8주차 비지도학습 - 군집화 발제

K-Means Clustering은 데이터 포인트와 중심점 사이의 거리를 기반으로 각 군집의 기준을 설정하는 방법이다.

- K는 군집의 개수를 의미하고 Means는 군집에 선택된 데이터 포인트들의 평균(중심점)을 의미한다.

1. 군집의 개수(K)를 설정한다. 군집의 개수를 어떻게 설정하느냐에 따라 결과가 상당히 다르게 나타날 수 있으나, 정답 레이블이 없기에 최적의 군집을 알기 어렵습니다. 그렇기에 여러 번 실험을 반복하면서 최적의 군집을 찾거나 군집 평가 지표를 활용해야 한다.
2. K개의 초기 중심점(Centroid, 무게중심)을 설정한다. 여기서도 마찬가지로 초기 중심점을 어떻게 설정하느냐에 따라 결과가 크게 달라질 수 있다.
3. 주어진 모든 데이터를 거리상 가장 가까운 군집의 중심점으로 할당한다. 간단하게 유클리디안 거리를 계산한다. (다차원에서의 두 점 사이의 거리를 계산하기 위해 차원별로 차이를 계산한 뒤 모두 제곱하여 더하고 루트를 취한 것)

$$\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

4. 모든 데이터의 군집 할당이 끝나면 군집의 내 데이터들의 평균점을 새로운 중심점으로 갱신한다.
5. 더 이상 중심점이 갱신되지 않을 때까지 3~4단계를 반복한다.

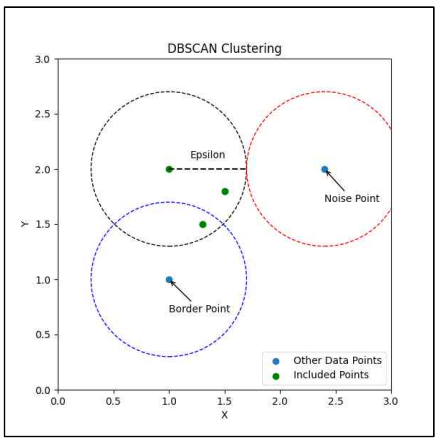
K-Means 알고리즘의 장점은 직관적으로 이해할 수 있고 처리 속도가 빨라 많은 양의 데이터를 다룰 수 있다는 점이다.

단점은 모든 특성들이 연속적이어야 가능하며, 군집의 개수 K, 초기 중심점 설정 등에 따라 결과가 크게 달라질 수 있다는 것이다. 군집의 밀도나 크기가 다른 경우, 데이터 분포가 특이한 경우 군집화가 잘 되지 않다.

DBSCAN은 Density-Based Spatial Clustering of Application with Noise의 약자로, 밀도를 기반으로 군집화하는 방식이다.

- DBSCAN에서 데이터 포인트를 기준으로 지정된 반경 내에 있는 데이터 포인트의 개수에 따라 군집이 형성된다. 여기서 설정할 수 있는 값은 반경(Epsilon)과 반경 내 최소

데이터 포인트 수(minPts)다.



1. 주어진 거리 범위(Epsilon) 내에서 포인트의 이웃들을 찾는다.
2. 도달 가능한 포인트를 추가하며 클러스터를 확장해나간다.
3. 모든 데이터 샘플에 대해서 클러스터를 할당한다.