 $y=ax$(a는 상수, $a\neq0$)의 그래프는 원점을 지나는 직선이다.





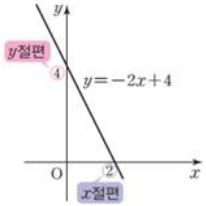


한 도형을 일정한 방향으로 일정한 거리만큼 이동하는 것을 평행이동이라고 한다.

또 일차함수 $y=ax+b(b\neq0)$의 그래프는 일차함수 $y=ax$의 그래프를 y축의 방향으로 b만큼 평행이동한 직선이다.

4. 함수의 그래프가 x축과 만나는 점의 x좌표를 그 그래프의 **x절편**, y축과 만나는 점의 y좌표를 그 그래프의 **y절편**이라고 한다. 이를테면 일차함수 $y=-2x+4$의 그래프의 x절편은 2이고, y절편은 4이다.

5. 일차함수 $y=ax+b$(a, b는 상수, $a\neq0$)에서 x의 값의 증가량에 대한 y의 값의 증가량의 비율은 항상 일정하며, 그 값은 x의 계수 a와 같다. 이 증가량의 비율 a를 일차함수 $y=ax+b$의 그래프의 **기울기**라고 한다.



확인 문제 2

일차함수 $y=3x-6$의 그래프의 x절편과 y절편을 구하시오.

| 수학으로 풀어 보기 |

$y=3x-6$에서
$y=0$일 때, $0=3x-6$, $x=2$
$x=0$일 때, $y=3\times0-6$, $y=-6$
따라서 x절편은 2이고, y절편은 -6이다.

답 x절편: 2, y절편: -6

머신러닝 구현 과정 예제 (3/4)

- **사람의 키를 입력했을 때 몸무게를 추측하는 작업을 머신러닝으로 구현하면?**

  ② 머신러닝 모델로 가정한 직선의 방정식은 아직 기울기 a와 y절편 b의 값이 결정되지 않은 상태로, 학습 데이터를 확보하여 최적화된 직선을 구함

  – 학습 결과, 직선의 방정식 [몸무게 = (0.5 × 키) – 20]이 해당 학습 데이터에 최적화된 함수임

		x	y
No.	이름	키 (cm)	몸무게 (kg)
1	김민성	100	30
2	박다인	120	40
3	윤이안	130	45
4	최서연	160	60
5	문진승	190	75

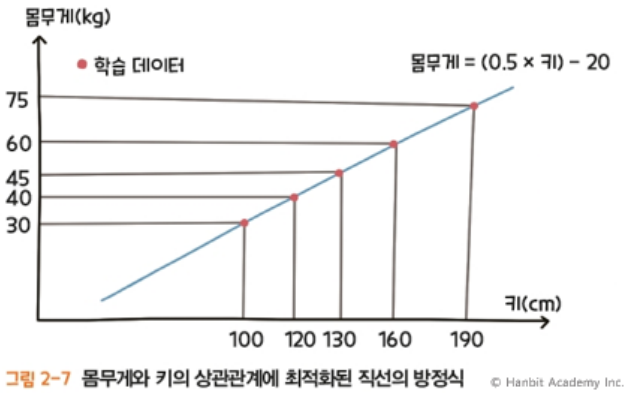


그림 2-7 몸무게와 키의 상관관계에 최적화된 직선의 방정식 © Hanbit Academy Inc.

$$\text{몸무게} = a \times \text{키} + b$$

$$y = a \times x + b$$

↓

최적화 (Optimization)

$$a = 0.5$$

$$b = -20$$

↓

$$y = 0.5 \times x - 20$$

머신러닝 구현 과정 예제 (4/4)

- **사람의 키를 입력했을 때 몸무게를 추측하는 작업을 머신러닝으로 구현하면?**

③ 최적화된 머신러닝 모델을 실제 문제에 적용해 확인

- **학습 데이터에는 없던 키 180cm를 방정식에 대입하면 예측되는 몸무게는 [70kg = (0.5×180) – 20]이 나옴**

		x	y
No.	이름	키 (cm)	몸무게 (kg)
1	김민성	100	30
2	박다인	120	40
3	윤이안	130	45
4	최서연	160	60
5	문진승	190	75

Q) 키가 180cm 인 사람은 몸무게가 몇 kg일까?

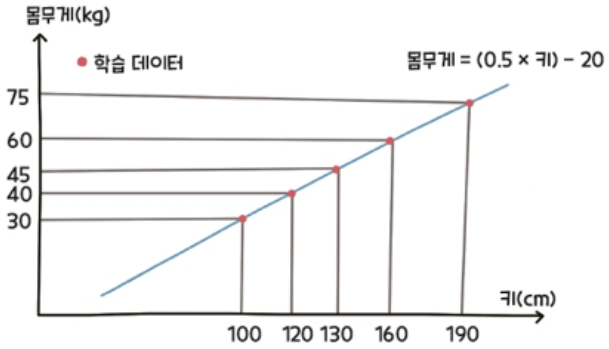


그림 2-7 몸무게와 키의 상관관계에 최적화된 직선의 방정식 © Hanbit Academy Inc.

$$y = 0.5 \times x - 20$$

↓

$$x = 180 \text{ 대입}$$

↓

$$y = 0.5 \times 180 - 20$$

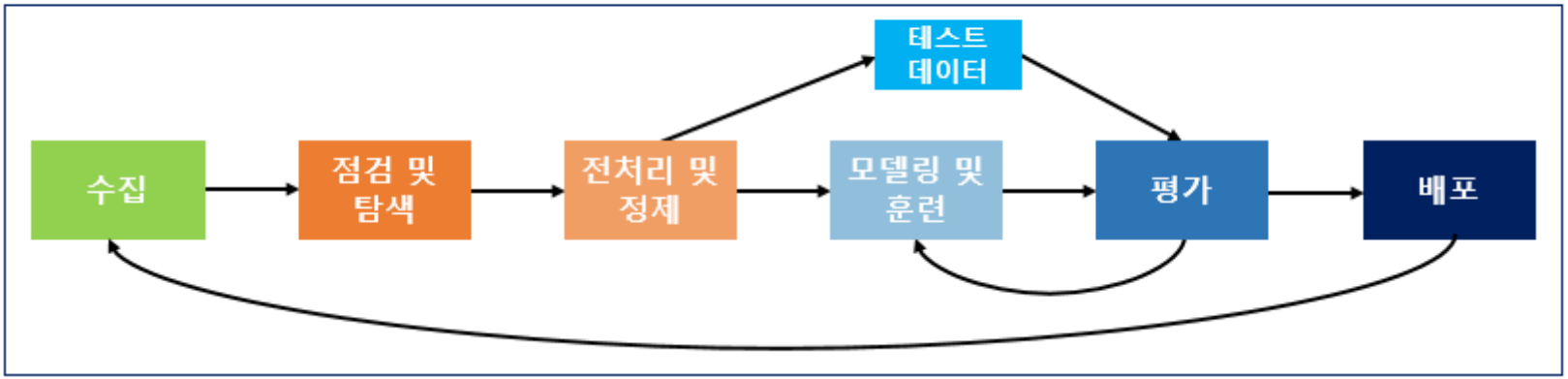
$$= 90 - 20$$

$$= 70$$

머신 러닝 워크플로

- 작업 과정

머신 러닝 워크플로우



회귀와 분류 (regression and classification)

DONGYANG MIRAE UNIVERSITY
Dept. of Artificial Intelligence

누구나 이해하는 인공지능

회귀와 분류

강환수 교수

AI

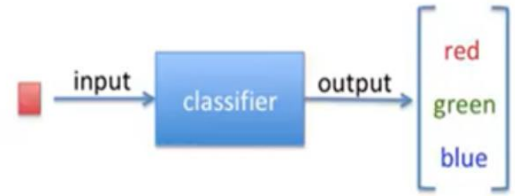
Ebs 수학
https://www.ebssw.kr/info/intron/infoTchmtrHeaderView.do?tabType=AI

DMUAi
동양미래대학교
인공지능소프트웨어학과

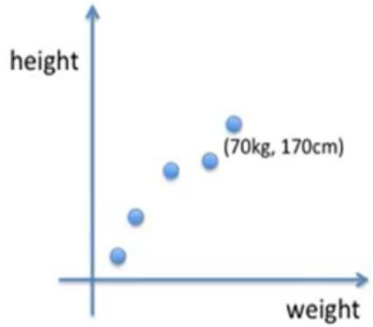
회귀(regresson)와 분류(classification)

