[2020 혁신성장 청년인재 집중양성 사업]

프로젝트 기반 데이터 과학자 양성과정

# 빅데이터 분석

- 11주차 -

**#군집 분석 #연관성 분석**

# Part 1

## 군집 분석

# 군집 분석

**군집화란?**

- 관측값 또는 개체를 의미 있는 몇 개의 부분 집단으로 나누는 과정

- 군집화의 기준 : 동일한 군집에 속한 개체들이 여러 속성이 유사하고, 서로 다른 군집에 속한 관찰치는 다른 속성을 갖도록 군집을 구성

- 군집화를 위한 변수 : 전체 개체의 속성을 판단하기 위한 기준

  - 인구 : 성별, 나이, 거주지, 직업, 소득 등

  - 구매 패턴 : 상품, 주기, 거래액 등

- 의미 있는 집단 : 같은 집단에 속한 관측값 또는 개체들이 서로 유사하고 다른 군집에 속한 개체 사이에는 유사성이 적은 것을 의미함

# 군집 분석

**군집 분석의 활용**

- 고객 세분화

- 고객이 기업의 수익에 기여하는 정도를 통한 고객세분화

  - 우수고객의 인구통계적 요인, 생활패턴 파악

  - 개별고객에 대한 맞춤 관리(추천시스템)

