Density Based Spatial Clustering (DBSCAN)

With Data Analysis

Kristina Pestaria Sinaga

Master in Information System Management
Bina Nusantara University

January 08, 2021

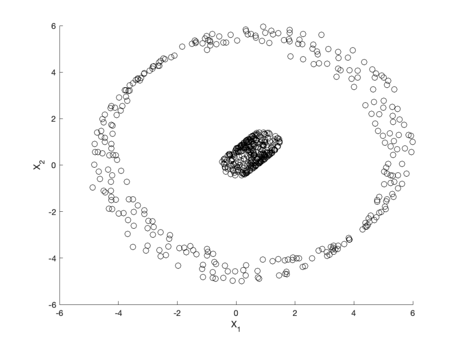
Summary

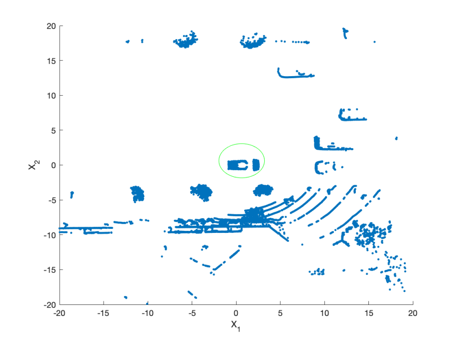
Why DBSCAN?

What is DBSCAN?

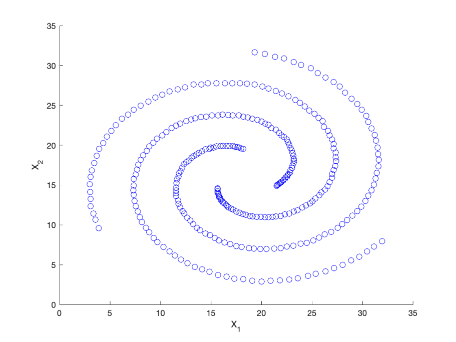
- DBSCAN is a clustering algorithm that defines clusters as continuous regions of high density and works well if all the clusters are dense enough and well separated by low-density regions.
  Source: www.mygreatlearning.com
- DBSCAN requires two parameters:
  1. ϵ (Epsilon): A distance measure that will be used to locate the points or to check the density in the neighbourhood of any point.
  2. n (minPts): the minimum number of points required to form a dense region
- Often used on non-linear or non-spherical datasets

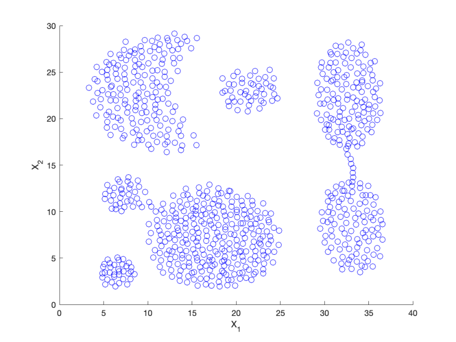
What is non-linear or non-spherical datasets?





What is non-linear or non-spherical datasets?





Advantages: DBSCAN

DBSCAN does not require one to specify the number of clusters in the data a priori, as opposed to k-Means, K-Medoids

$\Downarrow$

DBSCAN can sort data into clusters of varying shapes as well

$\Downarrow$

DBSCAN has a notion of noise, and is robust to outliers.

Advantages: DBSCAN

DBSCAN does not require one to specify the number of clusters in the data a priori, as opposed to k-Means, K-Medoids

$$\Downarrow$$

DBSCAN can sort data into clusters of varying shapes as well

$$\Downarrow$$

DBSCAN has a notion of noise, and is robust to outliers.

Advantages: DBSCAN

DBSCAN does not require one to specify the number of clusters in the data a priori, as opposed to k-Means, K-Medoids

$\Downarrow$

DBSCAN can sort data into clusters of varying shapes as well

$\Downarrow$

DBSCAN has a notion of noise, and is robust to outliers.

