동국대학교

dongguk UNIVERSITY

2020년 - 데이터 청년 캠퍼스

데이터사이언스 기반 지능소프트웨어 과정

데이터 사이언스 개론

3) 지도 학습 (Supervised Learning) - 1

학습 내용

- **비지도 학습 (Unsupervised Learning) : 4가지**
  - K-평균 군집화 (k-means Clustering)
  - ...
- **지도 학습 (Supervised Learning) : 6가지 중 2가지**
  - 회귀 분석 (Regression Analysis)
    - 기울기 하강법 (Gradient Descent)
    - 상관 계수 (Correlation Coefficient)
  - k-최근접 이웃 (k-Nearest Neighbors)
- **강화 학습 (Reinforcement Learning) : 1가지**
  - 멀티-암드 밴딧 (Multi-Armed Bandits)

지도 학습 (Supervised Learning)

- 데이터 셋 (Dataset) 의 패턴 (Pattern) 을 기반으로 새로운 데이터로부터 예측

- 예측 결과와 실제 결과를 비교하여 해당 알고리즘의 성능을 평가할 수 있음

- 지도 : 이미 존재하는 패턴을 바탕으로 예측하기 때문
  - 기존 COVID-19 환자들의 특징을 분석하여 검사자가 확진자인지 아닌지를 예측
  - 식료품점에서 생선을 구매한 사람들의 정보를 토대로 과일도 구매할지 혹은 과일을 몇 개 구매할지 예측
  - 과거 날씨의 변화를 토대로 내일 비가 올 것인지 예측

회귀 분석 (Regression Analysis)

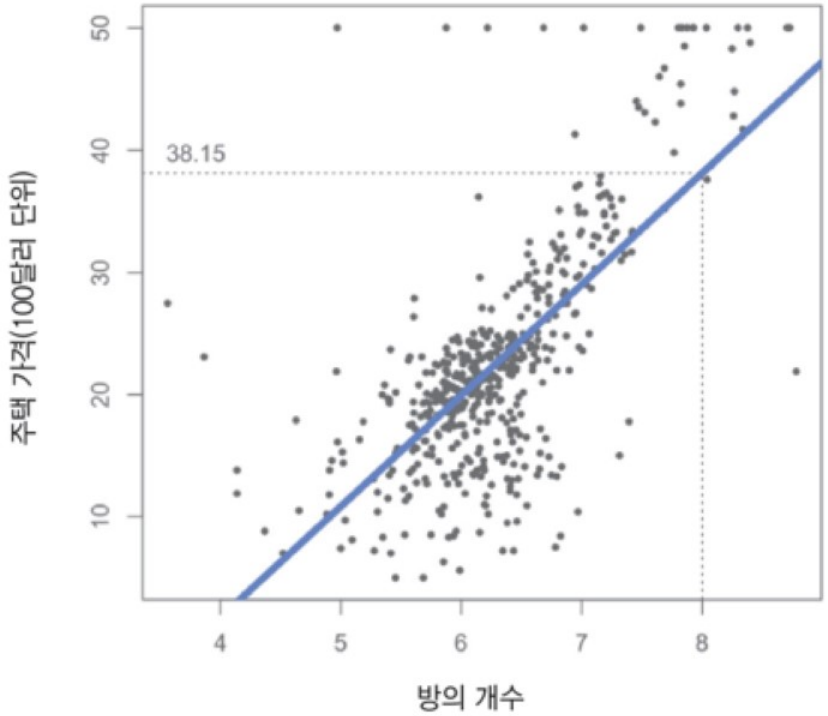
- 추세 / 추체선 (Trend / Trend Line) 은 생성하기 쉽고 이해하기 쉬워 예측에 널리 사용되는 도구