- 고객 구매 패턴에 따른 고객 세분화

  - 제품 포지셔닝, 목표 고객집단 구성

# 군집 분석

**군집 분석의 유형**

- 장점

  - 탐색적인 기법
  - 다양한 형태의 데이터에 적용 가능
  - 분석 방법의 적용 용이

- 단점

  - 가중치와 거리의 정의
  - 초기 군집의 설정
  - 결과 해석의 어려움

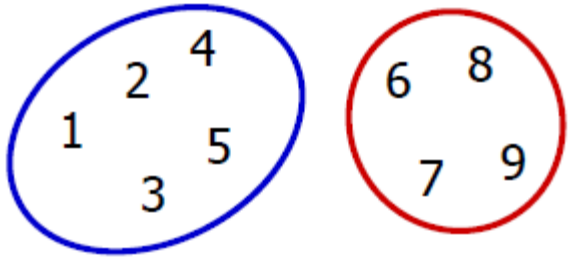
# 군집 분석

**군집 분석의 유형**

- 상호 배반적 유형

  - 개체의 유형이 다른 여러 군집 중, 오직 하나에만 속함

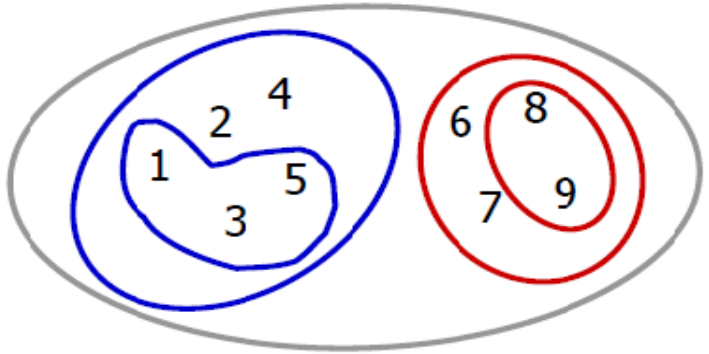
  - ex) 한국인, 중국인, 일본인




- 계층적 유형

  - 한 군집이 다른 군집의 내부에 포함되는 형태로 군집 간의 중복은 없으며 군집들이 단계마다 계층구조를 이룸

  - ex) 전자제품 > 주방용 > 냉장고

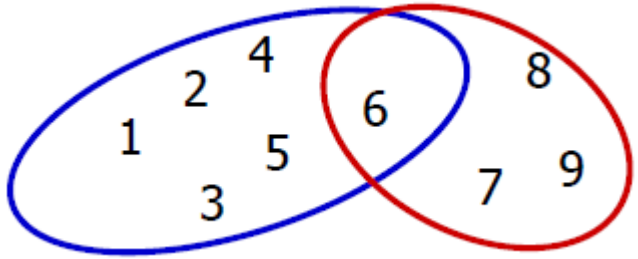


# 군집 분석

**군집 분석의 유형**

- 중복 유형

  - 두 개 이상의 군집에 한 관찰치가 동시에 소속되는 것을 허용




- 퍼지 유형

  - 관찰치가 소속되는 특정한 군집을 표현하는 것이 아니라