- **회귀 모델**
  - 연속적인 값(실수)을 예측
    - **캘리포니아의 주택 가격이 얼마인가요?**
    - **사용자가 이 광고를 클릭할 확률이 얼마인가요?**
    - **키 (Height) 정보가 주어졌을 때, 몸무게를 예측**
    - **공부한 시간 정보가 주어졌을 때, 시험 성적 예측**
    - **커피를 몇 잔 마셨는지에 대한 정보가 주어졌을 때, 수면 시간 예측**
    - **사과의 전년도 수확량과 날씨, 고용 인원 수 등으로 올해 수확량 예측**
- **분류 모델**
  - 불연속적인 값(유한개 이산 값)을 예측
    - **주어진 이메일 메시지가 스팸인가요, 스팸이 아닌가요?**
    - **이 이미지가 강아지, 고양이 또는 햄스터의 이미지인가요?**

Classification VS Regression



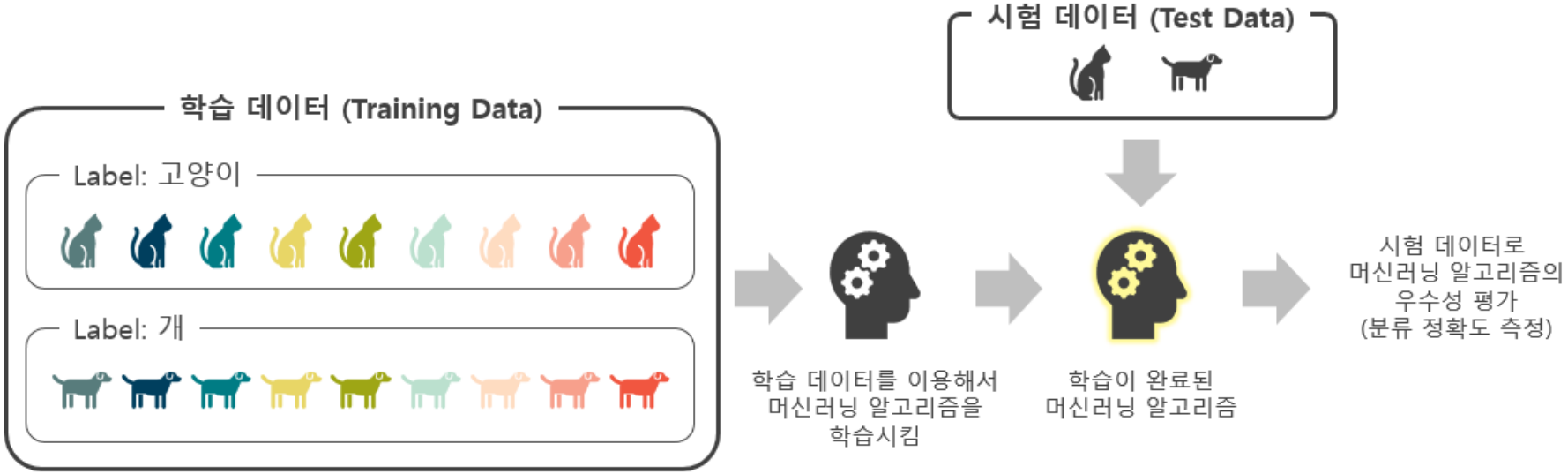
classify input into categorical output



how tall is he if his weight is 80kg?

분류 (Classification)

- **정답(레이블)이 포함된 데이터를 학습하여**
  - 유사한 성질을 갖는 데이터끼리 분류한 후
    - **새로 입력된 데이터가 어느 그룹에 속하는지를 찾아내는 기법**
  - 어떤 입력 데이터가 들어오더라도
    - **학습에 사용한 레이블(Label) 중 하나로 결과값을 결정**
  - 레이블(Label)이 이산적인 (Discrete) 경우 (즉, [0, 1, 2, 3, …]와 같이 유한한 경우)



이진분류와 다중분류

이진 분류 (Binary Classification)

- 데이터를 2개의 그룹 (Class)으로 분류

Label: 고양이

Label: 개

"고양이" or "개"
둘 중에 하나로 분류

다중 분류 (Multiclass Classification)

- 데이터를 3개의 그룹 (Class) 이상으로 분류

Label: 고양이

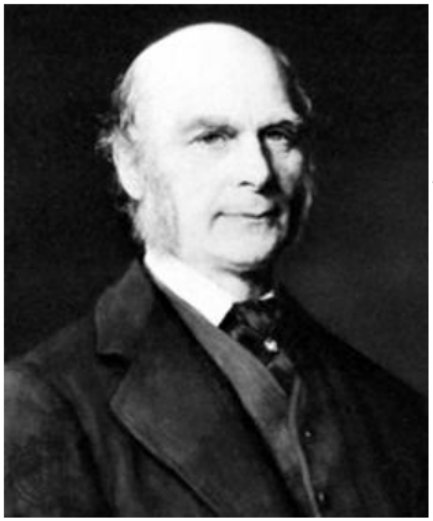
Label: 개

Label: 토끼

"고양이" or "개" or "토끼"
셋 중에 하나로 분류

회귀의 어원

- **회귀 분석(regression analysis)**
  - 관찰된 연속형 변수들에 대해 두 변수 사이의 모형을 구한 뒤 적합도를 측정해 내는 분석 방법
  - 회귀분석은 시간에 따라 변화하는 데이터나 어떤 영향, 가설적 실험, 인과 관계의 모델링 등의 통계적 예측에 이용



[사진출처] https://en.wikipedia.org/wiki/Francis_Galton

프랜시스 골턴

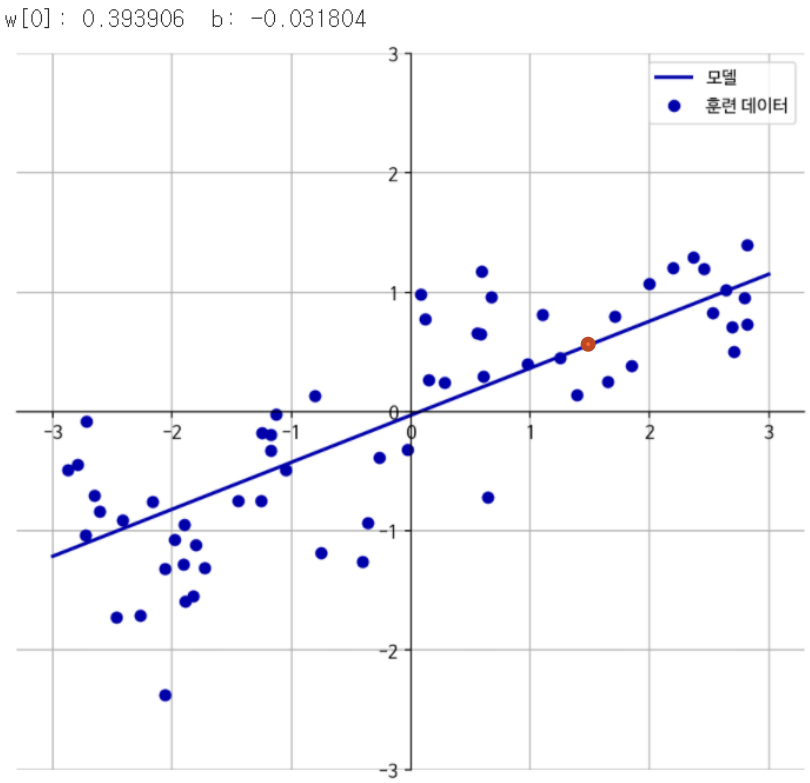
- 아버지와 자식의 키를 분석함
- 사람의 키 (Height)는 세대를 거듭할 수록 평균에 가까워지는 경향이 있다는 것을 발견
- 키가 큰 아버지의 자식은 아버지보다 키가 작고,
  키가 작은 아버지의 자식은 아버지보다 키가 크다
- 즉, 세대를 거듭할 수록 큰 키는 작아지고, 작은 키는 커지고 평균에 수렴한다
- 이를 프랜시스 골턴은 "평균으로 돌아간다 (=회귀)"라고 표현함

회귀(영어: regress 리그레스[*])의 원래 의미

- **변수 사이의 관계를 분석하는 방법을 역사적인 이유 때문에**
  - "회귀 (Regression)" 라고 부름
- **19세기, 통계학자이자 인류학자인 프랜시스 골턴 (Francis Galton)이 처음 사용**
  - 옛날 상태로 돌아가는 것을 의미
  - 프랜시스 골턴은 "평균으로의 회귀(regression to the mean)"
    - **부모의 키와 아이들의 키 사이의 연관 관계를 연구**
      - **부모와 자녀의 키 사이에는 선형적인 관계가 있고 키가 커지거나 작아지는 것보다는**
      - **전체 키 평균으로 돌아가려는 경향이 있다는 가설을 세웠으며**
      - **이를 분석하는 방법을 "회귀분석"이라고 함**
    - **이러한 경험적 연구 이후, 칼 피어슨은 아버지와 아들의 키를 조사한 결과를 바탕으로 함수 관계를 도출하여 회귀분석 이론을 수학적으로 정립**

