Clustering and DBSCAN – Density-Based Spatial Clustering of Applications with Noise

Thach Nguyen

Institute of Mathematic and Stochastic

Outline

- Introduction
- DBSCAN Algorithms
- Evaluation benchmark
- Demo

Introduction

- Why clustering?

- Limitations of K-mean and other algorithm.

- Density-based Clustering locates regions of high density that are separated from one another by regions of low density.
  - Density = number of points within a specified radius (Eps)

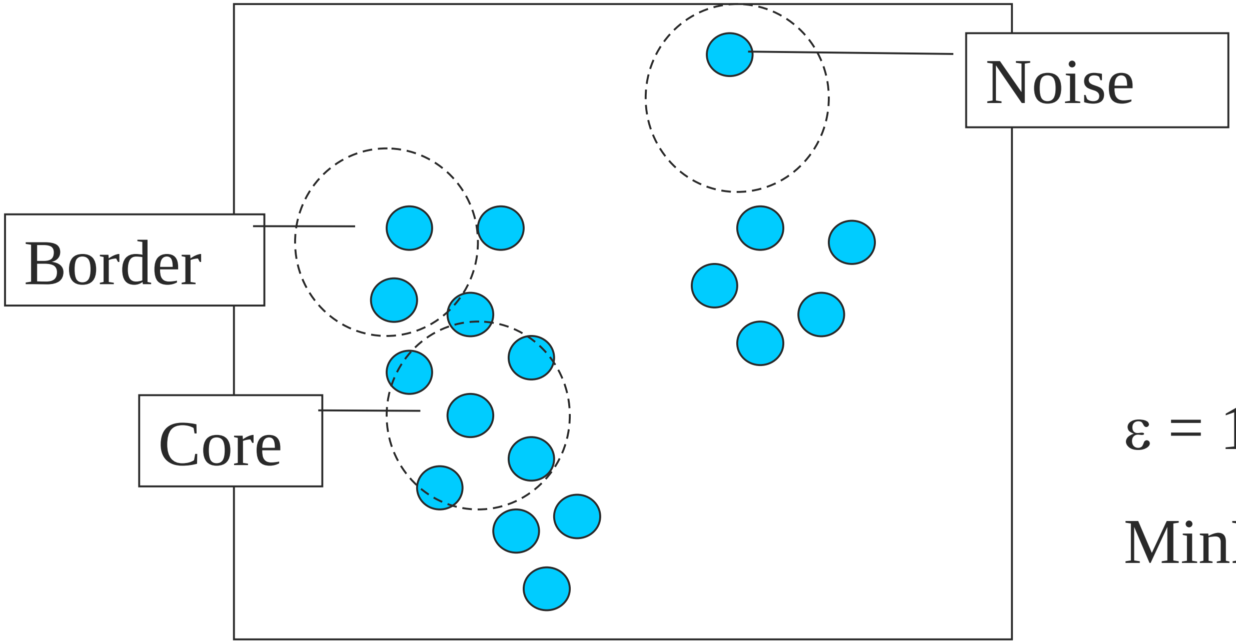
Some definitions

- Core point
- Border point
- Noise point

DBSCAN

- A noise point is any point that is not a core point or a border point.
- Any two core points are close enough– within a distance *Eps* of one another – are put in the same cluster
- Any border point that is close enough to a core point is put in the same cluster as the core point
- Noise points are discarded

Border & Core



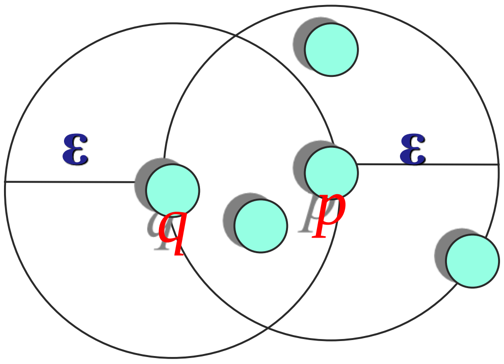
ε = 1unit

MinPts = 5

Concepts: Reachability

Directly density-reachable

- An object q is directly density-reachable from object p if q is within the ε-Neighborhood of p and p is a core object.

