31 Hierarchical Clustering: Building a Tree from the Data

31.1 The Big Picture: From Points to a Tree of Groups

31.1.1 Two Roads to Hierarchy: Bottom-Up and Top-Down Clustering

Hierarchical clustering arranges data into a tree of nested groups, letting you explore cluster structure at many scales. There are two complementary approaches: agglomerative (bottom-up) and divisive (top-down). In agglomerative clustering each sample starts as its own cluster and the algorithm repeatedly merges the closest pair of clusters until a single cluster remains.

Divisive clustering works in the opposite direction, beginning with one cluster that is successively split. The result of either approach is visualized with a dendrogram, a tree diagram that shows how and when clusters join or split. Figure 1 shows a dendrogram produced from the Iris dataset using its first two features (sepal length and sepal width). The vertical position where two branches join encodes the distance between those clusters, so cutting the tree at different heights gives different clusters.

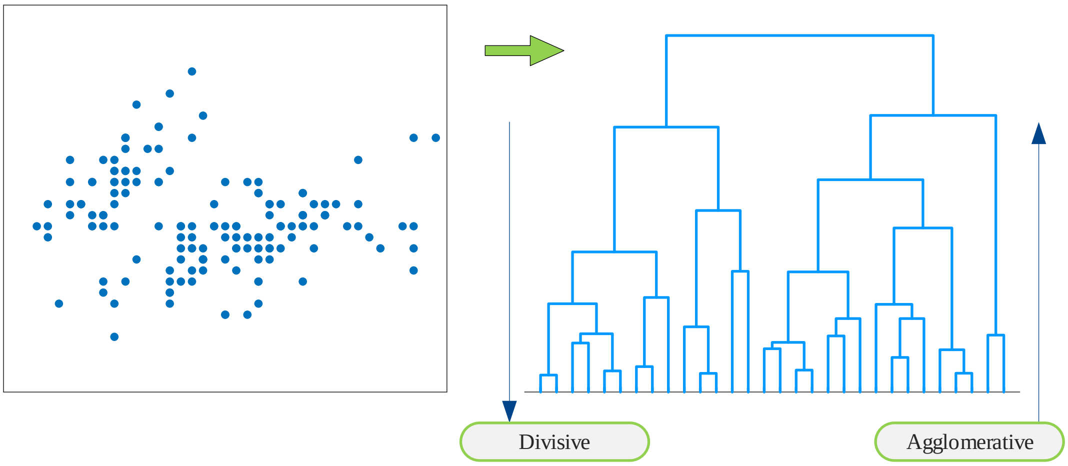


Figure 1. Dendrogram: Top-Down (Divisive) versus Bottom-Up (Agglomerative) Hierarchical Clustering

31.1.2 Distance and Linkage: The Rules That Shape Clusters

A key point in agglomerative clustering is how we measure distance. We need both a point-to-point distance (for example, Euclidean) and a rule for cluster-to-cluster distance, called a linkage method. Common linkages are single (nearest neighbor), complete (farthest neighbor), average, and Ward’s method (minimizes within-cluster variance). The choice of distance and linkage strongly affects the shape and number of clusters.

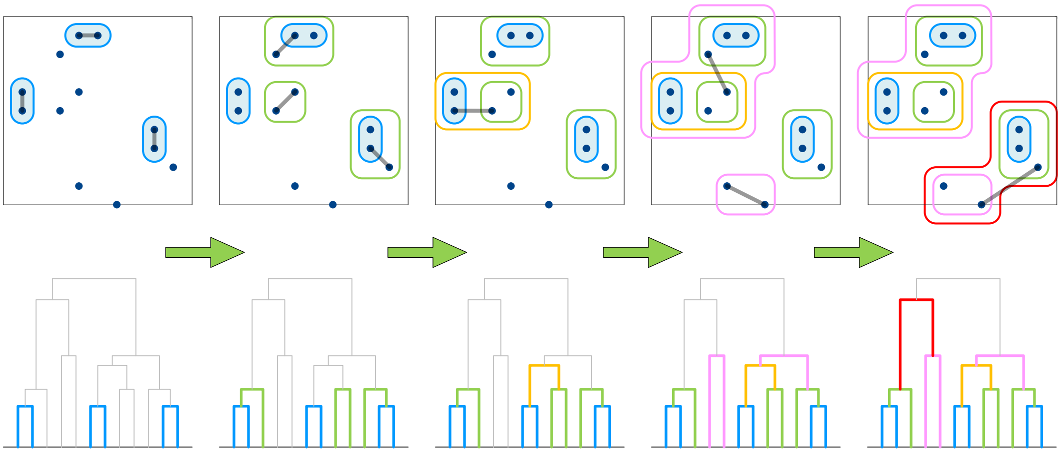


Figure 2. Agglomerative Merging: Start from Singletons and Merge Closest Pairs Repeatedly

Finally, hierarchical clustering is non-inductive: it does not produce a compact predictive model for assigning new, unseen points to clusters without re-computing (or using a chosen cut and a nearest-center rule). In this chapter we focus on the step-by-step construction of a dendrogram by bottom-up merging and on how different linkage rules change the cluster tree.

