Assignment 3

Durbasmriti Saha

EE708: FUNDAMENTALS OF DATA SCIENCE AND MACHINE INTELLIGENCE

March 7, 2025

1) a) * A metric similarity measure s satisfies $s(x, y) \geq 0$, hence $s(x, y) + a \geq 0$
* since s(x, y) = s(y, x) hence s(x, y) + a = s(y, x) + a
* from its identity property, $s(x, x) = \max_{u,v \in X} s(u, v)$ then s(x, x) + a = $s(x, x) = \max_{u,v \in X} s(u, v)$ + a = $s(x, x) = \max_{u,v \in X}(s(u, v) + a)$
Thus minimum value of s(u, v) + a is s(x, x) + a.

s(x, y) holds triangular inequality i.e. $s(x, z) \geq s(x, y)n + s(y, z) - \max_{u,v \in X} s(u, v)$.
Add a on both sides:
$s(x, z) + a \geq s(x, y)n + s(y, z) - \max_{u,v \in X} s(u, v) + a$
Hence, $s(x, z)+a \geq (s(x, y)+a)+(s(y, z)+a)-\max_{u,v \in X}(s(u, v)+a)$ $\because \max_{u,v \in X}(s(u, v) + a) = \max_{u,v \in X} s(u,$
Thus triangular inequality is being satisfied by s(x, y) + a.
Hence s(x, y) + a is also a metric similarity measure.

b) Since $d(x, y) \geq 0$ and $a \geq 0$, So, $d(x, y) + a \geq 0$
By symmetry property, d(x, y) = d(y,x) that means d(x,y) + a = d(y,x) + a
With its identity property, d(x,x) + a = 0+a since d(x,x) = 0.
But for d(x,y) to be a metric dissimilarity measure, d(x,x) + a must be equal to 0 means a = 0. Hence d(x,y) +a is a metric dissimilarity measure only when a = 0.

2) Triangular inequality states that $d(x, z) \leq d(x, y) + d(y, z)$ where d(x,y) is the euclidean distance between x and y given by $d(x, y) = \sum_{i=1}^{l}(x_i - y_i)^2$
The Minkowski inequality states that $\left(\sum_{i=1}^{l} |x_i + y_i|^p\right)^{\frac{1}{p}} \leq \left(\sum_{i=1}^{l} |x_i|^p\right)^{\frac{1}{p}} + \left(\sum_{i=1}^{l} |y_i|^p\right)^{\frac{1}{p}}$
where x = $[x_1, \ldots, x_l]^T$ and y = $[y_1, \ldots, y_l]^T$
For euclidean distance, p = 2:
Proof:
In $d(x, z) = \sqrt{\sum_{i=1}^{l}(x_i - z_i)^2}$, rewrite (x-z) as (x-z) = (x-y) + (y-z) and substitute this valuse in d(x,y)
$d(x, z) = \sqrt{\sum_{i=1}^{l}((x_i - y_i) + (y_i - z_i))^2}$

let u = x-y and v = y-z. then:

$d(x, z) = \sqrt{\sum_{i=1}^{l}(u_i + v_i)^2}$ and by Minkowski inequality with p = 2:

$$\sqrt{\sum_{i=1}^{l}(u_i + v_i)^2} \leq \sqrt{\sum_{i=1}^{l}(u_i)^2} + \sqrt{\sum_{i=1}^{l}(v_i)^2}$$

put u and v with their original values:

$$\sqrt{\sum_{i=1}^{l}(x_i - z_i)^2} \leq \sqrt{\sum_{i=1}^{l}(x_i - y_i)^2} + \sqrt{\sum_{i=1}^{l}(y_i - z_i)^2}$$

Means $d(x, z) \leq d(x, y) + d(y, z)$ — Hence proved!

3) The properties of a distance metric, d(x,y) are :
1) $d(x, y) \geq 0$ and d(x,y) = 0 when x=y.(Non-negativity)
2) d(x,y) = d(y,x) : symmetry
3) d(x,x) = 0 : Identity
4) Triangular inequality : $d(x, z) \leq d(x, y) + d(y, z)$.