Advantages: DBSCAN

$\Downarrow$

DBSCAN requires just two parameters and is mostly insensitive to the ordering of the points in the database. (However, points sitting on the edge of two different clusters might swap cluster membership if the ordering of the points is changed, and the cluster assignment is unique only up to isomorphism.)

$\Downarrow$

The parameters minPts and ϵ can be set by a domain expert, if the data is well understood.

Advantages: DBSCAN

$\Downarrow$

DBSCAN requires just two parameters and is mostly insensitive to the ordering of the points in the database. (However, points sitting on the edge of two different clusters might swap cluster membership if the ordering of the points is changed, and the cluster assignment is unique only up to isomorphism.)

$\Downarrow$

The parameters minPts and ϵ can be set by a domain expert, if the data is well understood.

Disadvantages: DBSCAN

1. DBSCAN struggles with clusters of similar density
2. Struggles with high dimensionality data

How Does DBSCAN Algorithm Works?

DBSCAN algorithm

Pick an arbitrary data point p as your first point.

⇓

Mark p as visited.

⇓

Extract all points present in its neighborhood (upto eps distance from the point), and call it a set nb

DBSCAN algorithm

Pick an arbitrary data point p as your first point.

$\Downarrow$

Mark p as visited.

$\Downarrow$

Extract all points present in its neighborhood (upto eps distance from the point), and call it a set nb

DBSCAN algorithm

Pick an arbitrary data point p as your first point.

$\Downarrow$

Mark p as visited.

$\Downarrow$

Extract all points present in its neighborhood (upto eps distance from the point), and call it a set nb

DBSCAN algorithm

If nb $\geq$ minPts, then Consider p as the first point of a new cluster, Consider all points within eps distance (members of nb) as other points in this cluster

$\Downarrow$

else label p as noise

$\Downarrow$

Repeat steps 1-5 till the entire dataset has been labeled ie the clustering is complete

DBSCAN algorithm

If nb $\geq$ minPts, then Consider p as the first point of a new cluster, Consider all points within eps distance (members of nb) as other points in this cluster

$\Downarrow$

else label p as noise

$\Downarrow$

Repeat steps 1-5 till the entire dataset has been labeled ie the clustering is complete

DBSCAN algorithm

If nb $\geq$ minPts, then Consider p as the first point of a new cluster, Consider all points within eps distance (members of nb) as other points in this cluster

$\Downarrow$

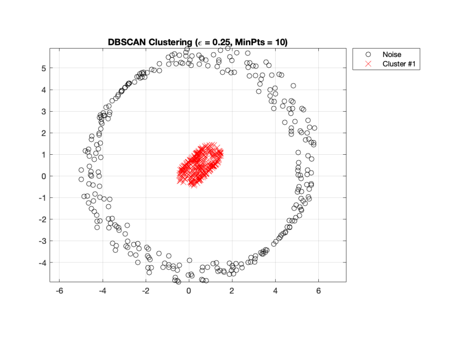
else label p as noise

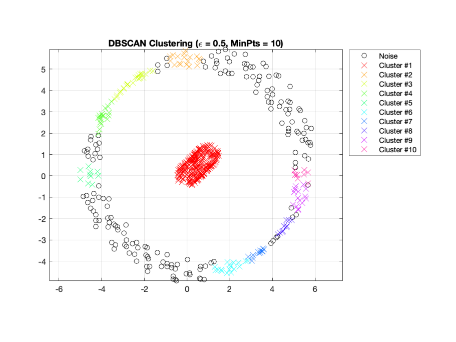
$\Downarrow$

Repeat steps 1-5 till the entire dataset has been labeled ie the clustering is complete

