HIERARCHICAL CLUSTERING

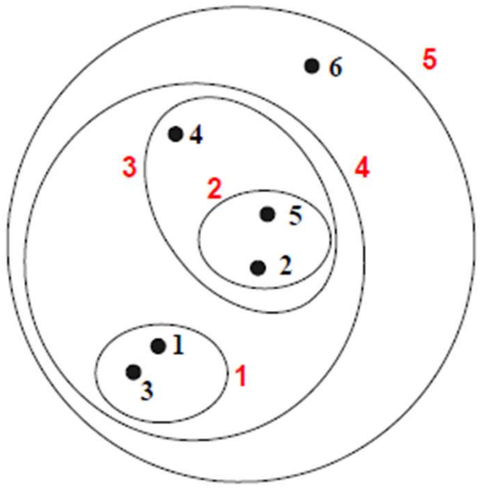
Sourav Karmakar

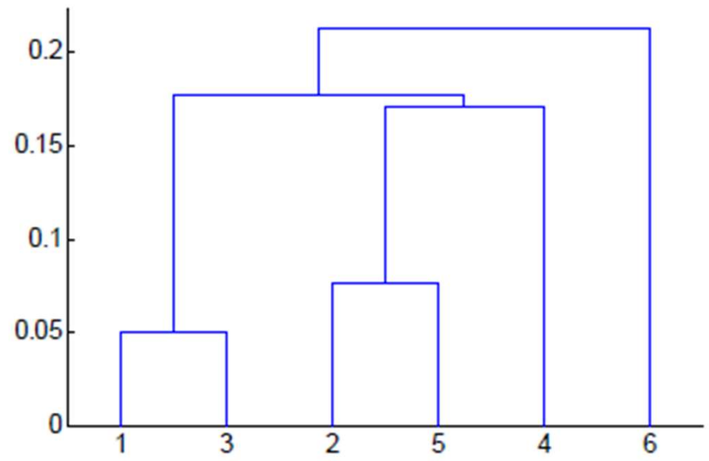
souravkarmakar29@gmail.com

HIERARCHICAL CLUSTERING

- Produces a set of nested clusters organized as a hierarchical tree
- Can be visualized as a dendrogram : A tree like diagram that records the sequences of merges or splits.

Datapoints





Dendrograms

- **Strengths of Hierarchical clustering:**

1. Do not have to assume any particular number of clusters : Any desired number of clusters can be obtained by ‘cutting’ the dendrogram at the proper level.
2. They may correspond to meaningful taxonomies. Example: book cataloguing, biological sciences (e.g. animal kingdom)