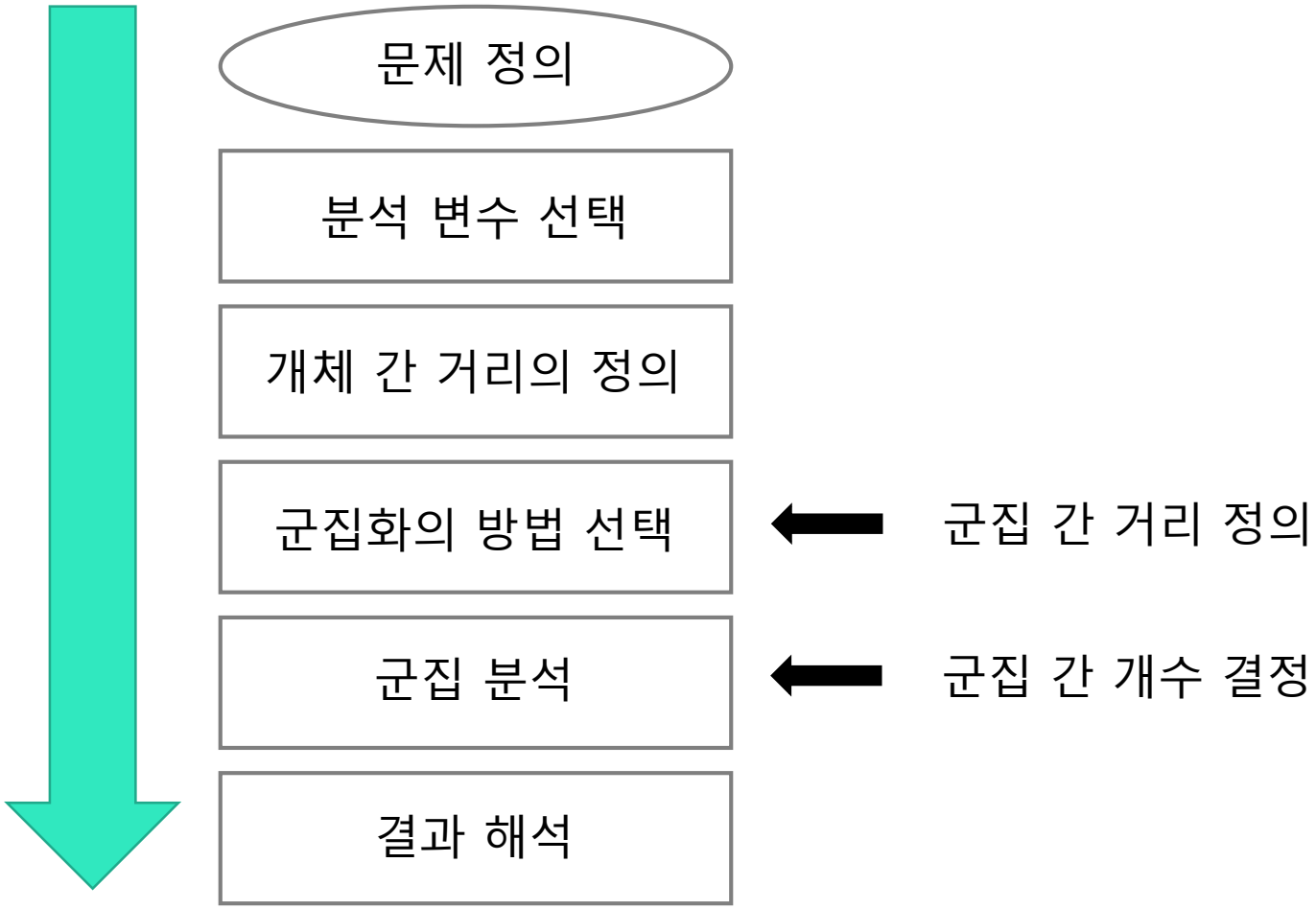
    각 군집에 속할 가능성을 표현

Prob ( 개체 1 ∈ 군집 $A$ ) = 0.7

Prob ( 개체 1 ∈ 군집 $B$ ) = 0.3

# 군집 분석

**군집 분석의 프로세스**



# 군집 분석

**거리 계산**

- 유클리디안 거리(Euclidian distance)

$$d(x_i, x_k) = \sqrt{\sum_{j=1}^{p} (x_{ij} - x_{kj})^2}$$

- 맨해튼 거리(Manhattan distance)

$$d(x_i, x_k) = \sqrt{\sum_{j=1}^{p} |x_{ij} - x_{kj}|}$$

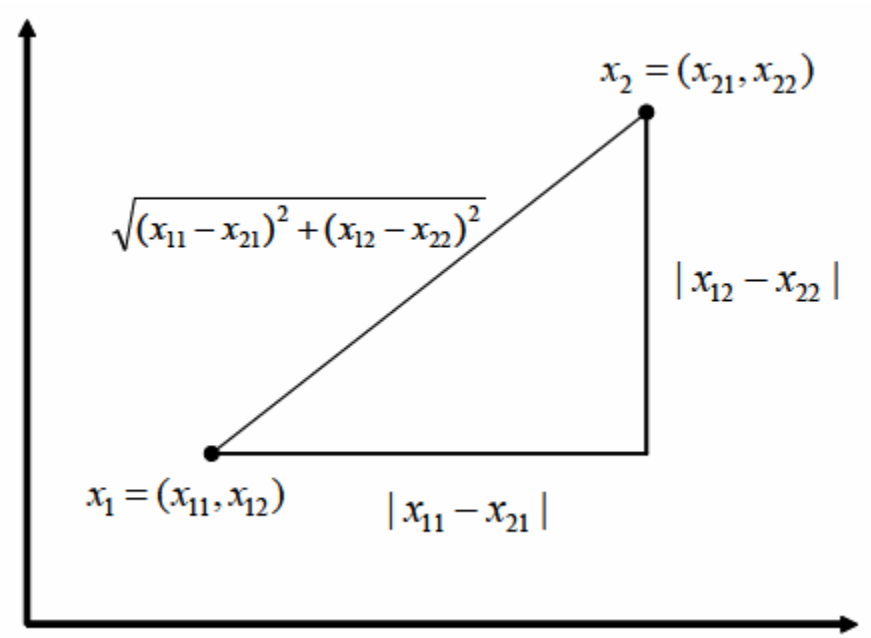
- 민코브스키 거리(Minkowski distance)

$$d(x_i, x_k) = \left( \sum_{j=1}^{p} |x_{ij} - x_{kj}|^m \right)^{1/m}$$

# 군집 분석

**유클리디안 거리와 맨해튼 거리 비교**

- 두 점 간에 유클리디안 거리와 맨해튼 거리 비교




- 거리는 계산하는 방법에 따라 군집화의 결과가 크게 변하지 않음. 일반적으로 유클리디안 거리가 가장 많이 사용되고 있는데, 맨해튼 거리보다 성능이 더 좋은 것으로 평가됨