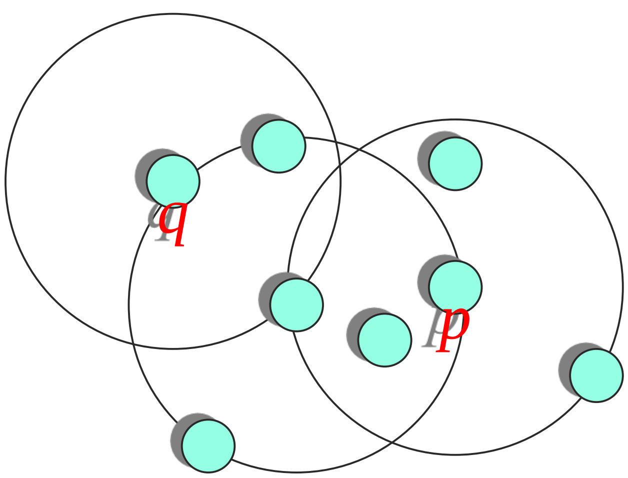


- q is directly density-reachable from p
- p is not directly density-reachable from q?

Concepts: Reachability

Density-reachable:

- An object p is density-reachable from q w.r.t ε and *MinPts* if there is a chain of objects $p_1,\ldots,p_n$, with $p_1=q$, $p_n=p$ such that p_{i+1} is directly density-reachable from p_i w.r.t ε and *MinPts* for all $1 <= i <= n$

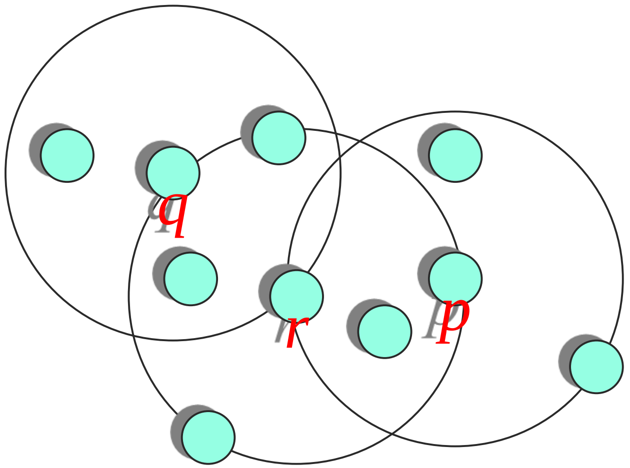


- q is density-reachable from p
- p is not density- reachable from q?
- asymmetric

Concepts: Connectivity

Density-connectivity

Object *p* is density-connected to object *q* w.r.t ε and *MinPts* if there is an object *o* such that both *p* and *q* are density-reachable from *o* w.r.t ε and *MinPts*



- *P* and *q* are density-connected to each other by *r*
- Density-connectivity is symmetric

Concepts: cluster & noise

- **Cluster**: a cluster ***C*** in a set of objects ***D*** w.r.t ε and *MinPts* is a non empty subset of ***D*** satisfying
  - Maximality: For all *p*, *q* if *p* ∈ ***C*** and if *q* is density-reachable from *p* w.r.t ε and *MinPts*, then also *q* ∈ ***C.***
  - Connectivity: for all *p*, *q* ∈ ***C***, *p* is density-connected to *q* w.r.t ε and *MinPts* in ***D.***
  - ***Note:*** cluster contains *core objects* as well as *border objects*
- **Noise:** objects which are not directly density-reachable from at least one core object.

DBSCAN: The Algorithm

- elect a point ***p***
- Retrieve all points density-reachable from ***p*** wrt ε and ***MinPts***.
- If ***p*** is a core point, a cluster is formed.
- If ***p*** is a border point, no points are density-reachable from ***p*** and DBSCAN visits the next point of the database.
- Continue the process until all of the points have been processed.

