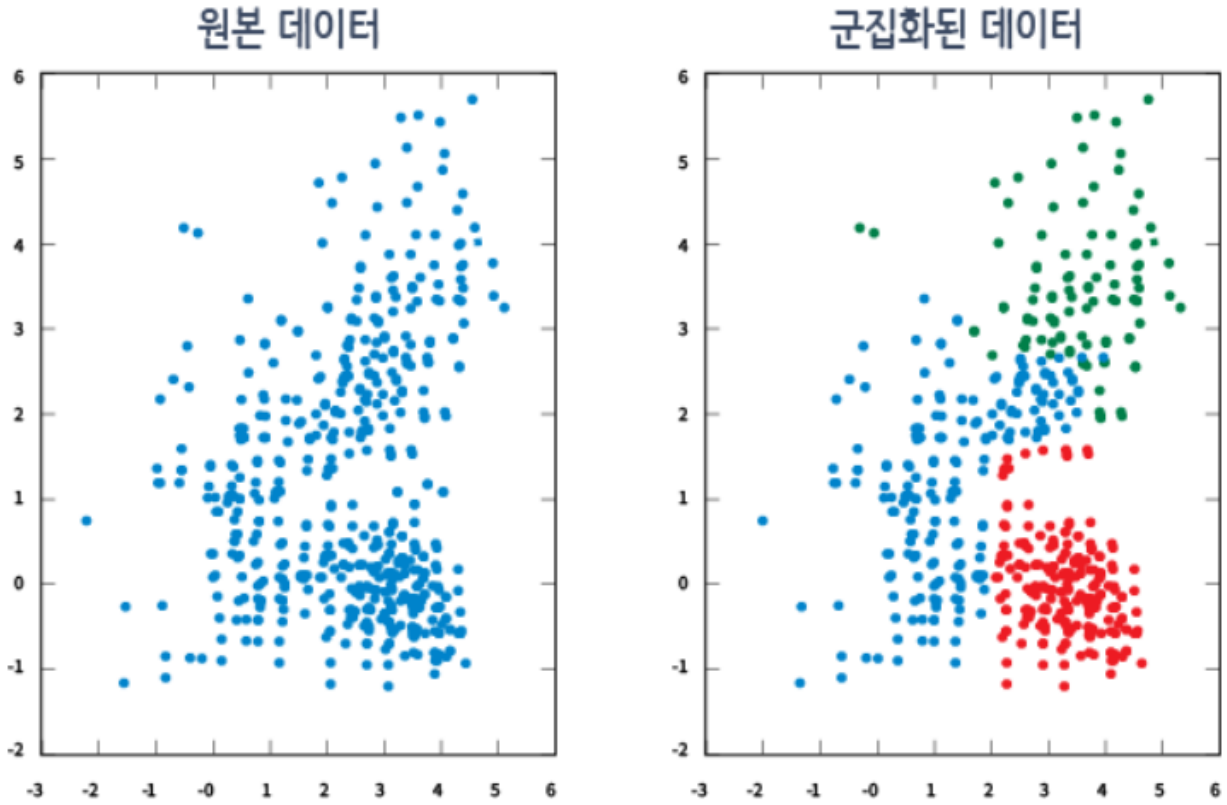
Clustering

컴퓨터학과 20191015 정수민

차례

군집화(Clustering)란?

