5.8 Unsupervised Learning Algorithms

박달님 03.26.2021

Unsupervised Algorithms

- 특정 입력 (input)에 대하여 올바른 정답이 없는 데이터 집합이 주어지는 경우의 학습
- “특징”만 경험하는 알고리즘
- 지도/비지도 알고리즘 구분 공식적으로 정의되지 않음
- 비공식적 정의: 예제에 주석을 달기 위해 사람의 노동이 필요하지 않은 분포에서 정보를 추출하려는 대부분의 시도
- 밀도 추정, 샘플 추철 학습, 데이터 노이즈 제거 학습, 데이터를 관련 예제 그룹으로 clustering

최상의 표현

- 비지도 학습 작업 = 데이터의 “최상의” 표현을 찾는 것
- x 자체가 표현을 더 간단하게 또는 더 쉽게 접근 할 수 있도록 하기 위한 몇가지 제약 조건을 준수하면서 가능한 한 많은 x에 대한 정보를 보존하는 표현을 찾는 것
- 더 간단한 표현으로 바꿔주는 가장 일반적인 세가지 방법:
  - Lower dimensional representations
  - Sparse representations
  - Independent representations

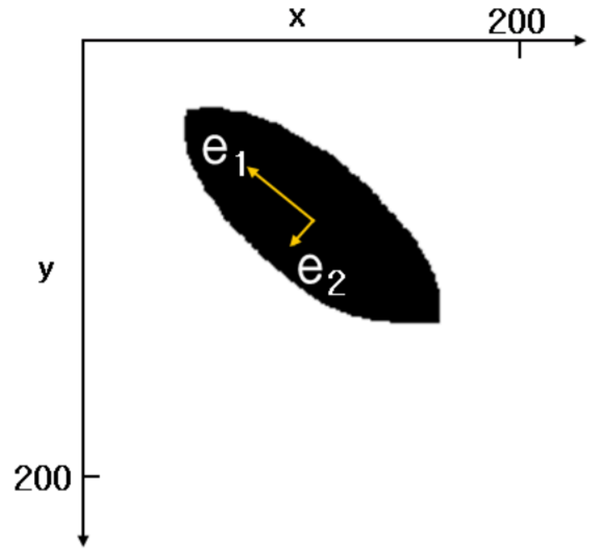
세가지 기준의 설명

- Low-dimensional representations
  - x에 대한 정보를 최대한 작게 압축
- Sparse representations
  - 대부분의 input에 대해 항목이 대부분 0인 representation
  - 차원을 늘려야하므로 표현이 대부분 0이 되더라도 너무 많은 정보를 버리지 않는다
  - 그 결과 - 표현 공간의 축을 따라 데이터를 분산하는 경향이 있는 표현의 전체 구조가 생성
- Independent representations
  - 표현의 차원이 통계적으로 독립적이 되는 데이터의 원인들을 각각의 차원으로 분리

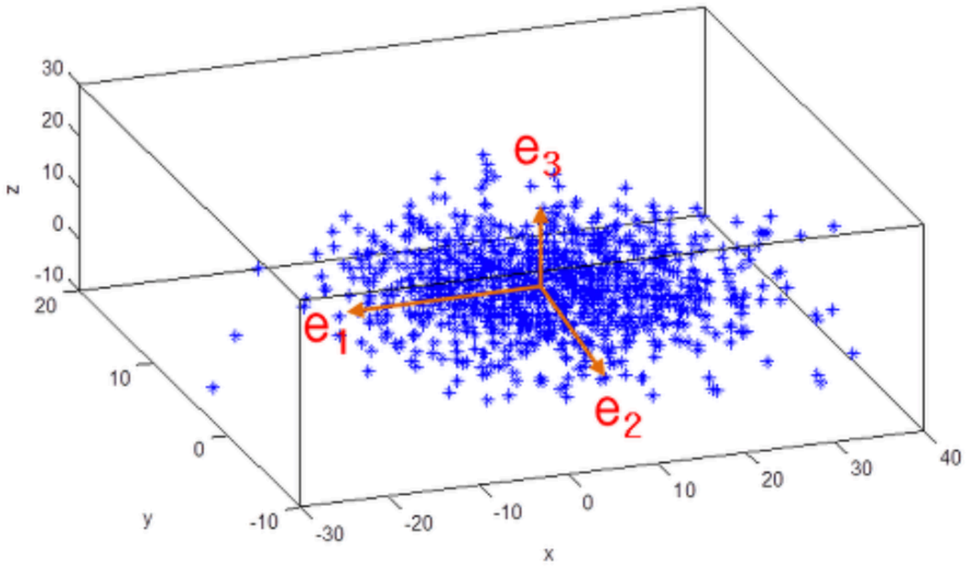
Principal Components Analysis
주성분 분석

- PCA - 데이터 표현을 학습하는 비지도 학습 알고리즘
- 두 가지 기준을 기반
  - 원래 입력보다 차원이 낮은 표현 학습
  - 요소가 서로 선형 상관 관계가 없는 표현 학습
    - 통계적으로 독립적인 학습 표현의 기준을 향한 첫번째 단계
    - 완전한 독립성 달성하려면 표현 학습 알고리즘이 변수 간의 비선형 관계도 제거

