

군집화

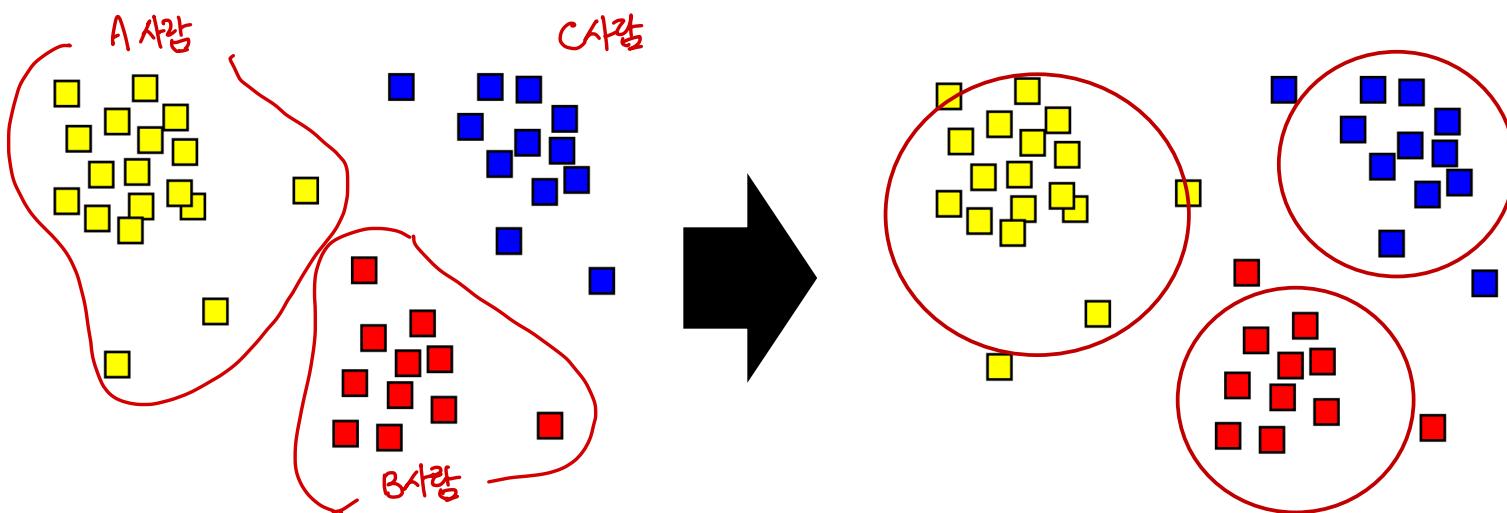
목차

- 군집화 소개
 - 계층형 군집화
 - 분리형 군집화(Kmeans) ☆
 - 분포 기반 군집화(DBSCAN)
 - 실습
- ☆) ↗ 최근가장 사용 多

군집화 개념

- 군집화(Clustering) 개념

- [유사한 속성을 갖는 데이터를 묶어] 전체 데이터를 몇개의 군집으로 나누는 것



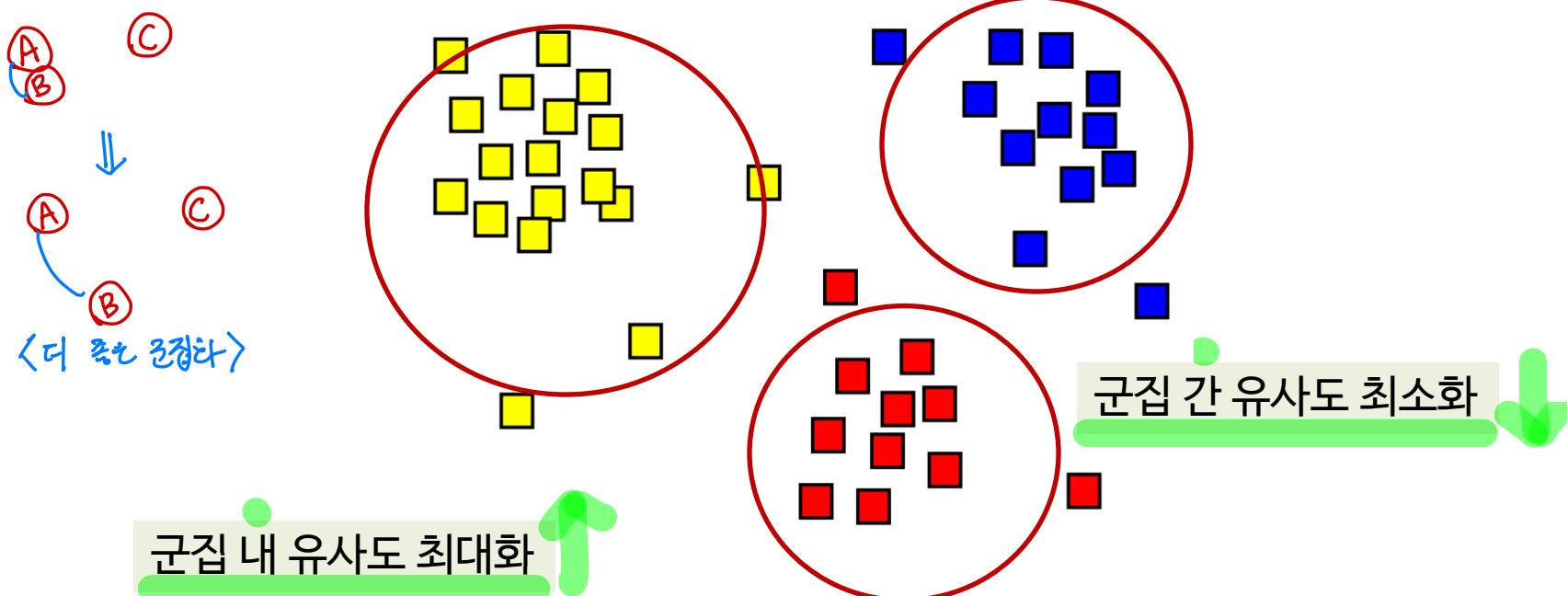
처음에는 몰랐다가 점점 유사한 속성끼리 묶어서 ⇒ 군집화!

군집화 개념

▪ 좋은 군집화(Clustering)의 기준

- 동일한 군집에 소속된 데이터는 서로 유사할 수록 좋음 (inter-class similarity)
- 상이한 군집에 소속된 데이터는 서로 다를 수록 좋음 (intra-class dissimilarity)

유사도 ↓



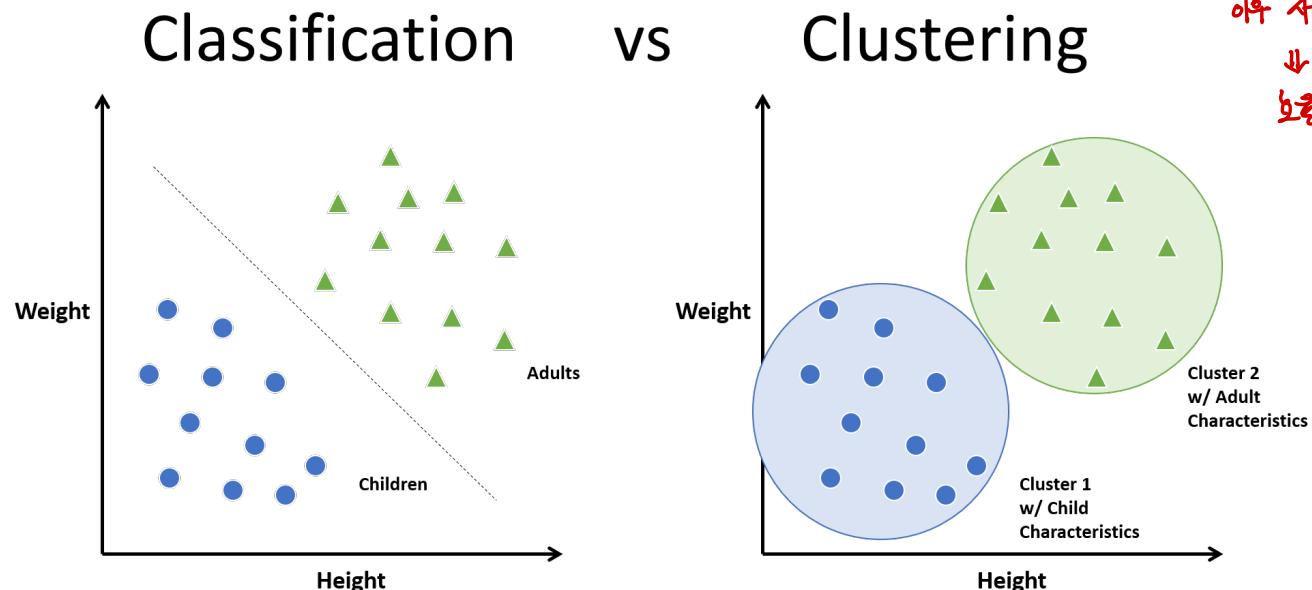
군집화 개념

- 분류(classification) vs 군집화(clustering)
▪ 분류 : (사전 정의된 범주가 있는 데이터)로부터 예측 모델을 학습하는 문제
 지도 학습 (Supervised learning)
▪ 군집화 : (사전 정의된 범주가 없는 데이터)로부터 최적의 그룹을 찾아가는 문제
 비지도 학습 (Unsupervised learning)

$$f(x) = Y$$

↳ 군집화는 사용해 먼저 1차 필터링

↓
이후 서왕이 직접 분류
↓
분류



군집화 적용 사례

ex)

