

교 육 세 미 나

ToBig's 9기 박송은

클러스터링 Clustering

Contents

Unit 01 | 군집화(Clustering)

Unit 02 | 계층적 군집화

Unit 03 | K-Means

Unit 04 | 모델 평가

Unit 05 | DBSCAN

Unit 01 | 군집화(Clustering)

Machine Learning Algorithms *(sample)*

	<u>Unsupervised</u>	<u>Supervised</u>
<u>Continuous</u>	<ul style="list-style-type: none"> Clustering & Dimensionality Reduction <ul style="list-style-type: none"> SVD PCA K-means 	<ul style="list-style-type: none"> Regression <ul style="list-style-type: none"> Linear Polynomial Decision Trees Random Forests
<u>Categorical</u>	<ul style="list-style-type: none"> Association Analysis <ul style="list-style-type: none"> Apriori FP-Growth Hidden Markov Model 	<ul style="list-style-type: none"> Classification <ul style="list-style-type: none"> KNN Trees Logistic Regression Naive-Bayes SVM

Unit 01 | 군집화(Clustering)

지도학습(Supervised Learning)

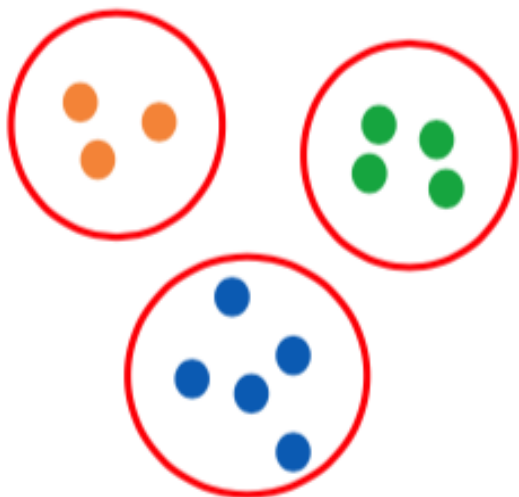
- 분류: 소속집단의 정보를 이미 알고 있는 상태에서, 비슷한 집단으로 묶는 방법
즉, **Label이 있는 Data를 나누는 방법**으로, Supervised Learning의 일종

비지도학습(Unsupervised Learning)

- 군집화: 소속집단의 정보가 없고, 모르는 상태에서 비슷한 집단으로 묶는 방법
즉, **Label이 없는 Data를 군집단위로 나누는 것**으로, Unsupervised Learning 의 일종

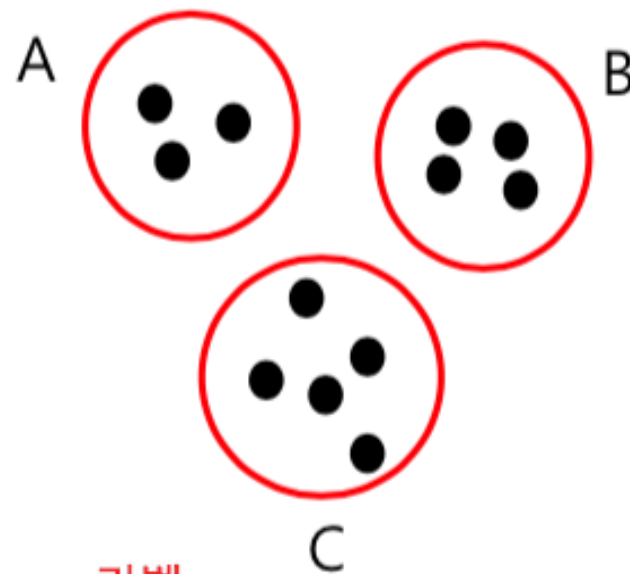
Unit 01 | 군집화(Clustering)

Classification



라벨0

Clustering



라벨x

Unit 01 | 군집화(Clustering)

군집 분석

- 각 개체의 유사성을 측정하여 높은 대상 집단을 분류하고, 군집에 속한 개체들의 유사성과 서로 다른 군집에 속한 개체 간의 상이성을 규명하는 통계분석방법
- 주어진 데이터셋 내에 존재하는 몇 개의 군집을 찾아내는 비지도 기법
- 생물학, 행동과학, 마케팅 및 의학분야에서 다양하게 사용됨

Unit 01 | 군집화(Clustering)

군집분석 방법

1. Hierarchical agglomerative clustering(계층적 군집화):

모든 데이터가 하나의 군집으로 병합될 때까지 군집들을 자연적인 계층 구조로 정렬하는 것
ex) single linkage, complete linkage, average linkage, centroid, Ward의 방법 등

2. Partitioning clustering(비계층적 군집화):

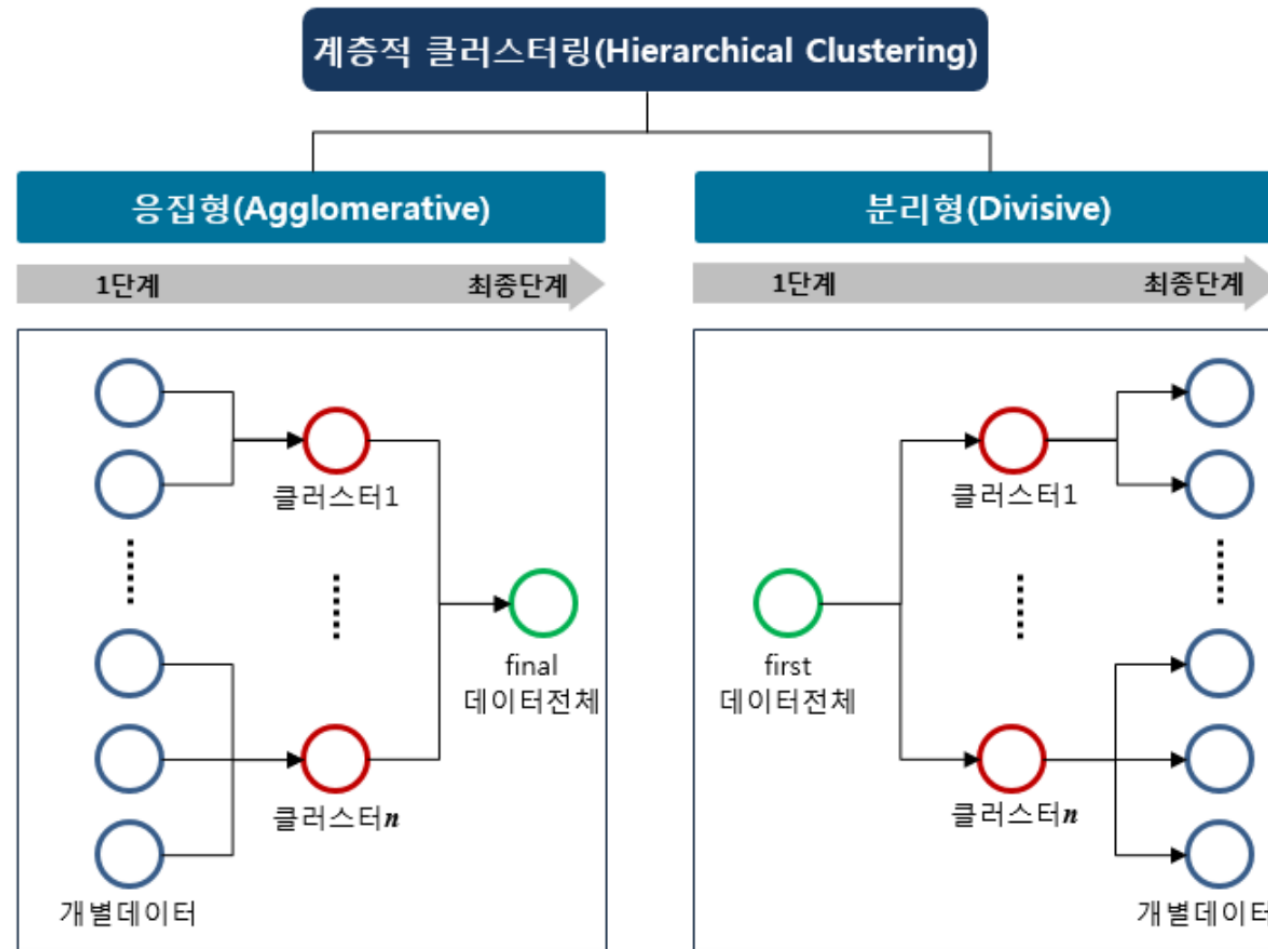
구하고자 하는 군집의 수를 정한 상태에서 설정된 군집의 중심에 가장 가까운 개체를 하나씩 포함해 가는 방식
ex) k-means, PAM(partitioning around medoids)

