KUBIG
Data Science and Machine Learning

Week 7. Unsupervised Lerning

Unsupervised Learning

Machine Learning

회귀 분류 최적화 SVM
Regularization Boosting.

기계학습
(Machine Learning)

지도학습
(Supervised Learning)
입력(input)과 결과값(label)을 이용한 학습

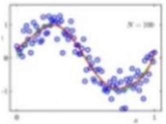
비지도학습
(Unsupervised Learning)
입력(input)만 가지고 학습

강화학습
(Reinforcement Learning)
- 보상 시스템으로 학습
- 의사 결정을 위한 최적의 액션 선택

분류
(classification)

- MNIST 필기체 인식
- 스팸 메일 분류

회귀
(regression)



- 주가 예측

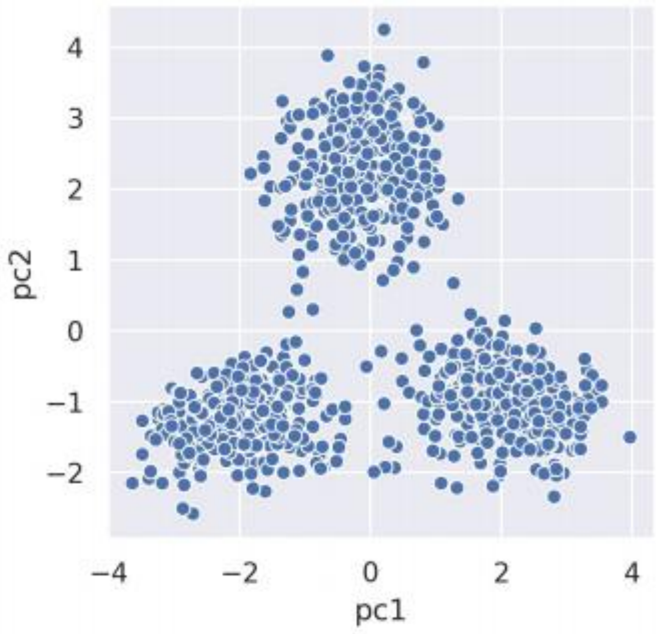
클러스터링
(clustering)

- 마케팅 고객 그룹화

PCA t-SNE

- Semi-Supervised Learning
- Self-Supervised Learning

Clustering

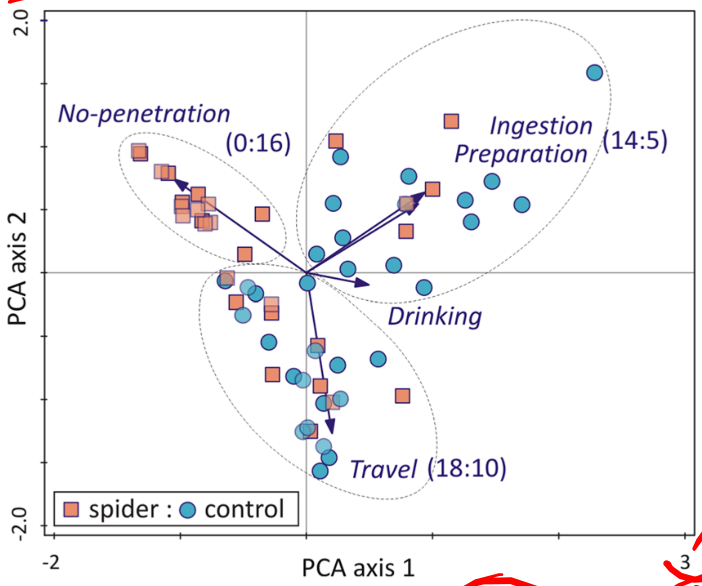


- Don’t have labels for each point
- Want to infer pattern without labels

Dimensionality Reduction

Curse of Dimensionality

$(X_1, \dots X_n) \to (Z_1, Z_2)$



feature selection

$\cdot\ (X_1, X_2, X_3, X_4) \to (Z_1, Z_2)$

$Z_1 = 2X_1 + X_2$, $Z_2 = X_3^2 + X_4$

- High dimensional data leads to high computational cost to perform learning

- Reduce correlation and complexity in data while preserving most of the relevant information in the data

PCA

선형

[illegible]

eigenvector

SVD

t-SNE

비선형 Dimensionality Reduction

Loss Function

Clustering

① Prototype-based Clustering

② Hierarchical Clustering

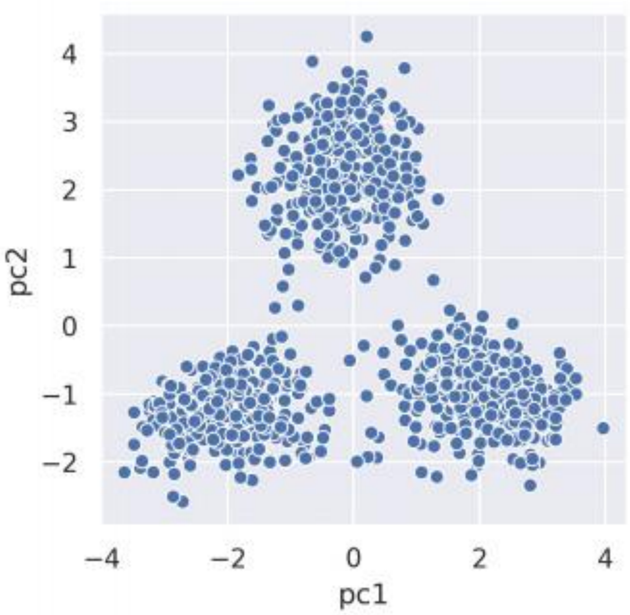
③ Density-Based Clustering.

① 모델 performance metric

② Distance Metric

Validity Index

- What is the best cluster for this data?



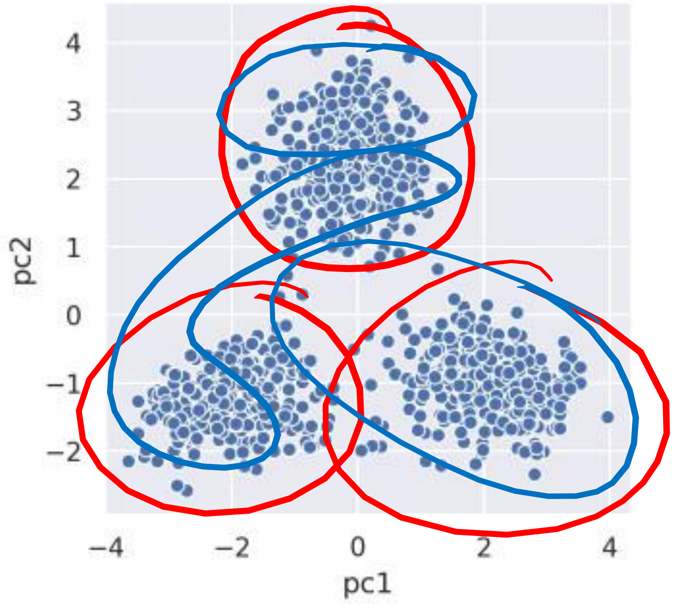
Validity Index

유효성 지표 / 성능 척도

Supervised: · Accuracy · Recall

- What is the best cluster for this data?

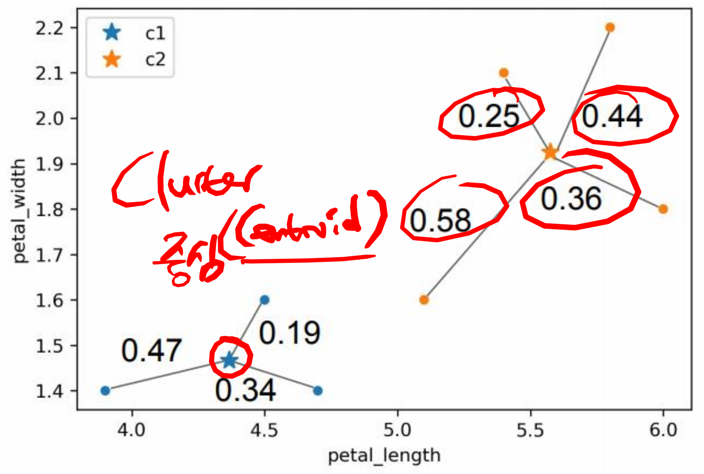
- High Intra-cluster similarity
- Low inter-cluster similarity



Validity Index

Two common loss functions:

SSE
- Inertia: Sum of squared distances from each data point to its center
- Distortion: Weighted sum of squared distances from each data point to its center



Example:

SSE
- Inertia: $0.47^2 + 0.19^2 + 0.34^2 + 0.25^2 + 0.58^2 + 0.36^2 + 0.44^2$
- Distortion: $(0.47^2 + 0.19^2 + 0.34^2)/3 + (0.25^2 + 0.58^2 + 0.36^2 + 0.44^2)/4$

Lower is Better

유사성 척도 : ESS(error sum of squares) 의 증분.

