DBSCAN Algorithm

가) DBSCAN Algorithm이란

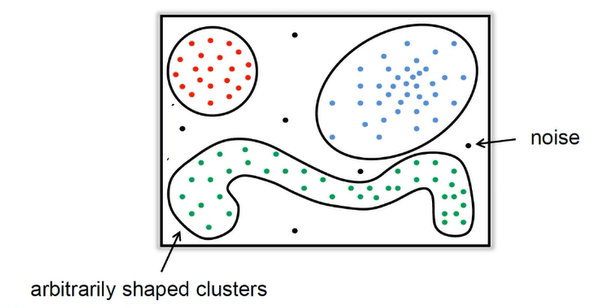
밀도기반 클러스터링 알고리즘으로, data point들의 밀도를 고려하여 클러스터링을 수행하는 알고리즘이다. 아래와 같은 특징을 지닌다.

1. 구형 형태의 클러스터만을 찾을 수 있는 k-means와 달리, 임의의 모양의 클러스터를 찾을 수 있다.

2. 특정한 하나의 개체를 반드시 클러스터에 할당해야 했던 기존과 달리, 어떤 클러스터에도 존재하지 않는 객체(=noise)를 만들 수 있다.

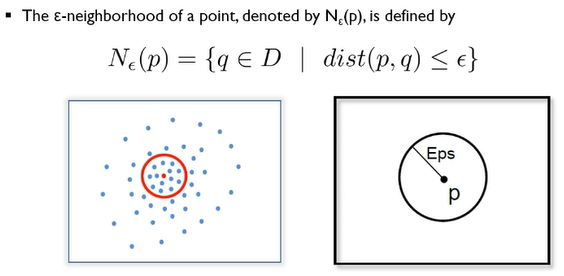
3. 클러스터는 data point들의 집합으로 높은 밀도를 가지는 반면, noise point 주위의 밀도는 매우 낮다.

→ 따라서 이러한 ‘클러스터’와 ‘노이즈 포인트’들의 특징들을 정형화해 유효한 클러스터의 집합을 찾는 것이 목적이다.



나) DBSCAN Algorithm의 정의들

정의 1. Nϵ(p): Eps-neighborhood of a point (=포인트(p)의 Eps-neighborhood)  
→ 포인트 p를 기준으로 반경 Eps안에 있는 모든 데이터 포인트 q들은 Eps-neighborhood이다.



✔ Naïve Approach: 특정 점을 기준으로, 각 데이터 포인터들의 Eps-neightborhood가 최소 포인트갯수(MinPts) 이상이면 이들의 집합은 유효한 클러스터일 것이다.

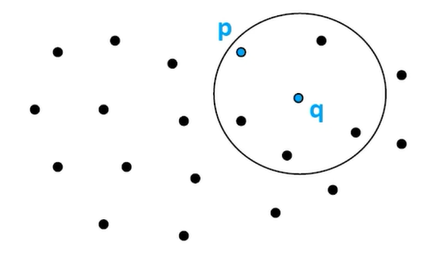
✔ Problem of Naïve Approach:

* 클러스터 내부에는 두 종류의 점이 존재한다.  
  (1) core points: Points inside of the cluster.  
  (2) border points: Points on the border of the cluster.
* 일반적으로 border point의 Eps-neighborhood는 core point에 비해 현저히 적은 점을 포함한다.

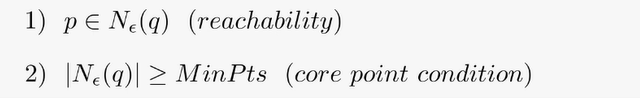
→ 따라서 동일 클러스터에 속하는 모든 포인트를 포함하기 위해선 MinPts를 상대적 낮게 설정해야 한다. 허나 해당 설정값은 각 클러스터에 특성화된 값이 아니라는 문제가 존재한다.

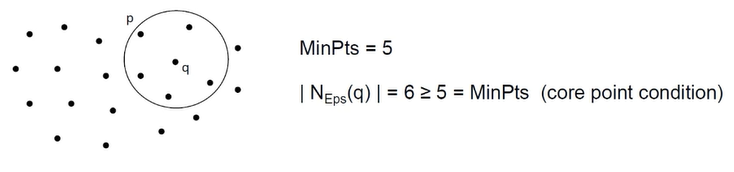
✔ Better Approach:

* 클러스터 C에 포함된 모든 점 p에 대하여, 기준점 q가 C에 존재한다. 즉, core point(q)를 중심으로 Eps-neightborhood 안에 점 p가 존재한다. (Border points are connected to core points)
* Nϵ(q)는 MinPts point이상의 점들을 가진다. (Core points = high density)



정의 2. Directly density-reachable: 아래의 경우, 포인트 p는 q에 directly density-reachable이다.