Unit 01 | 군집화(Clustering)

군집분석 단계

1. 알맞은 속성 선택 - 데이터를 군집화하는데 중요하다고 판단되는 속성들을 선택
2. 데이터 표준화 - 분석에 사용되는 변수들의 범위에 차이가 있는 경우 가장 큰 범위를 갖는 변수가 결과에 가장 큰 영향을 미치게 됨
3. 이상치 선별 - 많은 군집분석 방법은 이상치에 민감하기 때문에 군집 분석 결과가 왜곡됨
4. 군집 알고리즘 선택
5. 군집의 개수 결정

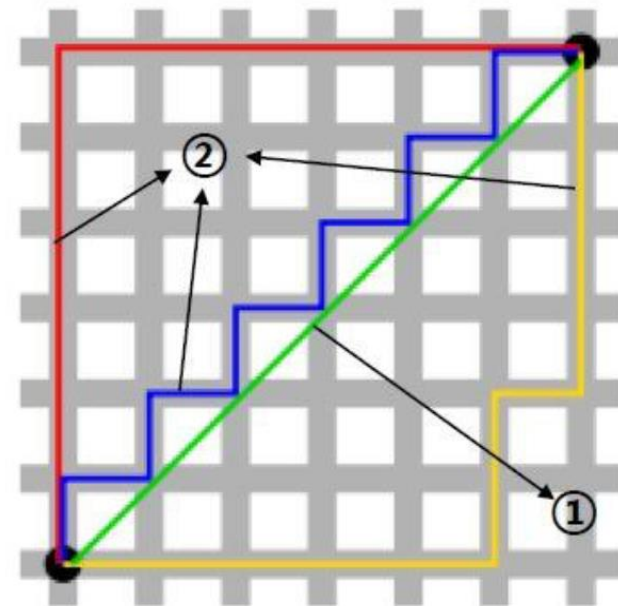
Unit 02 | 계층적 군집화(Hierarchical agglomerative clustering)



Unit 02 | 계층적 군집화(Hierarchical agglomerative clustering)

주로 사용되는 거리의 정의

- ① 유클리드 거리 (Euclidean) : $d(x,y) = (\sum_{i=1}^p (x_i - y_i)^2)^{1/2}$
- ② 맨하탄 거리 (Manhattan) : $d(x,y) = \sum_{i=1}^p |x_i - y_i|$
- ③ 표준화 거리 (Standardized) : $d(x,y) = (\sum_{i=1}^p (x_i - y_i)^2 / s_i^2)^{1/2}$
- ④ 민콥스키 거리 (Minkowski) : $d(x,y) = (\sum_{i=1}^p (x_i - y_i)^m)^{1/m}$



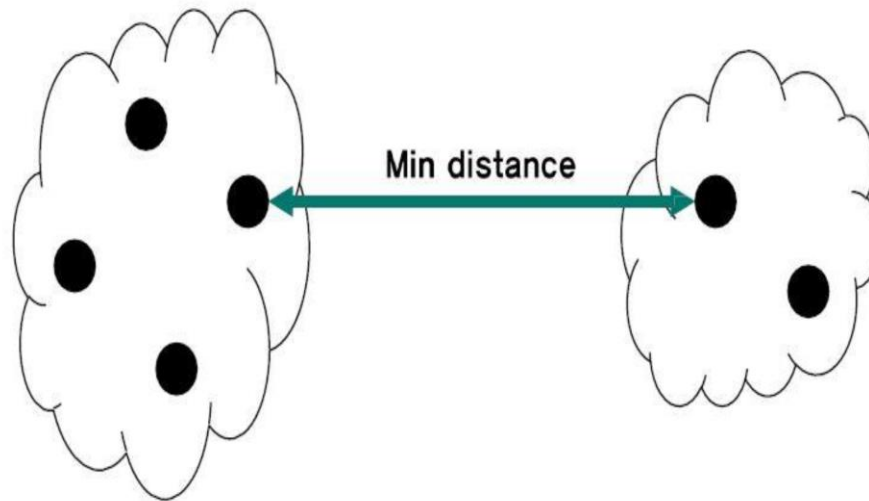
Unit 02 | 계층적 군집화(Hierarchical agglomerative clustering)

군집-군집 or 군집-개체 간 거리 측정 방법

1. 최단 연결법
2. 최장 연결법
3. 평균 연결법
4. 중심 연결법

Unit 02 | 계층적 군집화(Hierarchical agglomerative clustering)

I 최단 연결법 (Single Linkage Method)



$$d_{(UV)W} = \min(d_{UW}, d_{VW})$$


Unit 02 | 계층적 군집화(Hierarchical agglomerative clustering)

I 최단 연결법 (Single Linkage Method)

데이터	(x1, x2)		유클리드 제곱거리	A	B	C	D	E
A	(1, 5)	 <i>Dist(데이터)</i>	A	0				
B	(2, 4)		B	2	0			
C	(4, 6)		C	10	8	0		
D	(4, 3)		D	13	5	9	0	
E	(5, 3)		E	20	10	10	1	0

Unit 02 | 계층적 군집화(Hierarchical agglomerative clustering)

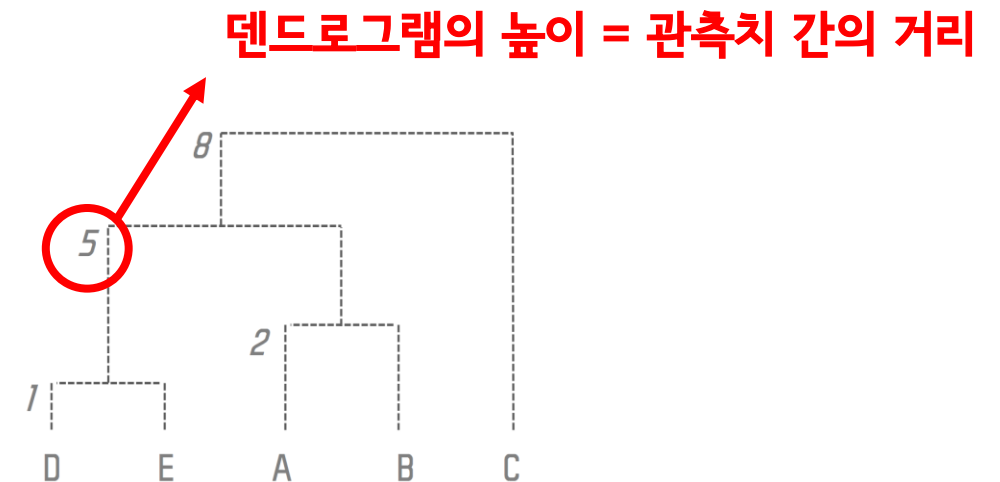
I 최단 연결법 (Single Linkage Method)

데이터	(x1, x2)		유클리드 제곱거리	A	B	C	(D, E)
A	(1, 5)	 <i>Dist(데이터)</i>	A	0			
B	(2, 4)		B	2	0		
C	(4, 6)		C	10	8	0	
D	(4, 3)		(D, E)	13	5	9	0
E	(5, 3)						