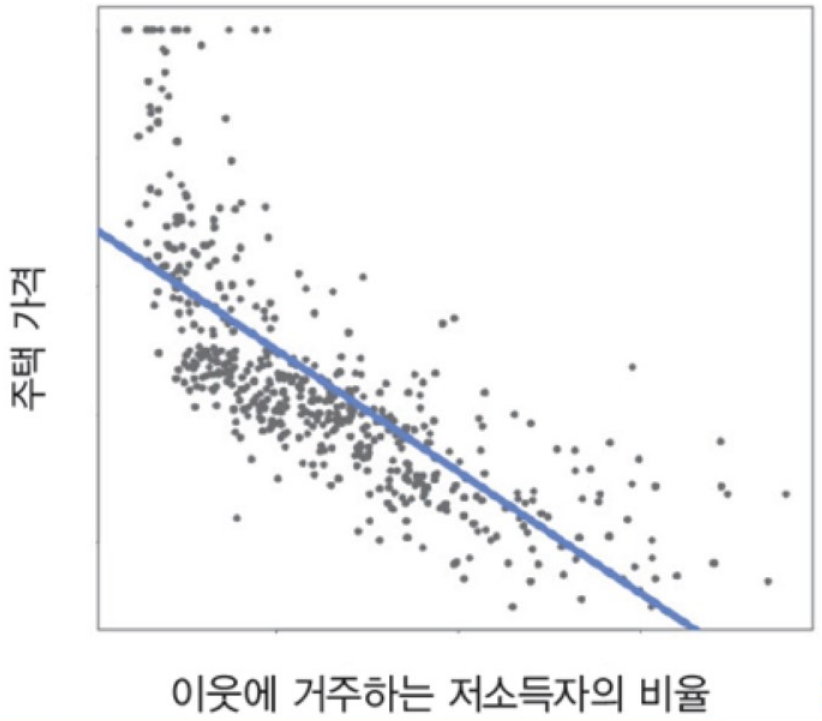
- 예측자 (Predictor) 를 기준으로 결과 (Outcome) 을 예측
  - 어떤 회사의 주식 (결과) 를 예측하기 위해 시간 (예측자) 를 사용
  - 정확도를 높이기 위해 영업이익 (새로운 예측자) 를 사용

- 예측자를 기준으로 데이터 셋을 이용하여 추체선을 생성하고, 이를 바탕으로 결과를 예측

회귀 분석 (Regression Analysis)

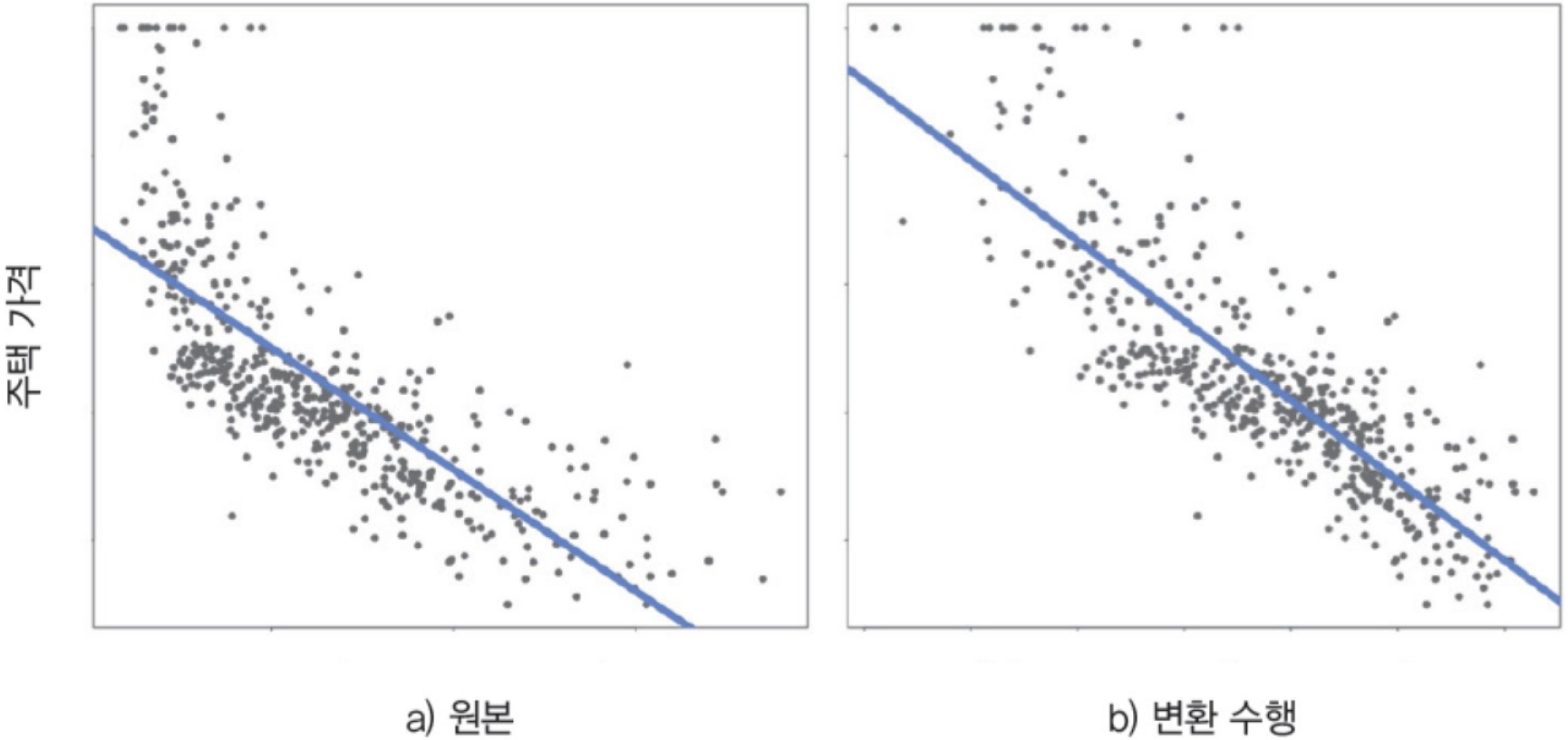
- **1970년대 보스턴 지역의 주택 가격 상승 데이터와 그에 관한 예측자**
  - 주택 가격과 가장 밀접한 예측자는
    - 방의 개수
    - 이웃에 거주하는 저소득자 비율





