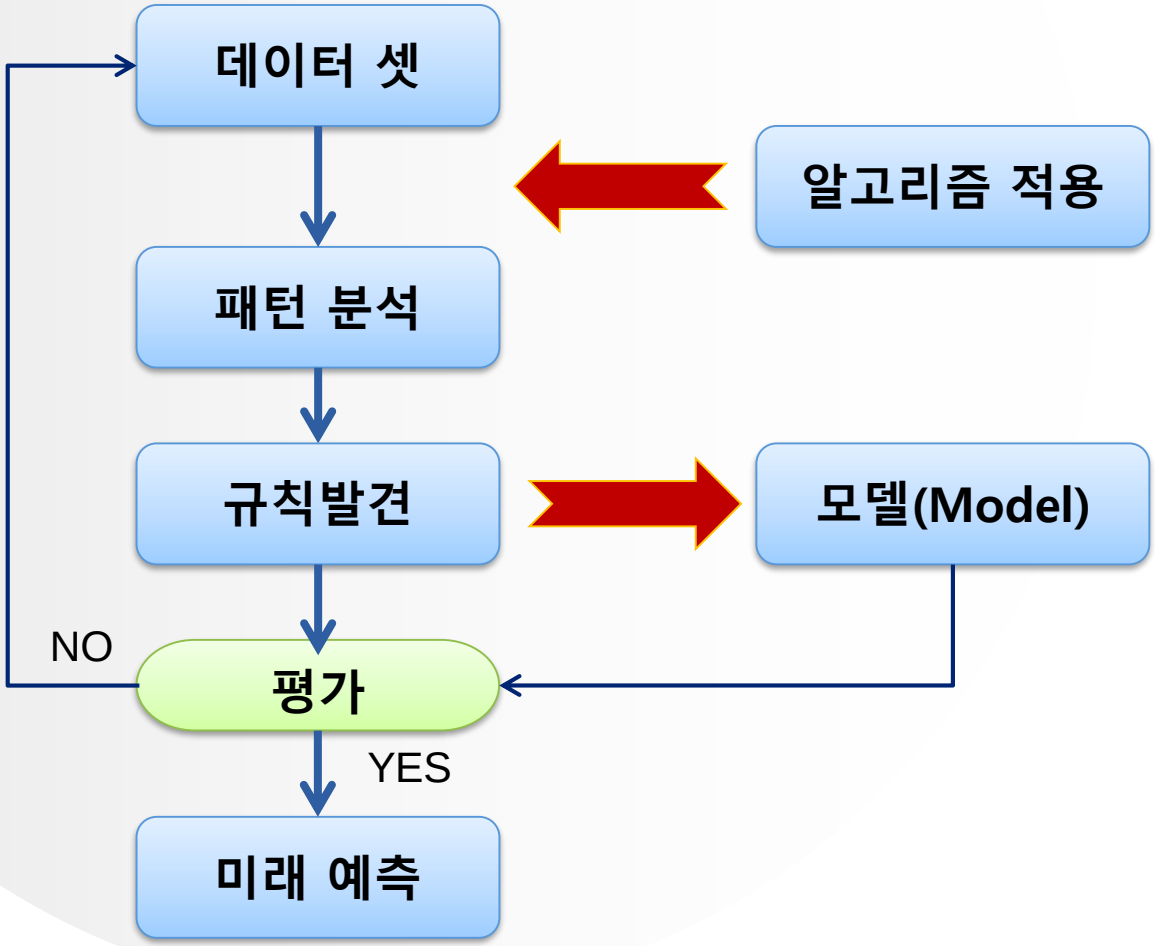
비지도학습

- 비지도학습(unSupervised Learning) 절차



18. 군집분석

Chap18_ClusteringAnalysis 수업내용

1) 군집분석 개요

2) 유클리드 거리

3) 계층적 군집분석

4) 비계층적 군집분석

1) 군집 분석 개요

- **군집 분석?**
  - 종속변수(y변수)가 없는 데이터 마이닝 기법
  - 유클리드 거리 기반 유사 객체 묶음
  - 고객 DB -> 알고리즘 적용 -> 패턴 추출(rule) -> 근거리 모형으로 군집형성
  - 계층적 군집분석(탐색적), 비계층적 군집분석(확인적)
  - 주요 알고리즘 : k-means, hierarchical