HIERARCHICAL CLUSTERING

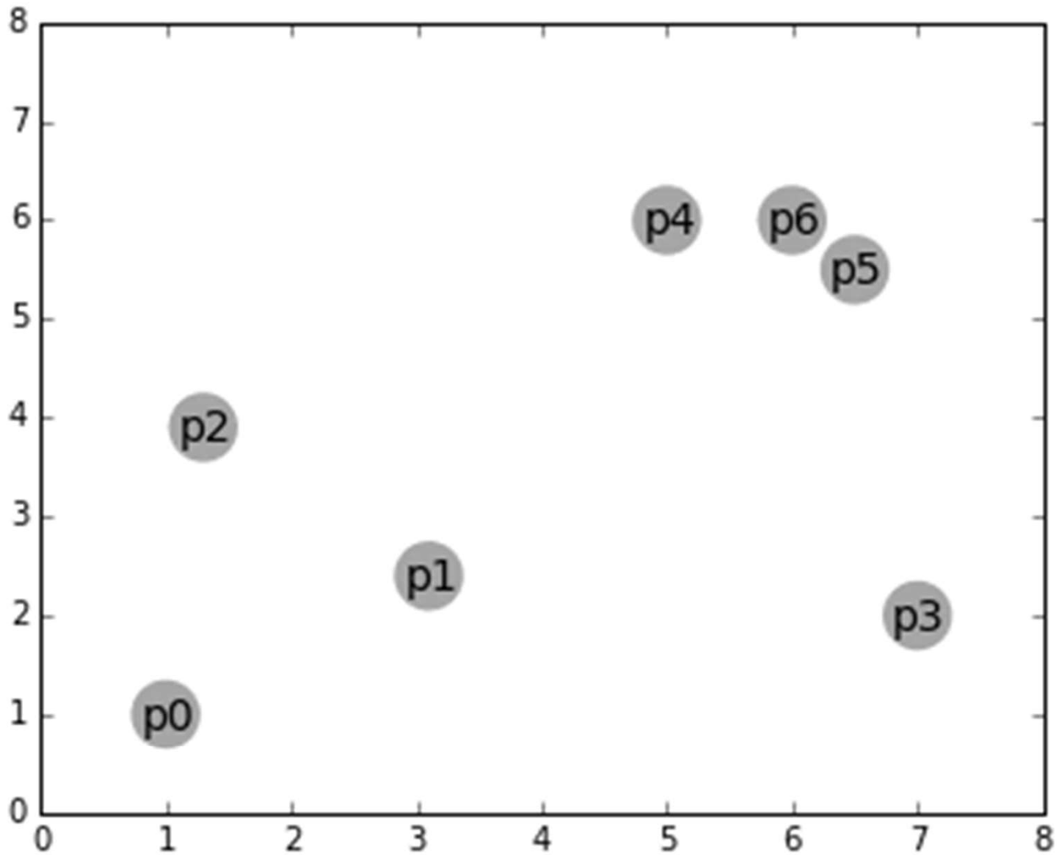
- **Two main types of Hierarchical Clustering**

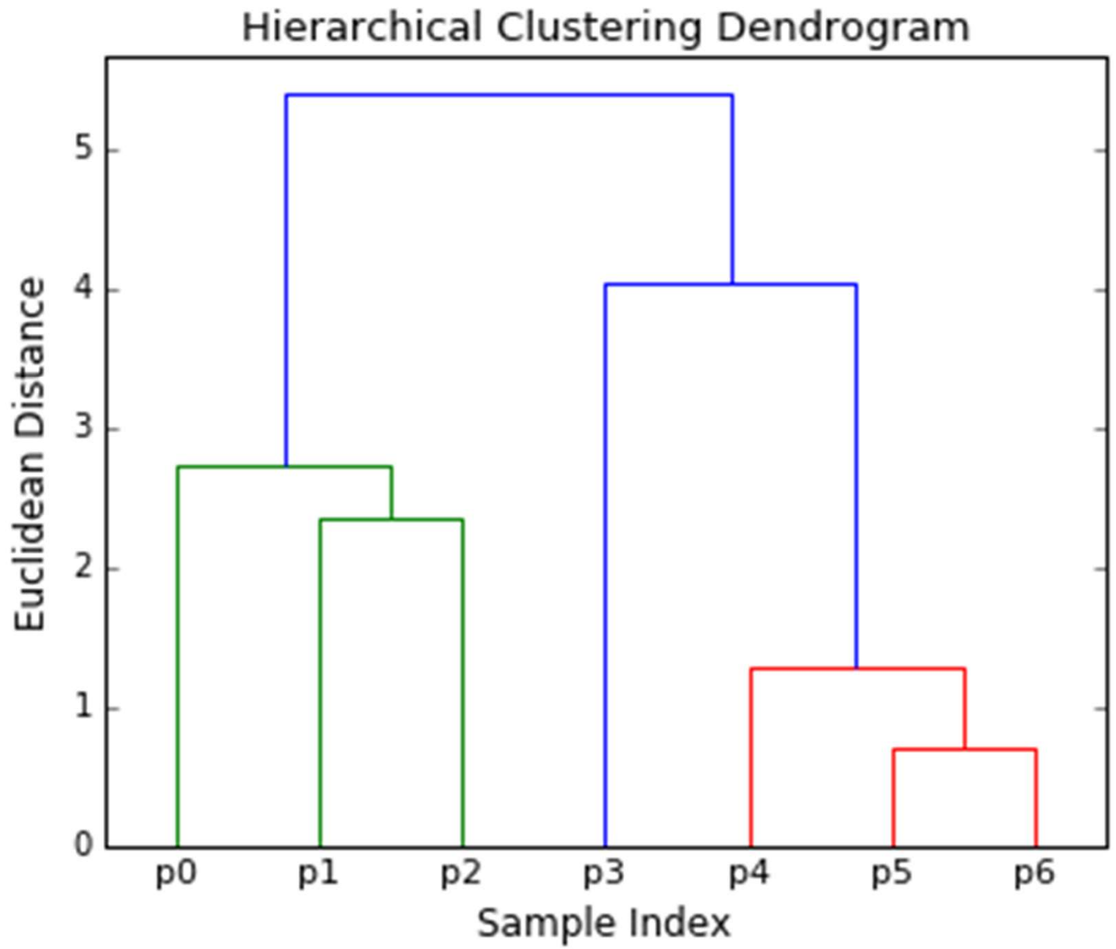
1. **Agglomerative: (Bottom Up Approach)**
    - Start with the points as individual clusters.
    - At each step, merge the closest pair of clusters until you are with only one cluster (or k clusters).

