[Data Science] 데이터 사이언스 개념 - 7.비지도 학습 — 나무늘보의 개발 블로그

노트북: 첫 번째 노트북
만든 날짜: 2021-01-21 오전 12:22
URL: https://continuous-development.tistory.com/217?category=833358

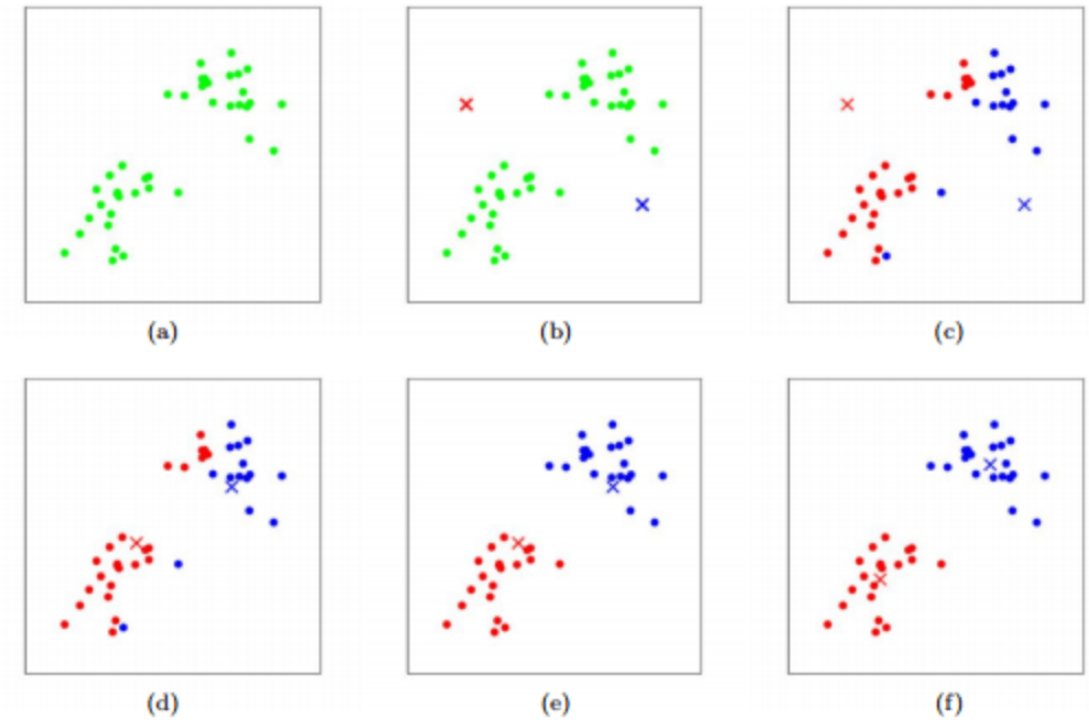
Data Science

[Data Science] 데이터 사이언스 개념 - 7.비지도 학습

2021. 1. 14. 05:26 수정 삭제 공개

비지도 학습

1.K-평균법



k평균법 - 같은 클러스터 내의 데이터 점끼리 거리가 짧아지도록 데이터를 주어진 수의 클러스터로 분류하는 것

비지도 학습의 일종으로 클러스터링이다.
위와 같이 데이터가 어느 그룹에 속할지 결정하는 것이 목표이다.

k 평균법 구현하는 방법

데이터를 몇 개의 클러스터로 나눌지 결정한다.

