DBSCAN 강의

출처

- [Korea University] Multivariate Data Analysis

- https://godongyoung.github.io/머신러닝/2019/07/15/HDBSCAN-이해하기-(with-python).html

Density-based Clustering

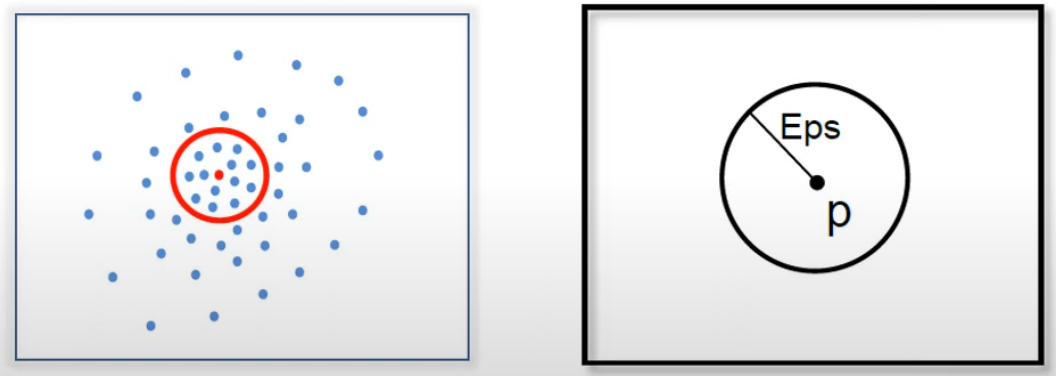
- 밀도 기반 클러스터링 특징
  - 임의의 모양의 클러스터 찾을 수 있음.
  - 어떤 클러스터에도 할당되지 않는 noise가 있음.
- Idea
  - 군집 내는 밀도가 높을 것임.
  - 노이즈의 밀도는 낮을 것임.
- Purpose
  - 유효한 클러스터를 찾기 위해서 clusters와 noise points의 특징을 정량화하자

정의

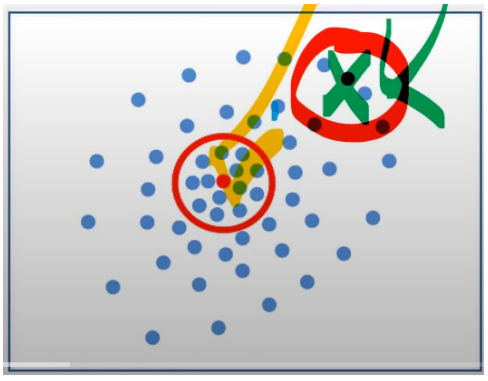
1. ϵ-neighborhood of a point

   - 각 포인트마다 입실론 이웃이 있음.
     ⇒
     $N_{\epsilon}(p)$ 로 표현

$$N_\epsilon(p) = \{q \in D \mid dist(p,q) \leq \epsilon\}$$



  - p점을 기준으로 일정 반경(Eps) 내에 존재하는 점들
- Naive Approach: 특정 클러스터에 속하기 위해서 입실론 네이버후드가 일정 개수 이상인 것들을 모아두면 군집이 될 것
  - 클러스터 내에 임의의 점을 찍어도 eps 내에 일정 개수 이상의 점들이 존재해야 한다는 것.
  - 문제는 core point는 쉽게 만족시킬 수 있지만 border points는 만족시키기 어려움.



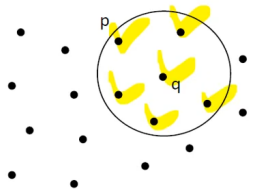
- Better idea
  - 클러스터 C에 속한 점p, q가 있음. q의 eps안에 p가 있음.
    Border points are connected to core points.
  - $N_\epsilon(q)$는 최소한의 포인트를 포함하고 있음
    Core points = high density