- 대표적인 비지도학습 알고리즘
- 유사한 속성들을 갖는 관측치들을 묶어 전체 데이터를 몇 개의 군집으로 나누는 것



cf) 비지도 학습

컴퓨터가 스스로 레이블 되어 있지 않은 데이터에 대해 학습하는 것

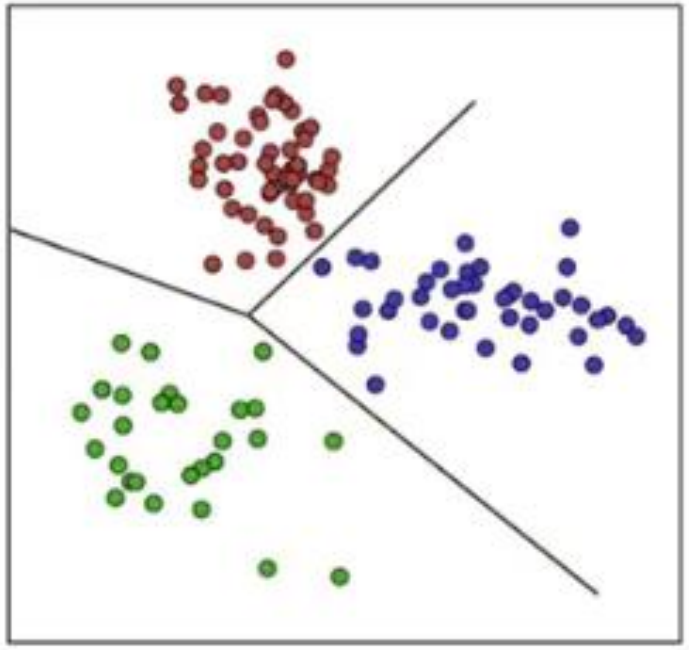
이미지출처 https://m.blog.naver.com/k0sm0s1/221863569856

군집화(Clustering)란?

분류(Classification) vs 군집화(Clustering)

