Clustering Analysis

Prayag Ranjan Sahu

NISER, Bhubaneswar

April 15, 2024

Overview

Clustering

- Clustering is an interesting problem of **unsupervised learning** $\rightarrow$ cluster analysis does not use category labels that tag objects with prior identifiers.
- Deals with **data structure partitioning** in space.
- Forms the basis of **exploratory data analysis (EDA).**
- The idea of clusters is intuitively accessible.

A cluster is comprised of a number of *similar* objects.

- It is interesting to see how one might go about formally defining clusters.

Defining Clusters

- A cluster is a set of entities which are **alike**, and entities from different clusters are not.
-

Everitt, B. S. (1974). *Cluster Analysis*. John Wiley & Sons, Inc., New York.

Defining Clusters

- A cluster is a set of entities which are **alike**, and entities from different clusters are not.
- A cluster is an aggregation of points such that the **distance** between any two points in the cluster is lesser than between any point in the cluster and any point not in it.
-

Everitt, B. S. (1974). *Cluster Analysis*. John Wiley & Sons, Inc., New York.

Defining Clusters

- A cluster is a set of entities which are **alike**, and entities from different clusters are not.
- A cluster is an aggregation of points such that the **distance** between any two points in the cluster is lesser than between any point in the cluster and any point not in it.
- Clusters may be described as connected regions of a multidimensional space containing a relatively **high density** of points separated from other such regions by a region containing a relatively low density of points.

Everitt, B. S. (1974). *Cluster Analysis*. John Wiley & Sons, Inc., New York.

Defining Clusters

- A cluster is a set of entities which are **alike**, and entities from different clusters are not.
- A cluster is an aggregation of points such that the **distance** between any two points in the cluster is lesser than between any point in the cluster and any point not in it.
- Clusters may be described as connected regions of a multidimensional space containing a relatively **high density** of points separated from other such regions by a region containing a relatively low density of points.

The last two definitions assume that objects to be clustered are represented as points in measurement space, and that this is the premise from now on.

Everitt, B. S. (1974). *Cluster Analysis*. John Wiley & Sons, Inc., New York.

Clustering Techniques

- Centroid-Based Techniques
- Density-Based Techniques

Distance Functions

$$\mathbf{X} = (x_1, x_2, x_3, \dots , x_n), \mathbf{Y} = (y_1, y_2, y_3, \dots , y_n) \in \mathbb{R}^n$$

City Block Distance

$$\mathcal{D}(\mathbf{X},\mathbf{Y}) = \sum_i^n |x_i - y_i|$$

Euclidean Distance

$$\mathcal{D}(\mathbf{X},\mathbf{Y}) = (\sum_i^n (x_i - y_i)^2)^{\frac{1}{2}}$$

Chebyshev Distance

$$\mathcal{D}(\mathbf{X},\mathbf{Y}) = \mathscr{M}(|x_i - y_i|)$$

Minkowski Distance

$$\mathcal{D}(\mathbf{X},\mathbf{Y}) = (\sum_i^n (x_i - y_i)^p)^{\frac{1}{p}}$$

Minkowski Distance

- $p = 1$ (City Block Distance)
- $p = 2$ (Euclidean Distance)
- $p \to \infty$ (Chebyshev Distance)

Clustering Algorithms

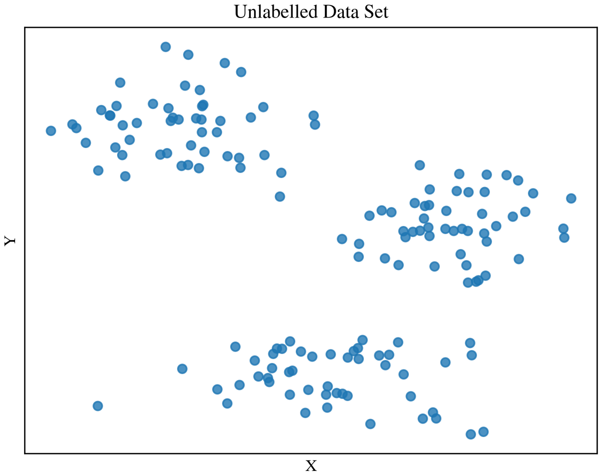
- K-Means Algorithm
- DBSCAN
- HDBSCAN

K-Means Algorithm[1]

