계층적 군집분석
(Hierarchical Clustering)

Key words

#병합적 방법 #단일연결법
#완전연결법 #평균연결법 #중심연결법
#와드연결법 #덴드로그램

군집분석 개요

→ 행압축, 시계열압축

군집분석이란

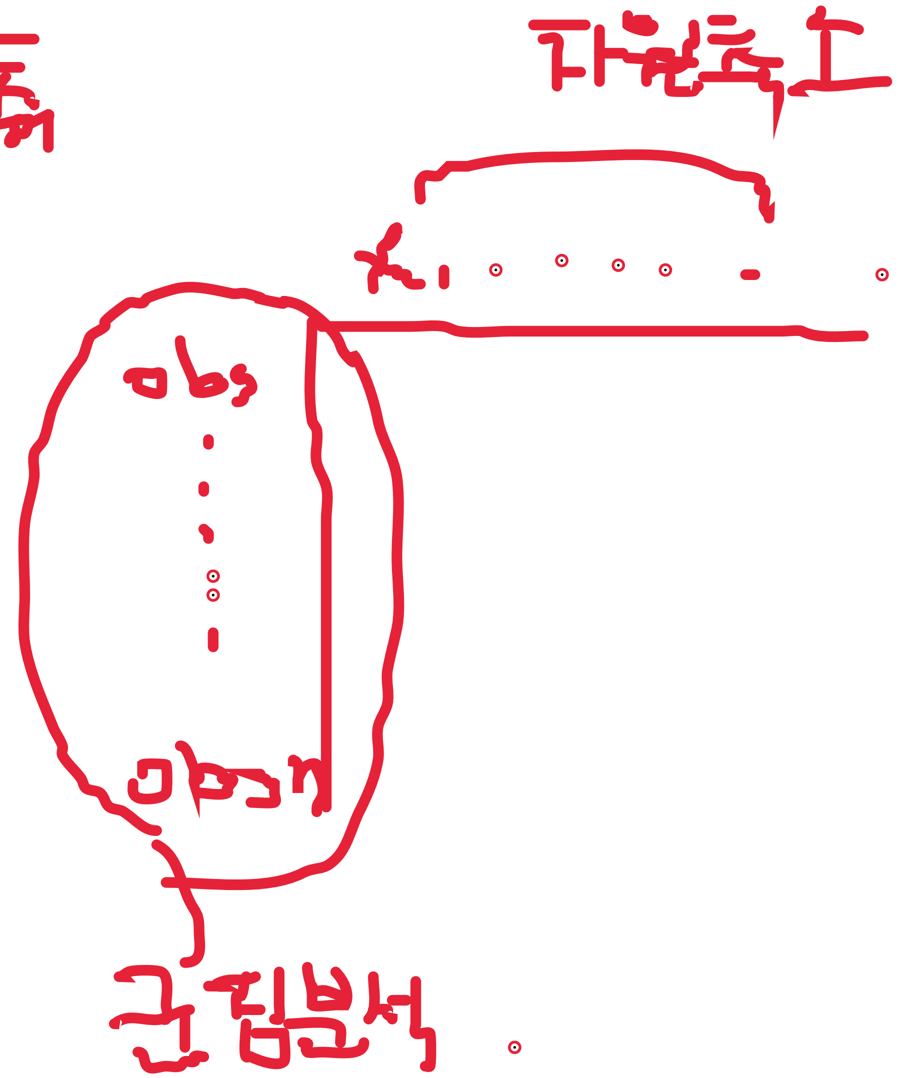
Y 없이 X로만

- 어떤 개체나 대상들을 밀접한 유사성(similarity) 또는 비유사성(dissimilarity)에 의하여 유사한 특성을 지닌 개체들을 몇 개의 군집으로 집단화하는 비지도학습법.

cluster / 정보군집

특성변수만 사용

- 각 군집의 특성, 군집간의 차이 등에 대한 탐색대상으로, 집단에 대한 심화된 이해가 목적.



- 특이 군집의 발견, 결측값의 보정 등에도 사용될 수 있음.