2. directly density-reachable

- p는 q로부터 directly density-reachable하다

1) $p \in N_\epsilon(q)$ $(reachability)$

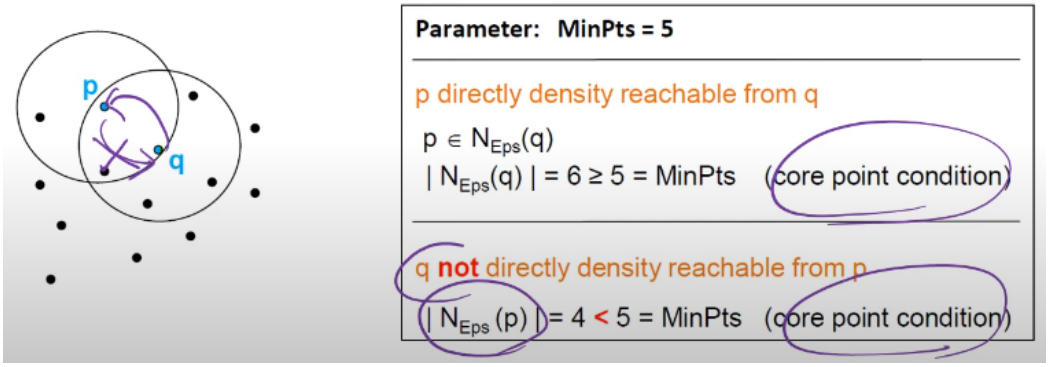
2) $|N_\epsilon(q)| \geq MinPts$ $(core\ point\ condition)$



MinPts = 5

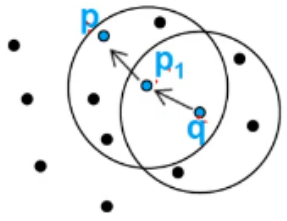
$| N_{Eps}(q) | = 6 \geq 5 = MinPts$ (core point condition)

- Property
  - core points쌍은 서로 direct density-reachable함.
  - core point와 border point 끼리는 그게 아님.



3. density-reachable

- p1이라는 연결고리가 있음. 즉, p1, p2, ..., ps으로 점들이 d.d.s 하게 이어져있다면 시작점p와 끝점q는 density-reachable하다는 내용



MinPts = 5

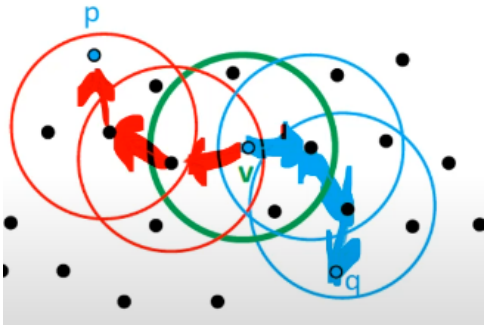
$| N_{Eps}(q) | = 5 = MinPts$ (core point condition)

$| N_{Eps}(p_1) | = 6 \geq 5 = MinPts$ (core point condition)

- p1은 q로부터 d.d.r
- p는 p1으로부터 d.d.r
- 질문1) p는 q로부터 d.d.r한가?→ No

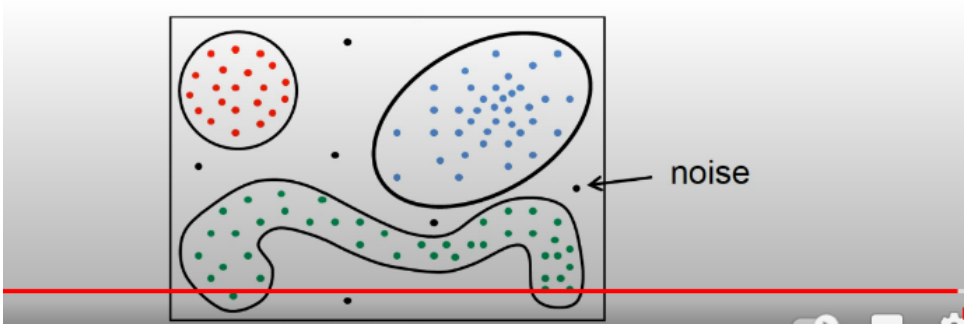
- 질문2) p는 q로부터 d.r 한가? →Yes

4. density-connected
    - p는 v로부터 d.r하고 q는 로부터 d.r할때 p와 q는 density-connected 함.
    - border point끼리 연결하기 위한 정의



5. Cluster
    - 임의의 점부터 시작하여 density-connected될 수 있는 점들을 모두 확장함. d.c 할 점이 없으면 클러스터를 종료시키고 새로운 점으로 이동하여 새로운 클러스터를 만듬.
    - 밀도가 낮은 지역에 홀로 남겨진 점들은 어떠한 점과도 density-connected되어 있지 않기 때문에 noise로 취급됨.

