BITAmin

13주차
정규 세션

비타민 10기 5조

최대상, 조은정, 조예진

세션 순서

PART 01

복습세션

-12주차 복습
-복습과제 리뷰

PART 02

진도세션

-Kmeans
-PCA

복습세션 목차

PART 01

군집화 기준

거리

PART 02

군집화 알고리즘

-BIRCH
-OPTICS

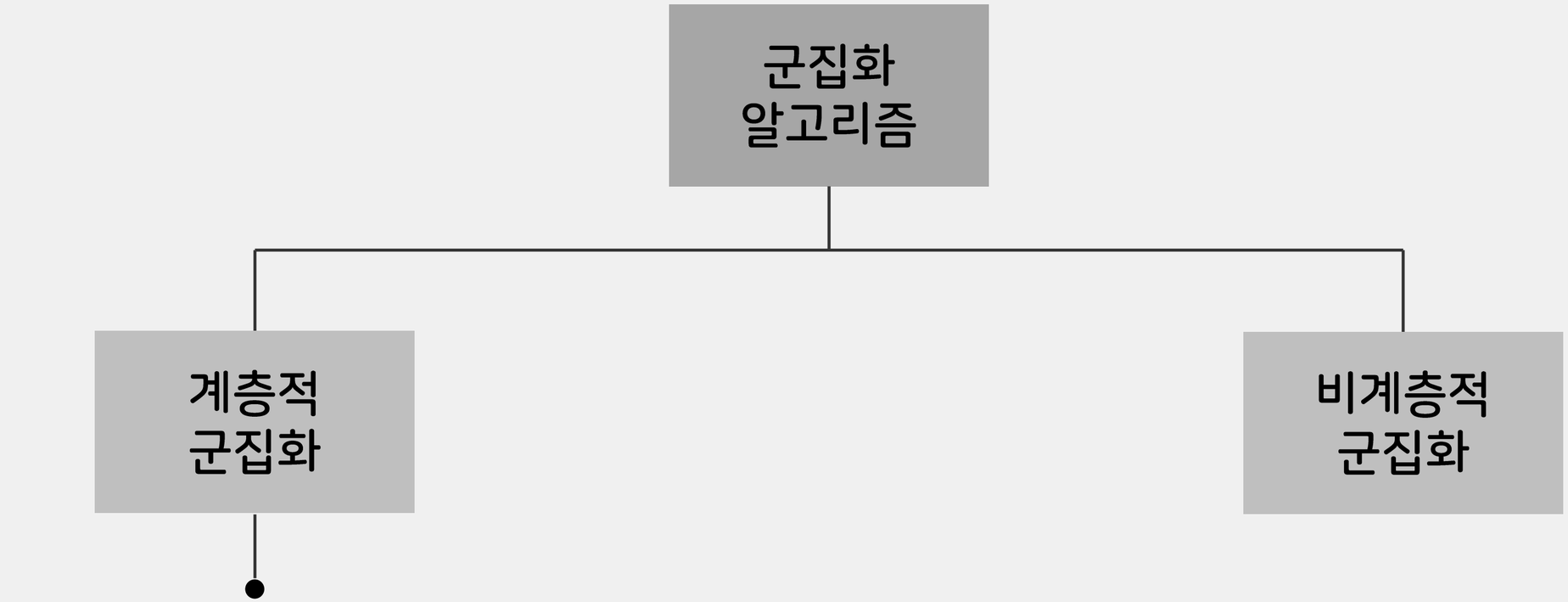
PART 03

군집화 성능 평가

-내부 평가
-외부 평가

PART
01

군집화 기준



군집 간 거리를 기준으로
응집 또는 분리하는 과정

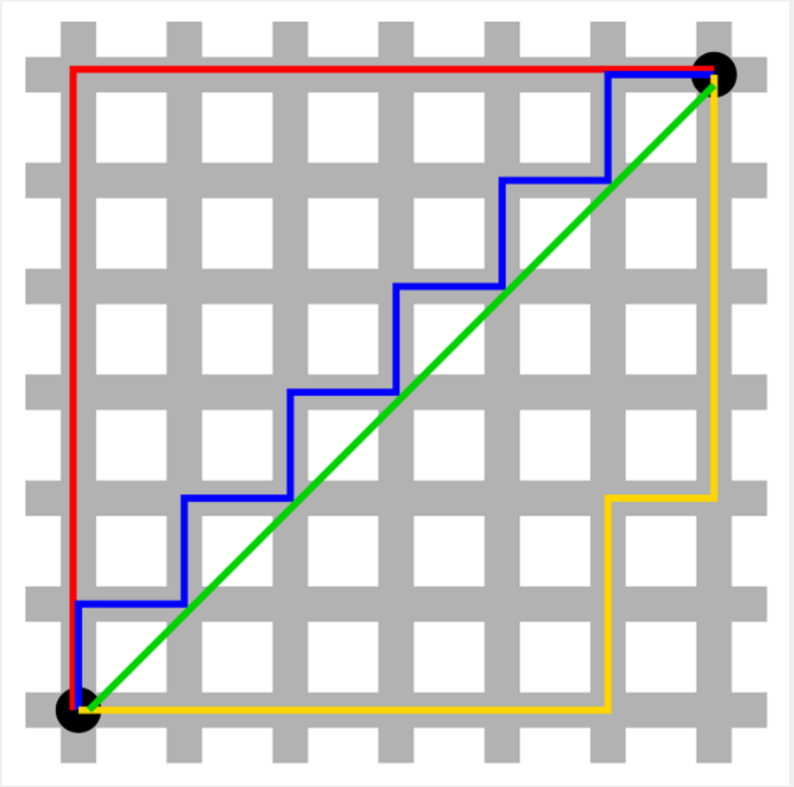
벡터 간 거리 척도(Distance Metric)

- 유클리드 거리(Euclidean Metric)
  - 두 데이터 간의 직선거리
  - 다차원 좌표에서 두 점 사이의 거리를 구할 때 사용
  - L2 distance

$$d(x,y) = \sum_{i=1}^{p} \sqrt{(x_i - y_i)^2} = \sqrt{(x-y)'(x-y)}$$

- 맨해튼 거리(Manhattan distance)
  - 두 데이터 간 절댓값 거리
  - L1 distance

$$d(x,y) = \sum_{i=1}^{p} |x_i - y_i|$$