유사 문서 군집화

- 스포츠, 경제, 기술, 엔터 등 (문서 내 단어를 이용한 유사 문서 군집화)

유사 영상 군집화

- 사람, 자동차, 오토바이 등 (영상 내 영상 특징을 이용한 유사 영상 군집화)

유사 고객 군집화

- 신한 카드는 고객정보와 카드 결제내역을 통해 남녀 고객을 총 18가지 그룹으로 분류하고 고객 타입마다 다른 카드 혜택을 제공하는 카드를 선보임



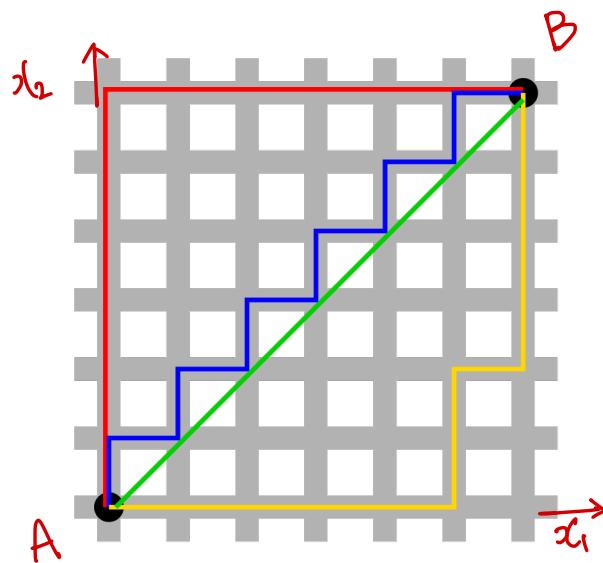
군집화 수행 시 주요 고려사항

~ $|x_1 - x_2|$ ~ 거리 어떤 종류로 구한 것인가?

- 어떤 거리 측도를 사용하여 유사도(similarity metric)를 측정 할 것인가?
- 어떤 군집화 알고리즘을 사용할 것인가? K-means, DBSCAN
- 어떻게 최적의 군집 수(K)를 결정할 것인가? (K) (ϵ, M)
- 어떻게 군집화 결과를 측정/평가할 것인가?

군집화: 유사도 척도

- 어떤 거리 측도를 사용하여 유사도(similarity metric)를 측정 할 것인가?
 - 유클리디언 거리 (Euclidean Distance)
▪ 일반적으로 사용하는 거리 척도, L_2 거리
 - 두 관측치 사이의 직선 거리를 의미함
- 맨하탄 거리 (Manhattan Distance)
▪ 택시 거리, L_1 거리



$$d(x, y) = \sqrt{\sum_{i=1}^p (x_i - y_i)^2}$$

$$d_{\text{Manhattan}(X,Y)} = \sum_{i=1}^p |x_i - y_i|$$

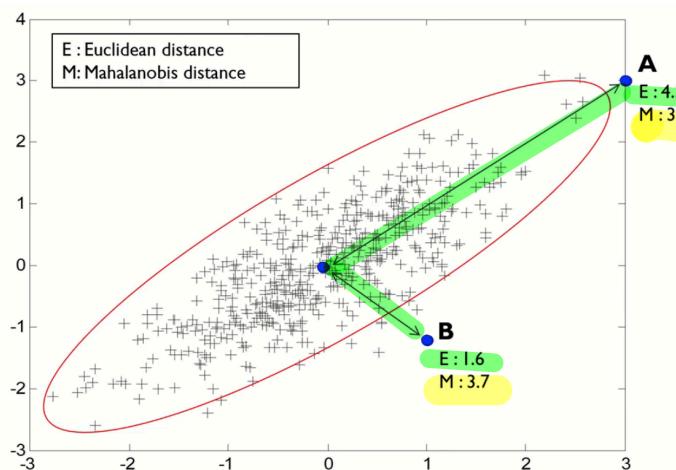
군집화: 유사도 척도

- 마할라노비스 거리 (Mahalanobis Distance)

$$d_{Mahalanobis}(X, Y) = \sqrt{(X - Y) \Sigma^{-1} (X - Y)}$$

where Σ^{-1} = inverse of covariance matrix

- 변수 내 분산, 변수 내 공분산을 모두 반영하여 X, Y 간 거리를 계산하는 방식
- 데이터의 Covariance Matrix 가 identity matrix 인 경우는 유clidean 거리와 동일



군집화: 알고리즘

▪ 군집화 알고리즘의 종류

▪ 계층적 군집화

- 개체들을 가까운 집단부터 차근차근 묶어 나가는 방식
- 군집화 결과 뿐만 아니라 유사한 개체들이 결합되는 덴드로그램 생성

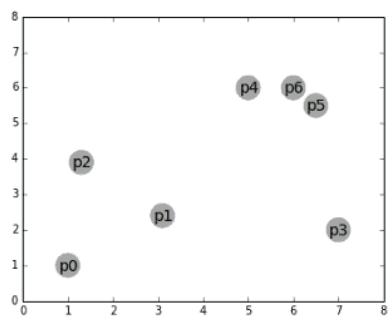
▪ 분리형 군집화 → Kmeans 대표적

↳ 어떻게 굽힐지 되는지 시작할 수 있음

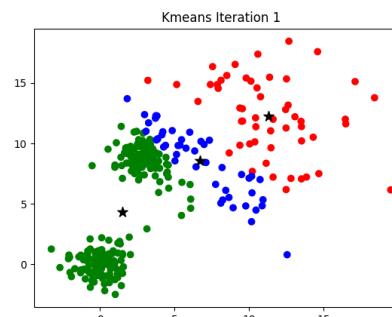
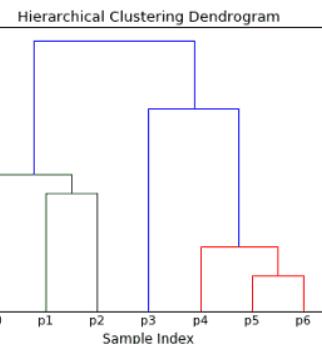
- 전체 데이터 영역을 특정 기준에 의해 동시에 구분
- 각 객체들은 사전에 정의된 개수의 군집 중 하나에 속하게 됨 $k=3$

▪ 분포 기반 군집화 → DBSCAN 대표적

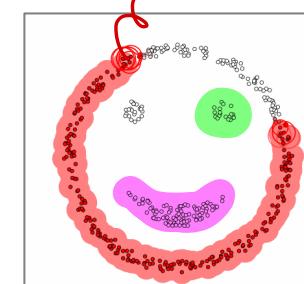
- 데이터의 분포를 기반으로 높은 밀도를 갖는 세부 영역들로 전체 영역을 구분



계층적 군집화



분리형 군집화



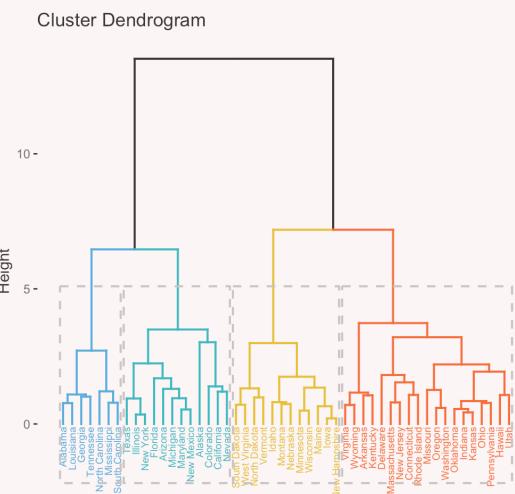
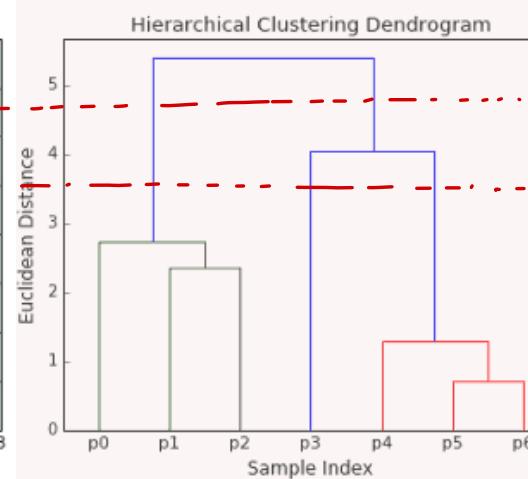
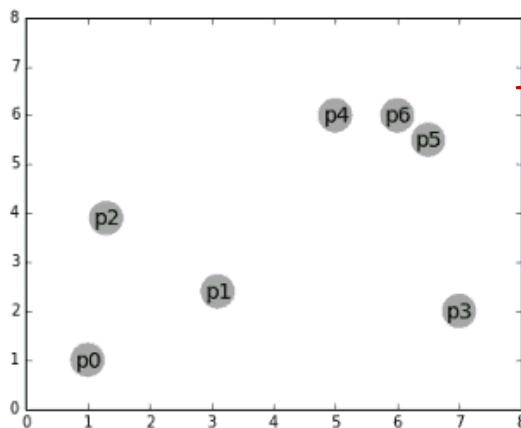
분포 기반 군집화