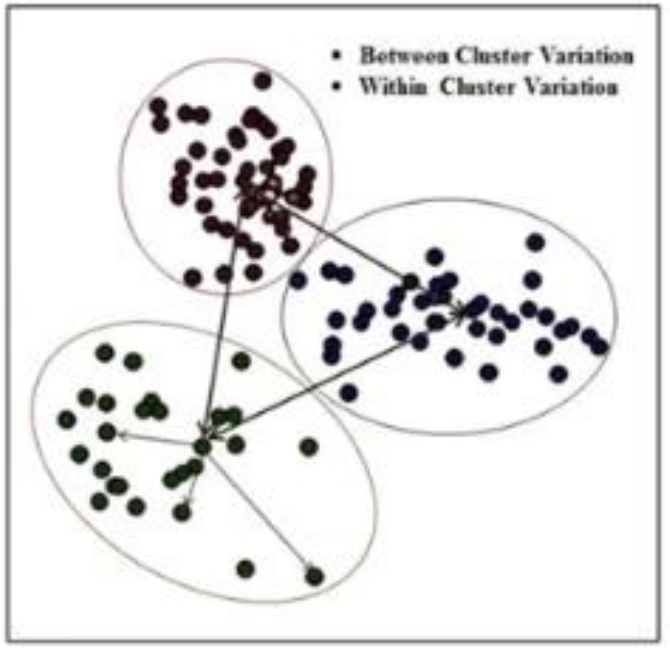
사전 정의된 범주 있음

지도 학습

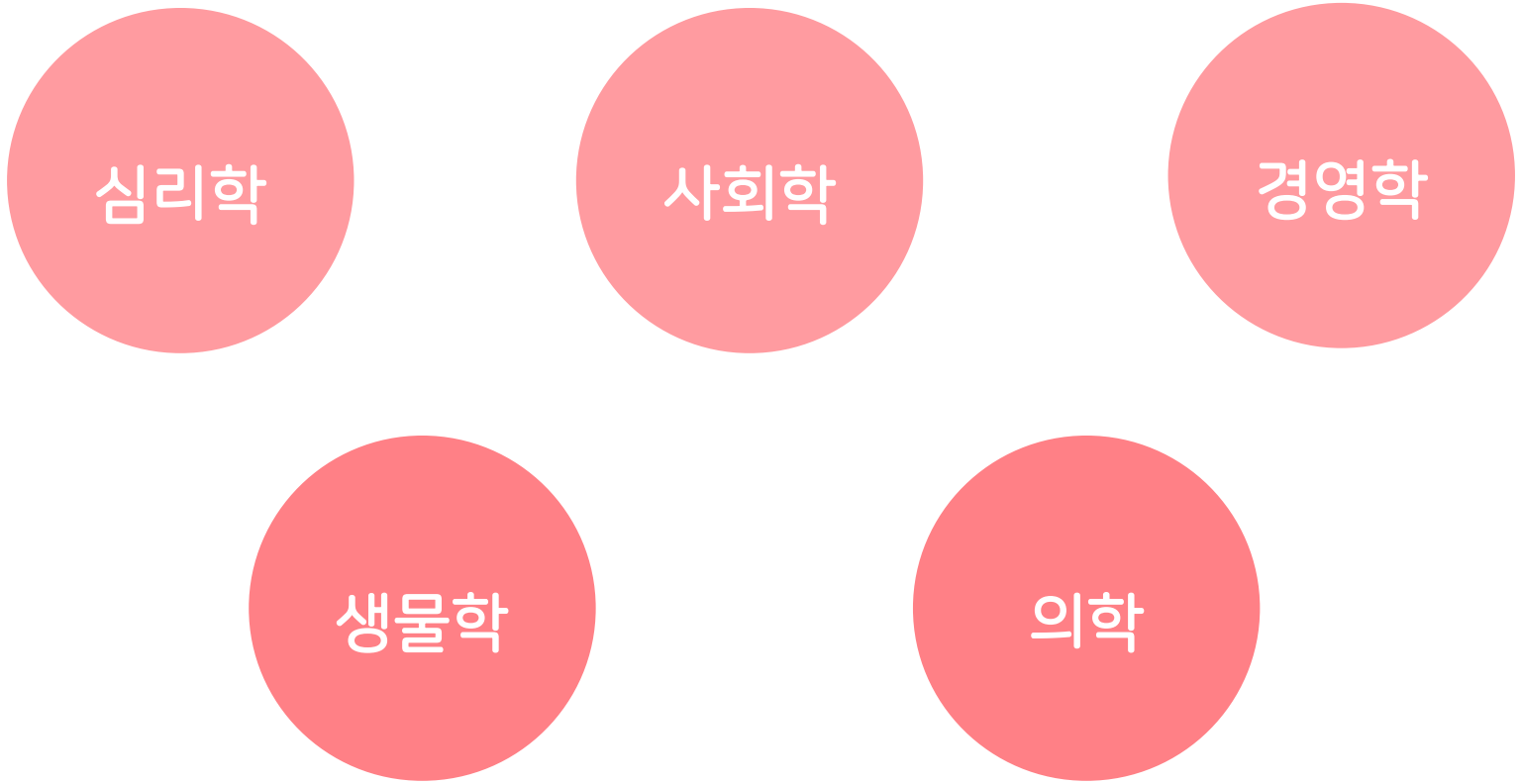


사전 정의된 범주 없음

비지도 학습



군집화 적용 사례



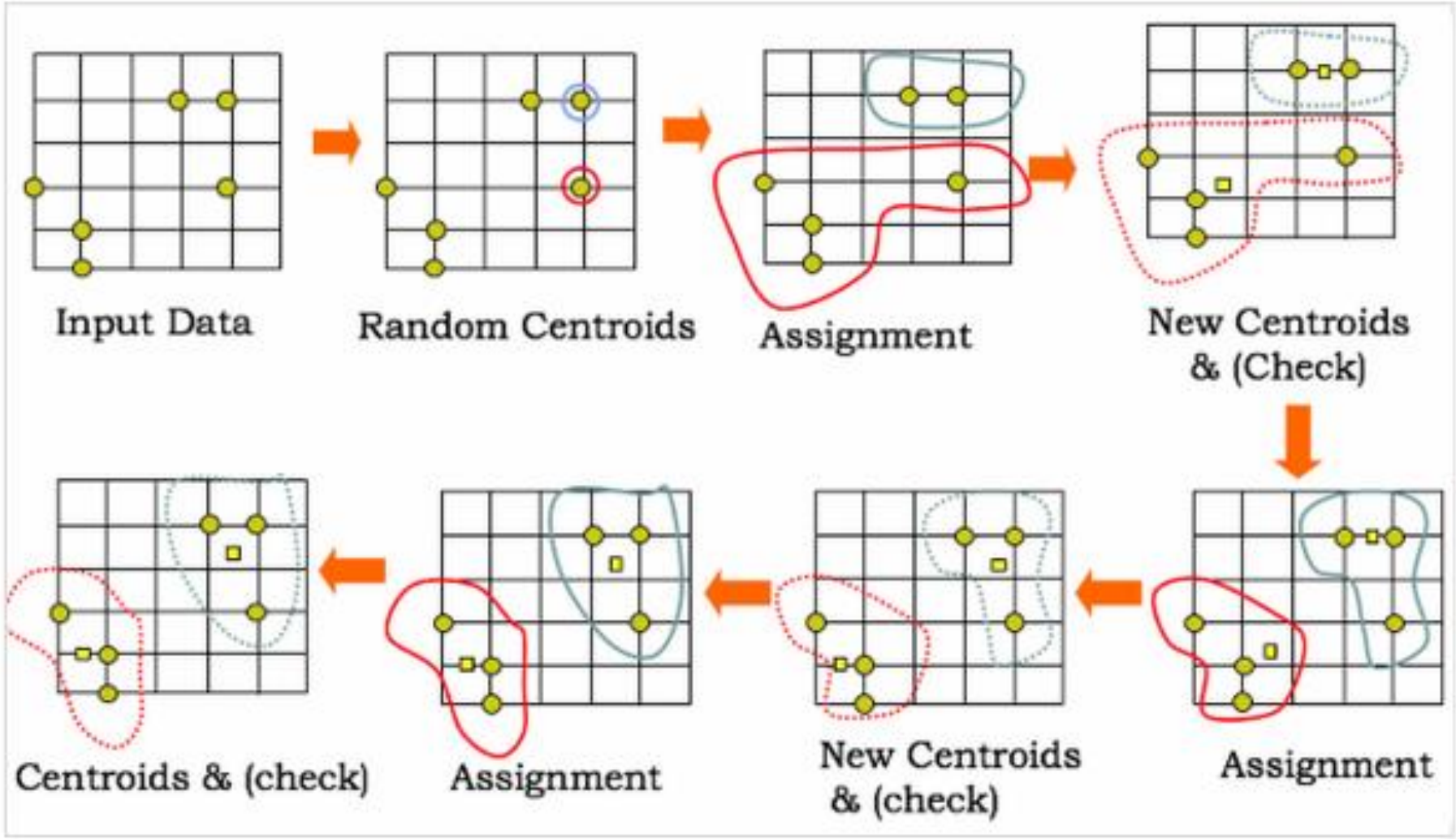
군집화의 종류

이미지 출처 https://muzukphysics.tistory.com/108

K-means

- 각 군집은 하나의 중심(centroid)을 가짐
- 각 개체는 가장 가까운 중심에 할당, 같은 중심에 할당된 개체들이 모여 하나의 군집을 형성
- 사전에 군집의 수 k가 정해져야 알고리즘을 실행할 수 있음