Unit 02 | 계층적 군집화(Hierarchical agglomerative clustering)

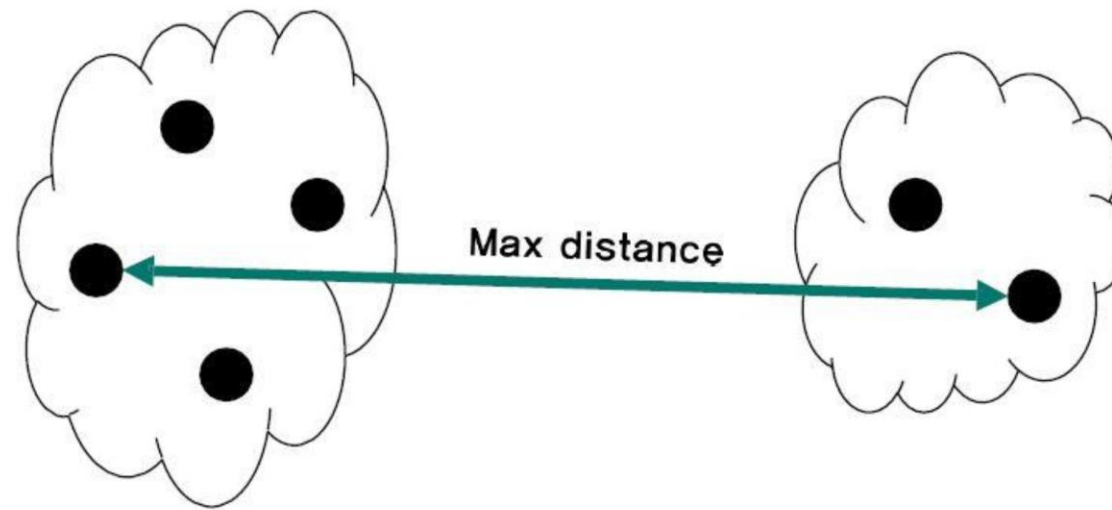
I 최단 연결법 (Single Linkage Method)

유클리드 제곱거리	(A, B)	C	(D, E)
(A, B)	0		
C	8	0	
(D, E)	5	9	0



Unit 02 | 계층적 군집화(Hierarchical agglomerative clustering)


② 최장 연결법 (Complete Linkage Method)



$$d_{(UV)W} = \max(d_{UW}, d_{VW})$$


Unit 02 | 계층적 군집화(Hierarchical agglomerative clustering)

I 최장 연결법 (Complete Linkage Method)

데이터	(x1, x2)		유클리드 제곱거리	A	B	C	D	E
A	(1, 5)	 <i>Dist(데이터)</i>	A	0				
B	(2, 4)		B	2	0			
C	(4, 6)		C	10	8	0		
D	(4, 3)		D	13	5	9	0	
E	(5, 3)		E	20	10	10	1	0

Unit 02 | 계층적 군집화(Hierarchical agglomerative clustering)

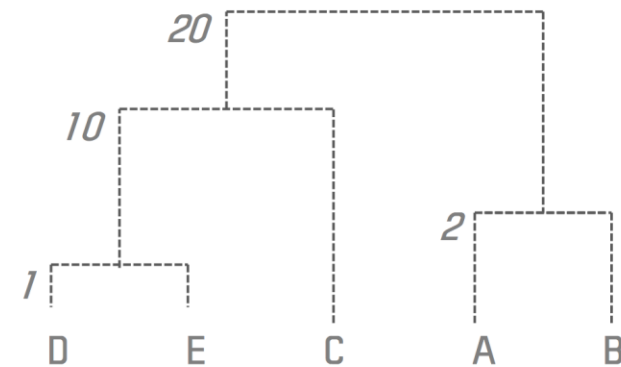
2 최장 연결법 (Complete Linkage Method)

데이터	(x1, x2)		유클리드 제곱거리	A	B	C	(D, E)
A	(1, 5)	 <i>Dist(데이터)</i>	A	0			
B	(2, 4)		B	2	0		
C	(4, 6)		C	10	8	0	
D	(4, 3)		(D, E)	20	10	10	0
E	(5, 3)						

Unit 02 | 계층적 군집화(Hierarchical agglomerative clustering)

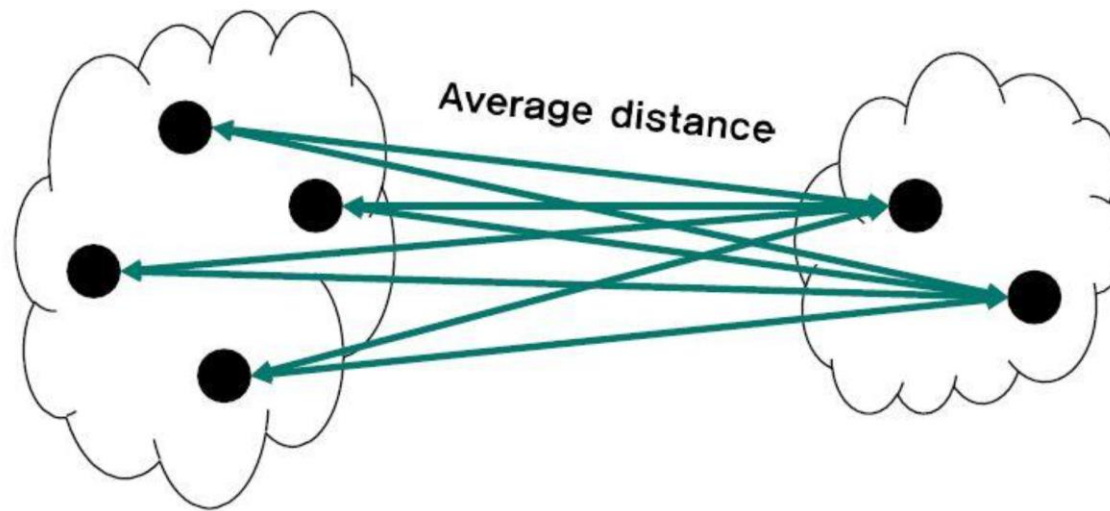
2 최장 연결법 (Complete Linkage Method)

유클리드 제곱거리	(A, B)	C	(D, E)
(A, B)	0		
C	10	0	
(D, E)	20	10	0



Unit 02 | 계층적 군집화(Hierarchical agglomerative clustering)

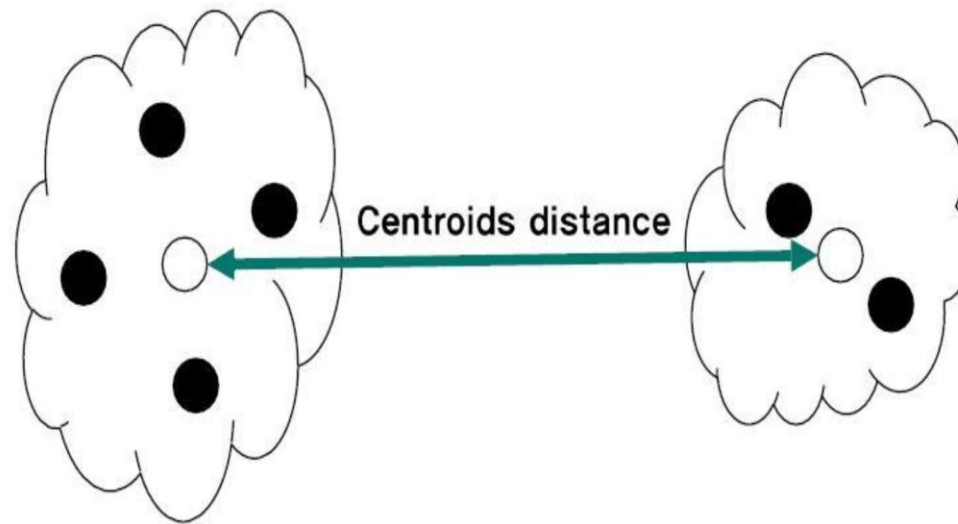
3 평균 연결법 (Average Linkage Method)



$$d_{(UV)W} = \frac{\sum_{x_i \in (U,V)} \sum_{x_j \in W} d(x_i, x_j)}{n_{(UV)}n_W}$$

