Part. 05

Clustering

| 최적의 k를 찾는 방법

FASTCAMPUS
ONLINE
머신러닝과 데이터분석 A-Z|

강사. 이경택

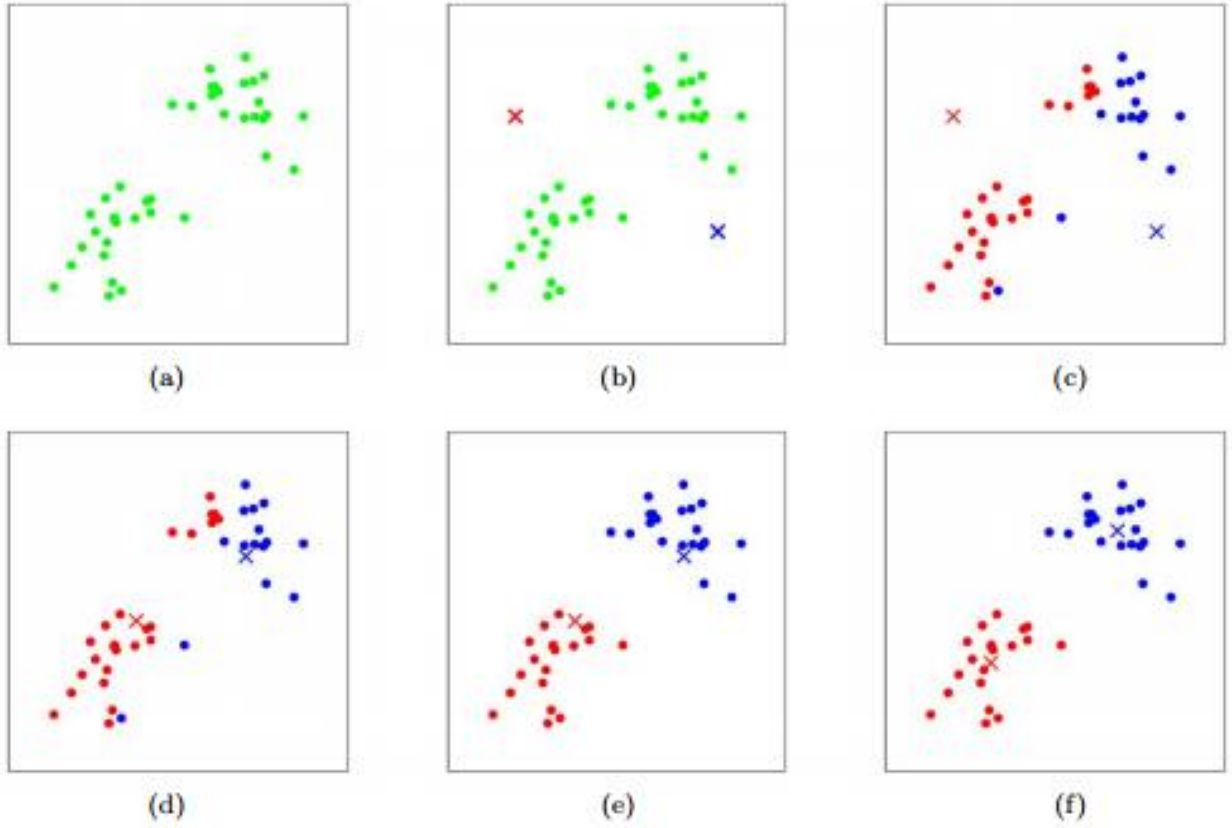
I 최적의 K를 찾는 방법

- K-means clustering
  - 각 군집에 할당된 포인트들의 평균 좌표를 이용해 중심점을 반복적으로 업데이트

  - 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 때까지 반복

I 최적의 K를 찾는 방법

- 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 최적의 K를 찾는 방법

- 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)

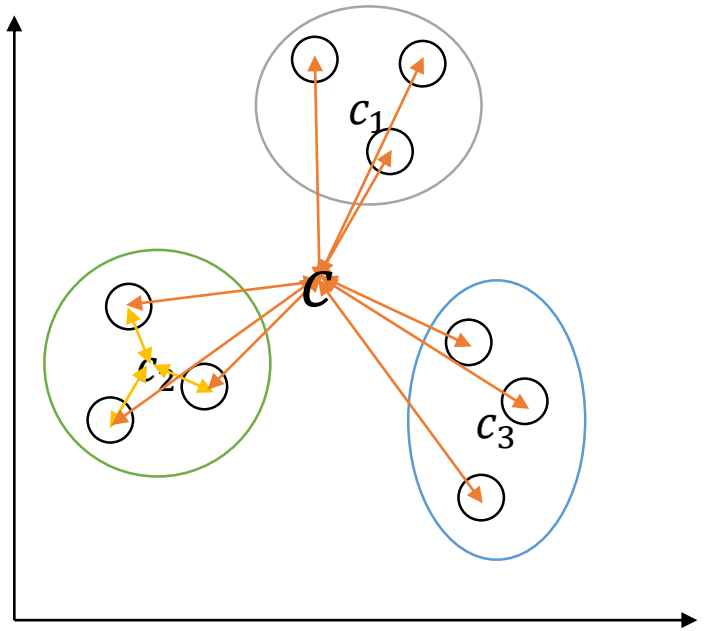
$$= \sum_{j=1}^{K} \sum_{i \in c_j} d(x_i, c_j)^2$$