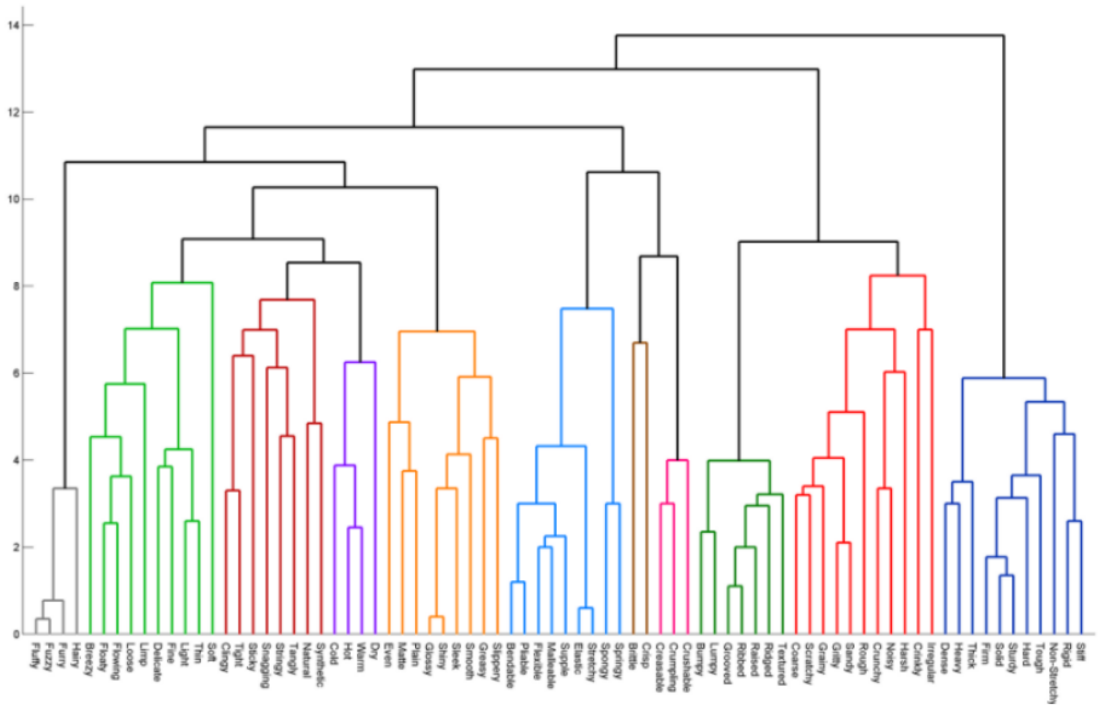
라벨을 랜덤으로 붙인다. 다음으로 각 라벨의 점의 중심을 계산해준다.

큰 라벨처럼 중심점이 정해진다.

다음으로 각 점에 가장 가까운 중심점과 같은 라벨을 다시 칠해준다.

이것을 반복하고 각 라벨의 갱신을 반복해간다.

2.계층적 클러스터링



계층적 클러스터링 - 하나하나의 데이터를 근접한 데이터와 결합함으로써 바텀업 방식으로 클러스터링하는 방법

계층적 클러스터링을 구현 하는 방법

클러스터수를 데이터 수와 같게 설정하고 하나하나의 데이터가 각 클러스터에 속해 있다고한다.

클러스터끼리의 거리를 모두 계산해, 가장 거리가 가까운 2개의 클러스터를 하나로 결합

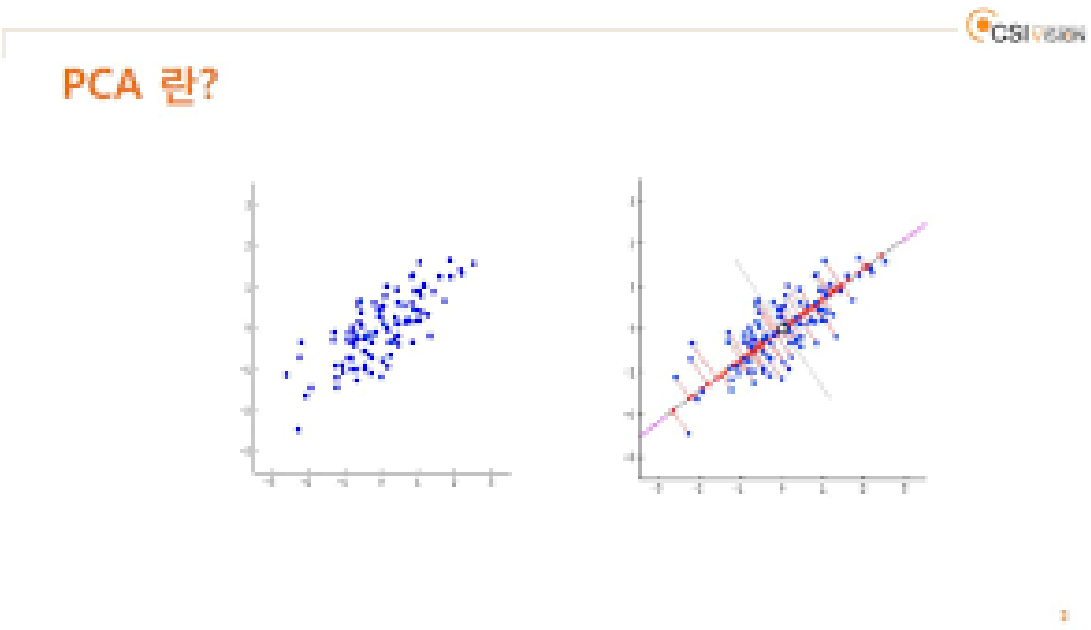
이때 거리를 높이로 해서 어느 클러스터를 결합했는지 기록한다.(덴드로그램)

