Chap.5 군집화 (Clustering)

방 수 식 교수
(bang@tukorea.ac.kr)

한국공학대학교 전자공학부

2024년도 1학기
머신러닝실습 & 인공지능설계실습1

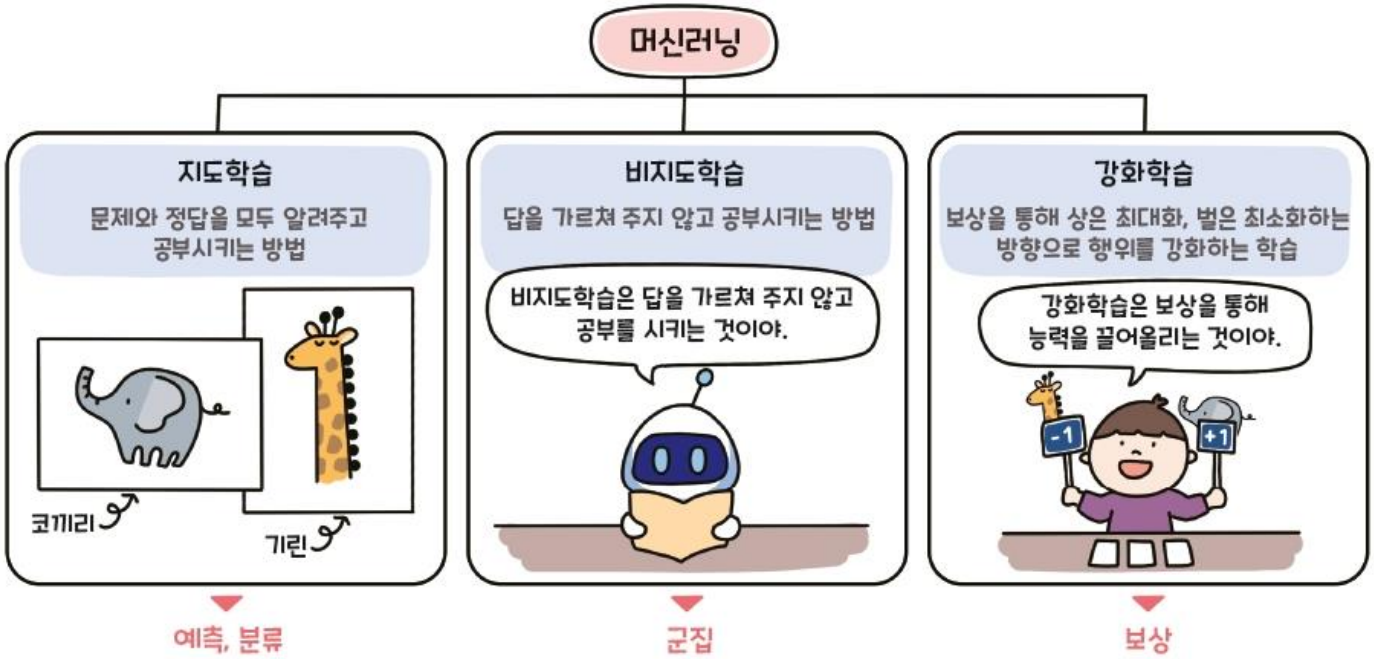
머신러닝의 분류

- **머신러닝의 분류**

  - **지도학습(Supervised Learning)** : 예측이나 분류를 위해 사용

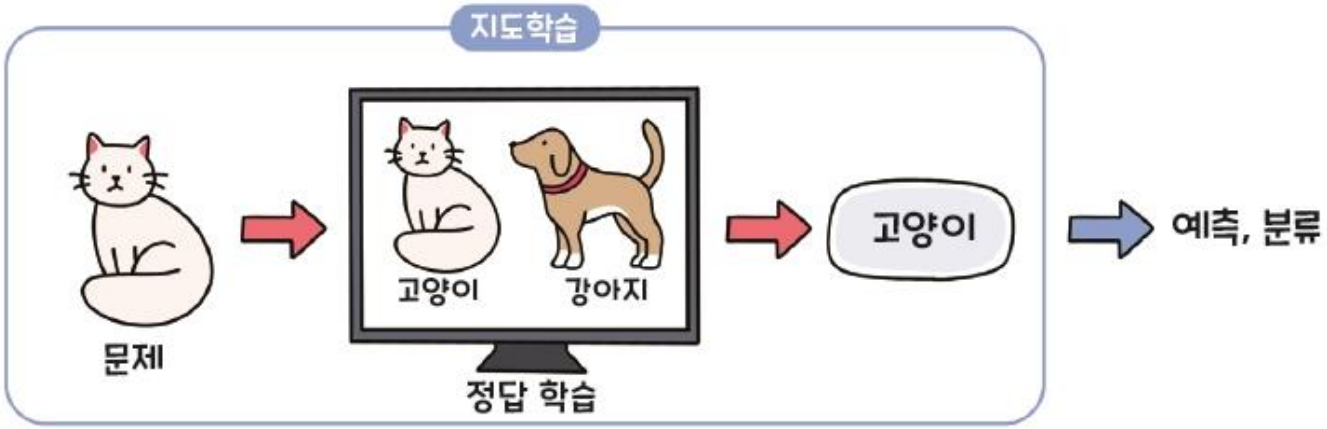
  - **비지도학습(Unsupervised Learning)** : 군집을 위해 사용

  - **강화학습(Reinforcement Learning)** : 환경에서 취하는 행동에 대한 보상을 이용하여 학습을 진행



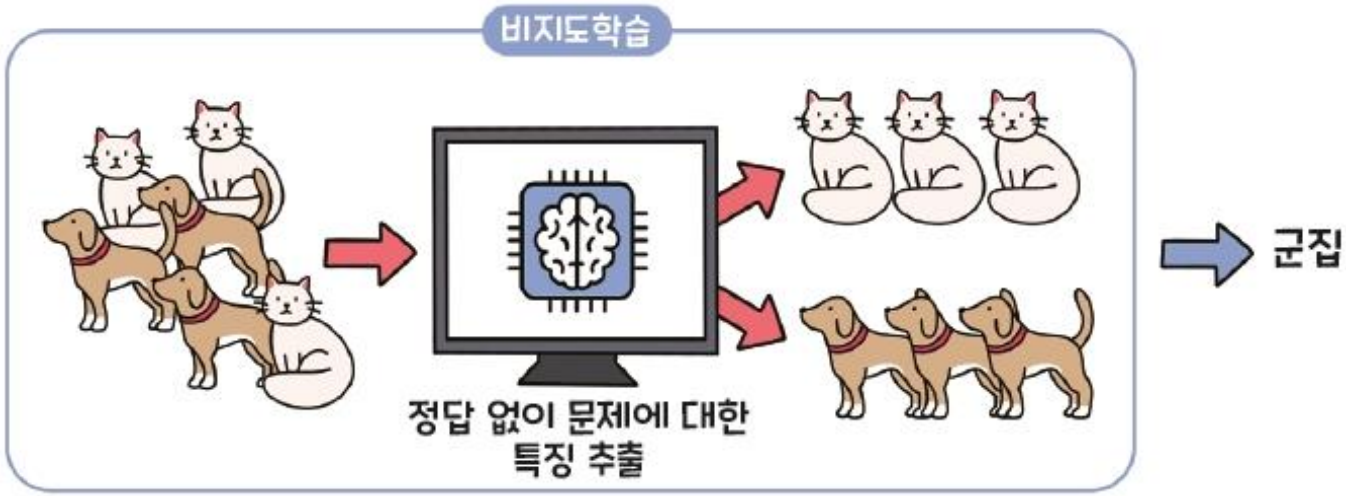
지도학습(Supervised Learning)

