

Machine Learning

## 12장 군집 (Clustering)

---

고려대학교 통계학과  
박유성



# Contents

**01** Introduction

**02** K-means Clustering

**03** Hierarchical Clustering (계층적 군집)

**04** DBSCAN

# 01 Introduction

## ■ 군집 (Clustering)

- 목적: 특성변수의 유사성을 토대로 관측치를 2개 이상의 그룹으로 구별
- 목표변수의 관측값  $y_i$  (class)가 주어지지 않음. → “비지도 학습 (Unsupervised learning)”

## ■ 비유사성 (dissimilarity) 측도

- 관측치  $i$ 와  $i'$ 의 유사성은 특성변수의 관측값  $x_i$ 와  $x_{i'}$ 의 비유사성 측도에 의해 측정

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

## 02 K-means Clustering

- 개요

- 가장 널리 사용되는 군집방법
- 군집의 개수  $K$ 는 미리 정해주어야 하는 숫자임 → K-means 군집의 최대 약점

- Algorithm

- [Step 1] 각 특성변수의 자료타입에 따라 자료를 변환한 후 특성변수를 표준화
- [Step 2] 학습데이터에서 임의의  $K$ 개 표본을 뽑아  $K$ 개 군집의 중심값  $\mu_l$  ( $l=1, \dots, K$ )로 놓음
- [Step 3] 각 관측치를 제공 유클리디안거리 기준으로 가장 가까운  $\mu_l$  군집에 편입
- [Step 4] 각 군집의 평균을 새로 구해 이를 새로운 중심값  $\mu_l$  ( $l=1, \dots, K$ )로 놓고 [Step 3] 실행
- [Step 5] 각 군집의 member가 변하지 않을 때 까지 (or 미리 정해진 최대 반복횟수에 도달할 때 까지) [Step 4] 반복

- $K$  의 선택

- Elbow method

# K의 선택

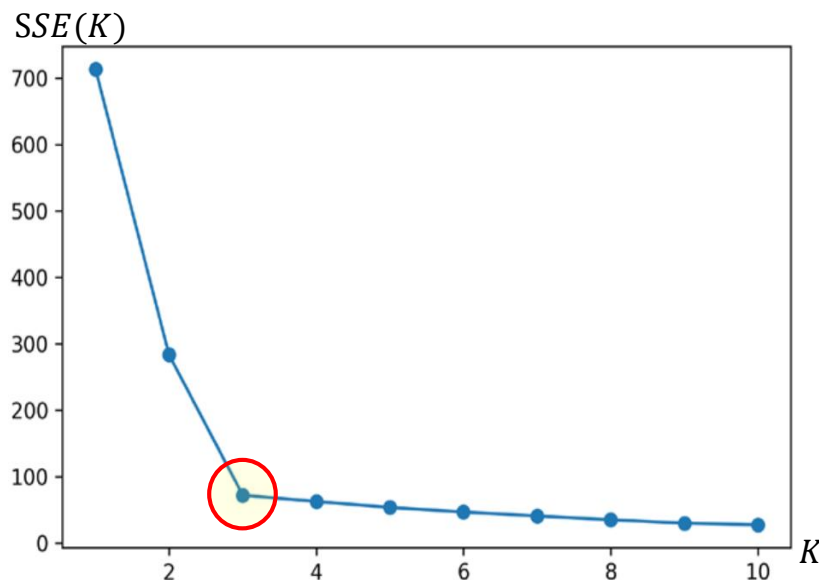
## ▪ Elbow Method (팔꿈치 방법)

- 목표:  $SSE(K) = \sum_{l=1}^K S_l^2$ 를 최소화하는  $K$ 를 찾는다.

▶  $S_l^2 = \sum_{i \in l} \sum_{j=1}^d (x_{ij} - \mu_{lj})^2$ : class  $l$ 에 포함된 관측치들의  $\mu_l$ 로부터의 제곱 유클리디안거리의 합

- But,  $SSE(K)$ 는  $K$ 에 대해 단조감소 → (대안)  $SSE(K)$ 의 감소속도가 확연히 줄어드는  $K$ 를 찾는다.