cluster로 판단가능

→ cluster → 다른변수 포함해서 예측

군집분석 개요

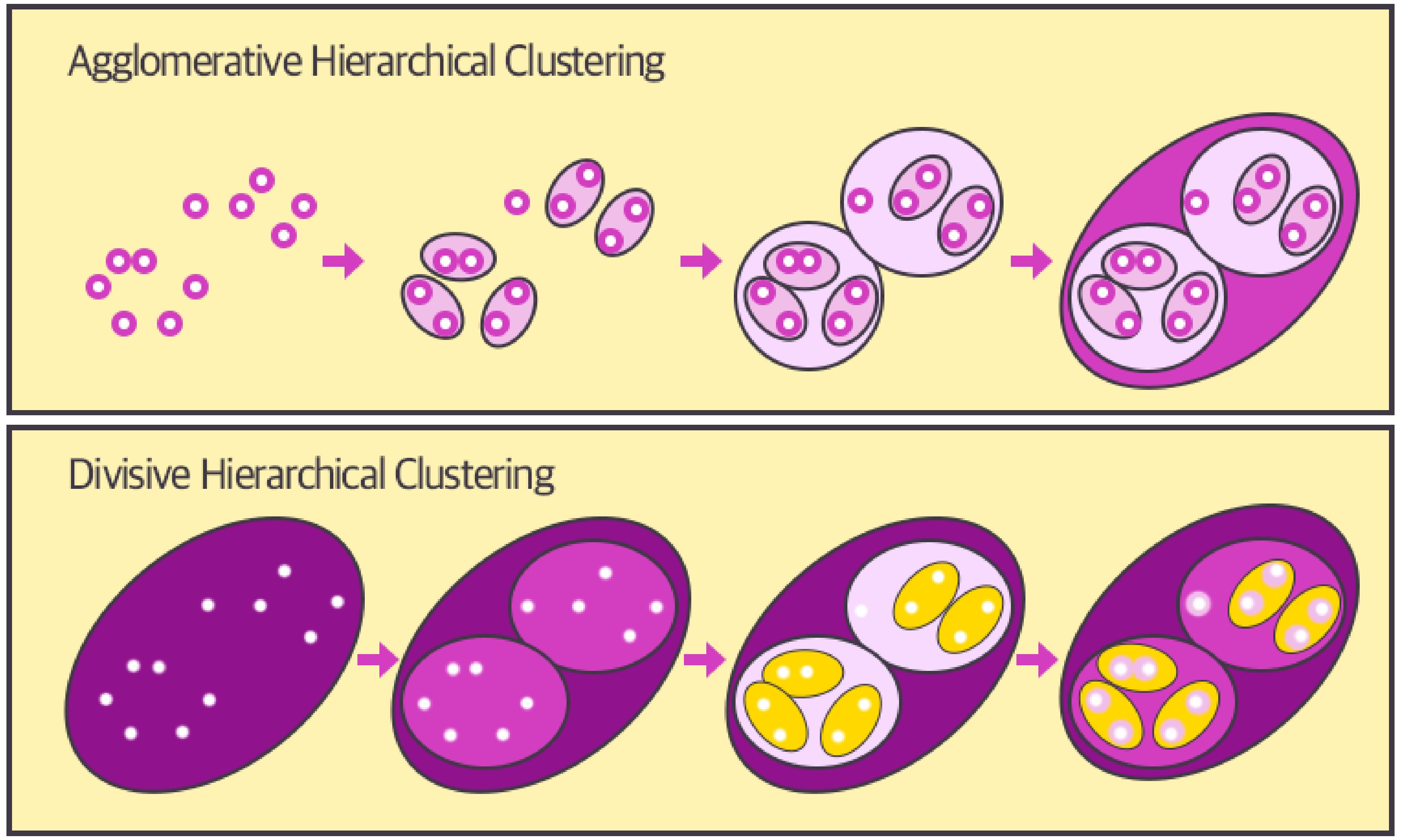
군집의 조건

- 동일 군집에 속한 개체끼리는 유사한 속성이 매우 많음.
- 다른 군집에 속하는 개체끼리는 유사한 속성이 매우 적음.

계층적 군집분석

계층적 군집분석 개요

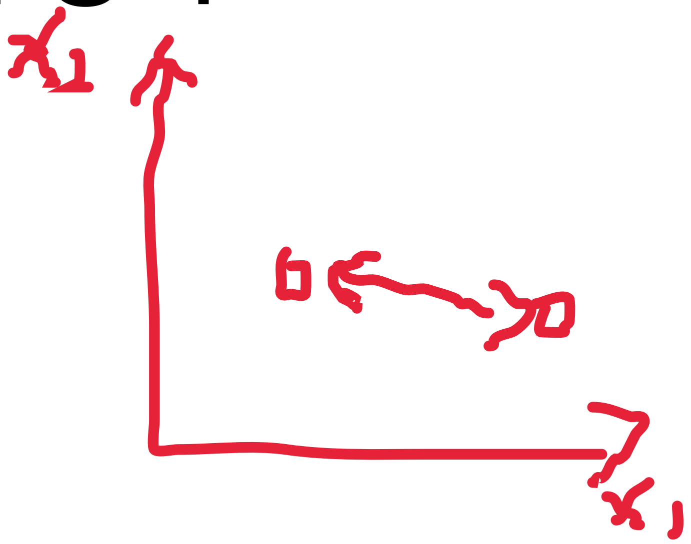
- 병합적(agglomerative) vs 분할적(divisive)
  - 병합적 : 개체 간 거리가 가까운 개체끼리 차례로 묶어주는 방법으로 군집을 정의.
  - 분할적 : 개체 간 거리가 먼 개체끼리 나누어 가는 방법으로 군집을 정의. 전체 1 → 쪼개기.
  - 계층적 군집분석에서는 병합적 방법이 주로 사용됨.