- **지도학습(Supervised Learning)**
  - 문제와 답을 함께 학습함으로써 미지의 문제에 대한 올바른 답을 예측하는 학습
  - 지도학습에서 사용하는 모델로는 크게 Regression과 Classification이 있음



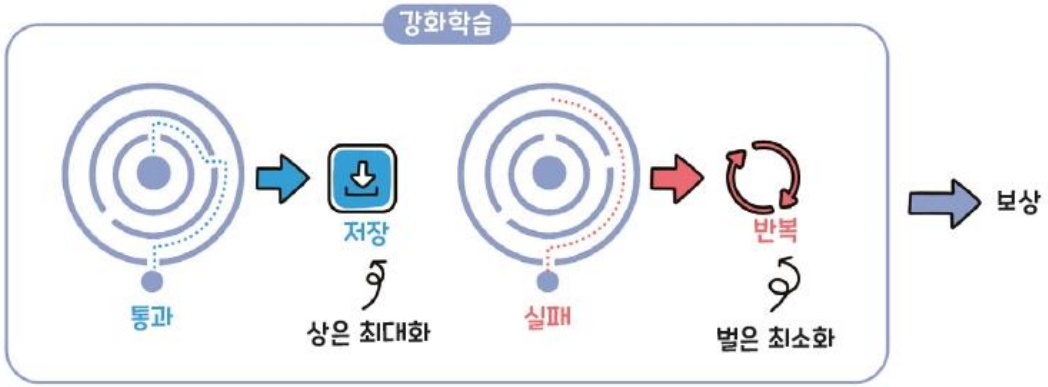
비지도학습(Unsupervised Learning)

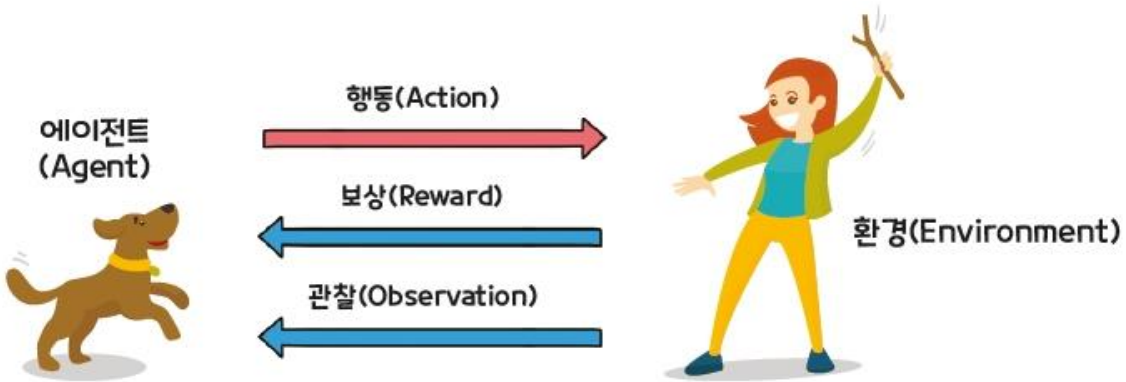
- **비지도학습(Unsupervised Learning)**
  - 지도학습과 다르게 조력자의 도움 없이 컴퓨터 스스로 학습하는 형태
  - 컴퓨터가 훈련 데이터를 이용하여 데이터들 간의 규칙성을 찾음



강화 학습(Reinforcement Learning)

- **강화 학습(Reinforcement Learning)**
  - 자신이 한 행동에 대해 보상(Reward)을 받으며 학습하는 것
  - 컴퓨터가 주어진 상태에 대해 최적의 행동을 선택하도록 학습하는 방법

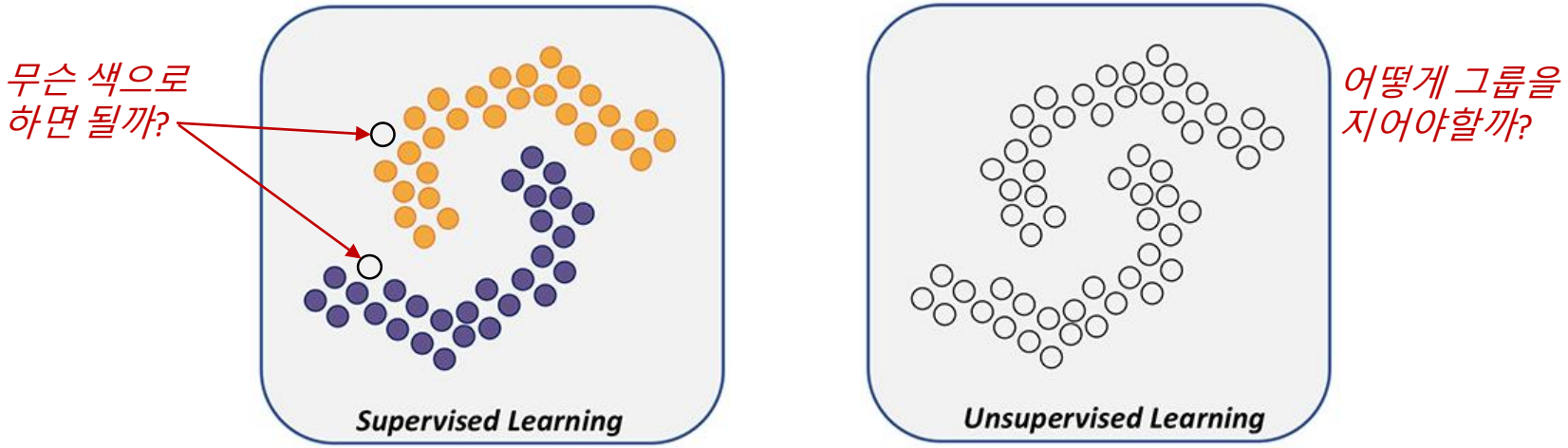




비지도학습(Unsupervised Learning)