최단 거리 = 유클리드 거리

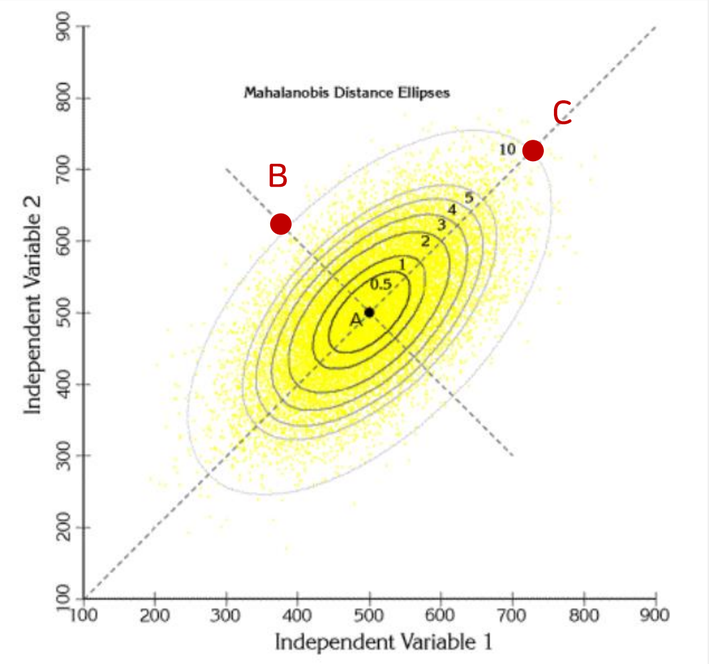
벡터 간 거리 척도(Distance Metric)

- 민코우스키 거리(Minkowski distance)
  - 유클리드 거리와 맨해튼 거리를 일반화한 거리
  - 값이 0에 가까울수록 유사성이 큼
  - m=1인 경우 맨해튼 거리, m=2인 경우 유클리드 거리

$$d(x,y) = \left[\sum_{i=1}^{p} |x_i - y_i|^m \right]^{\frac{1}{m}}$$

- 마할라노비스 거리(Mahalanobis distance)
  - 변수의 분산과 상관성을 고려한 거리 측정 방법
  - 값이 0에 가까울수록 유사성이 큼
  - 평균과의 거리가 표준편차의 몇 배인지 나타내는 값

$$d(x,y) = \sqrt{(x-y)' S^{-1} (x-y)}$$



d(A,B) > d(A,C)

벡터 간 거리 척도(Distance Metric)

(참고)

- 코사인 유사도(Cosine similarity)

- 두 개의 벡터값에서 코사인 각도를 구하는 방법, 1에 근접할수록 유사

$$\cos(\theta) = \frac{A \cdot B}{||A||||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