Unit 02 | 계층적 군집화(Hierarchical agglomerative clustering)

4 중심 연결법 (Centroid Linkage Method)



$$d(G_1, G_2) = \|\bar{x}_1 - \bar{x}_2\|$$

Unit 02 | 계층적 군집화(Hierarchical agglomerative clustering)

4 중심 연결법 (Centroid Linkage Method)

데이터	(x1, x2)		유클리드 제곱거리	A	B	C	D	E
A	(1, 5)	➔ <i>Dist(G O E)</i>	A	0				
B	(2, 4)		B	2	0			
C	(4, 6)		C	10	8	0		
D	(4, 3)		D	13	5	9	0	
E	(5, 3)		E	20	10	10	1	0

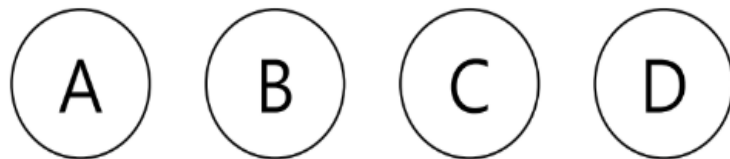
Unit 02 | 계층적 군집화(Hierarchical agglomerative clustering)

4 중심 연결법 (Centroid Linkage Method)

데이터	(x1, x2)		유클리드 제곱거리	A	B	C	(D, E)
A	(1, 5)		A	0			
B	(2, 4)		B	2	0		
C	(4, 6)		C	10	8	0	
D	(4, 3)		(D, E)	16.25	7.25	9.25	0
E	(5, 3)						

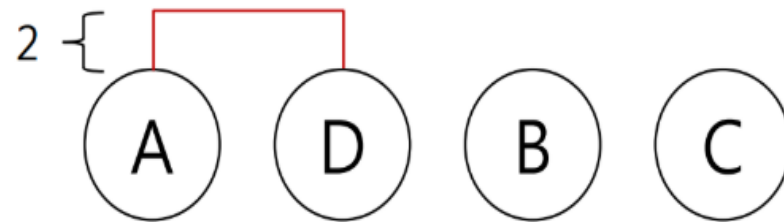
$\xrightarrow{\text{Dist}(\{D, E\})}$
 $\rightarrow (4.5, 3)$

Unit 02 | 계층적 군집화(Hierarchical agglomerative clustering)



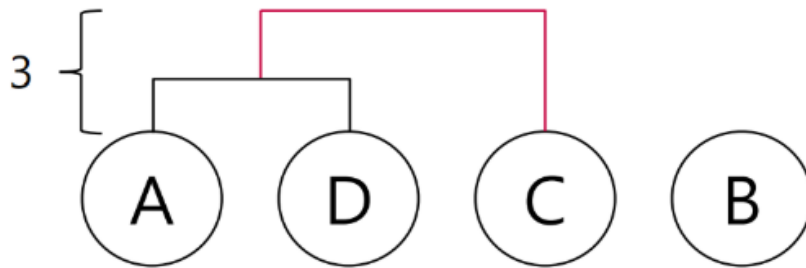
	A	B	C	D
A		20	7	2
B			10	25
C				3
D				

Unit 02 | 계층적 군집화(Hierarchical agglomerative clustering)



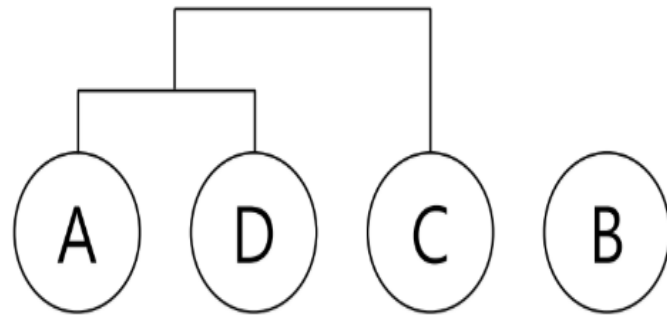
	A	B	C	D
A		20	7	2
B			10	25
C				3
D				

Unit 02 | 계층적 군집화(Hierarchical agglomerative clustering)



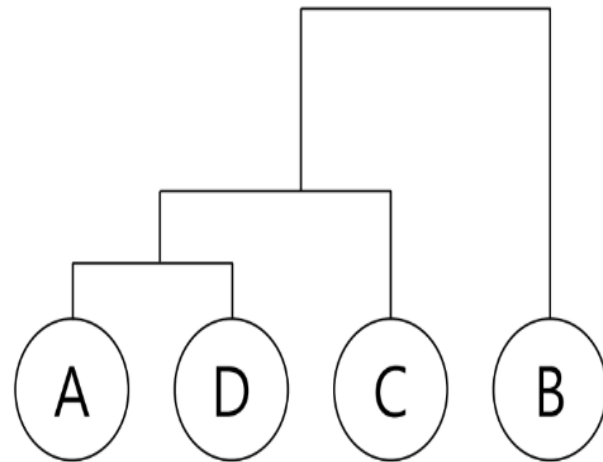
	AD	B	C	
AD		20	3	
B			10	
C				

Unit 02 | 계층적 군집화(Hierarchical agglomerative clustering)



	ADC	B		
ADC		10		
B				

Unit 02 | 계층적 군집화(Hierarchical agglomerative clustering)



	AD CB			
AD CB				

Unit 02 | 계층적 군집화(Hierarchical agglomerative clustering)

요약

1. 단일 데이터 간 거리를 정의하고

- 맨하탄 거리, 유클리드 거리 등

2. 군집-군집 or 군집-개체 간 거리를 정의하고

- 최단 연결법, 평균 연결법, 최장 연결법 등

3. 돌리자!