- 팔꿈치 그림:  $K$ 를 가로축에,  $SSE(K)$ 를 세로축에 놓고 그린 그림



▶  $K$ 가 커질수록  $SSE(K)$  감소

▶  $K=3$ 에서  $SSE(K)$ 의 감소속도(기울기)가 현저하게 줄어듦

▶  $K=3$  선택

→ “Elbow method”

- 단점: 군집 결과가 [Step 2]에서 임의로 뽑은 초기 중심값의 질에 의존 → (대안) “K-means++”

# K-means++

- 목적: 임의추출한  $K$ 개의 초기 중심값에 군집 결과가 의존하는 문제 완화
  - By 초기 중심값을 지정하는 방법 수정
- Algorithm
  - [Step 1]  $n$ 개 표본으로 구성된 학습데이터에서 임의표본 1개를 뽑아 이를 초기 중심값  $\mu_1$ 으로 놓음.
  - [Step 2] 나머지  $(n - 1)$ 개 자료에 대해  $\mu_1$ 으로부터의 제곱 유클리디안거리를 구하고 이 거리에 비례하는 확률을  $(n - 1)$ 개 자료 각각에 부여. 그리고 이 확률에 비례하여 1개의 임의표본을 뽑아 이를 두 번째 초기 중심값  $\mu_2$ 로 놓음.
  - [Step 3] 나머지  $(n - 2)$ 개 자료에 대해  $\min(d(x_i, \mu_1), d(x_i, \mu_2))$ 을 구하고 이 최소거리에 비례하는 확률을  $(n - 2)$ 개 자료 각각에 부여. 그리고 이 확률에 비례하여 1개의 임의표본을 뽑아 이를 세 번째 초기 중심값  $\mu_3$ 로 놓음.
  - [Step 4] [Step 1] - [Step 3]을 최종  $\mu_K$ 를 구할 때 까지 반복. 즉,  $(n - (k - 1))$ 개 자료에 대해  $\min(d(x_i, \mu_1), \dots, d(x_i, \mu_{k-1}))$ 을 구하고 이 최소거리에 비례하는 확률을  $(n - (k - 1))$ 개 자료 각각에 부여. 그리고 이 확률에 비례하여 1개의 임의표본을 뽑아 이를  $k$  번째 초기 중심값  $\mu_k$ 로 놓음.

## 03 Hierarchical Clustering (계층적 군집)

- 종류

- **결합적 (agglomerate) 군집**:  $n$ 개의 자료가 있을 때 각 자료를 한 개의 군집으로 보고, 비유사성 측도 기준으로 군집간 결합(agglomerate)을 최종적으로 하나의 군집이 될 때 까지 반복
- **분할적 (divisive) 군집**: 모든 자료를 하나의 동일한 군집으로 보고, 군집을 한 개씩 떼어냄.

- Notes

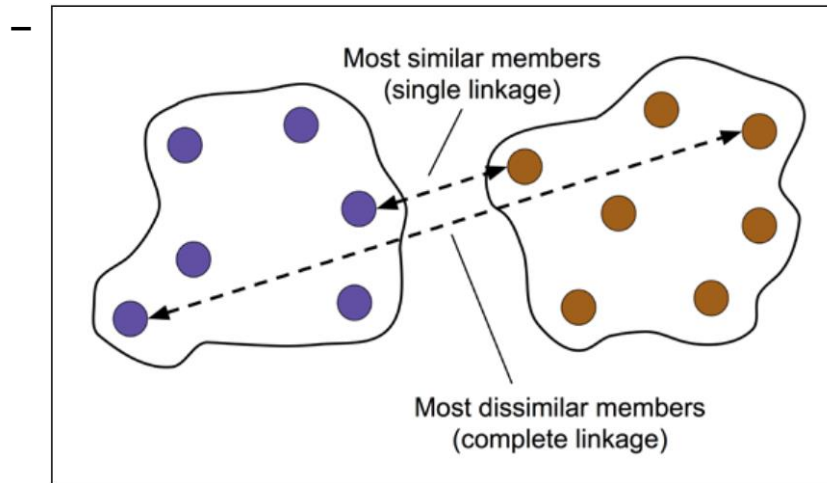
- 두 방법은 본질적으로 동일
- 군집의 개수  $K$ 를 미리 정할 필요 X → 계층적 군집의 장점
- 수업에서는 결합적 군집만 설명할 것임.

