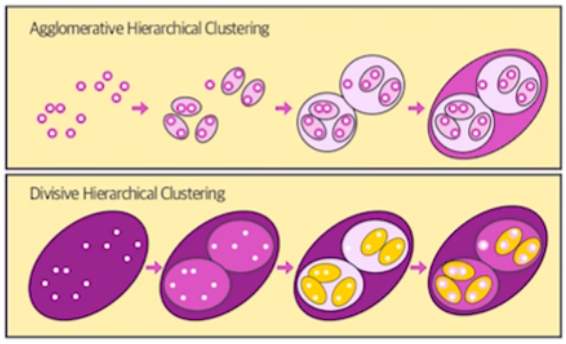
08.계층적 군집분석

- 군집분석 개요
  - 군집분석
    - 어떤 개체나 대상들을 밀접한 유사성(similarity) 또는 비유사성(dissimilarity)에 의하여 유사한 특성을 지닌 개체들을 몇 개의 군집으로 집단화하는 비지도 학습법
    - 각 군집의 특성, 군집간의 차이 등에 대한 탐색대상으로, 집단에 대한 심화된 이해가 목적
    - 특이 군집의 발견, 결측값의 보정 등에도 사용될 수 있음.
  - 군집의 조건
    - 동일 군집에 속한 개체끼리는 유시한 속성이 매우 많음
    - 다른 군집에 속하는 개체끼리는 유사한 속성이 매우 적음
- 계층적 군집분석 개요
  - 병합적(agglomerative) vs 분할적(divisive)
    - 병합적 : 개체 간 거리가 가까운 개체끼리 차례로 묶어주는 방법으로 군집을 정의
    - 분할적 : 개체간 거리가 먼 개체끼리 나누어 가는 방법으로 군집을 정의
    - 계층적 군집분석에는 병합적 방법이 주로 사용됨.

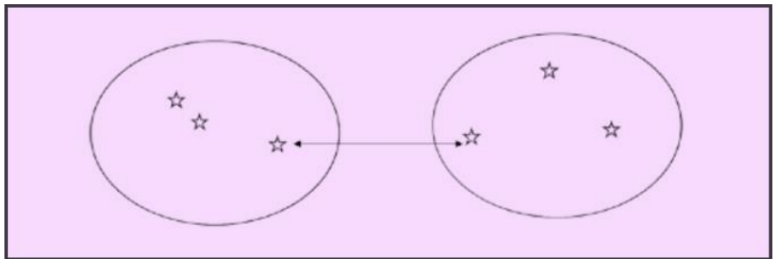


- 개체 간 거리 및 군집 간 거리의 정의
  - 개체 간 거리
    - 유클리디안 거리

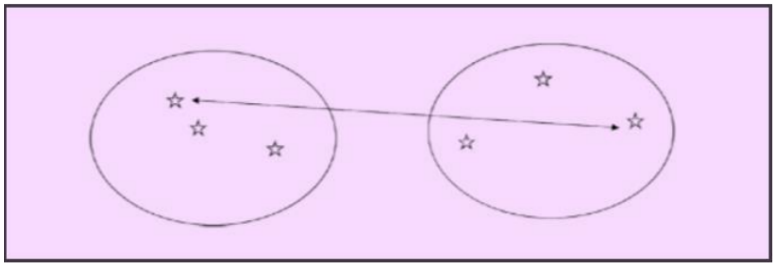
- 맨허튼 거리
- 민코우스키 거리

◦ 군집 간 거리

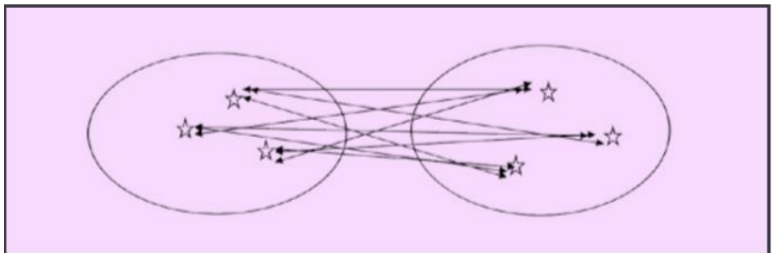
- 단일 연결법(최단 연결법, single linkage)
  - 두군집$C1$과$C2$의거리는$dC1C2 = min\ d\ x,y\ |x \in C1,y \in C2$ 로정의



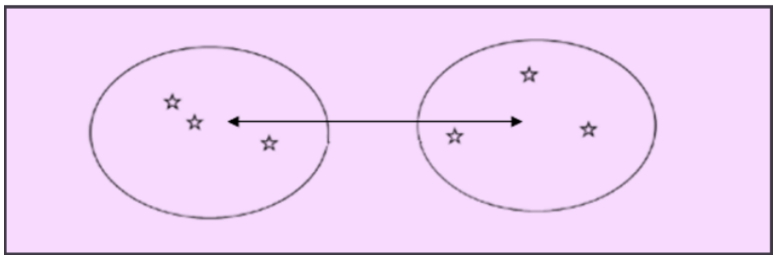
- 완전 연결법(최장 연결법, complete linkage)
  - 두군집$C1$과$C2$의거리는$dC1C2 = max\ d\ x,y\ |x \in C1,y \in C2$ 로정의



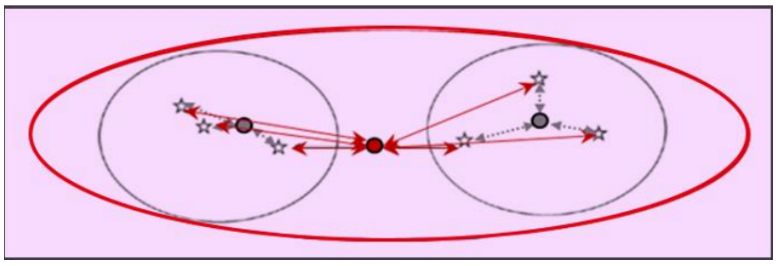
- 평균 연결법(average linkage)
  - 두군집$C1$과$C2$의거리는두군집의모든개체간거리들의평균으로정의



- 중심 연결법(centroid linkage)
  - 두군집$C1$과$C2$의거리는두군집의중심사이의거리로정의.



- 와드 연결법(ward linkage)
  - SSE_k를군집 k 의중심으로부터해당군집각개체간의 거리제곱합으로 정의한뒤, 총K 개의군집이있다면 $SSE = \sum_k^k SSE_k$로정의.(k=1)
  - K개중2 개의군집을하나의군집으로묶었을때 오차제곱합이증가하는정도를두군집간의거리로정의.



- **병합적 방법에서 단일 연결법 사용 군집분석 예시**
  - **이해 못함 다시 볼 것**