새로 만들어진 클러스터는 클러스터 내 데이터 중심점을 대표점으로 하여 새로 설정하고 남은 클러스터와 데이터 점과의 거리를 다시 계산

한다

반복

3.주성분 분석

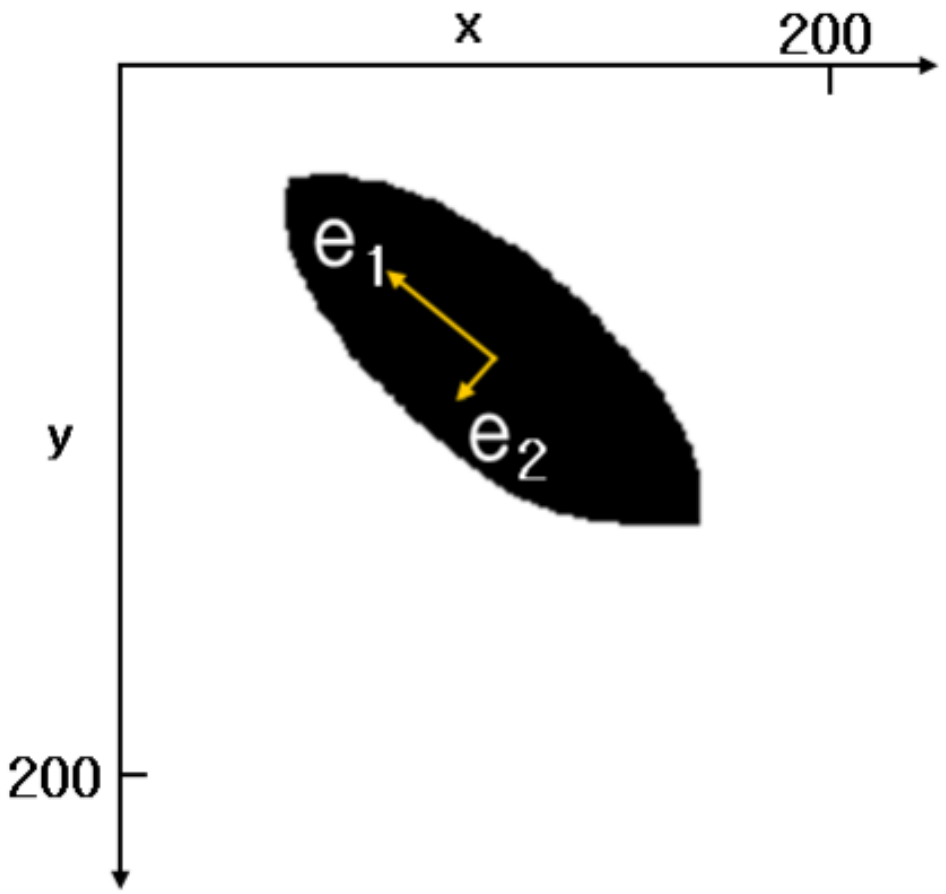


주성분 분석 - 다수의 변수를 소수로 줄여 데이터를 다시 표현

이것을 차원감소라고 부른다.

변수에 상관관계가 없으면 유효한 방법은 아니지만, 주가의 시계열 등 변수 개수와 비교해 분산을 낳는 주요인이 적을 때 매우 효과적인 방법이다.

주성분 분석의 경우 분산을 많이 설명하는 것이 좋은 표현이다.



이 e1과 e2 두개의 벡터로 데이터 분포를 설명하는 것
데이터들의 분산이 가장 큰 방향벡터를 의미한다.

4.주성분 분석과 특잇값 분해

SVD (Singular Value Decomposition)

- 어떤 $n \times m$ 행렬 A는 다음과 같은 형태의 세 가지 행렬의 곱으로 분해(decomposition)할 수 있다.

$$A_{n\times m} = U_{n\times n}\ \Sigma_{n\times m}\ V^T_{m\times m}$$

- $U : n \times n$ 직교행렬, $AA^T = U(\Sigma\Sigma^T)U^T$
- $V : m \times m$ 직교행렬, $A^TA = V(\Sigma^T\Sigma)V^T$
- $\Sigma : n \times m$ 직사각 대각행렬

$$\Sigma = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_m \\ 0 & & \end{bmatrix} (n > m), \qquad \Sigma = \begin{bmatrix} \sigma_1 & & & \\ & \ddots & & \\ & & \sigma_n & 0 \end{bmatrix} (n < m)$$

특잇값 분해 - 행렬 X에 대해서 행렬 분해를 해서 행렬을 대각화하는 방법이다.
주성분 분석과 특잇값 분해는 수학적으로 비슷한 문제를 해결한다.

본 내용은 그림으로 배우는 DataScience 데이터 과학을 참고한 내용입니다

출처: https://continuous-development.tistory.com/210?category=833358 [나무늘보의 개발 블로그]

'Data Science' 카테고리의 다른 글□

K 평균법 계층적 클러스터링 비지도 학습 주성분 분석 특잇값 분해

나아무늘보

혼자 끄적끄적하는 블로그 입니다.