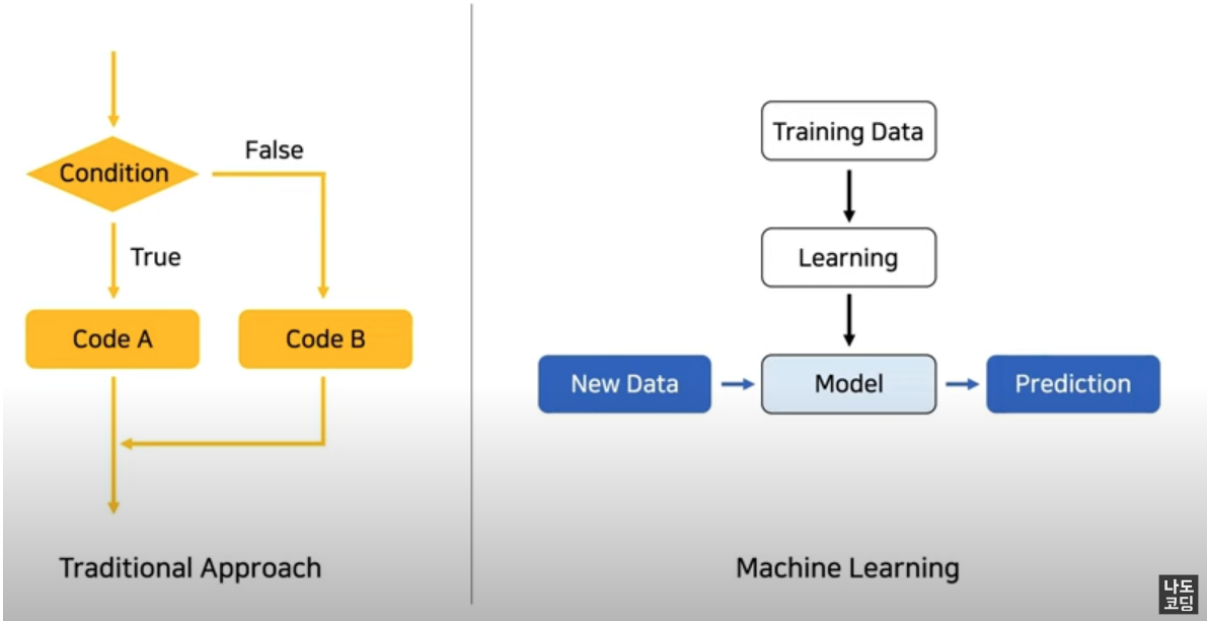
1주차

개요

인공지능의 한분야로 머신러닝,

머신러닝: 데이터를 통해서 기계가 스스로 학습하게끔

명시적으로 프로그램을 작성하지 않고, 컴퓨터가 스스로 규칙을 학습하는 연구 분야



- 머신러닝의 분류
  - Supervised Learning 지도학습
    - 정답 있음
    - 데이터 분류 / 올바른 결과 예측
  - Unsupervised Learning 비지도 학습
    - 정답 없음
    - 데이터의 유의미한 패턴, 구조 발견
  - Reinforcement Learning 강화학습
    - 행동에 대한 보상
    - 누적 보상을 최대화하는 의사결정