군집화: 알고리즘 (1)

▪ 계층적 군집화 (Hierarchical Clustering)

- 계층적 트리 모델을 이용하여 개별 개체들을 순차적/계층적으로 유사한 개체/군집과 통합
- 덴드로그램 (Dendrogram)을 통해 시각화 가능
 - 덴드로그램 : 개체들이 결합되는 순서를 나타내는 트리 형태의 구조
- 사전에 군집의 수를 정하지 않아도 수행 가능
 - 덴드로그램 생성 후 적절한 수준에서 자르면 그에 해당하는 군집화 결과 생성

▷ 덴드로그램을 보고 결정

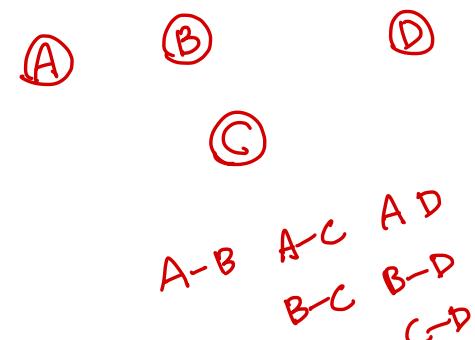


덴드로그램 (Dendrogram)

군집화: 알고리즘 (1)

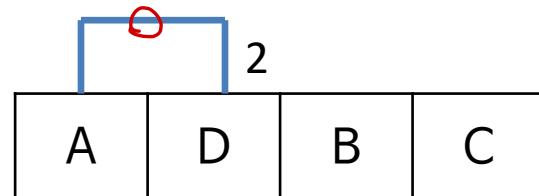
▪ 계층적 군집화 예시

- (1) 모든 데이터 사이의 거리에 대한 유사도 행렬 계산



	A	B	C	D
A		20	7	2
B			10	25
C				3
D				

거리 \downarrow 이 가장
유사도 \uparrow



군집화: 알고리즘 (1)

A D C B

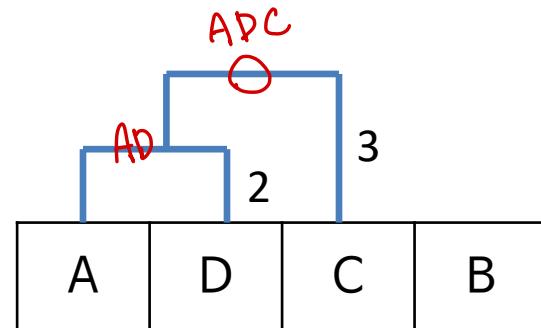
▪ 계층적 군집화 예시

- (1) 모든 데이터 사이의 거리에 대한 유사도 행렬 계산
- (2) 거리가 인접한 데이터끼리 군집 형성
- (3) 유사도 행렬 갱신

앞의 과정에서 A-D 거리 가장 가까웠으므로 뭉여짐



	AD	B	C	
AD		20	③	
B			10	
C				

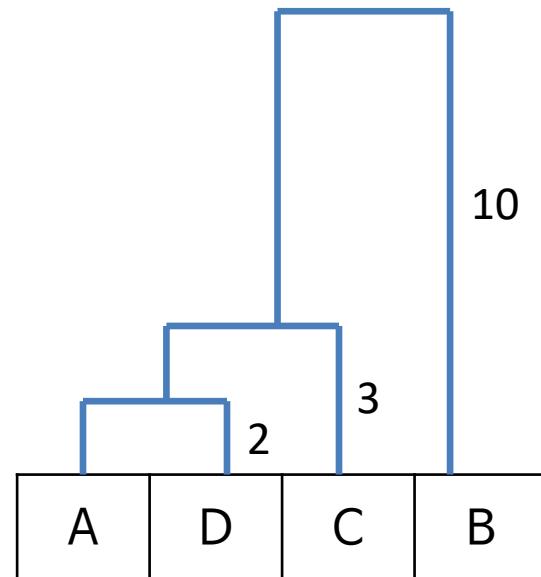


군집화: 알고리즘 (1)

▪ 계층적 군집화 예시

- (1) 모든 데이터 사이의 거리에 대한 유사도 행렬 계산
- (2) 거리가 인접한 데이터끼리 군집 형성
- (3) 유사도 행렬 갱신
- (1-3) 반복

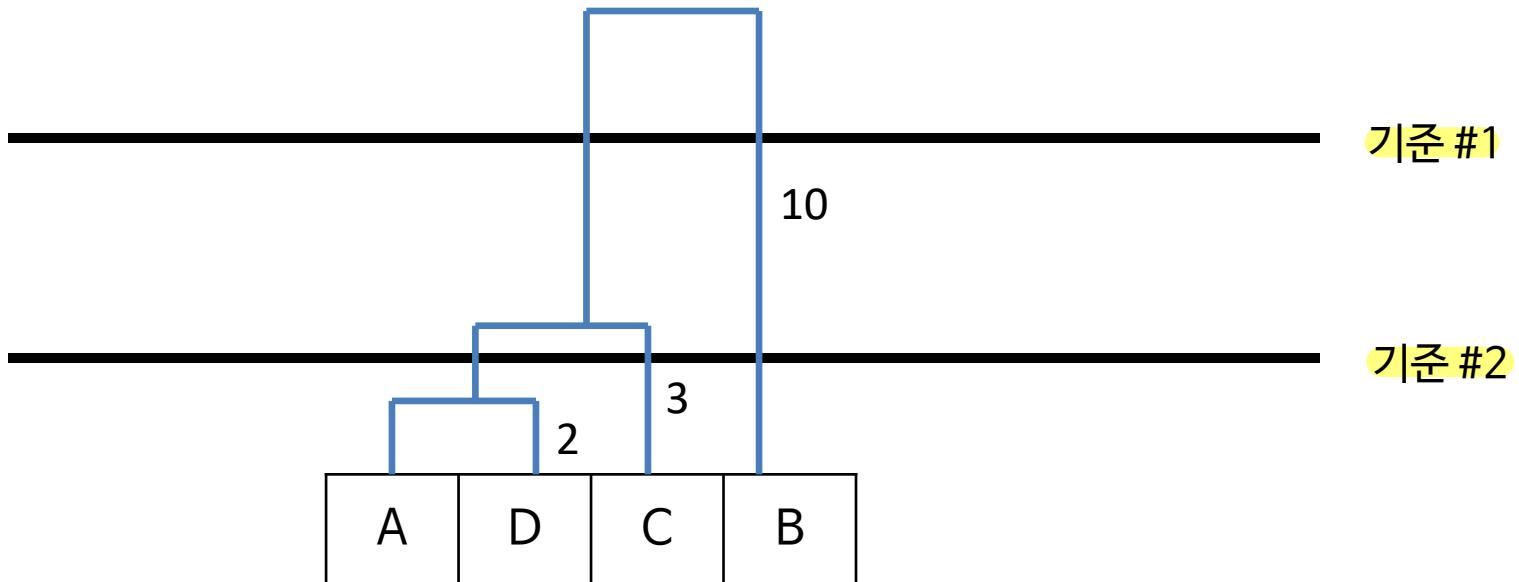
	ADC	B		
ADC		10		
B				



군집화: 알고리즘 (1)