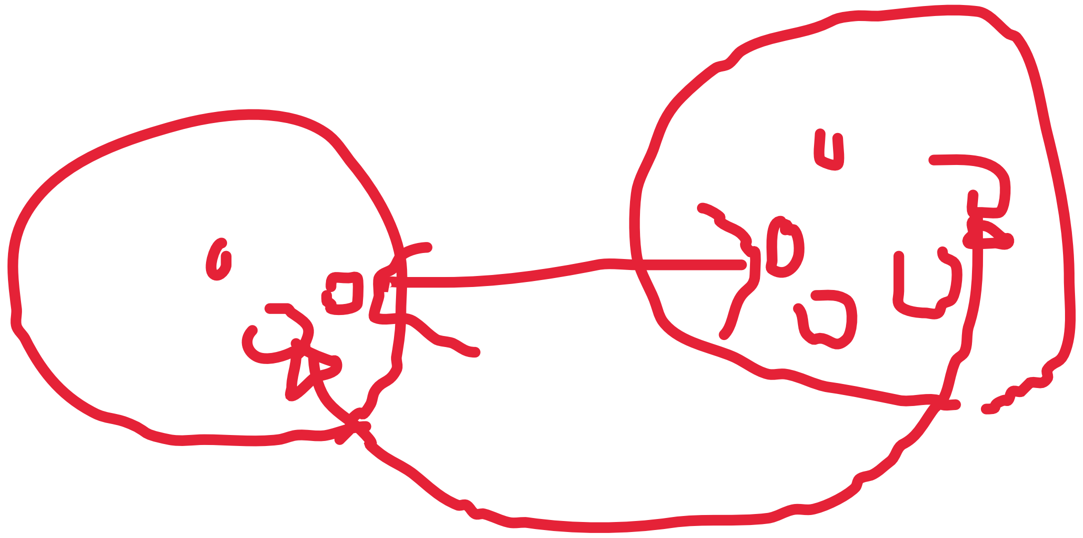


계층적 군집분석

개체 간 거리 및 군집 간 거리의 정의

- 개체 간 거리
  - 유클리디안 거리 — 물리적
  - 맨해튼 거리
  - 민코우스키 거리
- 군집 간 거리
  - 단일 연결법 (최단 연결법, single linkage)
  - 완전 연결법 (최장 연결법, complete linkage)
  - 평균 연결법 (average linkage)
  - 중심 연결법 (centroid linkage)
  - ward 연결방법 (ward linkage)





어느걸로랑 거리계산

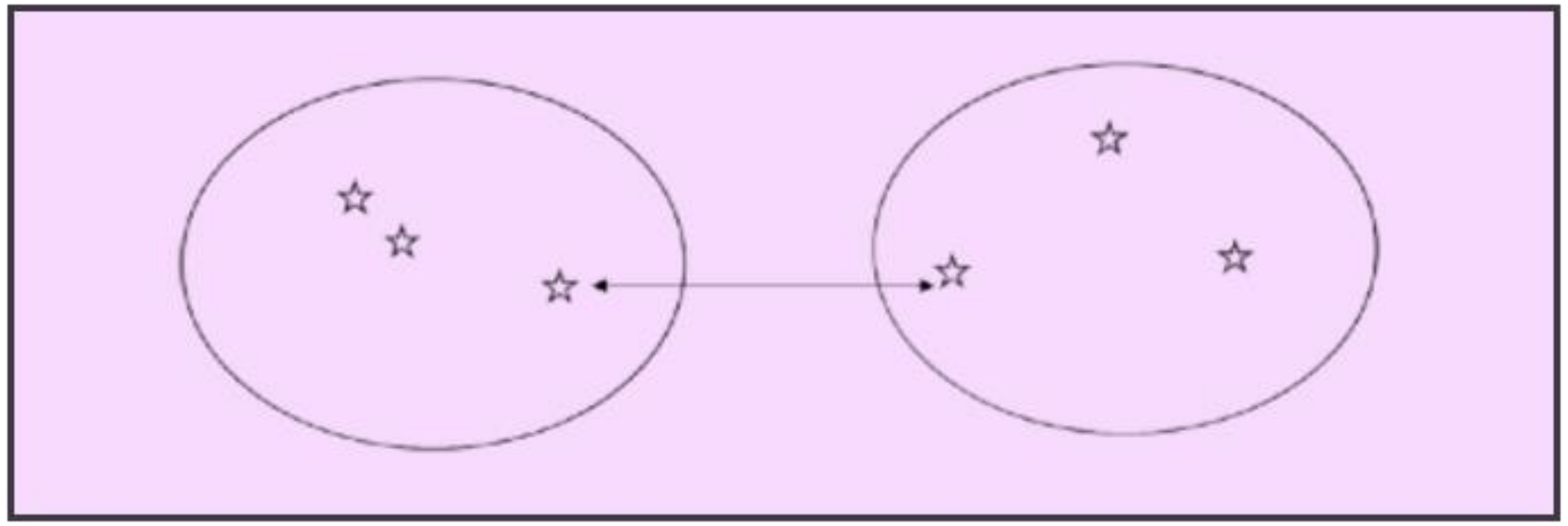
계층적 군집분석

군집 간 거리

- 단일 연결법(single linkage)

최소.

- 두 군집 C_1과 C_2의 거리는 $d_{C_1C_2} = min\{d(x,y)|x \in C_1, y \in C_2\}$로 정의.