Principal Components Analysis



- 분포된 데이터들의 주성분 (Principal Component)를 찾아주는 방법
- 2차원 좌표평면에 n개의 점 데이터 (x1, y1), (x2, y2), …, (xn, yn)들이 타원형으로 분포되어 있을 때 e1, e2 두 개의 벡터로 데이터 분포 설명
- 분포의 주 성분 분석 - 분산이 가장 큰 방향 벡터
- 3차원 점들 - 3개의 서로 수직인 주성분 벡터 반환



K-means Clustering

- 분리형 군집화 알고리즘의 하나
- 각 군집은 하나의 중심 (centroid)를 가짐
- 각 개체는 가장 가까운 중심에 할당, 같은 중심에 할당된 개체들이 모여 하나의 군집을 형성
- K - hyperparameter

$$X = C_1 \cup C_2 \ldots \cup C_K, \quad C_i \cap C_j = \phi$$

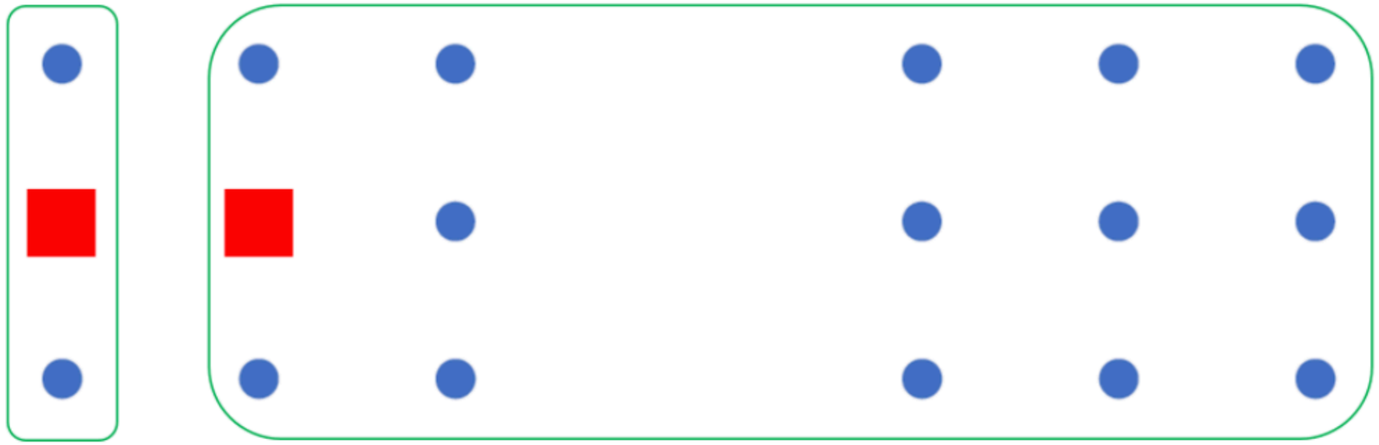
$$argmin_C \sum_{i=1}^{K} \sum_{x_j \in C_i} \|x_j - c_i\|^2$$

학습 과정

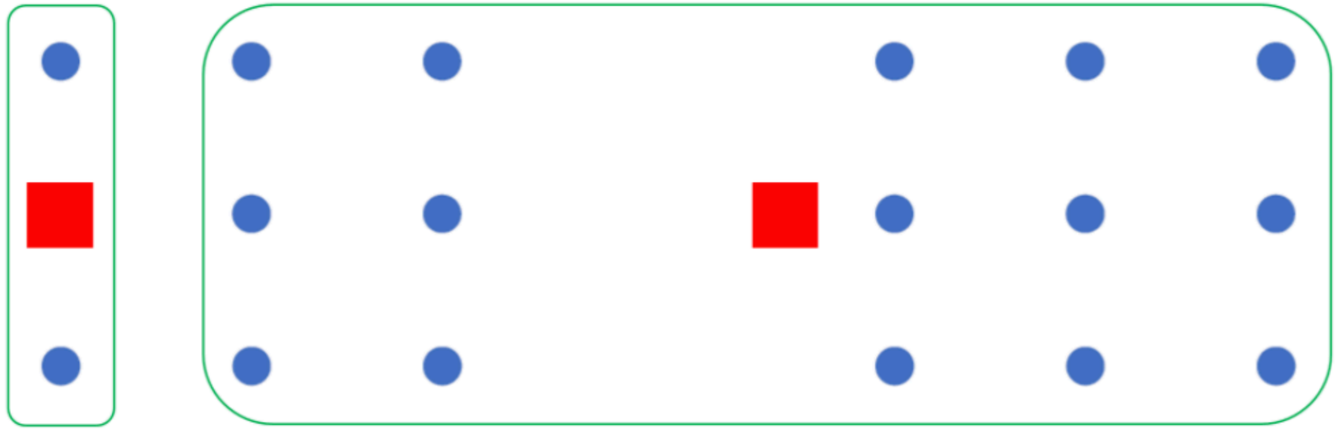
- EM 알고리즘 기반 (Expectation + Maximization)
- 수렴할 때까지 반복
- 1. 각 군집 중심의 위치
- 2. 각 개체가 어떤 군집에 속해야 하는지

