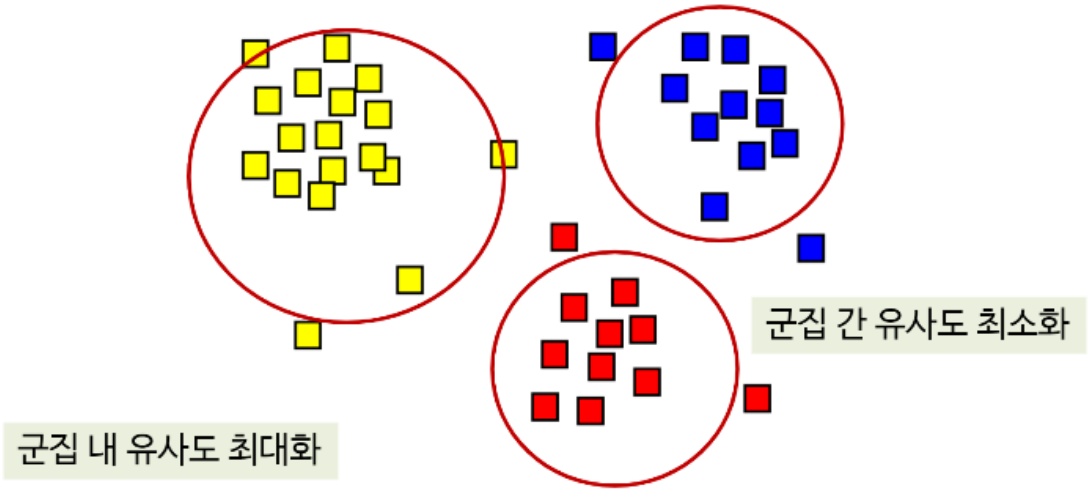
군집화 : 유사한 속성을 갖는 데이터를 묶어 전체 데이터를 몇 개의 군집으로 나누기

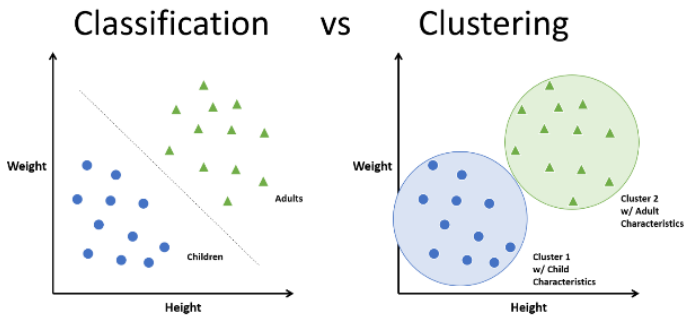
좋은 군집화 :

- 동일한 군집에 소속된 데이터는 서로 유사할수록 좋음(Inter-class similarity)
- 상이한 군집에 소속된 데이터는 서로 다를수록 좋음(Intra-class dissimilarity)



분류(Classification) vs 군집화(Clustering)

- **분류** : 사전 정의된 범주가 있는 데이터로부터 예측 모델을 학습하는 문제. **(Supervised learning)**
- **군집화** : 사전 정의된 범주가 없는 데이터로부터 최적의 그룹을 찾아가는 문제 **(Unsupervised learning)**

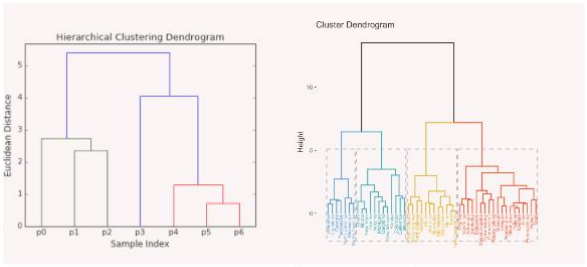


군집화 수행 시 주요 고려사항 :

- **어떻게 최적의 군집 수(K)를 결정할 것인가**
- **어떻게 군집화 결과를 측정/평가할 것인가**
- **어떤 거리 측도를 이용하여 유사도를 측정할 것인가**
  - 유클리디언 거리 (L2)
  - 맨하탄 거리 (L1)
  - 마할라노비스 거리 : 변수 내 분산, 공분산을 모두 반영하여 거리를 계산