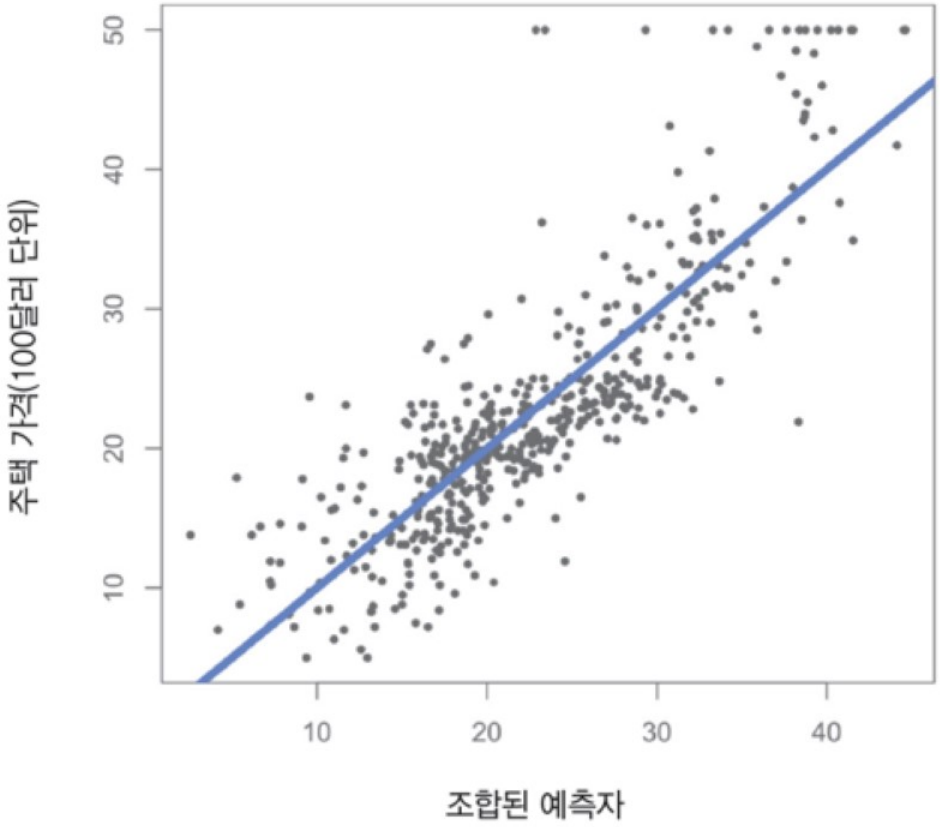
회귀 분석 (Regression Analysis)

■ 예측자를 변환하여 적절한 추체선으로 변경하기도 함



회귀 분석 (Regression Analysis)

- 예측자를 조합하여, 새로운 예측자로 만들기도 함
  - 영향력이 좀 더 큰 예측자에 가중치 (Weight) 를 주기도 함



- 그런데 최적의 가중치는 어떻게 찾는가? (=> 기울기 하강법)

회귀 분석 (Regression Analysis)