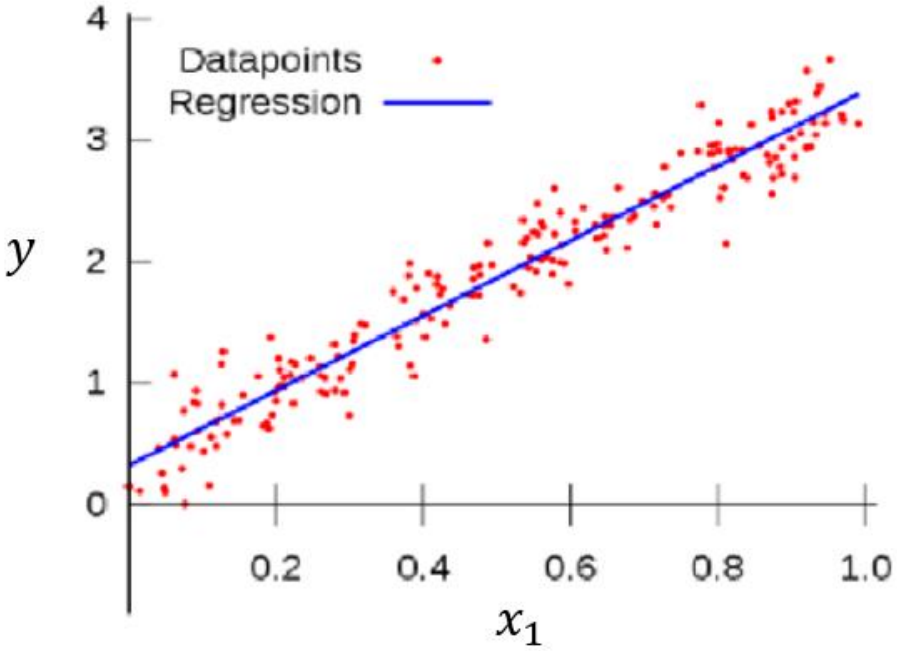
선형 회귀 (Linear Regression) (1/4)

- 데이터로 부터 예측
  - X로 y를 예측, 특징 수: 1
  - 직선으로 예측
- 테스트 데이터, 붉은 점의 y(타깃)?



선형 회귀 (Linear Regression) (2/3)

- **학습 데이터 (Training Data)로부터**
  - 종속 변수 y와 한 개 이상의 독립 변수 x와의 선형 상관 관계를 모델링하는 방법
    - **쉽게 이해하면, 학습 데이터를 잘 표현하는 직선 하나를 찾아내겠다는 의미**
      - **입력이 하나이면 "직선"이 됨**



[사진출처] https://bangu4.tistory.com/100

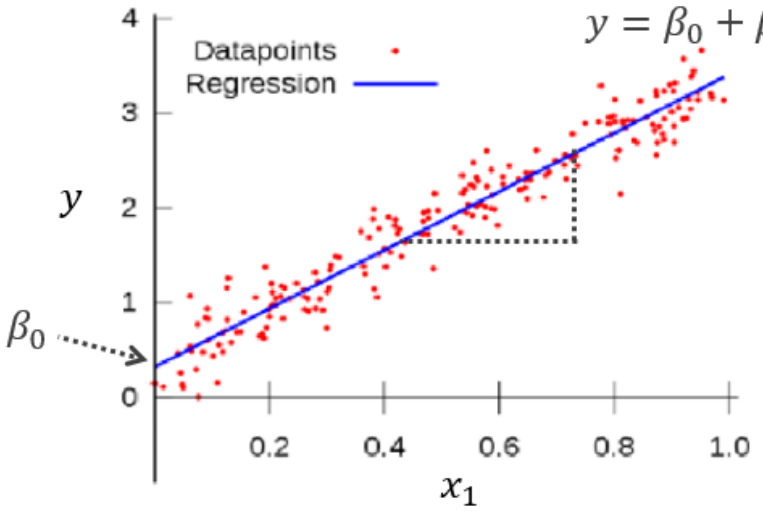
빨간색 점들이 학습 데이터이고,

이 데이터들을 잘 표현하는
파란색 직선을

찾아내는 것이 "선형 회귀"가 하는
역할임

선형 회귀 (Linear Regression) (3/4)

- **독립 변수 x의 개수에 따라 선형 회귀는 아래와 같이 분류됨**
  - 단순 선형 회귀 (Simple Linear Regression)
    - **독립 변수 x의 개수가 1개**
      - **샘플 문제 제외하곤 실제 문제에서는 거의 없는 경우**
  - 다중 선형 회귀 (Multiple Linear Regression)
    - **독립 변수 x의 개수가 2개 이상**



[사진출처] https://bangu4.tistory.com/100

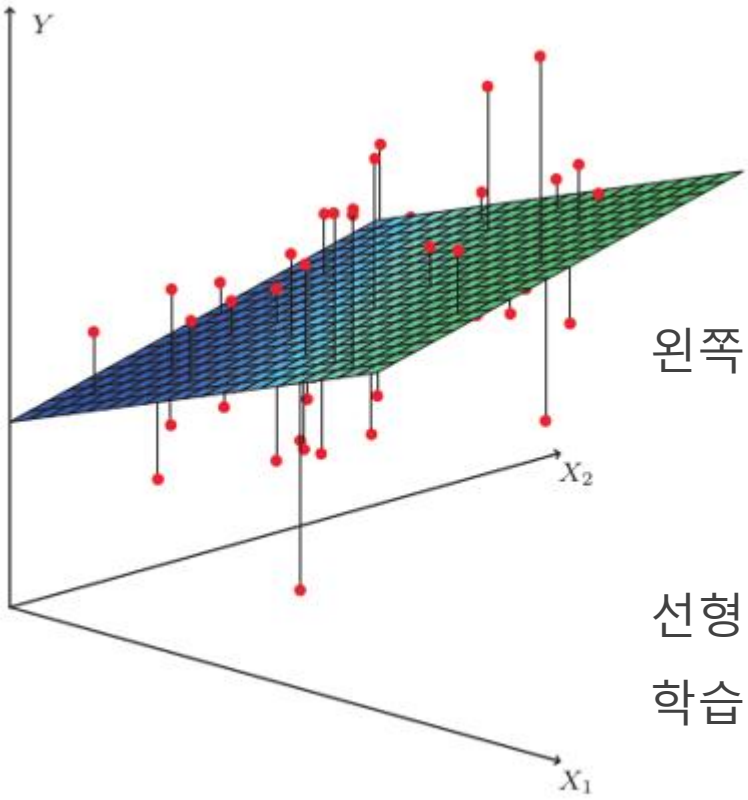
왼쪽 예제는 독립 변수의 개수가 1개인 경우임

$$y = \beta_0 + \beta_1 \times x_1$$

위 직선에서 기울기 $(= \beta_1)$ 값과 y절편 $(= \beta_0)$ 값을 찾는 것이 선형 회귀 알고리즘에서 수행하는 동작

선형 회귀 (Linear Regression) (4/4)

- **독립 변수 x의 개수에 따라 선형 회귀는 아래와 같이 분류됨**
  - 단순 선형 회귀 (Simple Linear Regression): 독립 변수 x의 개수가 1개
  - 다중 선형 회귀 (Multiple Linear Regression): 독립 변수 x의 개수가 2개 이상



[사진출처] https://p829911.github.io/2020/01/16/3.2.1/

독립 변수(입력의 차수)가
여러 개이면 그만큼 복잡

왼쪽 예제는 독립 변수의 개수가 2개 (x_1, x_2)인 경우임

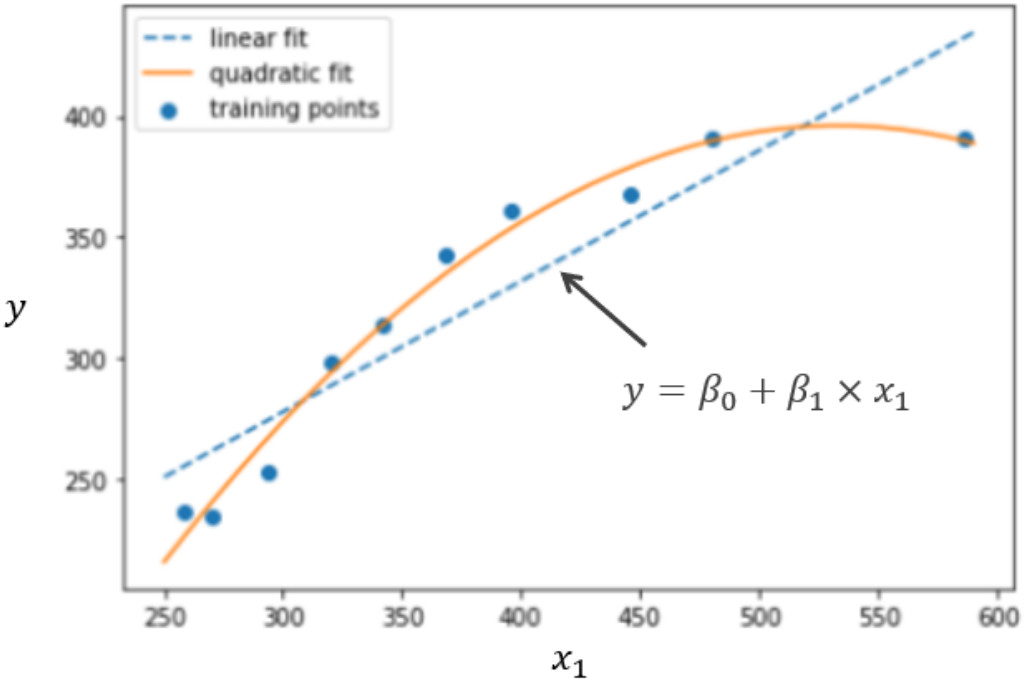
$$y = \beta_0 + \beta_1 \times x_1 + \beta_2 \times x_2$$