31.2 Seeing the Tree Grow: The Dendrogram in Action

31.2.1 From Coordinates to Distances

Figure 3 shows the positions of 12 sample points on a two-dimensional plane. We can compute a pairwise distance matrix to measure how far every point is from every other point. Using Euclidean distance, we obtain the heatmap in Figure 4, where darker colors indicate larger distances.

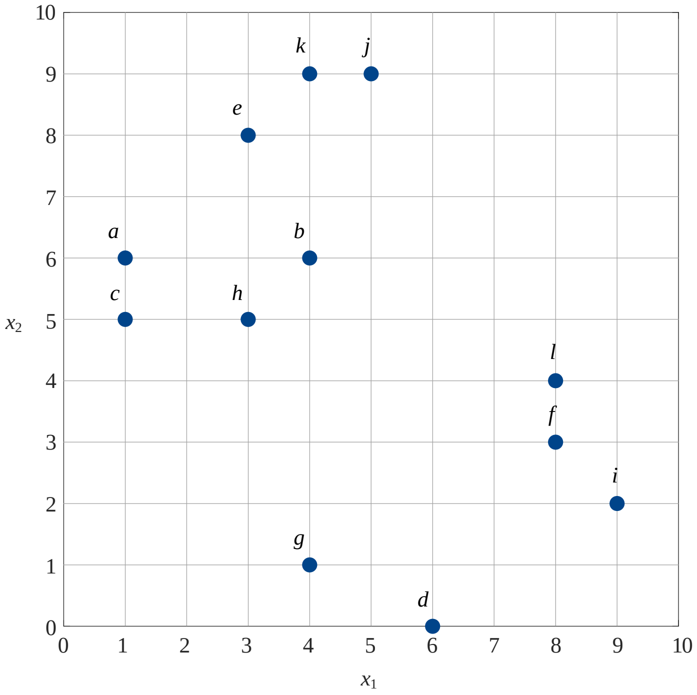
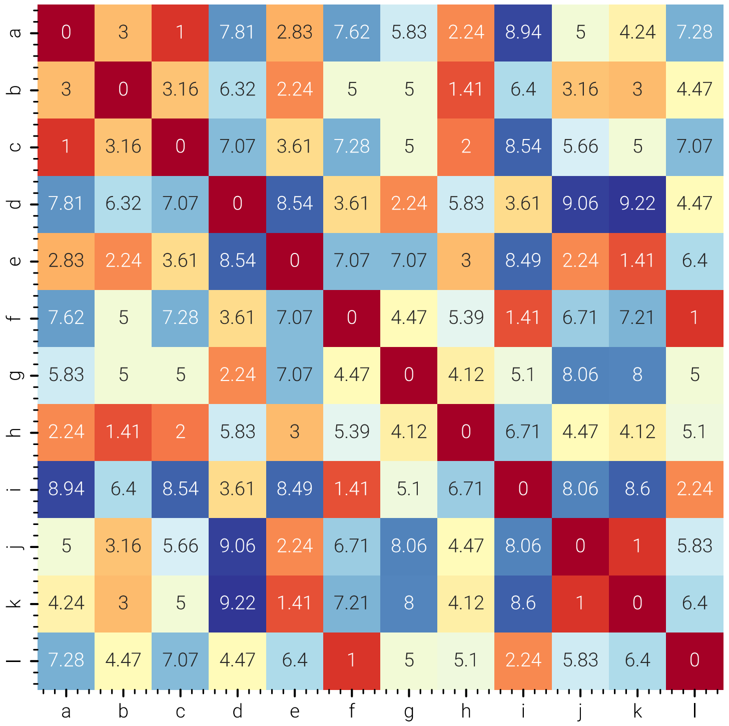


Figure 3. Spatial distribution of 12 sample points. Figure generated by Ch31_01_hierarchical_clustering.ipynb.

From the distance matrix, we can construct a dendrogram, shown in Figure 5. On this plot, the horizontal axis lists the sample indices, and the vertical axis represents the distance at which two points or clusters are merged. By drawing a horizontal cut across the tree, we can decide how many clusters the data should be divided into. For example, if we cut the dendrogram at height 2.5, the data in Figure 3 form three clusters. If we instead cut at height 4, we obtain two clusters. In the remainder of this section, we will show how such a tree is built step by step using the bottom-up (agglomerative) approach.



	a	b	c	d	e	f	g	h	i	j	k	l
a	0	3	1	7.81	2.83	7.62	5.83	2.24	8.94	5	4.24	7.28
b	3	0	3.16	6.32	2.24	5	5	1.41	6.4	3.16	3	4.47
c	1	3.16	0	7.07	3.61	7.28	5	2	8.54	5.66	5	7.07
d	7.81	6.32	7.07	0	8.54	3.61	2.24	5.83	3.61	9.06	9.22	4.47
e	2.83	2.24	3.61	8.54	0	7.07	7.07	3	8.49	2.24	1.41	6.4
f	7.62	5	7.28	3.61	7.07	0	4.47	5.39	1.41	6.71	7.21	1
g	5.83	5	5	2.24	7.07	4.47	0	4.12	5.1	8.06	8	5
h	2.24	1.41	2	5.83	3	5.39	4.12	0	6.71	4.47	4.12	5.1
i	8.94	6.4	8.54	3.61	8.49	1.41	5.1	6.71	0	8.06	8.6	2.24
j	5	3.16	5.66	9.06	2.24	6.71	8.06	4.47	8.06	0	1	5.83
k	4.24	3	5	9.22	1.41	7.21	8	4.12	8.6	1	0	6.4
l	7.28	4.47	7.07	4.47	6.4	1	5	5.1	2.24	5.83	6.4	0

Figure 4. Heatmap of the pairwise Euclidean distance matrix. Figure generated by Ch31_01_hierarchical_clustering.ipynb.