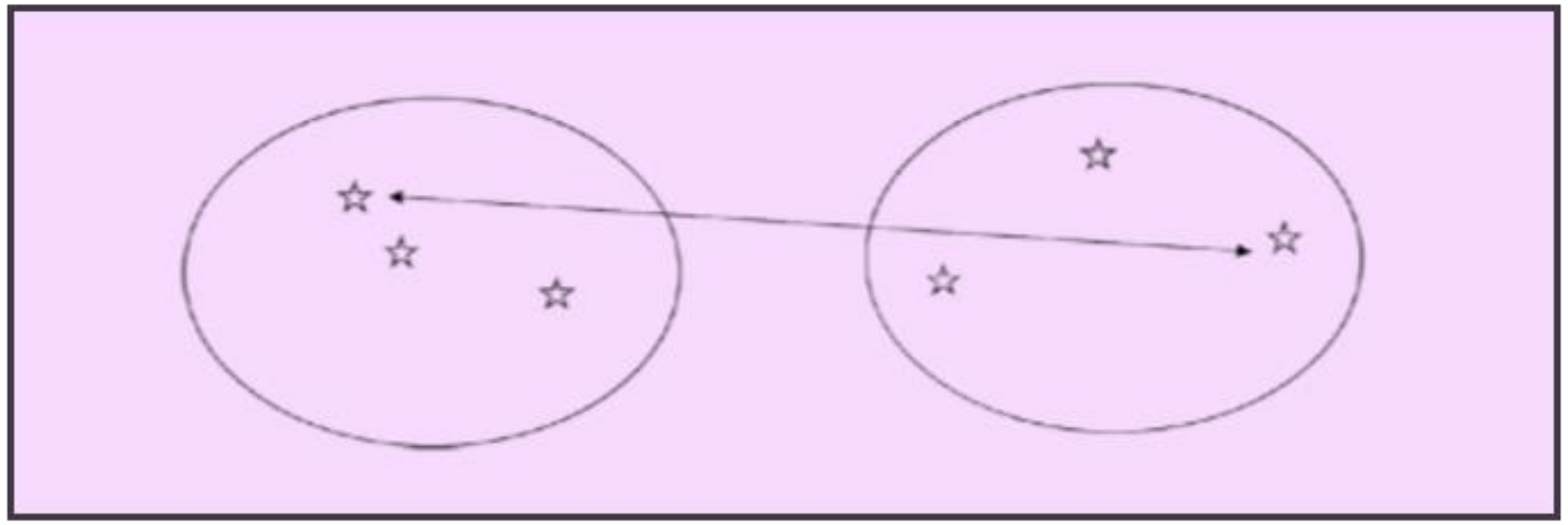


계층적 군집분석

군집 간 거리

- 완전 연결법(complete linkage)

  - 두 군집 C_1과 C_2의 거리는 $d_{C_1C_2} = max\{d(x,y)|x \in C_1, y \in C_2\}$로 정의.



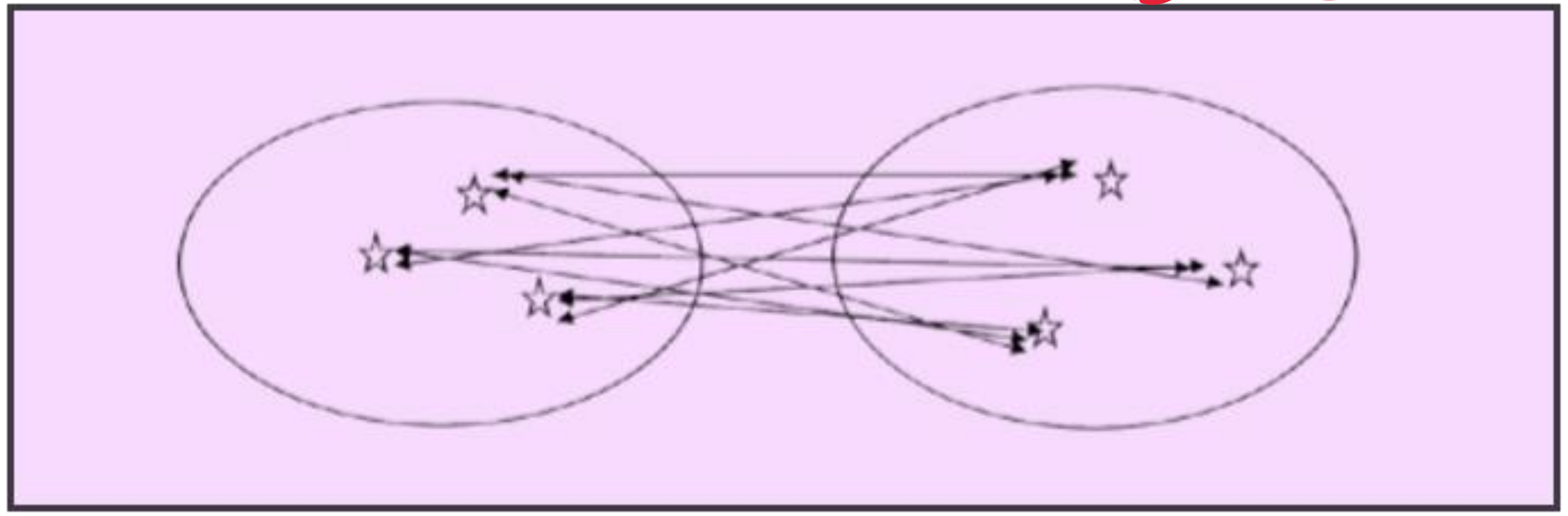
계층적 군집분석

군집 간 거리

- 평균 연결법(average linkage)

거리

- 두 군집 C_1과 C_2의 거리는 두 군집의 모든 개체간 거리들의 평균으로 정의.



