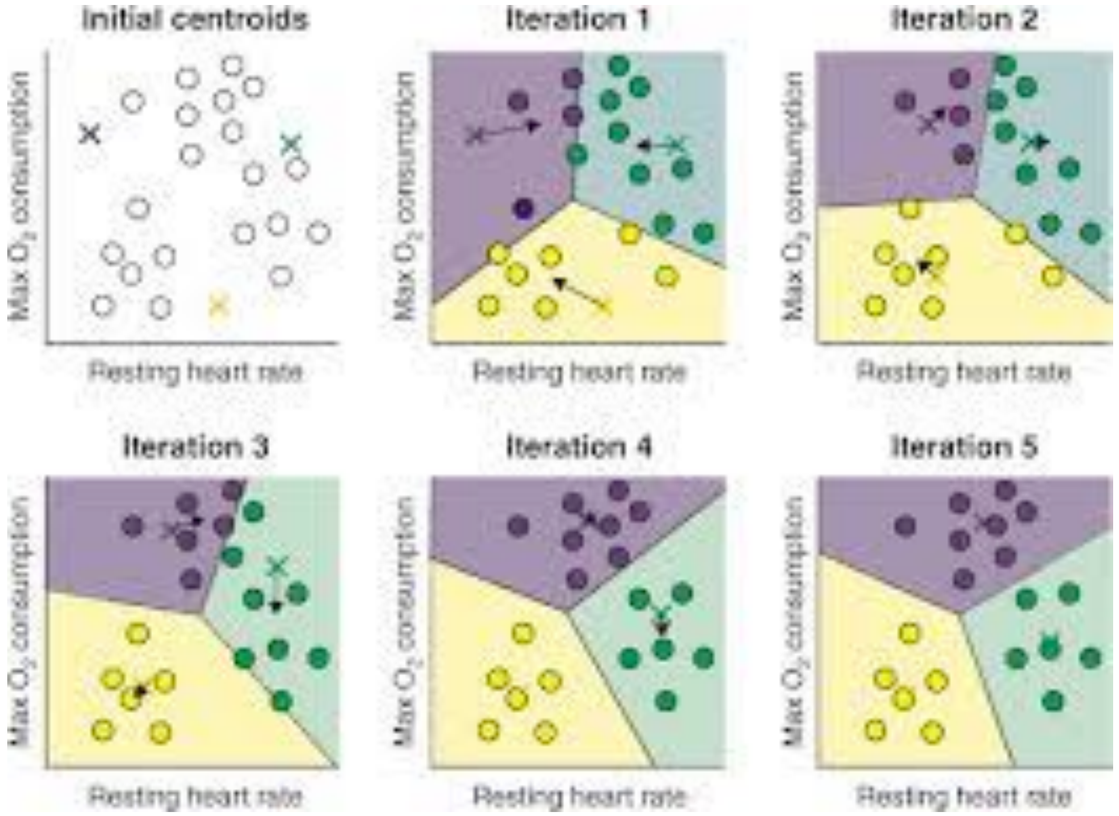
K-means Clustering

데이터 분석 모델링반 - ML1

K- means 클러스터링



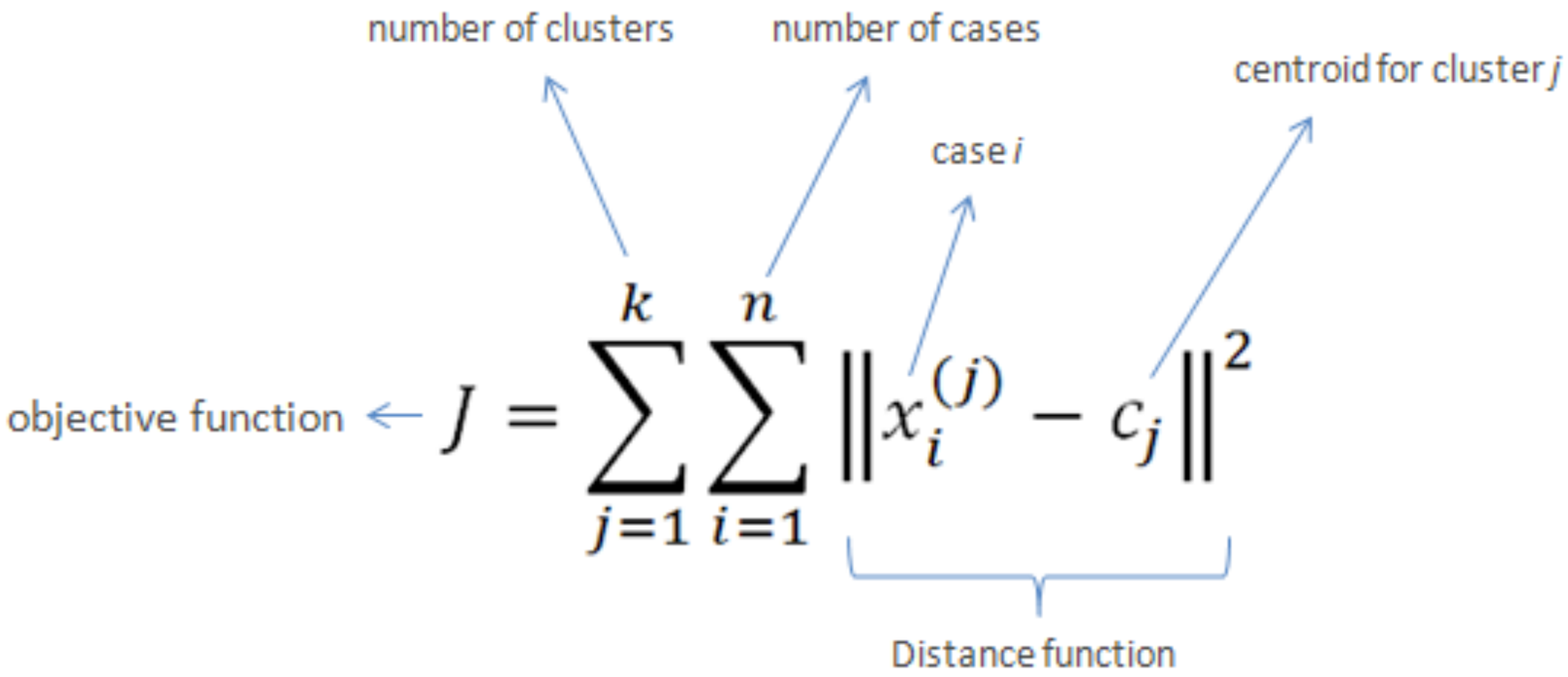
Algorithm 1: K-means clustering

Input : a given data $X = \{x_1, x_2, ..., x_n\}$,
the number of clusters k,
maximum number of iteration I

Output : clustering results r_{nk} for all n and k,
centroid of clusters C

Randomly initialize $C = \{c_1, c_2, ..., c_k\}$
for $t = 1 : I$ **do**
  // Assignment step
  for $n = 1 : N$ **do**
    $$r_{nk} = \begin{cases} 1, & \text{if } k = \underset{i}{\operatorname{argmin}} ||x_n - c_i||^2 \\ 0, & \text{otherwise} \end{cases}$$
  end
  // Update step
  for $k = 1 : K$ **do**
    $$c_k = \frac{1}{\sum_{n=1}^{N} r_{nk}} \sum_{n=1}^{N} r_{nk} x_n$$
  end
end

K-means 한계?


number of clusters
number of cases
centroid for cluster j
case i
objective function
$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$
Distance function

