07

군집분석 개념

[군집분석 개념]

군집분석의 개념은 컬럼들의 거리를 계산으로 유사성을

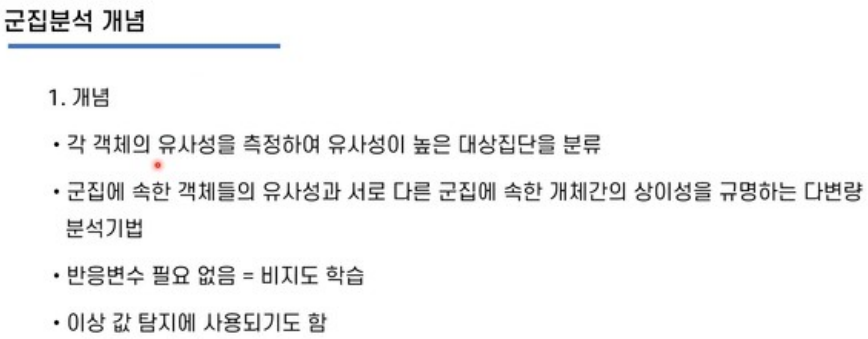
측정합니다.

그리고 유사하다면 같은 그룹, 유사하지 않다면 다른 그룹으로

나누어줄 수 있습니다.

- 유사하다는 것의 기준은 무엇일까요? 바로 거리입니다.

00:10



[군집분석과 차원축소의 차이]

군집분석은 추가적인 자료가 어디에 속할 것인가?

차원축소는 주어진 자료를 축소시키는 것

03:34

군집분석 개념

2. 비교

- 요인분석과 비교: 요인분석은 유사한 변수를 묶어 집단을 설명함
- 판별분석과 비교: 판별분석은 집단이 나누어져 있는 자료를 통해 새로운 데이터를 기존의 집단에 할당함

[거리 공식 종류]

#군집분석의 유사성을 판단하는 거리는 종류가 많습니다.

대표적으로 k-means에는 유클리디언, gmm은 마할라노비스 거리를 사용합니다.

군집분석 개념

다. 거리
관측 데이터 간 유사성이나 근접성을 거리로 판단함 → 어느 군집으로 묶을지 판단

거리종류	설명
유클리디언	데이터 간의 유사성을 측정. 통계적 개념이 내포되어있지 않음. 변수의 산포정도가 감안되지 않음
표준화	해당변수의 표준편차로 척도 변환 후 유클리디안 거리로 계산. 표준화덕에 왜곡을 피함
마할라노비스	통계적 개념이 포함된 거리. 변수들의 산포를 고려해 표준화됨. 그룹에 대한 사전지식 필요(두 벡터 사이의 거리를 표본공분산으로 나누어주어야 함)
체비셰프	두 점의 x좌표 차이와 y좌표 차이 중 큰 값을 갖는 거리
맨하탄	도시에서 건물을 가기 위한 최단거리를 구하기 위해 고안됨
캔버라	두 점 사이의 차이에 대한 절대값을 두 점의 합으로 나눈 값의 합으로 구함
민코우스키	맨하탄거리와 유클리디안 거리를 한번에 표현한 공식