계층적 군집분석

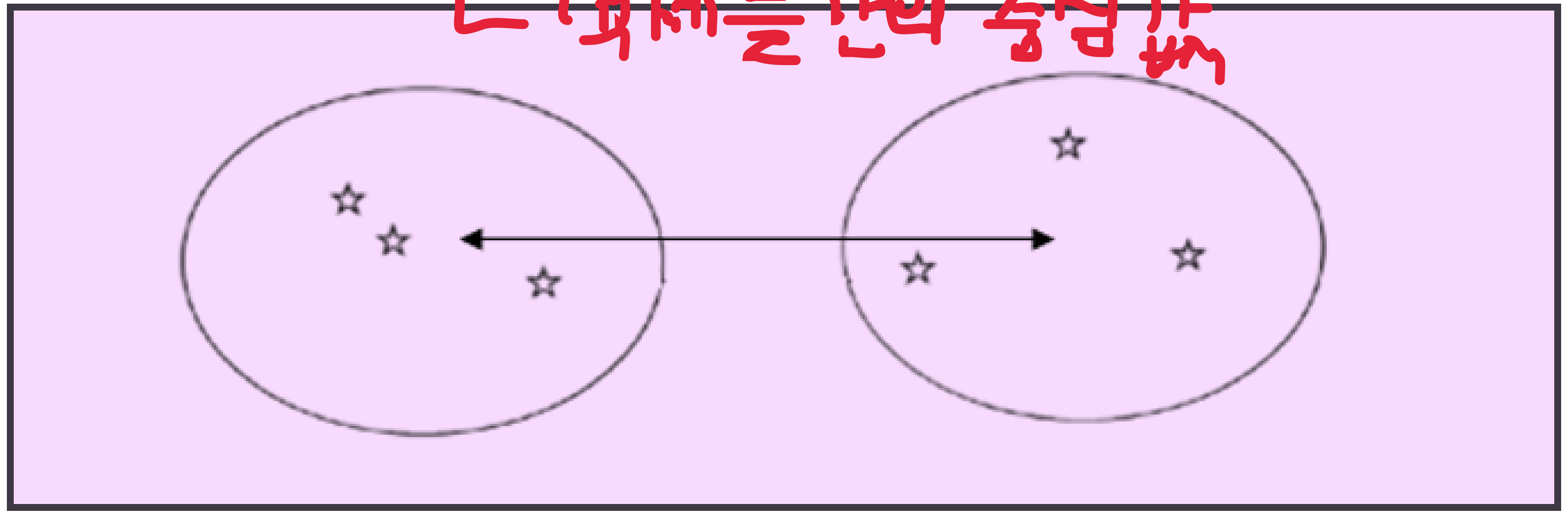
군집 간 거리

center들간의 거리.

- 중심 연결법 (centroid linkage)

  - 두 군집 C_1과 C_2의 거리는 두 군집의 중심 사이의 거리로 정의.

ㄴ개체들간의 중심값



계층적 군집분석

군집 간 거리

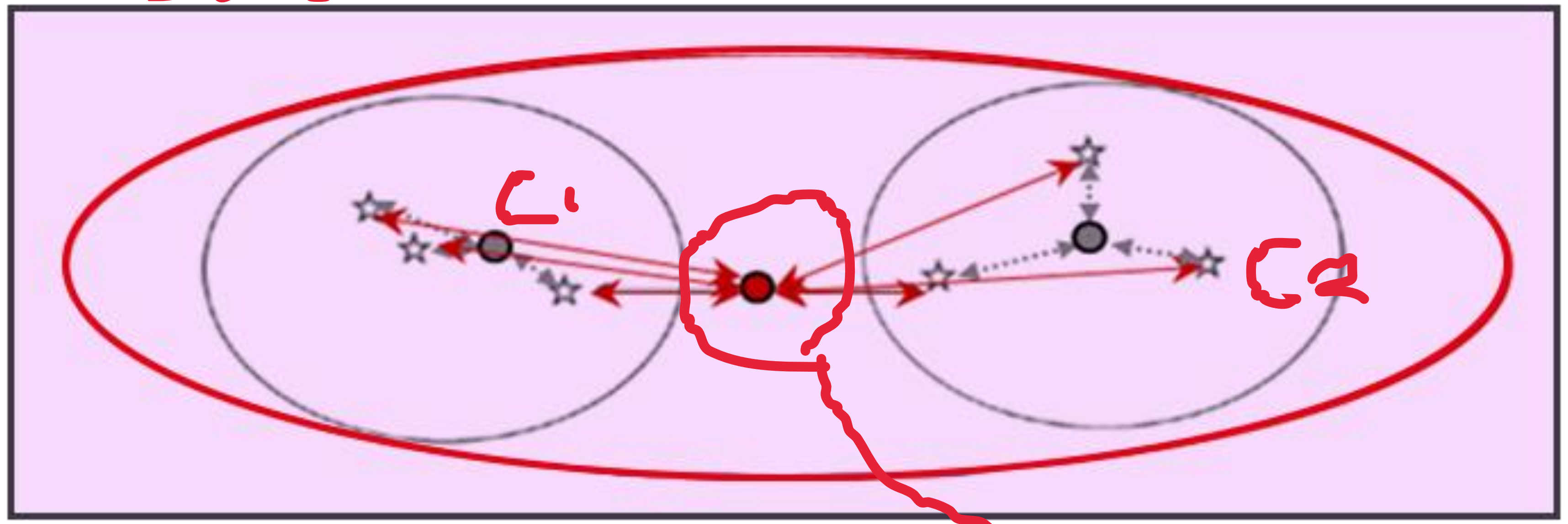
- 와드 연결법(ward linkage)
  - SSE_k를 군집 k 의 중심으로부터 해당 군집 각 개체 간의 거리 제곱 합으로 정의한 뒤, 총 K 개의 군집이 있다면 $SSE = \sum_{k=1}^{K} SSE_k$로 정의.
  - K 개 중 2 개의 군집을 하나의 군집으로 묶었을 때 오차제곱합이 증가하는 정도를 두 군집 간의 거리로 정의.

SSE

거리의 제곱합

$(SSE_1 + SSE_2)$

변동성



Sum of squares

증가 → 변동↑

빨간거의 총합 SSE ... 군집

클수록 거리가 멀다

계층적 군집분석

병합적 방법에서 단일연결법 사용 군집분석 예시

거리행렬

$$
D = \begin{array}{c c}
 & \begin{array}{ccccc} 1 & 2 & \boxed{3} & 4 & 5 \end{array} \\
\begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ \boxed{5} \end{array} &
\begin{bmatrix}
0 & & & & \\
9 & 0 & & & \\
3 & 7 & 0 & & \\
6 & 5 & 9 & 0 & \\
11 & 10 & \underline{2} & 8 & 0
\end{bmatrix}
\end{array}
$$

중복

자신과의 거리

➡ $d_{53} = 2$ 가 가장 작음. **개체 5와 3을 통합.**

계층적 군집분석

병합적 방법에서 단일연결법 사용 군집분석 예시

35 간의거리

$$D = \begin{array}{c|cccc} & (35) & 1 & 2 & 4 \\ \hline (35) & 0 & & & \\ 1 & 3 & 0 & & \\ 2 & 7 & 9 & 0 & \\ 4 & 8 & 6 & 5 & 0 \end{array}$$

$$\Rightarrow \begin{pmatrix} d_{(35)1} = \underline{min(d_{31}, d_{51}) = min(3,11) = 3} \\ d_{(35)2} = min(d_{32}, d_{52}) = min(7,10) = 7 \\ d_{(35)4} = min(d_{34}, d_{54}) = min(9,8) = 8 \end{pmatrix}$$