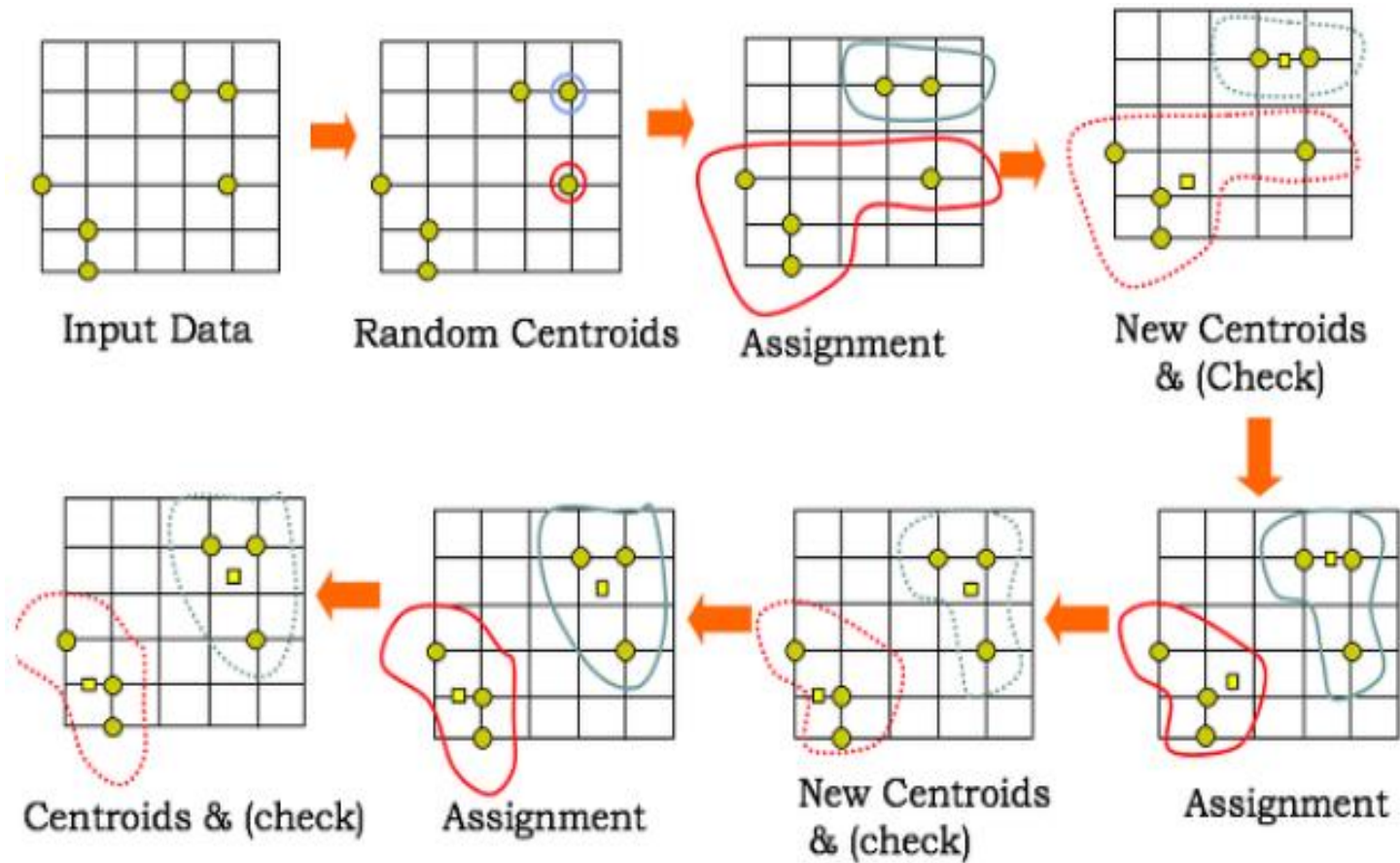
Unit 03 | K-Means

K-Means(비계층적 군집화)

1. 데이터 내 객체 중 임의로 K개의 군집 중심점(Centroid) 설정
2. 모든 객체에 대해 각 군집 중심점까지의 거리 계산
3. 모든 객체를 가장 가까운 군집 중심점이 속한 군집으로 할당
4. 각 군집의 중심점 재설정
5. 군집의 중심점이 변경되지 않을 때까지 1~4 반복

(또는 적당한 범위 내로 수렴하거나 적당한 반복회수에 도달할 때까지 반복)

Unit 03 | K-Means



Unit 03 | K-Means

K-Means의 주요 변수

1. 초기 군집 중심점(Centroid) 설정
2. 군집의 개수(K)

Unit 03 | K-Means

초기 centroid 설정

1. 무작위 분할 : \sqrt{n} , n = 데이터의 수
2. Forgy 알고리즘 : Chooses k objects at random and uses them as the initial centroids.

Unit 03 | K-Means

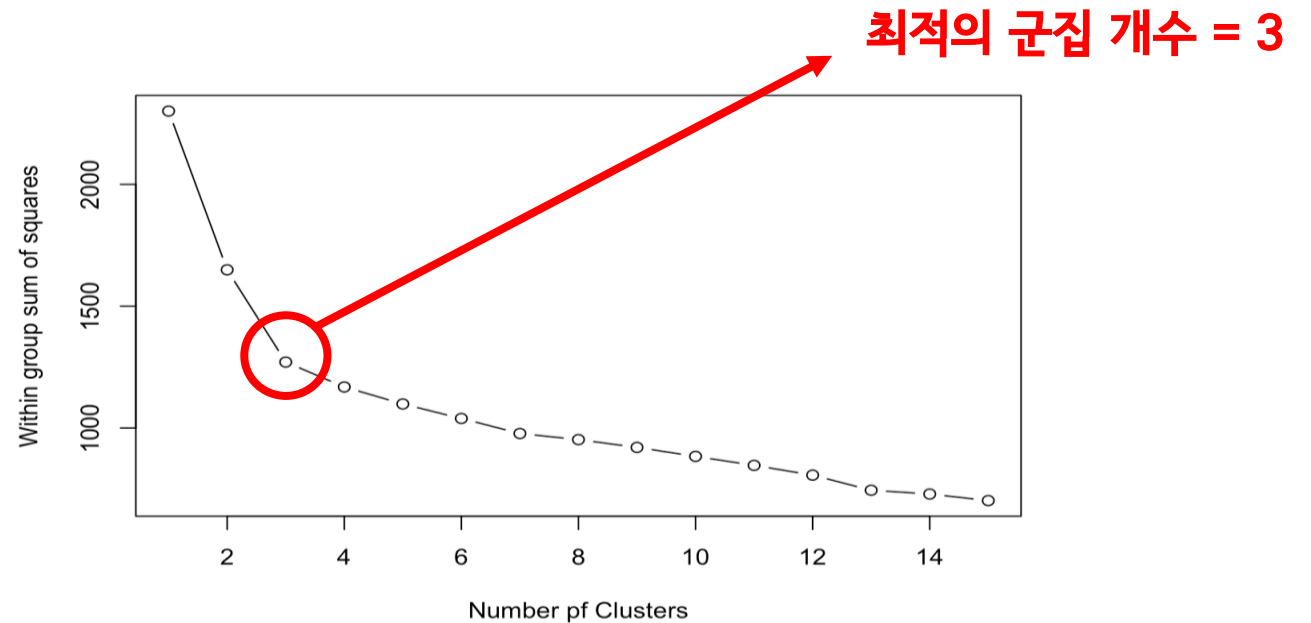
초기 centroid 설정

1. 무작위 분할 : \sqrt{n} , n = 데이터의 수
2. Forgy 알고리즘 ← 간단하고 좋은 성능

Unit 03 | K-Means

군집 개수의 선택

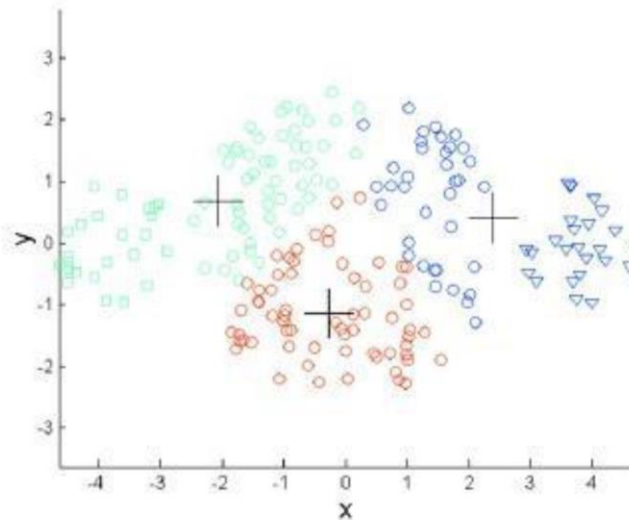
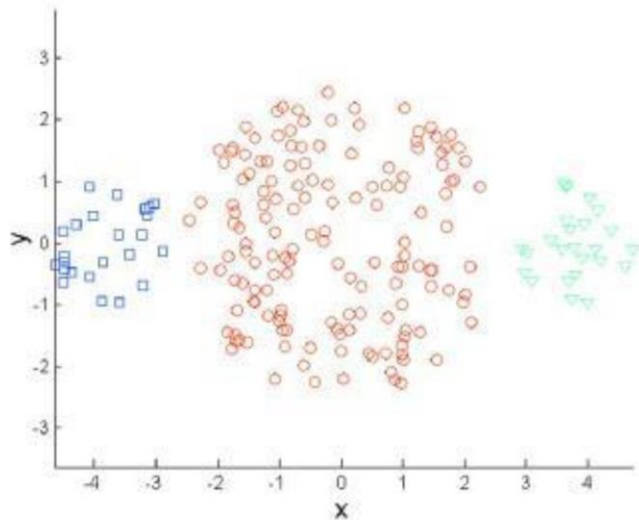
1. 경험적 방법 : \sqrt{n} , n = 데이터의 수
2. Elbow Point 기법(통계적 기법)



