

Part.05

Clustering

# | K-medoids clustering

FASTCAMPUS  
ONLINE

머신러닝과 데이터분석 A-Z

강사. 이경택

# I K-medoids clustering

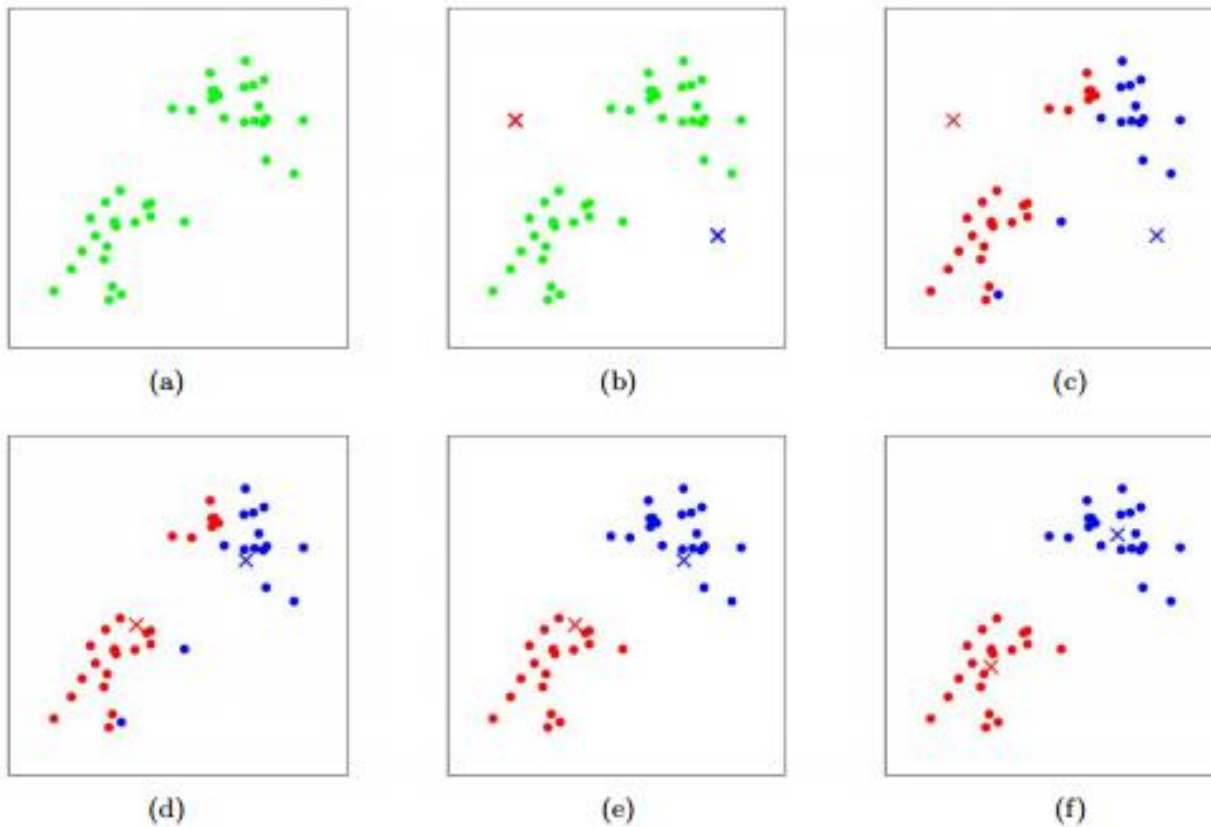
- K-means clustering

- 각 군집에 할당된 포인트들의 평균 좌표를 이용해 중심점을 반복적으로 업데이트
- Step1 – 각 데이터 포인트  $i$ 에 대해 가장 가까운 중심점을 찾고, 그 중심점에 해당하는 군집 할당
- Step2 – 할당된 군집을 기반으로 새로운 중심 계산, 중심점은 군집 내부 점들 좌표의 평균(mean) 으로 함
- Step3 – 각 클러스터의 할당이 바뀌지 않을 때까지 반복

