K-means 클러스터링

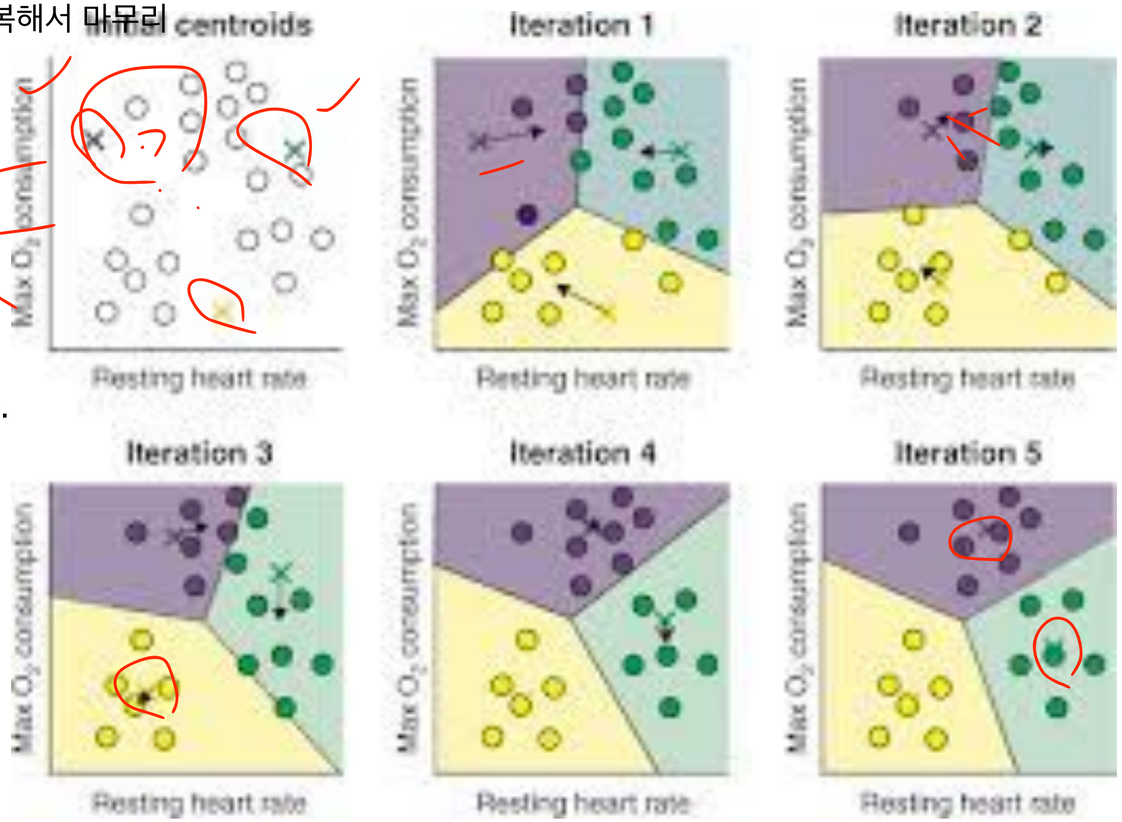
BDA_7기_파이썬문법응용반_수업자료

KNN
Kmeans
둘의 차이는 무엇인가?
지도학습과 비지도학습 차이

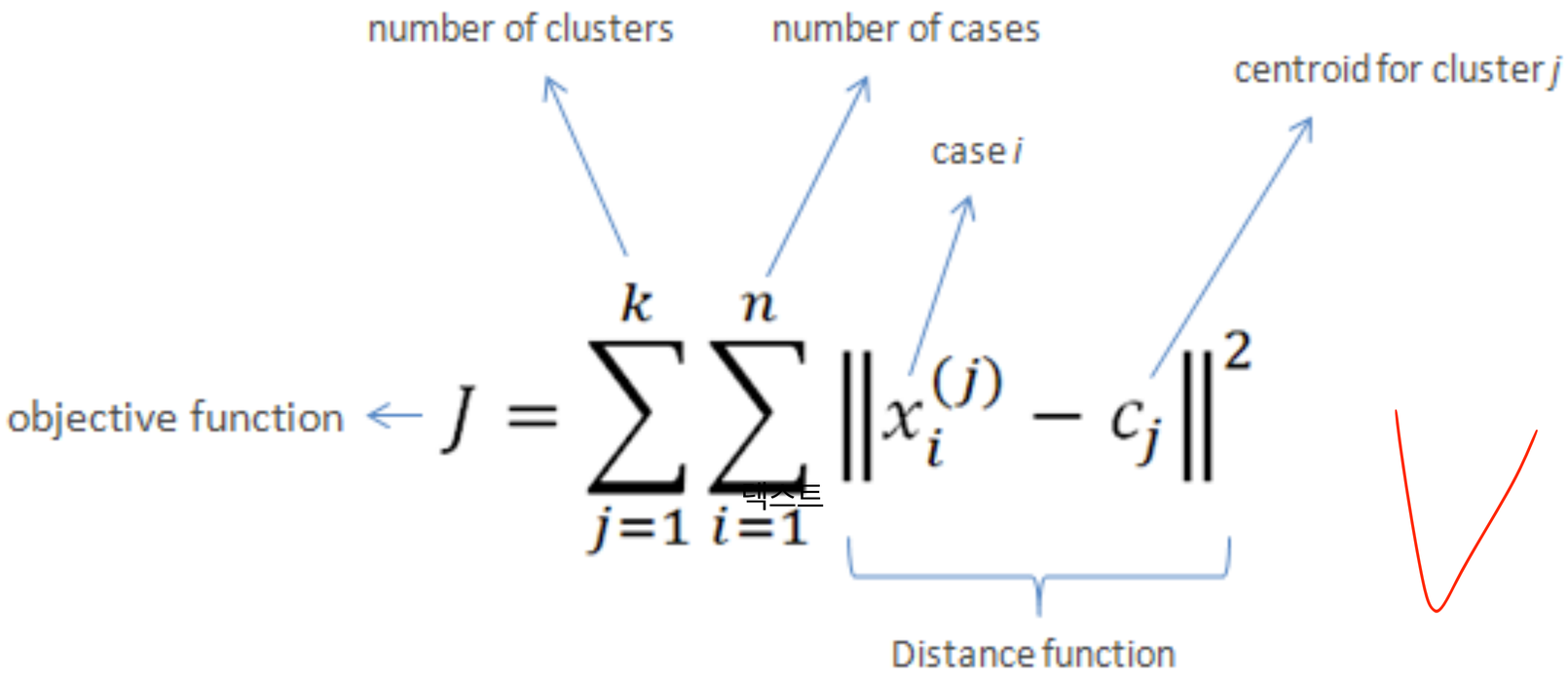
K- means 클러스터링

K개의 중심점을 임의의 데이터 공간에 선정
각 중심점 관측치들 거리 (유클리드) 계산
각 중심점과 거리가 가까운 관측치들을 해당 군집으로 할당
할당된 군집의 관측치들과 해당 중심점과의 거리를 계산
중심점을 군집의 중앙으로 이동 (군집의 관측치들간의 거리가 최소 지점)
중심점이 더 이상 움직이지 않을 때까지 반복해서 마무리

k-means-++ 방법을 이용해서 초기 시작점을 잡는다.



