Hierarchical Density-Based Clustering

Prayag Ranjan Sahu
2011117
CS660/CS460: Machine Learning, Spring 2024
National Institute of Science Education and Research
An OCC of Homi Bhabha National Institute
Bhubaneswar, Odisha, India

Clustering Analysis

- Clustering is an interesting problem of unsupervised learning → 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.)

Figure 1. Dataset in 2D space

Classification of Clustering Algorithms

- Flat clustering creates a set of clusters that hold no inherent relationship to one another.
- Hierarchical clustering creates a family of sets of clusters.
- Centroid-based/Parametric clustering initializes centroids around which clusters form.
- Density-based/Non-Parametric clustering prepares clusters by quantifying density.

Clustering Type	Flat	Hierarchical
Centroid	K-means	Ward Complete-Linkage
Density	DBSCAN	HDBSCAN

- K-means:
  - suffers from the choice of parameter K
  - makes an assumption about the data distribution: the Gaussian-ball assumption
- DBSCAN:
  - gets rid of the Gaussian-ball assumption
  - the resolution parameter is arbitrary though
- Ward Complete-Linkage
  - Gaussian-ball assumption creeps in; the hierarchical tree needs to be cut somewhere

Hierarchical Density-Based Spatial Clustering of Applications with Noise

The protocol for HDBSCAN is as follows:

- Transformation of the dataset to mutual reachability space
- Constructing of a minimum spanning tree (MST)
- Preparation of a dendrogram for the MST
- Pruning of the dendrogram based on minimum cluster size
- Extraction of clusters