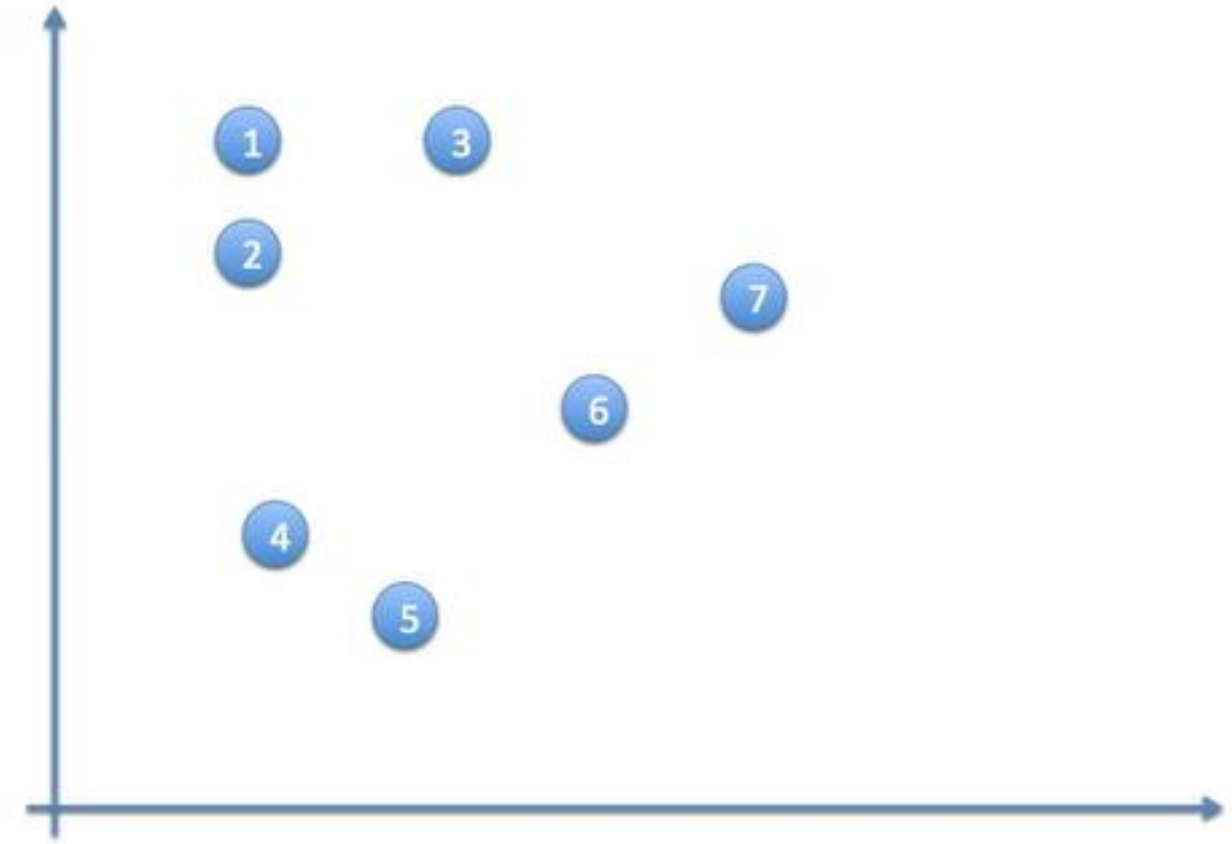
K-means


Input Data
Random Centroids
Assignment
New Centroids & (Check)
Assignment
New Centroids & (check)
Assignment
Centroids & (check)

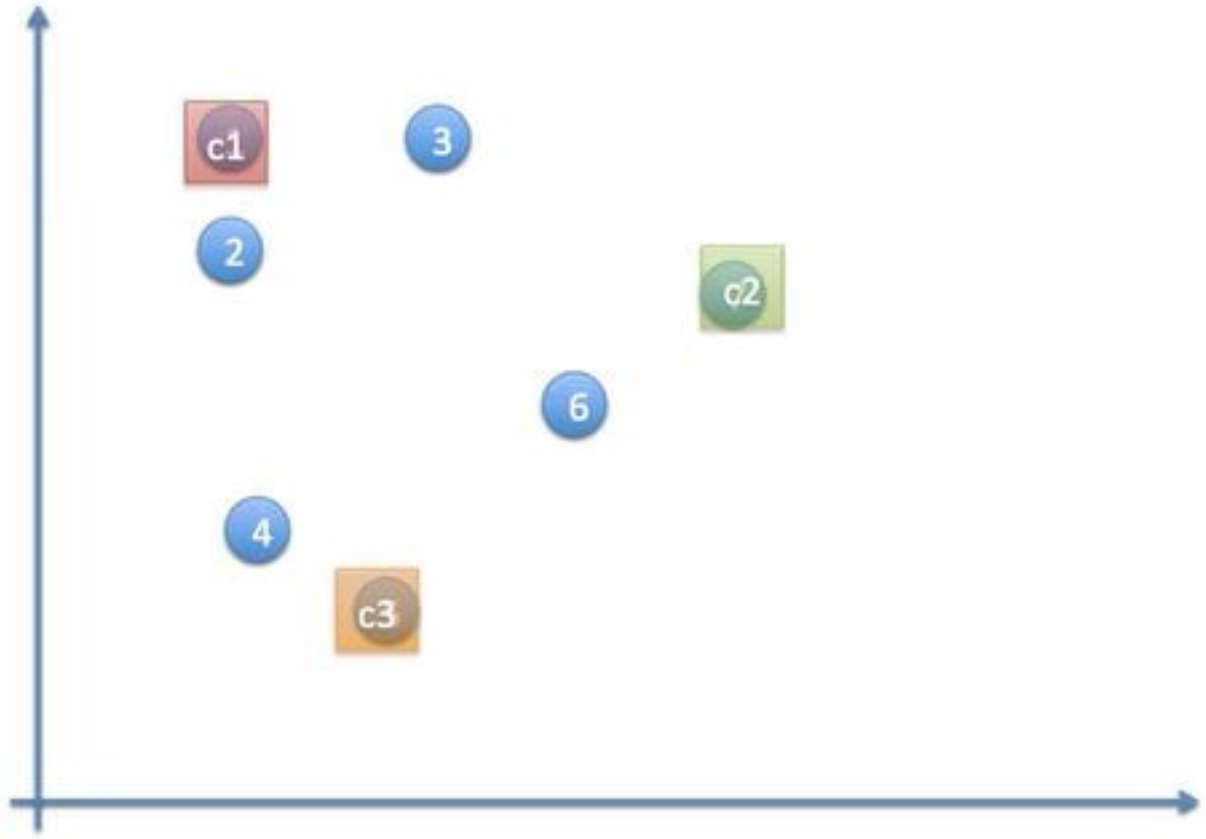
그런데 centroid를 어떻게 정할까?

