

13강 군집분석 II

통계·데이터과학과 장영재 교수



목차

01 군집분석 관련 R 함수

02 R 사용 예제



1. 군집분석 관련 R 함수



01 군집분석관련 R 함수

1 dist 함수

■ 함수의구조

`dist(x, method = "euclidean")`

■ 기능

행사이의 거리를 계산하는 기능을 수행하며, 통상 `dist`는 `hclust` 등 다른 명령문에 이용되므로 결과는 비유사성 행렬의 대각선 위부분이 출력됨. 즉, $n \times p$ 행렬에 대하여 $n(n-1)/2$ 개의 원소를 갖는 벡터가 생성

■ 옵션

- `x` : 행 사이의 거리를 계산하는 행렬
- `method` : 거리 계산방법 옵션. "euclidean"은 유클리디안 거리를 이용하여 거리를 계산하고, "manhattan"은 맨해튼 거리를 이용하여 거리를 계산. 기본값은 euclidean



01 군집분석관련 R 함수

2 hclust 함수

■ 함수의구조

`hclust(dist,method="complete")`

■ 기능

계층적군집분석중 한가지방법인 응집분석을 수행

■ 옵션

- `dist` : 거리 구조 또는 거리 행렬 객체. 보통 `dist` 함수 적용 결과를 이용
- `method` : 응집분석에 따른 계층적 군집화 방법 옵션. `single`은 단일 연결법, `complete`는 완전연결법, `average`는 평균연결법을 이용한 계층적 군집화를 수행. 기본값은 `complete`



01 군집분석관련 R 함수

3 diana 함수 (cluster 패키지 설치 및 호출 필요)

■ 함수의구조

diana 함수 (x, metric = "euclidean")

■ 기능

분할분석에따른 계층적 군집화를 수행

■ 옵션

- x : data frame 또는 data 행렬
- metric : 거리 계산방법 옵션. 예를 들어 "euclidean"은 유클리디안 거리를 이용하여 거리를 계산하고, "manhattan"은 맨해튼 거리를 이용하여 거리를 계산. 기본값은 euclidean



01 군집분석관련 R 함수

4 kmeans 함수

■ 함수의구조

`kmeans(x, centers, algorithm="Hartigan-Wong")`

■ 기능

K - 평균 군집분석을 수행

■ 옵션

- `x` : data 행렬
- `centers` : K - 평균 군집분석을 수행하기 위한 초기값을 가지고 있는 행렬. 각 행은 각 군집의 초기값을 가지고 있어야 하며, K - 평균 군집화를 위한 군집 수는 `centers`에서의 행의 수
- `algorithm` : K - 평균 군집분석에 사용되는 알고리즘. 기본값은 "Hartigan- Wong". 매퀸(1967)에 기초한 알고리즘을 사용하려면 "MacQueen"을 사용



01 군집분석관련 R 함수

5 plot 함수

- 함수의구조

plot(object)

- 기능

일반적으로 object에 대한 그림을 생성

- 옵션

- object : 명령문 hclust나 diana에 의하여 생성된 결과 객체(object)



01 군집분석관련 R 함수

6 cutree 함수

■ 함수의구조

`cutree(tree, k=)`

■ 기능

명령문 `hclust`나 `diana`에 의하여 생성된 결과 객체(object)를 가지고 주어진 군집수에 대하여 각 개체에 대한 id를 갖는 벡터를 생성

■ 옵션

- `tree` : 명령문 `hclust` 또는 `diana`의 결과를 가지고 있는 object
- `k` : 계층적 군집화로부터 얻기를 원하는 군집 수



01 군집분석관련 R 함수

7 table 함수

- 함수의구조
table(...)

- 기능
분할표를 생성

- 옵션

- ... : 범주형으로 해석될 수 있는 한 개 이상의 객체(object)



01 군집분석관련 R 함수

8 tapply 함수

■ 함수의구조

tapply(x, indices, FUN=)

■ 기능

자료에서 같은 범주에 속한 개체에 대하여 함수의 결과를 산출

■ 옵션

- x : 자료행렬
- indices : 범주를 가지고 있는 리스트
- FUN= : 함수의 이름을 가지고 있는 문자 string. K - 평균을 위해서는 mean 을 사용하면 되고 FUN=은 생략가능



2. R 사용 예제



다음시간 안내

14강. 연관성분석 I