2. **Divisive: (Top Down Approach)**
    - Start with one, all-inclusive cluster
    - At each step, split a cluster until each cluster contains a point (or there are k clusters)

Though both the approaches are perfectly alright in producing quality clusters, people prefer to use Agglomerative approach more than Divisive approach. This is because it is less computationally expensive to deal with smaller clusters as compared to the large all inclusive cluster.

AGGLOMERATIVE CLUSTERING





AGGLOMERATIVE CLUSTERING

- Traditional Hierarchical Clustering algorithms use a similarity or distance matrix to split or merge the cluster. This matrix is known as **Proximity Matrix.**

- The basic Agglomerative Clustering algorithm is as following**.**

1. Let each data point be a Cluster.
2. Compute the proximity matrix.
3. ***Repeat***
4. Merge the two closest clusters.
5. Update the proximity matrix.
6. ***Until*** only a single (or K many) cluster remains. (K is user provided number of clusters)

- Different approaches on defining the distance / proximity between two clusters distinguishes different algorithms. However the basic algorithm as shown above is followed everywhere in Agglomerative clustering.

AGGLOMERATIVE CLUSTERING

- How to define inter-cluster Distance / Proximity ?

- Following are the different ways to define the inter-cluster distance / proximity.
  - MIN or Single Linkage
  - MAX or Complete Linkage
  - Group Average
  - Distance Between Centroids

We shall now briefly discuss these methods.