Transformation of Space

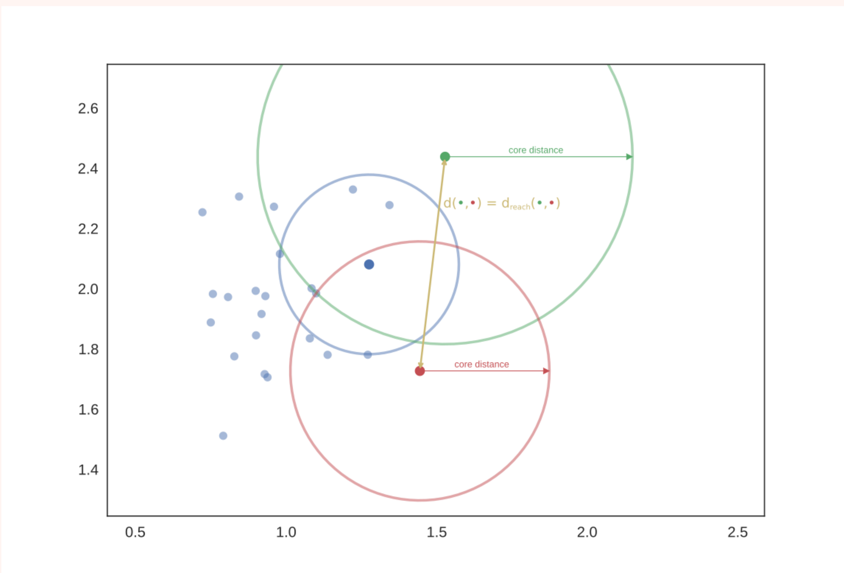


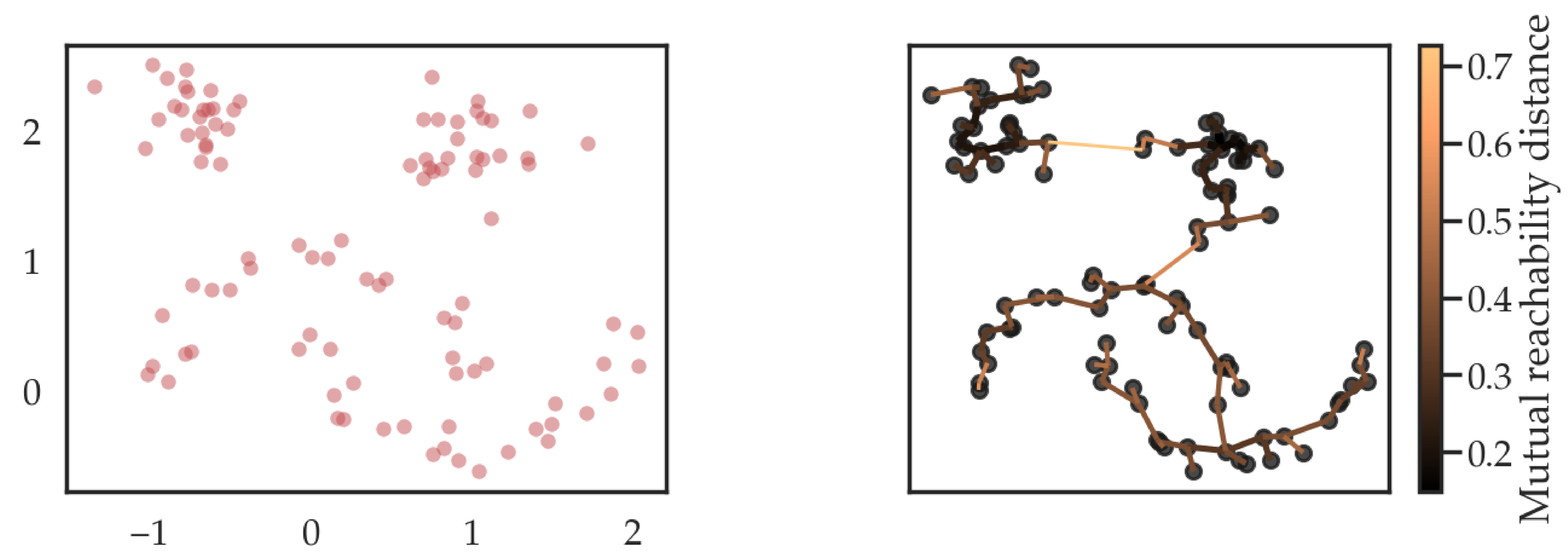
Figure 2. Visualization of the Distance Transformation

$$d_{mreach-k}(a,b) = max\{core_k(a), core_k(b), d(a,b)\}$$

- The entire dataset transformed to mutual reachability space by defining the distance between any two points as $d_{mreach-k}(a,b)$.
- This transformation has the effect of tightening clusters, rendering the algorithm more robust to noise.
- This transformation also has the effect of closely approximating the the hierarchy of level sets of whatever true density distribution the points were sampled from [1].

Preparation of the Minimum Spanning Tree

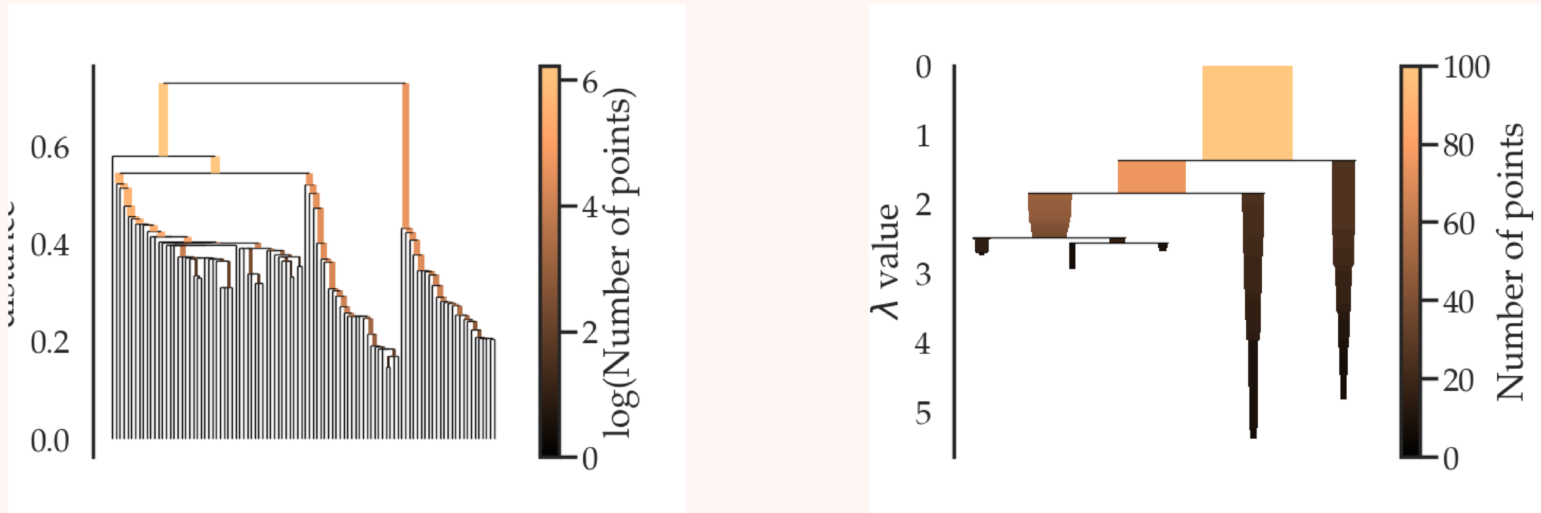
- A minimum spanning tree is a subset of the edges of a connected, edge-weighted undirected graph that connects all the vertices without cycles and ensures minimum possible total edge weight.
- Standard algorithms to do so include Prim's [2] and Kruskal's [3] algorithms.