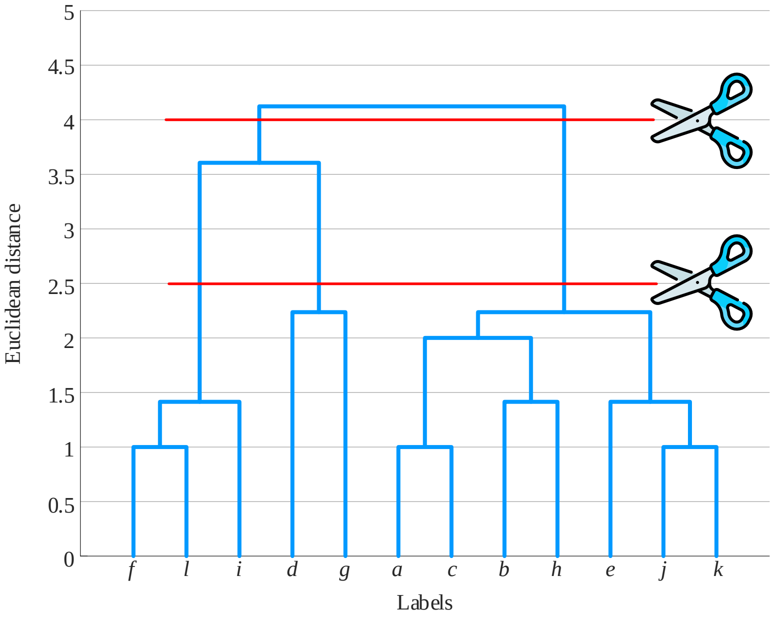


Figure 5. Dendrogram constructed from pairwise distances. Figure generated by Ch31_01_hierarchical_clustering.ipynb.

31.2.2 Step-by-Step Construction of the Tree

As shown in Figure 6, we begin by finding the closest pairs of individual points. In this dataset, the pairs (a, c), (k, j), and (f, l) each have distance 1, making them the earliest merges in the dendrogram. These locations appear as the darkest small values in the heatmap in Figure 8(a).

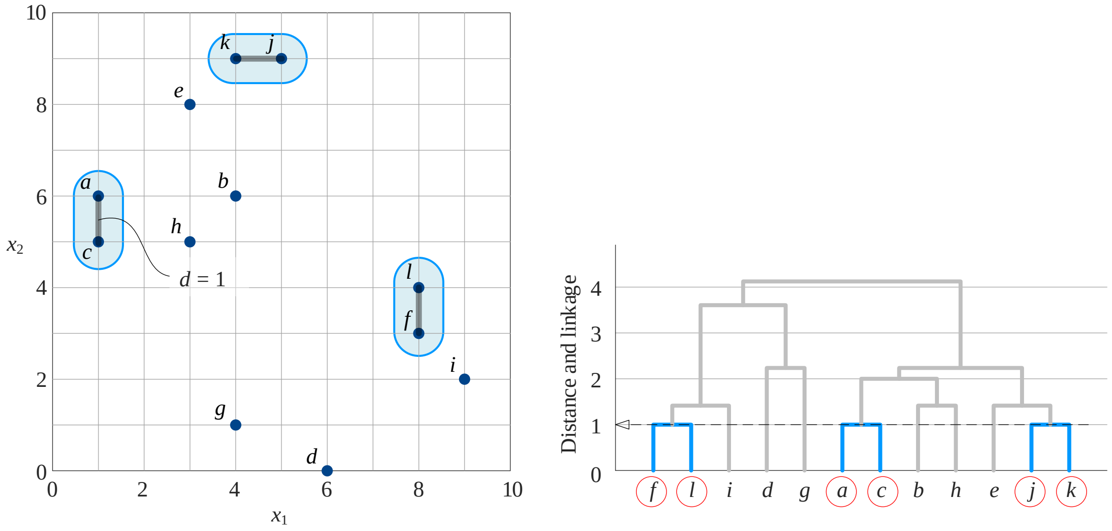


Figure 6. First merging step in the dendrogram

To build the second level, we now need the distance between clusters, not just between points. Here we use single linkage, which defines the distance between two clusters as the shortest distance between any pair of points across the two clusters.

From the first step, points (k, j) and (f, l) have already formed small clusters. We then observe that point e is closest to cluster (k, j), and point i is closest to cluster (f, l), each at a distance of square root of 2 (≈ 1.414). Meanwhile, points b and h also have the same distance. These observations produce the second layer of the dendrogram in Figure 7, and their positions in the heatmap are highlighted in Figure 8 (b).