- 계층적 군집화 예시
 - 최종결과

장점: 미리 군집의 개수를
지정할 필요 X

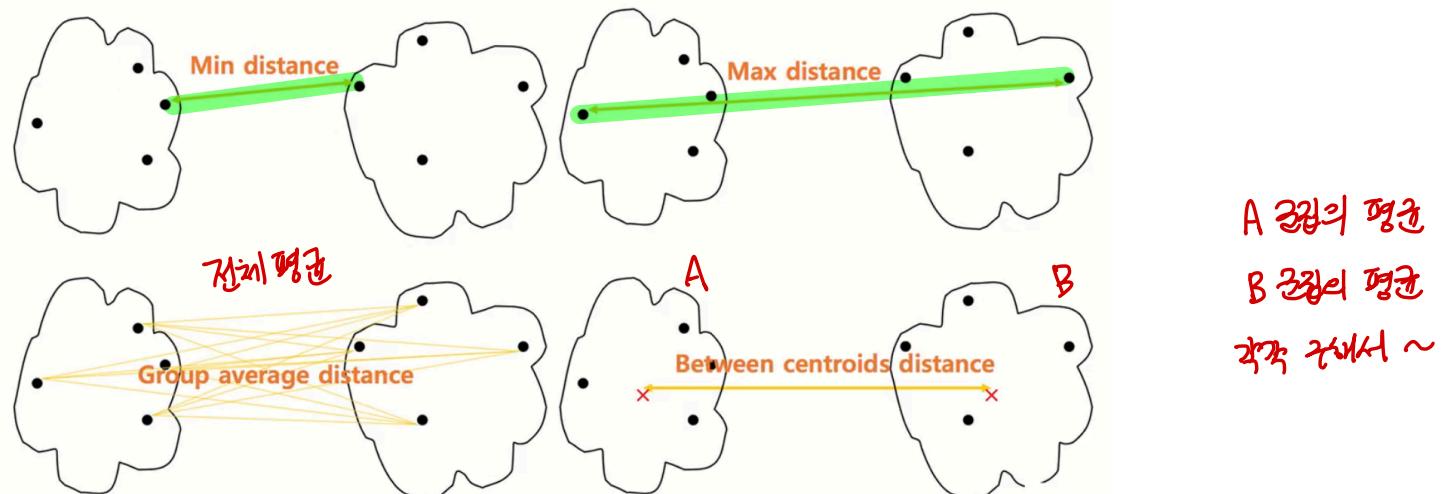


군집화: 알고리즘 (1)

▪ 계층적 군집화 예시

▪ 군집과 군집의 거리 계산 법

- 군집 내 데이터 간 거리를 모두 계산 하여 가장 작(min)은 거리 값을 선택
- 군집 내 데이터 간 거리를 모두 계산 하여 가장 큰(max) 거리 값을 선택
- 군집 내 데이터 간 거리를 모두 계산 하여 평균(avg) 거리 값을 선택
- 각 군집 내 데이터의 평균을 계산하여 평균과 평균 사이의 거리 값을 선택
- Ward 거리 계산법



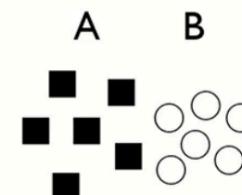
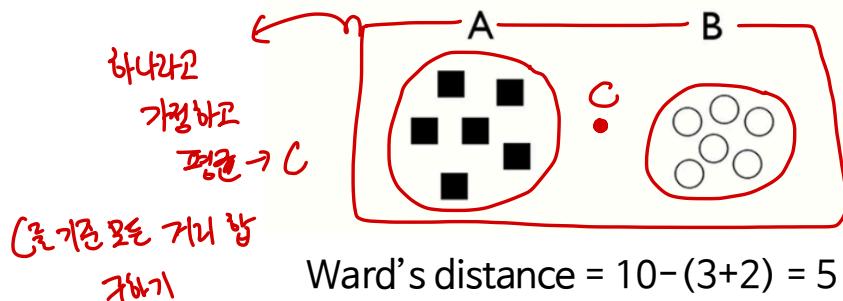
군집화: 알고리즘 (1)

▪ 계층적 군집화 예시

▪ Ward 거리 계산법 (ward linkage method)

- ① 서로 다른 군집에 해당하는 모든 데이터를 포함한 중심(Centroid)을 구하고, 구한 중심과 서로 다른 군집에 포함되는 모든 데이터 사이의 거리를 구한다.
- ② 각 군집에 해당하는 데이터를 통해 중심을 구하고, 구한 중심과 군집 내 데이터 사이의 거리를 구한다.
- ③ 최종 결과가 클 수록 서로 다른 군집은 유사도가 낮아 멀리 있고, 최종 결과가 작을 수록 서로 다른 군집의 유사도는 높아 가까이에 있다.

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



$$\text{Ward's distance} = 7 - (3+2) = 2$$

군집화: 알고리즘 (2)

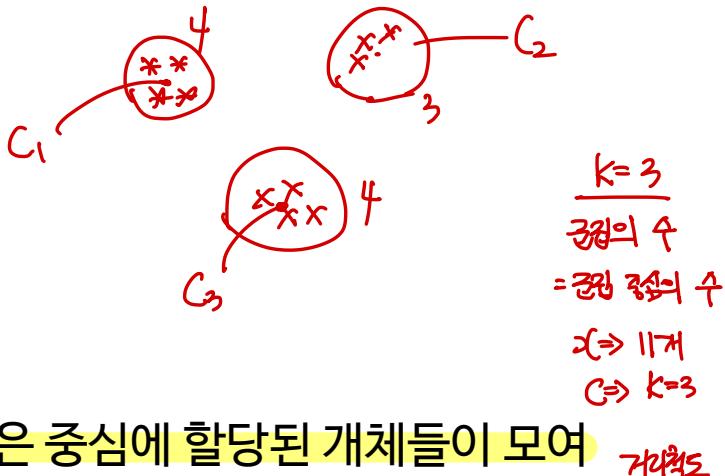
분리형 군집화 예시

K 평균 군집화 (K-means Clustering)

- 각 군집은 하나의 중심(centroid)을 가짐
- 각 객체는 가장 가까운 중심에 할당되며, 같은 중심에 할당된 개체들이 모여 하나의 군집을 형성
- 사전에 군집의 수 K가 정해져야 알고리즘을 수행할 수 있음

X

K ⇒ 군집수



$$\operatorname{argmax}_c \sum_{i=1}^K \sum_{x_j \in C_i} \|x_j - c_i\|^2$$

$$X = C_1 \cup C_2 \dots \cup C_k, C_i \cap C_j = \emptyset, i \neq j$$

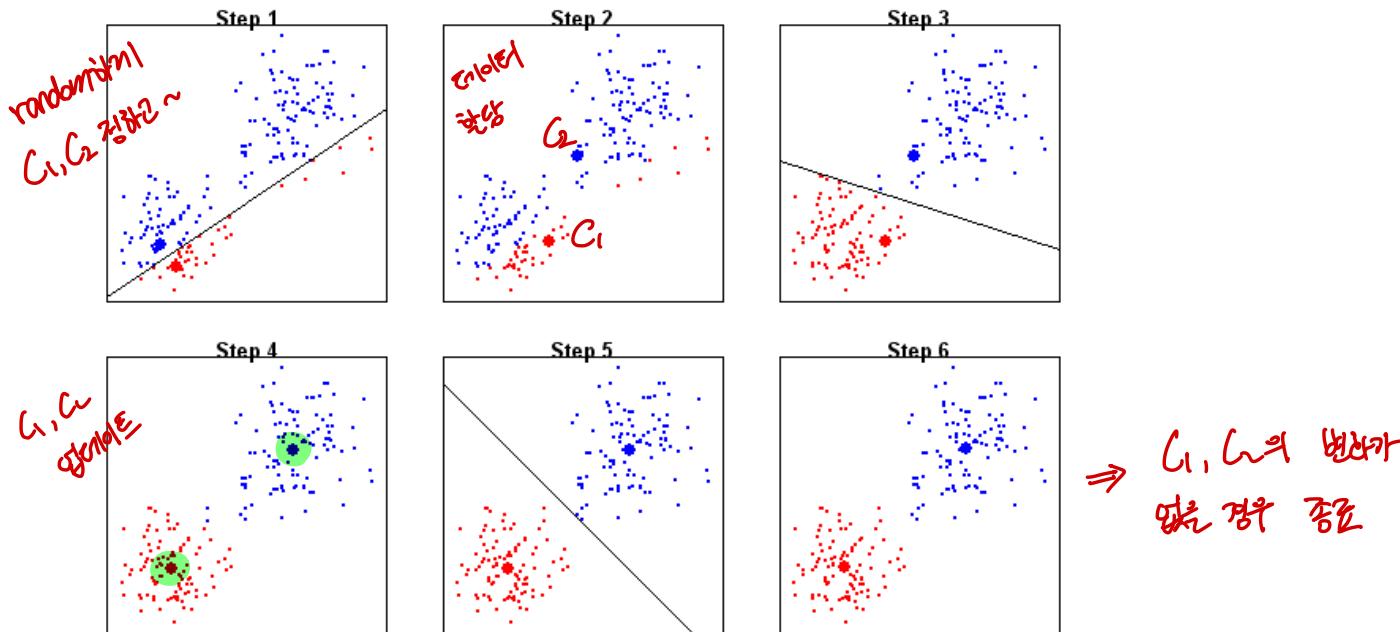
X : 데이터, C : 군집, c_i : 군집 당 중심

군집화: 알고리즘 (2)