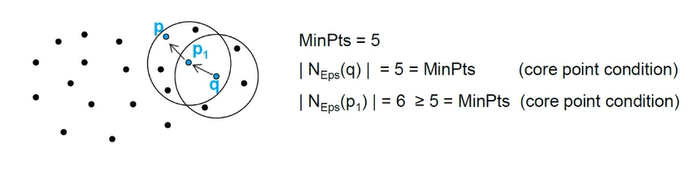




→ 포인트 p를 기준으로 반경 Eps안에 있는 모든 데이터 포인트 q들은 Eps-neighborhood이다.  
→ 이때, directly density-reachable은 core point 쌍에서는 대칭적이나 일반적으로는 비대칭적이다.

정의 3. Density-reachable: 아래의 경우, 포인트 p는 q에 density-reachable이다.



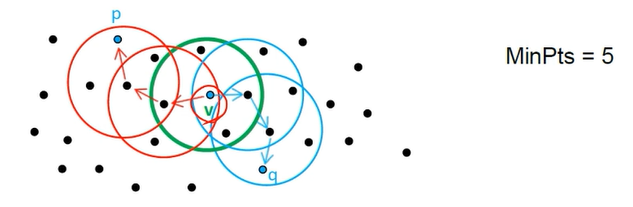


→ p1이 q로부터, p가 p1으로부터 direct-density-reachable이라면 p는 q에 density-reachable하다.

→ 이때, density-reachable은 전이적이지만 대칭적이지는 않다.

정의 4. Density-connected: 아래의 경우, 포인트 p는 q에 density-connected이다.





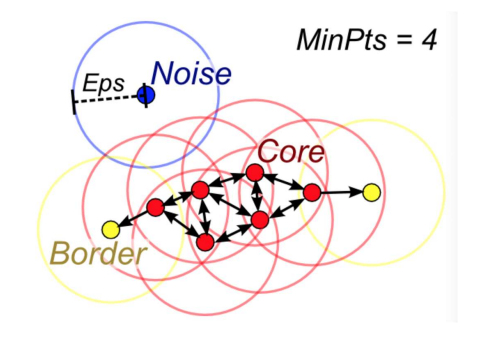
→ 점 p와 q가 density-reachable한 공유 점 v를 가질 때, p는 q에 density-connected이다.   
→ 이때, density-connected는 대칭적 관계를 가진다.

정의 5. Cluster: D를 포인터들의 데이터베이스라고 할 때, 클러스터 C는 아래와 같이 정의된다.

* 모든 p, q에 대해, p∈C이고 q가 p에 density-reachable하면 q∈C이다. (Maximality)
* 모든 p, q에 대해, p가 q에 density-connected이다. (Connectivity)

정의 6. Noise: C1, C2, ...Ck를 D의 클러스터라 할 때, 어떤 클러스터에도 속하지 않는 point들을 noise라 한다.

다) DBSCAN Algorithm의 핵심 용어 정리

- Eps (=Epsilon): 기준점으로부터의 반경 거리를 말한다.

- MinPts: 반경 Eps 내에 존재하는 점의 수를 말한다.

- Core: 기준점 p에서부터 Eps내의 점이 MinPts개 이상일 때, 점 p를 Core(=core point)라고 한다.

- Border: 클러스터에는 속하지만 스스로 core point가 안되는 점을 Border(=border point)라고 한다.

- Noise: 어느 클러스터에도 속하지 않는 outlier point를 말한다.

라) DBSCAN Algorithm 단계 과정

1. 임의의 오브젝트, 포인트 P 선택

2. Eps와 MinPts를 만족하는 P와 연결된 모든 포인트를 탐색

3-1. 만약 P가 Core라면 새로운 클러스터를 생성

3-2. 만약 P가 Border 또는 Noise라면, 데이터베이스 내의 다른 오브젝트 P로 변경

4. 모든 오브젝트를 방문할 때까지 반복 후, 없을 경우 과정 종료.

마) Eps, MinPts 변경 후 실행 결과 → Eps : 0.049, MinPts : 3로 설정하였음.