- **회귀 분석에서 가중치는 회귀 계수 (Regression Coefficients) 라고 함**
  - 특정 예측자에 대한 상대적 가치

- **키와 몸무게는 Cm 와 Kg 단위로 회귀 분석 가능**
  - 키 단위가 미터(m) 라면 : 키 * 100 혹은 몸무게 * 0.01
  - 몸무게 단위가 톤(t) 라면 : 키 * 0.001 혹은 몸무게 * 1000
  - 이를 표준화 (Standardization) 라고 함

- **표준화 후 계산된 회귀 계수는 베타 가중치 (Beta Weight) 라 함**

회귀 분석 (Regression Analysis)

■ 방의 개수와 이웃에 거주하는 저소득자 비율을 회귀 계수를 고려하여 가격을 계산하는 회귀 방정식

$$price = (2.7 \times \#\ of\ rooms) + (-6.3 \times rate\ of\ low\ income\ earner\ for\ neighborhood)$$
$$= (2.7 \times \#\ of\ rooms) - (6.3 \times rate\ of\ low\ income\ earner\ for\ neighborhood)$$

- *price* : 주택의 가격
- *# of rooms* : 방의 개수
- *rate of low income...* : 이웃에 거주하는 저소득자의 비율
- *2.7* : 양수, 방의 개수에 대한 회귀 계수
  - 양의 상관계수 (Positive Correlation)
- *-6.3* : 음수, 이웃에 거주하는 저소득자의 비율에 대한 회귀 계수
  - 음의 상관계수 (Negative Correlation)

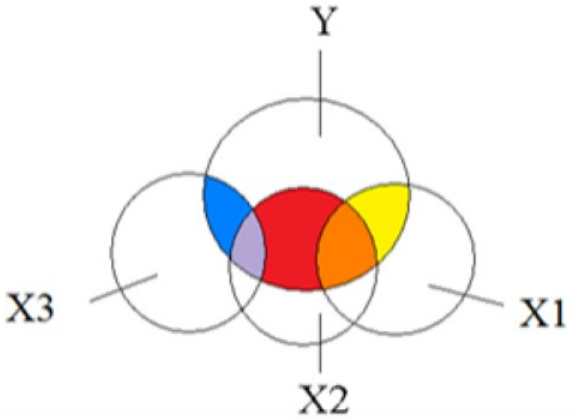
회귀 분석 (Regression Analysis)

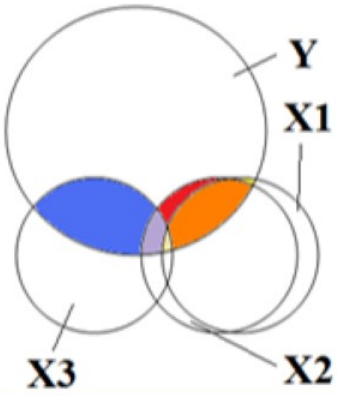
장점

- 사용하기 쉽고 회귀 분석 과정에 직관적임
- 결과값을 구하기 위한 계산이 빠름

단점

- 이상치 (Outlier) 에 대해서 민감
- 인과관계가 없음
- 이미 상관관계가 높은 예측자로 회귀 계수를 계산할 때, 왜곡된 결과가 나올 수 있음 (다중공선성 : Multicollinearity)