지도 학습

- **분류** - 범주형 변수

  주어진 데이터를 정해진 범주에 따라 분류

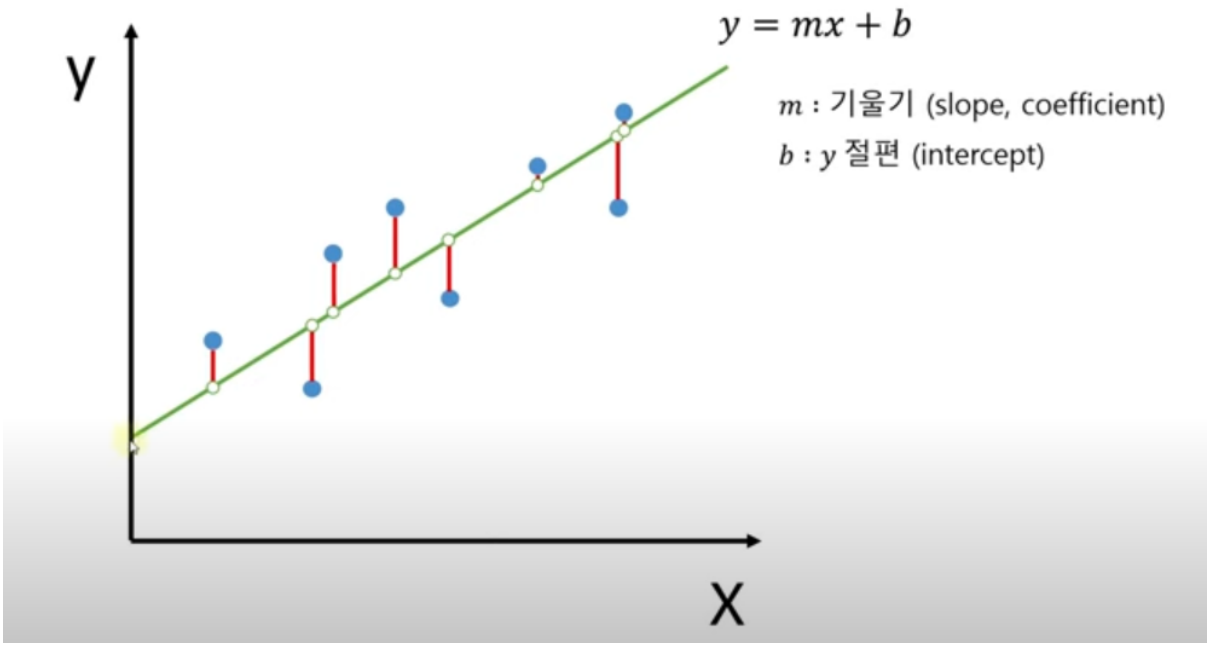
  예측 결과가 숫자가 아닐 때

- **회귀** - 연속형 변수

선형회귀

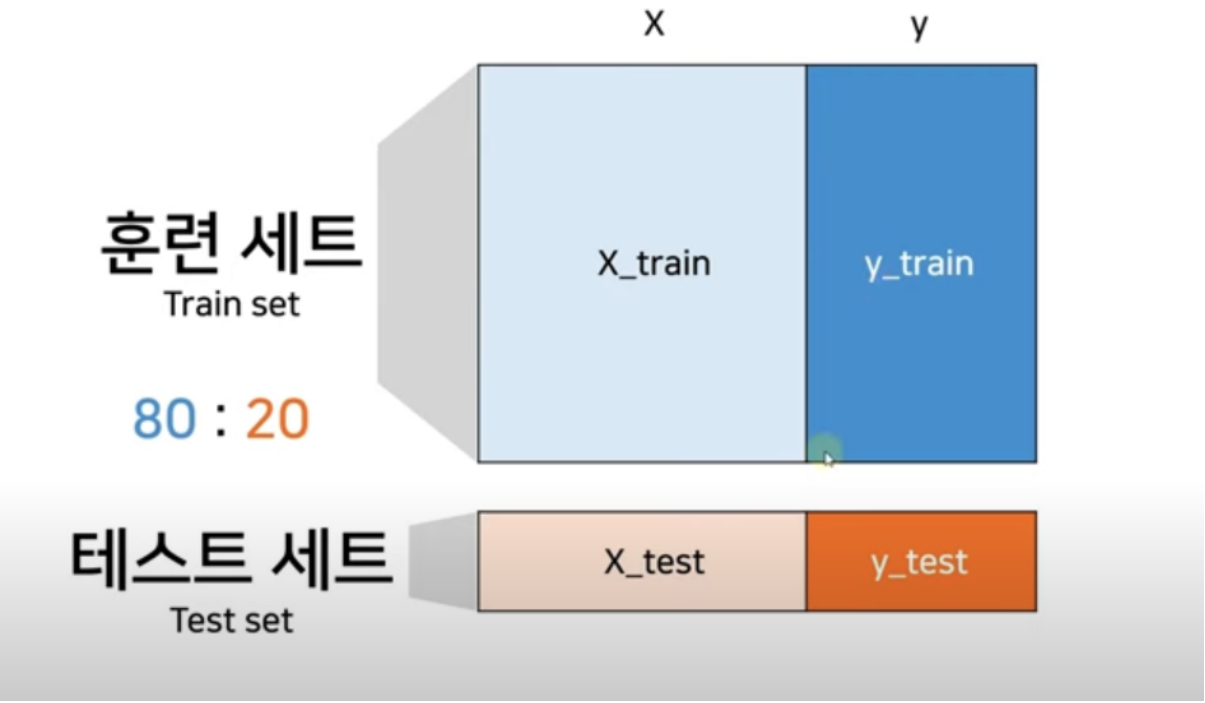
X	y
독립변수	종속변수
원인	결과
입력변수	출력변수
feature	target, label