1) 군집 분석 개요

- **군집분석 특징**
  - 전체적인 데이터 구조를 파악하는데 이용
  - 관측대상 간 유사성을 기초로 비슷한 것 끼리 그룹화(Clustering)
  - 유사성 = 유클리드 거리
  - 분석결과에 대한 가설 검정 없음(타당성 검증 방법 없음)
  - 분야 : 사회과학, 자연과학, 공학 분야
  - 척도 : 등간, 비율척도(연속적인 양)

- **유클리드 거리 계산식**

$$\sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \ldots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

관측대상 p와 q의 대응하는 변량값의 차가 작으면, 두 관측대상은 유사하다고 정의하는 식

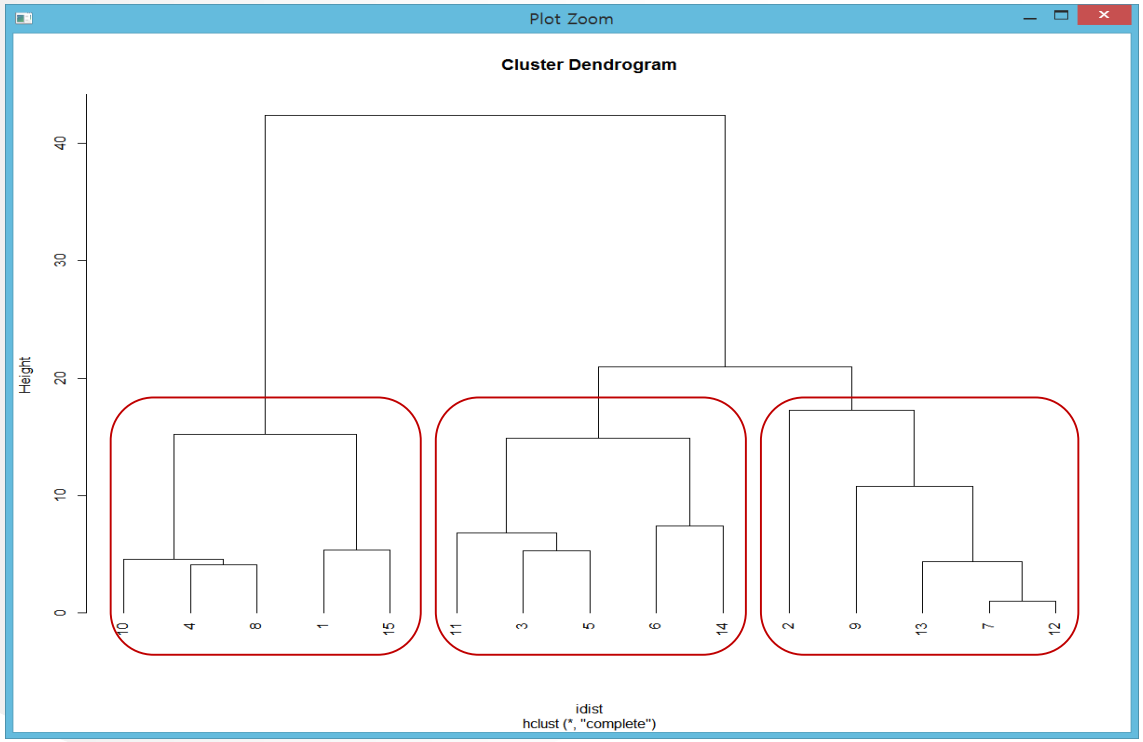
1) 군집 분석 개요

- **군집 구성법**
  - 그룹간의 유사성 계산 방법
  - 최단거리법, 최장거리법, 메디안법, 중심법, 그룹평균법

- **군집화방법**
  - 계층군집화 : 가장 가까운 대상끼리 순차적으로 묶음
  - 비계층군집화 : k-평균 군집법

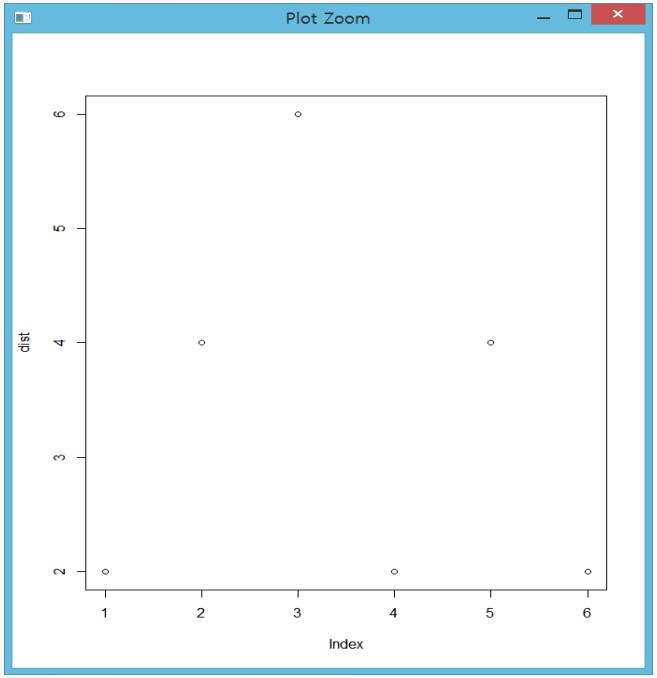
1) 군집 분석 개요

- **군집 분석 결과**
  - 평균결합방식을 적용한 덴드로그램(Dendrogram)
  - 가로축 : 학생번호, 세로축 : 상대적 거리
  - 군집수는 사용자가 정할 수 있음(2집단, 3집단 등)



2) 유클리드 거리

- **유클리드 거리(Euclidean distance)**
  - 두 점 사이의 거리를 계산하는 방법
  - 이 거리를 이용하여 유클리드 공간 정의



2) 유클리드 거리

- **유클리드 거리 실습**

(1) matrix 생성

```
x <- matrix(1:16, nrow=4)
```

```
     [,1] [,2] [,3] [,4]
[1,]    1    5    9   13
[2,]    2    6   10   14
[3,]    3    7   11   15
[4,]    4    8   12   16
```

(2) matrix 대상 유클리드 거리 생성 함수

```
# x : numeric matrix, data frame
dist <- dist(x, method="euclidean") # method 생략가능
```

```
#  1 2 3
#2 2
#3 4 2
#4 6 4 2    <- 가까운 객체 끼리 묶어줌
```