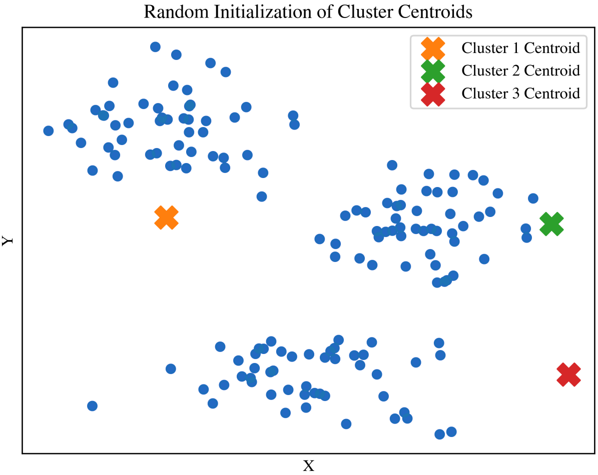
1. Randomly initialize K centroids.
2. Calculate distance of each point ($\mathbf{X}_i$) from each of the K centroids.
3. Assign each point ($\mathbf{X}_i$) to the centroid located at minimum distance.
4. Update the centroids by computing the mean of points assigned to each cluster.
5. Go to 2.

[1] S. Lloyd. "Least squares quantization in PCM". In: *IEEE Transactions on Information Theory* 28.2 (1982), pp. 129–137. DOI: 10.1109/TIT.1982.1056489.

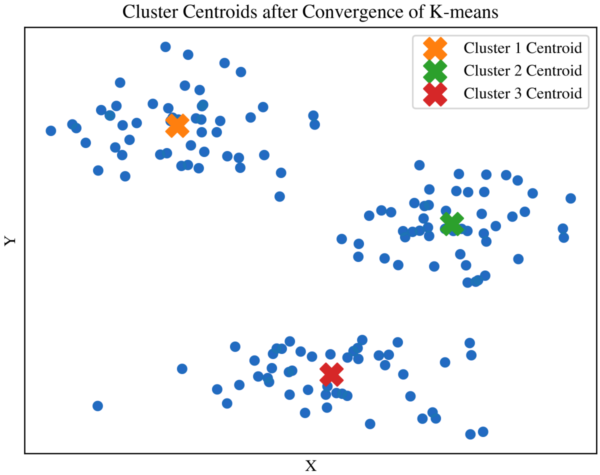
Visualizing the *K*-Means Algorithm



Visualizing the K-Means Algorithm

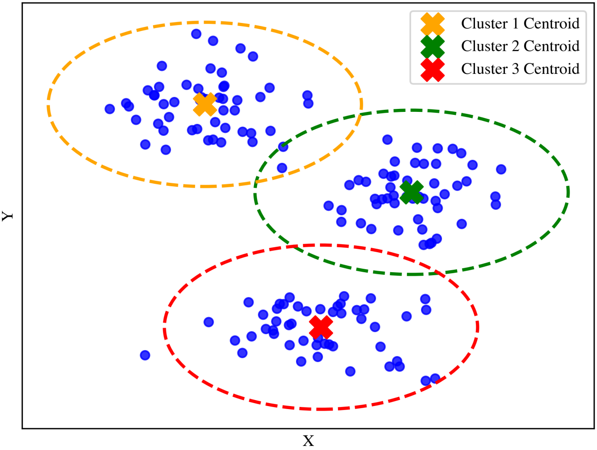


Visualizing the K-Means Algorithm



Visualizing the K-Means Algorithm

Final Clusters



The Fall of K-Means

1. What is K?
2. K-Means is sensitive to initial conditions.
3. K-Means can't handle "nested" clusters.

The Fall of K-Means

K-Means can't handle "nested" clusters.

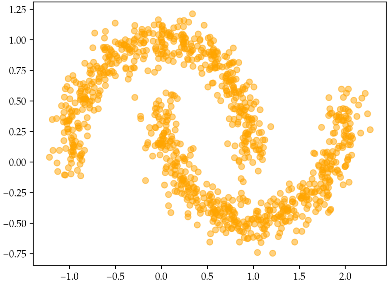


Figure: *Two Moons* Data Set

The Fall of *K*-Means

K-Means can't handle "nested" clusters.