선형 회귀 알고리즘 수행을 통해서,

학습 데이터를 가장 잘 표현 하는 β_0, β_1, β_2 값을 찾아냄

다항 회귀 (Polynomial Regression) (1/2)

- **선형 회귀의 단점**
  - 학습 데이터 내, 종속 변수 y와 독립 변수 x 사이의 상관 관계
    - **선형이 아닐 수 있음!**



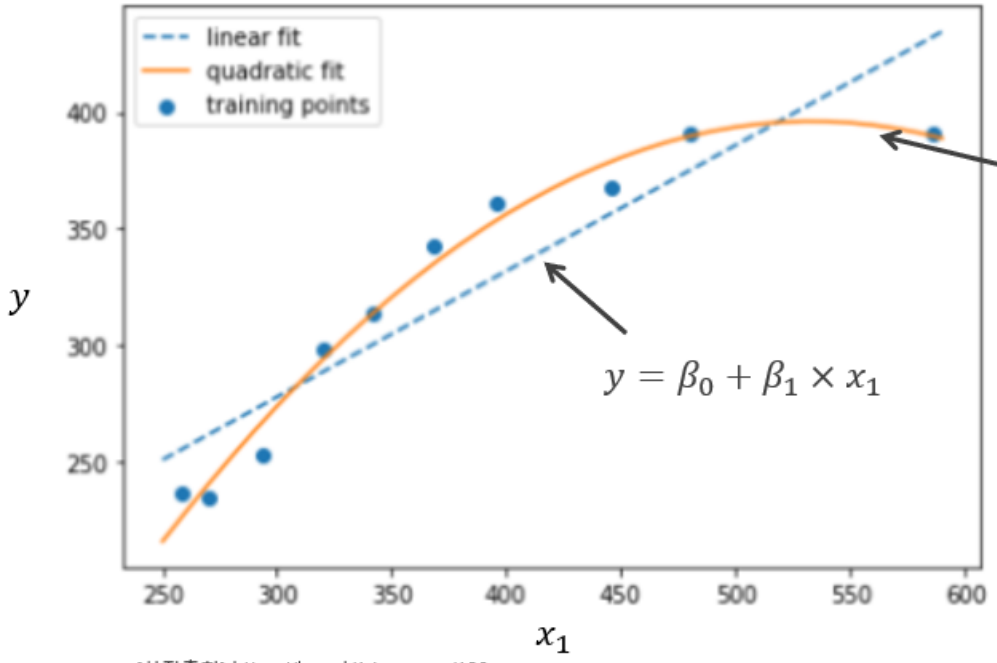
[사진출처] https://hongl.tistory.com/128

학습 데이터 ●가 선형 회귀 모형 (점선)으로 잘 표현되지 않고 있다 (오차가 크다)

오히려 주황색 실선이 학습 데이터 ●를 잘 표현하고 있다 (오차가 작다)

다항 회귀 (Polynomial Regression) (2/2)

- **각 독립 변수 x에 대한 고차원의 다항식을 이용**
  - 종속 변수 y의 관계를 비선형적(Non-linear)으로 모델링하는 방법



[사진출처] https://hongl.tistory.com/128

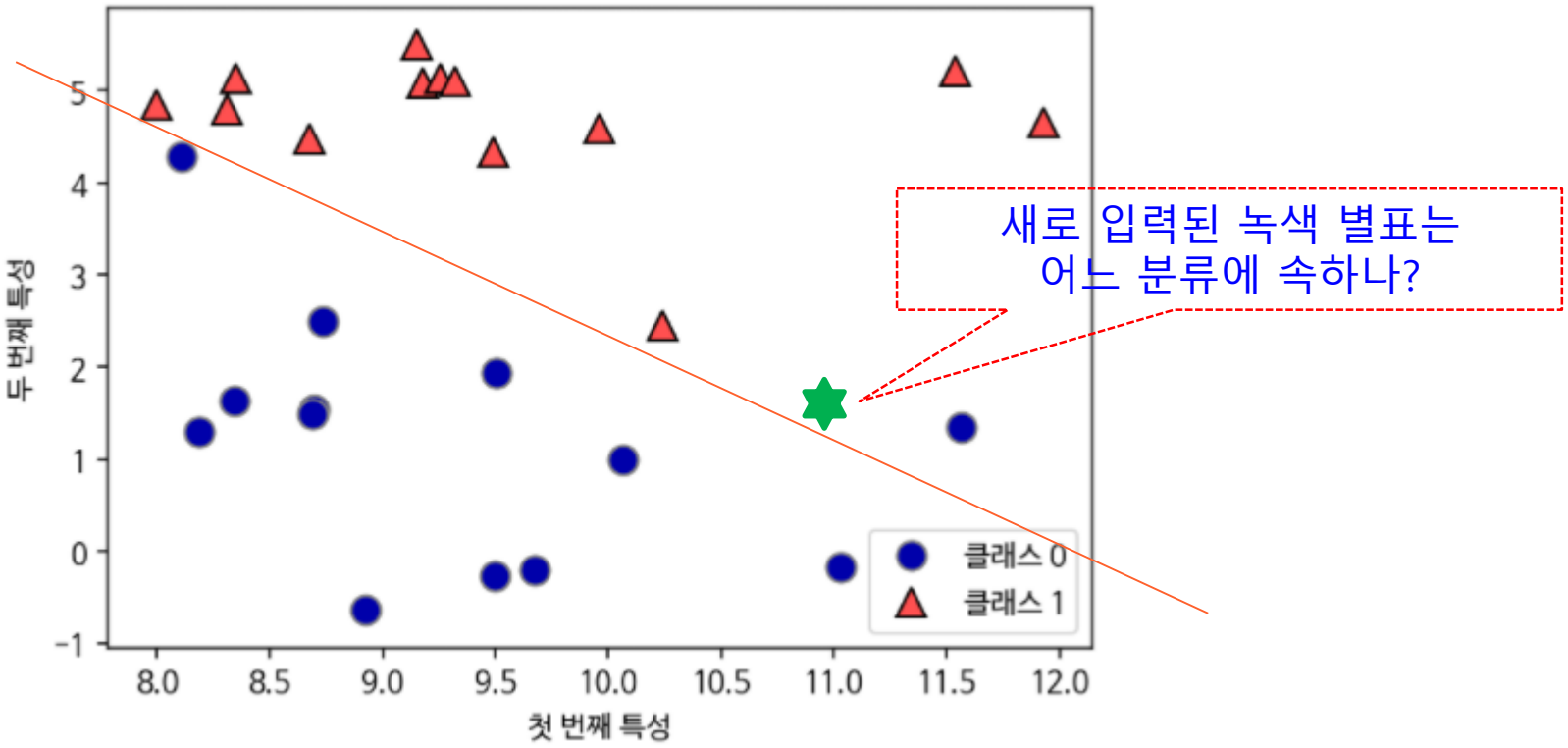
왼쪽 예제는 독립 변수의 개수가 1개인 경우임

$$y = \beta_0 + \beta_1 \times x_1 + \beta_2 \times x_1^2 + \cdots + \beta_n \times x_1^n$$

다항 회귀 알고리즘 수행을 통해서,
학습 데이터를 가장 잘 표현 하는
$\beta_0, \beta_1, \cdots, \beta_n$ 값을 찾아냄