$d_{(35)1} = 3$으로 가장 작음. **군집 (35)와 개체 1을 통합.**

계층적 군집분석

병합적 방법에서 단일연결법 사용 군집분석 예시

$$D = \begin{array}{c c} & \begin{array}{ccc} (135) & 2 & 4 \end{array} \\ \begin{array}{c} (135) \\ 2 \\ 4 \end{array} & \begin{bmatrix} 0 & & \\ 7 & 0 & \\ 6 & 5 & 0 \end{bmatrix} \end{array} \rightarrow \begin{pmatrix} d_{(135)2} = min\left(d_{(35)2}, d_{12}\right) = min(7,9) = 7 \\ d_{(135)4} = min\left(d_{(35)4}, d_{14}\right) = min(8,6) = 6 \end{pmatrix}$$

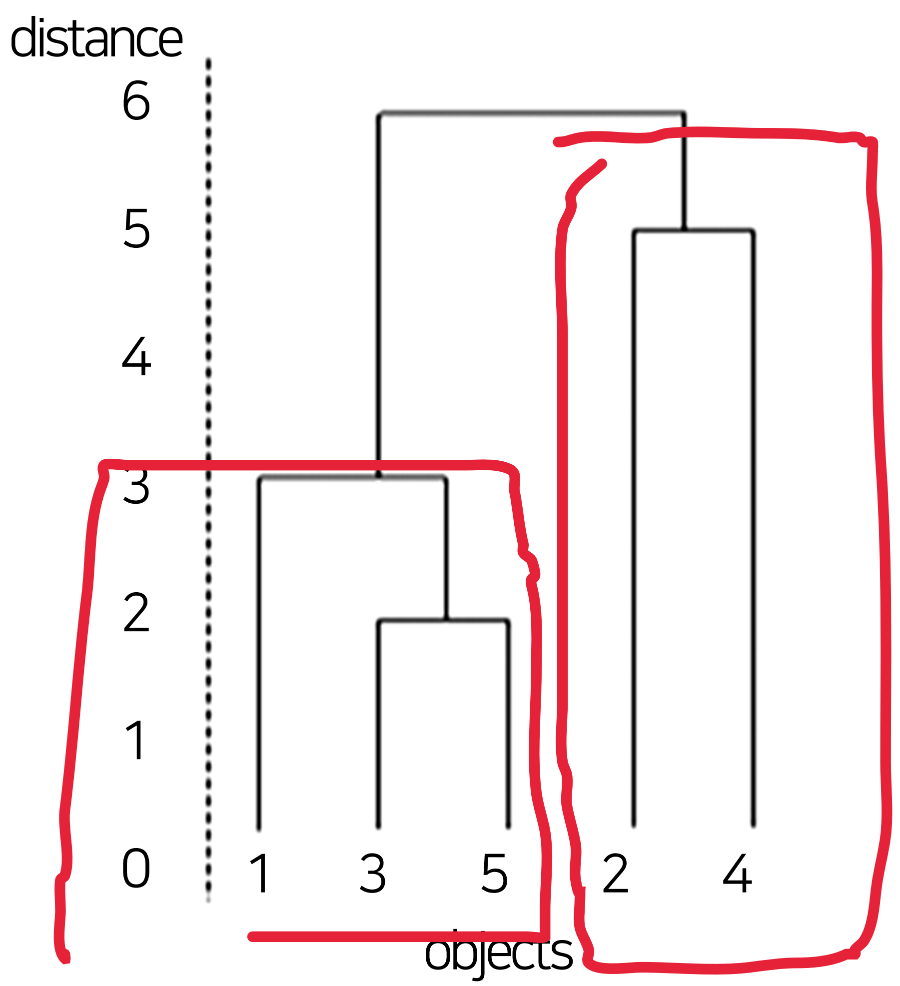
$d_{24} = 5$으로 가장 작음. **개체 2와 4를 통합.**

$$D = \begin{array}{c c} & \begin{array}{cc} (135) & (24) \end{array} \\ \begin{array}{c} (135) \\ (24) \end{array} & \begin{bmatrix} 0 & \\ 6 & 0 \end{bmatrix} \end{array} \rightarrow d_{(24)(135)} = min\left(d_{2(135)}, d_{4(135)}\right) = min(7,6) = 6$$

계층적 군집분석

병합적 방법에서 단일연결법 사용 군집분석 예시

- 군집분석의 결과를 Dendrogram으로 시각화 , 유사 정도에 따라 군집화. cluster 관계확인.
- 군집 간 거리가 멀고, 군집 내 거리가 가까워지도록 적절한 지점에서 절단하여 군집 수 결정



비계층적 군집분석
(K-means Clustering)