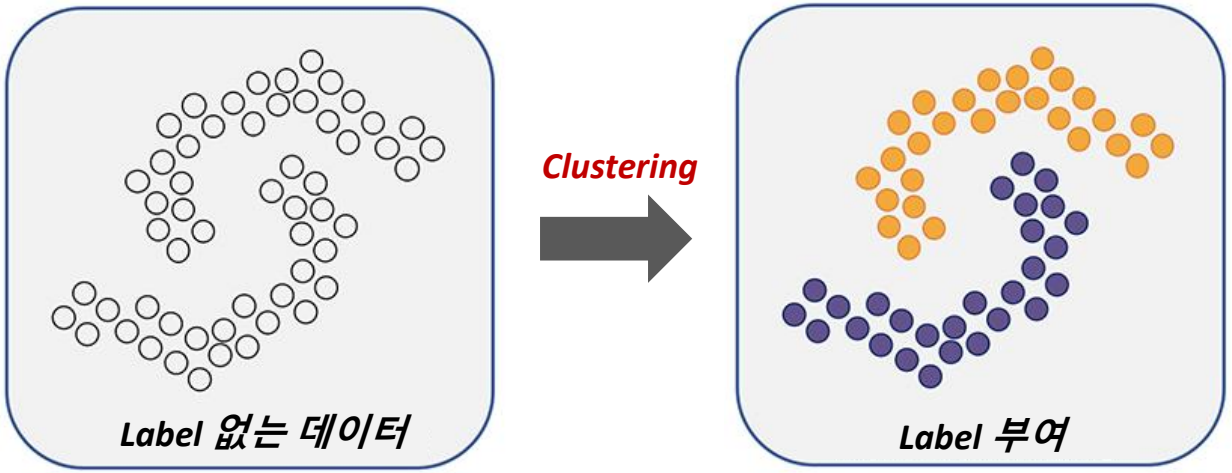
- **지도학습 VS 비지도학습**
  - **지도학습**: x(입력 데이터)와 y(지도학습에서 Target)의 관계를 파악
  - **비지도학습**: x(입력 데이터) 간의 관계 및 차이점을 스스로 파악함
  - **지도학습 vs 비지도학습**: y(레이블, Target)의 존재 유무 차이
  - 비지도학습에서 사용하는 대표적 모델로는 **군집화(Clustering)**가 있음

구분	지도학습	비지도학습
필요한 데이터 종류	x(학습 데이터), y(레이블)	x(학습 데이터)



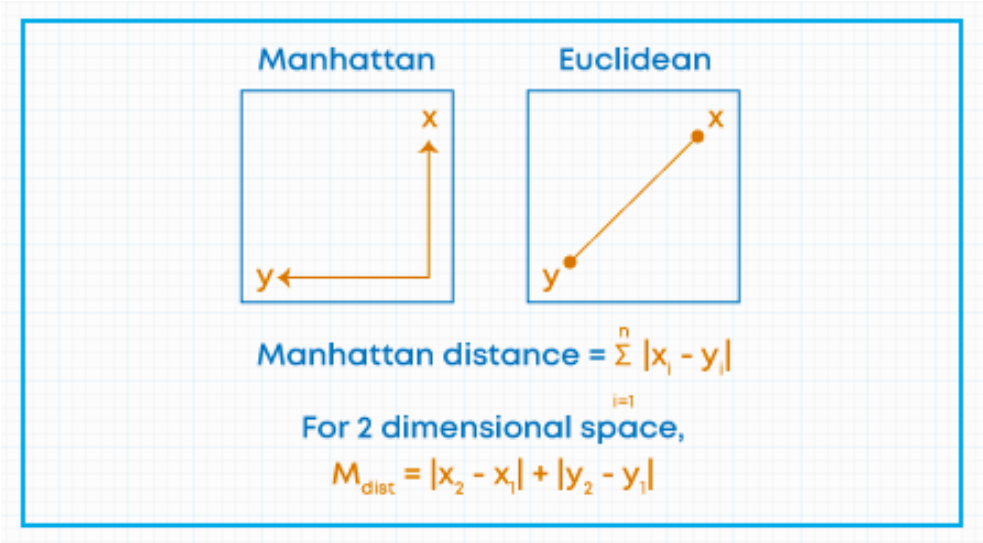
군집화(Clustering)

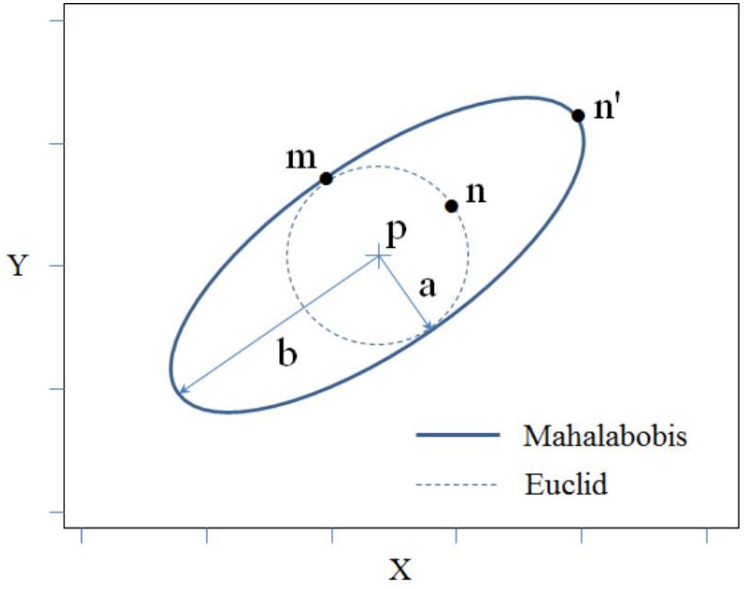
- **군집화(Clustering, 클러스터링)**
  - **군집(Cluster, 클러스터)**
    - 비슷한 특징을 가진 데이터들의 집단
  - **군집화(Clustering, 클러스터링)**
    - 데이터가 주어졌을 때 그 데이터들을 유사한 정도에 따라 군집으로 분류하는 것
    - Labeling이 되어 있지 않은 데이터에 Label을 부여하는 행위



Distance의 정의

- **다양한 Distance의 정의**
  - Manhattan(L1 Norm): 차원 간 거리의 총합으로 정의
  - Euclidean(L2 Norm): **일반적으로 가장 많이 사용**하는 정의
  - Mahalanobis: 타원 형태의 군집에서 사용하는 정의



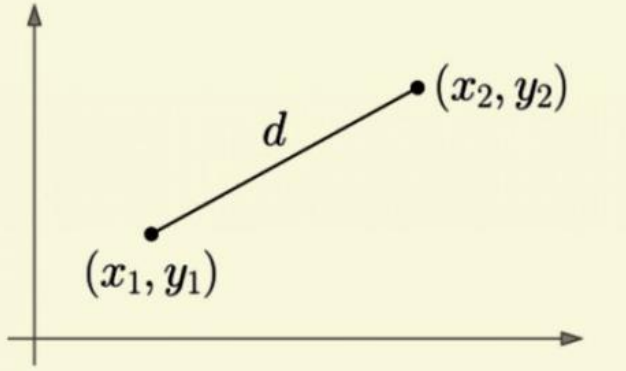


Distance의 정의

- **샘플 간의 거리**
  - 데이터 샘플을 하나의 벡터로 해석
  - 샘플 사이의 거리를 측정하는 객관적 방법 필요
  - $dist(\boldsymbol{x}_i, \boldsymbol{x}_j)$: 두 벡터 사이의 거리