▪ 분리형 군집화 예시

▪ K 평균 군집화 예시 ($K=2$) *random*

- (1) 두개의 중심을 임의로 생성
- (2) 생성된 중심을 기준으로 모든 데이터에 군집 할당 (가장 가까운 군집에 할당)
- (3) 각 군집의 중심 다시 계산 *
- (4) 중심이 변하지 않을 때까지 (2-3)번 반복



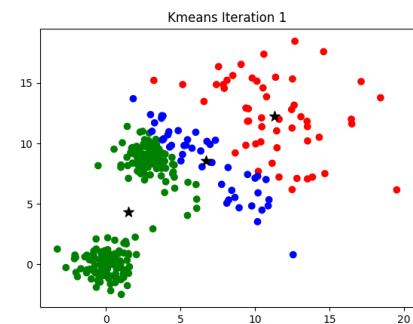
군집화: 알고리즘 (2)

▪ 분리형 군집화 예시

▪ K 평균 군집화 예시

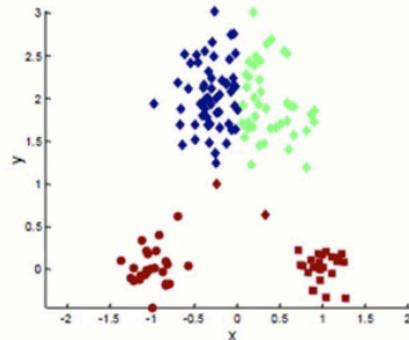
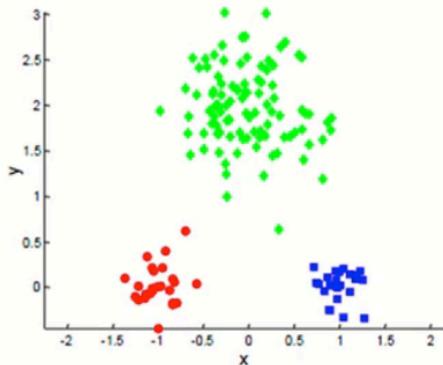
- (1) K개의 중심을 임의로 생성
- (2) 생성된 중심을 기준으로 모든 데이터에 군집 할당
- (3) 각 군집의 중심 다시 계산
- (4) 중심이 변하지 않을 때까지 (2-3)번 반복

c_1, c_2, \dots, c_k



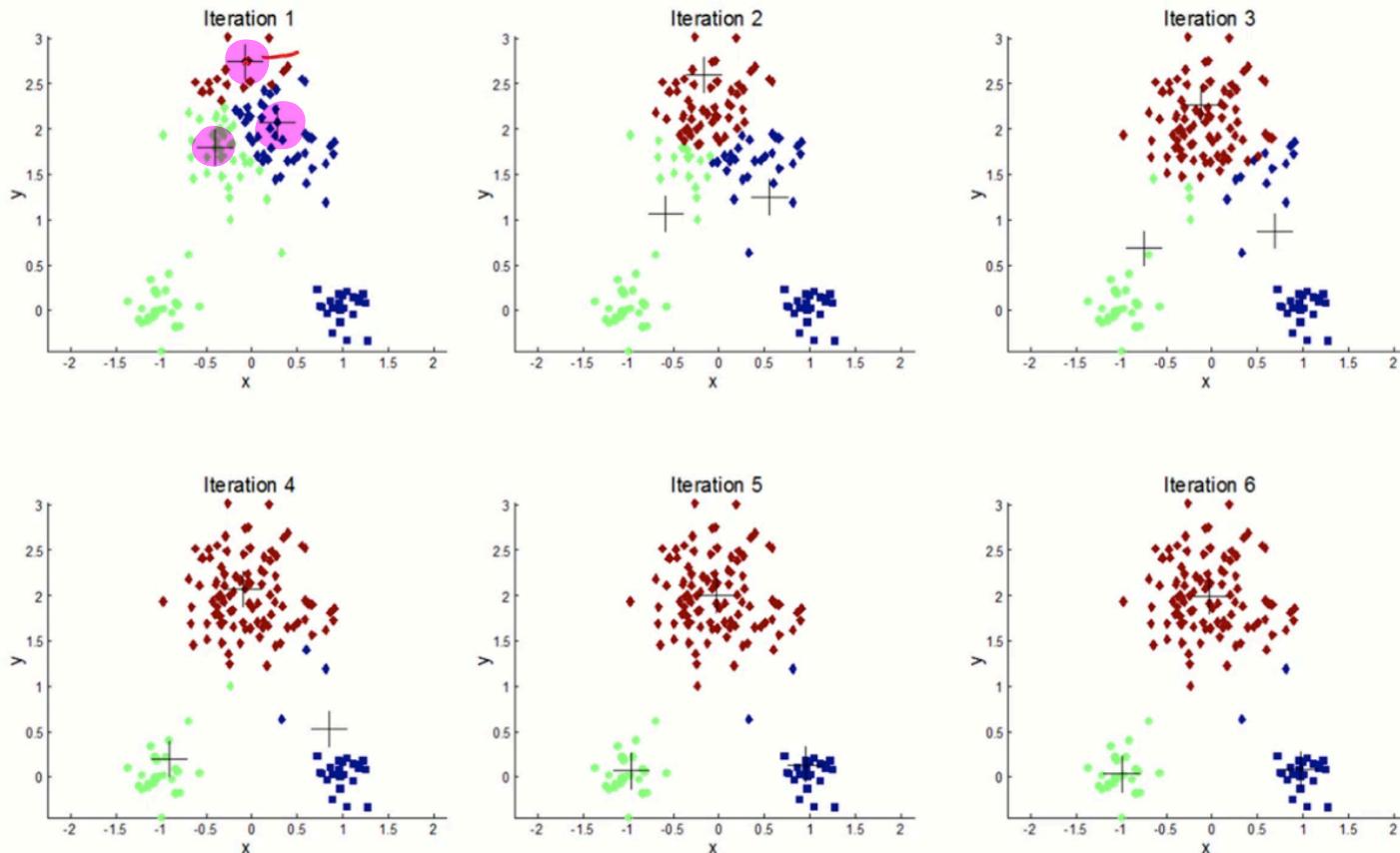
▪ 초기 중심 설정은 최종 군집화 결과에 영향을 미칠 수 있음

- (좌) 좋은 결과 (우) 안좋은 결과



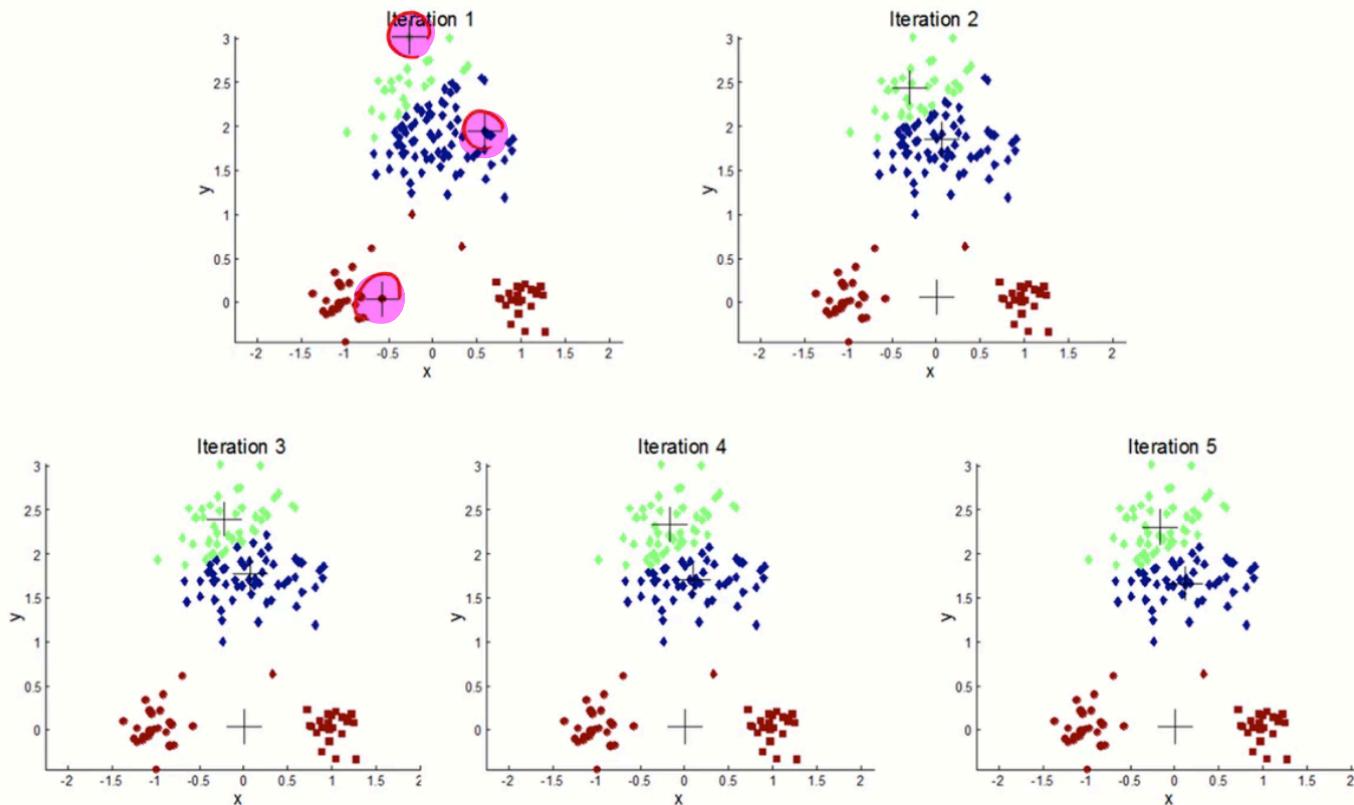
군집화: 알고리즘 (2)