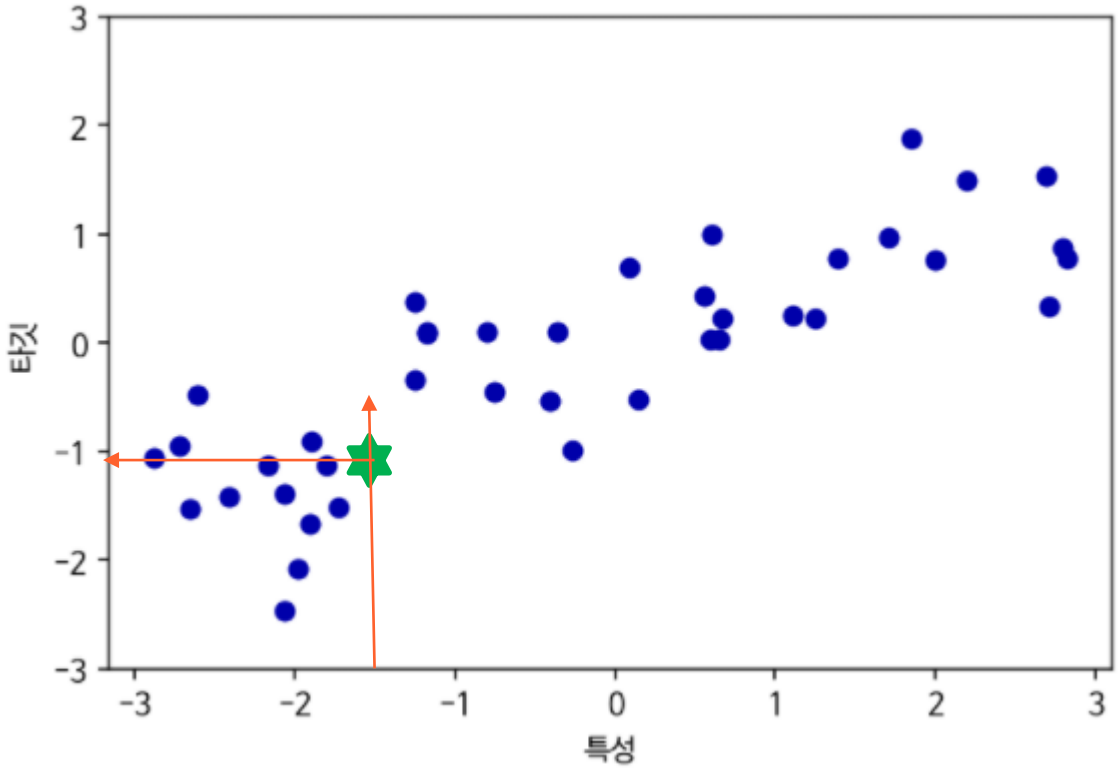
분류 문제

- 특성이 2개이며, # of classes=2인 자료



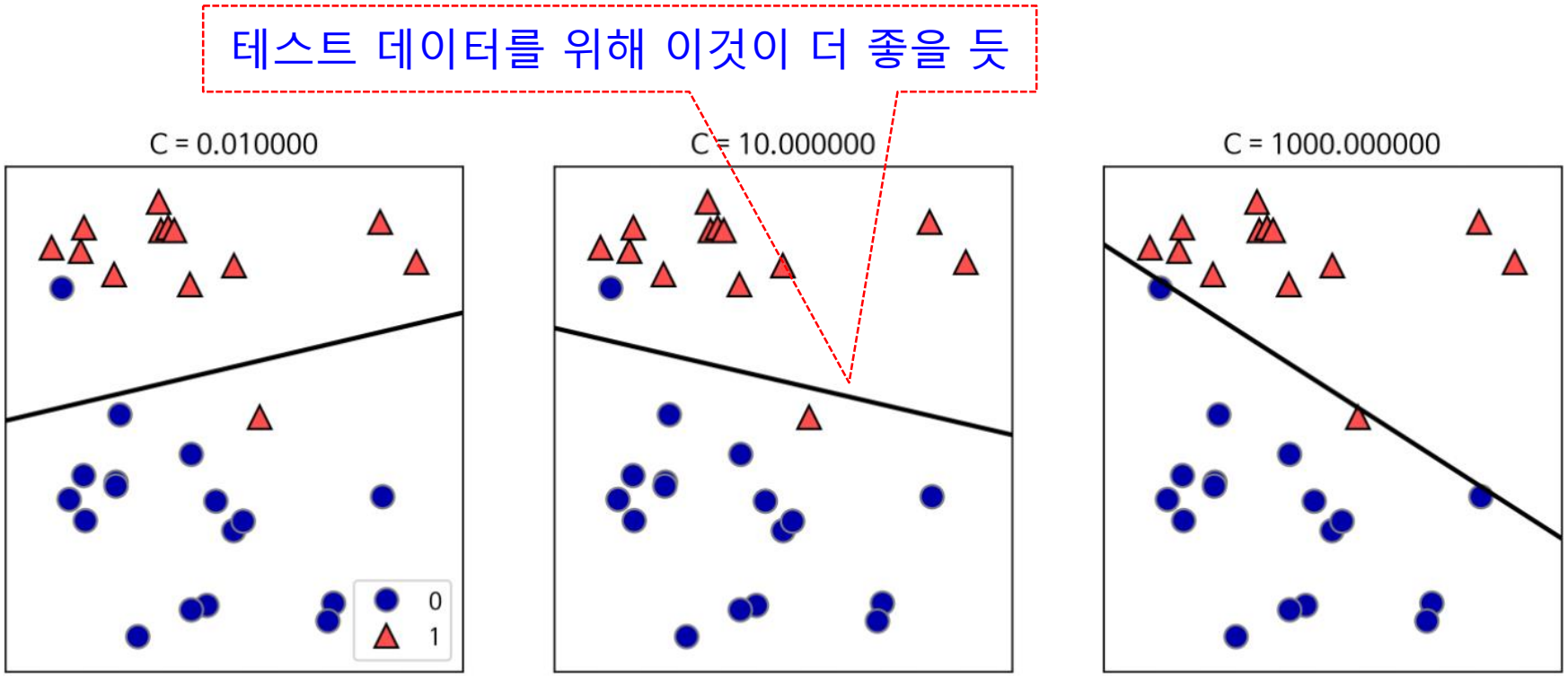
회귀 문제

- 특성이 1개(X축)이며, 타겟(label)이 연속된 값으로 Y축으로 표시



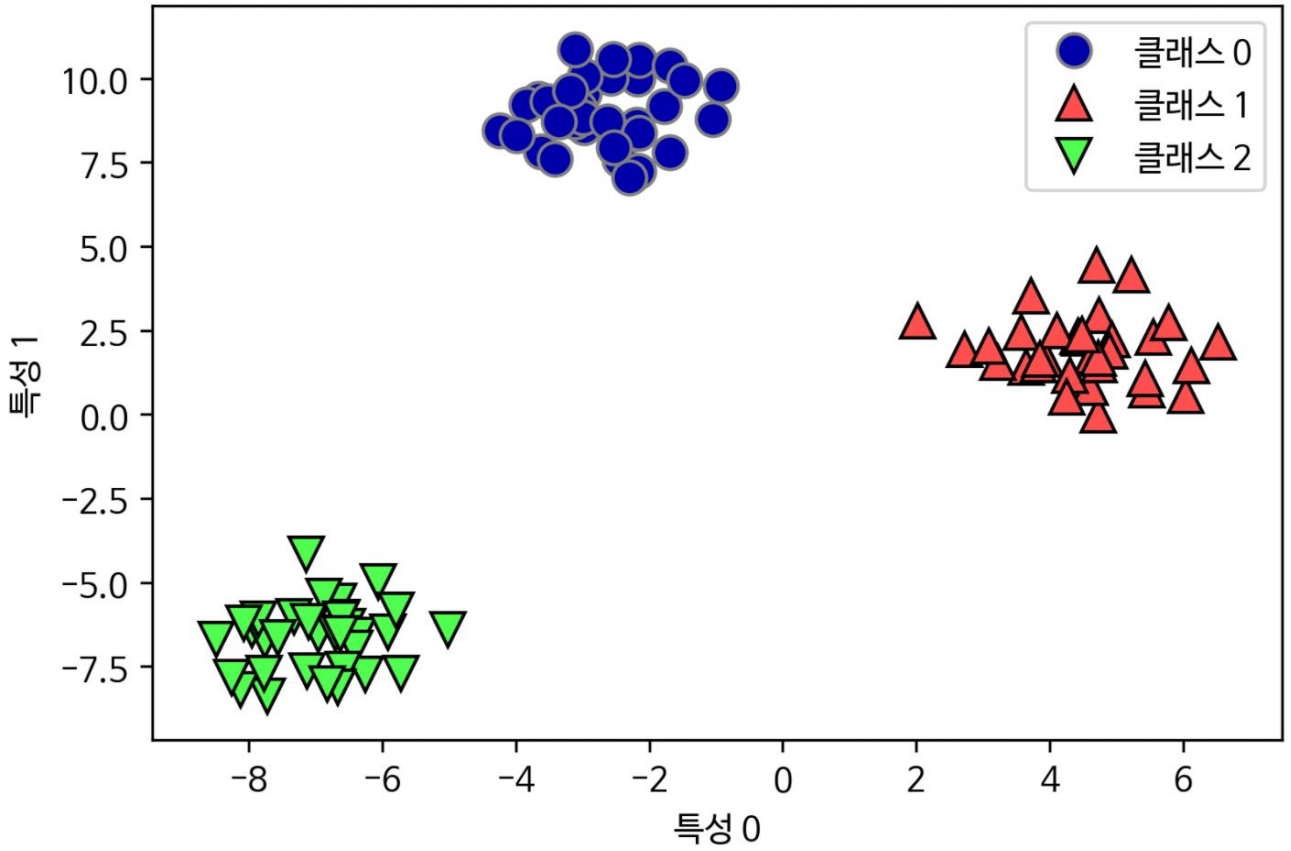
이진 분류

- 직선으로 분류하는 방법



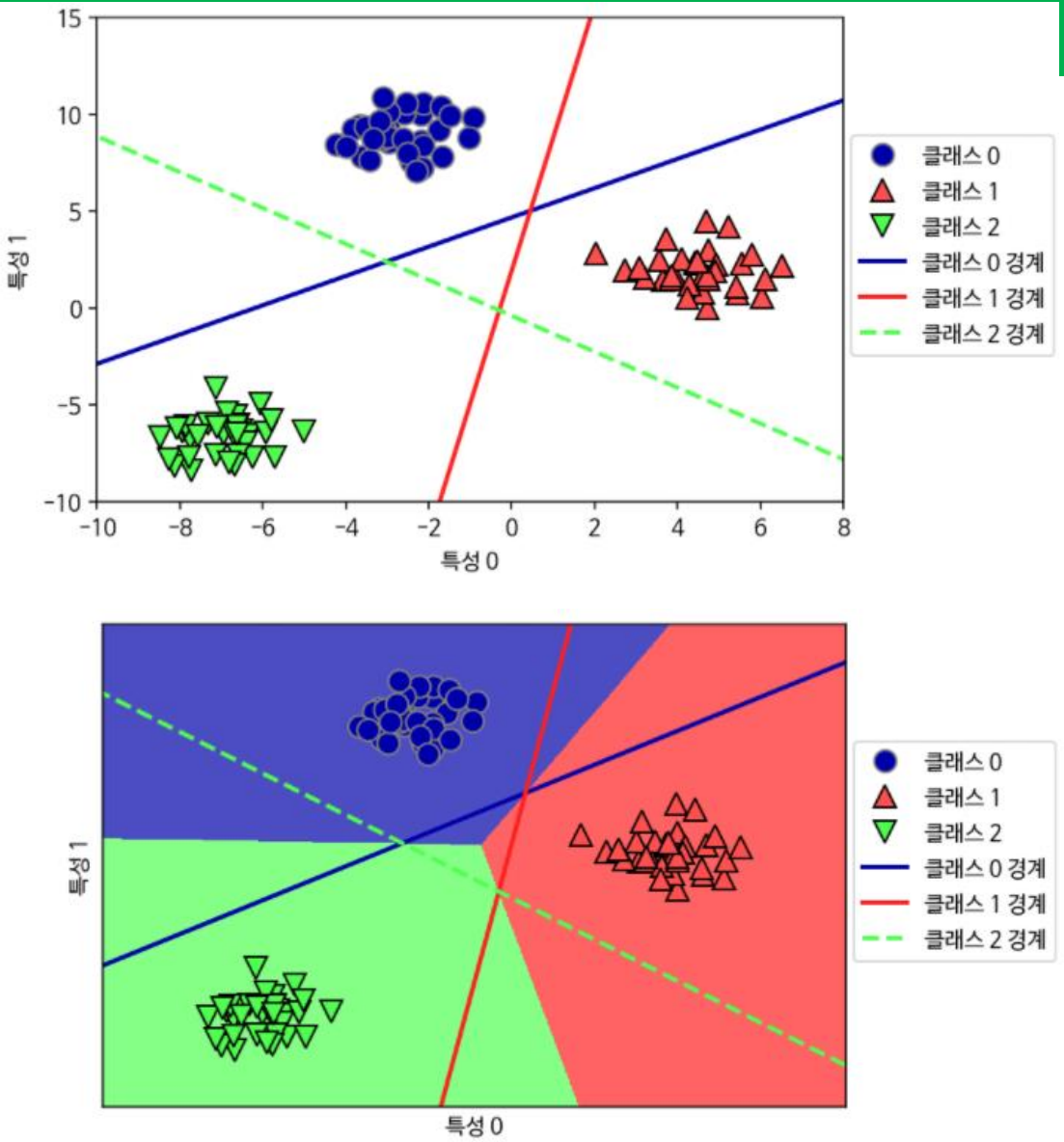
다중 클래스 분류

- **삼중 클래스 분류**
  - 이진 분류를 3 번 이용



삼중 클래스 분류

- **직선 3개로 분류**
  - 영역으로 구분



DONGYANG MIRAE UNIVERSITY
Dept. of Artificial Intelligence

누구나 이해하는 인공지능

머신러닝 절차

강환수 교수

Ai

DMUAi
동양미래대학교
인공지능소프트웨어학과