Key words

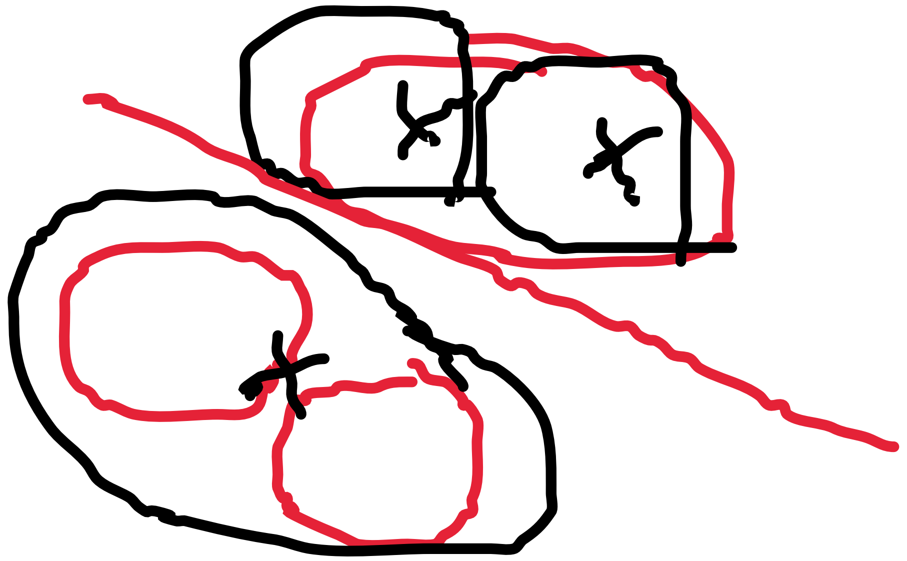
#K평균 군집분석
#군집 오차제곱합(SSE)
#Elbow차트

K-평균 군집분석 개요

K-평균 군집분석 → 큰 용량 + 빠른계산가능

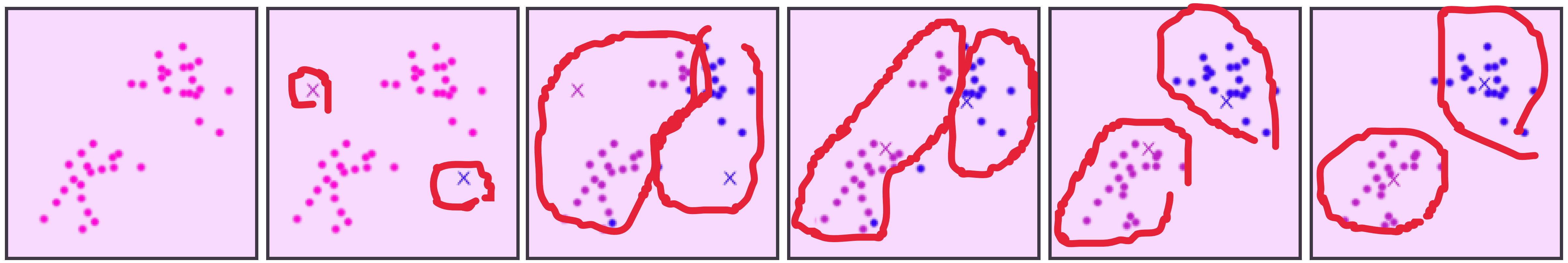
- 사전에 결정된 군집 수 k에 기초하여, 전체 데이터를 상대적으로 유사한 k 개의 군집으로 구분.
- 계층적 방식에 비하여 계산량이 적고, 대용량 데이터를 빠르게 처리함.
- 사전에 적절한 군집 수 k에 대한 예상이 필요.
- 초기에 군집 중심이 어디로 지정되는지에 따라 최종 결과가 영향을 많이 받음.
- 잡음이나 이상치의 영향을 많이 받음.

K-평균 군집분석



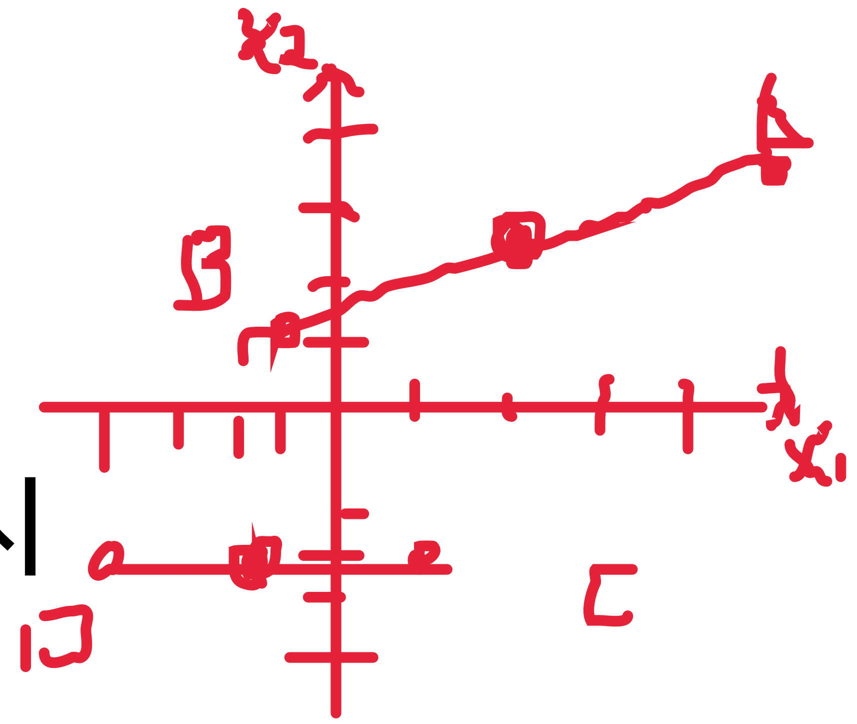
K-평균 군집분석 알고리즘

- 개체를 k 개의 초기 군집으로 나눈다.
- 각 군집의 중심(centroid)을 계산한 뒤 모든 개체들을 각 군집의 중심에 가장 가까운 군집에 할당시킨다.
- 새로운 개체를 받아들이거나 잃은 군집의 중심을 다시 계산한다.
- 위 과정을 더 이상의 재배치가 생기지 않을 때까지 반복한다. 이런식으로 반복.



K-평균 군집분석

K-평균 군집분석 (k -means clustering Method) 예시



관찰치	x_1	x_2
A	5	3
B	-1	1
C	1	-2
D	-3	-2

① 임의로 k=2 개의 군집 (AB), (CD)로 분할.
② 각 군집의 중심을 계산.