- 분리형 군집화 예시
 - K 평균 군집화 초기화의 바른 예시

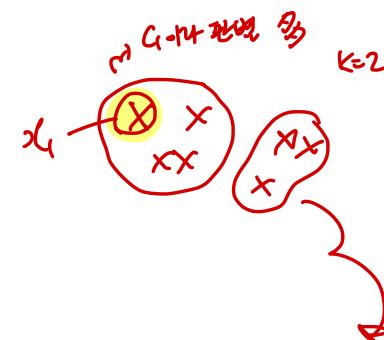


군집화: 알고리즘 (2)

- 분리형 군집화 예시
 - K 평균 군집화 초기화의 잘못된 예시



군집화: 알고리즘 (2)



▪ 분리형 군집화 예시

▪ K 평균 군집화의 랜덤 초기화의 단점 극복 방법

- 여러 번 Kmeans 군집화를 수행하여 가장 여러 번 나타나는 군집을 사용

▪ ↗ 양상블 (ensemble) 결과 통합 //

▪ 데이터 분포 정보를 활용한 초기화 선정

▪ 데이터가 Gaussian 분포라면, 중심을 초기 값으로 선정

▪ (데이터가 많을 때) 샘플링 데이터를 활용하여 계층적 군집화를 수행한 뒤 초기 군집 중심으로 사용 $\rightarrow N=1000$

▪ But, 많은 경우 초기 중심이 최종 결과에 영향을 미치지 않음 //

\therefore 초기 중심에 대한 고민 끝이 X

거의 사용 X
데이터의 분포는 보다는
가장 많이 주변
하는 것에 ~
☆ 주제 A/B

1 번째 Kmeans

$x_1 = 1$

2 번째 Kmeans

$x_1 = 2$

3 번째 Kmeans

$x_1 = 1$

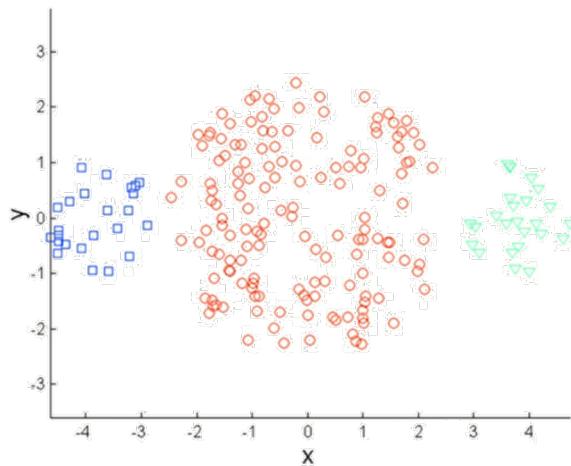
군집화: 알고리즘 (2)

- 분리형 군집화 예시

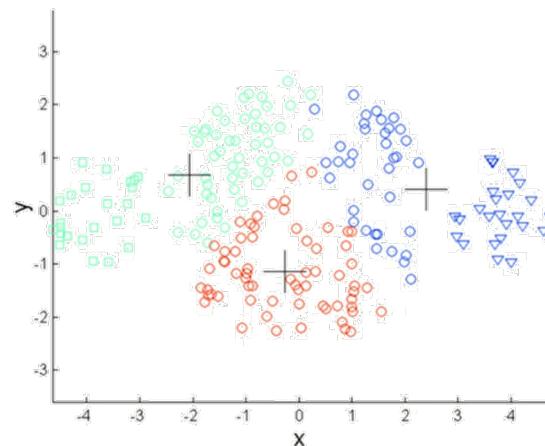
- K 평균 군집화의 단점

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

정답



K-평균 군집화 결과

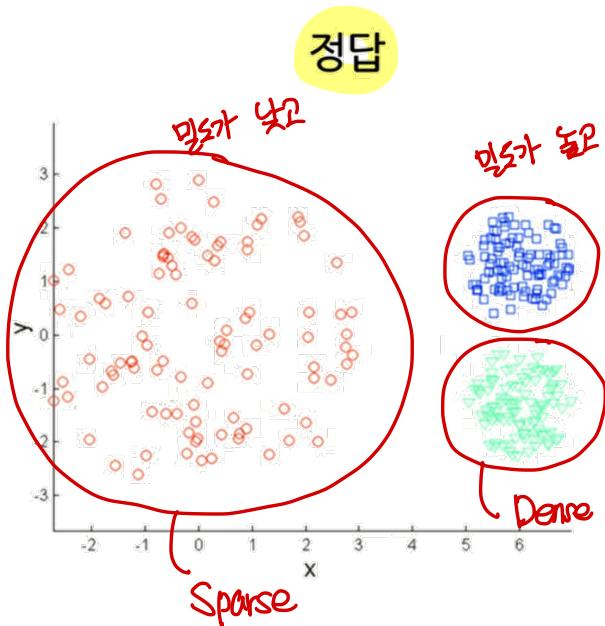


군집화: 알고리즘 (2)

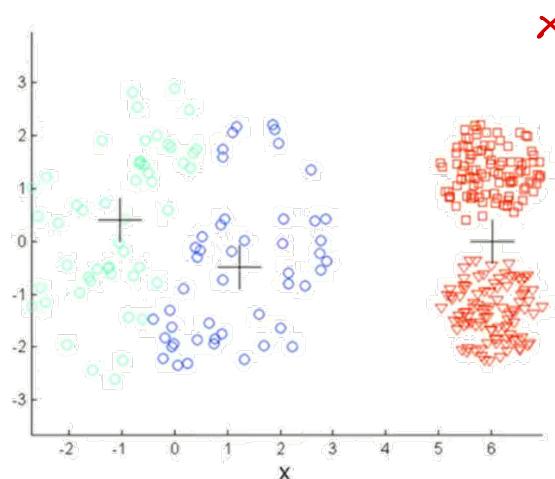
- 분리형 군집화 예시

- K 평균 군집화의 단점

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



K-평균 군집화 결과

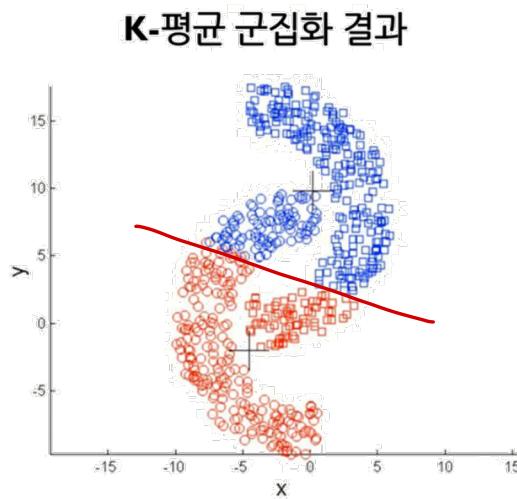
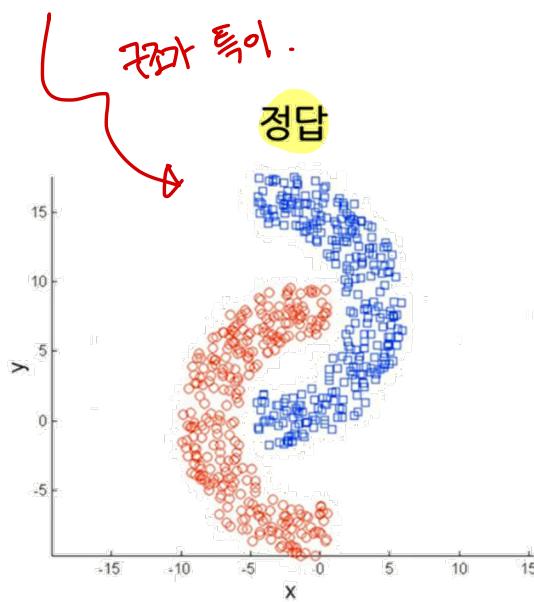