(3) 유클리드 거리 계산 식

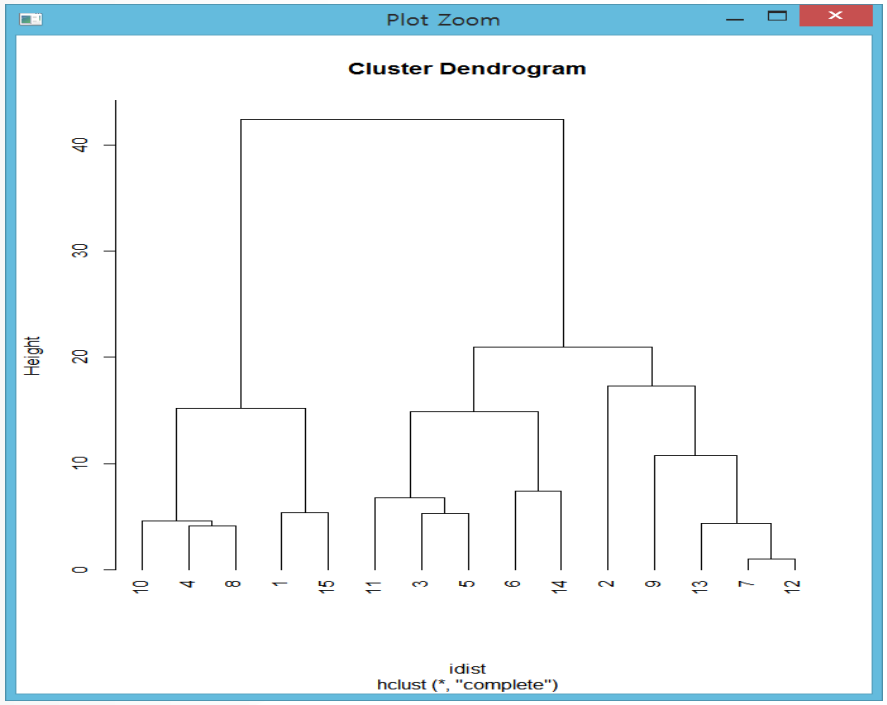
```
sqrt(sum((x[1,]-x[4,])^2)) # 6
```

<유클리드거리 계산법>

1. 두 벡터의 차이를 구한다.
2. 원소를 제곱해서 더한다.
3. 제곱근을 취한다.

3) 계층적 군집 분석

- **계층적 군집분석**
  - 유클리드 거리를 이용한 군집분석 방법
  - cluster 패키지에서 제공되는 hclust() 함수 이용
  - 계층적(hierarchical)으로 군집 결과 도출
  - 탐색적 군집분석



3) 계층적 군집 분석

- **계층적 군집분석 절차**

(1) 군집분석(Clustering)분석을 위한 패키지 설치
install.packages("cluster") # hclust() : 계층적 클러스터 함수 제공
library(cluster) # 일반적으로 3~10개 그룹핑이 적정

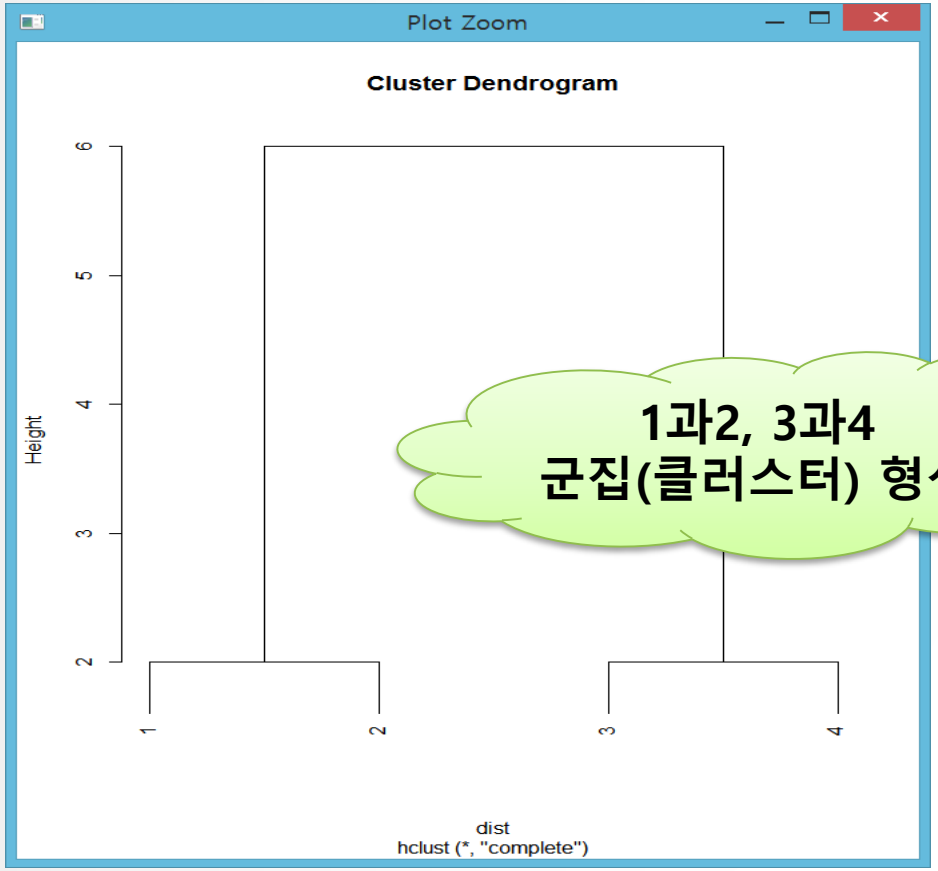
(2) 데이터 셋 생성
x <- matrix(1:16, nrow=4)