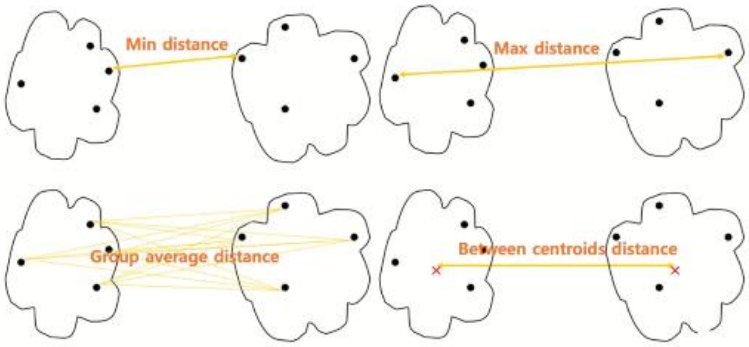
$$d_{Mahalanobis}(X,Y) = \sqrt{(X-Y)\sum^{-1}(X-Y)}$$
$$\text{where } \sum^{-1} = \text{inverse of covariance matrix}$$

- **어떤 군집화 알고리즘을 사용할 것인가**
  - **계층적 군집화** : 가까운 집단부터 차근차근 묶어 나가는 방식
    - 유사한 개체들이 결합되는 덴드로그램 생성 후 적절한 수준에서 자르기



덴드로그램 (Dendrogram)

    - 데이터 간 거리를 계산하는 방법 :
      - Min, max, avg, centroid, Ward



◆ Ward 거리 계산법 : 모든 데이터의 거리 평균(Centroid)를 구하고 이 중심과 모든 데이터 사이의 거리를 구한다. 그리고 각 군집의 평균, 중심을 구한다. 각 군집의 데이터들과 해당 군집의 중심의 거리 평균을 전체 중심 거리 평균에서 뺀다. 이 값이 크다는 것은 각 군집은 서로 멀고 군집 내의 데이터들은 모여 있다는 뜻이다.

$$Ward\ Distance = \sum_{i \in A \cup B} \|x_i - m_{A \cup B}\|^2 - \left\{ \sum_{i \in A} \|x_i - m_A\|^2 + \sum_{i \in B} \|x_i - m_B\|^2 \right\}$$

Ward's distance = 10−(3+2) = 5　　Ward's distance = 7−(3+2) = 2

■ **분리형 군집화** : 전체 데이터 영역을 특정 기준에 의해 동시에 구분. 각 개체들은 사전에 정의된 개수의 군집 중 하나에 속하게 된다.

◆ **K-means Clustering :** 각 군집은 하나의 중심을 가짐. 각 객체는 가장 가까운 중심에 할당되며, 같은 중심에 할당된 개체들이 모여 하나의 군집을 형성. 사전에 군집의 수 K가 정해져야 알고리즘을 수행할 수 있다.

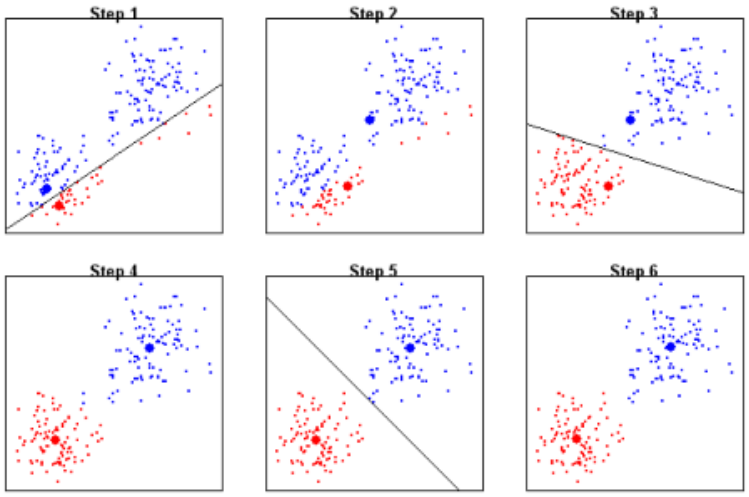
예시(K=2)