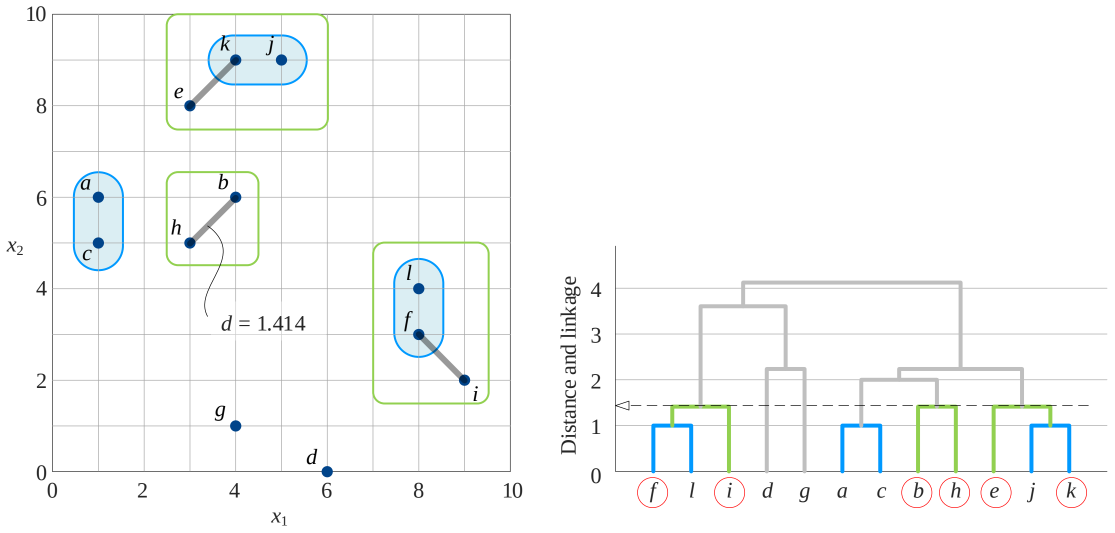


Figure 7. Second merging step in the dendrogram

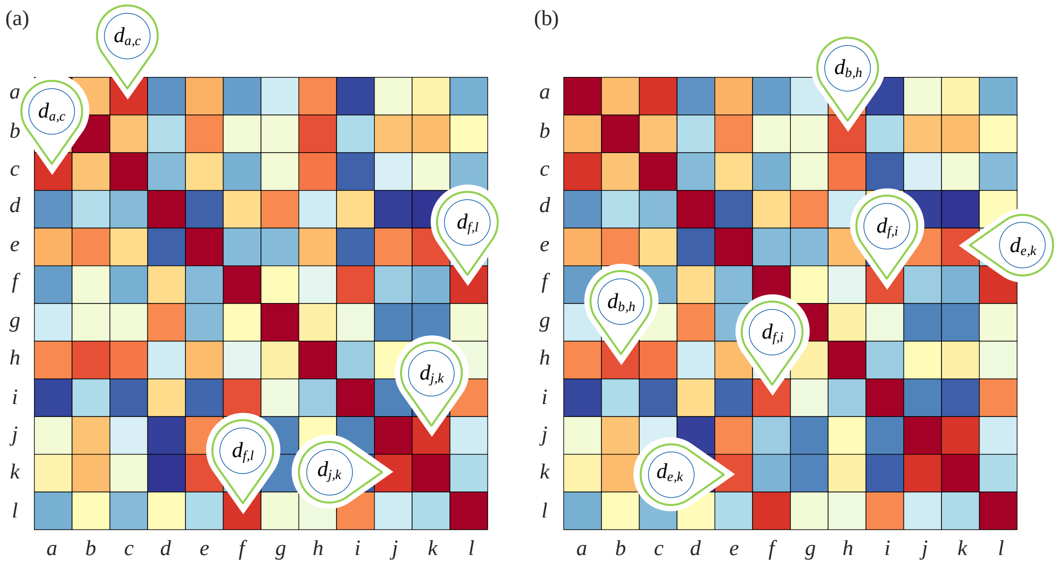


Figure 8. Heatmap locations of first- and second-level merges

Using the same rule, we continue merging clusters. The third-level merge combines earlier clusters based on the shortest distance between them, as illustrated in Figure 9, with its heatmap location shown in Figure 11(a).

The fourth level, shown in Figure 10, uses a distance of square root of 5 (≈ 2.226), which is the first point in the process where every data sample has joined one of three larger clusters. The corresponding heatmap positions are marked in Figure 11(b).