초기화

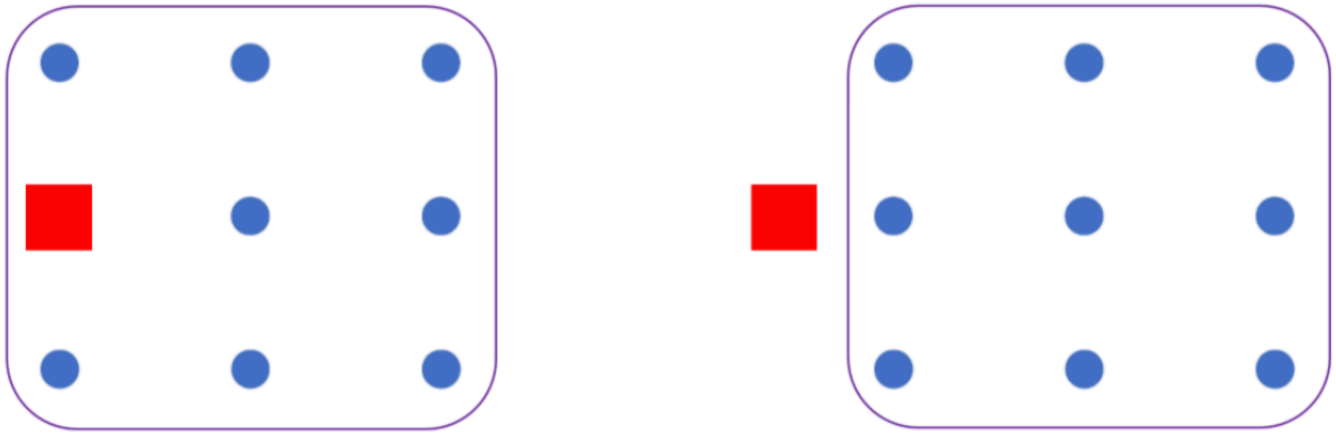
Expectation Step



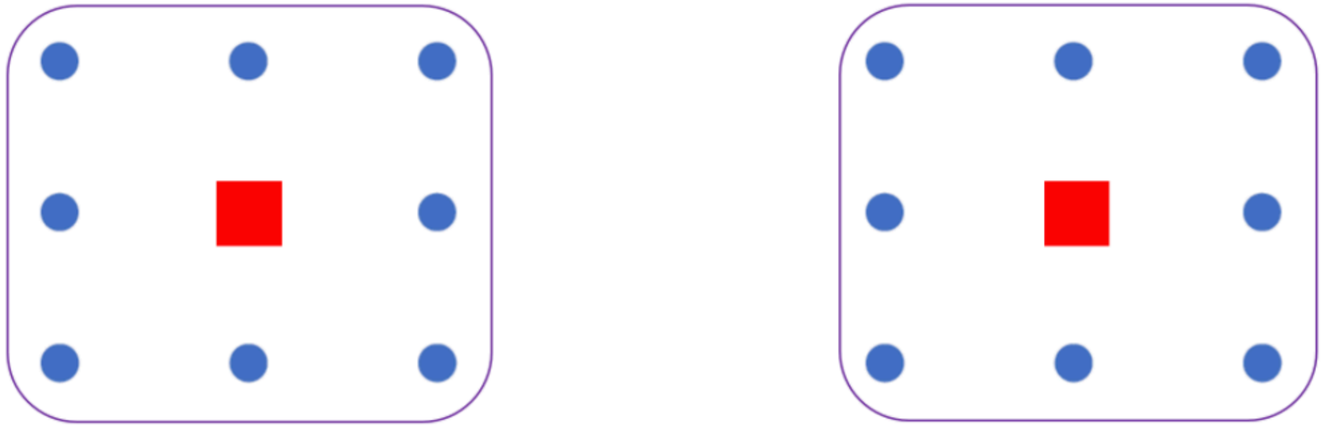
Maximization Step



Expectation Step 2



Maximization Step 2



KC의 단점

- 초기값 위치에 따라 원하는 결과가 나오지 않을 수도 있음
- 군집의 크기/밀도가 다를 경우 제대로 작동하지 않을 수 있음
- 데이터 분포가 특이한 케이스에도 군집이 잘 이루어지지 않음