Now the give distance metric is : $d(x, y) = |x - y|^2$
1) Non-negativity:
since $|x - y| \geq 0$ means $|x - y|^2 \geq 0$ So, $d(x, y) \geq 0$.
2)Symmetry :
Since $|x - y| = |y - x|$, So $|x - y|^2 = |y - x|^2$ means d(x,y) = d(y,x)
3) Identity: d(x,x) = $|x - x|^2 = 0$
4) Triangular Inequality:
To check triangular inequality let's take an example x = 0, y = 1, z = 2: then
d(x,z) = $|0 - 2|^2 = 4$
d(x,y) = $|0 - 1|^2 = 1$
d(y,z) = $|1 - 2|^2 = 1$
here $4 \leq 1 + 1$ means $4 \leq 2$.
Clearly d(x,y) doesn't satisfy triangular inequality.
Hence d(x,y) = $|x - y|^2$ is not a valid distance metric.

4) First we have to calculate $d(x, x_i)$ for $x_i = x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8$

$$x_1 = \begin{bmatrix} 1.5 \\ 1.5 \end{bmatrix}$$

$$\delta(x, x_1) = (6 - 1.5)^2 + (4 - 1.5)^2 = 4.5^2 + 2.5^2 = 20.25 + 6.25 = 26.5 \approx 5.15$$

$$x_2 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

$$\delta(x, x_2) = (6 - 2)^2 + (4 - 1)^2 = 4^2 + 3^2 = 16 + 9 = 25 = 5$$

$$x_3 = \begin{bmatrix} 2.5 \\ 1.75 \end{bmatrix}$$

$$\delta(x, x_3) = (6-2.5)^2 + (4-1.75)^2 = 3.5^2 + 2.25^2 = 12.25 + 5.0625 = 17.3125 \approx 4.16$$

$$x_4 = \begin{bmatrix} 1.5 \\ 2 \end{bmatrix}$$

$$\delta(x, x_4) = (6-1.5)^2 + (4-2)^2 = 4.5^2 + 2^2 = 20.25 + 4 = 24.25 \approx 4.92$$

$$x_5 = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$$

$$\delta(x, x_5) = (6-3)^2 + (4-2)^2 = 3^2 + 2^2 = 9 + 4 = 13 \approx 3.61$$

$$x_6 = \begin{bmatrix} 1 \\ 3.5 \end{bmatrix}$$

$$\delta(x, x_6) = (6-1)^2 + (4-3.5)^2 = 5^2 + 0.5^2 = 25 + 0.25 = 25.25 \approx 5.02$$

$$x_7 = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$$

$$\delta(x, x_7) = (6-2)^2 + (4-3)^2 = 4^2 + 1^2 = 16 + 1 = 17 \approx 4.12$$

$$x_8 = \begin{bmatrix} 3.5 \\ 3 \end{bmatrix}$$

$$\delta(x, x_8) = (6-3.5)^2 + (4-3)^2 = 2.5^2 + 1^2 = 6.25 + 1 = 7.25 \approx 2.69$$

Now,

$$D_{\min}(x, C) = \min_{v \in C} \{\delta(x, v)\}$$

$$\boxed{D_{\min}(x, C) = \min\{5.15, 5, 4.16, 4.92, 3.61, 5.02, 4.12, 2.69\} = 2.69}$$

$$D_{\max}(x, C) = \max_{v \in C} \{\delta(x, v)\}$$

$$\boxed{D_{\max}(x, C) = \max\{5.15, 5, 4.16, 4.92, 3.61, 5.02, 4.12, 2.69\} = 5.15}$$

$$D_{\text{avg}}(x, C) = \frac{1}{|C|} \sum_{v \in C} \delta(x, v).$$

and

$$\sum_{v \in C} \delta(x, v) = 5.15 + 5 + 4.16 + 4.92 + 3.61 + 5.02 + 4.12 + 2.69 = 34.67.$$

and

$$\boxed{D_{\text{avg}}(x, C) = \frac{34.67}{8} \approx 4.33.}$$

5)

