Assignment 4 3DM3

Jasraj Singh Johal

16 March 2025

Assignment 4.1 (30 marks)

Part (a)

The Single-Link Agglomerative Hierarchical Clustering method follows these steps to merge clusters based on the smallest distance between any pair of points in two clusters. The process of clustering the given data set $\{1, 2, 3, 4, 5\}, \{8, 9, 10, 11, 12\}, \{24, 28, 32, 36, 45\}$ is outlined below:

Step 1: Merge Cluster $\{1, 2, 3, 4, 5\}$ **Pairwise**
Since the distance between consecutive points in this set is 1, they are merged step-by-step:

- Merge $\{1\}$ and $\{2\}$.
- Merge the resulting cluster with $\{3\}$.
- Merge the new cluster with $\{4\}$.
- Finally, merge with $\{5\}$.

This forms a cohesive cluster that accurately reflects the natural grouping of these points.

Step 2: Merge Cluster $\{8, 9, 10, 11, 12\}$ **Pairwise**
Similarly, the distance between consecutive points is also 1 in this set:

- Merge $\{8\}$ and $\{9\}$.
- Merge the resulting cluster with $\{10\}$.
- Merge with $\{11\}$.
- Finally, merge with $\{12\}$.

This results in a tightly connected cluster similar to the previous step.

Step 3: Merge Cluster $\{24, 28, 32, 36\}$ **Pairwise**
Here, the distance between consecutive points is 4:

- Merge $\{24\}$ and $\{28\}$.
- Merge the resulting cluster with $\{32\}$.

- Merge with $\{36\}$.

This creates a compact cluster, though with slightly larger gaps between points compared to the previous clusters.

Step 4: Merge Cluster $\{24, 28, 32, 36\}$ with $\{45\}$
Since the smallest distance between the last point (36) and 45 is 9, they are merged next.

Step 5: Merge Cluster $\{1, 2, 3, 4, 5\}$ with $\{8, 9, 10, 11, 12\}$
The minimum distance between these two clusters is 3 (between points 5 and 8), forming a larger cluster.

Step 6: Merge the Remaining Two Clusters
Finally, the distance between the two remaining clusters (12 and 24) is 12, so they are merged to complete the dendrogram.

Dendrogram

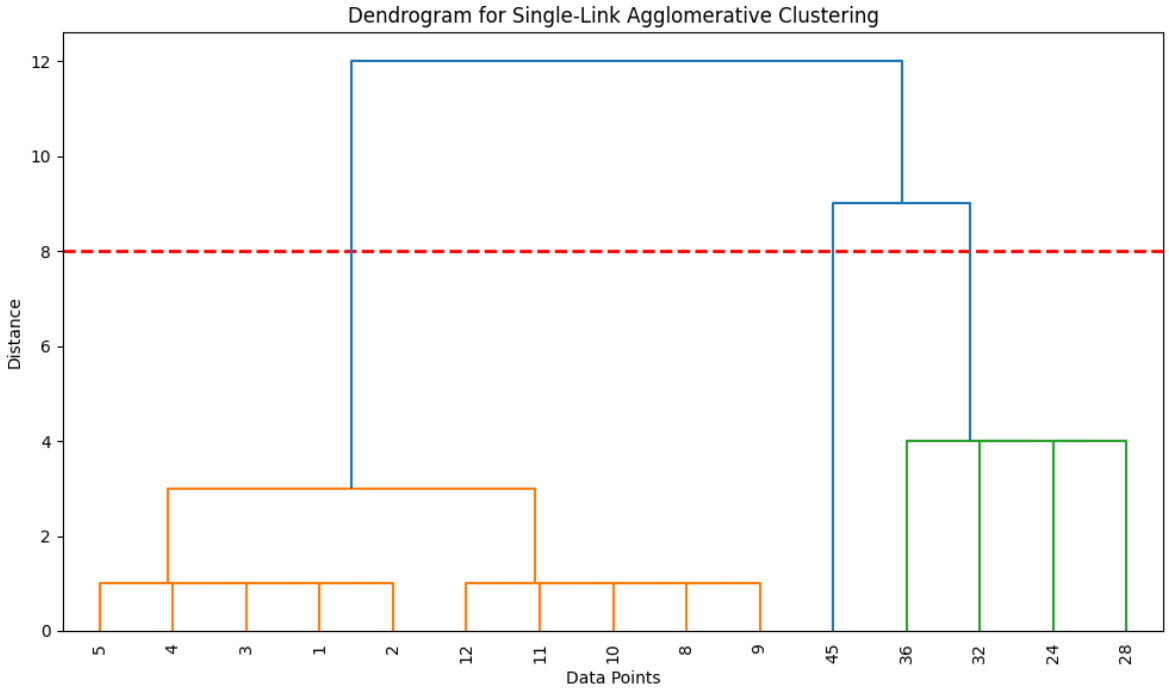

Figure 1:

Part (b)

To obtain three clusters, we cut the dendrogram at a height between 4 and 9. For practical purposes, we choose a height of 5. This cut divides the dataset into three distinct clusters as follows:

- **Cluster 1:** $\{1, 2, 3, 4, 5, 8, 9, 10, 11, 12\}$
  This cluster contains the first two natural groups of consecutive points. Since the distance between the clusters $\{1, 2, 3, 4, 5\}$ and $\{8, 9, 10, 11, 12\}$ is minimal (3

between points 5 and 8), they are merged into a single larger cluster. Despite combining two natural clusters, this grouping still preserves reasonable proximity between points.