객체 x_i와 군집 j의 중심 c_j와의 거리 제곱합

TSS (Total Sum of Squares)

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

객체 x_i와 전체 데이터의 중심c와의 거리 제곱합

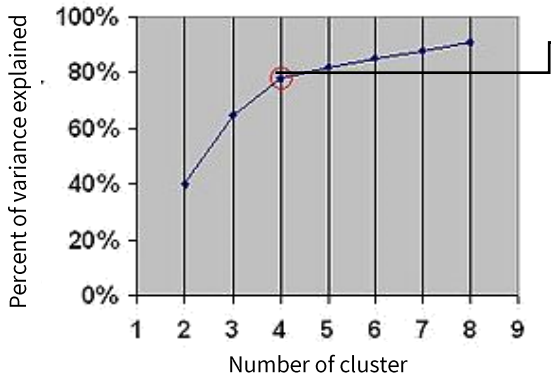


I 최적의 K를 찾는 방법

- K 값을 설정하는 방법
  - Elbow method
    - 군집 간 분산과 전체 분산의 비율

$$ratio = \frac{BSS}{TSS} = \frac{TSS - WSS}{TSS}$$

    - 비율의 한계 비용(marginal cost)이 줄어드는 지점이 최적의 클러스터 개수



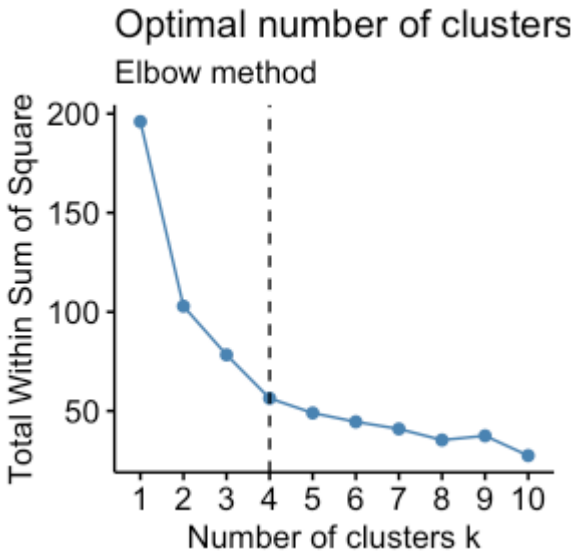
분산 비율의 증가분이 줄어드는 지점인 k=4가 최종 클러스터 개수

I 최적의 K를 찾는 방법

- K 값을 설정하는 방법
  - Elbow method
    - ➢ WWS를 통해 비율의 한계 비용(marginal cost)이 줄어드는 지점이 최적의 클러스터 개수

WSS (Within cluster Sum of Squares)

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

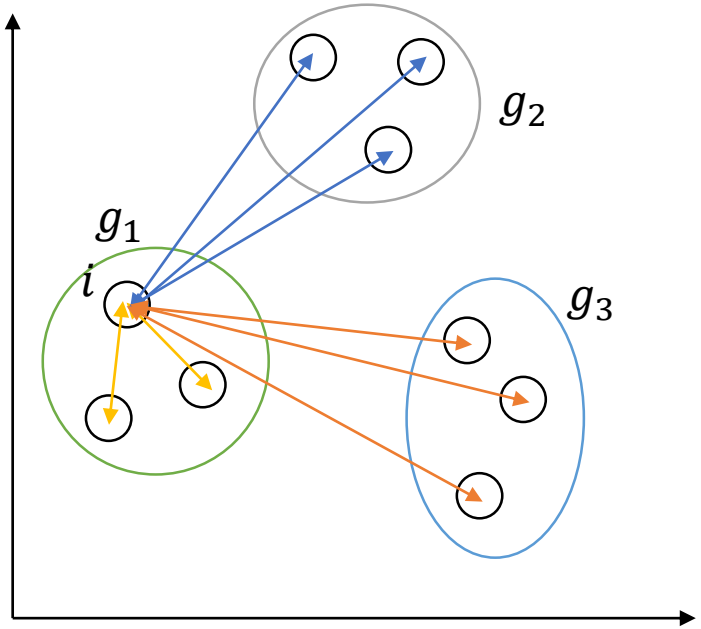


I 최적의 K를 찾는 방법

- K 값을 설정하는 방법
  - Silhouette method
    - 객체와 그 객체가 속한 군집의 데이터들과의 비 유사성(dissimilarity)을 계산하는 방법으로, elbow method 에 비해 상대적으로 간단함
    - $a(i)$: 객체 i 와 그 객체가 속한 군집의 데이터들과의 비 유사성

