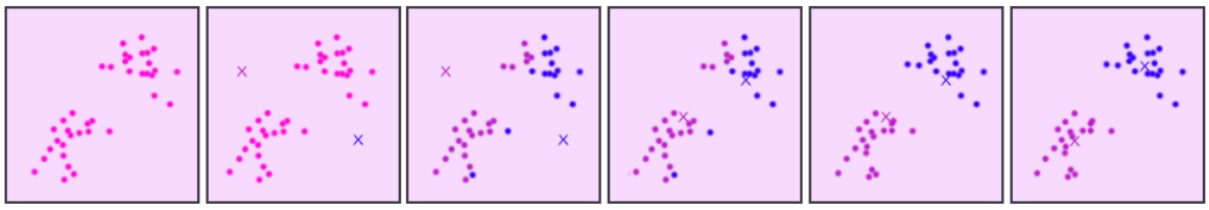
09.비계층적 군집분석

- K-평균 군집분석
  - 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 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



- (k -means clustering Method) 예시
  - ①임의로k=2 개의군집(AB), (CD)로분할.
  - ②각군집의중심을계산.
    - (AB)의 중심: x 1 = 2, x 2 = 2
    - (CD)의 중심: x 1 = −1, x 2 = −2
  - ③각 개체에 대하여, 각 군집 중심과의 거리를 계산.

- <A> : (AB)에 더가까움.
  d A,(AB) = $\sqrt{(5-2)^2+(3-2)^2} = \sqrt{10}$
  d A,(CD) =$\sqrt{(5+1)^2+(3+2)^2} = \sqrt{61}$
- **<B> : (CD)에더가까움.**
  d **B,(AB) =** $\sqrt{(-1-2)^2+(1-2)^2} = \sqrt{10}$
  d **B,(CD) =**$\sqrt{(-1+1)^2+(1+2)^2} = \sqrt{9}$
- **<C>: (CD)에 더가까움.**
  d **C,(AB) =** $\sqrt{17}$
  d **C,(CD) =** $\sqrt{4}$
- **<D>: (CD)에 더가까움.**
  d **D,(AB) =** $\sqrt{41}$
  d **D,(CD) =** $\sqrt{4}$

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

◦ ⑤ 다시 군집의 중심값 계산.

- (A)의 중심 :x 1 = 5, x 2 = 3
- (BCD)의 중심 :x 1 = −1, x 2 = −1

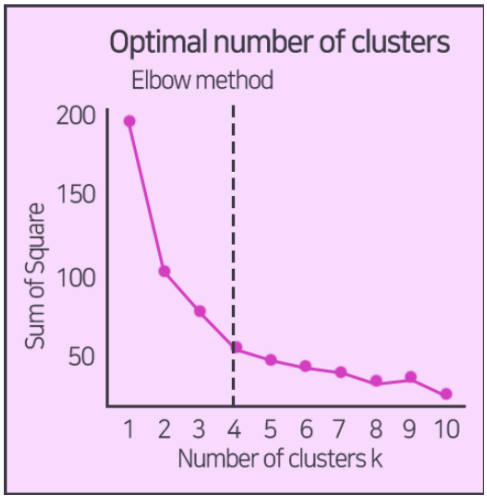
◦ ⑥ 군집중심에서각개체간의거리를계산

	A	B	C	D
A	$\sqrt{0}$	$\sqrt{40}$	$\sqrt{41}$	$\sqrt{89}$
BCD	$\sqrt{52}$	$\sqrt{4}$	$\sqrt{5}$	$\sqrt{5}$

◦ ⑦ 다른군집중심에더가까운개체가없으므로종료. 최종군집은(A)와(BCD)가 됨.

- K-평균 군집분석에서 적절한 군집 수의 결정
  ◦ 오차제곱합(SSE, sum of squared error)
  - 각군집내개체들과해당군집중심점과의거리를제곱한값들의합.
  - 오차제곱합이작을수록군집내유사성이높아잘응집된것임.

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


Optimal number of clusters
Elbow method
Sum of Square
200
150
100
50
1 2 3 4 5 6 7 8 9 10
Number of clusters k