- **Cluster 2:** $\{24, 28, 32, 36\}$
  This cluster is formed by merging the points that are spaced by 4 units each. These points naturally group together due to their uniform spacing and relatively close proximity compared to other points.

- **Cluster 3:** $\{45\}$
  The point 45 forms a single-point cluster as it is distant from the nearest cluster (36) by a distance of 9. This large gap prevents merging, making it an isolated cluster.

Cutting the dendrogram at a height of 8 is justified because it efficiently balances the natural grouping of clusters while avoiding forced merging of distant points. The choice of cut height should aim to:

- Preserve natural clusters without forming overly large groups.
- Maintain compact clusters that exhibit cohesion among their points.
- Avoid merging clusters separated by significant distances.

Had the dendrogram been cut at a greater height (e.g., 12), all clusters would be combined into one, losing the original natural grouping. Lower cuts (e.g., height 3) would produce unnecessary fragmentation. Therefore, a height of 8 provides an optimal balance for this dataset.

Part (c)

One of the most significant drawbacks of the Single-Link clustering method is its sensitivity to the chaining effect. This effect occurs when clusters are merged solely based on the shortest pairwise distance, even if the overall clusters are far apart.

- The chaining effect can cause clusters to be connected by a series of points that form an elongated, loosely related cluster, even when the main groups should remain separate.
- This can result in merging clusters that are naturally distinct if intermediate points connect them through a chain of short distances.
- In more complex or noisy datasets, the method may falsely combine distinct clusters into a single, elongated one, resulting in distorted cluster boundaries.
- For instance, if a few additional points were placed between clusters {1, 2, 3, 4, 5} and {8, 9, 10, 11, 12}, the method might merge these clusters incorrectly, forming one large cluster instead of preserving the natural grouping.

The Single-Link clustering method is therefore useful for capturing elongated and loosely connected clusters but fails when distinct clusters are connected by intermediate points, leading to misleading results.

Assignment 4.2 (70 marks)

Given a set $S = \{\bar{x}_1, \ldots, \bar{x}_n\}$ of n d-dimensional points, the CF-value is defined as:

$$\mathrm{CF}(S) = \langle n, \mathrm{LS}, \mathrm{SS} \rangle$$

where:

- n is the number of points in S,
- $\mathrm{LS} = (\sum_{i=1}^{n} x_{i1}, \sum_{i=1}^{n} x_{i2}, \ldots, \sum_{i=1}^{n} x_{id})$ is the linear sum,
- $\mathrm{SS} = \sum_{i=1}^{n} \sum_{j=1}^{d} x_{ij}^2$ is the sum of squares.

Part (a)

We need to prove that the following clustering metrics can be computed using only $\mathrm{CF}(S)$ without knowing the set S.

1. Centroid (10 marks)

The centroid $\bar{\mu}$ of the set S is defined as:

$$\bar{\mu} = \frac{1}{n} \left(\sum_{i=1}^{n} x_{i1}, \ldots, \sum_{i=1}^{n} x_{id} \right)$$

From the definition of LS:

$$\mathrm{LS} = \left(\sum_{i=1}^{n} x_{i1}, \sum_{i=1}^{n} x_{i2}, \ldots, \sum_{i=1}^{n} x_{id} \right)$$

So, we can write the centroid as:

$$\bar{\mu} = \frac{1}{n} \mathrm{LS}$$

Since n and LS are parts of $\mathrm{CF}(S)$, we can compute $\bar{\mu}$ directly from $\mathrm{CF}(S)$.

2. Radius R (10 marks)

The radius R is defined as:

$$R = \sqrt{\frac{\sum_{i=1}^{n} (\bar{x}_i - \bar{\mu})^2}{n}}$$

Here, $(\bar{x}_i - \bar{\mu})^2$ is the squared Euclidean distance between a point $\bar{x}_i$ and the centroid $\bar{\mu}$. In d-dimensions, this becomes:

$$(\bar{x}_i - \bar{\mu})^2 = \sum_{j=1}^{d} (x_{ij} - \mu_j)^2$$

Thus:

$$\sum_{i=1}^{n}(\bar{x}_i - \bar{\mu})^2 = \sum_{i=1}^{n}\sum_{j=1}^{d}(x_{ij} - \mu_j)^2$$