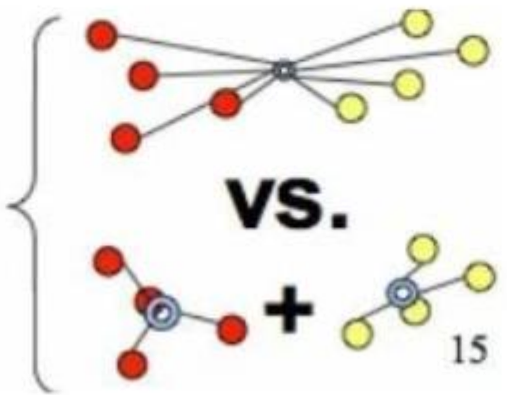
Validity Index

- Ward's Method

$$d(i+j,k)=\frac{\left\|\mu_{i+j}-\mu_k\right\|^2}{\frac{1}{n_1}+\frac{1}{n_2}}$$

$$ESS=\sum_{k=1}^{K}\sum_{x_i\in C_k}\sum_{j=1}^{n}(x_{ij}-\bar{x}_{kj})^2$$

* ESS : error sum of squares
** k : number of clusters (1 ~ K)
x_i : elements of cluster C_k
j : number of variables (1 ~ n)

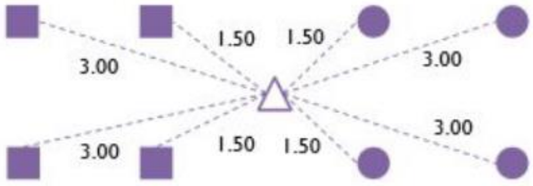


10번째

- SSE before merge: $1^2+1^2+1^2+1^2+1^2+1^2+1^2+1^2=8$

11번째

- SSE after merge: $4\times1.5^2+4\times3^2=45$



- Ward distance: 45-8 = 37

Validity Index

external index

internal index

Data $D = \{x_1, x_2, \dots, x_m\}$

Clusters $C = \{C_1, C_2, \dots, C_k\}$

Reference Clusters $C^* = \{C_1^*, C_2^*, \dots, C_s^*\}$

Cluster Labels: $\boldsymbol{\lambda}$, $\boldsymbol{\lambda}^*$

- External Index (Using Reference Model)

$$a = |SS|, \quad SS = \{(\boldsymbol{x}_i, \boldsymbol{x}_j) \mid \boldsymbol{\lambda}_i = \boldsymbol{\lambda}_j, \boldsymbol{\lambda}_i^* = \boldsymbol{\lambda}_j^*, i < j)\}$$

$$b = |SD|, \quad SD = \{(\boldsymbol{x}_i, \boldsymbol{x}_j) \mid \boldsymbol{\lambda}_i = \boldsymbol{\lambda}_j, \boldsymbol{\lambda}_i^* \neq \boldsymbol{\lambda}_j^*, i < j)\}$$

$$c = |DS|, \quad DS = \{(\boldsymbol{x}_i, \boldsymbol{x}_j) \mid \boldsymbol{\lambda}_i \neq \boldsymbol{\lambda}_j, \boldsymbol{\lambda}_i^* = \boldsymbol{\lambda}_j^*, i < j)\}$$

$$d = |DD|, \quad DD = \{(\boldsymbol{x}_i, \boldsymbol{x}_j) \mid \boldsymbol{\lambda}_i \neq \boldsymbol{\lambda}_j, \boldsymbol{\lambda}_i^* \neq \boldsymbol{\lambda}_j^*, i < j)\}$$

Jaccard Coefficient

CV (IoU)

$$JC = \frac{a}{\boxed{a + b + c}}$$

Fowlkes and Mallows Index

$$FMI = \sqrt{\frac{a}{a+b} * \frac{a}{a+c}}$$

Rand Index

$$RI = \frac{2(a+d)}{m(m-1)}$$

Validity Index

Model Score

Data $D = \{x_1, x_2, \dots, x_m\}$

Clusters $C = \{C_1, C_2, \dots, C_k\}$

- Internal Index

$$avg(C) = \frac{2}{|C|(|C|-1)} \sum_{1 \le i < j \le |C|} dist(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

$$\text{diam}(C) = max_{1 \le i < j \le |C|}\, dist(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

$$d_{min}(C_i, C_j) = min_{\boldsymbol{x}_i \in C_i, \boldsymbol{x}_j \in C_j}\, dist(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

$$d_{cen}(C_i, C_j) = dist(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j)$$

$$\boldsymbol{\mu} = \frac{1}{|C|} \textstyle\sum_{1 \le i \le |C|} \boldsymbol{x}_i$$

Daives-Bouldin Index

$$DBI = \frac{1}{k} \sum_{i=1}^{k} max_{j \ne i}\left(\frac{avg(C_i) + avg(C_j)}{d_{cen}(C_i, C_j)}\right)$$

Dunn Index

$$DI = \frac{1}{|C|} \min_{1 \le i \le k} \left\{ \min_{j \ne i}\left(\frac{d_{min}(C_i, C_j)}{\max\limits_{1 \le l \le k} diam(C_l)} \right) \right\}$$

Distance Measure

정규화
표준화

몸무게 (kg)

2차원 Euclidean

- Distance measure

키 (mm)

p=2 $$d(\mathbf{u},\mathbf{v}) = \left(\sum |u_i - v_i|^2\right)^{\frac{1}{2}} = ||\mathbf{u}-\mathbf{v}||_2$$ *Euclidean (L2 norm)*

p=1 $$d(\mathbf{u},\mathbf{v}) = \sum |u_i - v_i| = ||\mathbf{u}-\mathbf{v}||_1$$ *Manhattan (L1 norm)*

$$d(\mathbf{u},\mathbf{v}) = \left(\sum |u_i - v_i|^p\right)^{\frac{1}{p}} = ||\mathbf{u}-\mathbf{v}||_p$$ *Minkowski (Lp norm)*

$$d(\mathbf{u},\mathbf{v}) = \sqrt{(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})}$$ *Mahalanobis Distance*

상관계수 고려

1-2 cor.

Distance Measure

매우불만족 불만족 보통 만족 매우만족

1 2 3 4 5

1. 1

- Distance measure

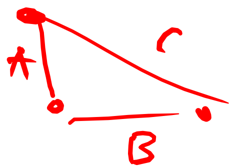
연속 숫자 (유사)

특성변수	비유사성 척도	
연속형	제곱 유클리디안 거리 (squared Euclidean distance)	$d(x_i, x_{i'}) = \sum_{j=1}^{m} (x_{ij} - x_{i'j})^2 = (x_i - x_{i'})^T (x_i - x_{i'})$
	맨해튼 ~~1차 유클리디안~~ 거리 (L1 Euclidean distance)	$d(x_i, x_{i'}) = \sum_{j=1}^{m} \lvert x_{ij} - x_{i'j} \rvert$
	마할라노비스 거리 (Mahalanobis distance)	$d(x_i, x_{i'}) = (x_i - x_{i'})^T \Sigma^{-1} (x_i - x_{i'})$ (Σ: 특성변수들의 분산-공분산 행렬)
순서형 (ordinal)	특성변수를 $\frac{k-1/2}{M}$ ($k = 1, \cdots, M$, M: 순서의 크기)로 변환 후 연속형 비유사성 척도 적용	
범주형 (categorical)	두 관측치가 같은 범주에 속하면 0, 아니면 1로 값 부여	

1 2 3 4 5

dummy variance

distance VDM



Prototype – based Clustering

Centroid

K-Nearest Neighbors

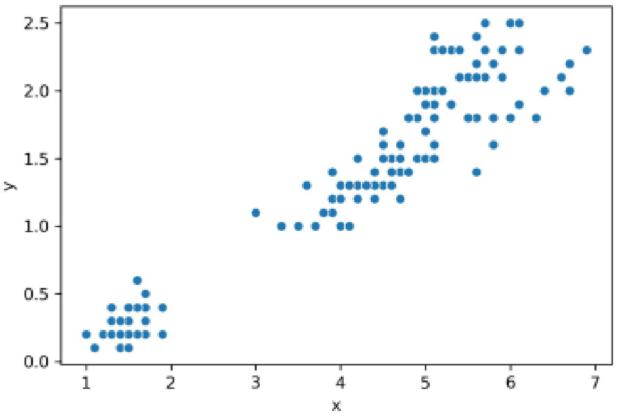
K-Means Clustering

Most popular clustering approach: K-Means

K= 클러스터 개수

- Pick an arbitrary k, and randomly place k “centers”, each a different color
- Repeat until convergence:
    - Color points according to the closest center
    - Move center for each color to center of points with that color

Raw data ⟶



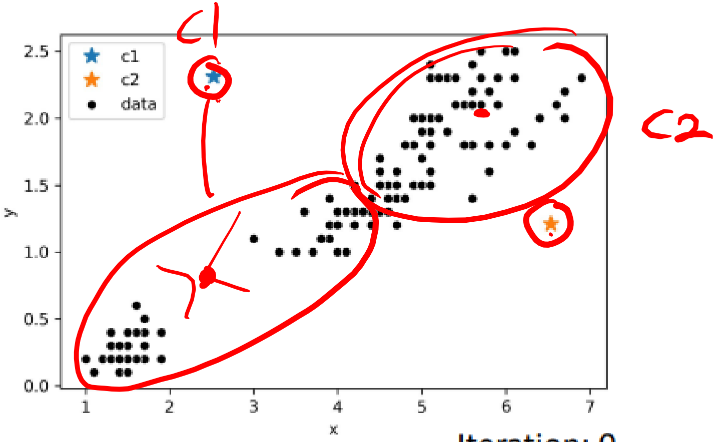
KUBIG

K-Means Clustering

Most popular clustering approach: K-Means.

- Pick an arbitrary k, and **randomly place k "centers",** each a different color
- Repeat until convergence:
  - Color points according to the closest center
  - Move center for each color to center of points with that color

Initial random placement of two centers



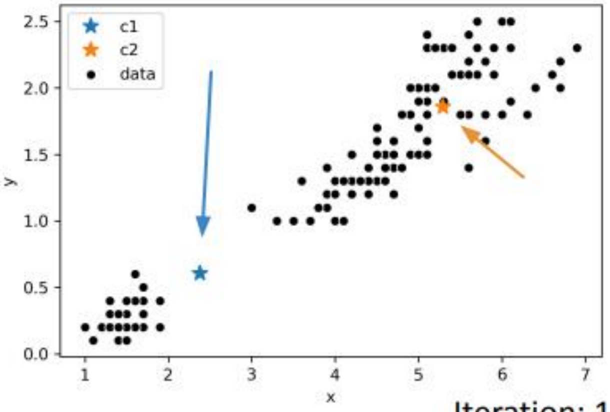
Iteration: 0

K-Means Clustering

Most popular clustering approach: K-Means

- Pick an arbitrary k, and randomly place k "centers", each a different color
- Repeat until convergence:
    - Color points according to the closest center
    - **Move center for each color to center of points with that color**