# 군집 분석

**유사도 계산**

- 유클리안 거리 방법으로 계산하라

| 개체 | 변수1 | 변수2 |
|---|---|---|
| 1 | 1 | 1 |
| 2 | 2 | 2 |
| 3 | 3 | 3 |
| 4 | 4 | 5 |
| 5 | 5 | 4 |
| 6 | 6 | 7 |
| 7 | 7 | 5 |

```
similarity1 <- dist(ex1, method = "euclidean")
similarity2 <- dist(ex1, method = "manhattan")
```

# 군집 분석

**계층적 군집화**

- 계층적 군집

  - 응집분석 : 각 개체를 하나의 군집으로 보고 가까운 군집끼리 합하여 결국에는 하나의 군집까지 합하는 기법

  - 분할분석 : 전체의 개체를 하나의 군집으로 보고 각 군집을 두 개의 군집으로 계속 나누는 기법

- 계층적 군집화는 사전지식이 필요 없음

- 종종 탐색적인 의미로 사용됨

# 군집 분석

**응집분석**

- 응집분석

  - 각 개체를 하나의 군집으로 하여 전체 n개의 군집을 형성

  - 각 군집간의 거리를 계산하여 가장 가까운 두 개의 군집을 합침

  - 전 개체가 하나의 군집이 될 때까지 군집을 계속 합함

- 군집 간 거리계산 방법

  - 단일연결법 : 최단연결법

  - 완전연결법 : 최장연결법

  - 평균연결법

# 군집 분석

**응집분석**

- 단일 연결법 : 개체 간의 가장 작은 거리를 두 군집 간의 거리로 함

$$d(P,Q) = \min_{i \in P, k \in Q} d(i,k)$$


- 완전 연결법 : 개체 간의 가장 먼 거리를 두 군집 간의 거리로 함

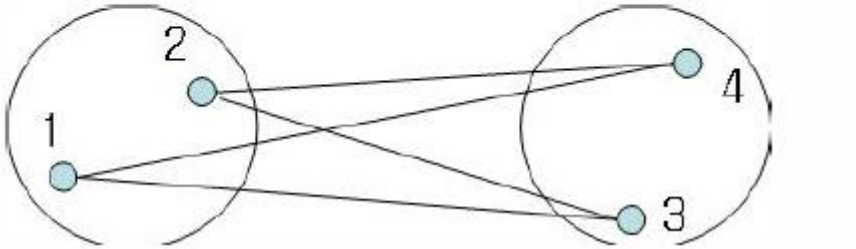
$$d(R,Q) = \max_{i \in R, k \in Q} d(i,k)$$

# 군집 분석

**응집분석**

- 평균 연결법 : 개체 간의 모든 거리의 평균

$$d(P,\ Q) = \frac{\sum_{i \in P, k \in Q} d(i,k)}{(\text{군집 } P\text{에서의 개체수}) \times (\text{군집 } Q\text{에서의 개체수})}$$




- 단일 연결법과 완전 연결법의 절충안

# 군집 분석

**응집분석**

- 군집 간의 거리를 계산하라

| 개체 | 변수1 | 변수2 |
|---|---|---|
| 1 | 1 | 1 |
| 2 | 1 | 2 |
| 3 | 3 | 4 |
| 4 | 5 | 5 |
| 5 | 7 | 5.5 |

```
single <- hclust(dist(ex2, method = "manhattan"),
method = "single")
plclust(single)

complete <- hclust(dist(ex2, method = "manhattan"),
method = "complete")
plclust(complete)

average <- hclust(dist(ex2, method = "manhattan"),
method = "average")
plclust(average)
Cutree(average, 2)
```

