# STSA6823 Assignment 4 Part 2

Madimetja Maredi 2014095653

19 September 2025

## Contents

## Introduction

We are given a distance matrix for five items (a, b, c, d, e) and asked to perform hierarchical clustering using both the **average linkage** and **single linkage** methods, and to draw the corresponding dendrograms for each.

The provided distance matrix is as follows:

$$D = \left(\begin{array}{cccccc} & a & b & c & d & e \\ a & 0 & 4 & 6 & 1 & 6 \\ b & 4 & 0 & 9 & 7 & 3 \\ c & 6 & 9 & 0 & 10 & 5 \\ d & 1 & 7 & 10 & 0 & 8 \\ e & 6 & 3 & 5 & 8 & 0 \end{array}\right)$$

First, we will load the necessary libraries and create the distance object in R.

```
##    a  b  c  d
## b  4
## c  6  9
## d  1  7 10
## e  6  3  5  8
```

### 1. Average Linkage Hierarchical Procedure

For average linkage, the distance between two clusters is the average of the distances between all pairs of items where one item is in the first cluster and the other is in the second.

### Step-by-Step Analysis

1. **Initial State**: The initial clusters are $\{a\}, \{b\}, \{c\}, \{d\}, \{e\}$. The smallest distance in the matrix is $d(a, d) = 1$. We merge these two items.

2. **Merge {a} and {d}**: The new cluster is $\{a, d\}$. The height of this merge is **1**. We calculate the distances from this new cluster to the others:

   - $d(\{a, d\}, b) = (d(a, b) + d(d, b))/2 = (4 + 7)/2 = 5.5$
   - $d(\{a, d\}, c) = (d(a, c) + d(d, c))/2 = (6 + 10)/2 = 8.0$
   - $d(\{a, d\}, e) = (d(a, e) + d(d, e))/2 = (6 + 8)/2 = 7.0$

3. **Merge {b} and {e}**: The smallest distance in the updated matrix is now $d(b, e) = 3$. We merge these two items.

4. **Merge {b} and {e}**: The new cluster is $\{b, e\}$. The height of this merge is **3**. We calculate the distances from this cluster to the remaining clusters:

   - $d(\{b, e\}, c) = (d(b, c) + d(e, c))/2 = (9 + 5)/2 = 7.0$
   - $d(\{b, e\}, \{a, d\}) = (d(b, a) + d(b, d) + d(e, a) + d(e, d))/4 = (4 + 7 + 6 + 8)/4 = 6.25$

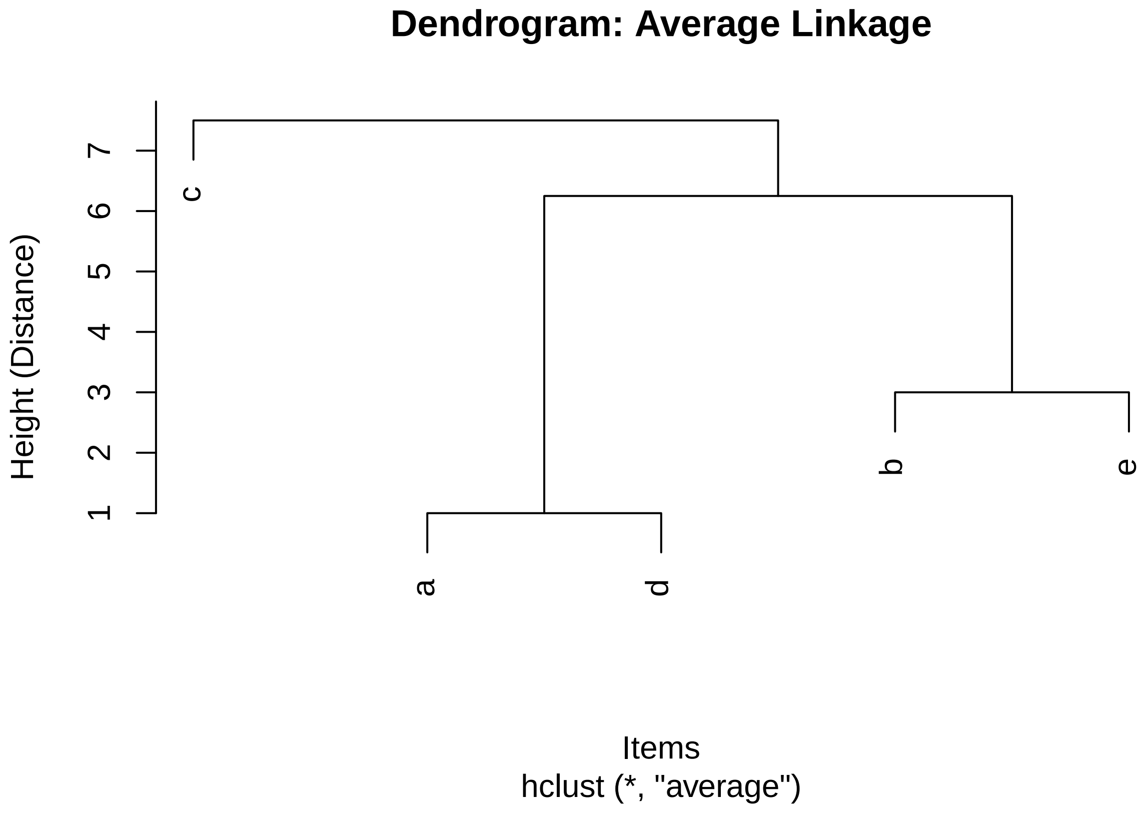
5. **Merge {a,d} and {b,e}**: The minimum distance is now between the two existing clusters, $d(\{a, d\}, \{b, e\}) = 6.25$. We merge them.