    - $\boldsymbol{x}_i = \left[x_{i0}\ x_{i1}\ \cdots\ x_{i(N-1)}\right]^T$
    - $\boldsymbol{x}_j = \left[x_{j0}\ x_{j1}\ \cdots\ x_{j(N-1)}\right]^T$

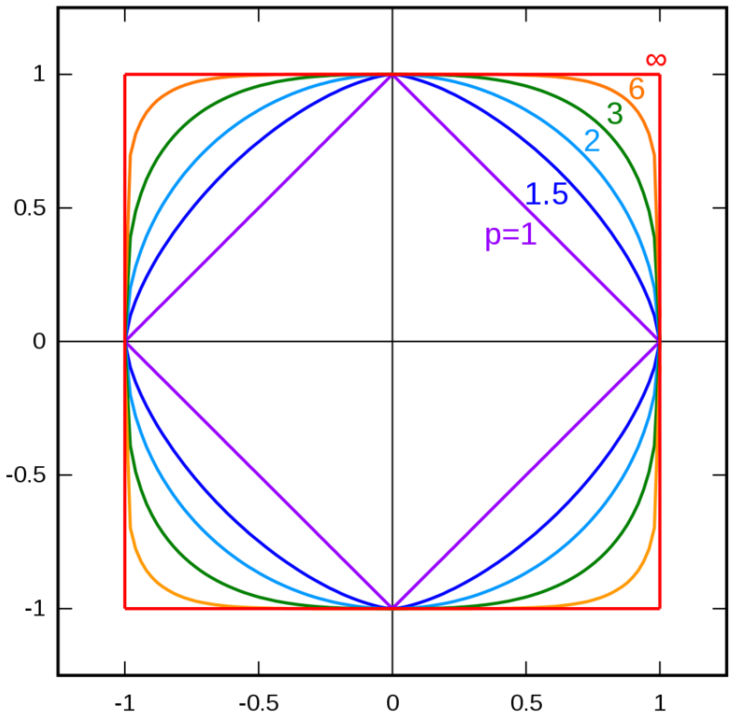
$$Distance = \sqrt{\sum_{u=0}^{N-1} |x_{iu} - x_{ju}|^2}$$



$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

L1 Norm vs L2 Norm

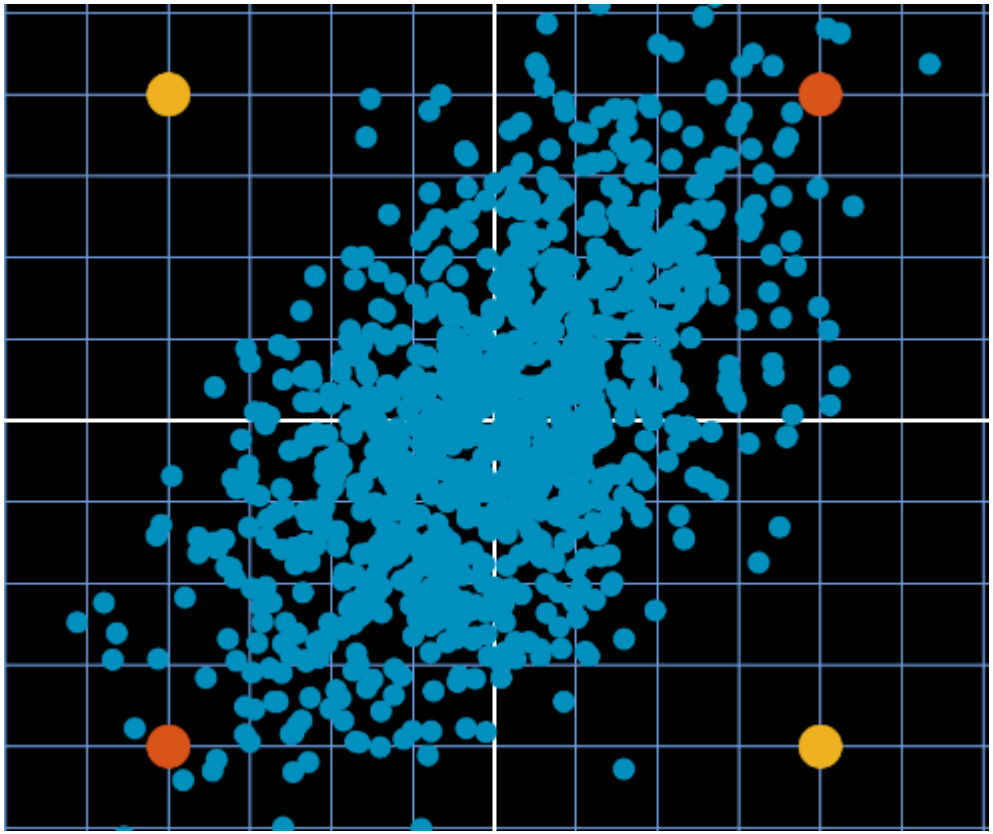
- **L1 Norm vs L2 Norm**
  - L1 Norm: $\|\boldsymbol{v}\|_1 = |v_1| + |v_2| + \cdots + |v_n|$
  - L2 Norm: $\|\boldsymbol{v}\|_2 = \sqrt{v_1^2 + v_2^2 + \cdots + v_n^2}$
  - Lp Norm: $\|\boldsymbol{v}\|_p =$



Manhattan Distance, Euclidean Distance

샘플 간의 거리

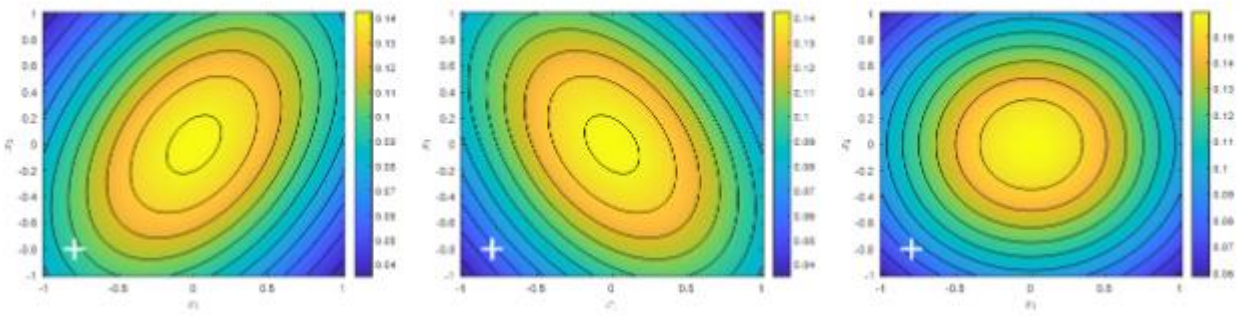
- **데이터 분포를 고려하였을 때,**
  - 데이터 분포 중심에서 노란점까지의 거리
  - 데이터 분포 중심에서 빨간점까지의 거리