Expand the squared term:

$$(x_{ij} - \mu_j)^2 = x_{ij}^2 - 2x_{ij}\mu_j + \mu_j^2$$

Summing over all points and dimensions:

$$\sum_{i=1}^{n}\sum_{j=1}^{d}(x_{ij} - \mu_j)^2 = \sum_{i=1}^{n}\sum_{j=1}^{d}x_{ij}^2 - 2\sum_{j=1}^{d}\mu_j\sum_{i=1}^{n}x_{ij} + \sum_{j=1}^{d}\mu_j^2\sum_{i=1}^{n}1$$

Now substitute the known terms:

- $\sum_{i=1}^{n}\sum_{j=1}^{d}x_{ij}^2 = \text{SS}$,
- $\sum_{i=1}^{n}x_{ij} = \text{LS}_j$ (the j-th component of LS),
- $\sum_{i=1}^{n}1 = n$,
- $\mu_j = \frac{1}{n}\text{LS}_j$ (since $\bar{\mu} = \frac{1}{n}\text{LS}$).

Plugging these in:

$$\sum_{i=1}^{n}\sum_{j=1}^{d}(x_{ij} - \mu_j)^2 = \text{SS} - 2\sum_{j=1}^{d}\left(\frac{1}{n}\text{LS}_j\right)\text{LS}_j + n\sum_{j=1}^{d}\left(\frac{1}{n}\text{LS}_j\right)^2$$

Simplify:

$$= \text{SS} - \frac{2}{n}\sum_{j=1}^{d}\text{LS}_j^2 + \frac{1}{n}\sum_{j=1}^{d}\text{LS}_j^2$$

$$= \text{SS} - \frac{1}{n}\sum_{j=1}^{d}\text{LS}_j^2$$

So:

$$R = \sqrt{\frac{\text{SS} - \frac{1}{n}\sum_{j=1}^{d}\text{LS}_j^2}{n}}$$

Since SS, LS, and n are all in CF(S), we can compute R using only CF(S).

3. Average Pairwise Distance D within a Cluster (20 marks)

The average pairwise distance D is defined as:

$$D = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}(\bar{x}_i - \bar{x}_j)^2}{n(n-1)}}$$

Expand the squared distance:

$$(\bar{x}_i - \bar{x}_j)^2 = \sum_{k=1}^{d}(x_{ik} - x_{jk})^2 = \sum_{k=1}^{d}(x_{ik}^2 - 2x_{ik}x_{jk} + x_{jk}^2)$$

So:

$$\sum_{i=1}^{n}\sum_{j=1}^{n}(\bar{x}_i - \bar{x}_j)^2 = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{d}(x_{ik}^2 - 2x_{ik}x_{jk} + x_{jk}^2)$$

Break it into three parts:

1. $\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{d} x_{ik}^2 = \sum_{k=1}^{d}\sum_{i=1}^{n} x_{ik}^2 \sum_{j=1}^{n} 1 = n \cdot \text{SS}$,
2. $-2\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{d} x_{ik}x_{jk} = -2\sum_{k=1}^{d}\left(\sum_{i=1}^{n} x_{ik}\right)\left(\sum_{j=1}^{n} x_{jk}\right) = -2\sum_{k=1}^{d}\text{LS}_k^2$,
3. $\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{d} x_{jk}^2 = n \cdot \text{SS}$ (same as the first term).

Combine them:

$$\sum_{i=1}^{n}\sum_{j=1}^{n}(\bar{x}_i - \bar{x}_j)^2 = n \cdot \text{SS} - 2\sum_{k=1}^{d}\text{LS}_k^2 + n \cdot \text{SS} = 2n \cdot \text{SS} - 2\sum_{k=1}^{d}\text{LS}_k^2$$

Thus:

$$D = \sqrt{\frac{2n \cdot \text{SS} - 2\sum_{k=1}^{d}\text{LS}_k^2}{n(n-1)}}$$

Since n, SS, and LS are in CF(S), we can compute D using only CF(S).

Part (b) (30 marks)

Clustering algorithms such as k-means and **agglomerative hierarchical clustering** may produce different results when applied to the **CF-values** in the leaf nodes of a **CF-tree** compared to the original dataset. This difference stems from the approximation caused by summarizing data points into **micro-clusters**. Rather than working with individual points, these algorithms use **summary statistics**, typically **centroids**, which can alter **distance computations**. In k-means, for example, this affects **cluster assignments** since the distances between centroids may not match the distances between original points, potentially leading to less accurate **cluster boundaries**.

CF-values do not completely represent the **geometric properties** of the original point distribution, which can impact algorithms sensitive to **cluster shapes**, particularly for **non-convex clusters**. In hierarchical clustering, the **merging or splitting** of clusters relies on distances between CF-values, which may differ from distances between individual points, thus altering the resulting **hierarchy**. The structure of the **CF-tree**, determined by parameters like the **branching factor** and **radius threshold** also shapes the micro-clusters and influences the final clustering.