Unit 03 | K-Means

K-Means의 한계점

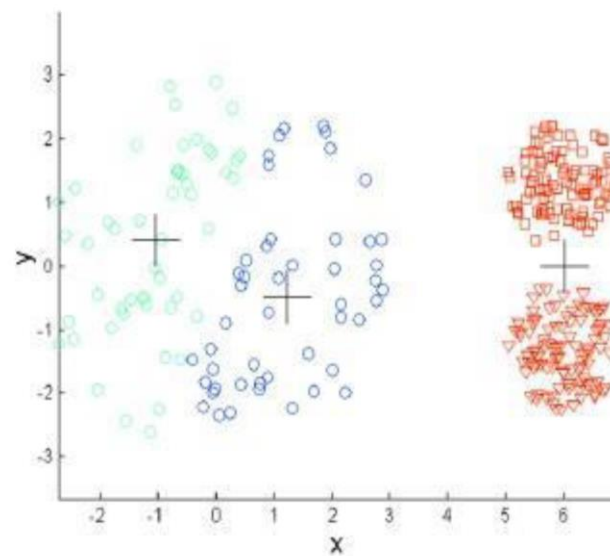
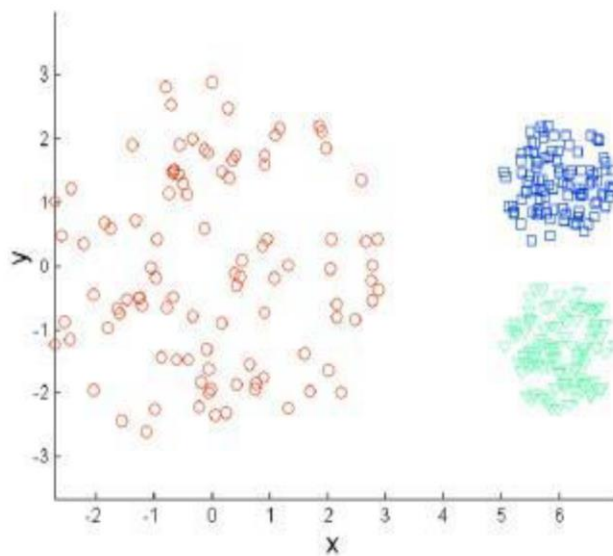
- 문제점 1: 서로 다른 크기의 군집을 잘 찾아내지 못함



Unit 03 | K-Means

K-Means의 한계점

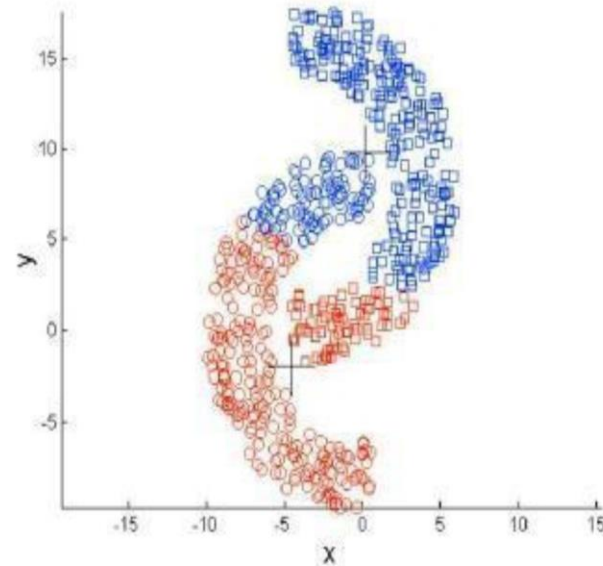
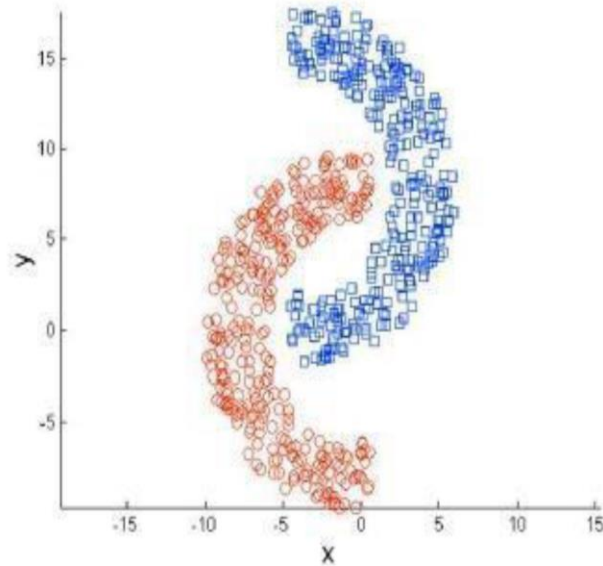
- 문제점 2: 서로 다른 **밀도**의 군집을 잘 찾아내지 못함



Unit 03 | K-Means

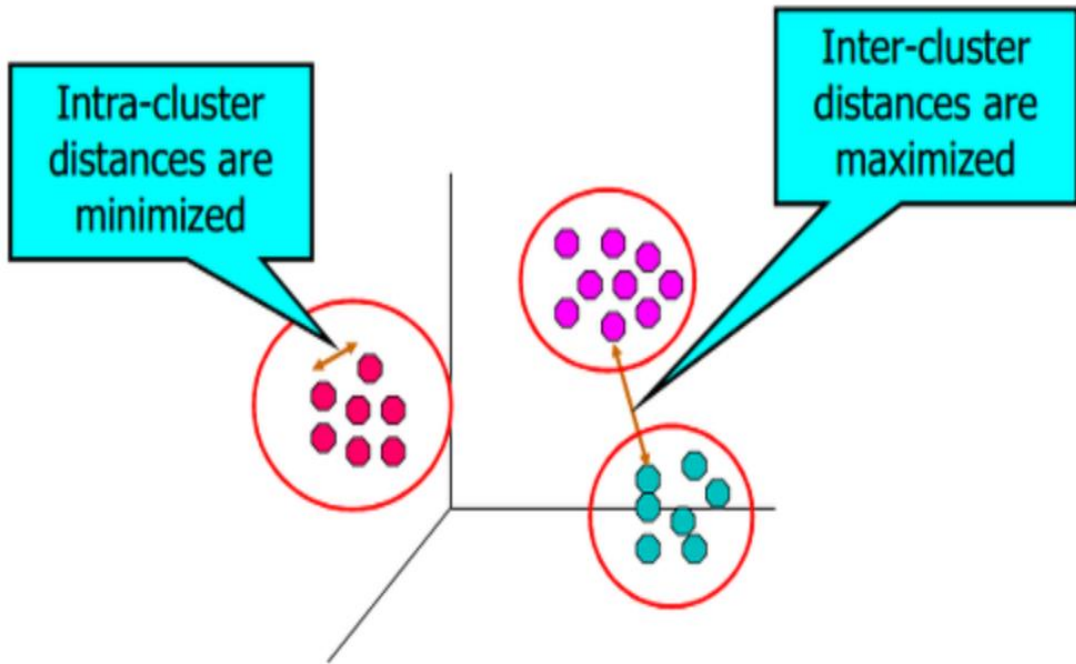
K-Means의 한계점

- 문제점 3: 구형이 아닌 형태의 군집을 판별하기 어려움



Unit 04 | 모델 평가

모델 평가



- Inter-cluster distance
클러스터 간의 거리를 최대로
- Intra-cluster distance
각 클러스터 내 데이터의 거리는 작게