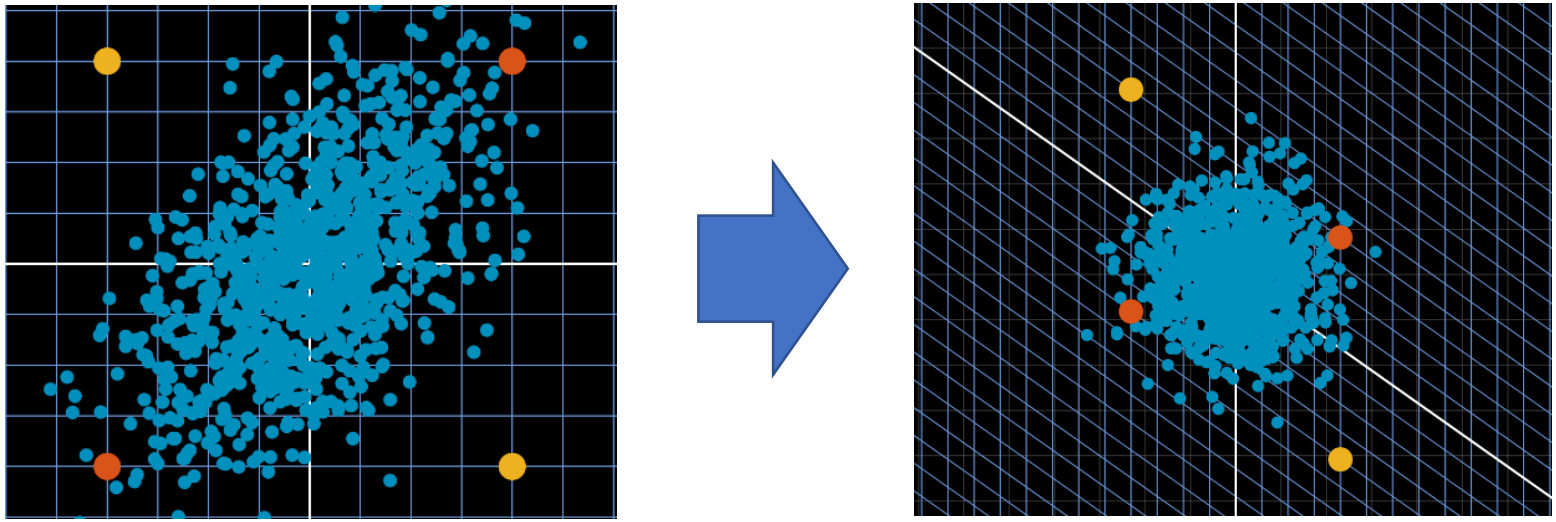


Mahalanobis Distance

- **Mahalanobis(마할라노비스) Distance**
  - 데이터 분포 방향 (공분산 행렬, Σ)을 활용하여, 거리를 보정

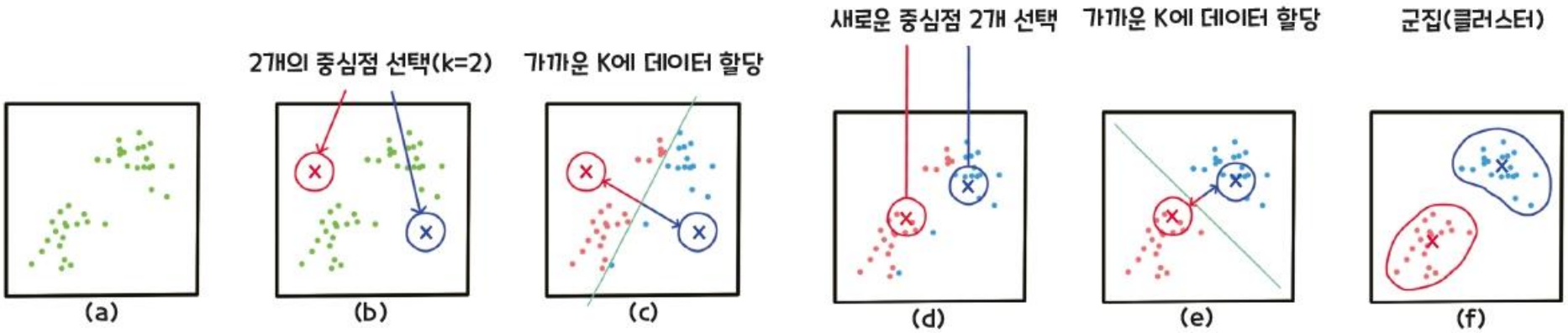
$\Sigma = \begin{bmatrix} 1.0 & 0.5 \\ 0.5 & 1.5 \end{bmatrix}$ $\Sigma = \begin{bmatrix} 1.0 & -0.5 \\ -0.5 & 1.5 \end{bmatrix}$ $\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$





K-평균 군집화(K-means Clustering)

k-평균 군집화(K-Means Clustering)



- (a) : 일반적인 데이터 분포입니다.
- (b) : 데이터셋에서 K개의 중심점을 임의로 지정하는데, 여기에서는 K=2의 값으로 중심점 2개를 설정했습니다.
- (c) : 데이터들을 가장 가까운 중심점에 할당합니다.
- (d) : (c)에서 할당된 결과를 바탕으로 중심점을 새롭게 지정합니다.
- (e) : 중심점이 더 이상 변하지 않을 때까지 (c)~(d) 과정을 반복합니다.
- (f) : 최종적인 군집이 형성됩니다.

몇 개의 군집 (k)으로 나눌지 설정하는 것이 핵심

K-평균 군집화(K-means Clustering)

- **Step 1) K를 선택하고, 랜덤으로 초기 중심 선택**

번호	밀도	당도	번호	밀도	당도	번호	밀도	당도
1	0.697	0.460	11	0.245	0.057	21	0.748	0.232
2	0.774	0.376	12	0.343	0.099	22	0.714	0.346
3	0.634	0.264	13	0.639	0.161	23	0.483	0.312
4	0.608	0.318	14	0.657	0.198	24	0.478	0.437
5	0.556	0.215	15	0.360	0.370	25	0.525	0.369
6	0.403	0.237	16	0.593	0.042	26	0.751	0.489
7	0.481	0.149	17	0.719	0.103	27	0.532	0.472
8	0.437	0.211	18	0.359	0.188	28	0.473	0.376
9	0.666	0.091	19	0.339	0.241	29	0.725	0.445
10	0.243	0.267	20	0.282	0.257	30	0.446	0.459

$$\mu_1 = (0.403, 0.237)$$

$$\mu_2 = (0.343, 0.099)$$

$$\mu_3 = (0.478, 0.437)$$

K-평균 군집화(K-means Clustering)