6. **Merge {a,d} and {b,e}**: The new cluster is $\{a, d, b, e\}$. The height of this merge is **6.25**. The final distance to cluster $\{c\}$ is:

   - $d(\{a, d, b, e\}, c) = (d(a, c) + d(d, c) + d(b, c) + d(e, c))/4 = (6 + 10 + 9 + 5)/4 = 7.5$

7. **Final Merge**: The last remaining clusters $\{a, d, b, e\}$ and $\{c\}$ are merged at a height of **7.5**.

**R Implementation and Dendrogram** We use the `hclust` function with `method = "average"` to perform the clustering and plot the resulting dendrogram.



The dendrogram visually confirms our step-by-step analysis. Items ‘a’ and ‘d’ merge first at height 1, followed by ‘b’ and ‘e’ at height 3. The cluster {a,d} then merges with {b,e} at a height of 6.25, and finally, this larger cluster merges with ‘c’ at height 7.5.

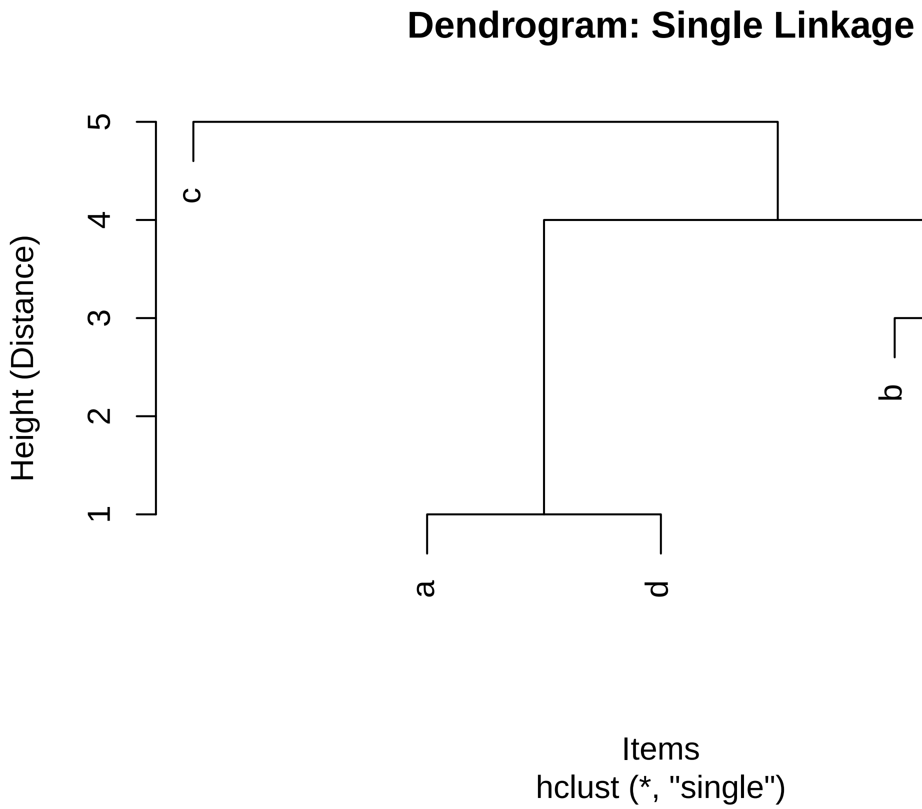
### 2. Single Linkage Hierarchical Procedure

For single linkage (also known as the nearest neighbour method), the distance between two clusters is the minimum distance between an item in the first cluster and an item in the second cluster.