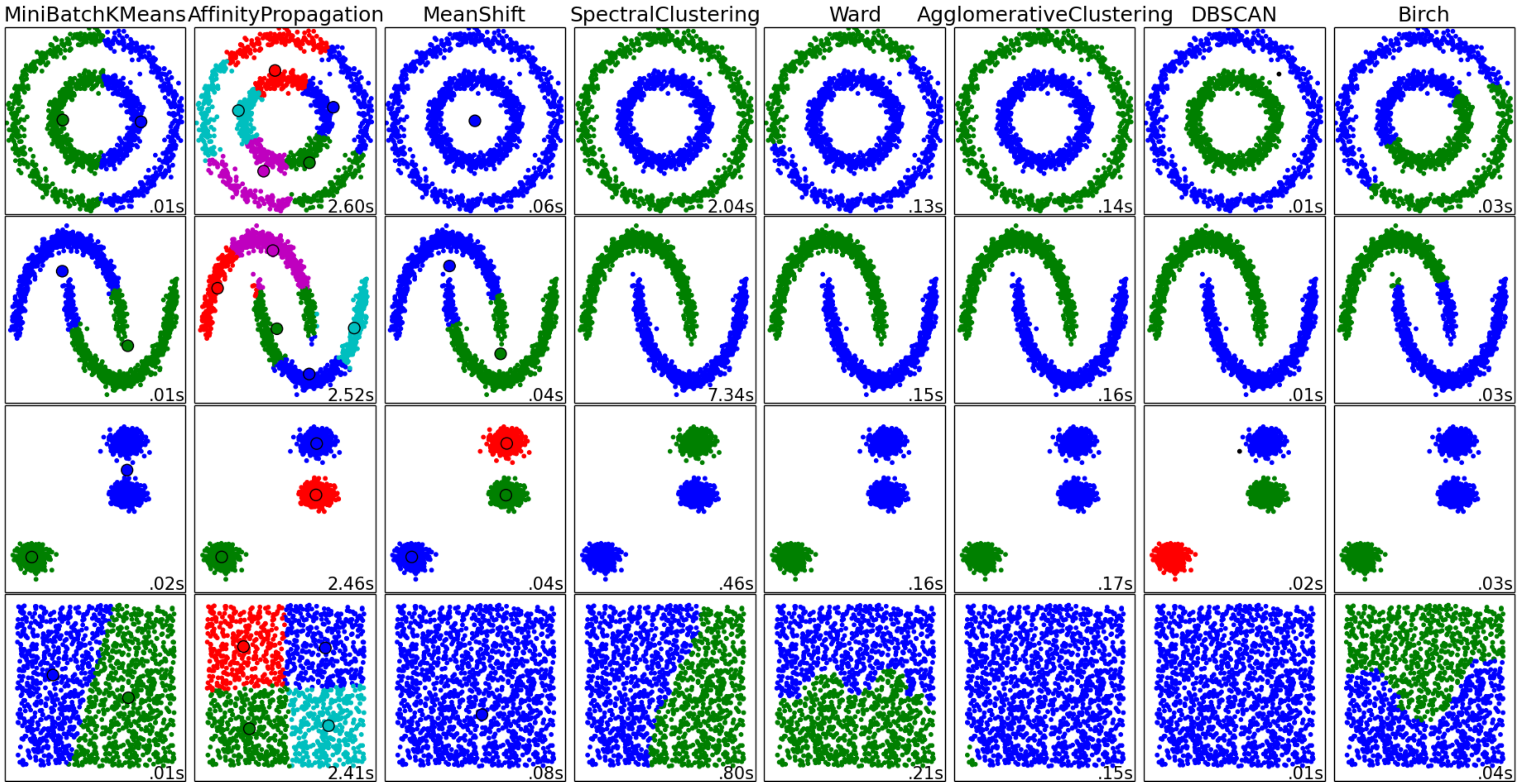
DBSCAN: Determining EPS and MinPts

- Histogram analysis
- Supervised validation based on a training set
- OPTIC algorithm

Evaluation benchmark

- **Homogeneity:** metrics.homogeneity_score(labels_true, labels))
- **Completeness:** metrics.completeness_score(labels_true, labels))
- **V-measure:** metrics.v_measure_score(labels_true, labels))
- **Adjusted Rand Index:** metrics.adjusted_rand_score(labels_true, labels)
- **Adjusted Mutual Information:** metrics.adjusted_mutual_info_score(labels_true, labels)
- **Silhouette Coefficient:** metrics.silhouette_score(X, labels))

Evaluation


MiniBatchKMeans
AffinityPropagation
MeanShift
SpectralClustering
Ward
AgglomerativeClustering
DBSCAN
Birch
.01s
2.60s
.06s
2.04s
.13s
.14s
.01s
.03s
.01s
2.52s
.04s
7.34s
.15s
.16s
.01s
.03s
.02s
2.46s
.04s
.46s
.16s
.17s
.02s
.03s
.01s
2.41s
.08s
.80s
.21s
.15s
.01s
.04s

Demo

Reference

[1] Sander, Jörg, et al. "Density-based clustering in spatial databases: The algorithm gdbscan and its applications." Data mining and knowledge discovery 2.2 (1998): 169-194.

[2] scikit-learn.org

[3]http://home.iitk.ac.in/~arpanm/cs365/ProjectPPT_arpan/DBSCAN.ppt