

✓ 군집분석(cluster analysis)

9강. 계층적 군집분석

- 응집형/분리형 군집화
- 자카드 유사도와 코사인 유사도
- 실루엣 계수

+ 코드

+ 텍스트

■ 비지도학습의 종류



■ 계층적 군집분석

계층적으로 각 데이터를 유사한 군집으로 묶어가며 진행하는 방법



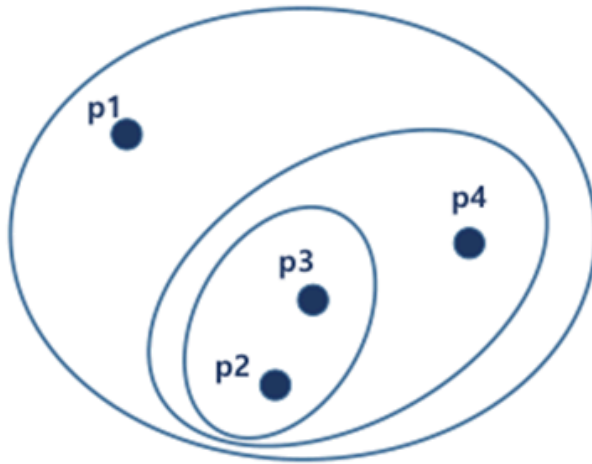
■ 계층적 군집분석

계층적으로 각 데이터를 유사한 군집으로 묶어가며 진행하는 방법

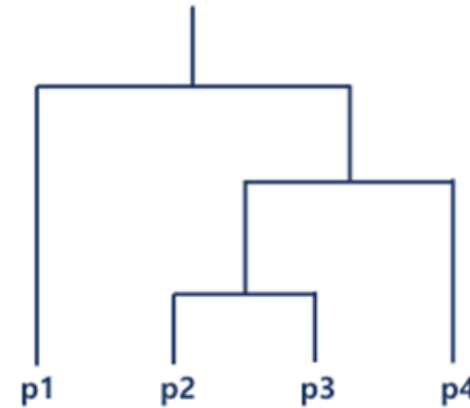
응집형

하나의 데이터를 하나의 군집으로 보고 가까운 군집끼리 병합해 나가는 방식

분리형



[부분 군집 다이어그램]



[덴드로그램]