1. Randomly choose
2. Manually assign init centroid
3. K-mean++

3. K-mean++



3. K-mean++



K-means 장·단점

장점

알고리즘이 단순

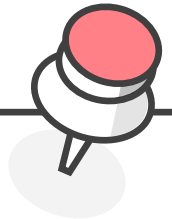
거의 모든 형태의 데이터에 적용이 가능

단점

잡음이나 이상값에 영향을 받기 쉬움

사전에 군집의 수를 정해주어야 함

‘U’형태의 군집일 경우 성능이 떨어짐



군집화 측정/평가 방법

어떻게 군집화 결과를 측정/평가할 것인가?

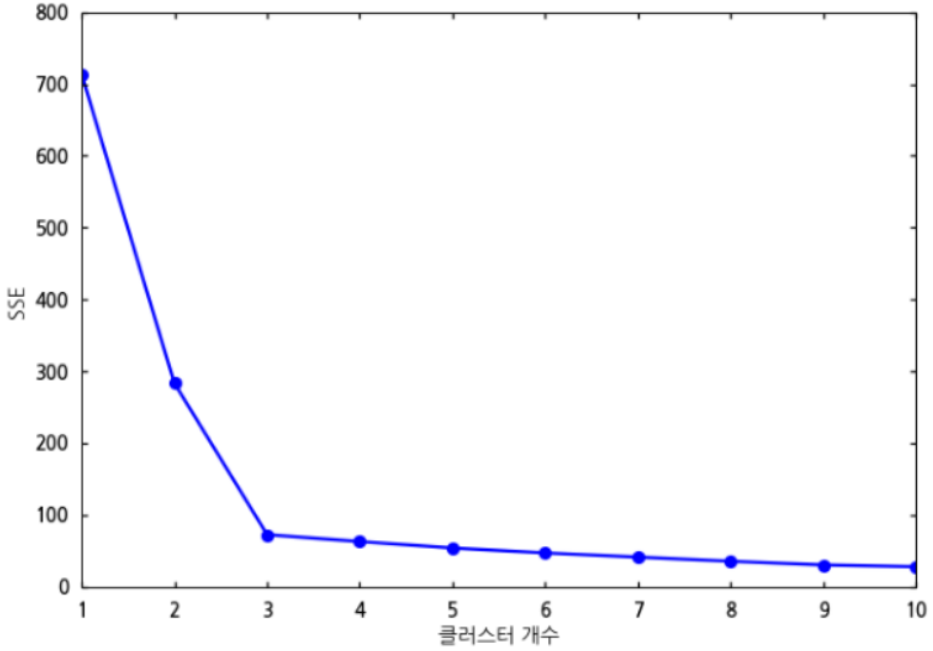
- 정답(목표 변수)이 없기 때문에 일반적인 머신러닝 알고리즘처럼 단순정확도(Accuracy)와 같은 지표로 평가할 수 없다