Centers moved to their new homes



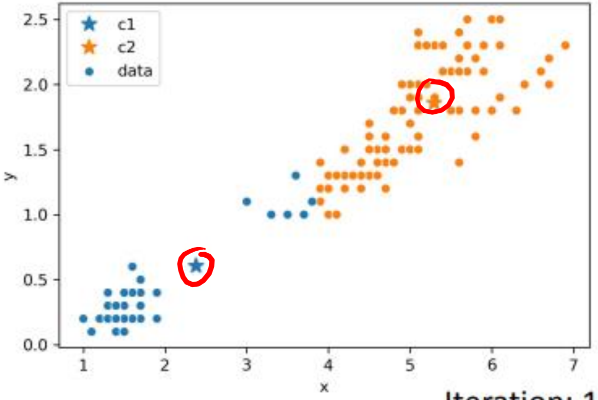
Iteration: 1

K-Means Clustering

Most popular clustering approach: K-Means

- Pick an arbitrary k, and randomly place k “centers”, each a different color
- Repeat until convergence:
    - **Color points according to the closest center**
    - Move center for each color to center of points with that color

Data colored by closest center (in new position)



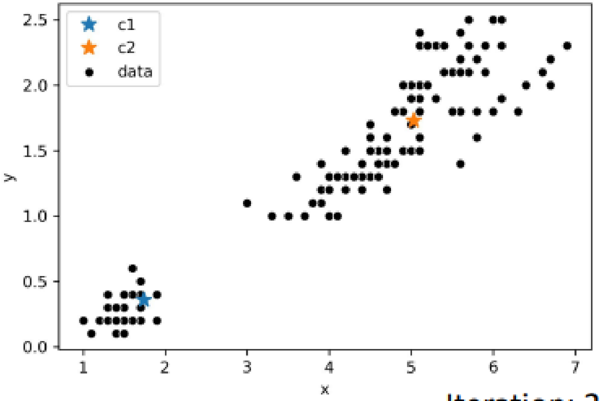
Iteration: 1

K-Means Clustering

Most popular clustering approach: K-Means

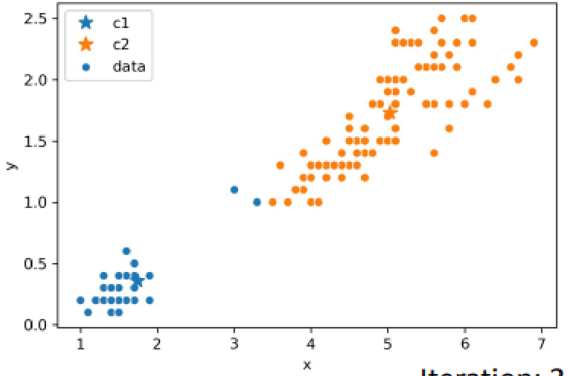
- Pick an arbitrary k, and randomly place k “centers”, each a different color
- Repeat until convergence:
  - Color points according to the closest center
  - **Move center for each color to center of points with that color**

Centers moved to new position

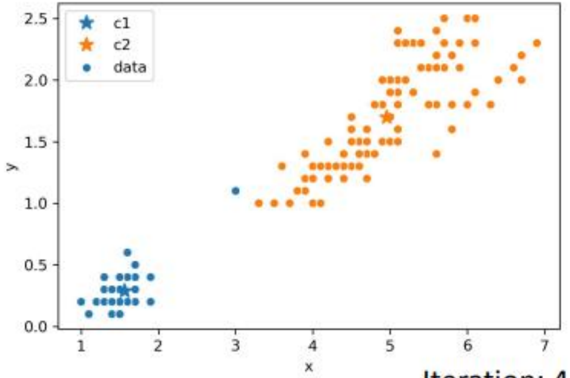


Iteration: 2

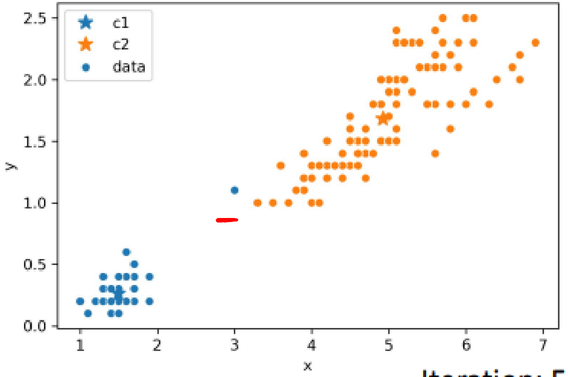
K-Means Clustering



Iteration: 3



Iteration: 4



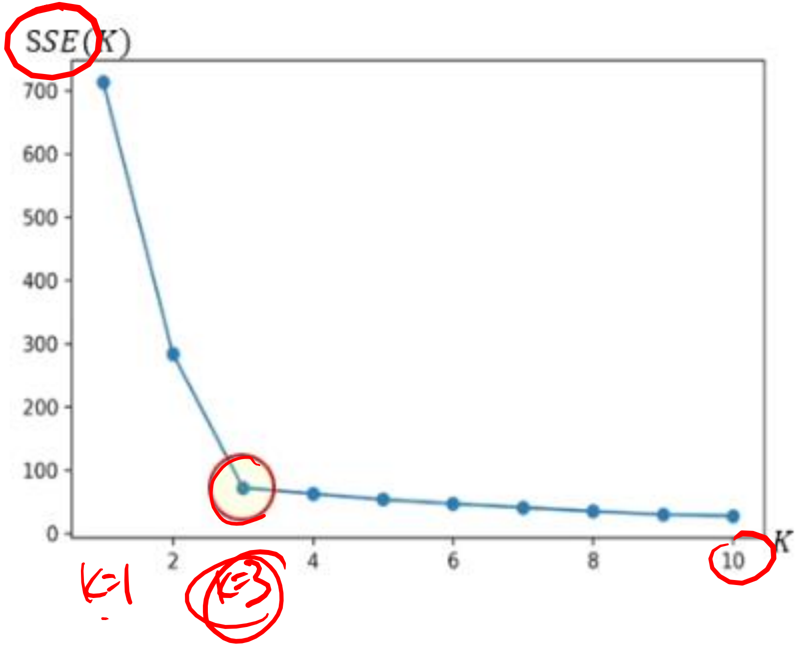
Iteration: 5

K-Means Clustering

$$S_l^2 = \sum_{i \in l} \sum_{j=1}^{d} \left((x_{ij} - \mu_{lj}\right)^2$$

$$SSE(K) = \sum_{l=1}^{K} S_l^2$$

K-Means Clustering



- ▶ K가 커질수록 $SSE(K)$ 감소
- ▶ K=3에서 $SSE(K)$의 감소속도(기울기)가 현저하게 줄어듦
- ▶ K=3 선택

➔ "Elbow method"

K-Means Clustering

- Downfalls of K-Means Clustering
  - Output is inconsistent
  - Performance depends on hyperparameter K

K-Means ++

1. N개의 표본으로 구성된 학습데이터로부터 1개의 임의표본을 뽑는다. 이를 초기 중심값 μ_1으로 한다.
2. 나머지 (n-1) 개의 자료에 대해 μ_1으로부터의 유클리디안 거리를 구하고 이 거리에 비례한 확률을 (n-1)개의 자료 각각에 부여한 후, 이 확률에 비례하여 1개의 임의표본을 뽑는다. 이를 두번째 초기 중심값 μ_2로 한다.
3. (n-2)개의 자료에 대해 $\min(d(x_i, \mu_1), d(x_i, \mu_2))$를 구해 이 최소거리에 비례한 확률을 (n-2)개 자료 각각에 부여한 후, 이 확률에 비례하여 1개의 임의표본을 뽑아 이를 두번째 초기 중심값 μ_3로 한다.
4. 앞의 방법을 계속 반복

K-Means ++

1. N개의 표본으로 구성된 학습데이터로부터 1개의 임의표본을 뽑는다. 이를 초기 중심값 μ_1으로 한다.
2. 나머지 (n-1) 개의 자료에 대해 μ_1으로부터의 유클리디안 거리를 구하고 이 거리에 비례한 확률을 (n-1)개의 자료 각각에 부여한 후, 이 확률에 비례하여 1개의 임의표본을 뽑는다. 이를 두번째 초기 중심값 μ_2로 한다.
3. (n-2)개의 자료에 대해 $\min(d(x_i, \mu_1), d(x_i, \mu_2))$를 구해 이 최소거리에 비례한 확률을 (n-2)개 자료 각각에 부여한 후, 이 확률에 비례하여 1개의 임의표본을 뽑아 이를 두번째 초기 중심값 μ_3로 한다.
4. 앞의 방법을 계속 반복

Other Prototype Based Clustering

이미 클러스터가 있다고 가정

- Centroid
- 분포

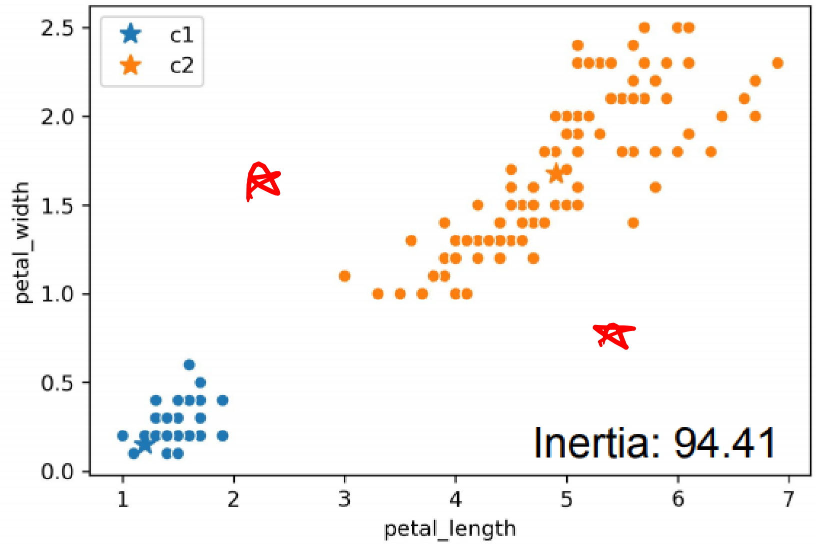
- Learning Vector Quantization (LVQ)
- Mixture-of-Gaussian Clustering

