Clustering with DBSCAN

Charlie Rehder, Luis Rivera, Parth Patel, Marjea Mckoy

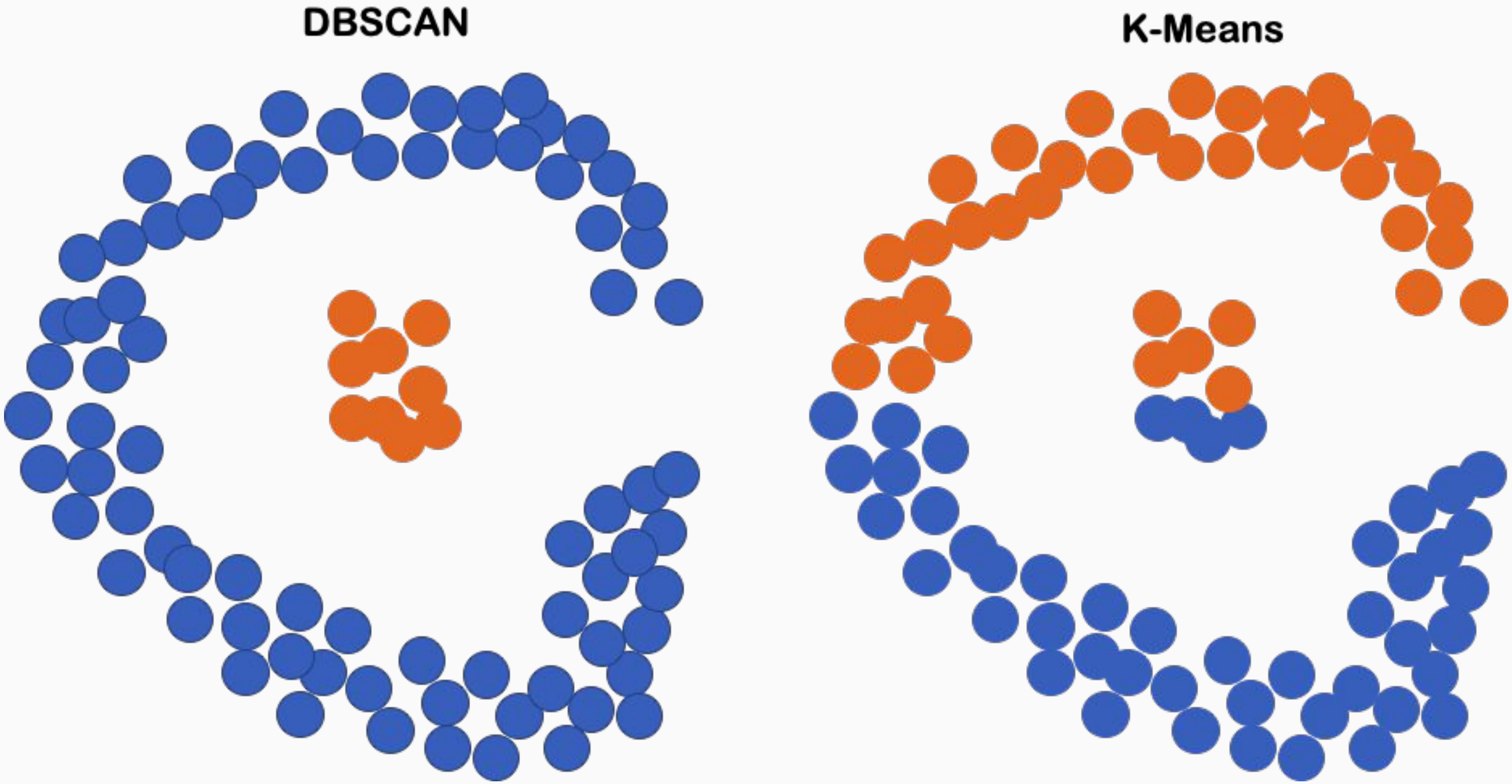
Overview

- **D**ensity-**B**ased **S**patial **C**lustering of **A**pplications with **N**oise
- Takes two parameters Epsilon and Minimum Points
- Finds all the Core Points, Border Points, and Outliers
- Creates clusters based on the Core and Border Points and Epsilon