(a) Dataset
(b) Minimum Spanning Tree

Figure 3. Conversion of the Dataset to the Minimum Spanning Tree

Condensing the Cluster Tree



(a) Dendrogram before Pruning
(b) Dendrogram after Pruning

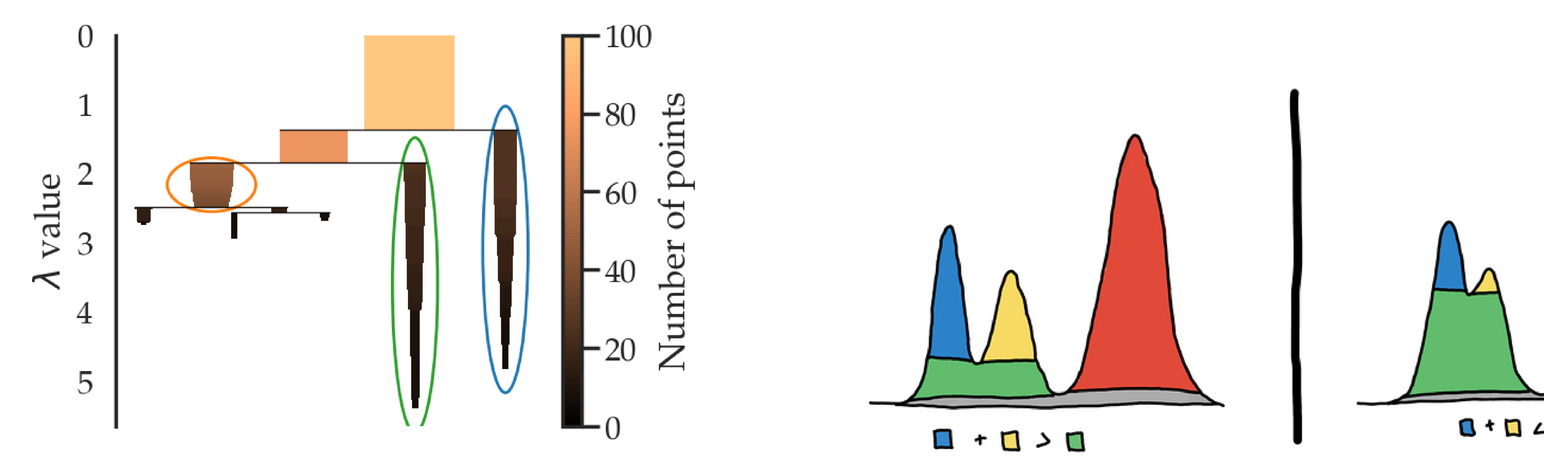
Figure 4. Pruning the Dendrogram based on Minimum Cluster Size

This step condenses down the large and complicated cluster hierarchy into a smaller tree.

- To do this, the algorithm takes in a parameter: the minimum number of points that constitute a cluster (min_cls_size.)
- Starting from the root, it is checked if one of the new clusters created by a split has fewer points than min_cls_size:
  - If yes,the larger cluster retains the cluster identity
  - If no, it is a true cluster split

Extraction of Clusters

We want to choose clusters that have a long lifetime.