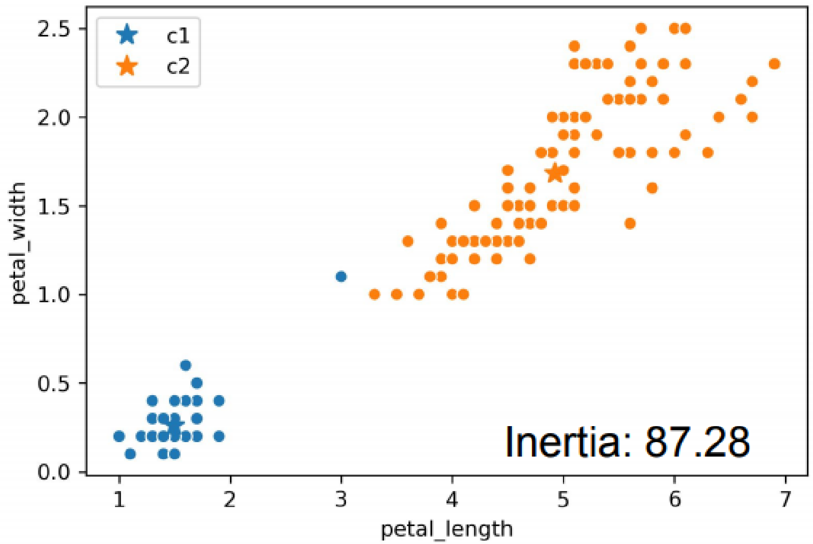
(분포) ← EM 알고리듬

Hierarchical Clustering

다변량 통계분석

Problems of Prototype – Based Clustering



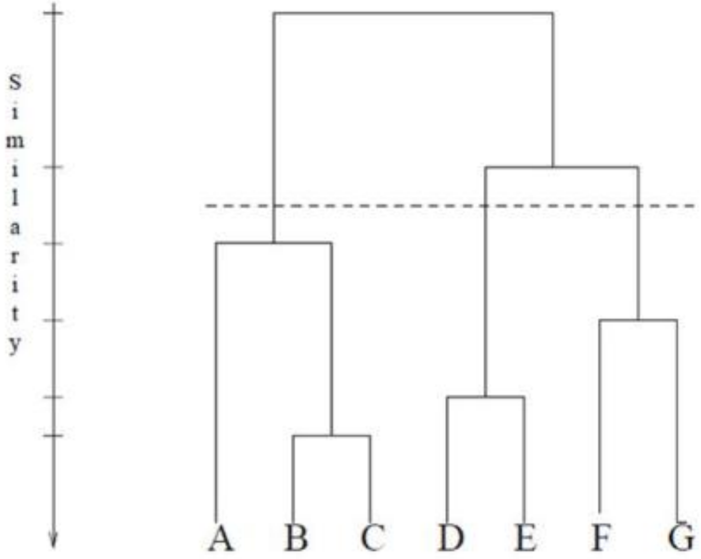


Hierarchical Clustering

동물

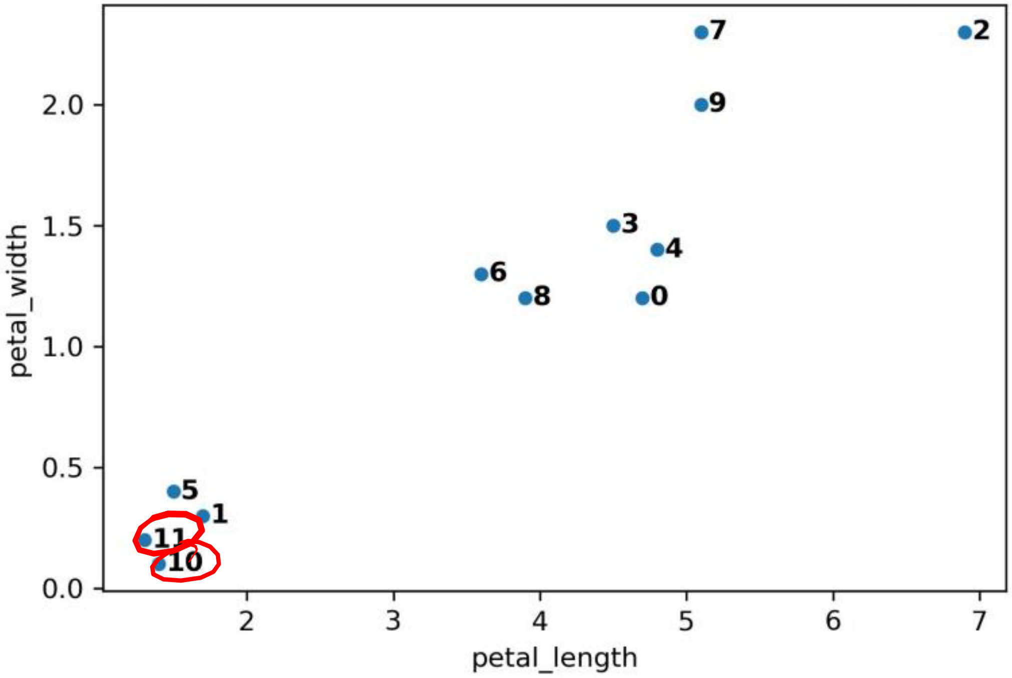
개

진돗개, 셰퍼드, 요크셔테리어, 푸들, 물소, 젖소 민들레

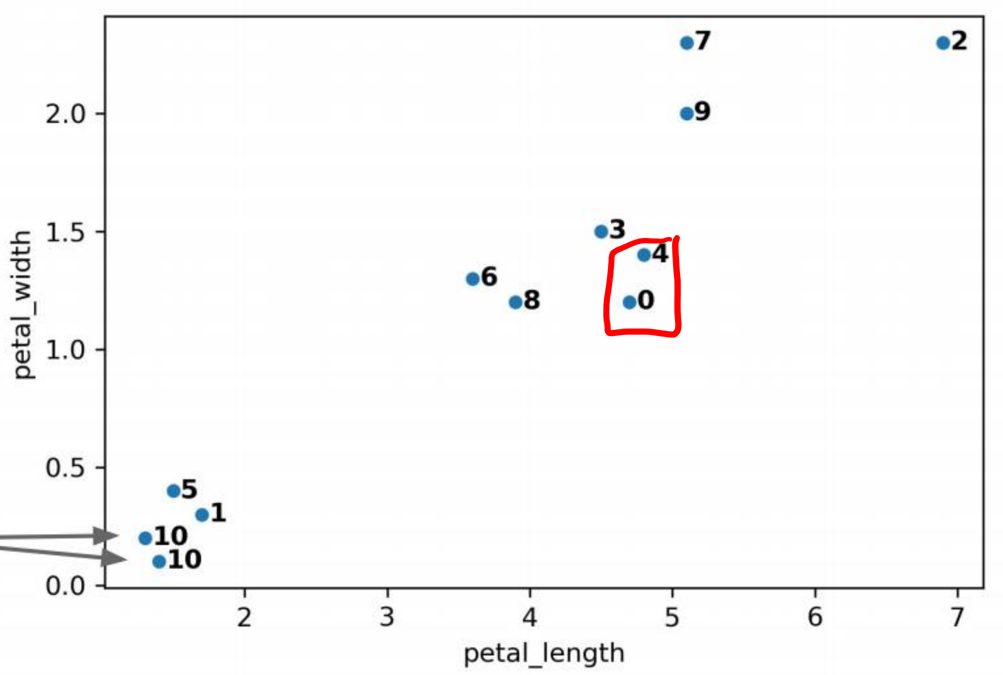


- Direction
  - Agglomerative Clustering (Bottom-Up Approach)
  - Divisive Clustering
- Distance
  - Complete-Linkage Clustering
  - Single-Linkage Clustering
  - Average-Linkage Clustering