- 쿨백-라이블러 발산(KL divergence)

- 두 확률 분포의 차이 확인, 두 확률분포가 동일하면 0이 됨

$$D_{KL}(P\|Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$$

$$= \sum_i P(i) \log P(i) - \sum_i P(i) \log Q(i)$$

$$= -H(P) + H(P, Q)$$

- 해밍거리(Hamming distance)

- 거리가 정확히 같은 지 확인

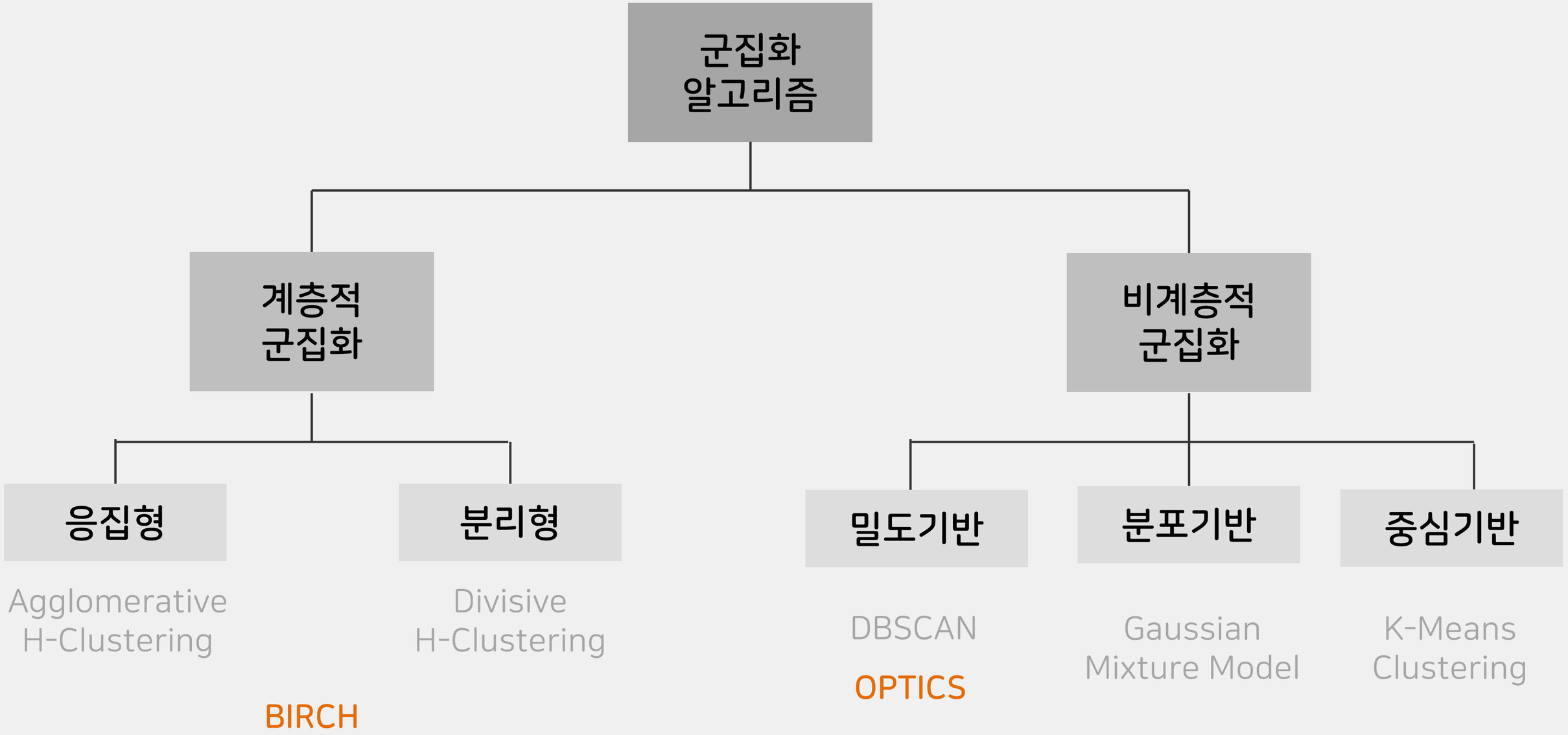
- 맞춤법 검사에 주로 사용

$$d_H(s,t) = \left|\left\{i \in \{0, 1, \ldots, n-1\} : s_i \neq t_i\right\}\right| \in \{0, 1, \ldots, n\}$$