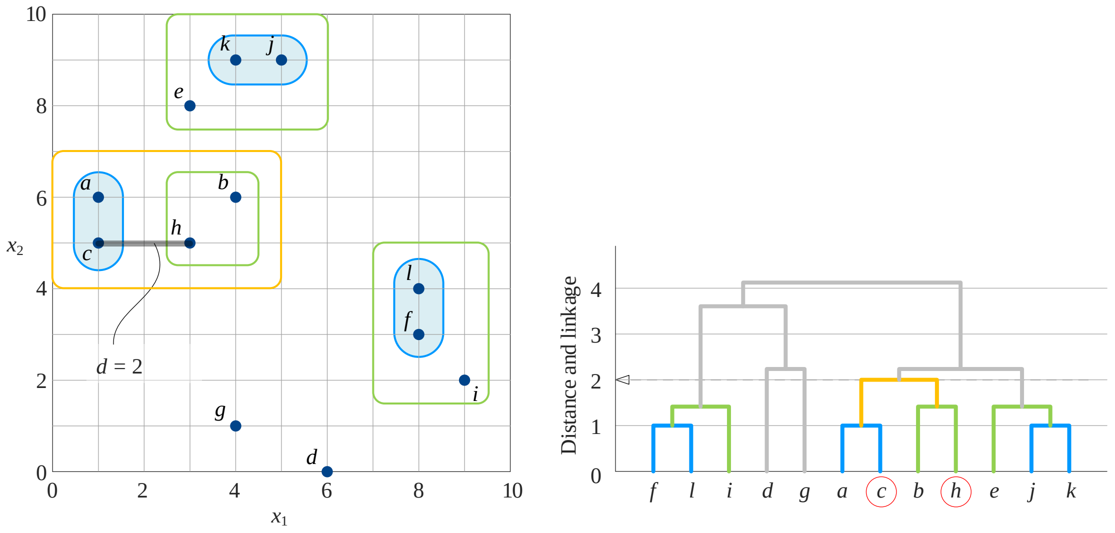


Figure 9. Third merging step

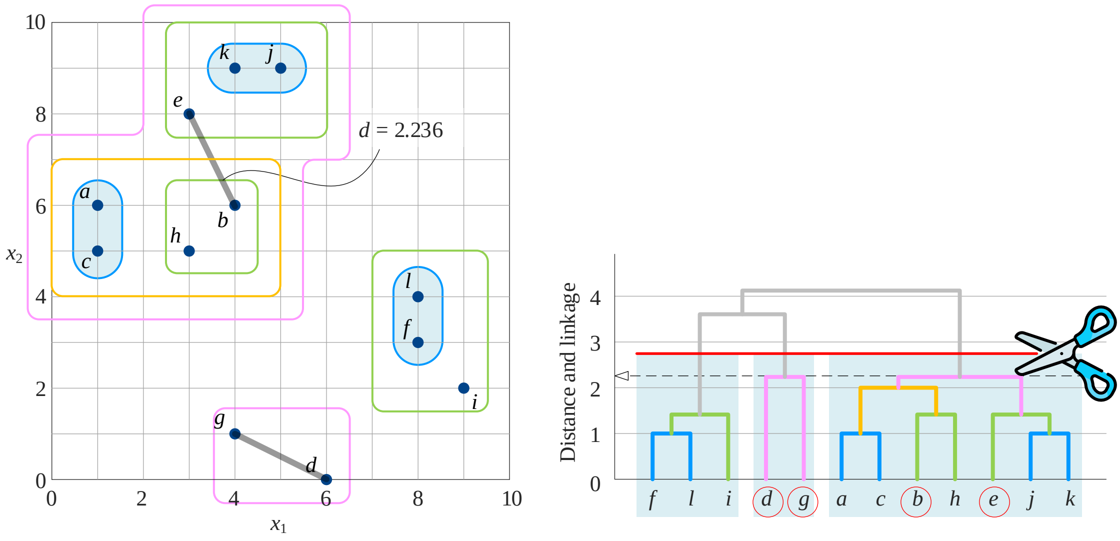


Figure 10. Fourth merging step

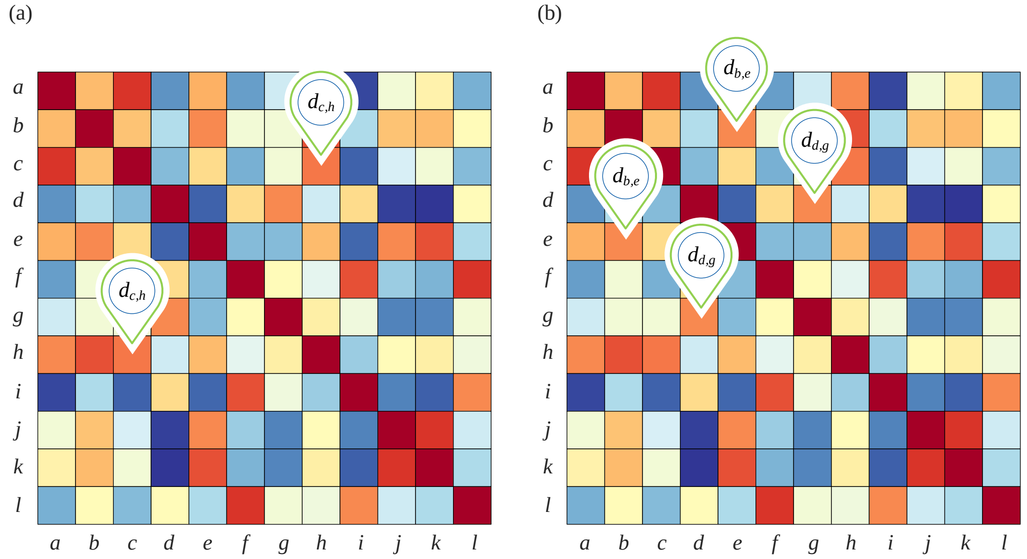


Figure 11. Heatmap locations for the third and fourth levels

At the fifth merge (Figure 12), the dendrogram shows two dominant clusters. One further merge would complete the tree and form a single cluster at the very top. After the tree is constructed, the sample order can be rearranged based on the dendrogram structure.