$$a(i) = \frac{1}{|g(x_i)| - 1} \sum_{j \in g(x_i)} d(x_i, x_j)$$

*$g(x_i)$: x_i가 속한 군집



- $b(i)$: 그 객체가 속하지 않은 다른 군집의 모든 데이터들과의 비 유사성의 최솟값 (가장 가까운 군집)

$$b(i) = \min_k \left(\frac{1}{|g_k|} \sum_{j \in g_k} d(x_i, x_j) \right)$$

*g_k: x_i가 속하지 않은 다른 군집 k

I 최적의 K를 찾는 방법

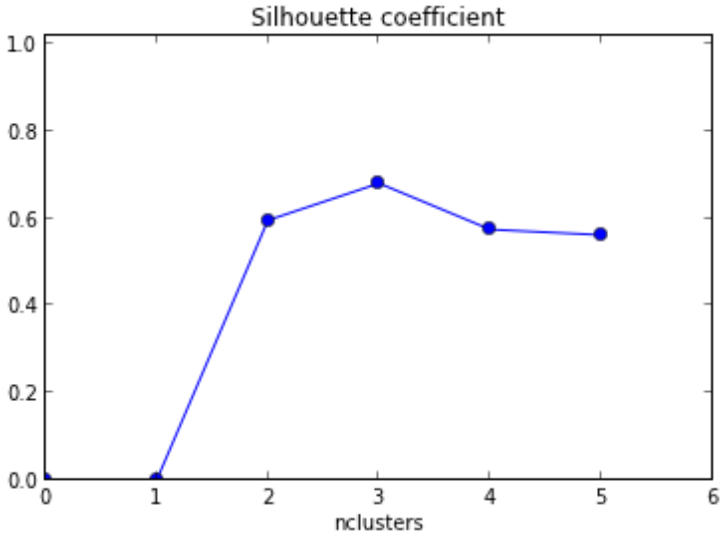
- K 값을 설정하는 방법
  - Silhouette method
    - $a(i)$와 $b(i)$가 정의되었을 때, 실루엣 $s(i)$는 다음과 같이 계산함

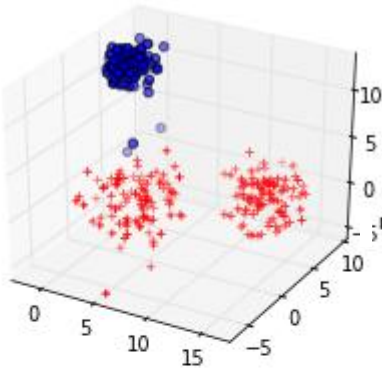
$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \qquad \text{where } -1 \leq s(i) \leq 1$$

    - $s(i)$의 값이 1에 가까울수록 객체 i 는 올바른 클러스터에 분류된 것
    - k를 증가시켜가며 평균 실루엣 값 $(\frac{1}{N}\sum_{i=1}^{N} s(i))$, 다른 말로 silhouette coefficient)이 최대가 되는 k를 선택

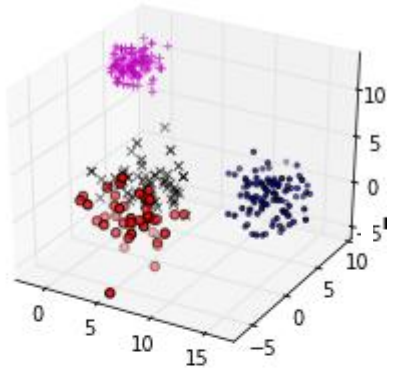
I 최적의 K를 찾는 방법

- K 값을 설정하는 방법
  - Silhouette method
    - ➢ 간단한 예시는 아래와 같음

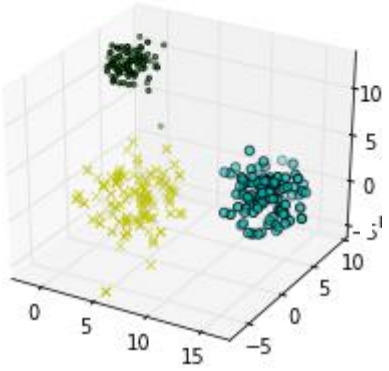




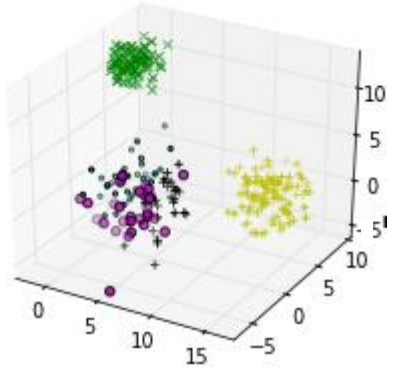
nclusters=2



nclusters=4



nclusters=3



nclusters=5

Part. 05

Clustering

| K-medoids clustering 소개

FASTCAMPUS

ONLINE

머신러닝과 데이터분석 A-Z|

강사. 이경택