1) 두개의 중심을 임의로 생성

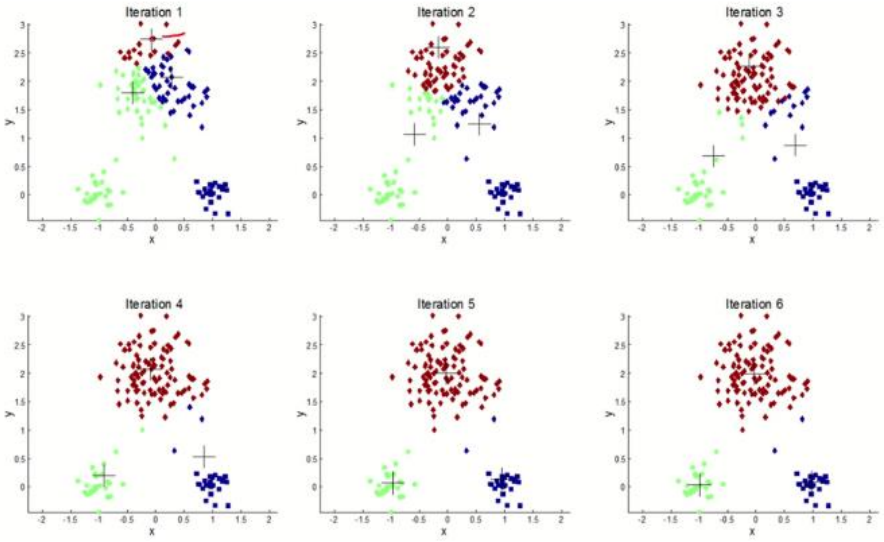
2) 생성된 중심을 기준으로 모든 데이터에 군집 할당

3) 각 군집의 중심 다시 계산

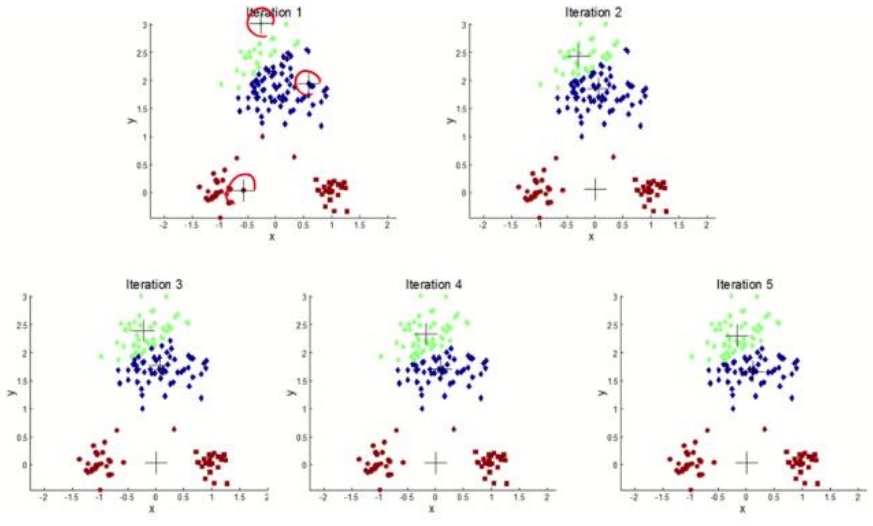
4) 중심이 변하지 않을 때 까지 (2)-(3) 반복



좋은 예시 vs 나쁜 예시



VS



◆ 이러한 Random 초기화의 단점 극복 방법 :

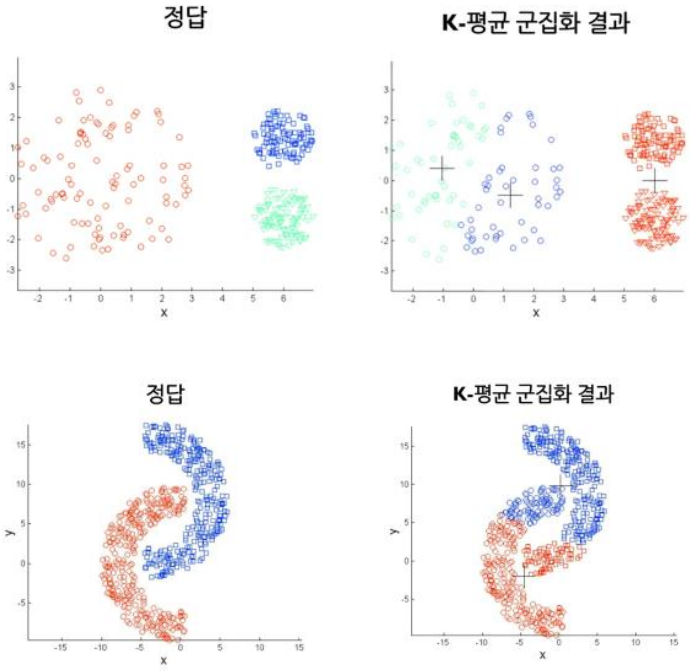
1) 여러 번 Kmeans 군집화를 수행하여 가장 많이 나타나는 결과를 사용 (Ensemble)