MIRICOMPANY

PART
02

군집화 알고리즘


군집화
알고리즘
계층적
군집화
비계층적
군집화
응집형
분리형
밀도기반
분포기반
중심기반
Agglomerative
H-Clustering
Divisive
H-Clustering
DBSCAN
OPTICS
Gaussian
Mixture Model
K-Means
Clustering
BIRCH

BIRCH

Balanced Iterative Reducing and Clustering Using Hierarchies

등장 배경

BEFORE

- 입출력 작업 비용 ↑
- 데이터 특성 반영 X
  (데이터 포인트 간 거리에 가중치 부여 X)
- 대규모 데이터셋에서 효율성 ↓

BIRCH

- 입출력 작업 비용 ↓
- 데이터 특성 반영
  (데이터 포인트 간 거리에 가중치 부여)
- 전체 데이터셋이 필요하지 않은 증분 방식
  => 대규모 데이터셋에서 효율성 ↑

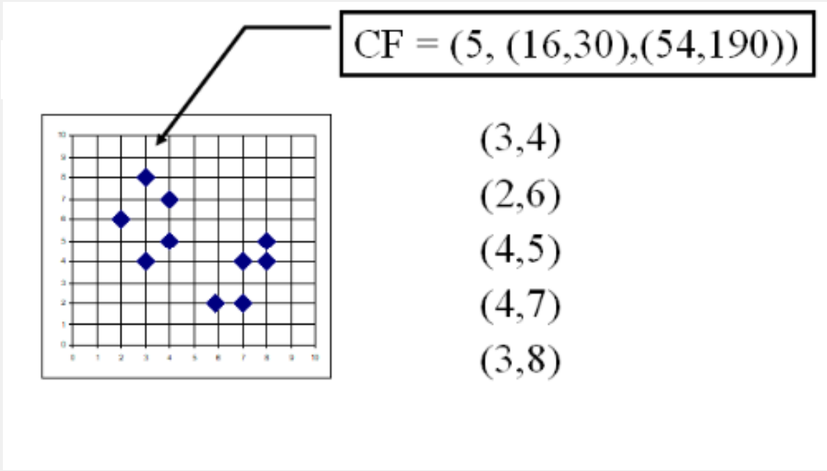
BIRCH

Balanced Iterative Reducing and Clustering Using Hierarchies

개념

- 계층적 군집화 알고리즘
- CF(Clustering Feature) tree와 계층적 데이터 구조를 증분 방식으로 구성

Clustering Feature (CF): CF = (N,LS, SS)
-N: 데이터 포인트의 수
-LS: 데이터 포인트의 합
-SS: 데이터 포인트의 제곱합



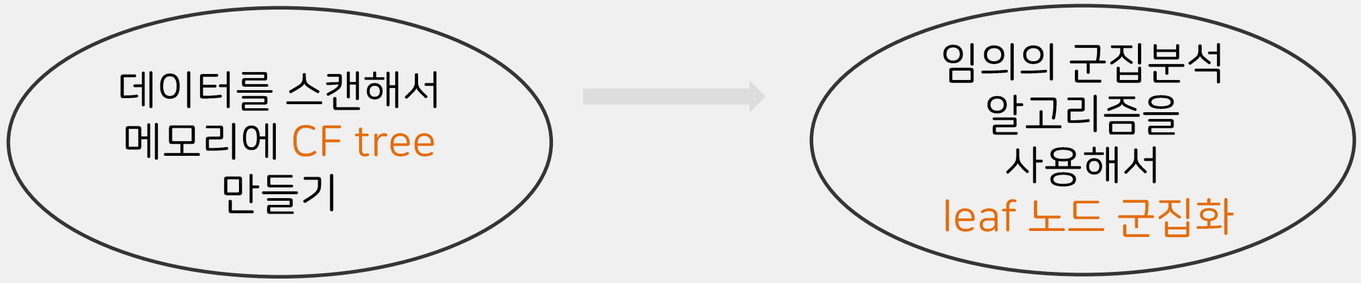
CF Tree

- 높이 균형 트리
- 파라미터
  - 분기 계수(branching factor): 자식 노드 수의 최댓값
  - 임계값(threshold): 서브 클러스터의 최대 지름을 leaf 노드에 저장 (cluster 크기 조절)

BIRCH

Balanced Iterative Reducing and Clustering Using Hierarchies

과정



입력 데이터의 각 포인트에서
- 가까운 leaf entry 찾기
- leaf entry에 포인트 추가하여 CF 업데이트
- 만약 'entry 지름 > 임계값' -> leaf 분할