(3) matrix 대상 유클리드 거리 생성 함수
dist <- dist(x, method="euclidean") # method 생략가능

(4) 유클리드 거리 matrix를 이용한 클러스터링
hc <- hclust(dist) # 클러스터링 적용
plot(hc) # 클러스터 플로팅

3) 계층적 군집 분석

- 계층적 군집분석 결과 : 벤드로그램(dendrogram)



1과2, 3과4
군집(클러스터) 형성

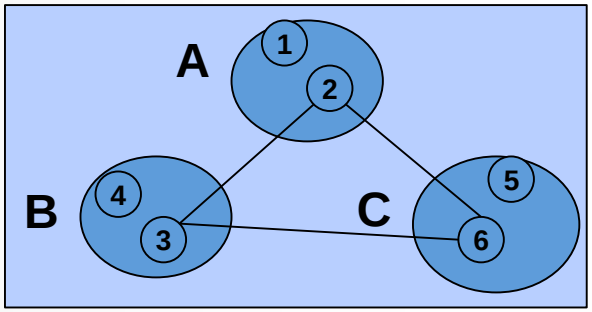
2. 알고리즘에 따른 분류

군집화 방식

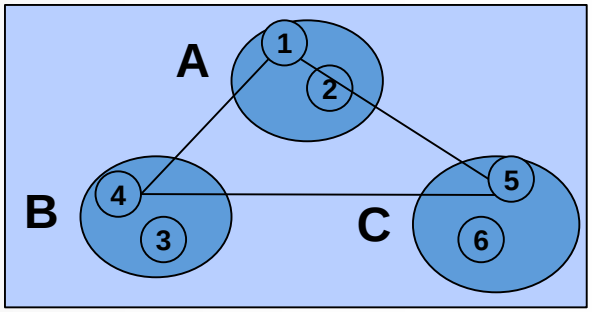
단일기준결합방식 :
각 군집에서 중심으로부터 거리가 가까운 것(2,3,6) 1개씩 비교하여 가장 가까운 것 끼리 군집화

완전기준결합방식 :
각 군집에서 중심으로부터 가장 먼 대상(1,4,5) 끼리 비교하여 가장 가까운 것 끼리 군집화

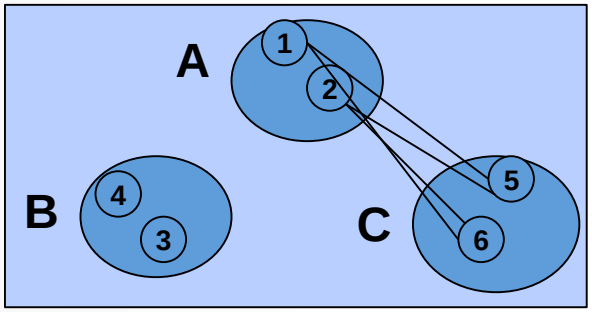
평균기준결합방식 :
한 군집 안에 속해 있는 모든 대상과 다른 군집에 속해있는 모든 대상의 쌍 집합에 대한 거리를 평균 계산하여 가장 가까운 것 끼리 군집화
(1 -> 5,6 평균, 2 -> 5, 6 평균)