2) 데이터 분포 정보를 활용한 초기화 선정

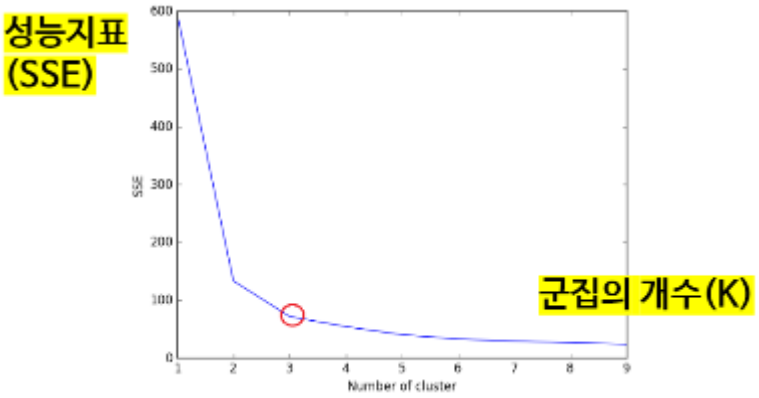
(데이터가 Gaussian 분포라면, 중심을 초기 값으로 선정)

◆ K-means 군집화의 단점 :

- 서로 다른 크기의 군집을 잘 찾아내지 못함
- 서로 다른 밀도의 군집을 잘 찾아내지 못함
- 지역적 패턴이 존재하는 군집을 판별하기 어려움



◆ K 값 선정 방법 : 성능 평가 지표를 통해 최적ㄴ의 군집 수(K) 선택. 일반적으로 Elbow Point에서 최적의 군집 수 결정. 군집화는 지도학습기반 분류 문제처럼 모든 상황에 적용 가능한 평가 지표가 부재하다.

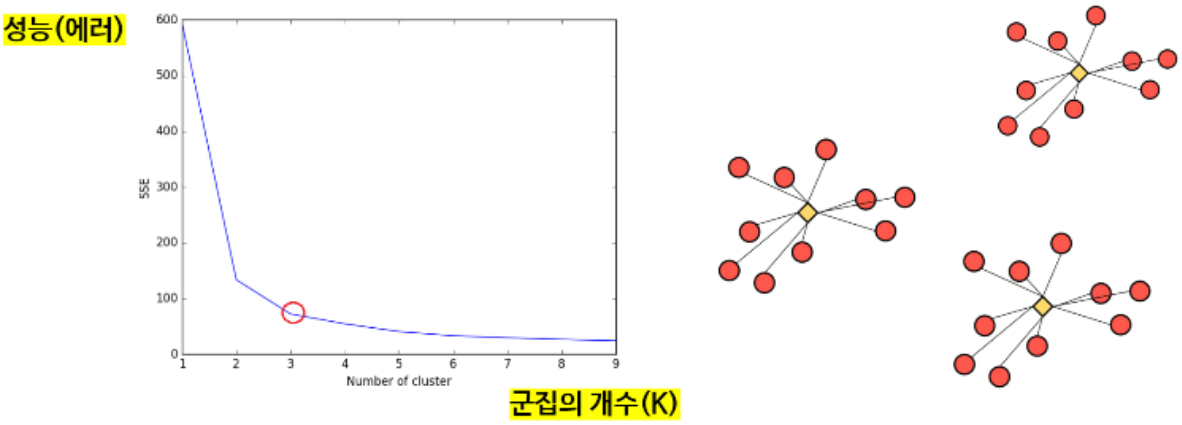


Elbow point

◆ 군집화 평가 지표

- Sum of Squared Error (SSE) : 군집 내 거리 최소화, 군집 간 거리 최대화.

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



- Silhouette 통계량 :

  a(i) = i번째 데이터와 같은 군집 내에 있는 모든 데이터 사이의 평균 거리. 작을수록 유사한 데이터가 잘 모여 있다는 의미

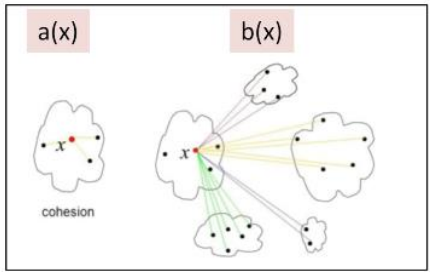
  b(i) = i번째 데이터와 다른 군집 내에 있는 모든 데이터 사이의 최소 거리. 클수록 서로 다른 데이터가 잘 흩어져 있다는 의미