#### Step-by-Step Analysis

1. **Initial State**: The initial clusters are $\{a\}, \{b\}, \{c\}, \{d\}, \{e\}$. The smallest distance is $d(a, d) = 1$. We merge them.
2. **Merge {a} and {d}**: The new cluster is $\{a, d\}$. The height of this merge is **1**. We calculate the distances from this new cluster to the others using the minimum distance:
   - $d(\{a, d\}, b) = \min(d(a, b), d(d, b)) = \min(4, 7) = 4$
   - $d(\{a, d\}, c) = \min(d(a, c), d(d, c)) = \min(6, 10) = 6$
   - $d(\{a, d\}, e) = \min(d(a, e), d(d, e)) = \min(6, 8) = 6$
3. **Merge {b} and {e}**: The smallest distance in the updated matrix is $d(b, e) = 3$. We merge these two items.
4. **Merge {b} and {e}**: The new cluster is $\{b, e\}$. The height of this merge is **3**. We calculate the distances from this cluster to the remaining clusters:
   - $d(\{b, e\}, c) = \min(d(b, c), d(e, c)) = \min(9, 5) = 5$
   - $d(\{b, e\}, \{a, d\}) = \min(d(b, a), d(b, d), d(e, a), d(e, d)) = \min(4, 7, 6, 8) = 4$
5. **Merge {a,d} and {b,e}**: The minimum distance is now $d(\{a, d\}, \{b, e\}) = 4$. We merge them.
6. **Merge {a,d} and {b,e}**: The new cluster is $\{a, d, b, e\}$. The height of this merge is **4**. The final distance to cluster $\{c\}$ is:
   - $d(\{a, d, b, e\}, c) = \min(d(a, c), d(d, c), d(b, c), d(e, c)) = \min(6, 10, 9, 5) = 5$
7. **Final Merge**: The last remaining clusters $\{a, d, b, e\}$ and $\{c\}$ are merged at a height of **5**.

**R Implementation and Dendrogram** We use the `hclust` function with `method = "single"` to perform the clustering and plot the resulting dendrogram.



The dendrogram for single linkage shows a different structure compared to average linkage. While the first two merges ({a,d} at height 1 and {b,e} at height 3) are the same, the subsequent merge heights differ. The cluster {a,d} merges with {b,e} at height 4, and this group finally merges with 'c' at height 5. This "chaining" effect, where clusters are merged based on the single closest pair of points, is characteristic of the single linkage method.

## R Code

```
# --- Hierarchical Clustering Analysis ---
# Assignment 4, Part 2

# --- 1. Data Setup ---

# Create the full symmetric matrix from the provided distances.
# The items are 'a', 'b', 'c', 'd', 'e'.
dist_matrix <- as.matrix(data.frame(
  a = c(0, 4, 6, 1, 6),
  b = c(4, 0, 9, 7, 3),
  c = c(6, 9, 0, 10, 5),
  d = c(1, 7, 10, 0, 8),
  e = c(6, 3, 5, 8, 0)
))

# Set the row and column names to match the items.
```

```r
rownames(dist_matrix) <- colnames(dist_matrix)

# Convert the matrix into a 'dist' object, which is the required
# format for R's clustering functions.
dist_object <- as.dist(dist_matrix)

# Print the distance object to verify it was created correctly.
print("--- Initial Distance Matrix ---")
print(dist_object)


# --- 2. Average Linkage Clustering ---

# Perform hierarchical clustering using the average linkage method.
hc_average <- hclust(dist_object, method = "average")

# Plot the resulting dendrogram for average linkage.
# The `main` argument sets the title of the plot.
plot(hc_average,
     main = "Dendrogram: Average Linkage",
     xlab = "Items",
     ylab = "Height (Distance)",
     sub = "") # Removes the default subtitle


# --- 3. Single Linkage Clustering ---

# Perform hierarchical clustering using the single linkage method.
hc_single <- hclust(dist_object, method = "single")

# Plot the resulting dendrogram for single linkage.
plot(hc_single,
     main = "Dendrogram: Single Linkage",
     xlab = "Items",
     ylab = "Height (Distance)",
     sub = "") # Removes the default subtitle
```