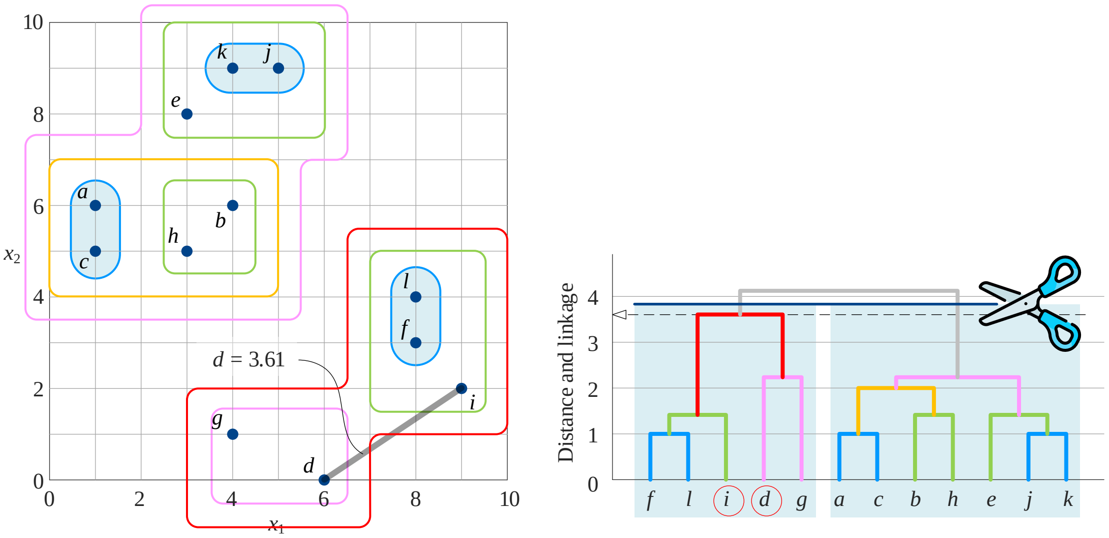


Figure 12. Fifth merging step in the dendrogram

Using this new order, a reordered distance heatmap is shown in Figure 13, where the block patterns clearly reveal which samples belong together. This visual separation is the essence of hierarchical clustering.

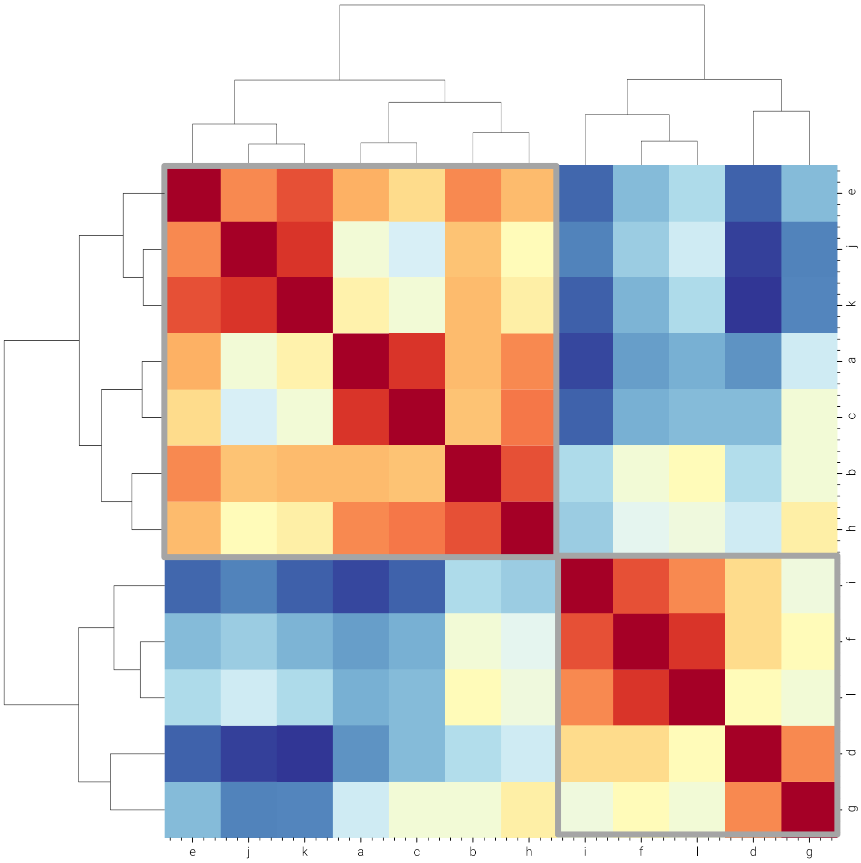


Figure 13. Reordered heatmap after clustering. Figure generated by Ch31_01_hierarchical_clustering.ipynb.