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

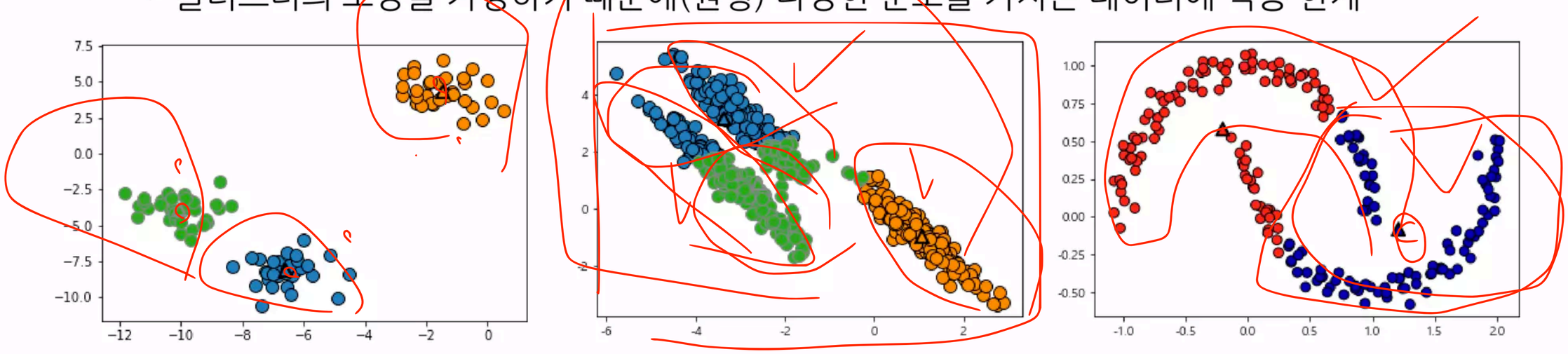
3 K-means clustering: 장단점

장점

- 알고리즘이 간단하고 큰 데이터에도 손쉽게 사용 가능

단점

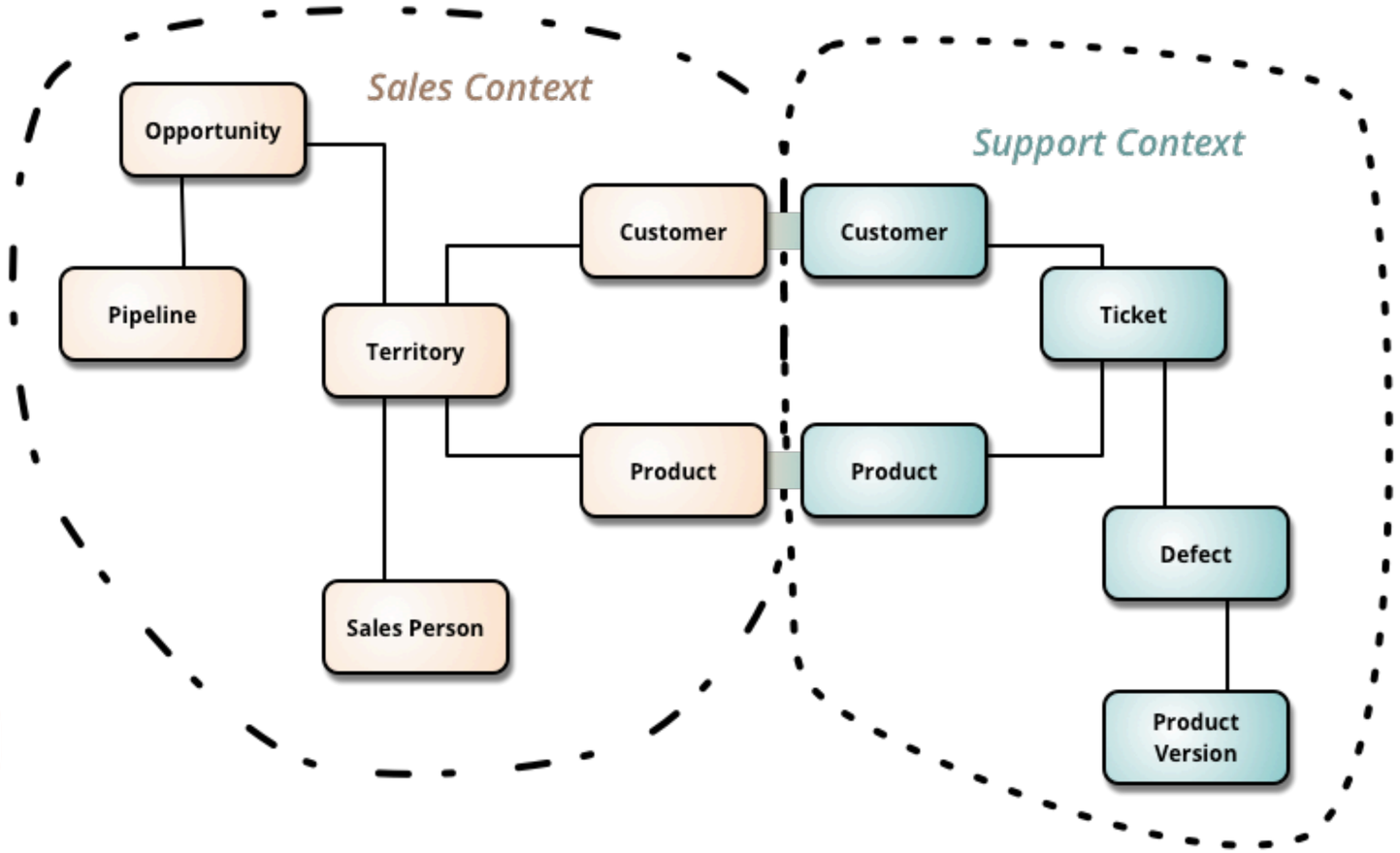
- 연속형 변수에 가장 최적
- 결과가 초기에 지정한 클러스터 중심의 위치에 따라 달라질 수 있어 반복 필요
- 클러스터의 개수를 지정해야 함
- 클러스터의 모양을 가정하기 때문에(원형) 다양한 분포를 가지는 데이터에 적용 한계