(AB)의 중심 : $\bar{x}_1 = 2,\ \bar{x}_2 = 2$
(CD)의 중심 : $\bar{x}_1 = -1,\ \bar{x}_2 = -2$

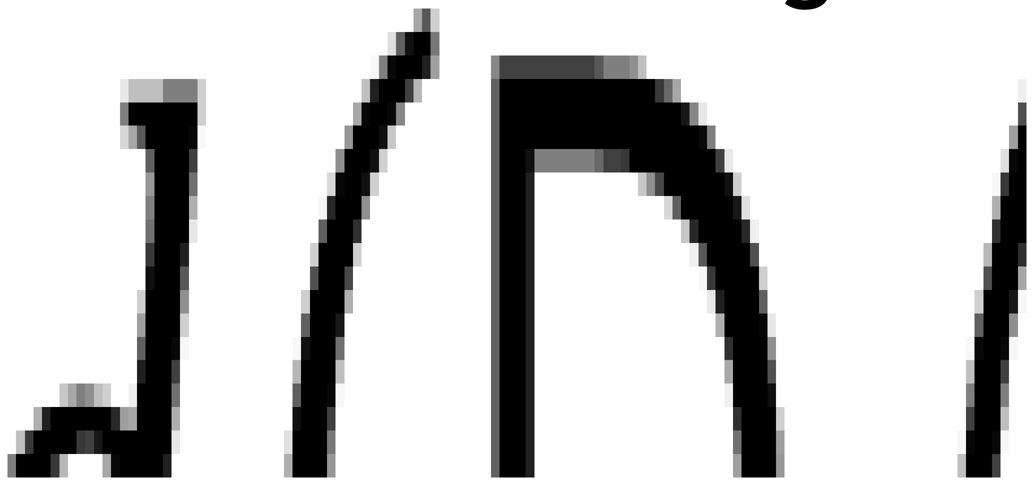
$a = a_1\ a_2$

$b = b_1\ b_2$

$$d(a,b) = \sqrt{(a_1-b_1)^2 + (a_2-b_2)^2}$$

K-평균 군집분석

K-평균 군집분석 (k -means clustering Method) 예시



조

유클리드거리

③ 각 개체에 대하여, 각 군집 중심과의 거리를 계산.

<A> : (AB)에 더 가까움.

$$d(\mathrm{A},(\mathrm{AB})) = \sqrt{(5-2)^2 + (3-2)^2} = \sqrt{10}$$
$$d(\mathrm{A},(\mathrm{CD})) = \sqrt{(5+1)^2 + (3+2)^2} = \sqrt{61}$$

<B> : (CD)에 더 가까움.

$$d(\mathrm{B},(\mathrm{AB})) = \sqrt{(-1-2)^2 + (1-2)^2} = \sqrt{10}$$
$$d(\mathrm{B},(\mathrm{CD})) = \sqrt{(-1+1)^2 + (1+2)^2} = \sqrt{9}$$

<C> : (CD)에 더 가까움.

$$d(\mathrm{C},(\mathrm{AB})) = \sqrt{17}$$
$$d(\mathrm{C},(\mathrm{CD})) = \sqrt{4}$$

<D> : (CD)에 더 가까움.

$$d(\mathrm{D},(\mathrm{AB})) = \sqrt{41}$$
$$d(\mathrm{D},(\mathrm{CD})) = \sqrt{4}$$

K-평균 군집분석

K-평균 군집분석 (k -means clustering Method) 예시

관찰치	x_1	x_2
A	5	3
B	-1	1
C	1	-2
D	-3	-2

④B는 군집 (CD)에 더 가까우므로, B를 (CD)에 통합하여 (BCD) 군집으로 정의. 나머지 개체는 변화가 없으므로 변동 없음. 내보내기!

⑤ 다시 군집의 중심값 계산. 재소속.

(A)의 중심 :$\bar{x}_1 = 5,\ \bar{x}_2 = 3$

(BCD)의 중심 :$\bar{x}_1 = -1,\ \bar{x}_2 = -1$

K-평균 군집분석

K-평균 군집분석 (k -means clustering Method) 예시

관찰치	x_1	x_2
A	5	3
B	-1	1
C	1	-2
D	-3	-2

⑥ 군집 중심에서 각 개체간의 거리를 계산

	A	B	C	D
(5, 3) A	$\sqrt{0}$	$\sqrt{40}$	$\sqrt{41}$	$\sqrt{89}$
BCD (-1, -1)	$\sqrt{52}$	$\sqrt{4}$	$\sqrt{5}$	$\sqrt{5}$

K-평균 군집분석

K-평균 군집분석 (k -means clustering Method) 예시

관찰치	x_1	x_2
A	5	3
B	-1	1
C	1	-2
D	-3	-2

⑦ 다른 군집 중심에 더 가까운 개체가 없으므로 종료.
최종 군집은 (A)와 (BCD)가 됨.

K-평균 군집분석

k=3

SSE1 SSE2 SSE3

K-평균 군집분석에서 적절한 군집 수의 결정

미리 지정

- 오차제곱합(SSE, sum of squared error)

  - 각 군집 내 개체들과 해당 군집 중심점과의 거리를 제곱한 값들의 합.
  - 오차제곱합이 작을수록 군집 내 유사성이 높아 잘 응집된 것임.

$SSE = SSE_1 + SSE_2 + SSE_3$

- 군집수 k에 따른 SSE의 변화를 Elbow 차트로 시각화한 뒤, SSE가 급격히 감소하다가 완만해지기 시작하는 시점의 k를 적정 군집수로 판단함.

무조건적 x // Elbow

└ 꺾기는 지점 찾기.