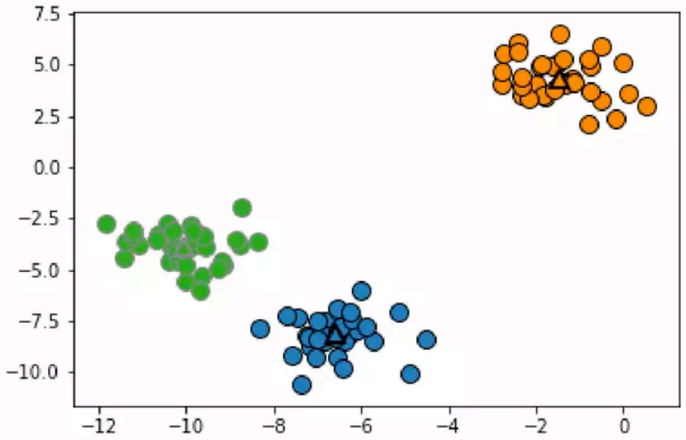
3 K-means clustering: 장단점

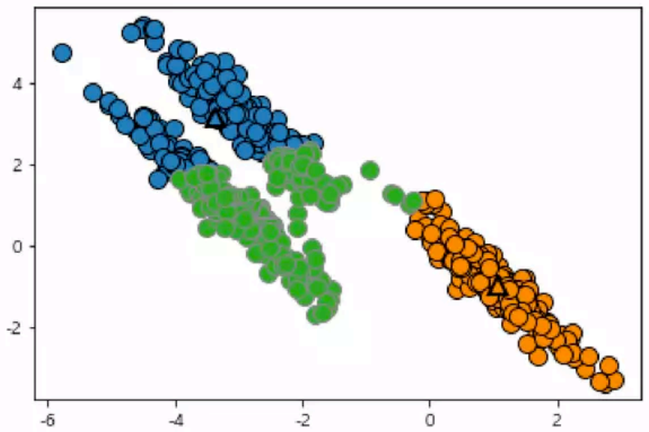
장점

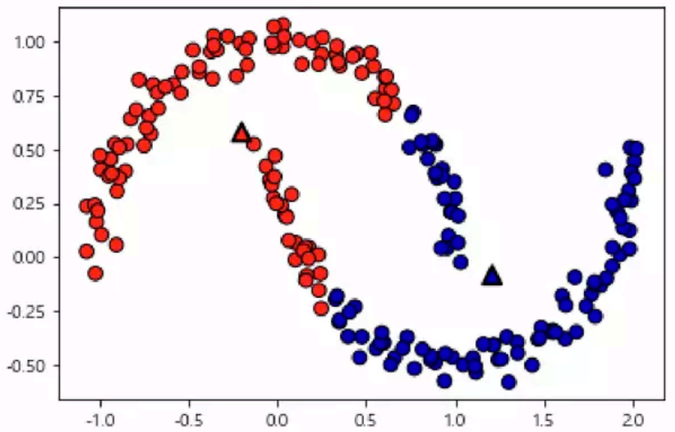
- 알고리즘이 간단하고 큰 데이터에도 손쉽게 사용 가능

단점

- 연속형 변수에 가장 최적
- 결과가 초기에 지정한 클러스터 중심의 위치에 따라 달라질 수 있어 반복 필요
- 클러스터의 개수를 지정해야 함
- 클러스터의 모양을 가정하기 때문에(원형) 다양한 분포를 가지는 데이터에 적용 한계