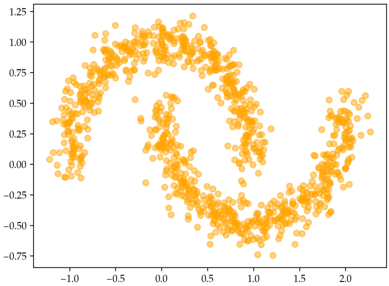


Figure: *Two Moons* Data Set

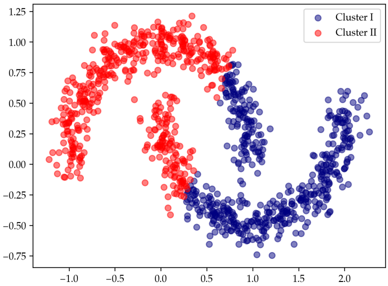


Figure: *K*-Means on *Two Moons*

The Fall of K-Means

K-Means can't handle "nested" clusters.

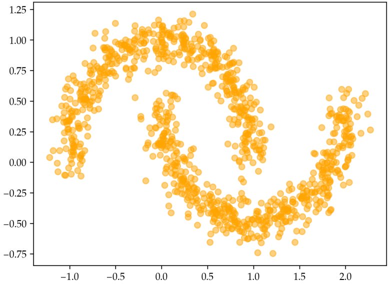


Figure: *Two Moons* Data Set

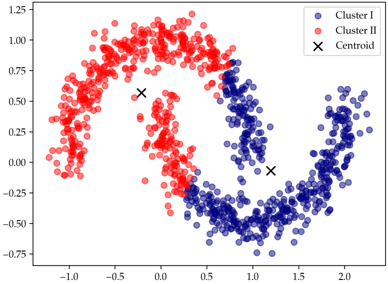


Figure: K-Means on *Two Moons*

DBSCAN[1] (Density-Based Spatial Clustering of Applications with Noise)

Identifying Clusters by Visual Inspection

Clusters are defined by high density regions. **Outliers** are defined by low density regions.

1. For each point in the data, check if there are at least η points around it at ϵ distance from it. Every point that satisfies this criterion is said to be a **Core**. Others are **Non-Cores**.
2. Start with a random core point. Add itself and all the cores around it that are at least ϵ distance from it to one cluster.
3. Let the clusters grow until there are only cores in each cluster. After that, add all non-cores that are at least ϵ distance from any of the cores to the respective clusters. These are **Boundary Points**.
4. The remaining points are labelled as outliers.

[1] Martin Ester et al. "A density-based algorithm for discovering clusters in large spatial databases with noise". In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. 1996, pp. 226–231.

DBSCAN in Action

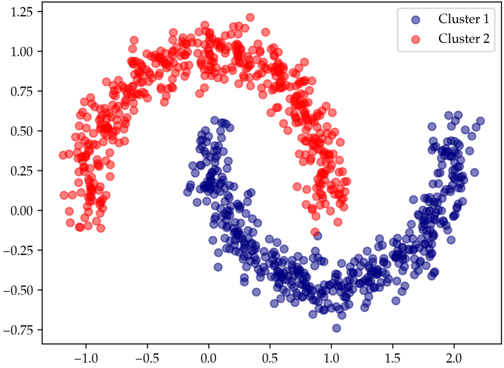


Figure: DBSCAN on *Two Moons* ($\eta = 4$, $\epsilon = 0.1$)

The Fall of DBSCAN

1. What are η & ϵ?
2. Does not do well with real-world data that is affected by noise.

The Fall of DBSCAN

Does not do well with real-world data that is affected by noise.

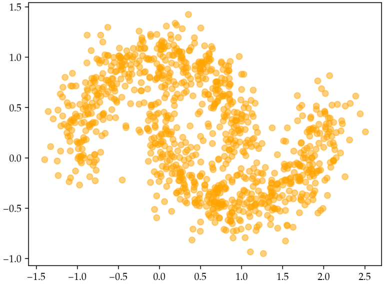


Figure: *Two Moons* Data Set (noise = 0.18)

The Fall of DBSCAN

Does not do well with real-world data that is affected by noise.