Agglomerative Clustering

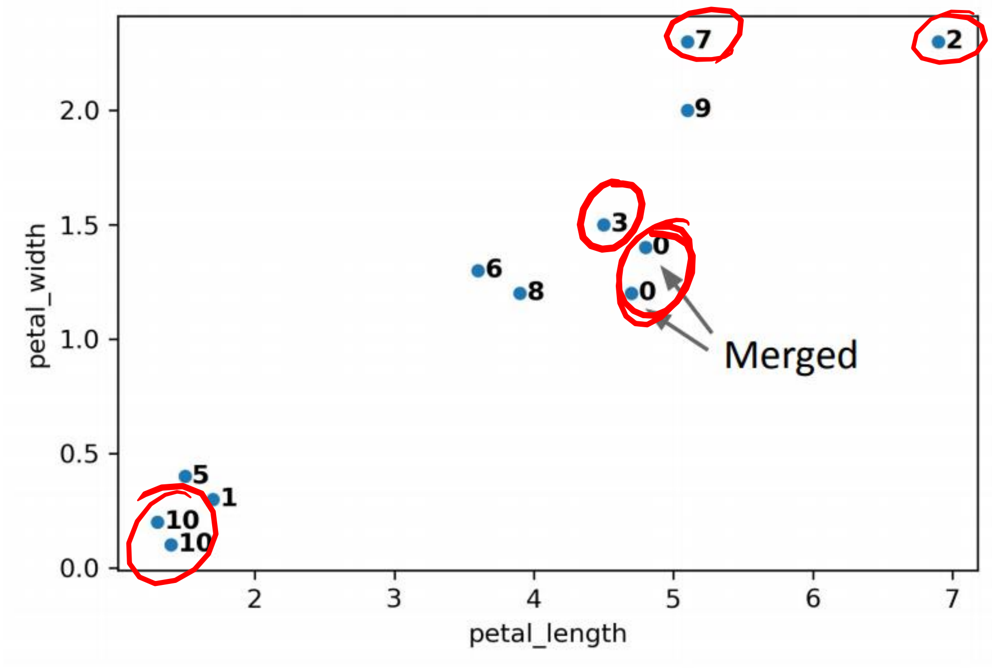


Agglomerative Clustering

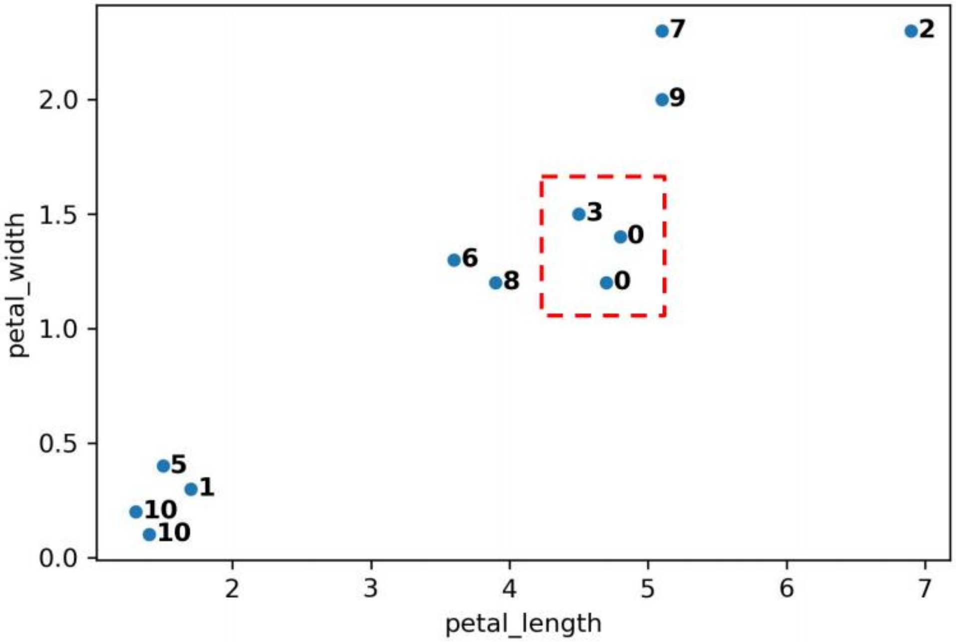


Two points in
the same cluster

Agglomerative Clustering



Agglomerative Clustering



Agglomerative Clustering

Cluster A: $\boldsymbol{x}_1, \dots, \boldsymbol{x}_{n_A}$
Clusters B : $\boldsymbol{z}_1, \dots, \boldsymbol{z}_{n_B}$

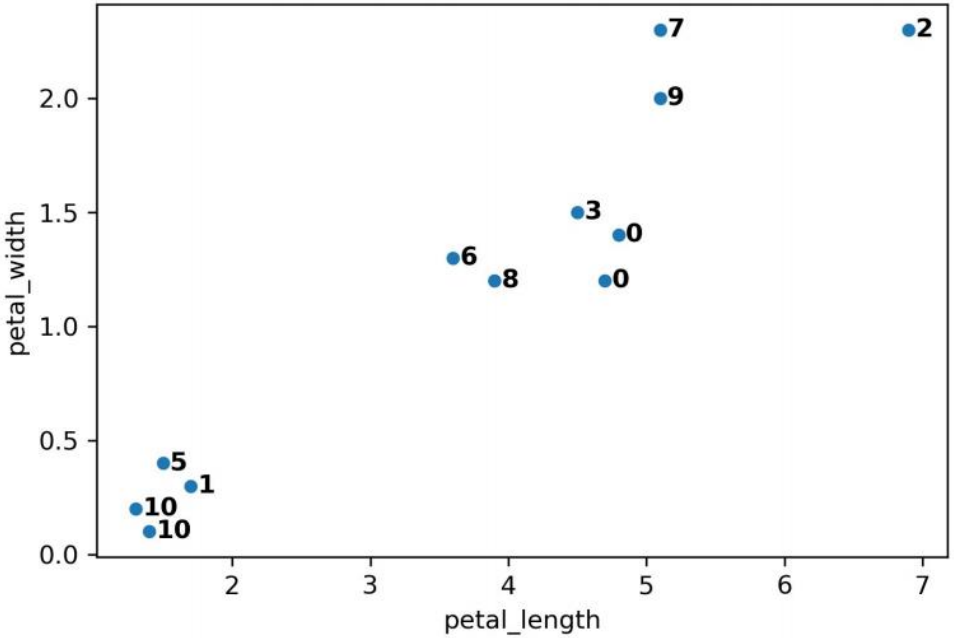
- Distance between Clusters

$\boxed{Complete\ - Linkage\ Clustering}$ $\quad \max\{d(\boldsymbol{x}_i, z_j);\ i = 1, \dots, n_A, j = 1, \dots, n_B\}$

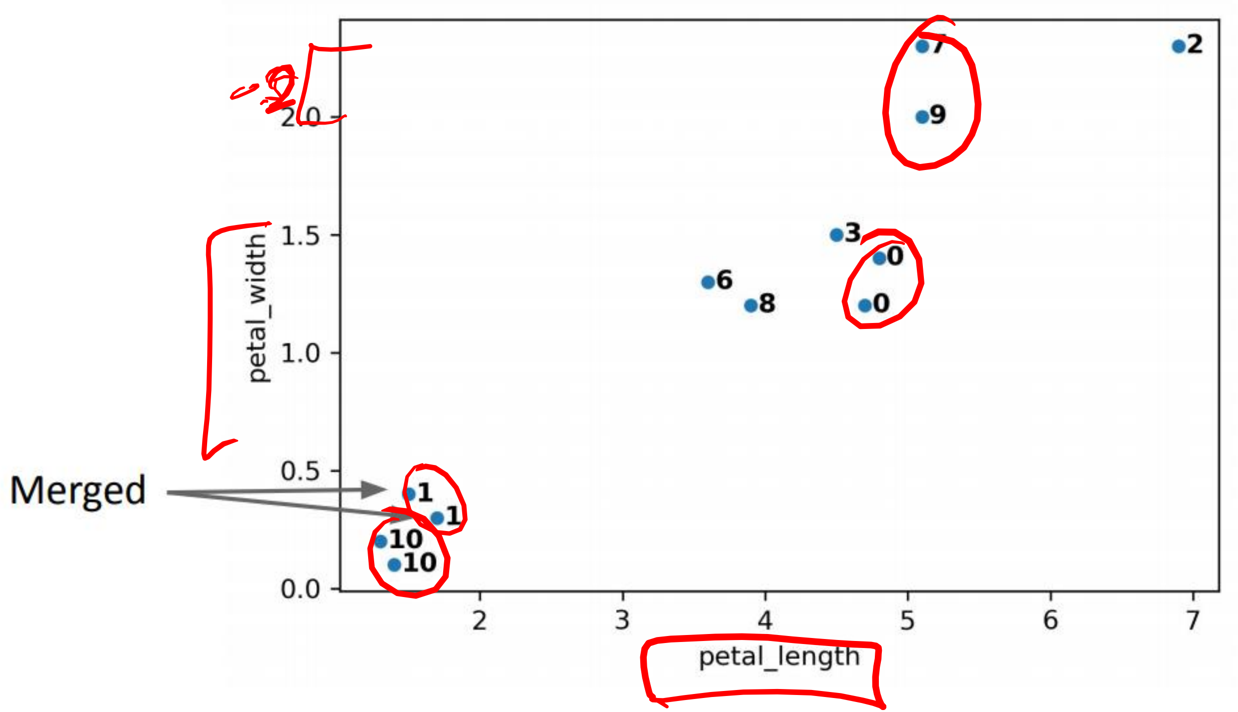
$Single\ - Linkage\ Clustering$ $\quad \min\{d(\boldsymbol{x}_i, z_j);\ i = 1, \dots, n_A, j = 1, \dots, n_B\}$

$Average\ - Linkage\ Clustering$ $\quad \dfrac{1}{n_A n_B} \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} d(\boldsymbol{x}_i, z_j)$