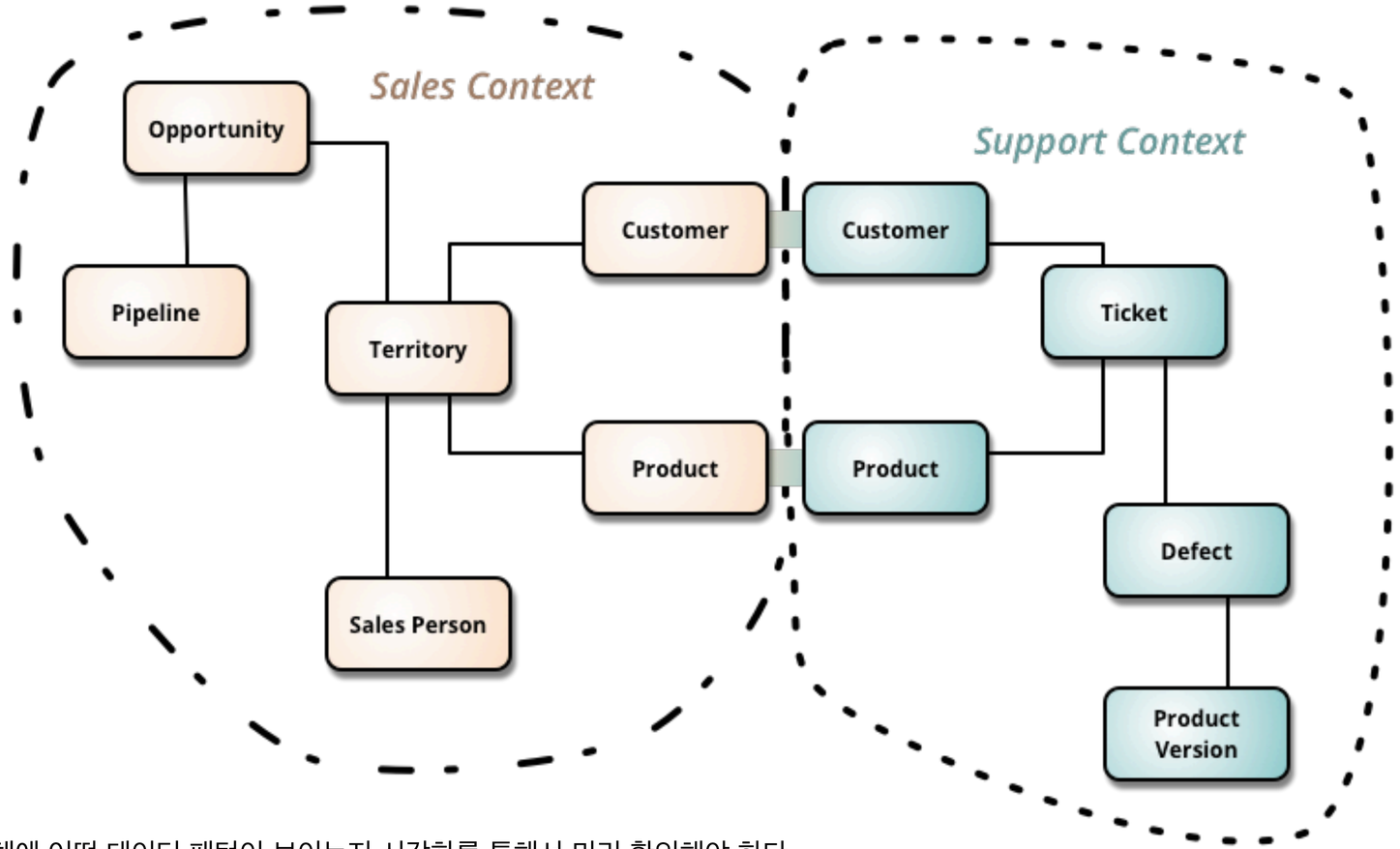




안드레아스 뮐러, 세라 가이도 (2019) 파이썬 라이브러리를 활용한 머신러닝, 박해선 옮김, 한빛미디어

K-mean 군집 평가

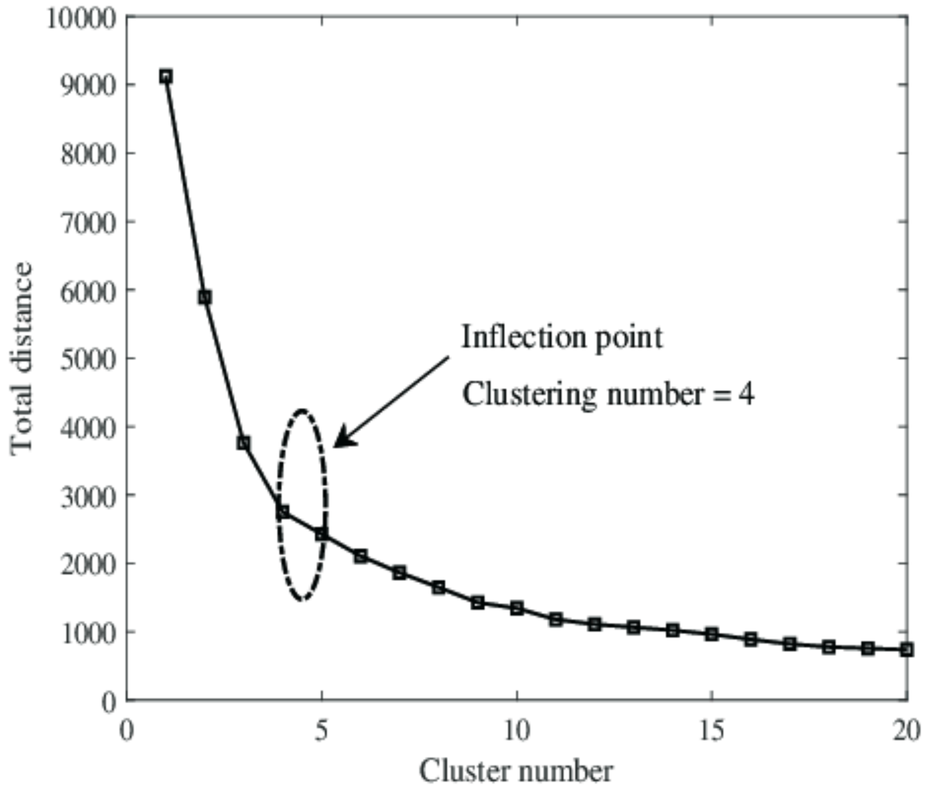
비즈니스 도메인 지식



1. Kmeans에 들어갈 피처 자체에 어떤 데이터 패턴이 보이는지 시각화를 통해서 미리 확인해야 한다.

흔히 말하는 차원축소나, 다른 군집화를 사용하는 것에 대한 기준점으로 현재 피처들이 어떤 데이터 패턴을 보이는지가 정말 중요합니다.

K-mean 군집 평가



Elbow method

K-means 군집 평가