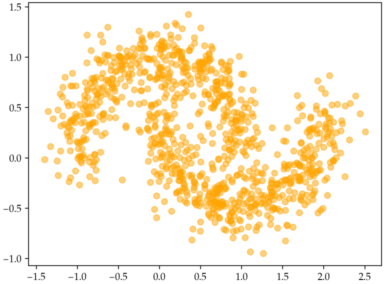


Figure: *Two Moons* Data Set (noise = 0.18)

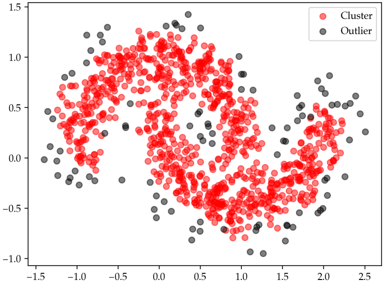


Figure: DBSCAN on noisy *Two Moons*

HDBSCAN[2]

(Hierarchical Density-Based Spatial Clustering of Applications with Noise)

1. Lower the "sea" level using: $\mathrm{d}_k(\mathrm{a}, \mathrm{b}) = \max\{\mathrm{core}_k(\mathrm{a}), \mathrm{core}_k(\mathrm{b}), \mathrm{d}_{a,b}\}$[1]. k denotes the k^{th} nearest neighbour.
2. Consider the data as a weighted graph with the data points as vertices and an edge between any two points with weight equal to the d_k of those points.
3. Make a dendrogram, starting with each point as a single cluster and ending with one large cluster of all points.
4. Prune the dendrogram whenever there there are less than m number of points in a cluster.

[1] Justin Eldridge, Mikhail Belkin, and Yusu Wang. "Beyond Hartigan Consistency: Merge Distortion Metric for Hierarchical Clustering". In: *Proceedings of The 28th Conference on Learning Theory*. Vol. 40. Proceedings of Machine Learning Research. Paris, France, Mar. 2015, pp. 588–606.

[2] Leland McInnes, John Healy, and S. Astels. "hdbscan: Hierarchical density based clustering". In: *J. Open Source Softw.* 2 (2017), p. 205.

HDBSCAN in Action

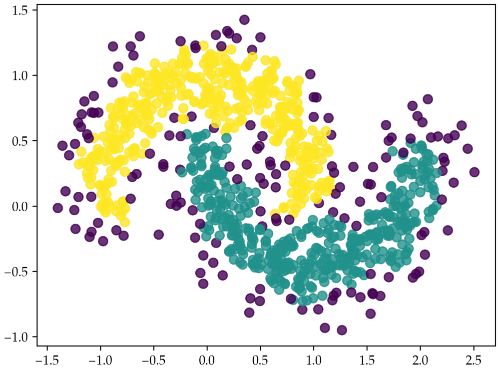


Figure: HDBSCAN on *Two Moons* ($\eta = 4$, $m = 400$)

Summary

- Clustering is an important problem concerning unsupervised learning algorithms.
- Distance metrics are important for discerning similarity or dissimilarity.
- K-Means is a centroid-based algorithm. It requires a judicious choice of K. Further, it makes assumptions about the nature of the shape of clusters $\rightarrow$ the Gaussian "ball" assumption.
- DBSCAN is a density-based algorithm. It performs poorly on data sets containing clusters of varying densities.
- HDBSCAN is a hierarchical density-based algorithm. It improves upon DBSCAN.

References

[1] S. Lloyd. "Least squares quantization in PCM". In: *IEEE Transactions on Information Theory* 28.2 (1982), pp. 129–137. DOI: 10.1109/TIT.1982.1056489.

[2] Martin Ester et al. "A density-based algorithm for discovering clusters in large spatial databases with noise". In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. 1996, pp. 226–231.

[3] Justin Eldridge, Mikhail Belkin, and Yusu Wang. "Beyond Hartigan Consistency: Merge Distortion Metric for Hierarchical Clustering". In: *Proceedings of The 28th Conference on Learning Theory*. Vol. 40. Proceedings of Machine Learning Research. Paris, France, Mar. 2015, pp. 588–606.

[4] Leland McInnes, John Healy, and S. Astels. "hdbscan: Hierarchical density based clustering". In: *J. Open Source Softw.* 2 (2017), p. 205.

Thank you.

Thank you.

(You are welcome.)