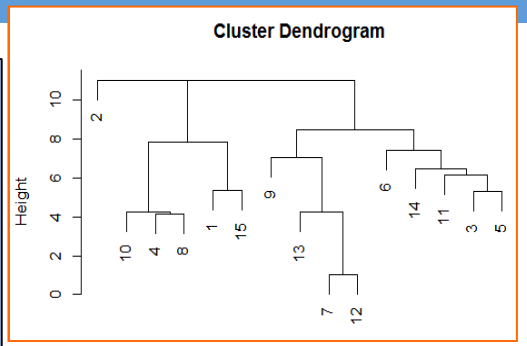
단일기준결합방식

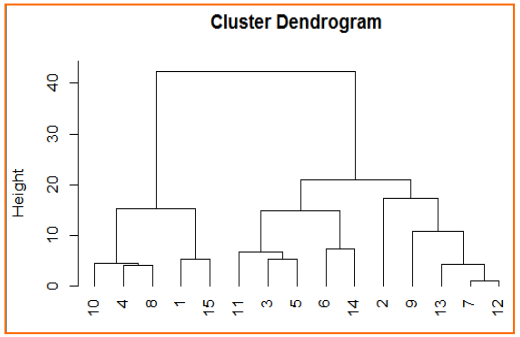


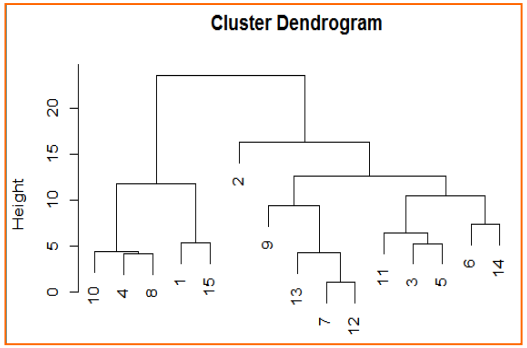
완전기준결합방식



평균기준결합방식







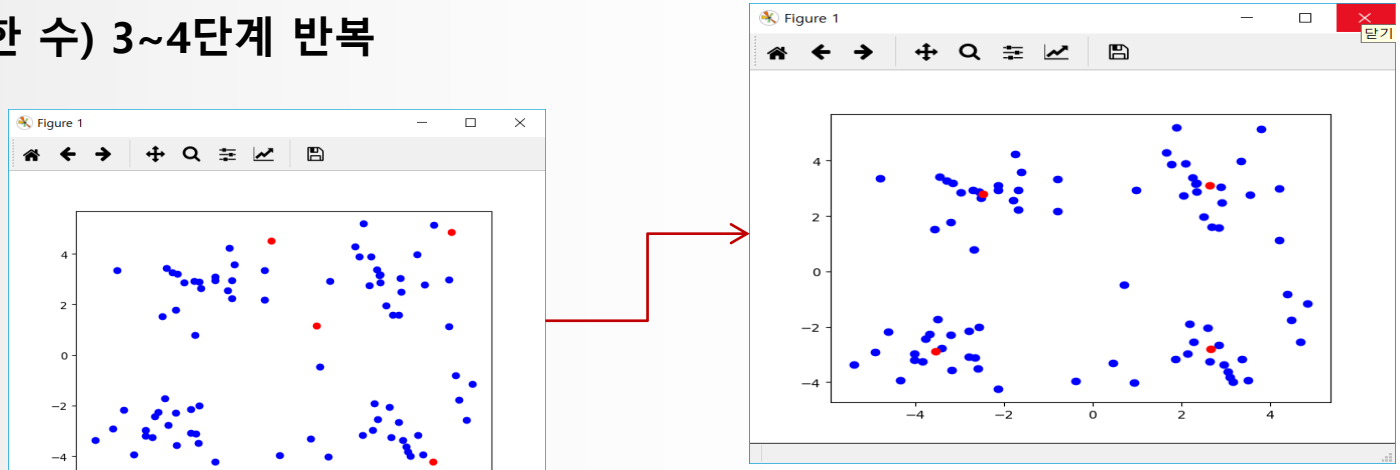
R의 덴드로그램

4) 비 계층적 군집 분석

- **비계층적 군집 분석(k-means)**
  - 확인적 군집분석 방법
  - 계층적 군집분석법 보다 속도 빠름
  - 군집의 수를 알고 있는 경우 이용
  - K는 미리 정하는 군집 수
  - 계층적 군집화의 결과에 의거하여 군집 수 결정
  - 순차적 군집분석법(군집과정 반복)
  - 변수 보다 관측대상 군집화에 많이 이용
  - 군집의 중심(Cluster Center) 사용자가 정함

4) 비 계층적 군집 분석

- **k-평균 군집분석 알고리즘**
  - 단계1. k값을 초기값으로 k개 centroid 선정
  - 단계2. 각 데이터 포인터를 가장 가까운 centroid에 할당
  - 단계3. centroid에 할당된 모든 데이터 포인트의 중심 위치 계산(centroid 재조정)
  - 단계4. 재조정된 centroid와 가장 가까운 데이터 포인트 할당
  - 단계5. centroid 재조정이 발생되지 않을 때 까지 (or 지정한 수) 3~4단계 반복



군집분석