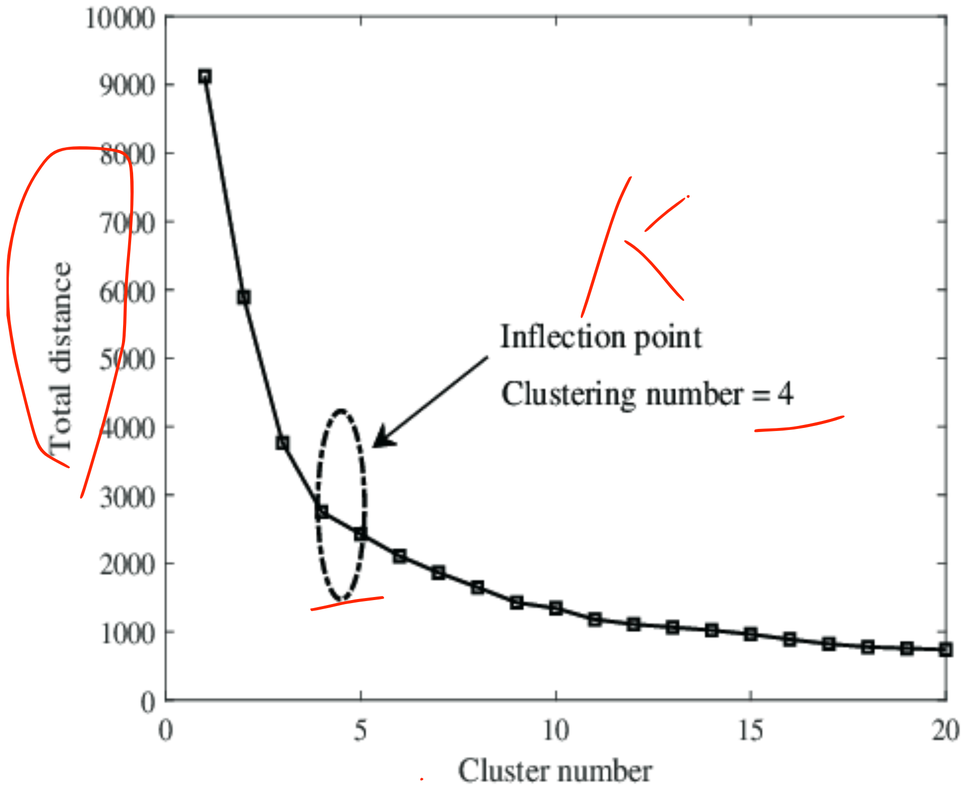
안드레아스 뮐러, 세라 가이도 (2019) 파이썬 라이브러리를 활용한 머신러닝, 박해선 옮김, 한빛미디어

K-mean 군집 평가

비즈니스 도메인 지식



K-mean 군집 평가



군집 내 중심점과 관측치 간의 거리 합으로 계산하여 급감하는 k 개수를 선정하는 방법

Elbow method

K-means 군집 평가

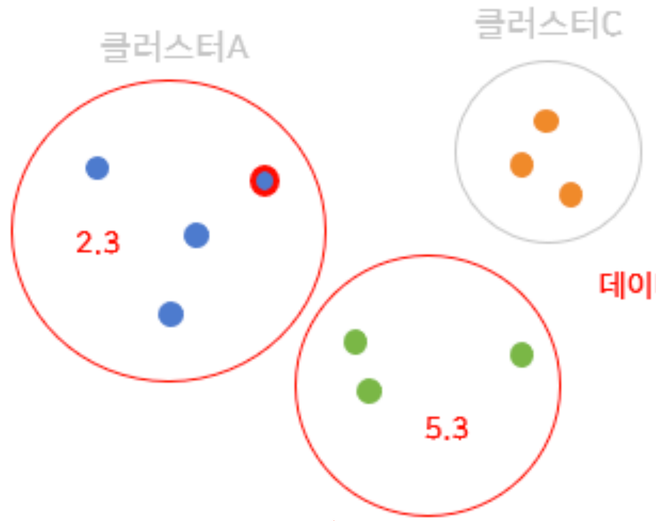
실루엣 계수(Silhouette coefficient)

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

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