Figure 14 illustrates the result of hierarchical clustering applied to the dataset.

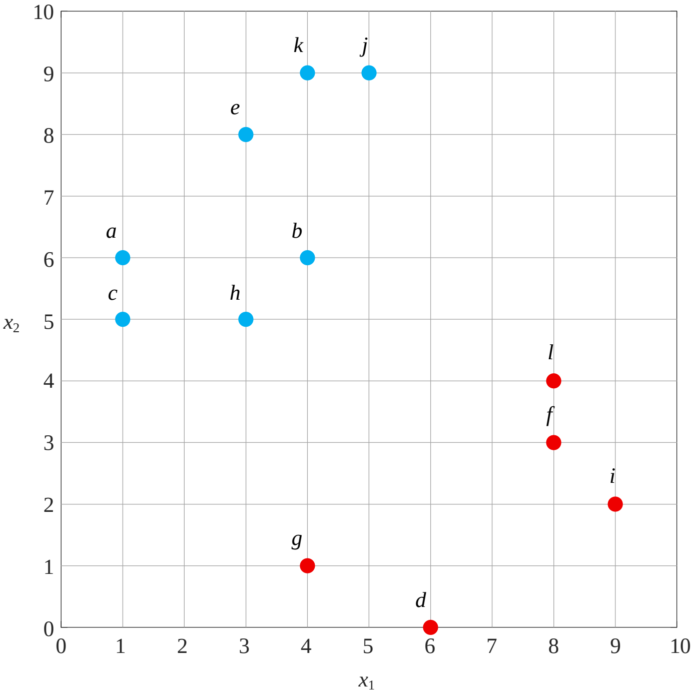


Figure 14. Data grouped into two clusters. Figure generated by Ch31_01_hierarchical_clustering.ipynb.

31.3 How Clusters "See" Each Other: Linkage Methods Explained

In hierarchical clustering, the result depends not only on how we measure the distance between individual data points, but also on how we define the distance between clusters. In the previous section, we used the "nearest point distance" (also called single linkage) as an example. In fact, there are several widely used ways to measure inter-cluster distance, each leading to different clustering behaviors and dendrogram shapes. The choice of linkage affects how clusters grow, how sensitive the algorithm is to noise, and whether the final dendrogram is balanced or skewed. Below, we introduce the most common linkage definitions.

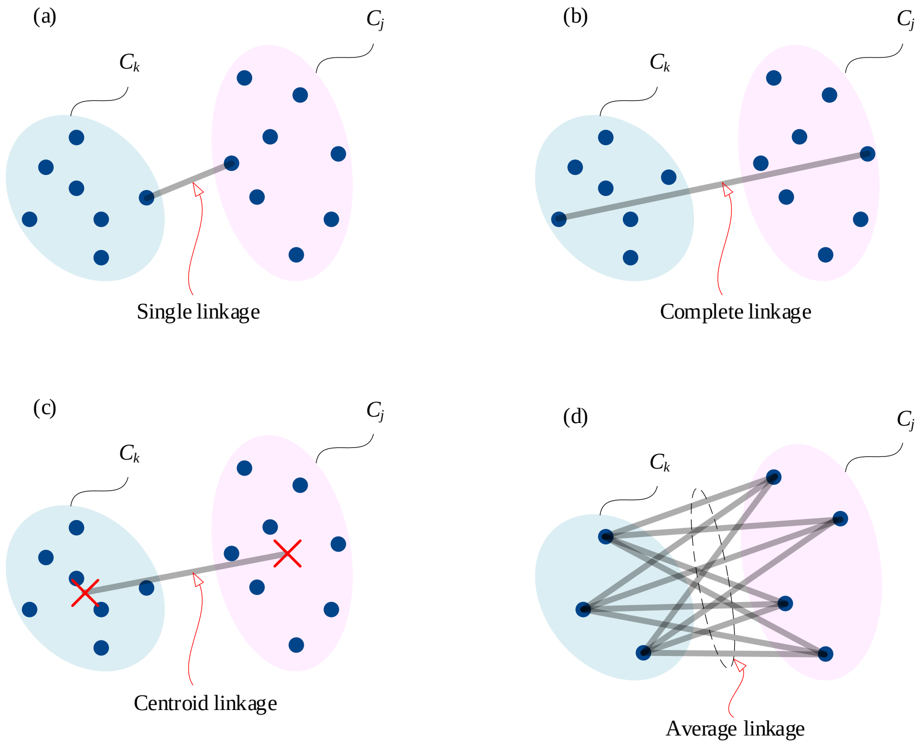


Figure 15. Distance between clusters defined

31.3.1 Single Linkage — The "Nearest Neighbor" Rule

As shown in Figure 16 (a), single linkage defines the distance between two clusters as the smallest distance between any pair of points across the two clusters:

$$d\left(C_k, C_j\right) = \min_{\boldsymbol{x}\in C_k,\ \boldsymbol{z}\in C_j} \left(\operatorname{dist}\left(\boldsymbol{x}, \boldsymbol{z}\right)\right) \tag{1}$$

This method tends to form long, chain-shaped clusters and is sensitive to points that create "bridges" between groups. Single linkage often produces uneven branches and less ideal clustering results.