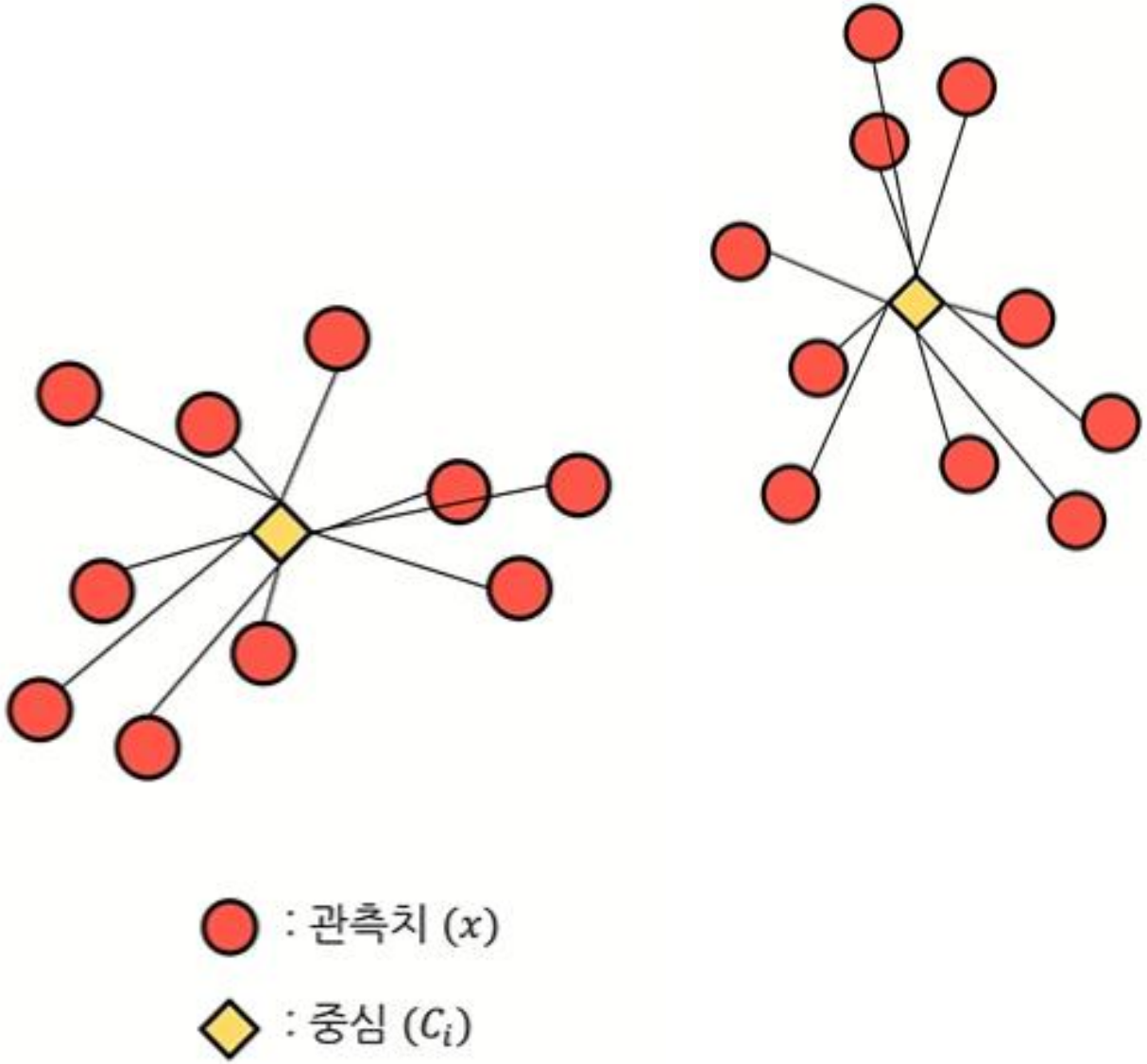
- 군집 타당성 지표(Clustering Validity Index)
  - 군집 간 거리, 군집의 지름, 군집의 분산 등을 고려
  - SSE, Dunn Index, Silhouette 등