- **Step 2) 각 데이터 샘플마다 중심과의 거리를 계산하여 가장 가까운 중심점에 대한 군집으로 할당**

번호	밀도	당도	번호	밀도	당도	번호	밀도	당도
1	0.697	0.460	11	0.245	0.057	21	0.748	0.232
2	0.774	0.376	12	0.343	0.099	22	0.714	0.346
3	0.634	0.264	13	0.639	0.161	23	0.483	0.312
4	0.608	0.318	14	0.657	0.198	24	0.478	0.437
5	0.556	0.215	15	0.360	0.370	25	0.525	0.369
6	0.403	0.237	16	0.593	0.042	26	0.751	0.489
7	0.481	0.149	17	0.719	0.103	27	0.532	0.472
8	0.437	0.211	18	0.359	0.188	28	0.473	0.376
9	0.666	0.091	19	0.339	0.241	29	0.725	0.445
10	0.243	0.267	20	0.282	0.257	30	0.446	0.459

$$\|x_1 - \mu_1\|_2 = 0.369$$

$$\|x_1 - \mu_2\|_2 = 0.506$$

$$\boxed{\|x_1 - \mu_3\|_2 = 0.220}$$

1번 샘플은 3번 클러스터 소속

$$C_1 = \{x_3, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{13}, x_{14}, x_{17}, x_{18}, x_{19}, x_{20}, x_{23}\}$$

$$C_2 = \{x_{11}, x_{12}, x_{16}\}$$

$$C_3 = \{x_1, x_2, x_4, x_{15}, x_{21}, x_{22}, x_{24}, x_{25}, x_{26}, x_{27}, x_{28}, x_{29}, x_{30}\}$$

K-평균 군집화(K-means Clustering)

- **Step 3) 각 군집에 대한 새로운 중심점을 계산하여, step 2 반복**

각 샘플에 대한 평균

번호	밀도	당도	번호	밀도	당도	번호	밀도	당도
1	0.697	0.460	11	0.245	0.057	21	0.748	0.232
2	0.774	0.376	12	0.343	0.099	22	0.714	0.346
3	0.634	0.264	13	0.639	0.161	23	0.483	0.312
4	0.608	0.318	14	0.657	0.198	24	0.478	0.437
5	0.556	0.215	15	0.360	0.370	25	0.525	0.369
6	0.403	0.237	16	0.593	0.042	26	0.751	0.489
7	0.481	0.149	17	0.719	0.103	27	0.532	0.472
8	0.437	0.211	18	0.359	0.188	28	0.473	0.376
9	0.666	0.091	19	0.339	0.241	29	0.725	0.445
10	0.243	0.267	20	0.282	0.257	30	0.446	0.459

$$\boldsymbol{\mu}_1' = \frac{1}{|C_1|}\sum_{\boldsymbol{x}\in C_1}\boldsymbol{x} = (0.493, 0.207)$$

$$\boldsymbol{\mu}_2' = \frac{1}{|C_2|}\sum_{\boldsymbol{x}\in C_2}\boldsymbol{x} = (0.394, 0.066)$$

$$\boldsymbol{\mu}_3' = \frac{1}{|C_3|}\sum_{\boldsymbol{x}\in C_3}\boldsymbol{x} = (0.602, 0.396)$$

$$C_1 = \{x_3, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{13}, x_{14}, x_{17}, x_{18}, x_{19}, x_{20}, x_{23}\}$$

$$C_2 = \{x_{11}, x_{12}, x_{16}\}$$

$$C_3 = \{x_1, x_2, x_4, x_{15}, x_{21}, x_{22}, x_{24}, x_{25}, x_{26}, x_{27}, x_{28}, x_{29}, x_{30}\}$$

K-평균 군집화(K-means Clustering)

- **Step 4) step 2와 step 3을 반복하다가 더 이상 군집의 요소가 변하지 않을 경우 종료**

$$C_1 = \{x_3, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{13}, x_{14}, x_{17}, x_{18}, x_{19}, x_{20}, x_{23}\}$$

$$C_2 = \{x_{11}, x_{12}, x_{16}\}$$

$$C_3 = \{x_1, x_2, x_4, x_{15}, x_{21}, x_{22}, x_{24}, x_{25}, x_{26}, x_{27}, x_{28}, x_{29}, x_{30}\}$$

4회 반복 후 최종 클러스터링 결과

$$C_1 = \{x_5, x_6, x_7, x_9, x_{13}, x_{14}, x_{16}, x_{17}, x_{21}\}$$

$$C_2 = \{x_6, x_8, x_{10}, x_{11}, x_{12}, x_{15}, x_{18}, x_{19}, x_{20}\}$$

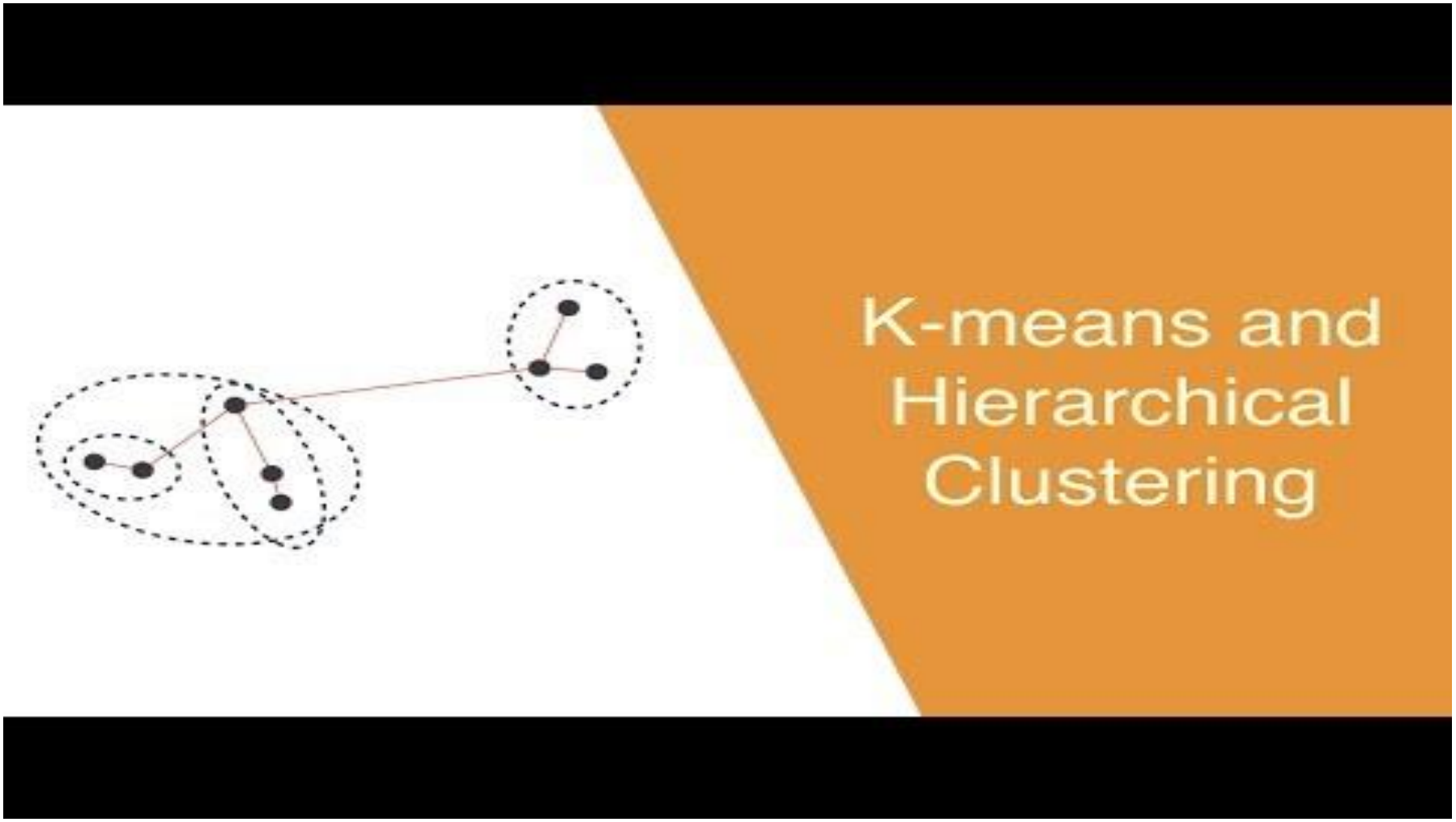
$$C_3 = \{x_1, x_2, x_4, x_{22}, x_{23}, x_{24}, x_{25}, x_{26}, x_{27}, x_{28}, x_{29}, x_{30}\}$$