실제 값과 예측 값 차이(잔차)의 제곱의 합을 최소화

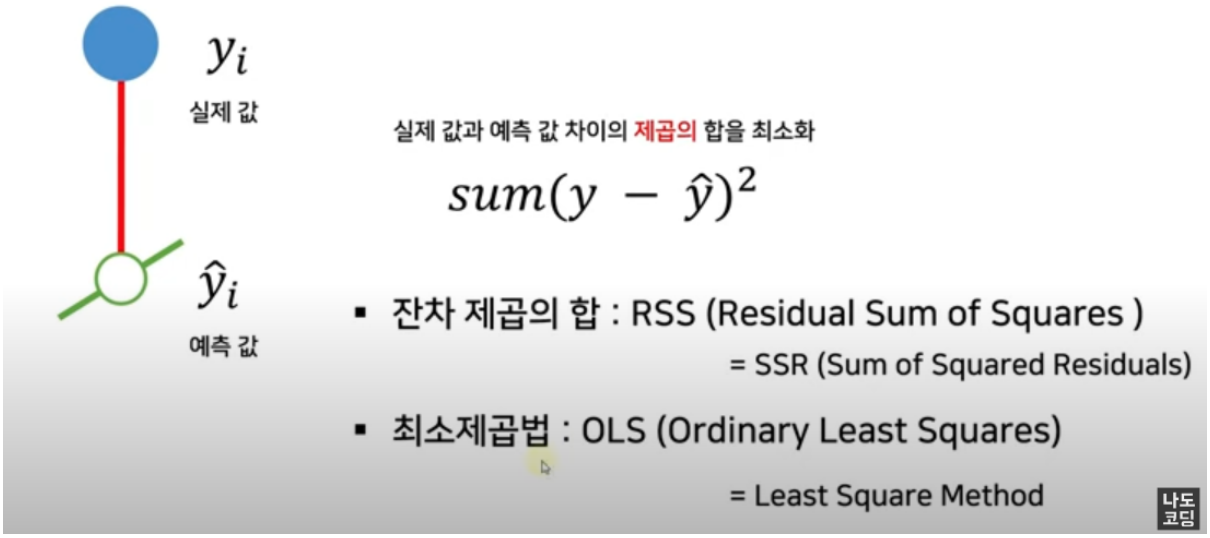


데이터 분리

- 훈련 세트 80
- 테스트 세트 20



경사하강법



최소 제곱법 → 잔차제곱의 합을 최소로 하는 회귀식을 찾는 방법

최소 제곱법의 단점) 노이즈에 취약함, 이상치에 취약함, 독립변수가 많아지면 컴퓨터 자원이 많이 필요로 함

경사하강법 Gradient Descent

목표) 잔차 제곱의 합이 가장 작아지도록 함

학습률(Learning rate)