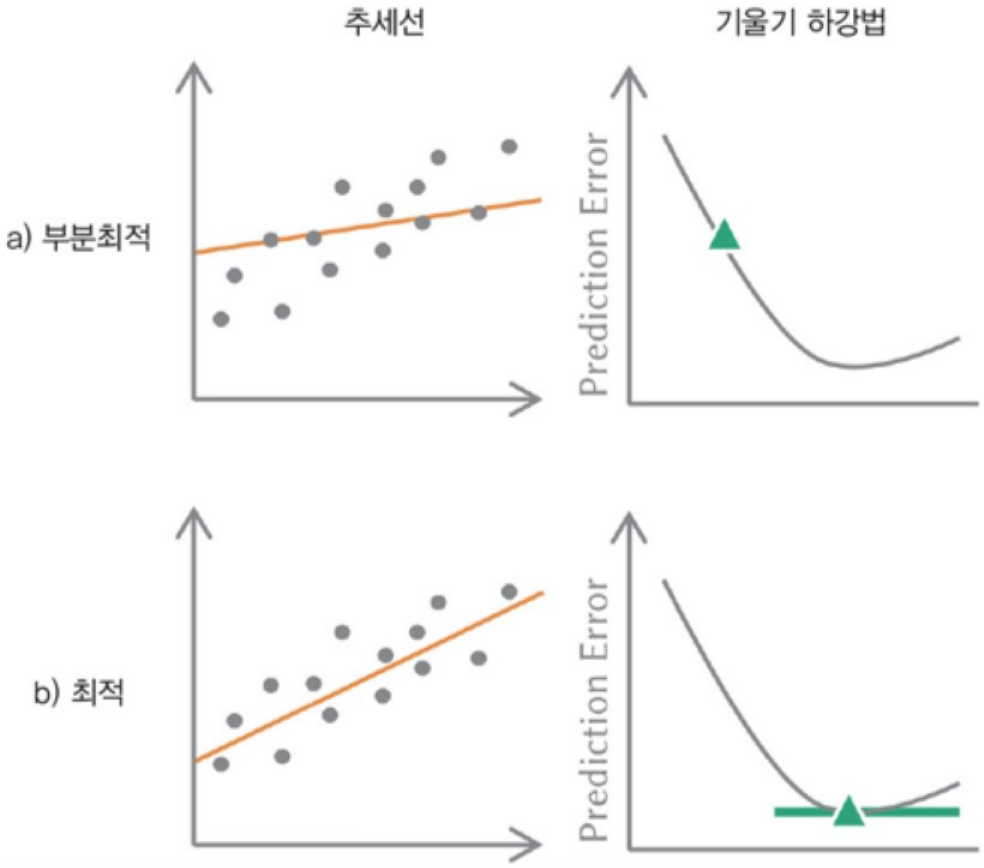




기울기 하강법 (Gradient Descent)

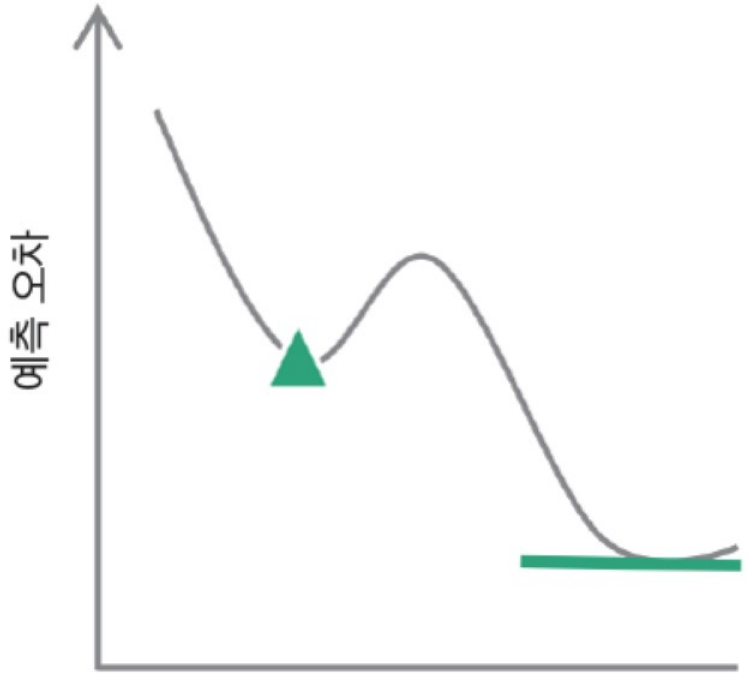
■ 최적의 가중치를 찾기 위해서는?

- 비례식이나 방정식 등을 이용하여 찾음
- 하지만 예측자가 너무 많을 때는?



기울기 하강법 (Gradient Descent)

- 기울기 하강법은 서포트 벡터 머신 혹은 신경망에서도 활용
  - 하지만 어떻게 시작해야 할까?
  - 어렵게 찾은 가중치 값이 정말 최적일까?



- ~~자알 찾는 수 밖에...~~

상관 계수 (Correlation Coefficient)

dongguk UNIVERSITY

서로 다른 예측자에 대한 관련도 (r)

- -1 에서 1 사이의 값으로 방향성을 가짐
- 방향 (Direction) 에 따라 양 / 음의 상관관계를 나타냄
  - 비례 방향 (좌하단 / 우상단) : 양의 상관관계
  - 반비례 방향 (좌상단 / 우하단) : 음의 상관관계
- 강도 (Strength) 에 따라 더 중요한 예측자를 결정할 수 있음
  - 0 에 가까울수록 : 덜 중요한 예측자 (관계가 없을 수 있음)
  - -1 혹은 1 에 가까울수록 : 더 중요한 예측자