Advantages and Disadvantages

Advantages	Disadvantages
- Does a great job separating clusters of high and low density	- It struggles with clusters of similar density.
- Identifies outliers and noise while clustering	- Suffers with high dimensionality
- Does not require a specification on the number of clusters	- Very sensitive with EPS and minimum points that you set (this heavily influences clustering)
- Can work with nested wrapping and arbitrarily shaped clusters	- It fails in identifying clusters if the density varies and if the dataset is too sparse

DBSCAN vs K-means



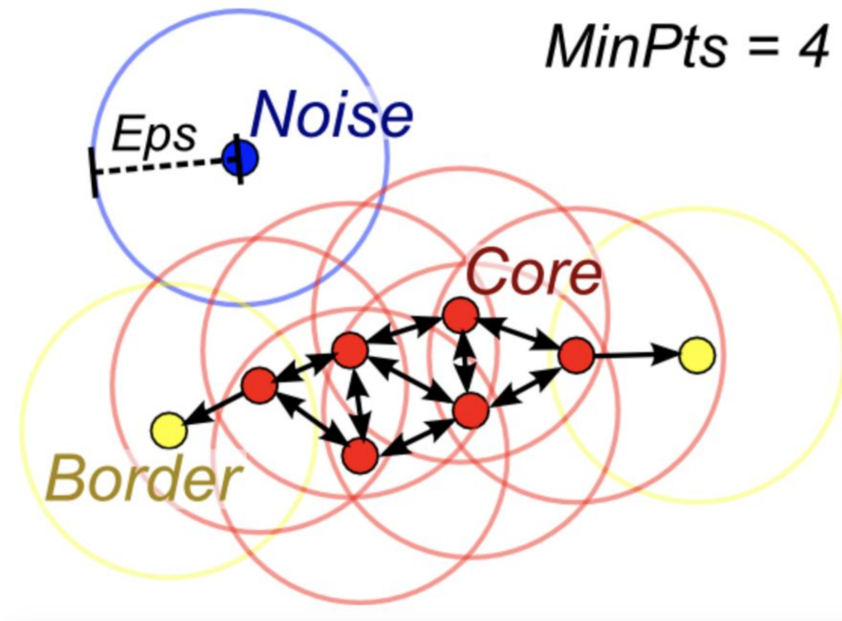
DBSCAN

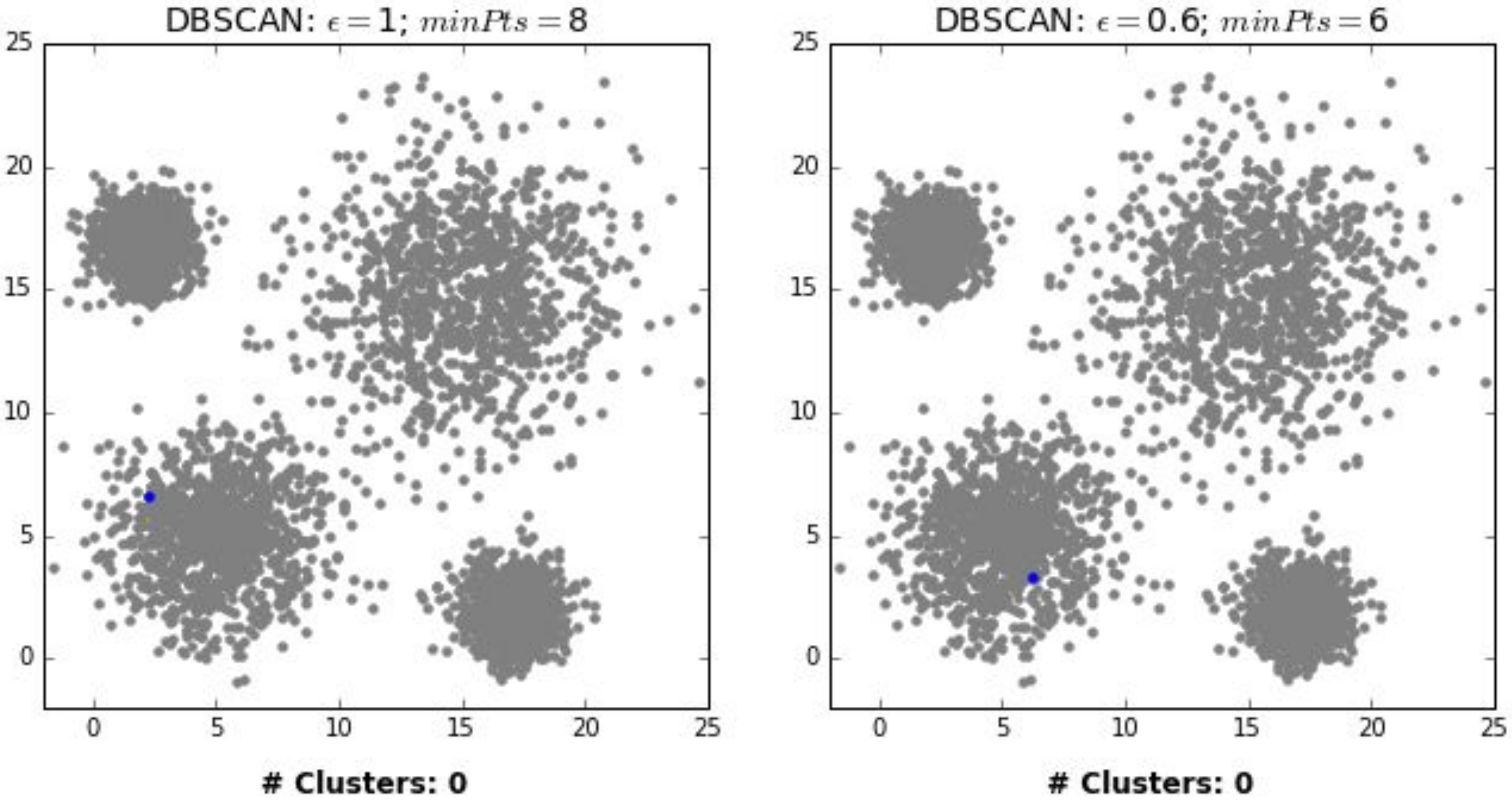
K-Means

Hyperparameters

- **Epsilon (ε)** - radius for all points
- **Minimum Points** - minimum number of data points to define a cluster

DBSCAN is highly sensitive to the values of these parameters.





DBSCAN: $\epsilon = 1$; $minPts = 8$

DBSCAN: $\epsilon = 0.6$; $minPts = 6$

Data Processing Steps

- Generally, when clustering, standardization is helpful, but there are times when it is not.
- Missing values must be removed or imputed; DBSCAN cannot handle missing values.
- DBSCAN is robust against outliers and noise; not necessary to handle these in preprocessing.
- All data must be numeric; however, avoid dummy variables as adding extra, sparse features can lead to poor performance.
- Suffers from the Curse of Dimensionality.