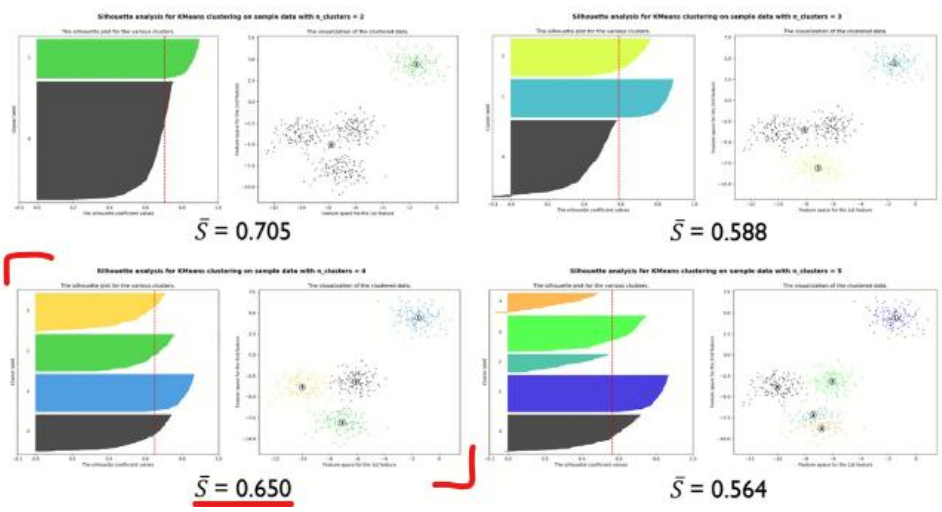
  일반적으로 S 값이 0.5보다 크면 군집 결과가 타당하다고 판단. 1에 가까울수록 Good, -1에 가까울수록 군집화 Bad. K=2인 경우에는 두 번째로 높은 K 값을 선택하는 것이 좋다.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

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

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





■ **분포 기반 군집화** : DBSCAN (Density Based Clustering)

◆ 높은 밀도를 가지고 모여 있는 데이터들을 그룹으로 분류

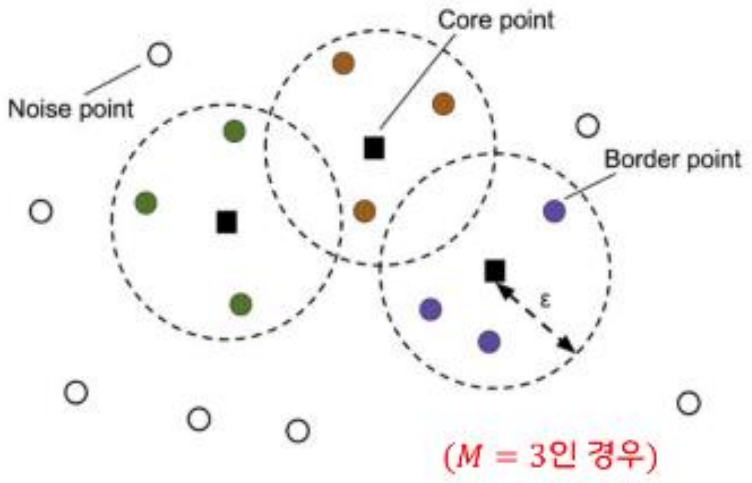
◆ 낮은 밀도를 가지고 있는 데이터는 이상치 또는 잡음으로 분류

◆ 데이터의 ε-neighborhood가 M개 이상의 데이터를 포함하는지 고려하여 분류.

◆ 핵심자료(Core Point) : ε-neighborhood가 M개 이상의 데이터를 포함하는 자료

◆ 주변자료(Border Point) : 핵심자료는 아니지만 ε-neighborhood에 핵심자료를 포함하는 자료

◆ 잡음자료(Noise Point) : 핵심자료도 주변자료도 아닌 자료



◆ DBSCAN 진행 과정

1) 임의 데이터 선택 후 군집1 부여

2) 임의 데이터의 ε-NN을 구하고 NN 데이터의 수가 M 보다 작으면 잡음자료 부여

3) M보다 크면 ε-NN 모두 군집1 부여, 군집1 모든 데이터의 ε-NN 크기가 M보다 큰 것이 없을 때까지 반복

4) 군집2에 대해 동일하게 반복. 즉, 모든 데이터에 군집이 할당되거나 잡음으로 분류될 때까지 절차 (1-3) 반복

◆ 파라미터

- ε : 너무 작으면 많은 데이터가 잡음으로 분류되고 너무 크면 군집의 개수가 적음. HDBSCAN의 경우 ε 자동 설정
- M : 일반적으로 feature의 개수 + 1을 사용