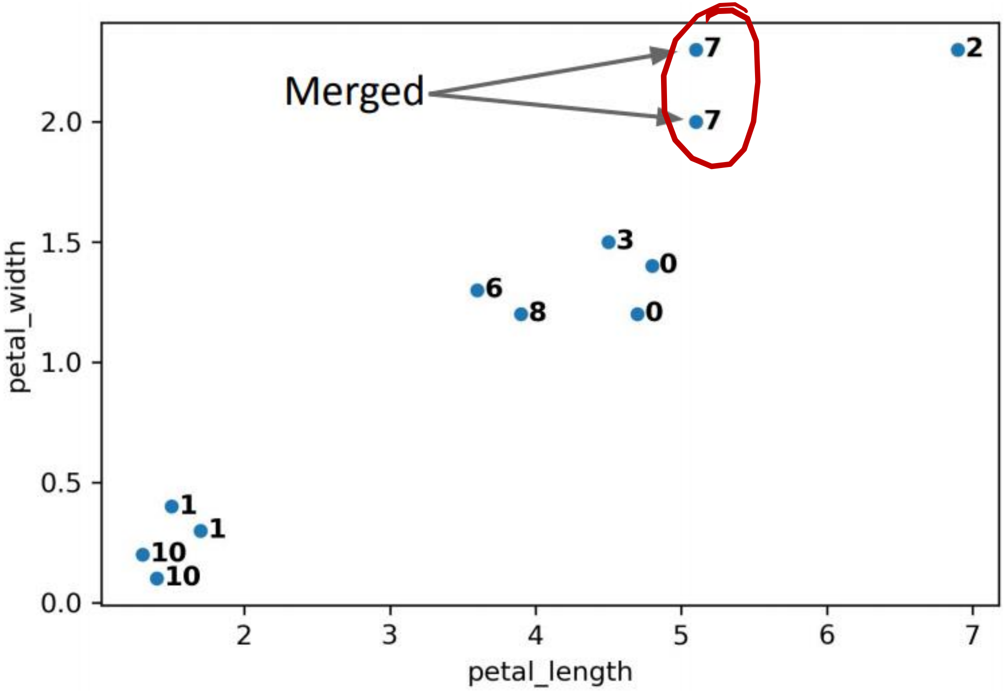
Agglomerative Clustering



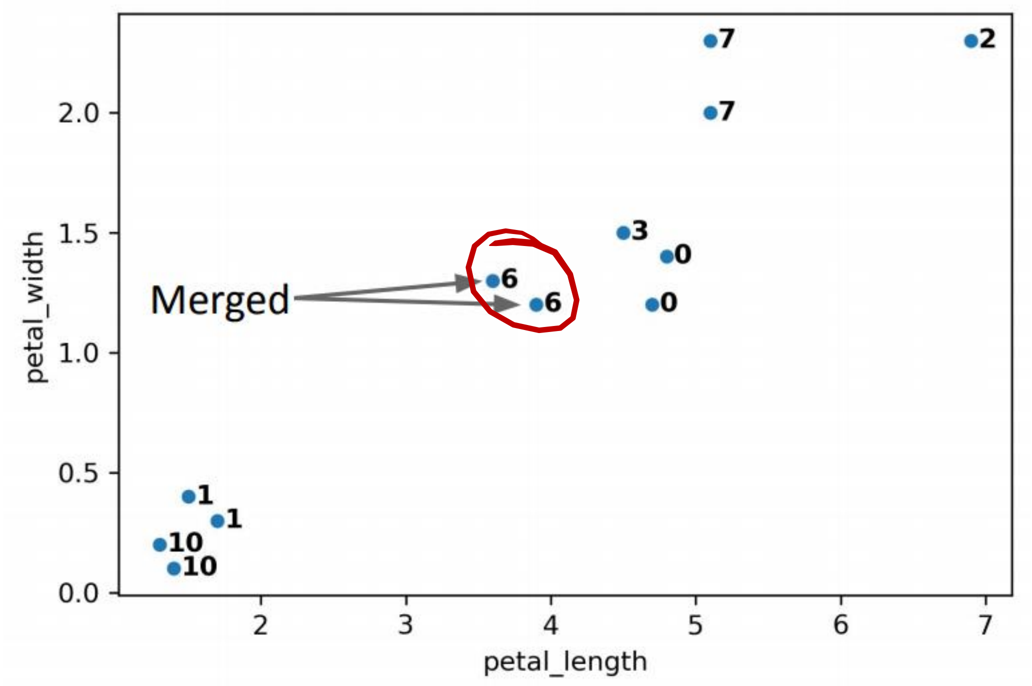
Agglomerative Clustering



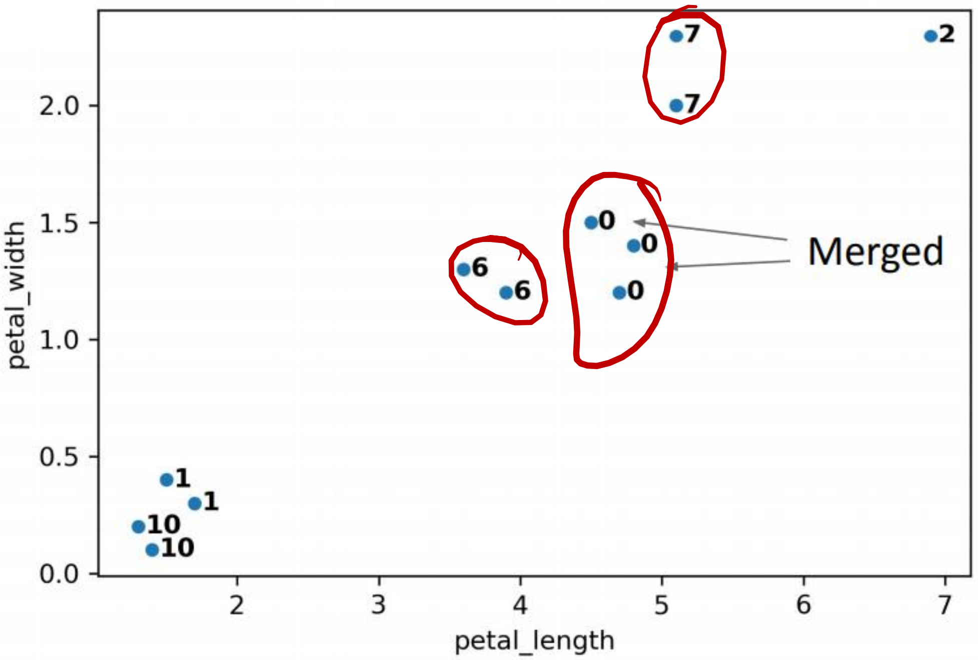
Agglomerative Clustering



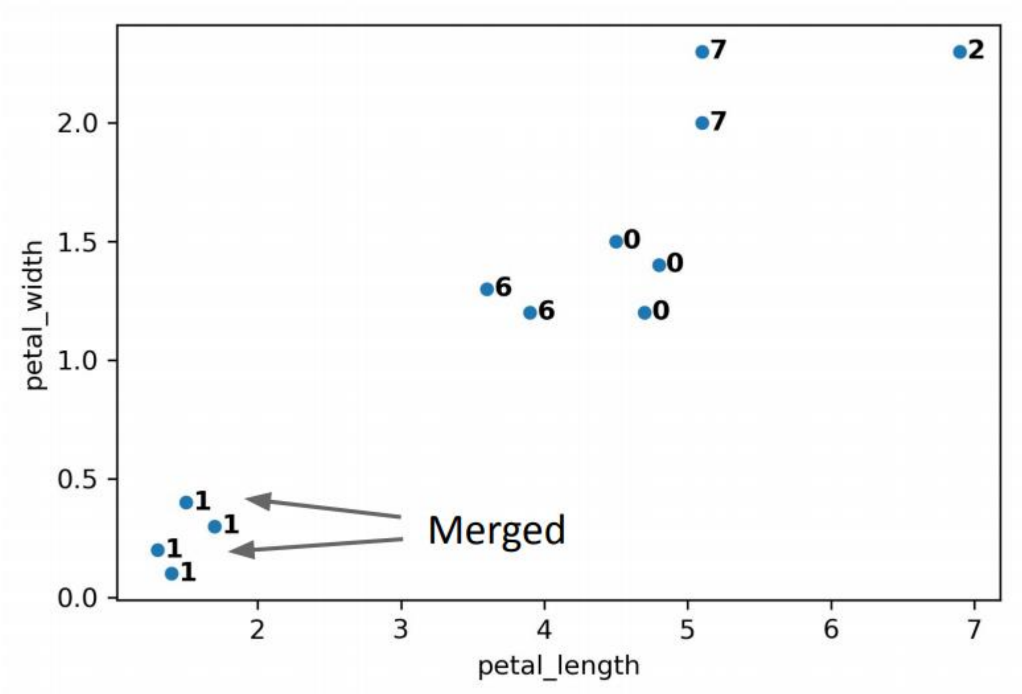
Agglomerative Clustering



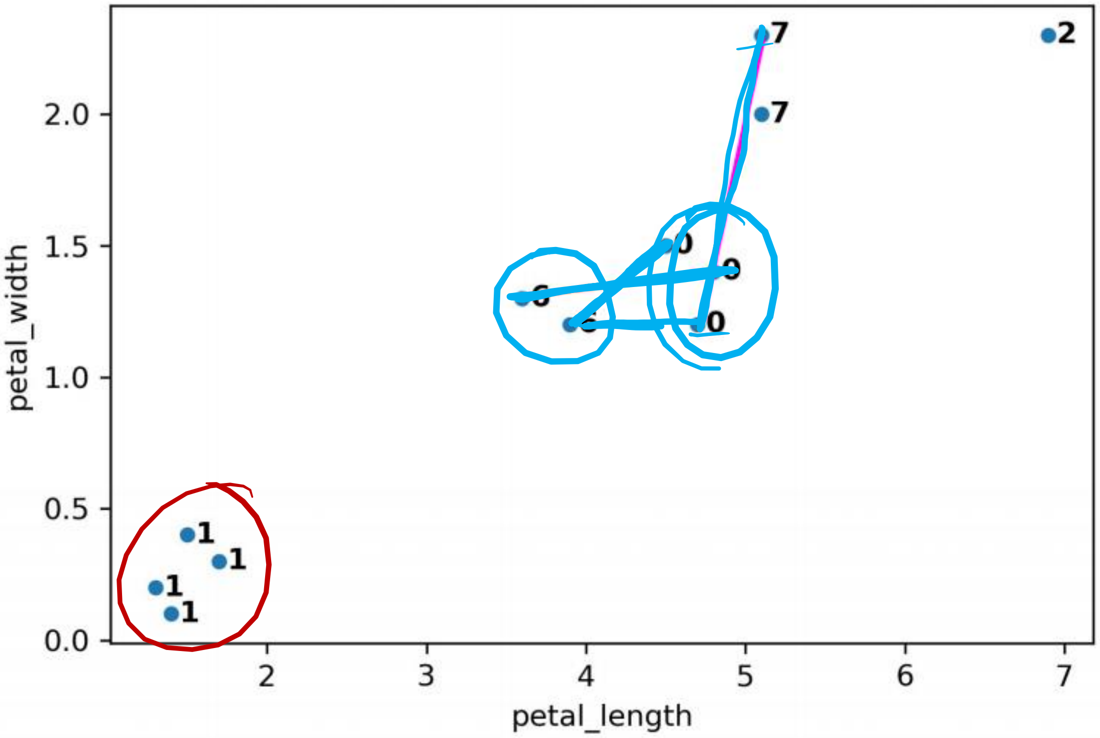
Agglomerative Clustering



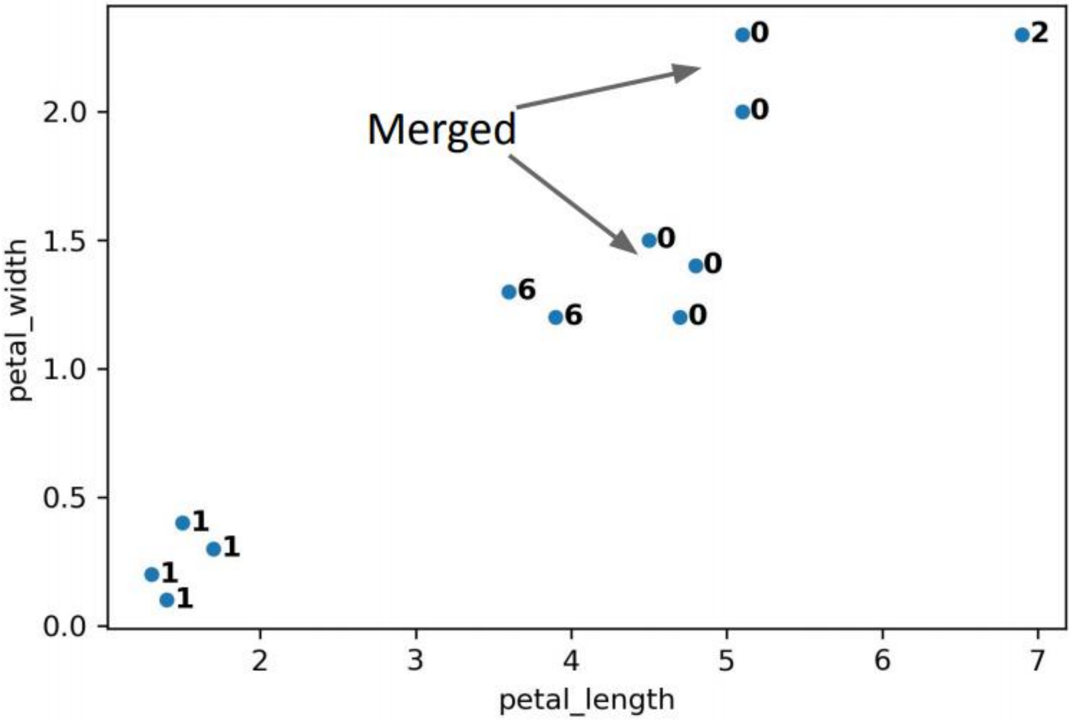
Agglomerative Clustering



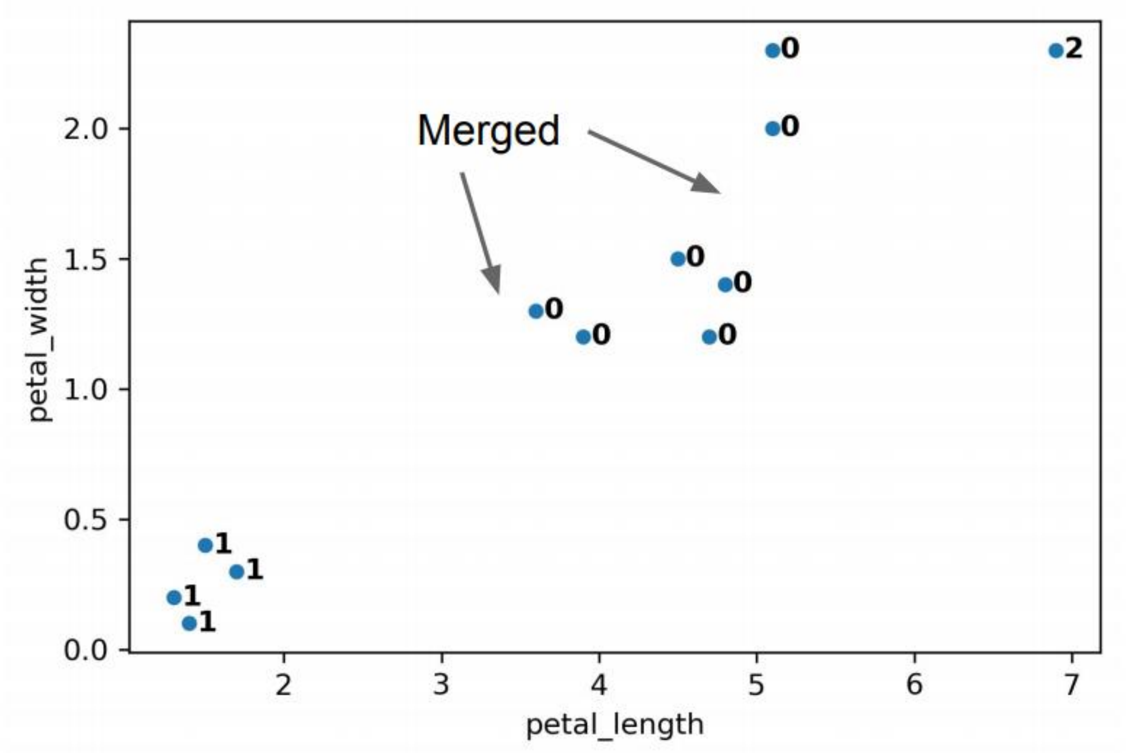
Agglomerative Clustering



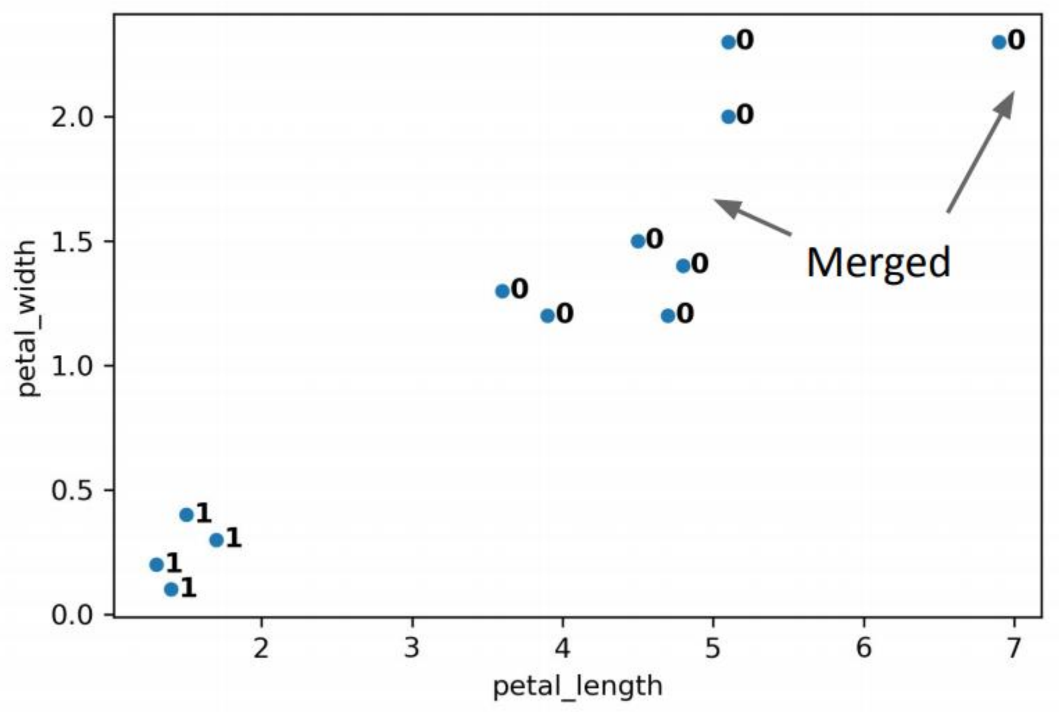
Agglomerative Clustering



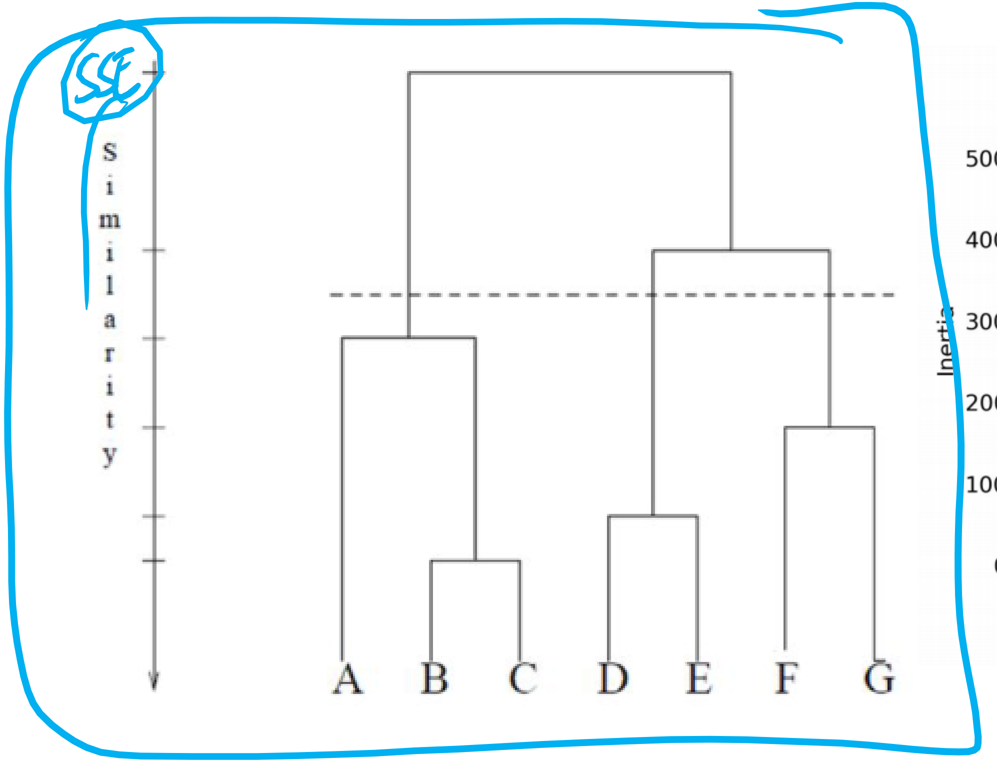
Agglomerative Clustering

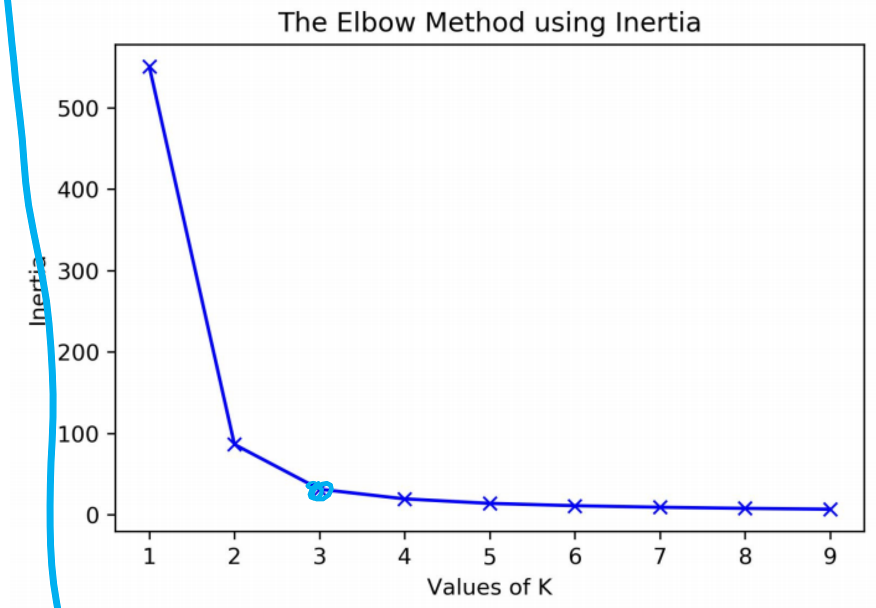


Agglomerative Clustering

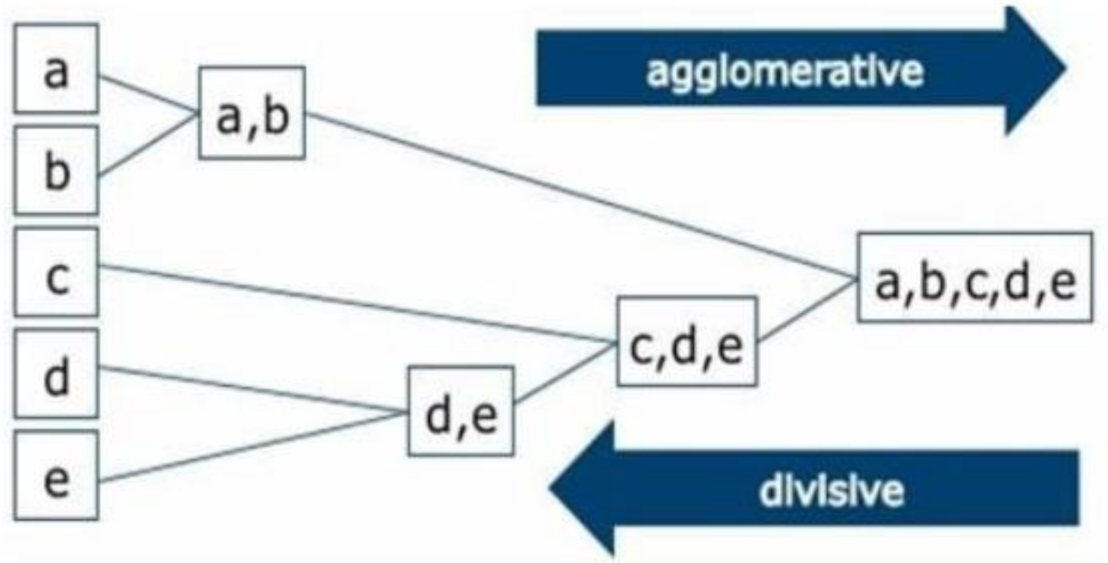


Agglomerative Clustering