# 군집 분석

**분할분석**

- 분할분석

  - 큰 군집을 둘로 나누는 계층적 군집화 방법

- DIANA 알고리즘

  - 개체와 군집과의 거리를 구할 때 평균 연결법을 사용

  - 분할분석은 전체를 하나의 군집으로 하고 하나의 군집이 둘로 나눠지는 과정을 거쳐 계층적 군집분석을 진행

# 군집 분석

**분할분석**

- 군집 간의 거리를 계산하라

| 개체 | 변수1 | 변수2 |
|---|---|---|
| 1 | 1 | 1 |
| 2 | 1 | 2 |
| 3 | 3 | 4 |
| 4 | 5 | 5 |
| 5 | 7 | 5.5 |

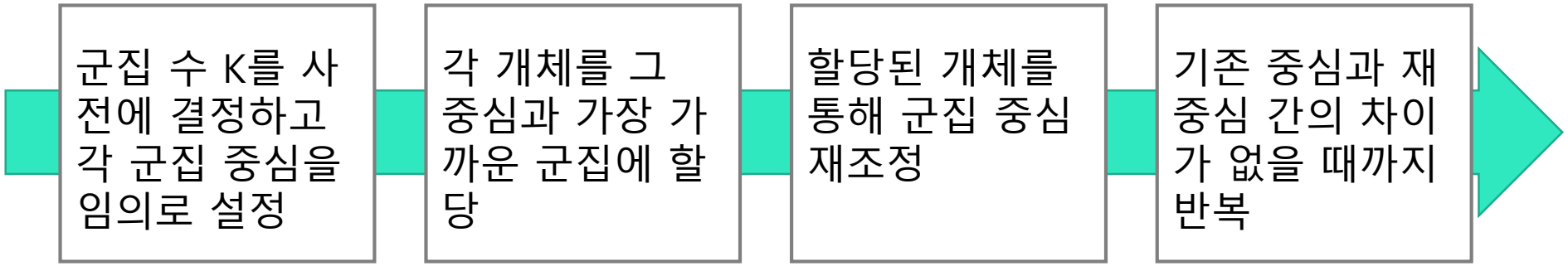
```
library(cluster)
diana <- diana(ex2, metric = "manhattan")
plot(diana)
cutree(diana)
```

# 군집 분석

## K-means Clustering

- 특징

  - 각 관찰치를 상호 배반적인 K개의 군집을 형성
  - 초기에 부적절한 분리가 일어났을 때 회복가능
  - 군집의 수를 사전에 정의
  - 대용량 데이터에 유리



# Part 2

## 연관성 분석

# 연관성 분석

**연관성 분석이란?**

- 연관규칙 : 어떤 사건이 얼마나 자주 동시에 발생하는가를 표현하는 규칙 또는 조건
- 연관규칙 표시 : "A가 발생하면 B가 발생한다"는 규칙을 "A => B"로 표현
- 연관성 분석 : 연관규칙을 이용해 하나의 사건이나 거래에 포함되어 있는 둘 시상의 품목 간 상호연관성을 발견해 내는 분석방법
- 장바구니 분석이라고도 함
- 군집분석과 함께 비지도 학습의 대표적인 분석방법

# 연관성 분석

**연관성 분석 특징**

- 확률과 기대값에 대한 개념

- 도출되는 결과를 직접적인 인과관계로 판단해서는 안되며, 두 개 또는 그 이상의 품목들 사이의 상호 관련성으로 해석해야 함

- 대용량 데이터로 얻어지는 연관성 규칙들이 모두 유용한 내용이 아닐 수 있음

"기저귀를 사러 온 고객들은 맥주도 함께 사간다"

"새로 문을 연 건축 자재점에서는 변기 덮개가 가장 많이 팔린다."

"이전에 동일한 제조사의 전자제품을 주로 구매했던 고객은 신제품 구매에서도 동일한 회사의 제품을 구매한다."

