군집화; 비지도학습

다양한 데이터 형태에 맞는 클러스터링을 진행하기 위해 그만큼의 군집화 알고리즘을 알 필요;

레이블이 없는 데이터 집합을 유사한 데이터들의 그룹으로 나누는 것.

- 동 클러스터 내 객체들의 유사성 요구
- 서로 다른 클러스터 내 객체들과는 이질성 필요

하드클러스터링 : 완전구분 vs 소프트 클러스터링 : 완전 구분까지는 아님.

Algorithms; flat Algorithms

Hierarchical algorithms

DBSCAN

KMEANS Clustering: 주어진 데이터 k개의 클러스터를 중심으로 묶는 알고리즘

1) 임의로 K개의 중심점(centroid)을 설정
2) 각 개체는 가장 가까운 중심에 할당되어 하나의 클러스터를 형성
3) 각 클러스터에 할당된 포인트들의 평균 좌표를 이용해 중심점을 반복적으로 업데이트

각 객체와 그룹 또는 그룹의 중심 간 유클리디안 거리 합이 최소가 되는 방향으로 군집화

EM Algorithm을 기반으로 작동

- EM algorithm: 잠재변수가 포함된 우도함수를 최적화하여 모수(parameter)의 최대
  - 잠재변수: 직접적으로 관찰 또는 측정이 불가능한 변수
- Expectation Step 및 Maxmization Step 반복
- Log likelihood의 기댓값을 계산하는 단계, 기댓값을 최대화하는 모수의 추정값을 구하는 단계

각 클러스터 중심의 위치 및 개체가 속하는 클러스터 탐색 요구(잠재변수 존재)

~ input을 클러스터에 할당한느 작업은 MLE와 같음.

유클리디안 거리로 거리를 측정하되 너무 차원이 높으면 클러스터링 성능 저하

(차원의 저주 : 차원이 증가하면서 학습 데이터 수가 차원 수보다 적어져 성능이 저하)

Kmeans는 outlier에 민감하므로 주의(평균의 한계)

KNN vs KMeans : 두 방식 모두 K개의 점을 지정해 거리를 기반으로 구현하는 알고리즘,.

But 둘은 목적부터 다르다.

KNN : 해당 데이터와 가까이 있는 K개의 데이터를 확인한 뒤 더 많은 데이터가 포함된 범주

KmeansClsutering : 주어진 데이터를 K개의 클러스터 중심으로 군집화

Hierarchical Clustering

- 계층적 트리 모형을 이용해 개별 객체들을 유사한 객체와 통합하는 알고리즘
- 덴드로그램(개체들이 결합되는 순서를 나타내는 트리 형태 구조) 시각화 기능
- 사전에 군집의 수를 정하지 않음.

Agglomerative Hierarchical Clustering

- 모든 개체들 사이의 거리에 대한 유사도 행렬 계산
- 거리가 인접한 관측치끼리 클러스터 형성
- 유사도 행렬 업데이트

Bottom up: 각 데이터 포인트를 순차적 병합하여 클러스터링

Top-Down: 전체를 하나의 클러스터를 보고 분할해 나가는 클러스터링

Distance in Similarity Matrix

1) Euclidean : single, complete, Average, Centrodis
2) Ward's Lickage : 두 군집이 합쳐졌을 때의 오차제곱합의 증가분을 기반하는 계산

Features : 군집의 수를 미리 정하지 않아도 되고, random point에서 시작하지 않아 항상 동일한 결과(클러스터 수는 덴드로그램을 자르는 위치에 따라 상이). 전체적인 군집 파악 간으

But, 데이터가 크면 연산시간이 지나치게 길다.

DBSCAN: 점이 세밀하게 몰려있어 밀도가 높은 부분을 클러스터링하는 알고리즘

점 p에서 거리e내에 점이 minPTs개 있으면 하나의 군집으로 인식하도록 함.

- 어느 군집에도 속하지 않은 outlier는 noise point

Heuristic Approach to Determine eps and minPts

1) minPts의 개수를 k개로 하면, 하나의 점에서 k번째로 가까운 점과의 거리
2) k-dist를 내림차순으로 정렬하며 데이터베이스의 밀도 분포에 대한 정보
3) eps를 k-dist, minPts K로 설정하면 K-dist보다 작거나 같은 모든 코어포인트
4) x축은 모든 포인트에 대해 k-dist를 내림차순 정렬한 포인트, y축은 각 포인트에 대한 k-dist 값

if minPts are too small -> noise도 유효한 포인트로 잘못 구분될 수 있음.

By 논문 -> 2차원 데이터에 대해서는 minPts가 4가 적절

Features

1. 군집의 수를 미리 정하지 않아도 됨
2. 객체의 밀도에 다라 클러스터를 서로 연결하기 때문에 기하학적인 모양의 군집도 생성 가능
3. Noise에 강함
4. 단, 밀도 기반의 군집화

클러스터링에 있어서의 문제 – 척도, 개수, 결과 평가 지표

Proximity measures = 데이터 포인트 간 비유사성 계산 지표

1) Eucildean Distance
2) Manhattan Distance
3) Cosine Similarity

유효성 검증?

Elbow method – 꺾인 점.에서 최적 클러스터 수 k 결정₩

Dunn Index : 클러스터 내 최대 거리에 대한 클러스터 간 최소 거리비

SSE : 각 군집 중심에서 해당 군집 관측치들의 거리 제곱 합

군집 간 분산 최대화를 위해 Silhouette Coefficient(elbow point가 없는 경우)

- 각 군집 간의 거리가 얼마나 효율적으로 분리되어 있는지를 나타내는 지표

Feature Normalization : 데이터를 특정 구간으로 바꾸는 척도법... 비지도학습을 위한 정규화

-