- (1) For all p, q ∈ D: If p ∈ C and q is density-reachable from p with regard to the parameters ε and MinPts, then q ∈ C (Maximality)
- (2) For all p, q ∈ C:The point p is density-connected to q with regard to the parameters ε and MinPts (Connectivity)



Algorithm

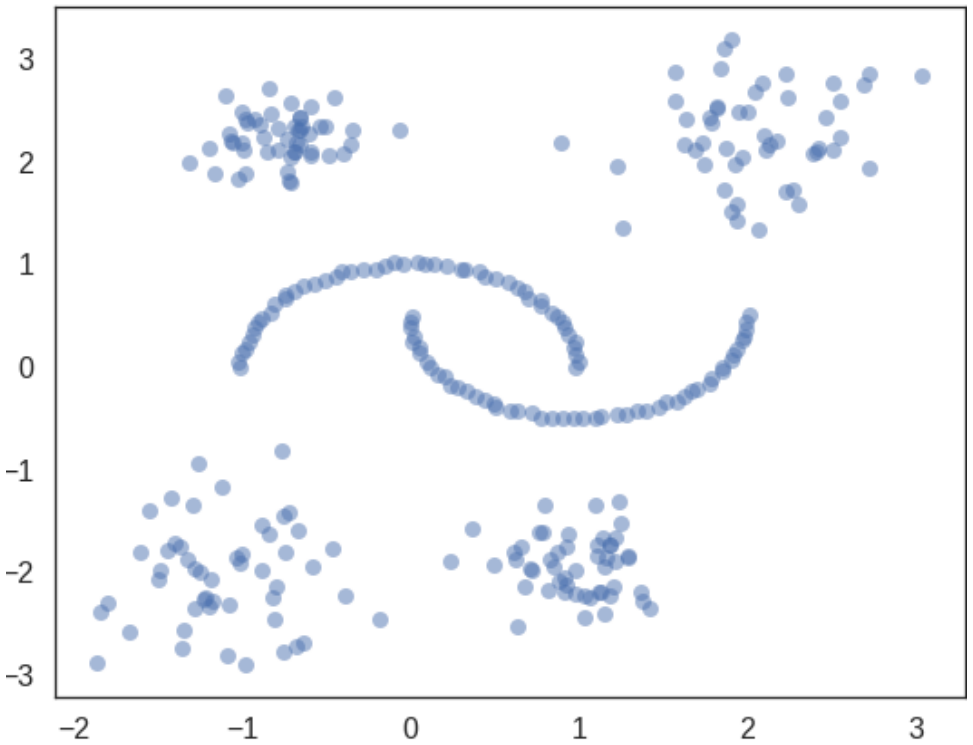
- input : N개의 object
    - 하이퍼파라미터 : ϵ (얼마의 반경을 볼 것인지), MinPts (클러스터내 최소 점개수)

        - 시각화를 통한 정성적 판단 ok, 실루엣지표 및 cluster 지표를 통해 정량적 기준도 ok
- output: Cluster
- 어느 포인트로 시작해도 똑같음(랜덤성 없음)
- p 점으로부터 density-reachable한 모든 점들을 수집하는데,
    - p가 core point라면 클러스터가 형성되고
    - p가 border point라면 density-reachable한 포인트 없이 다른 점으로 이동하게 됨.
- 모든 point가 끝날때까지 이 과정이 진행됨.

DBSCAN의 단점

- local density에 대한 정보를 반영해줄 수 없음(지역적 밀집 정도를 세밀하게 반영하는데 한계가 있다는 의미)
  → 밀도가 점진적으로 변화하는 데이터셋에서는 DBSCAN이 클러스터의 경계를 명확하게 정의하는 데 어려움
- 데이터들의 계층적 구조를 반영한 clustering이 불가능

⇒ 이를 개선한 알고리즘 HDBSCAN



- 반원형 데이터들은 **매우 오밀 조밀**하고,
  좌측 상단과 우측 하단의 원들은
  밀도가 낮은 타원

  우측상단과 좌측하단는 **더더욱 밀도가 낮은 타원형태**(분산이 4배)
- 이러한 데이터는 dense가 각기 달라, 만약 반원의 기준에 맞추게 되면 타원 데이터들은 모두 noise로 처리가 되거나 이상한 클러스터에 속하게 됨.