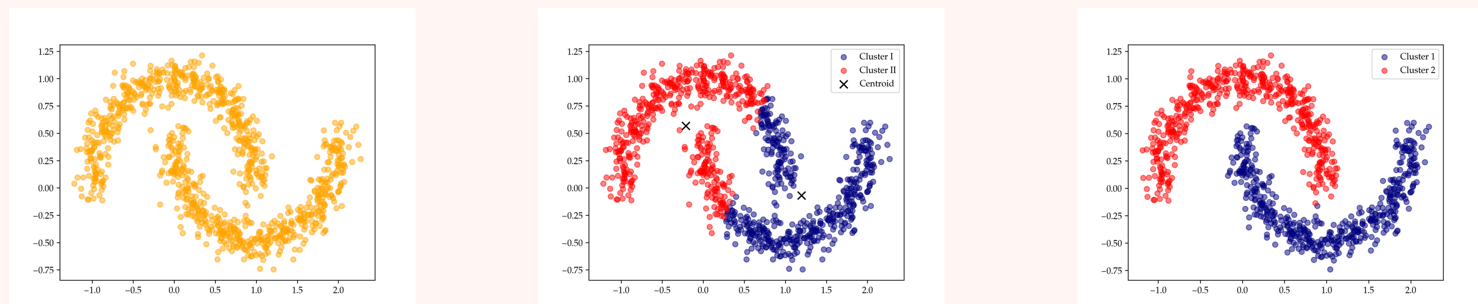
(a) Extracted clusters
(b) Three clusters v/s Two clusters

Figure 5. Extracting stable clusters

The stability of a cluster is defined as: $S = \sum_{p \in cluster} (\lambda_p - \lambda_{birth})$

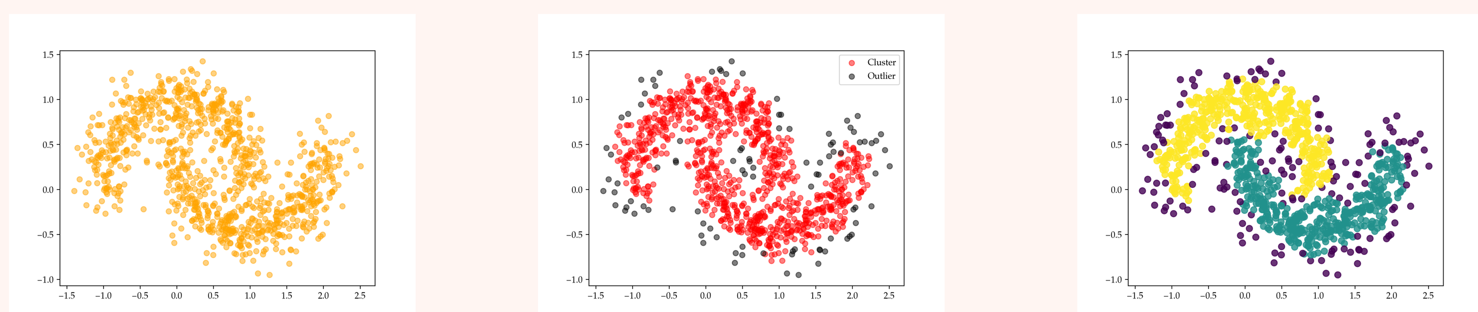
- λ denotes $\frac{1}{distance}$
- λ_p denotes the λ value when the point fell out of the cluster.
- λ_{birth} denotes the λ value when the cluster split off and became independent.
- Starting from the leaf, it is checked if $S^i_{left} + S^i_{right} > S^{i-1}$:
  - If yes, the children are true clusters
  - Otherwise, the parent cluster is true

HDBSCAN in Action: An Application to Noisy, Nested Data



(a) *Two Moons*
(b) K-means
(c) DBSCAN (η = 4, ϵ = 0.1)

Figure 6. Performance of K-means & DBSCAN on *Two Moons*



(a) Noisy *Two Moons*
(b) DBSCAN (η = 4, ϵ = 0.1)
(c) HDBSCAN (min_cls_size = 400)

Figure 7. Performance of DBSCAN & HDBSCAN on Noisy *Two Moons*

Closing Remarks

- K-means fails to cluster nested datasets due to the Gaussian-ball assumption.
- DBSCAN handles nested datasets well. However, it is not robust to noise.
- HDBSCAN can handle noisy, nested data. It also performs well for clusters of varying densities.

References

[1] J. Eldridge, M. Belkin, and Y. Wang, "Beyond hartigan consistency: Merge distortion metric for hierarchical clustering," in *Proceedings of The 28th Conference on Learning Theory*, vol. 40 of *Proceedings of Machine Learning Research*, pp. 588–606, PMLR, 2015.

[2] R. C. Prim, "Shortest Connection Networks And Some Generalizations," *Bell System Technical Journal*, vol. 36, pp. 1389–1401, Nov. 1957.

[3] J. B. Kruskal, "On the shortest spanning subtree of a graph and the traveling salesman problem," *Proceedings of the American Mathematical society*, vol. 7, no. 1, pp. 48–50, 1956.