K=3 이 아닌 경우에는 어떻게 될까?
K는 어떻게 설정하는 것이 좋을까?

K-평균 군집화(K-means Clustering)

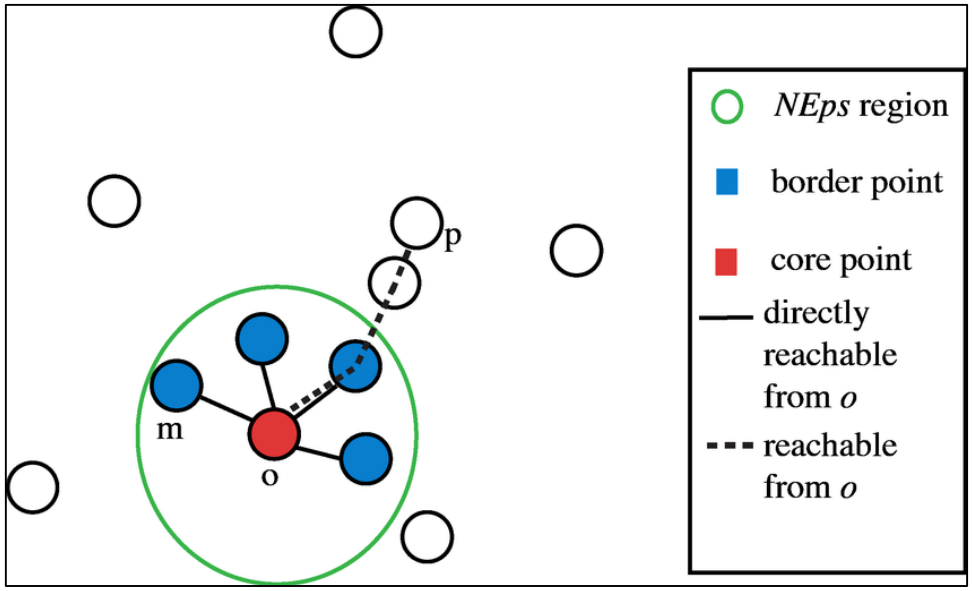
- **최적의 k 결정 방법론**
  - Elbow Method

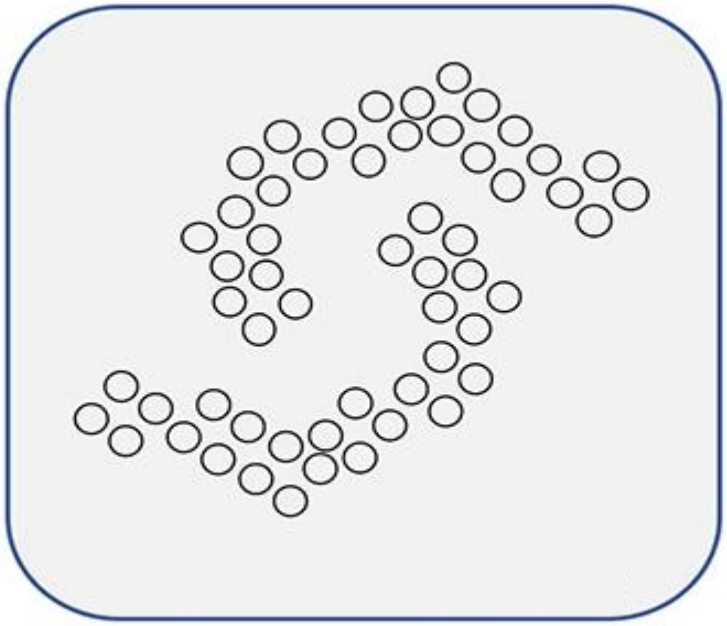
https://www.youtube.com/watch?v=QXOkPvFM6NU
: 7분 47초



밀도 기반 군집화(DBSCAN)