Unit 04 | 모델 평가

클러스터링 평가 척도

내부 평가

- 군집화한 그 자체를 놓고 평가하는 방식

1. **Dunn Index**
2. **실루엣(Silhouette)**
3. **Davies - Bouldin Index**

외부 평가

- 군집화에 사용되지 않은 데이터로 평가하는 방식

1. **Rand Measure**
2. **F - Measure**
3. **Jaccard Index**

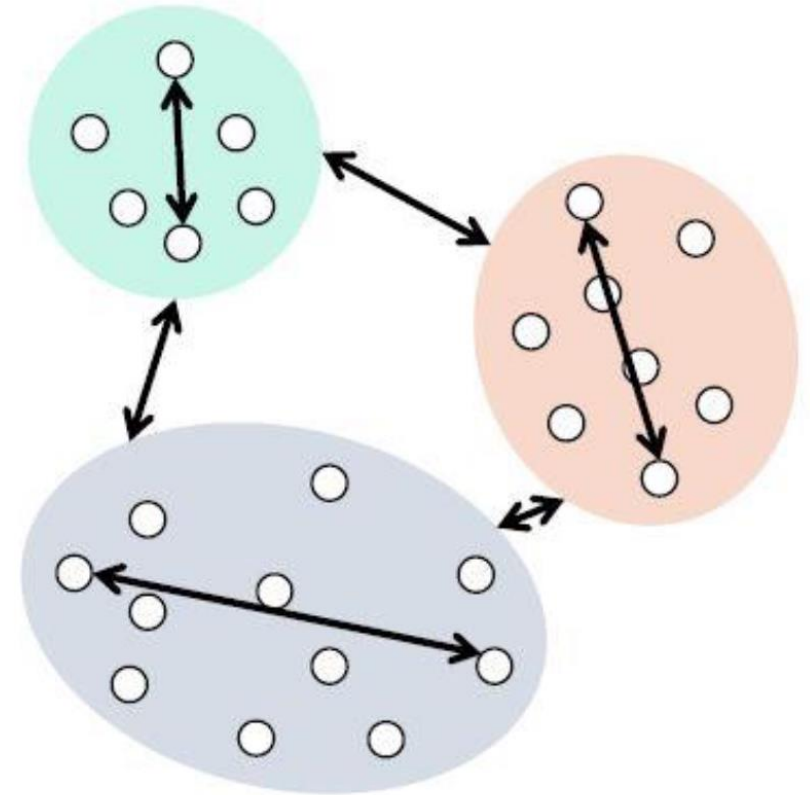
Unit 04 | 모델 평가

내부 평가

1. Dunn Index

$$DI = \frac{\text{군집과 군집 사이의 거리 중 최소값}}{\text{군집 내 객체 간 거리 중 최대값}}$$

군집과 군집 사이의 거리가 클수록,
군집 내 객체 간 거리가 작을수록
= DI가 큰 모델
좋은 모델



Unit 04 | 모델 평가

내부 평가

2. 실루엣(Silhouette)

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

$$s(i) = \begin{cases} 1 - a(i)/b(i), & \text{if } a(i) < b(i) \\ 0, & \text{if } a(i) = b(i) \\ b(i)/a(i) - 1, & \text{if } a(i) > b(i) \end{cases}$$

$a(i)$: 객체 i 로부터 같은 군집 내
모든 다른 객체들 사이 평균 거리
(클러스터내 데이터 응집도를 나타내는 값)

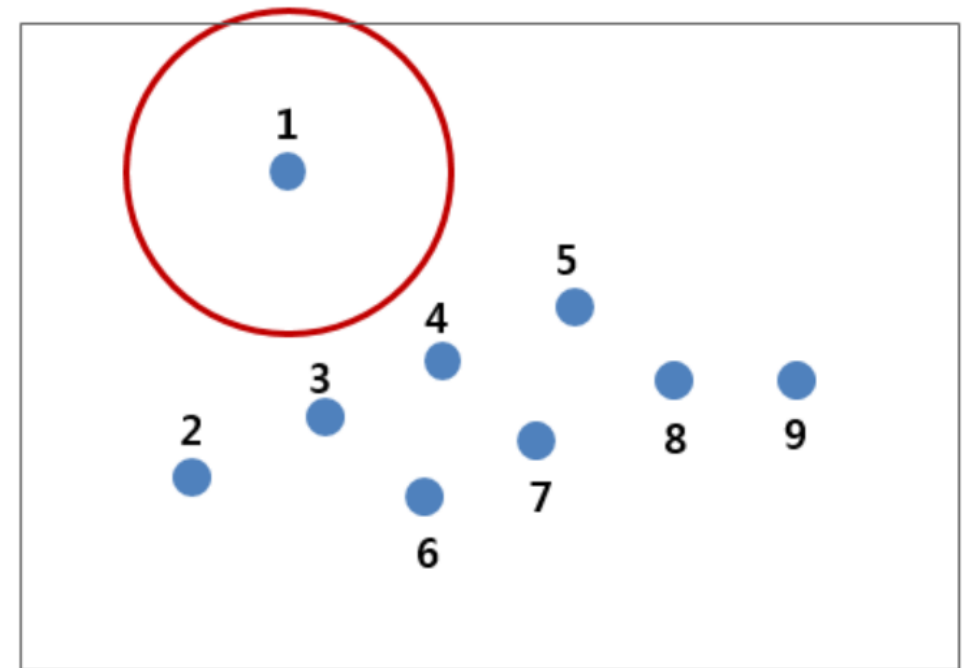
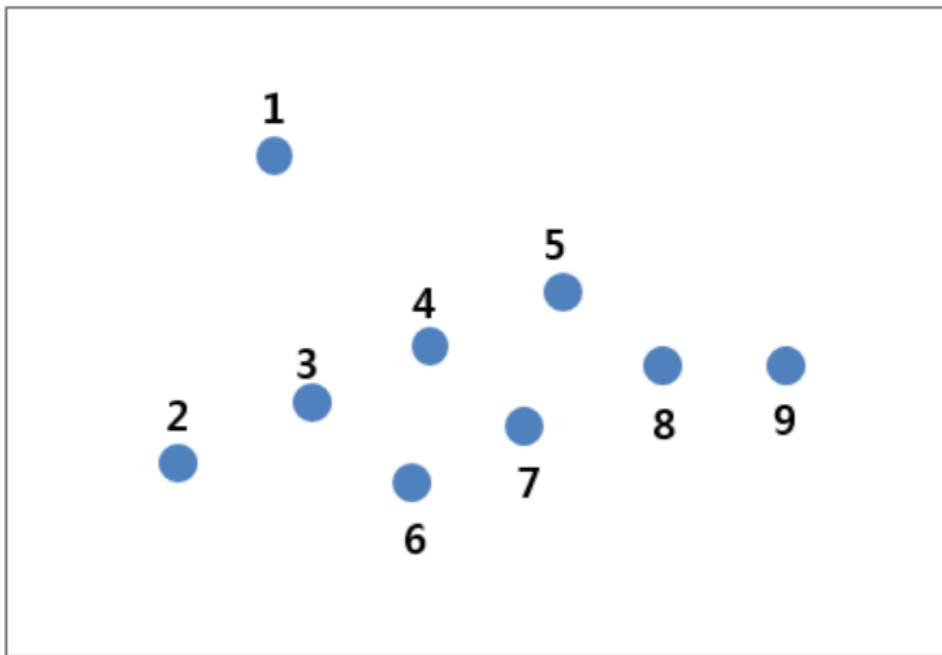
$b(i)$: 객체 i 로부터 다른 군집 내
객체들 사이의 평균 거리 중 최소값
(클러스터간 분리도를 나타내는 값)