# Agglomerate Clustering (결합적 군집)

## ■ 군집 간 거리 측도에 따른 결합적 군집의 분류

- (기호)  $x_1, \dots, x_{n_A}$ : 군집 A에 속한 관측치,  $z_1, \dots, z_{n_B}$ : 군집 B에 속한 관측치

결합적 군집	군집간 거리 측도
완전연계군집 (Complete-linkage clustering)	$\max\{d(x_i, z_j)   i = 1, \dots, n_A, j = 1, \dots, n_B\}$
단순연계군집 (Single-linkage clustering)	$\min\{d(x_i, z_j)   i = 1, \dots, n_A, j = 1, \dots, n_B\}$
평균연계군집 (Average-linkage clustering)	$\frac{1}{n_A n_B} \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} d(x_i, z_j)$



▶ 완전연계: 가장 멀리 떨어진 멤버들 간의 거리

▶ 단순연계: 가장 가까운 멤버들 간의 거리

▶ 평균연계: 모든 멤버 조합의 거리들의 평균



# Agglomerate Clustering (결합적 군집, 계속)

- Algorithm

- [Step 1] 전체  $n$ 개 표본에서 거리가 가장 가까운 두 개 관측치를 한 개의 군집으로 합침.
- [Step 2]  $k = n - 1, \dots, 2$ 에 대하여 적절한 연계군집을 통해  $k$ 개 군집간의 거리를 구하여, 가장 가까운 거리를 가진 두 개의 군집을 하나의 군집으로 합침.

## 04 DBSCAN (Density-based clustering of approximations with noises)

### ■ 밀도기반 군집 (Density-based clustering)

- 공간상에 높은 밀도 (high density)를 가지고 모여 있는 관측치들을 하나의 그룹으로 간주하고, 낮은 밀도를 가지고 홀로 있는 관측치는 이상치 또는 잡음 (noise)으로 분류

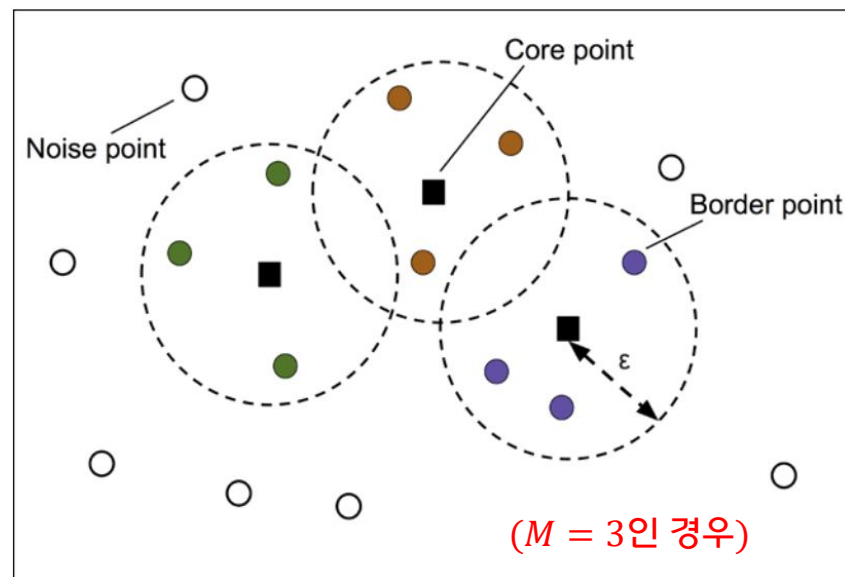
### ■ 관측치의 유형 분류

- 관측치의  $\varepsilon$ -neighborhood가  $M$ 개 이상의 다른 관측치를 포함하는지 여부를 고려하여 분류  
→ 즉,  $\varepsilon$ 과  $M$ 은 초모수 (hyper parameter)가 됨.

① 핵심자료 (core point):  $\varepsilon$ -neighborhood에  $M$ 개 이상의 다른 관측치를 포함하는 관측치

② 주변자료 (border point): 핵심자료는 아니지만  $\varepsilon$ -neighborhood에 핵심자료를 포함하는 관측치

③ 잡음자료 (noise point): 핵심자료도 주변자료도 아닌 관측치



# 수행 절차

- Algorithm
  - [Step 1] 각 핵심자료의  $\varepsilon$ -neighborhood에 있는 관측치들로 하나의 군집을 형성
    - ▶ 핵심자료 간의 거리가  $\varepsilon$  이하인 경우, 대응하는 군집들은 하나의 군집으로 통합
  - [Step 2] 주변자료는 가장 가까운 거리에 있는 핵심자료가 포함된 군집으로 편입
- 초모수
  - $\varepsilon$ : 너무 작으면 많은 관측치가 잡음자료로 분류되고, 너무 크면 군집의 개수가 적어짐.
  - $M$ : 일반적으로 '특성변수의 개수 + 1'을 사용

Q & A