에포트(Epoch) - 최적의 훈련값(파라미터)를 찾기 위해서 모든 훈련 세트를 다 사용함

확률적 경사 하강법 Stochastic Gradient Descent

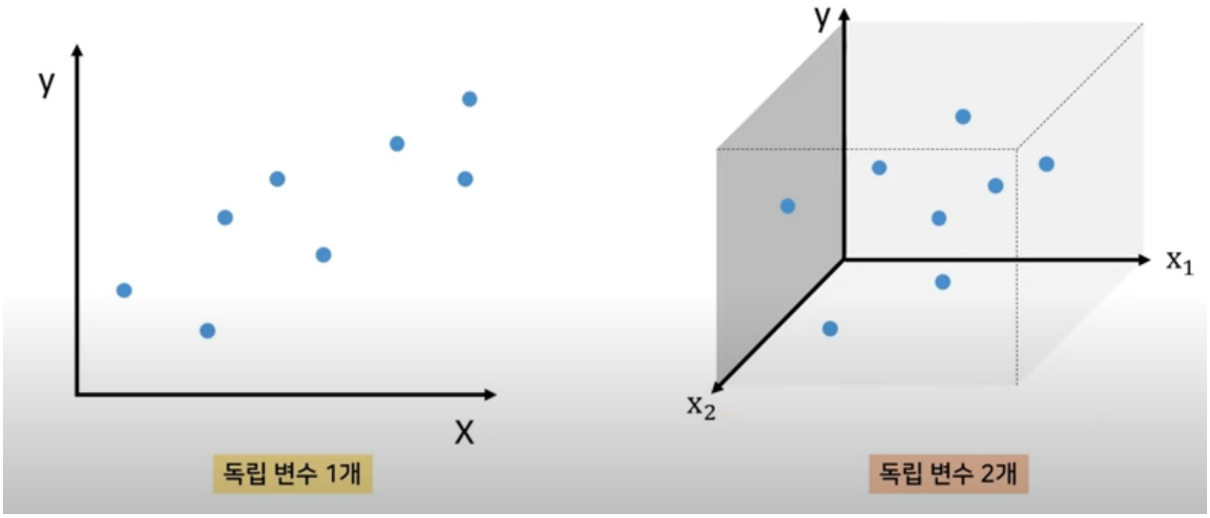
매 단계마다 무작위로 하나의 데이터를 선택하고 그 데이터에 대해서만 기울기를 계산함

다중 선형회귀

Multiple Linear Regression

y = b+m1x1 + m2x2 +…+mnxn

종속변수에 영향을 주는 독립변수가 많기 때문에..!



- 원-핫 인코딩 one-hot encoding

  표현하고 싶은 값만 1로, 나머지는 모두 0으로

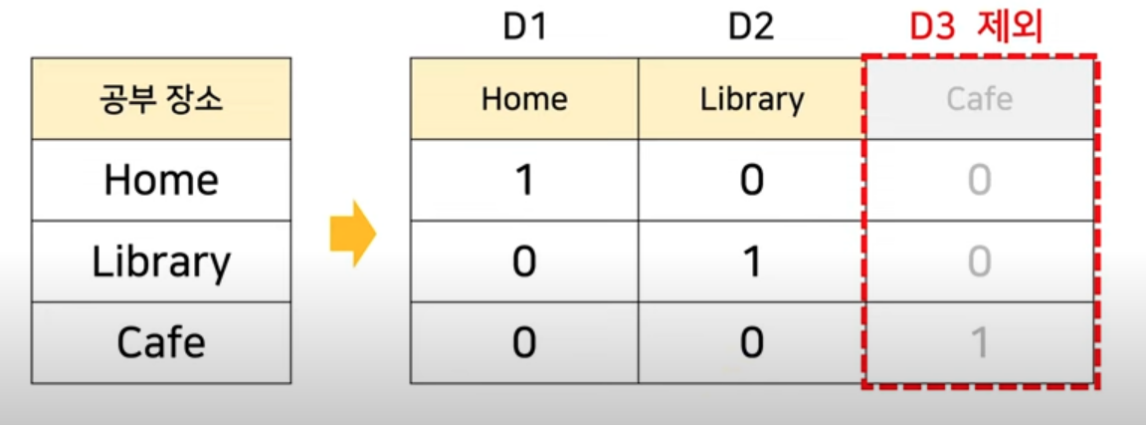
- 다중공선성 Multicollinearity

  독립 변수들 간에 서로 강한 상관관계를 가지면서 회귀계수 추정의 오류가 나타나는 문제

  하나의 피처가 다른 피처에 영향을 미침

  강한 상관관계는 문제가 될 수 있기 때문에 이를 해결해주기 위해서 한 칼럼을 아예 없앰

$$D1 + D2 + D3 = 1$$
$$D3 = 1 - (D1 + D2)$$



즉 dummy column 이 n개면 n-1개만 사용 → dummy variable trap