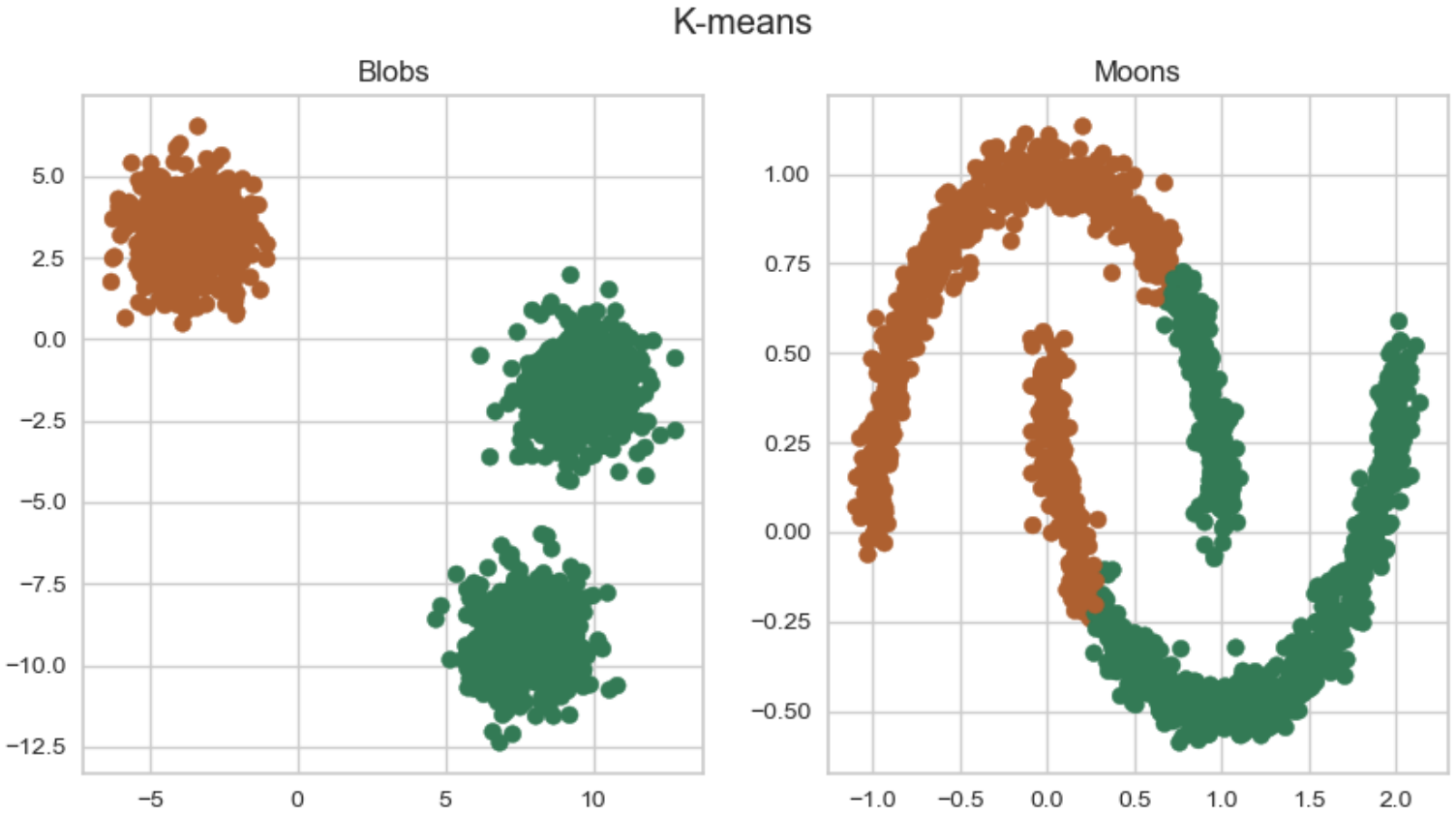
- **DBSCAN**
  - 클러스터 수를 지정할 필요가 없음.
  - K-평균 군집화가 찾을 수 없는 군집을 찾을 수 있음.
  - Hyper-parameter 설정이 매우 중요



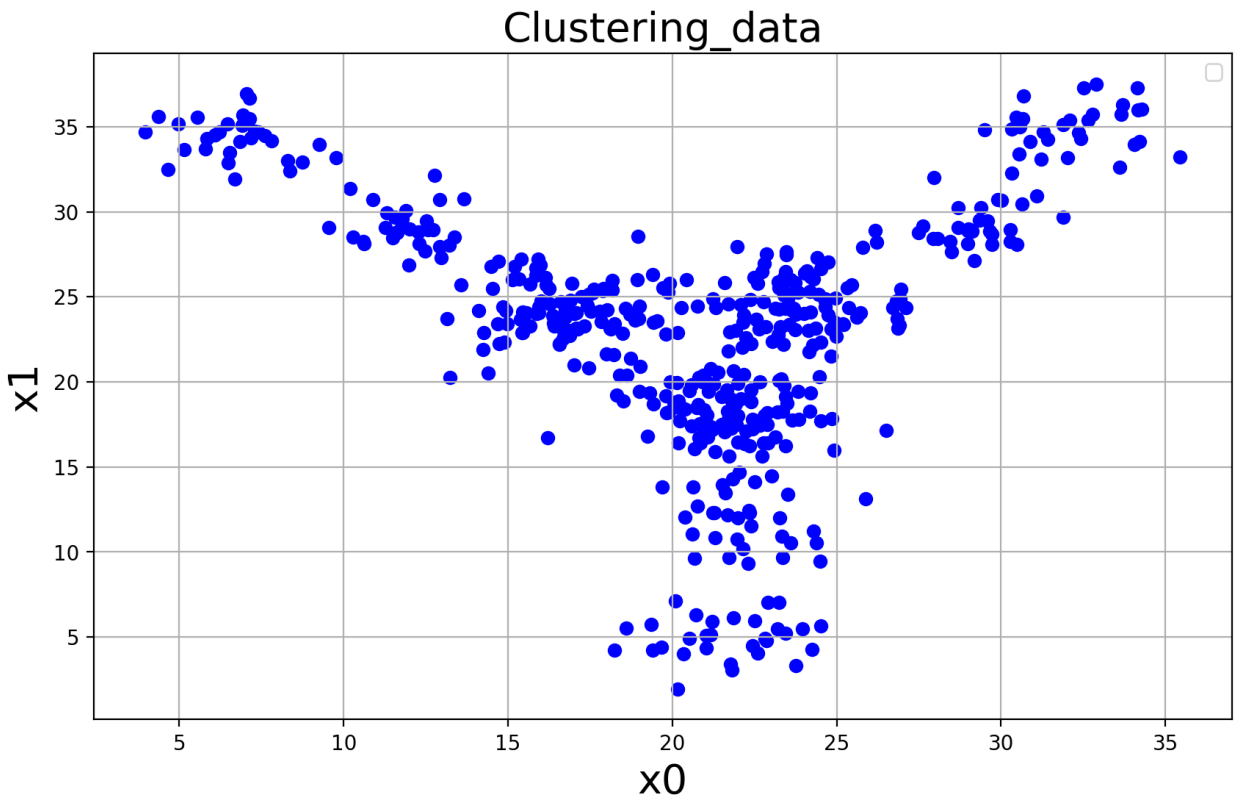


군집화(K-means Clustering) 실습

- **Sklearn 기반 실습**



Chap.5 실습 데이터



Target이 없는 데이터 총 480개의 데이터

Chap.5 K-means Clustering 실습

각 실습문제에 대해 code(+주석), 그래프, 분석 내용 등은 필수적으로 포함시킬 것

1) 초기 중심 선택

- "Clustering_data.csv" 데이터 불러오기 (첫번째 행은 header값)
- K = 3으로 선택하고, 불러온 데이터 중 랜덤으로 초기 중심을 선택
- 초기 중심과 불러온 데이터를 그래프로 출력

2) K-means Clustering 알고리즘 구현 (L2 Norm 이용)

- 본 강의자료 13 ~ 17 page를 참고하여 K-means Clustering을 사용자 지정함수로 구현
- 데이터, K를 함수의 입력으로 지정해 군집이 변하지 않으면 학습이 완료되는 함수로 구현 (K = 3)
- 최종 Clustering 결과 그래프로 출력

Pass 기준
1) 초기 중심과 불러온 데이터를 그래프로 출력
2) K=3일 때, 최종 Clustering 결과와 중심점을 한 그래프로 출력

3) 비교 분석

- K = 1~10일 때에 대한 초기 중심점에 따른 학습 결과 비교 분석 및 최적의 K 도출
- 본 강의자료 15 page step 2의 거리 계산 시 거리 종류에 대한 결과 비교 분석 (Euclidean vs Manhattan)