■ 계층적 군집분석

계층적으로 각 데이터를 유사한 군집으로 묶어가며 진행하는 방법

응집형

분리형

데이터간 거리
측정 방식

· 수치형 데이터

유클리디안 거리

맨하탄 거리

체비셰프 거리

마할라노비스 거리

좌표 차이 중 가장 큰 값을 갖는 거리

두 점 사이의 확률분포를 고려한 거리

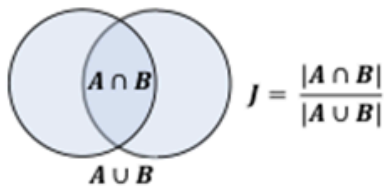
· 범주형 데이터

자카드 유사도

코사인 유사도

두 벡터의 교집합과 합집합의 비

두 벡터의 코사인 각도

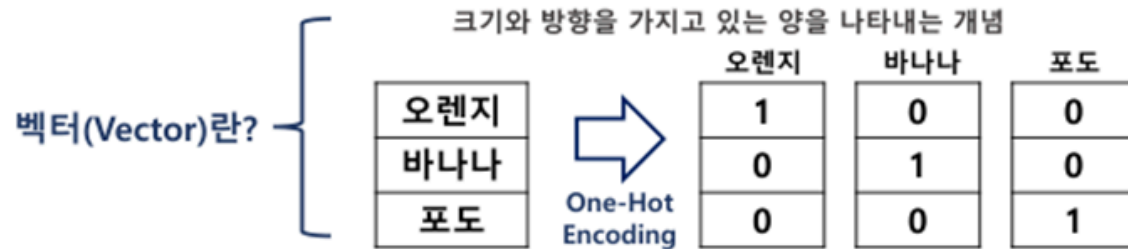


$$\cos\theta = \frac{A \cdot B}{|A||B|}$$

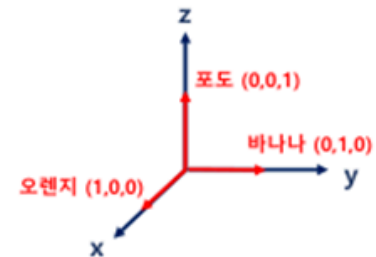
■ 계층적 군집분석

계층적으로 각 데이터를 유사한 군집으로 묶어가며 진행하는 방법

응집형



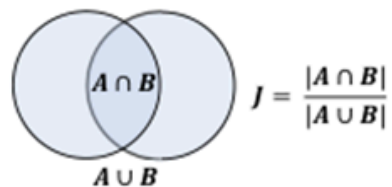
분리형



· 범주형 데이터

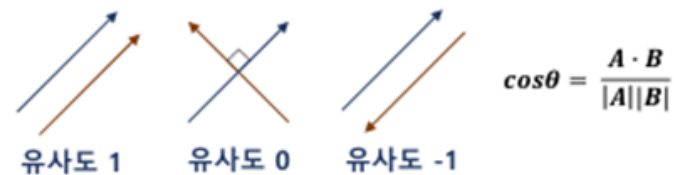
자카드 유사도

두 벡터의 교집합과 합집합의 비

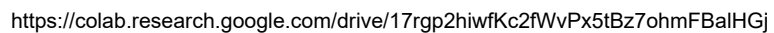


코사인 유사도

두 벡터의 코사인 각도



계층적으로 각 데이터를 유사한 군집으로 묶어가며 진행하는 방법



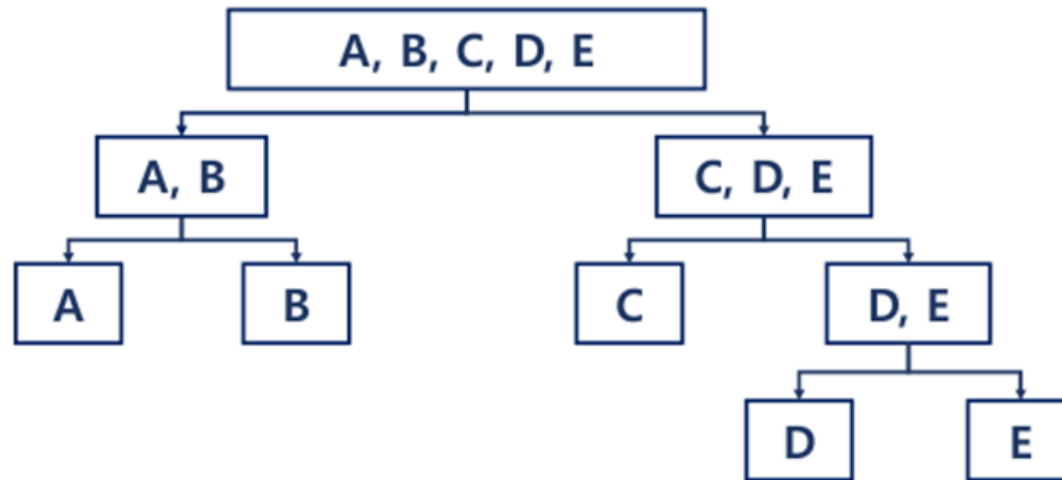
■ 계층적 군집분석

계층적으로 각 데이터를 유사한 군집으로 묶어가며 진행하는 방법

응집형

분리형

전체 데이터를 하나의 군집으로 보고 분할해 나가는 방식
대표적으로 **다이나나(DIANA) 알고리즘** 사용



■ 계층적 군집분석

계층적으로 각 데이터를 유사한 군집으로 묶어가며 진행하는 방법

덴드로그램(Dendrogram)



■ 군집분석 평가지표

실루엣 계수(Silhouette Coefficient)

각각의 데이터별로 그 데이터가 속한 군집 내의 유사도와 인접한 군집의 유사도를 비교하는 지표

[군집 B]

$$b(i) = a(i)$$