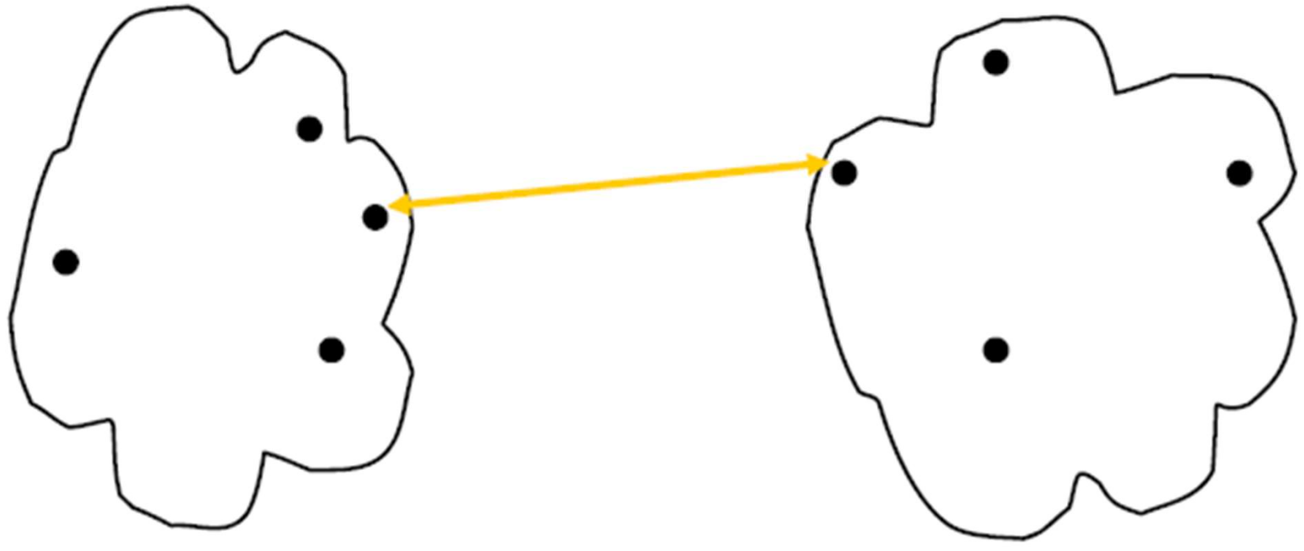
AGGLOMERATIVE CLUSTERING

MIN or Single Linkage:

Here *minimum* of the all inter-cluster distances is considered for proximity calculation.

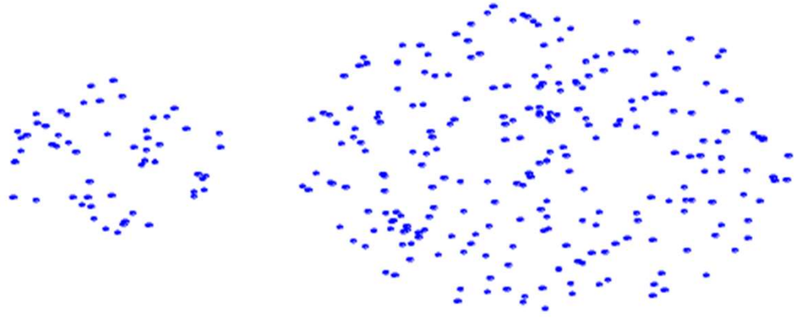
$$d_{C_i,C_j} = minimum\left\{ dist(\vec{\boldsymbol{x}}, \vec{\boldsymbol{y}}) : \vec{\boldsymbol{x}} \in C_i \; and \; \vec{\boldsymbol{y}} \in C_j \right\}$$

Here, d_{C_i,C_j} denotes representative distance between cluster C_i and C_j.