Agglomerative Clustering and Divisive Clustering

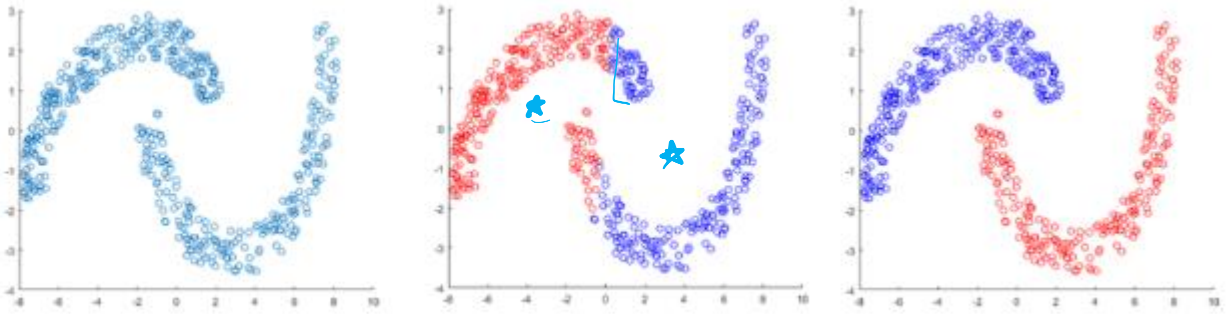


Downfalls of Hierarchical Clustering

- Noise와 Outlier에 민감하다 (Single Linkage)
- 한번 두 cluster을 묶으면 그 결정을 되돌릴 수 없다.

Density - based Clustering

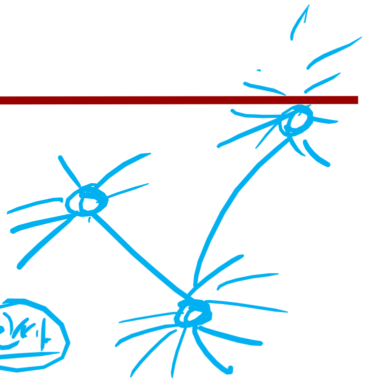
Limitations of Distance-based Clustering



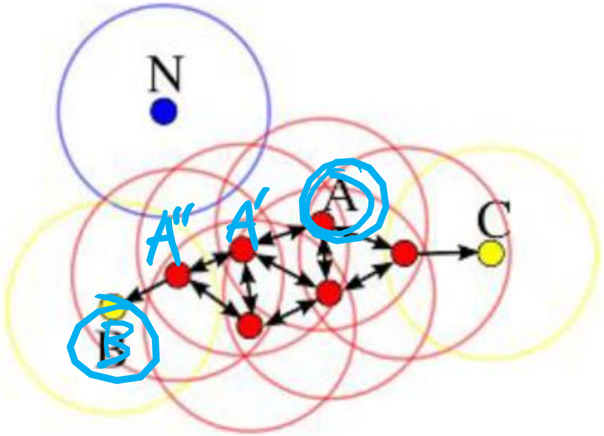
(a) 원본 데이터 (b) k-means clustering의 결과 (c) DBSCAN의 결과

DBSCAN: Density-Based Spatial Clustering of Approximations with Noise

노이즈

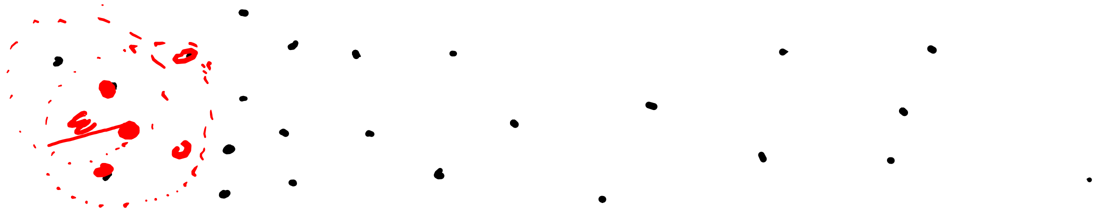


- Cluster: 접근 가능 관계로 유도해 낸 최대의 밀도 연결 샘플 집합
- Hyperparameter
  - M: Minimum # of samples in a neighborhood for core point
  - $\epsilon - Nearest\ Neighbors$: x와의 거리가 ϵ 보다 작은 샘플들



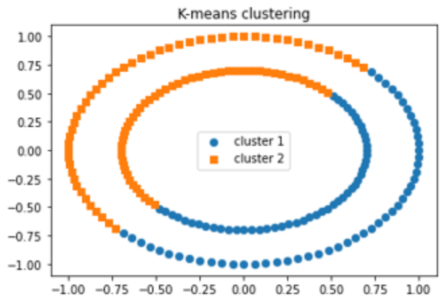
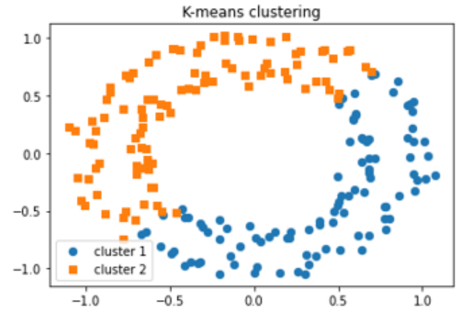