# I K-medoids clustering

## ■ K-means clustering

- Step1 – 각 데이터 포인트 i에 대해 가장 가까운 중심점을 찾고, 그 중심점에 해당하는 군집 할당
- Step2 – 할당된 군집을 기반으로 새로운 중심 계산, 중심점은 군집 내부 점들 좌표의 평균(mean) 으로 함
- Step3 – 각 클러스터의 할당이 바뀌지 않을 때 까지 반복



# I K-medoids clustering

## ■ K 값을 설정하는 방법

- 군집의 개수 K는 사용자가 임의로 정하는 것이기 때문에 데이터에 최적화된 k를 찾기 어려움
- K를 설정하는 대표적인 방법은 Elbow method, Silhouette method 등이 있음
- Elbow method

➤ 군집 간 분산(BSS; Between cluster Sum of Squares)과 전체 분산( $TSS = BSS + WSS$ )의 비율

**WSS** (Within cluster Sum of Squares)

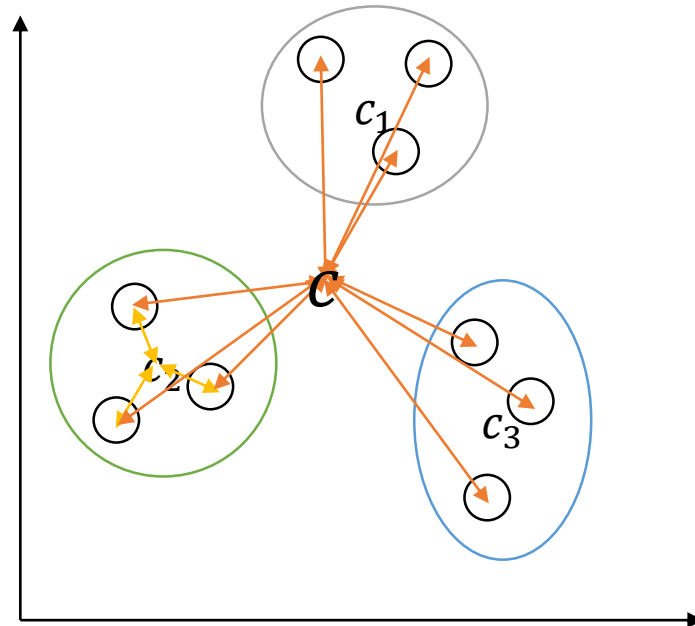
$$= \sum_{j=1}^K \sum_{i \in c_j} d(x_i, c_j)^2$$