단점

- 데이터 삽입 순서에 민감
- Leaf 노드의 크기를 고정하기 때문에 군집이 부자연스러울 수 있음 (둥근 형태로 제한)

OPTICS

Ordering Points to Identify the Clustering Structure

등장 배경

실제 데이터 집합의 주요 특성을 전역 밀도 변수로 특징지을 수 없는 한계 극복
-> 밀도 기반의 클러스터 순서화(density-based cluster ordering) 제안

- 데이터 공간의 다른 영역에 존재하는 클러스터를 발견하기 위해서 다양한 지역적 밀도 계산
- 매개변수 " minPts " 에 대해 고밀도에 관한 밀도 기반 클러스터는 저밀도에 관한 밀도기반 클러스터에 완전히 포함됨
- 일관성 있는 결과를 위해 특정 순서 지정

개념

- 밀도 기반의 클러스터를 탐색하기 위한 알고리즘
- DBSCAN을 확장한 형태 -> 다양한 밀도의 클러스터 처리 개선
- DBSCAN과 달리 차별화된 밀도 개념 적용
- 거리 매개변수를 동시에 처리 -> 다른 밀도의 클러스터를 동시에 탐색

OPTICS

Ordering Points to Identify the Clustering Structure

장점

- 밀도와 순서 모두 고려하기 때문에 더 효과적인 클러스터링 가능
- 데이터를 엄격하게 구분하지 않아서 사용하기 편리함

단점

- DBSCAN에 비해 비용 ↑
- DBSCAN보다 쿼리가 복잡 -> 속도 ↓

MIRICOMPANY

PART
03

군집화
성능 평가 ✨

군집화 성능 평가

군집화 성능 지표

내부 평가
Internal evaluation

- 데이터 집합을 클러스터링한 결과 그 자체로 평가하는 방식
- 군집 내 높은 유사도 (high intra-cluster similarity), 군집 간 낮은 유사도 (low inter-cluster similarity)를 높게 평가함
- 단점: 평가 점수가 높다고 실제 참값 (ground truth) 에 가까운 것은 아님
- 종류: 실루엣 계수, Dunn Index, Davies-Bouldin Index

외부 평가
External evaluation

- 클러스터링에 사용되지 않은 데이터로 평가
  => 클러스터링의 결과물을 미리 정해진 모범 답안 또는 평가 기준을 이용해서 정확도 평가
- 클러스터링 결과와 미리 정해진 결과물 간의 유사도 측정
- 종류: 랜드 측정, F 측정, 자카드 지수

실루엣 계수
Silhouette Coefficient

실루엣 계수 s(i)

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

- a(i) : i와 같은 군집에 속한 원소들의 평균 거리
- b(i) : i와 다른 군집 중 가장 가까운 군집까지의 평균 거리

특징

- -1와 1 사이의 값을 가짐
- 1에 가까울수록 데이터가 올바른 클러스터에, -1에 가까울수록 잘못된 클러스터에 분류된 것
- 1에 가까울수록 군집 간의 거리가 멀고, 0에 가까울수록 군집 간의 거리가 가까움

Dunn Index

$$D = \frac{\min_{1 \le i < j \le n} d(i,j)}{\max_{1 \le k \le n} d'(k)}$$

d(i, j): 클러스터 간의 거리
d'(k): 클러스터 내 거리

특징

- 군집 간 최소 거리와 군집 간 최대 거리의 비율
- D값이 클수록 성능 좋음

랜드 측정
Rand measure

랜드 지수 (Rand Index, RI)

$$\text{Rand Index} = \frac{a+b}{{}_N C_2}$$

a : 클러스터 내에서 동일하게 짝지어진 쌍의 개수
b : 클러스터 내에서 동일하게 짝지어지지 않은 쌍의 개수
N: 데이터 개수

특징

- 0부터 1 사이의 값을 가짐
- 1에 가까울수록 좋음

한계

- 클러스터 수가 커지면 b 값이 커짐 -> 성능 평가가 정확하지 않음
- 무작위로 클러스터링한 경우에도 랜드 지수가 높게 나옴

랜드 측정
Rand measure

조정 랜드 지수 (Adjusted Rand Index, ARI)

$$\mathrm{ARI} = \frac{\mathrm{RI} - \mathrm{E}[\mathrm{RI}]}{\max(\mathrm{RI}) - \mathrm{E}[\mathrm{RI}]}$$

무작위 군집화에서 생기는 랜드 지수의 기댓값과 분산 재조정

특징

- 0부터 1 사이의 값을 가짐
- 1에 가까울수록 좋음

감사합니다