단 하나의 예측자와 결과의 관련도

- = 베타 가중치

k-최근접 이웃 (k-Nearest Neighbors)

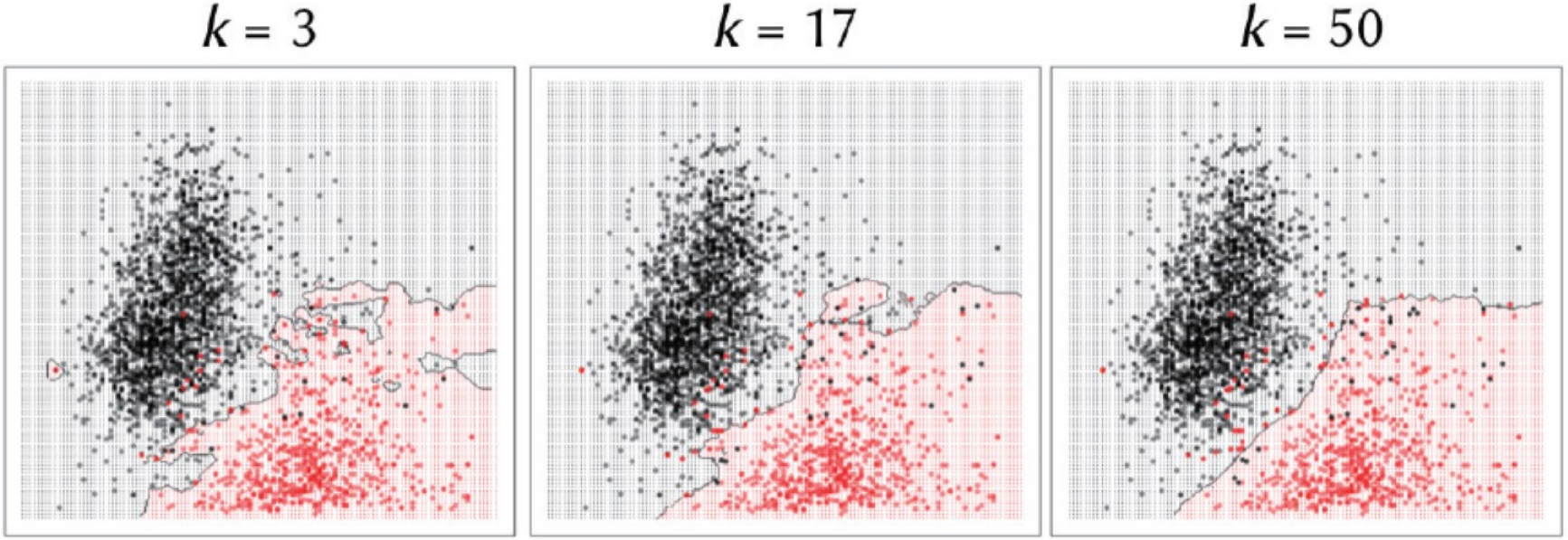
- **레드 와인과 화이트 와인의 차이점**
  - 레드 와인 : 숙성할 때, 포도 껍질을 넣음
  - 화이트 와인 : 숙성할 때, 포도 껍질을 뺌

- **만약 와인의 성분에서 포도 껍질로 인한 화학적 변화가 발견되었다면, 레드 와인이라고 예측 가능**

- **특정 지점 / 좌표의 데이터 포인트를 분석하여, 과반수에 따라 분류**
  - A 지점 10개의 데이터 포인트가 레드7 : 화이트3 → 레드 와인
  - B 좌표 5개의 데이터 포인트가 레드1 : 화이트4 → 화이트 와인

k-최근접 이웃 (k-Nearest Neighbors)

- 몇 개의 데이터 포인트 (k) 로 투표를 진행하는가?
  - 데이터 포인트의 영역은 보통 원형
  - 너무 많은 데이터 포인트 : 멀리 있는 데이터 포인트에 반응
  - 너무 적은 데이터 포인트 : 바로 옆 데이터 포인트에만 반응
  - 여러 번의 실험으로 적절한 k 를 결정 (보통 홀수)



a) 과적합　　b) 최적합　　c) 부적합

k-최근접 이웃 (k-Nearest Neighbors)

장점

- 데이터 포인트들의 시각화에 유리함

단점

- 분류 영역이 여러 개이고, 데이터 포인트 개수의 차가 심하다면, 정확한 분류가 어려움
- 예측자가 많다면, 차원 또한 증가하여 계산량이 증가함


Q & A