Skeleton Clustering : A Dimension free Density-Aided Clustering

Kaustav Paul
Sourav Biswas

Indian Statistical Institute, Kolkata

2024-10-29

Traditional Clustering Methods

- **k-means clustering:**
  - Unable to detect non-convex clusters.
  - The center of a non-convex cluster falls outside the cluster itself and may come close to observations from a different cluster.
  - In high dimension k-means algorithm may assign all the points to a single cluster.
- **Density Based Clustering:**
  - To estimate the underlying PDF and detect clusters based on the PDF.
  - The rate of convergence for the density estimates is $\mathcal{O}_{\mathbb{P}}(n^{-\frac{1}{d+4}})$
- **Hierarchical Clustering:**
  - Problem with non-convex clusters persists.
  - If any pair of the points in two different clusters lie very close to each other, the two clusters may get merged in this method.

Skeleton Clustering Framework.

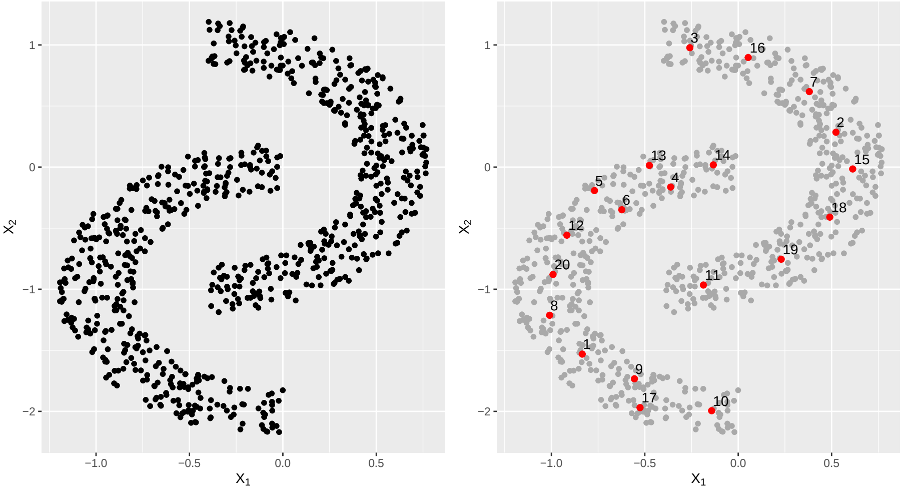
Input : Observations $X_1, X_2, \dots, X_N$, final number of clusters S.

1. **Knot construction** : Perform $k-$means clustering with a large number k; the centers are the knots.
2. **Edge construction** : Apply approximate Delaunay triangulation to the knots. Generally we choose $k = \lfloor \sqrt{n} \rfloor$
3. **Edge weights construction** : Add weights to each edge using either Voronoi density, Face density or Tube density similarity measure.
4. **Knots segmentation** : Use linkage criterion to segment knots into S groups based on the edge weights.
5. **Assignment of labels** : Assign a cluster label to each observation based on which knot group the nearest knot belongs to.

Knot construction

- Some knots are constructed to give a concise representation of the data structure.
- In practice we use k-Means to choose $k = \lfloor \sqrt{n} \rfloor$ knots, where n is the number of samples.
- Empirically robustness performance with sufficient number of knots.

Knot Construction

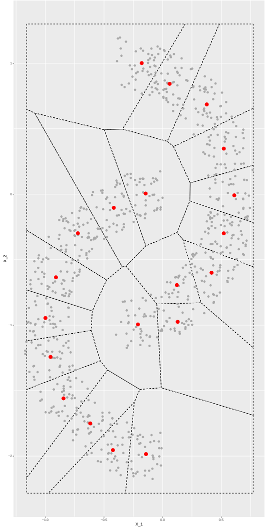


Edge construction

Let $c_1, c_2, \ldots, c_k$ be the given knots and we use $\mathcal{C} = \{c_1, c_2, \ldots, c_k\}$ to denote their collection of them.

- The Voronoi cell, or Voronoi region, $\mathbb{C}_j$ associated with a knot c_j is the set of all points in $\mathcal{X}$ whose distance to c_j is the smallest compared to other knots. That is, $\mathbb{C}_j = \{\boldsymbol{x} \in \mathcal{X} : d\left(\boldsymbol{x}, c_j\right) \leq d\left(\boldsymbol{x}, c_\ell\right) \forall \ell \neq j\}$ where $d\left(\boldsymbol{x}, \boldsymbol{y}\right)$ is the usual Euclidean distance.

Edge Construction



Edge Construction

- We add an edge between a pair of knots if they are neighbors, with the neighboring condition being that the corresponding Voronoi cells share a common boundary.
- Such resulting graph is the Delaunay Triangulation of the set of knots $\mathcal{C}$ and we denote is as $DT\left(\mathcal{C}\right)$.
- But in case of high dimensional data, it becomes computationally expensive. Therefore, in practice we approximate the exact Delaunay Triangulation with $\widehat{DT}\left(\mathcal{C}\right)$ by examining the 2-nearest knots of the sample data points.