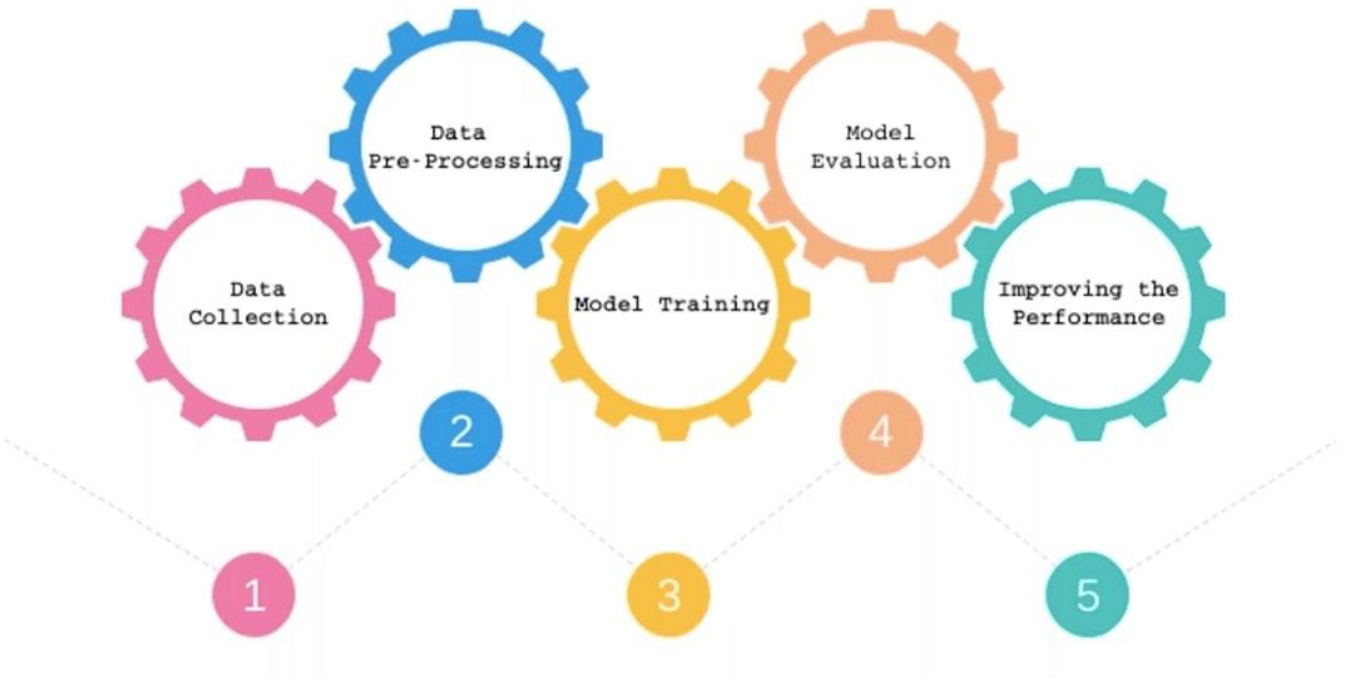
머신러닝 수행과정

그림 6.27 ▸ 머신러닝 수행 과정

머신 러닝 절차

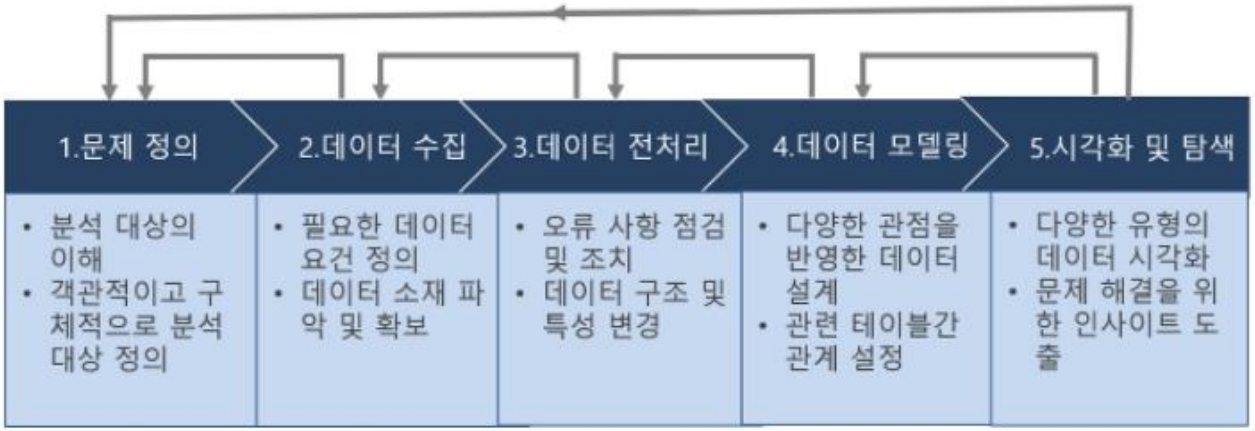
General Machine learning algorithm flow

altud



인터넷 자료

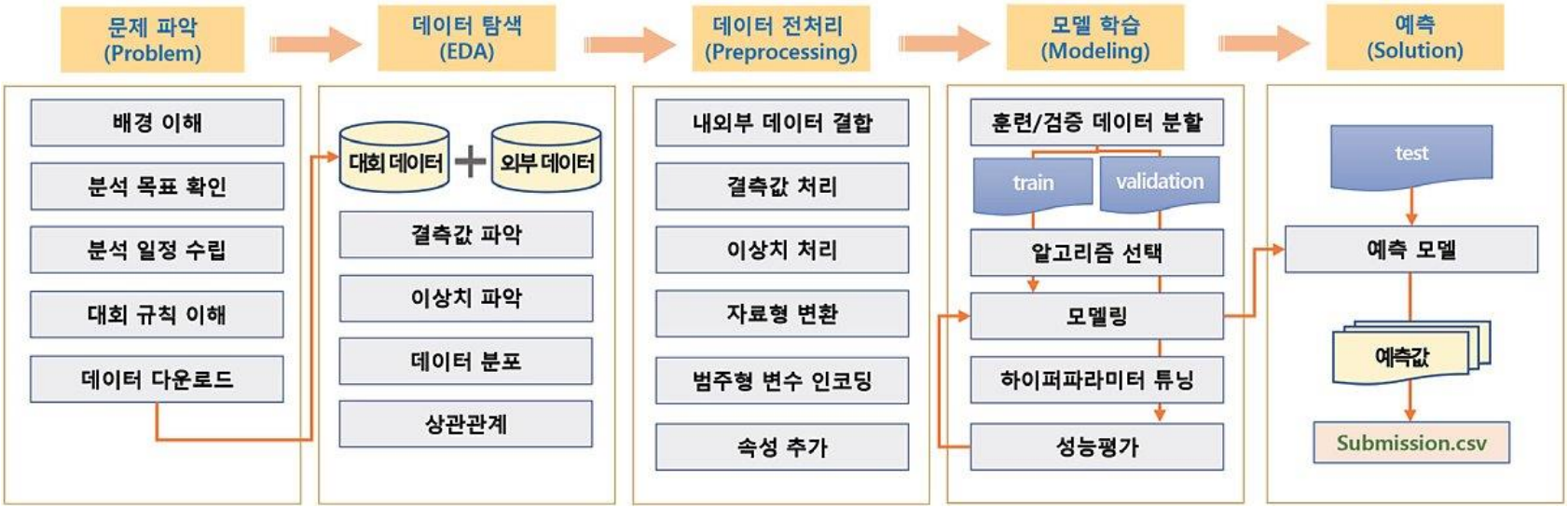
- https://brunch.co.kr/@data/10



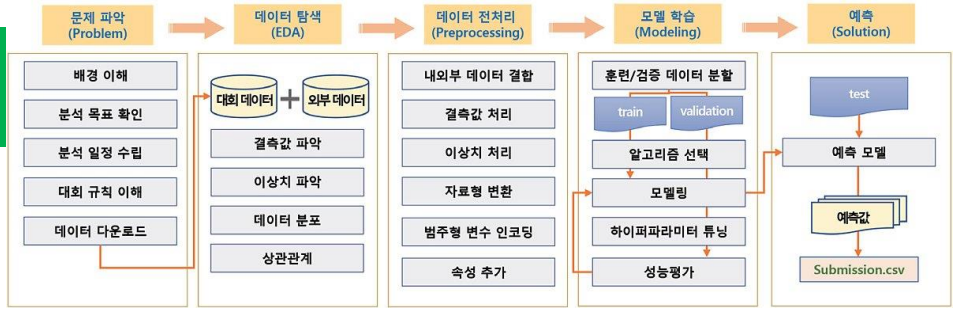
공공데이터 분석 절차

교재

- https://m.post.naver.com/viewer/postView.naver?volumeNo=30304207&memberNo=15488377



문제 파악과 데이터 탐색

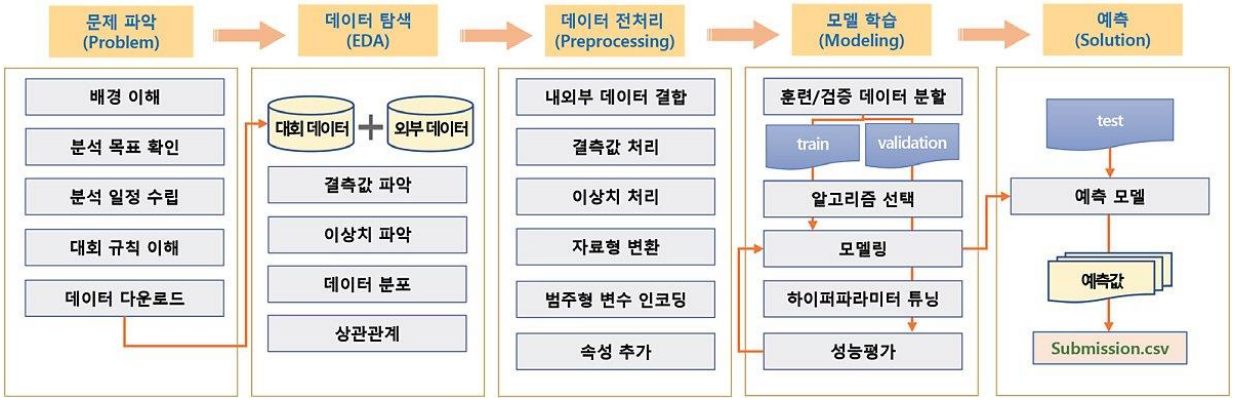


- **1 문제 파악**