# 연관성 분석

**연관성 분석 장점**

- [탐색적인 기법]
  "조건 => 반응"의 규칙 형태를 가지고 있어 이해가 쉽고 적용이 용이함

- [비지도 학습]
  목적변수 없이도 적용이 쉬움

- [쉬운 데이터 형태]
  특별한 전처리 없이 사용 가능한 데이터 구조

- [계산의 용이성]
  분석을 위한 계산이 아주 간단함

# 연관성 분석

**연관성 분석 단점**

- [상당한 수의 계산과정]

  연관성을 관찰하고자 하는 항목(Item)이 증가하면 계산의 수가 크게 증가함

- [적절한 항목의 결정]

  불필요한 항목이 존재

- [항목의 비율 차이]

  거래량이 적은 항목의 경우, 연관성 규칙 과정에서 제외될 가능성이 있음

# 연관성 분석

**지지도**

- 품목들 간의 연관성의 정도를 평가하는 도구

- 지지도(Support) : 전체 자료에서 관련성이 있다고 판단되는 품목들을 포함하고 있는 거래나 사건의 확률 => 두 개의 항목이 동시에 일어날 확률

$$A \rightarrow B\text{의 지지율} = \frac{A\text{와}\ B\text{를 포함한 거래수}}{\text{전체 거래수}}$$

ex) 대형할인점의 1백만 건의 거래 중에서 1만 건의 거래가 A와 b를 모두 포함한 경우 :

연관규칙 A -> B의 지지율은 1%

- 지지율은 자주 발생하지 않은 규칙을 제거하는데 사용될 수 있음

- A->B와 B->A가 같은 지지율을 갖기 때문에 차이를 알 수 없으므로 다른 평가지표가 필요

# 연관성 분석

**신뢰도**

- 신뢰도(Confidence) : 항목 A를 구매하였을 경우 항목B를 구매하는 확률은 얼마인가?

$$A \rightarrow B\text{의 신뢰도} = \frac{A\text{와}B\text{를 포함한거래수}}{A\text{를 포함한거래수}}$$

ex) 어느 슈퍼마켓의 1백만 건의 거래 중에서 A를 포함한 거래가 5만 건이고 A를 포함한 거래 중에서 B를 포함한 거래가 1만 건인 경우 : 신뢰도는 20%

- A->B의 신뢰도는 A가 발생했을 때 B가 발생하는 조건부 확률과 같음

- 신뢰도는 지지율에 비하여 상품의 연관성을 측정하기에 더 적합한 평가척도

# 연관성 분석

**향상도**

- 신뢰도는 연관규칙이 실제로 유용한지 아니면 임의로 나타난 결과인지도 알 수 없다는 단점이 있다

  ex) "빵 -> 우유 " 의 신뢰도가 30%이고 전체 거래 중 30%가 우유를 포함한다면

  연관규칙 "빵 -> 우유 " 는 유용하지 못한 규칙일 가능성이 높음. 빵을 통한 우유에 대한 마케팅이 일반적인 우유에 대한 마케팅에 비하여 높은 수익을 올릴 수 없기 때문.

- 향상도(Lift) : 연관규칙 A -> B의 신뢰도를 B를 포함한 거래비율로 나눈 값

$$\text{향상도} = \frac{A \rightarrow B\text{의 신뢰도}}{B\text{를 포함한 거래비율}} = \frac{P(B|A)}{P(B)}$$

- 향상도는 규칙을 모를 때에 비하여 규칙을 알 때에 얼마나 판매가 향상되는가를 나타냄. 즉, 향상도는 상품B를 연관규칙과 관계없이 판매하는 것에 비하여 연관규칙을 알고 A를 구매한 고객에 대하여 B를 판매하는 경우 얼마나 판매가 증가하는가를 나타냄

# 연관성 분석

**향상도**

- 향상도 = 1

  A와 B가 확률적으로 독립에 가까운 것을 의미

- 향상도 > 1

  A를 구매하는 경우 B를 구매할 가능성이 높다는 것을 의미하므로 양의 상관관계

- 향상도 < 1

  A를 구매하는 경우 B를 구매할 가능성이 낮다는 것을 의미하므로 음의 상관관계

# Part 3

## 실습