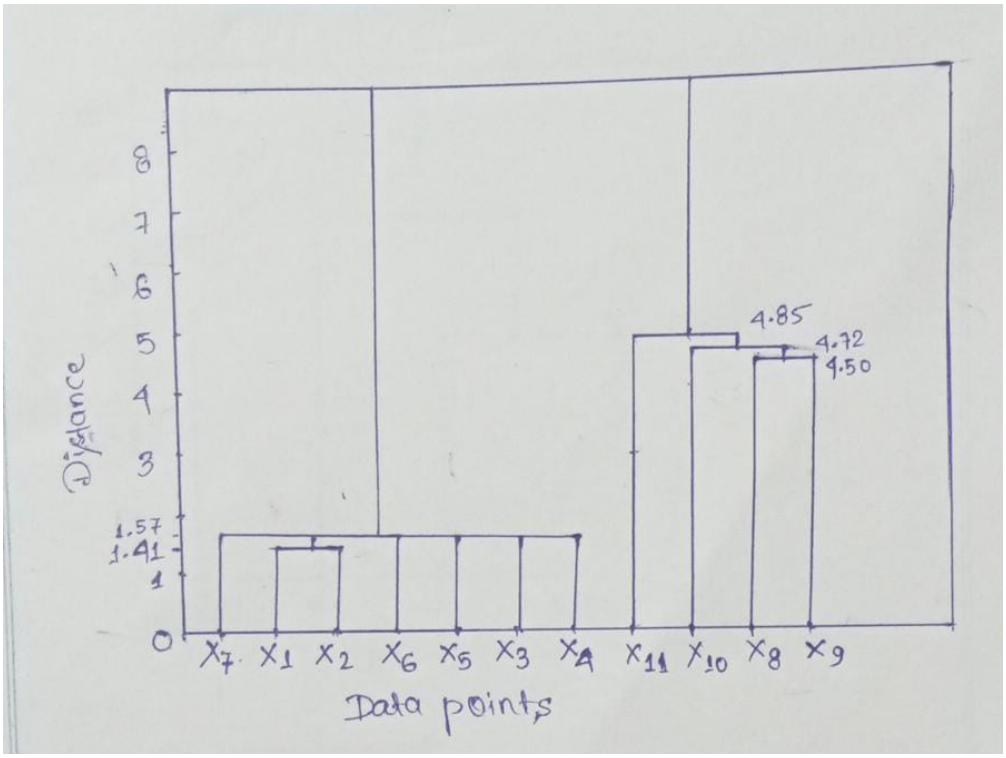

Figure 1: Dendrodram

7) Yes, the dendrogram may be pruned to achieve a given number of clusters or trim the hierarchical tree at a preferable level of detail. Pruning a dendrogram is about trimming the tree at some height or distance level, the threshold of which dictates the resultant number of formed clusters.

Steps to Prune a Dendrogram:

1)Select a Cutting Threshold

2)Choose a distance threshold or height at which to cut the dendrogram. The threshold sets the level of granularity for clustering. Otherwise we can also choose the number of clusters we desire then Cut the Dendrogram:

Then we Plot a horizontal line at the selected threshold across the dendrogram.

The intersections of this line with the vertical lines of the dendrogram mark the clusters.

3) Form Clusters:

All the leaf nodes (data points) connected below the cutting line are of the same cluster.

The number of clusters is the same as the number of intersections between the horizontal line and the dendrogram branches.

8) The k-means clustering algorithm is vulnerable to outliers because it employs the mean of points in a cluster as the center. Outliers can heavily manipulate the mean so that cluster allocations are poor. To make k-means insensitive to outliers, various approaches may be used:
1) Use of k-medoids (PAM) Instead of k-means:
Medoids are less sensitive to outliers since they use real data points and not averages.
2) Removing outliers:
We can use IQR range to cutoff all the outliers.
We can also use algorithms like DBSCAN to detect and remove outliers.
Z-score can also be used to remove points having z-score higher than a threshold.
3)Use of better Distance metrics:
Distance metrics like **Manhattan Distance** and **Mahalanobis distance** for better handling of outliers.
4) Use of Trimmed K-means:
Trimmed k-means removes a certain percentage of the extreme o=points.
5) Use of Robust clustering algorithms:
Algorithms like **DBSCAN**, **OPTICS** and **GMM** are better at handling outliers.