객체  $x_i$ 와 군집  $j$ 의 중심  $c_j$ 와의 거리 제곱합

**TSS** (Total Sum of Squares)

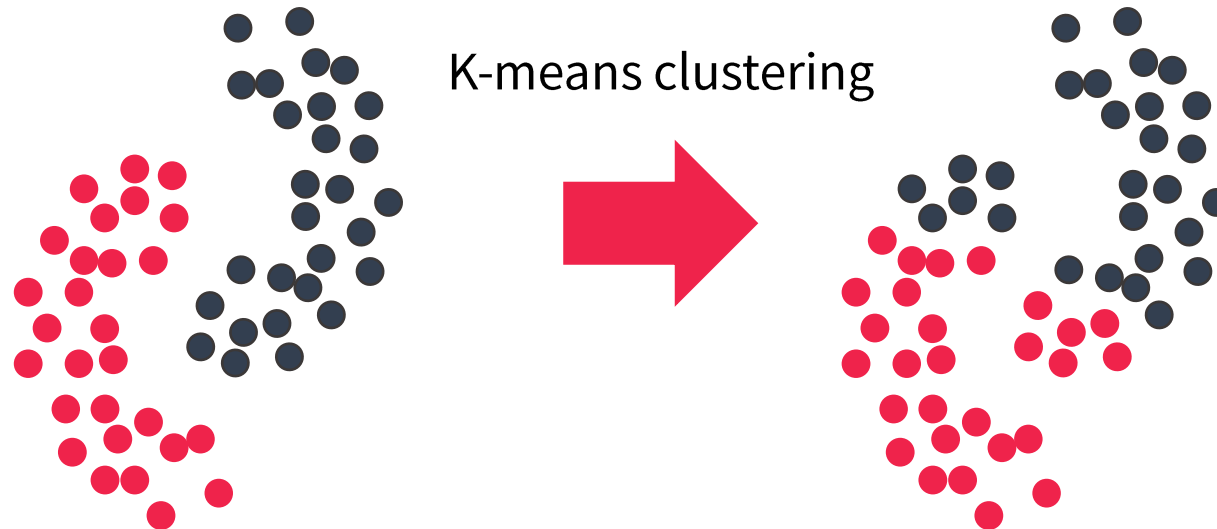
$$= \sum_{i=1}^N d(x_i, c)^2$$

객체  $x_i$ 와 전체 데이터의 중심  $c$ 와의 거리 제곱합



# I K-medoids clustering

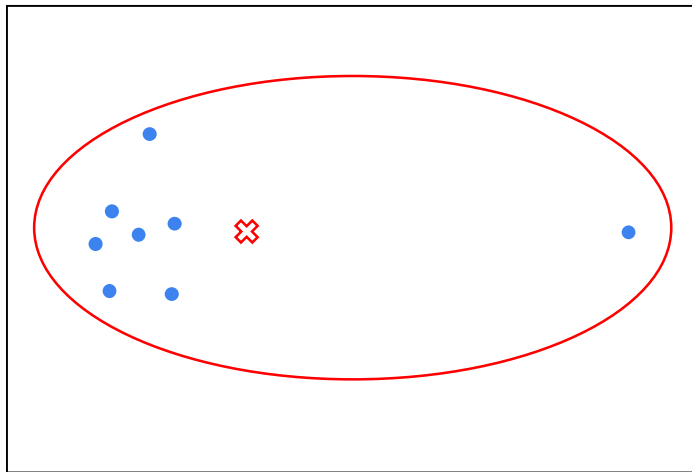
- K-means clustering의 단점
  - 초기 중심 값에 민감한 반응을 보임
  - 노이즈와 아웃라이어에 민감함
  - 군집의 개수 K를 설정하는 것에 어려움



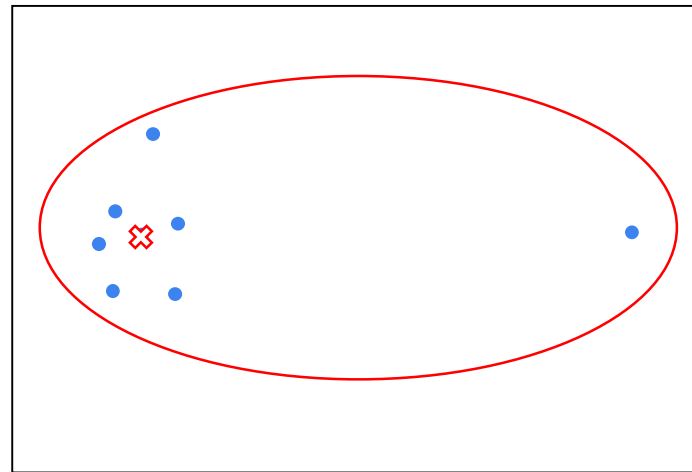
# I K-medoids clustering

## ■ K-medoids clustering

- K-means clustering의 변형으로, 군집의 무게 중심을 구하기 위해 데이터의 평균 대신 중간점(medoids)을 사용 (K-means보다 이상치에 강건한 성능을 보임)
- 아래 그림의 결과를 보면 K-medoids의 중앙점이 더 명확함 (이는 더 좋은 군집을 형성하게 될 가능성을 높임)



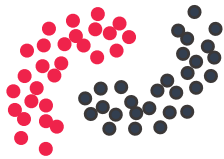
(a) Mean



(b) Medoid

# I K-medoids clustering

- K-means vs K-medoids

	K-means	K-medoids
중심	군집의 평균 값	군집 내 중앙 데이터
이상치	이상치가 전체 거리 평균 값에 영향을 주어 이상치에 민감함	K-means보단 덜 민감함
계산 시간	상대적으로 적은 시간이 소요	데이터 간 모든 거리 비용을 반복하여 계산해야 하므로 상대적으로 많은 시간이 소요
파라미터	군집의 개수 k, 초기 중심점	
군집 모양	원형의 군집이 아닌 경우 군집화를 이루기 어려움(아래 그림 참조) 	



Part.05

Clustering

# | Hierarchical clustering

FASTCAMPUS  
ONLINE

머신러닝과 데이터분석 A-Z

강사. 이경택