군집화: 알고리즘 (2)

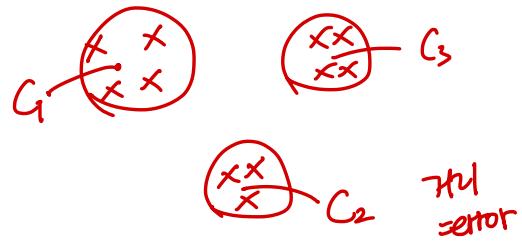
- 분리형 군집화 예시

- K 평균 군집화의 단점

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



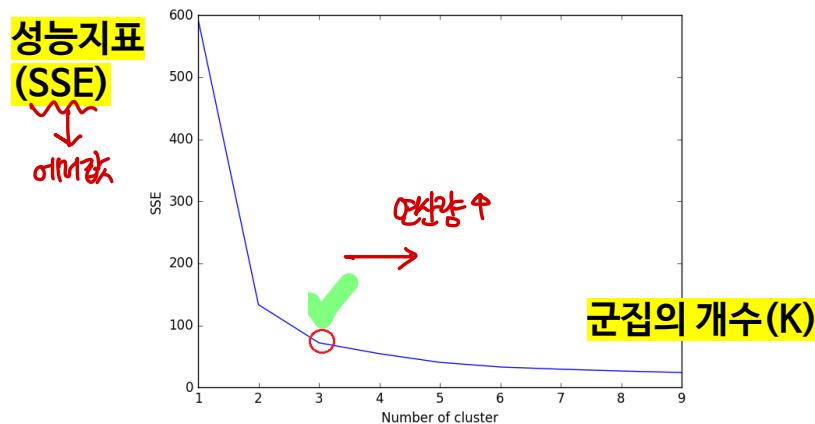
군집화: 알고리즘 (2)



- 분리형 군집화 예시

- K 평균 군집화의 K 값 선정 방법

- 다양한 군집 수에 대한 성능 평가 지표를 통해 최적의 군집 수(K) 선택
 - 일반적으로 Elbow point에서 최적의 군집 수 결정



- 군집화 평가 방법

- 지도학습기반 분류 문제처럼 모든 상황에 적용 가능한 평가 지표 부재
 - 내부 평가 지표: Dunn Index, Sum of Squared Error (SSE)
 - 외부 평가 지표: Rand Index, Jaccard Coefficient 등

군집화: 알고리즘 (2)

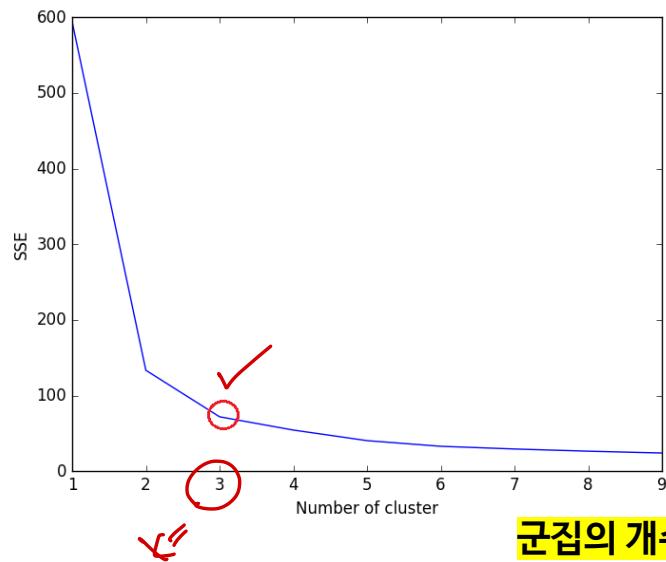
- 분리형 군집화 예시

- 군집화 평가 지표: Sum of Squared Error (SSE)**

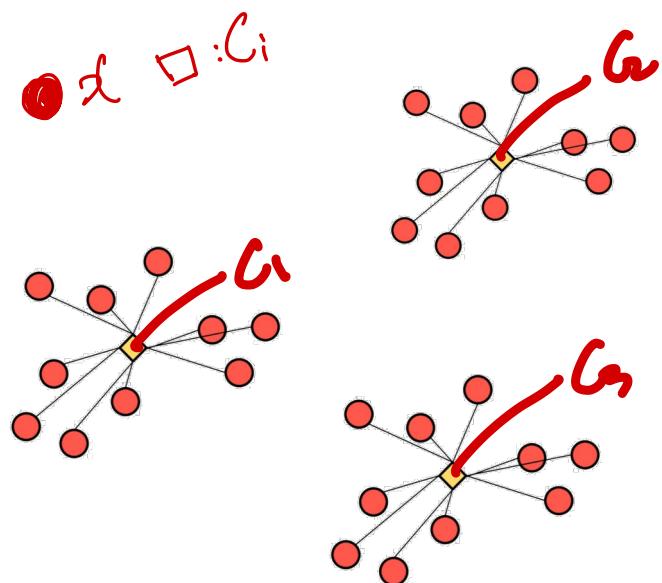
- 군집 내 거리 최소화 (만족), 군집 간 거리 최대화 (불만족)
- (식)

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

성능(에러)



군집의 개수(K)



군집화: 알고리즘 (2)

▪ 분리형 군집화 예시

▪ 군집화 평가 지표: silhouette 통계량

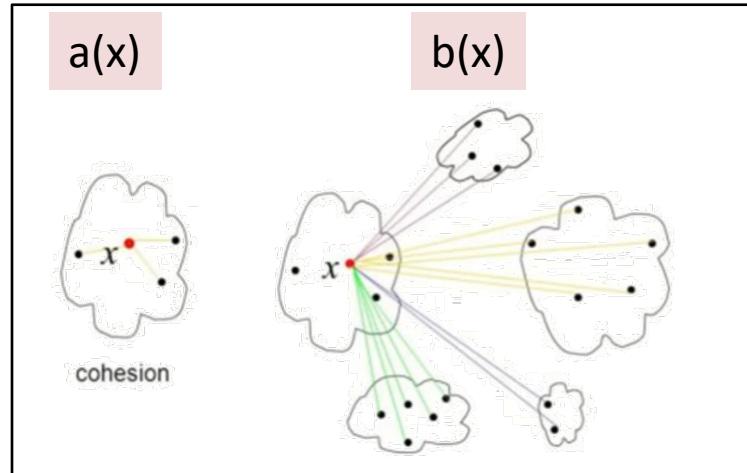
- a(i): i번째 데이터와 같은 군집 내에 있는 모든 데이터 사이의 평균 거리
 - 작을 수록 유사한 데이터가 잘 모여 있다는 의미 ok.
- b(i): i번째 데이터와 다른 군집 내에 있는 모든 데이터 사이의 최소 거리
 - 클수록 서로 다른 데이터가 잘 흩어져 있다는 의미 ok
- 일반적으로 S 값이 0.5보다 크면 군집 결과가 타당하다고 판단
 - S값이 1에 가까울 수록 군집화 Good, -1에 가까울 수록 군집화 Bad
 - K=2 인 경우에 통계량이 Best인 경우가 많아서 차 순위 K를 선정하는 것이 일반적임

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

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

안좋다 \leftarrow 군집화의 결과 \rightarrow 좋다

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



* 주의사항 : k=2일 때 → 항상 silhouette 통계량 ↑

∴ 우리는 Best score 말고

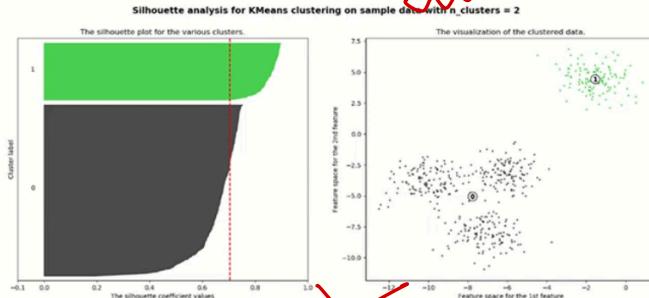
☞ 교수님께서는 이 방법 SSE 뿐 아니라 조금 더 낫다고 하셨다.
[이 가까운 k 설정!!]