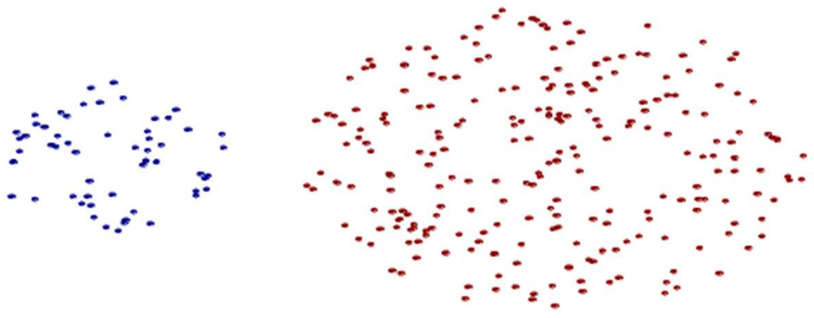


AGGLOMERATIVE CLUSTERING

Strength of MIN or Single Linkage



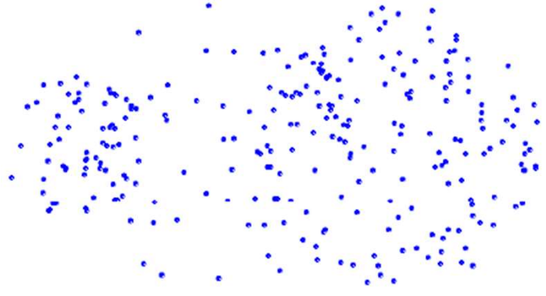
Original Points



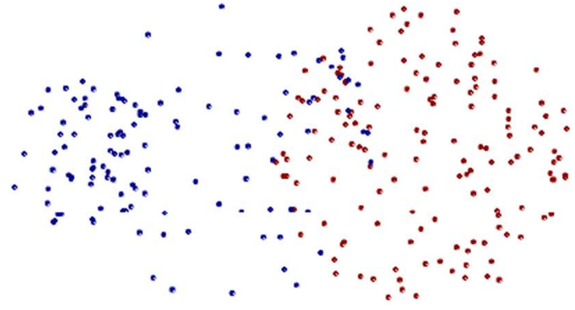
Clustered Points

Single Linkage can handle shapes of different sizes

Limitation of MIN or Single Linkage



Original Points



Clustered Points

Single Linkage is susceptible to noise and outliers

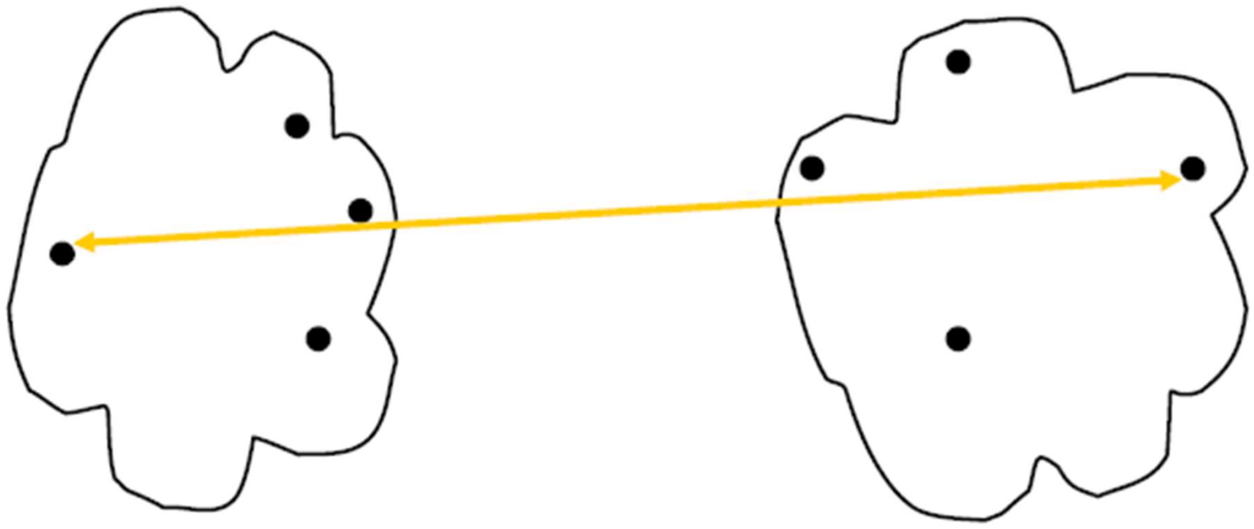
AGGLOMERATIVE CLUSTERING

MAX or Complete Linkage:

Here *maximum* of the all inter-cluster distances is considered for proximity calculation.

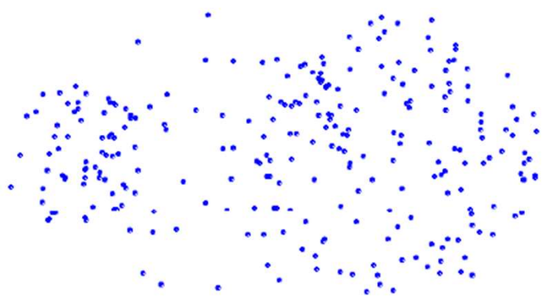
$$d_{C_i,C_j} = maximum\left\{ dist(\vec{\boldsymbol{x}}, \vec{\boldsymbol{y}}) : \vec{\boldsymbol{x}} \in C_i \; and \; \vec{\boldsymbol{y}} \in C_j \right\}$$

Here, d_{C_i,C_j} denotes representative distance between cluster C_i and C_j.