6)

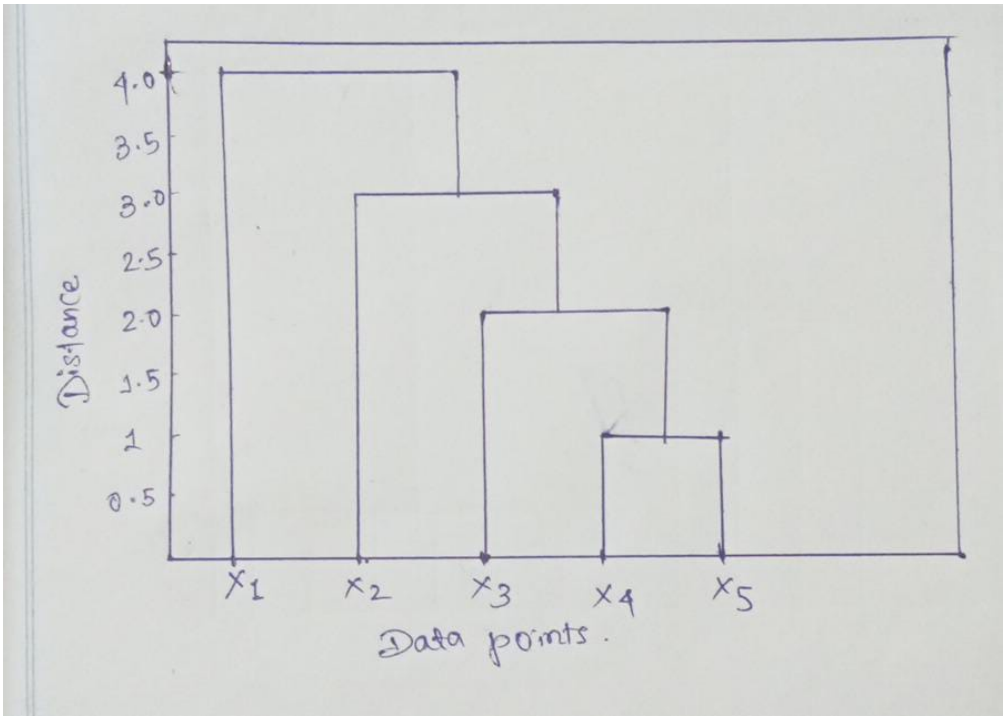

Figure 2: Single Linkage Dendrogram

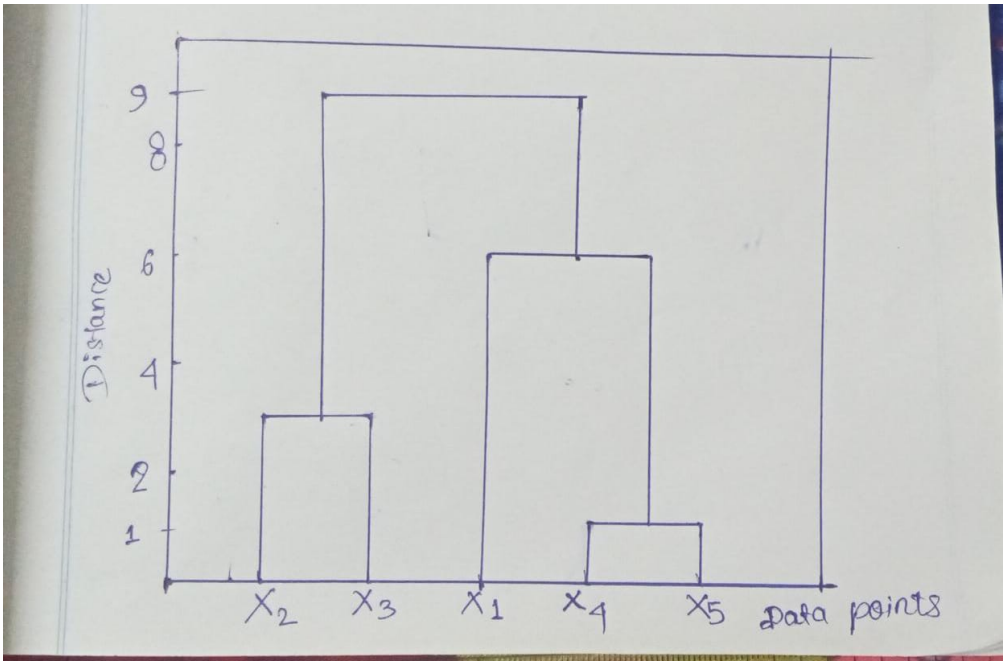

Figure 3: Complete Linkage Dendrogram