군집화: 알고리즘 (2)

- 분리형 군집화 예시
 - 군집화 평가 지표: Silhouette 통계량

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

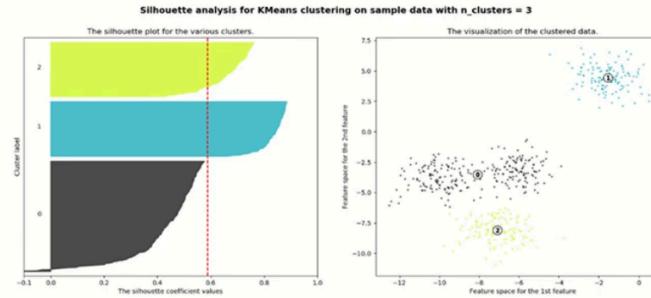
k=2



$$\bar{s} = 0.705$$

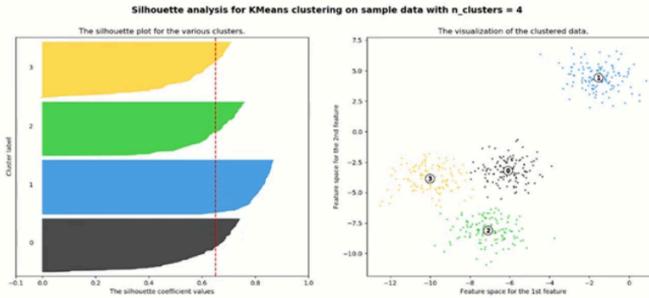
k=4

k=3

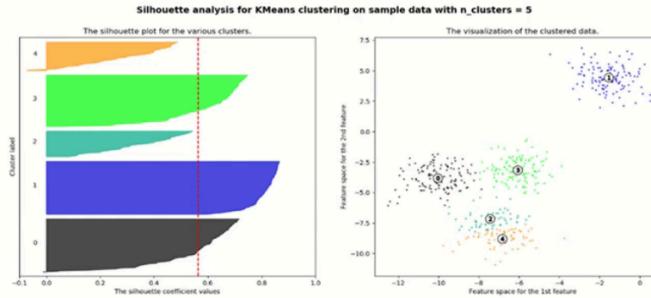


$$\bar{s} = 0.588$$

k=5



$$\bar{s} = 0.650$$



$$\bar{s} = 0.564$$

군집화: 알고리즘 (3)

▪ 분포 기반 군집화

▪ DBSCAN (Density Based Clustering) $\rightsquigarrow k$ 라는 파라미터 정한 필요 X But, 다른 파라미터

- 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
 - ϵ 과 M은 분포 기반 군집화의 파라미터

▪ 핵심자료 (core point)

- ϵ -neighborhood가 M개 이상의 데이터를 포함하는 자료

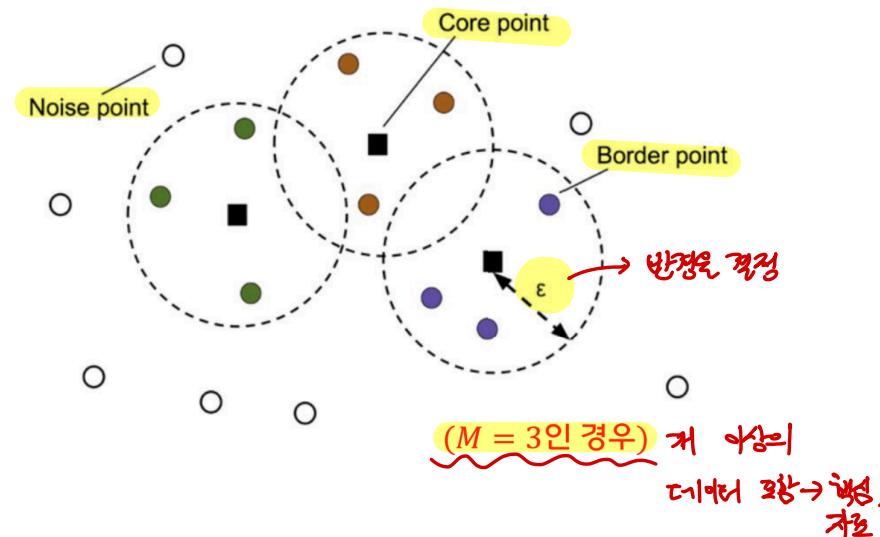
▪ 주변자료 (border point)

- 핵심자료는 아니지만 ϵ -neighborhood에 핵심자료를 포함하는 자료

▪ 잡음자료 (noise point)

- 핵심자료도 주변자료도 아닌 자료

DBSCAN은 분포 기반 \therefore 흩어져 있으면
↳ 밀도 \downarrow '잡음' 처리

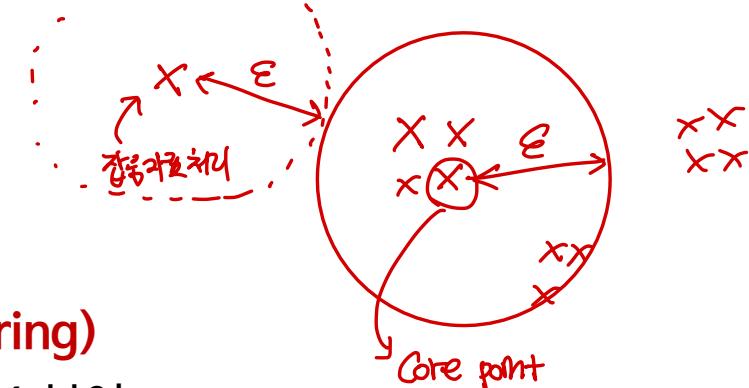


군집화: 알고리즘 (3)

▪ 분포 기반 군집화

▪ DBSCAN (Density Based Clustering)

- (1) 임의 데이터 선택하고 군집 1 부여
- (2) 임의 데이터의 ϵ -NN 을 구하고 데이터의 수가 M보다 작으면 잡음자료 부여
~~반복~~
- (3) M 보다 크면 ϵ -NN 모두 군집1 부여, 군집 1 모든 데이터의 ϵ -NN 의 크기가 M보다 큰 것이 없을 때까지 반복
- (4) 군집 2에 대해 동일하게 반복
즉, 모든 데이터에 군집이 할당되거나 잡음으로 분류될 때까지 절차 (1-3) 반복

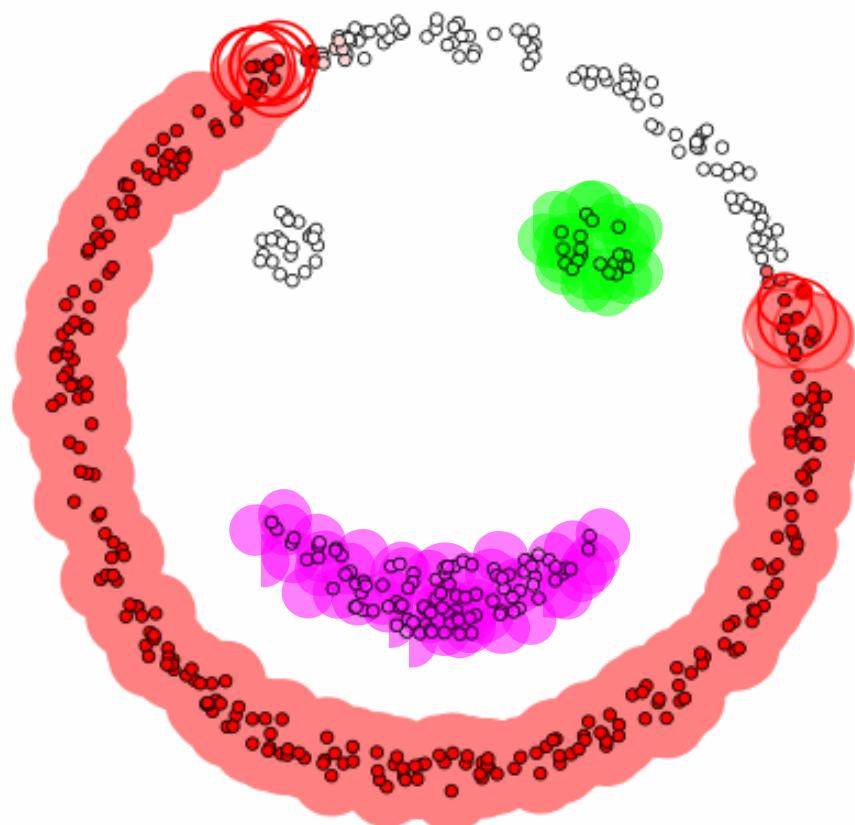


▪ 파라미터 설정

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

군집화: 알고리즘 (3)

- 분포 기반 군집화
 - DBSCAN (Density Based Clustering) 예시



군집화: 알고리즘 (3)

- 분포 기반 군집화

- DBSCAN (Density Based Clustering) 예시**

Kmeans
DBScan ↗ 기여 ::

ϵ : 너무 작으면 많은 데이터가 잡음으로 분류되고 너무 크면 군집의 개수가 적음