Algorithm Steps

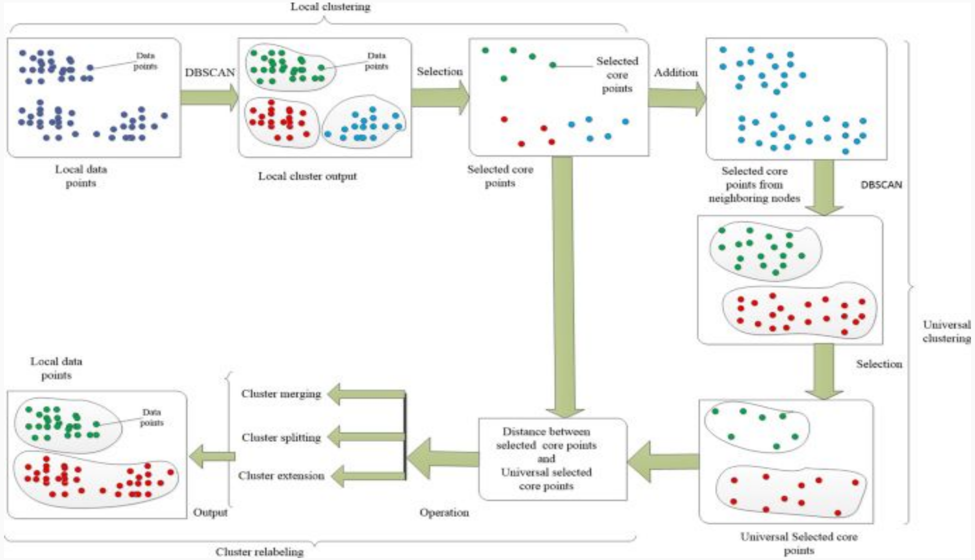
1. Classify the points.

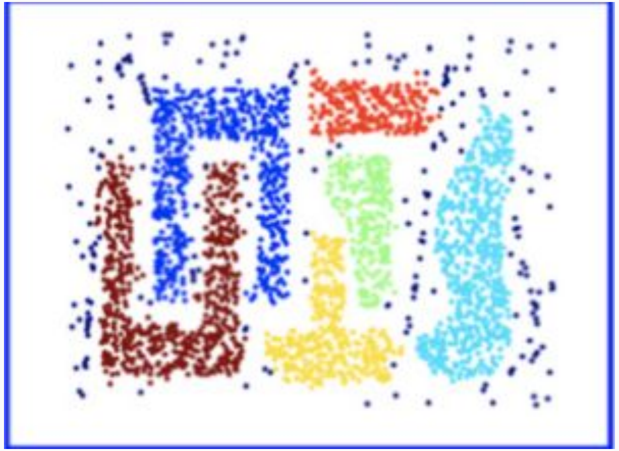
2. Discard noise.

3. Assign cluster to a core point.

4. Color all the density connected points of a core point.

5. Color boundary points according to the nearest core point.





Appendix

Overview

- https://towardsdatascience.com/dbscan-clustering-explained-97556a2ad556
- https://shritam.medium.com/how-dbscan-algorithm-works-2b5bef80fb3
- https://elutins.medium.com/dbscan-what-is-it-when-to-use-it-how-to-use-it-8bd506293818
- https://towardsdatascience.com/a-practical-guide-to-dbscan-method-d4ec5ab2bc99

Documentation

- https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html

Videos

- https://www.youtube.com/watch?v=RDZUdRSDOok

Articles with Code Snippets

- https://www.analyticsvidhya.com/blog/2020/09/how-dbscan-clustering-works/
- https://elutins.medium.com/dbscan-what-is-it-when-to-use-it-how-to-use-it-8bd506293818#:~:text=DBSCAN%20
- https://www.youtube.com/watch?v=RDZUdRSDOok&ab_channel=StatQuestwithJoshStarmer
- https://www.tutorialspoint.com/what-is-the-difference-between-k-means-and-dbscan
- https://machinelearningknowledge.ai/tutorial-for-dbscan-clustering-in-python-sklearn/
- https://datascience-enthusiast.com/Python/DBSCAN_Kmeans.html

Visualizations

- https://www.naftaliharris.com/blog/visualizing-dbscan-clustering/

Example Code(KMean, Hierarchical, DBSCAN)

- https://github.com/charlierehder/ml-assessment-group-6/blob/master/Assessment-%20DBSCAN.ipynb

HOLLYWOOD
Beverly Hills
CITY LIMIT
ELEV 268
THIS IS
THE END
STARZ