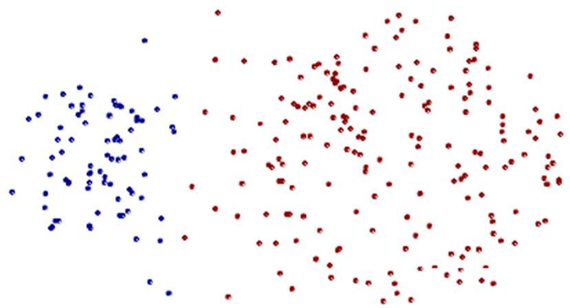


AGGLOMERATIVE CLUSTERING

Strength of MAX or Complete Linkage



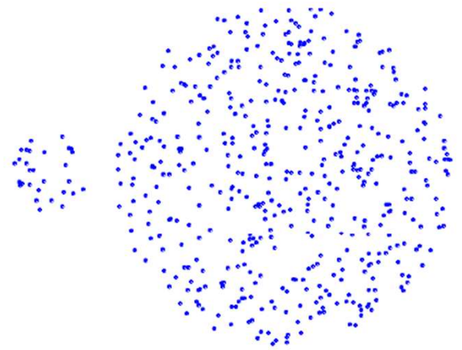
Original Points



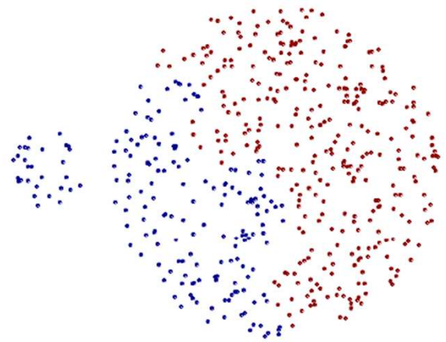
Clustered Points

Complete Linkage is less susceptible to noise

Limitation of MAX or Complete Linkage



Original Points



Clustered Points

Complete Linkage tends to break large clusters

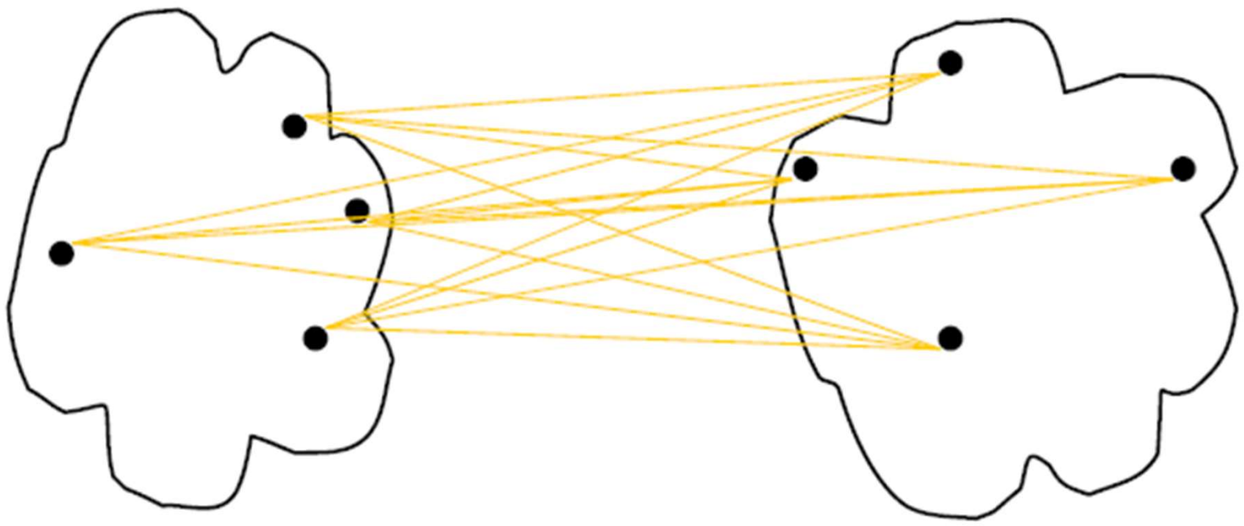
AGGLOMERATIVE CLUSTERING

Group Average:

Here *Average* of the all inter-cluster distances is considered for proximity calculation.

$$d_{C_i,C_j} = Average\left\{ dist(\vec{\boldsymbol{x}}, \vec{\boldsymbol{y}}) : \vec{\boldsymbol{x}} \in C_i \; and \; \vec{\boldsymbol{y}} \in C_j \right\} = \frac{1}{|C_i||C_j|} \sum_{\vec{\boldsymbol{x}} \in C_i, \vec{\boldsymbol{y}} \in C_j} dist(\vec{\boldsymbol{x}}, \vec{\boldsymbol{y}})$$

Here, d_{C_i,C_j} denotes representative distance between cluster C_i and C_j.



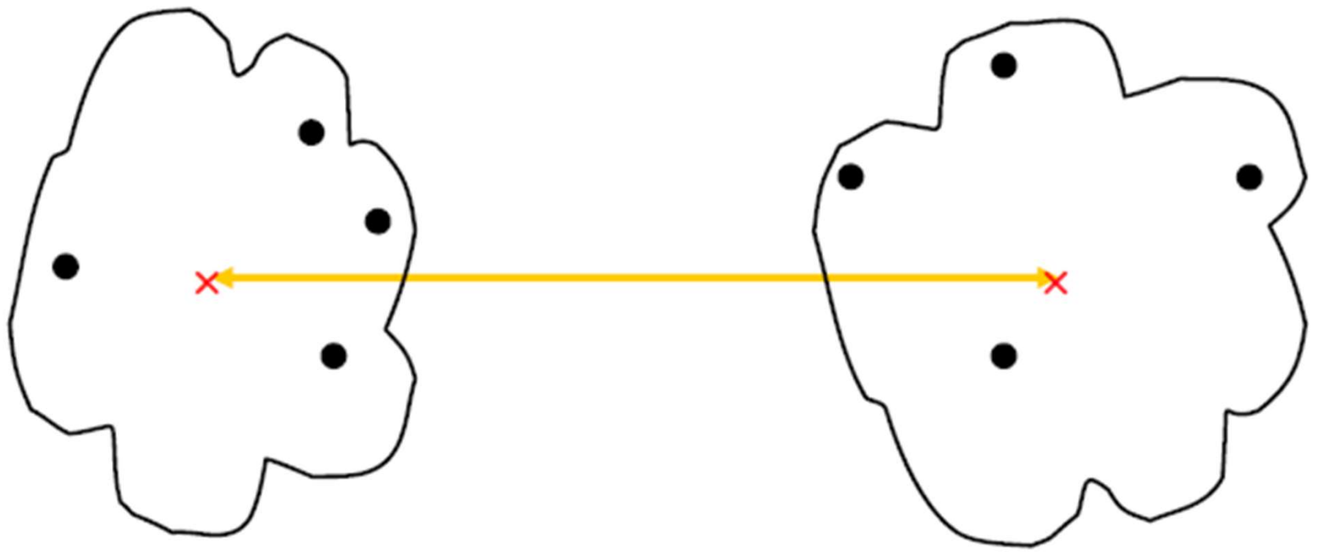
AGGLOMERATIVE CLUSTERING

Distance Between Centroids:

Centroid of cluster C_i is computed as $\vec{\boldsymbol{\mu}}_i = \frac{1}{|C_i|} \sum_{\vec{\boldsymbol{x}} \in C_i} \vec{\boldsymbol{x}}$, Similarly we compute centroid $\vec{\boldsymbol{\mu}}_j$ of cluster C_j.

$$d_{C_i,C_j} = dist\left(\vec{\boldsymbol{\mu}}_i, \vec{\boldsymbol{\mu}}_j\right)$$

Here, d_{C_i,C_j} denotes representative distance between cluster C_i and C_j.



Thank You