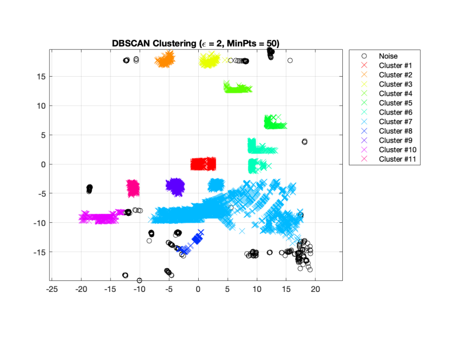
Case Analysis
Non-Spherical Artificial Data Segmentation

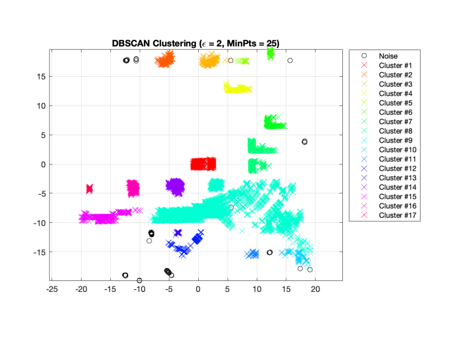
A 2-D circular Dataset



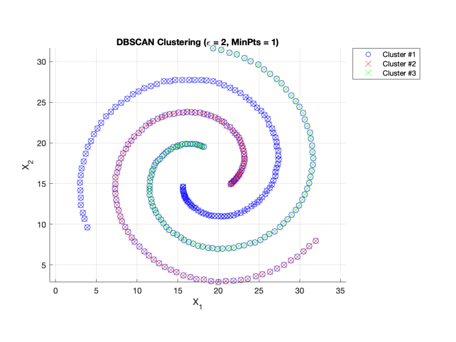


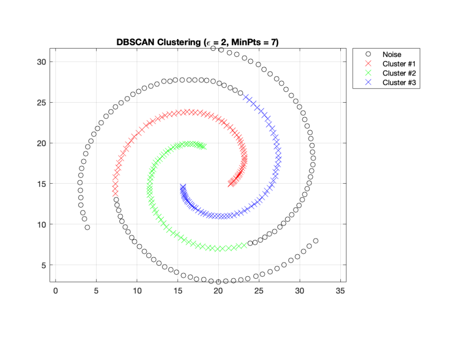
A Lidar Scan Dataset



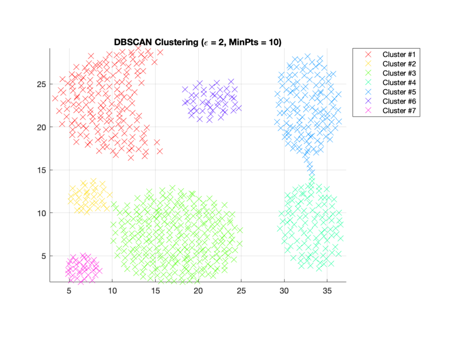


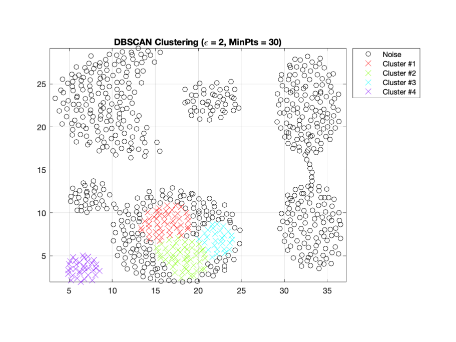
A Spiral Artificial Dataset





An Aggregation Artificial Dataset





References

- Asuncion, A. & Newman, D. *UCI machine learning repository.* 2007.
- Cheng, C.-H. & Chen, Y.-S. Classifying the segmentation of customer value via RFM model and RS theory. *Expert systems with applications* **36**, 4176–4184 (2009).
- Park, H.-S. & Jun, C.-H. A simple and fast algorithm for K-medoids clustering. *Expert systems with applications* **36**, 3336–3341 (2009).
- You, Z. *et al.* A decision-making framework for precision marketing. *Expert Systems with Applications* **42**, 3357–3367 (2015).

The End