군집화 측정/평가 방법

SSE(Sum of Squared Error)

$$SSE = \sum_{i=1}^{K} \sum_{x \epsilon C_i} dist(x, C_i)^2$$





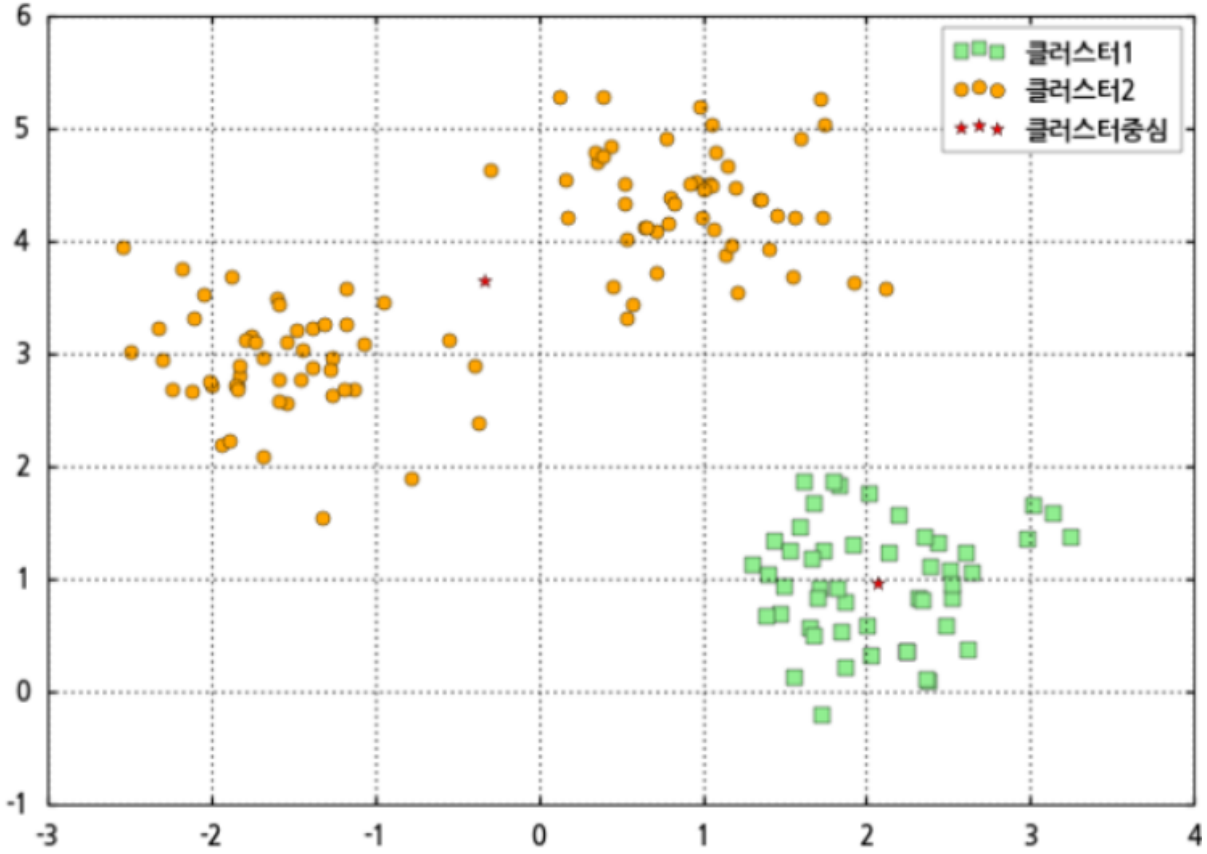
군집화 측정/평가 방법

Silhouette

$$s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \quad (i=1,2,\ldots n)$$

$$-1 \le s(i) \le 1$$

$$\bar{S} = \frac{1}{n}\sum_{i=1}^{n} S(i)$$



참고 자료

- 김성범[소장/인공지능공학연구소]

https://www.youtube.com/watch?v=8zB-_LrAraw

- 이수안 컴퓨터 연구소

https://www.youtube.com/watch?v=jn2HNDJmBZ8&t=2091s -

- Minsuk Heo

https://www.youtube.com/watch?v=9TR54u08lGU

- 군집화의 분류 이미지 출처

https://muzukphysics.tistory.com/108

- K-means 알고리즘 이미지 출처

https://gigle.tistory.com/121