  - 경진 대회에서 주어지는 문제를 파악
    - **해당 분야의 도메인 지식을 습득하는 것이 중요**
  - 데이터 분석의 목표를 설정하고, 분석 방법과 일정을 수립
  - 대회 규칙을 자세히 읽고 반드시 지켜야 함
    - **입상 순위에 들었더라도 규칙 위반으로 실격 처리될 수 있기 때문**

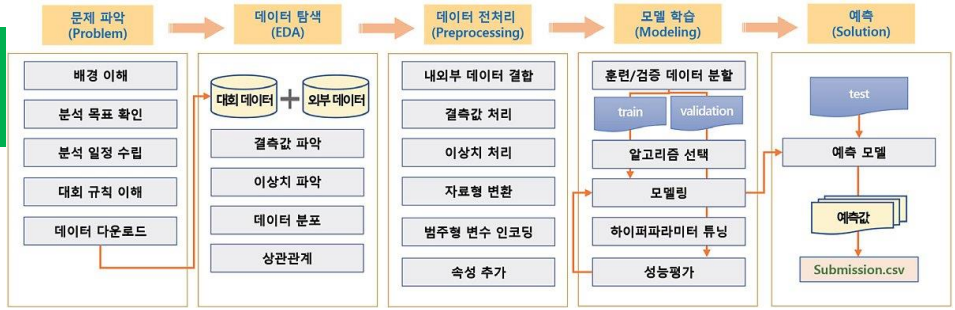
- **2 데이터 탐색**
  - 내부 데이터
    - **대회에서 제공하는 데이터를**
  - 외부 데이터
    - **공공데이터포털 등 외부에서 수집하는 데이터**
    - **대회에서 허용하는 범위에서 외부 데이터를 적극적으로 활용**
  - 데이터를 읽어 들이고 데이터의 이상 유무를 확인
    - **데이터가 누락된 결측값(missing value)**
    - **정상 범위를 벗어난 이상치(outlier)가 있는지 확인**
  - 데이터 구조 및 특성을 파악하고 데이터의 분포와 상관관계를 탐색