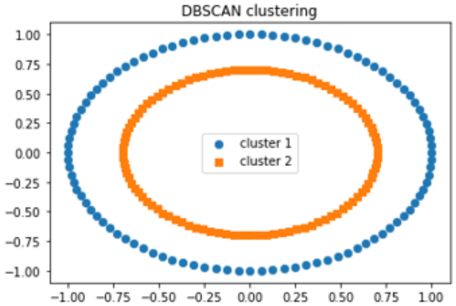
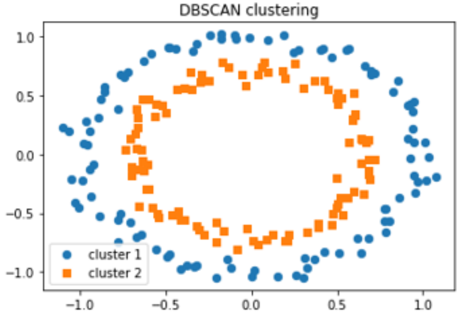
DBSCAN: Density-Based Spatial Clustering of Approximations with Noise

1. 임의 관측치를 선택하고, 군집 1 부여
2. 이 관측치의 $\epsilon - NN$을 구하고, $\epsilon - NN \ < M$이면 Noise나 Outlier로 표시 (아싸)
3. $\epsilon - NN \ \geq M$이면 모두에게 군집 1을 할당
4. 새로 군집 1이 부여된 NN에 대하여 그들의 $\epsilon - NN$ 이 M 보다 크면 이들 $\epsilon - NN$ 도 군집 1을 할당한다.
5. 군집 1의 어느 관측치도 M개 이상의 $\epsilon - NN$가 존재하지 않을 때까지 반복한다.
6. 군집 2를 형성하기 위해 (2) ~ (5) 반복
7. 모든 관측치가 군집 소속으로 분류되거나 잡음으로 분류될 때까지 반복



M=3 M..2

DBSCAN: Density-Based Spatial Clustering of Approximations with Noise

	Without Noise	With Noise
K-MEANS++		
DBSCAN		

Other ~~Prototype~~ Density Based Clustering

- HDBSCAN: Hierchical DBSCAN)
- OPTICS
- DENCLUE

DBSCAN & HDBSCAN