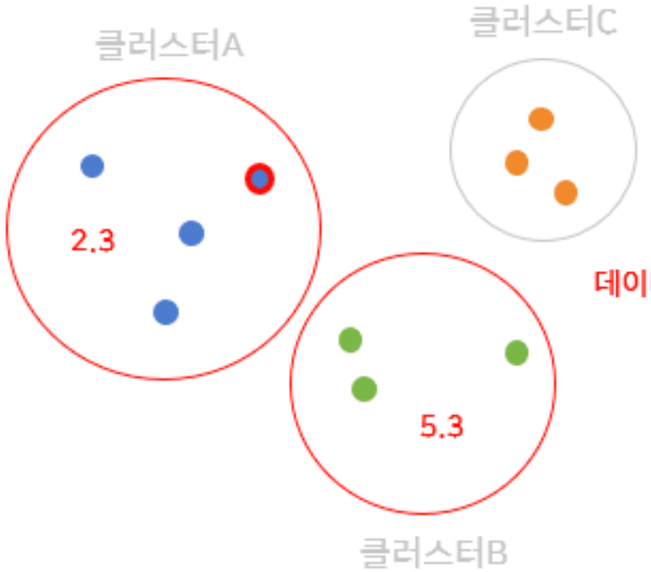
실루엣 계수(Silhouette coefficient)

i 번째 데이터 포인트와 동일한 클러스터에 속한 데이터 포인트들 간 거리들의 평균

i 번째 데이터 포인트와 다른 클러스터에 속한 데이터 포인트들 간 거리들의 평균을 클러스터 별로 구하는데 이들 중 가장 작은 값

$$S_i = \frac{b_i - a_i}{\max\{a_i,\ b_i\}}$$

i 번째 데이터 포인트의 실루엣 스코어



데이터 포인트 i의 실루엣 계수

$$\frac{5.3 - 2.3}{\max(5.3,\ 2.3)} = \frac{3}{5.3} = 0.57$$