데이터 전처리

- **3 데이터 전처리 :**
  - 데이터 탐색이 끝나면 모델 학습이 가능한 형태로 데이터를 정리하는 단계가 필요
    - **내부 데이터와 외부 데이터를 병합하고,**
    - **데이터 탐색 단계에서 확인한 결측값과 이상치를 처리**
  - 필요한 경우 자료형을 변환
    - **특히 머신러닝 및 딥러닝 모델은 숫자형 데이터를 입력으로 받기 때문에**
      - **문자열(범주형) 데이터나 이미지 데이터를 숫자형으로 변환하는 작업이 필요**
  - 새로운 속성을 추가하거나 불필요한 속성을 제거



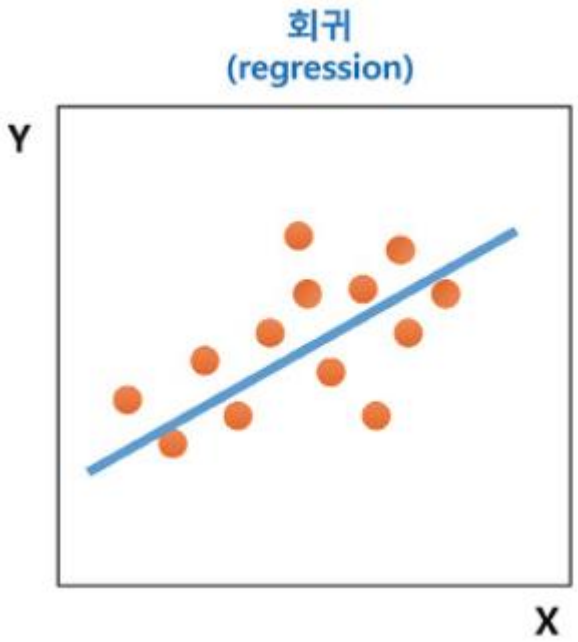
모델 학습과 예측



- **4 모델 학습**
  - 훈련 데이터(train data)와 검증 데이터(validation data)를 구분
    - **모델 학습에 필요한 훈련 데이터(train data)와 모델 성능을 평가하기 위한 검증 데이터(validation data)를 구분**
  - 예측 알고리즘을 선택하고 모델을 설계
    - **훈련 데이터를 입력하여 모델을 학습**
    - **검증 데이터를 입력하여 학습을 마친 모델의 성능을 평가**
  - 모델 성능 점검
    - **성능을 높일 수 있도록 하이퍼파라미터(hyperparameter)를 튜닝**
    - **최종 모델을 선택**
- **5 예측**
  - 예측해야 하는 테스트 데이터(test data)를 모델에 입력
    - **모델이 예측한 값을 제출용(submission) 파일의 형식에 맞게 정리**
  - 파일을 제출하면 주최 측에서 산출해 주는 평가 점수를 리더보드에서 확인
    - **로컬 환경에서 산출한 모델 성능 점수와 리더보드 점수를 비교하여 모델의 일반화 성능을 확인**
    - **리더보드 점수를 올릴 수 있는 방향으로 모델을 수정하고, 다시 제출하는 과정을 반복**
  - 최종 파일을 선택해서 제출

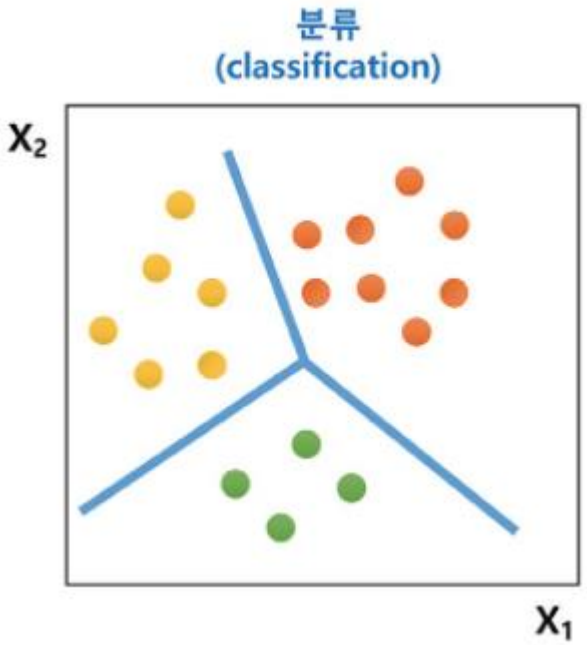
복습: 회귀와 분류 (1/3)

- **회귀 문제**
  - 설명 변수(X)와 목표 변수(Y) 사이의 회귀관계식을 찾는 문제
    - **목표 변수는 연속적인 값을 갖는 숫자형 데이터**
  - 예
    - **과거의 주가 데이터를 가지고 미래 주가를 예측**
    - **자동차 배기량이나 연식 등 중고차 정보를 이용하여 가격을 예측하는 문제**



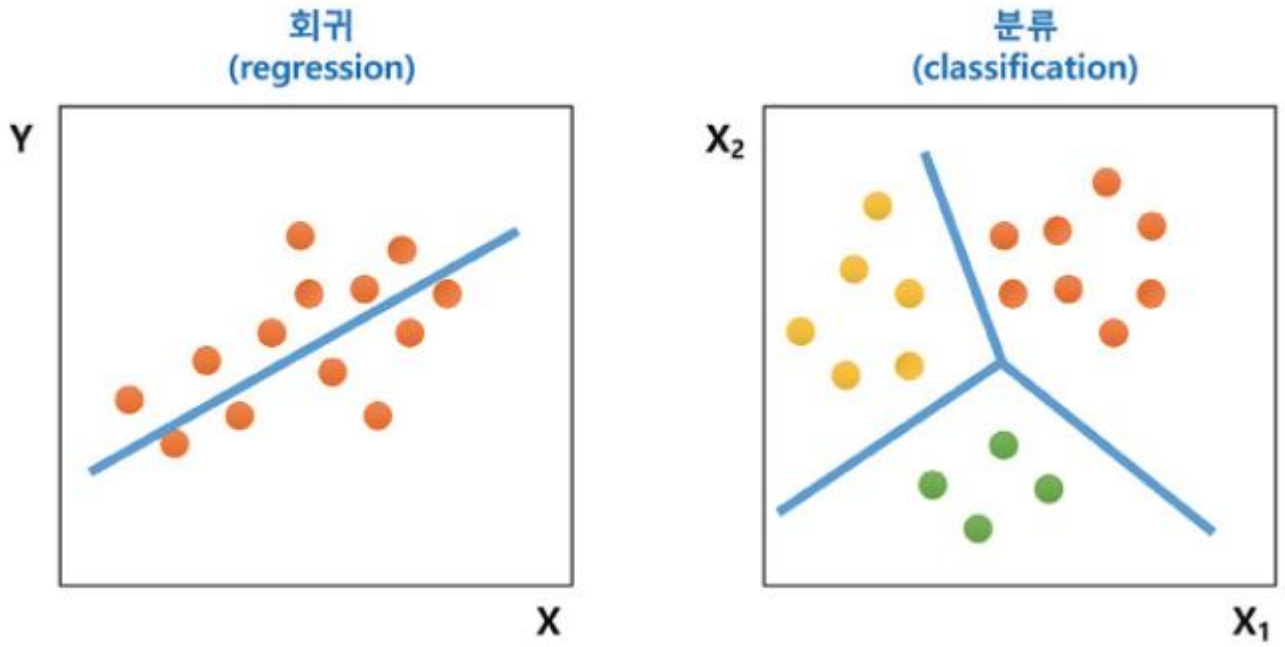
복습: 회귀와 분류 (2/3)

- **분류**
  - 설명 변수(X)와 목표 변수(Y) 사이의 관계를 찾지만
    - **예측하려는 목표 레이블이 연속적이지 않고 0, 1, 2와 같이 이산적인 값을 갖는 경우**
      - **클래스 0 또는 1 중에서 선택하는 이진 분류(binary classification) 문제**
      - **3개 이상의 클래스 중에서 하나를 선택하는 다중 분류(multi classification) 문제**
- **분류 예**
  - 개 또는 고양이를 찍은 사진을 읽어서 개(클래스 0)인지 고양이(클래스 1)인지 분류
  - 0~9 중에 하나의 숫자를 기록한 숫자 카드를 읽어서 어떤 숫자인지 판독하는 문제
    - **이 경우 분류의 목표가 되는 레이블이 10가지 종류인 다중 클래스(0, 1, 2, 3, 4, 5, 6, 7, 8, 9) 문제**



복습: 회귀와 분류 (3/3)

- **회귀**
  - 데이터의 분포를 가장 잘 설명할 수 있는 X, Y 사이의 함수식을 찾는 것
- **분류**
  - 섞여 있는 데이터들 중에서 목표 레이블을 가장 잘 구분할 수 있는 경계를 나타내는 함수식을 찾는 것



[그림] 회귀 vs. 분류