$S(i)$ 가 1에 가까울 수록 좋은 모델

Unit 05 | DBSCAN

DBSCAN- 밀도 기반 클러스터링

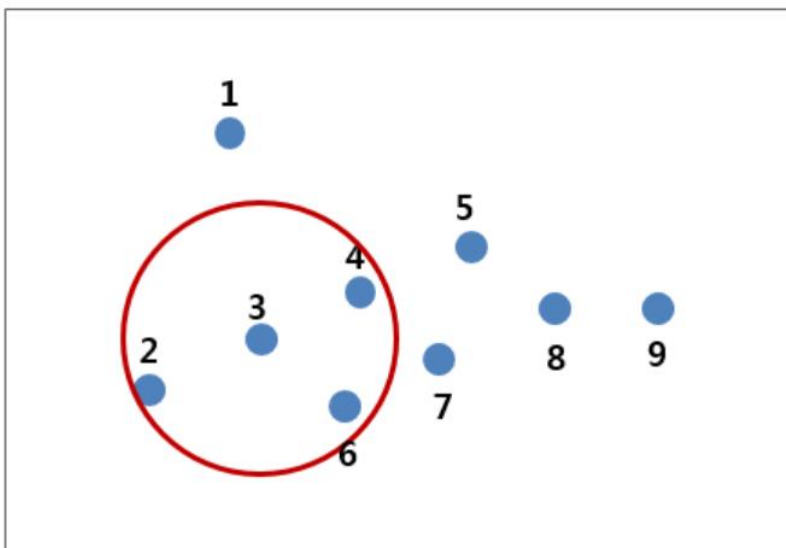
: 점 P에서부터 거리 $e(\epsilon)$ 내에 점이 $m(\text{minPts})$ 개 있으면 하나의 군집으로 인식



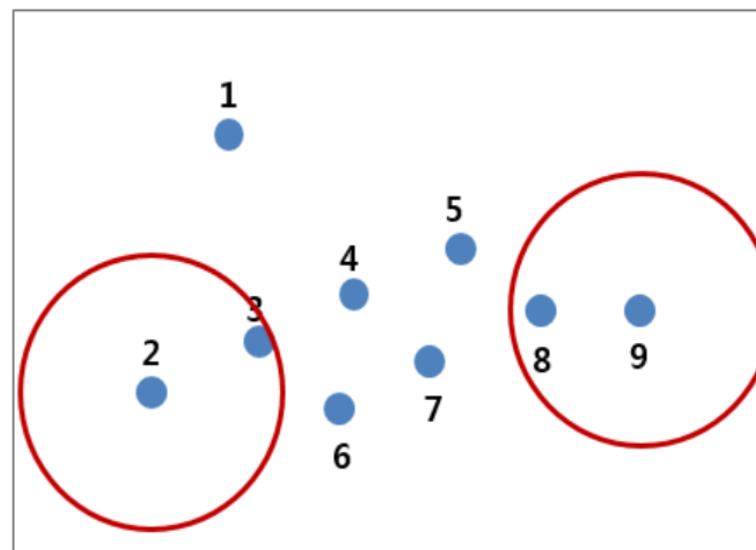
노이즈 데이터(noise point)

Unit 05 | DBSCAN

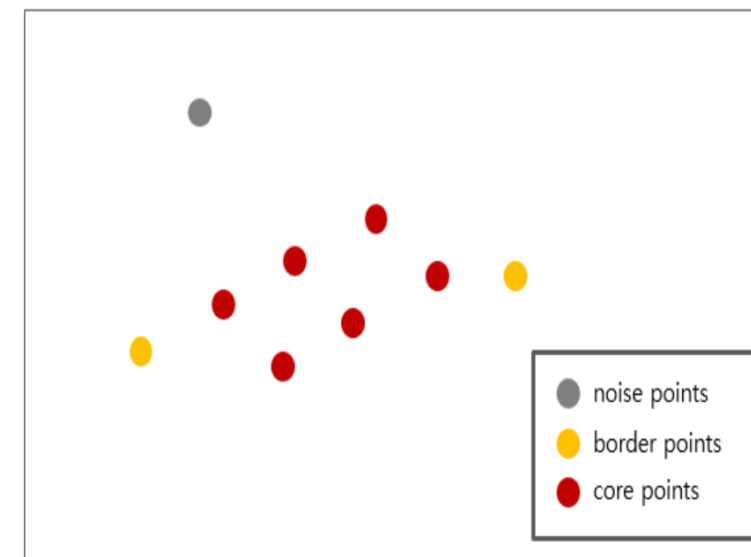
코어 데이터(core point)



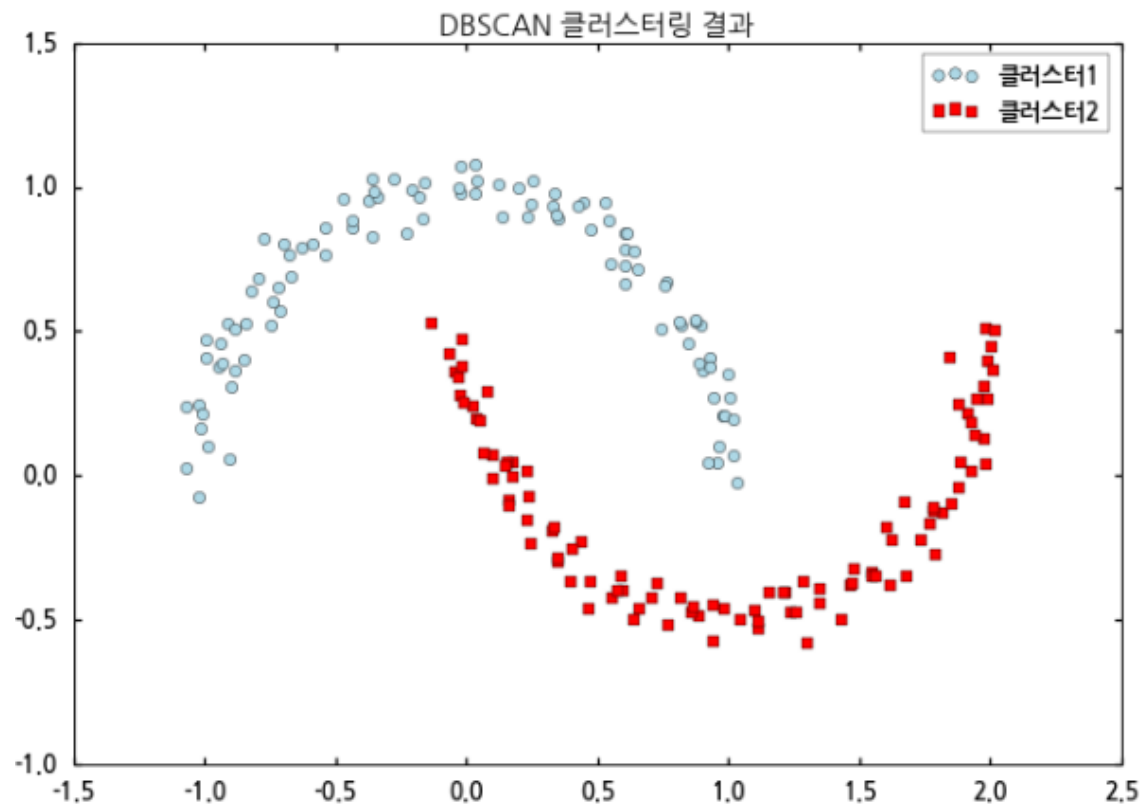
경계 데이터(border point)



최종 결과



Unit 05 | DBSCAN



<특징>

- K- means와 같이 **클러스터의 수**를 정하지 않아도 됨
- 비선형 경계의 군집을 구하는 것도 가능
(밀도에 따라 클러스터를 서로 연결하기 때문)
- 노이즈 데이터를 따로 분류하여 노이즈 데이터들이
군집에 영향을 주지 않음

Q & A

들어주셔서 감사합니다.