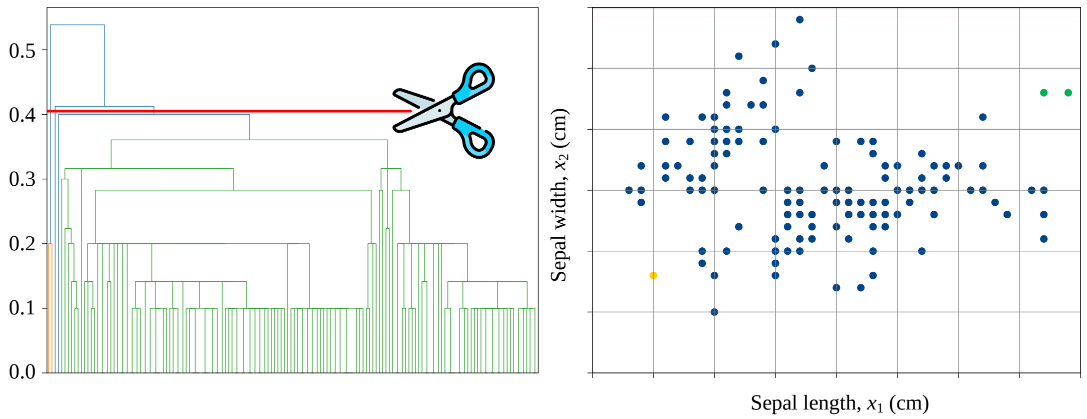


Figure 16. Hierarchical clustering on the Iris dataset using single linkage. Figures generated by Ch31_02_hierarchical_clustering.ipynb.

31.3.2 Complete Linkage — The "Farthest Neighbor" Rule

As shown in Figure 16 (b), complete linkage takes the largest pairwise distance between the two clusters:

$$d\left(C_k, C_j\right) = \max_{\boldsymbol{x}\in C_k,\ \boldsymbol{z}\in C_j} \left(\operatorname{dist}\left(\boldsymbol{x},\boldsymbol{z}\right)\right) \tag{2}$$

This approach encourages compact, evenly shaped clusters, but it can be overly sensitive to outliers—one noisy point can stretch the distance.

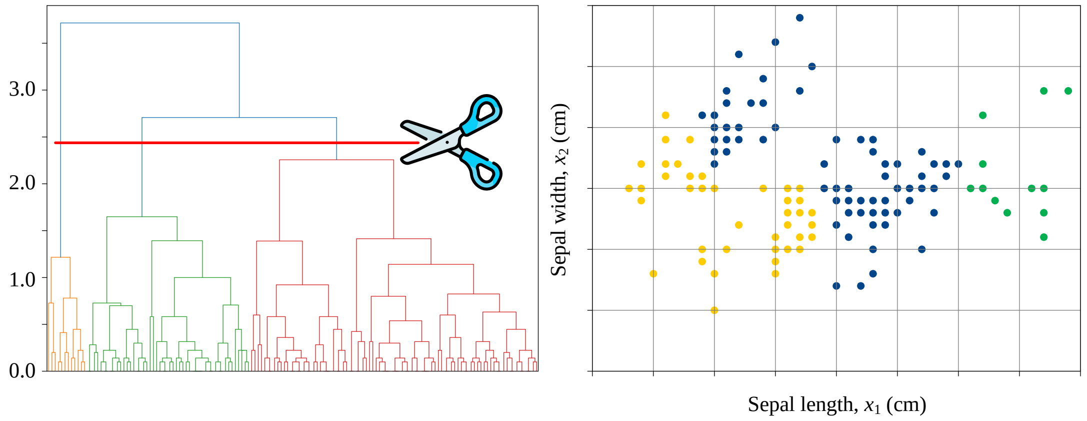


Figure 17. Hierarchical clustering on the Iris dataset using complete linkage. Figures generated by Ch31_02_hierarchical_clustering.ipynb.

31.3.3 Centroid Linkage — Meeting in the Middle

As shown in Figure 16 (c), centroid linkage measures the distance between the two cluster centroids:

$$d\left(C_i, C_j\right) = d\left(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j\right) \tag{3}$$

where $\boldsymbol{\mu}_i$ and $\boldsymbol{\mu}_j$ are the mean vectors of clusters C_i and C_j. This method is intuitive, but it may produce inversions in the dendrogram (a visual inconsistency where a merge appears at a smaller distance than earlier merges).

31.3.4 Average Linkage — Striking a Balance

As shown in Figure 16 (d), average linkage computes the mean of all pairwise distances between the two clusters:

$$d\left(C_k, C_j\right) = \underset{\boldsymbol{x}\in C_k,\ \boldsymbol{z}\in C_j}{\text{mean}} \left(\text{dist}\left(\boldsymbol{x}, \boldsymbol{z}\right)\right) = \frac{\sum_{\boldsymbol{x}\in C_k,\ \boldsymbol{z}\in C_j} \text{dist}\left(\boldsymbol{x}, \boldsymbol{z}\right)}{\text{count}\left(C_k\right)\cdot\text{count}\left(C_j\right)} \tag{4}$$

This method is a compromise between single and complete linkage, striking a balance between chaining and compactness.

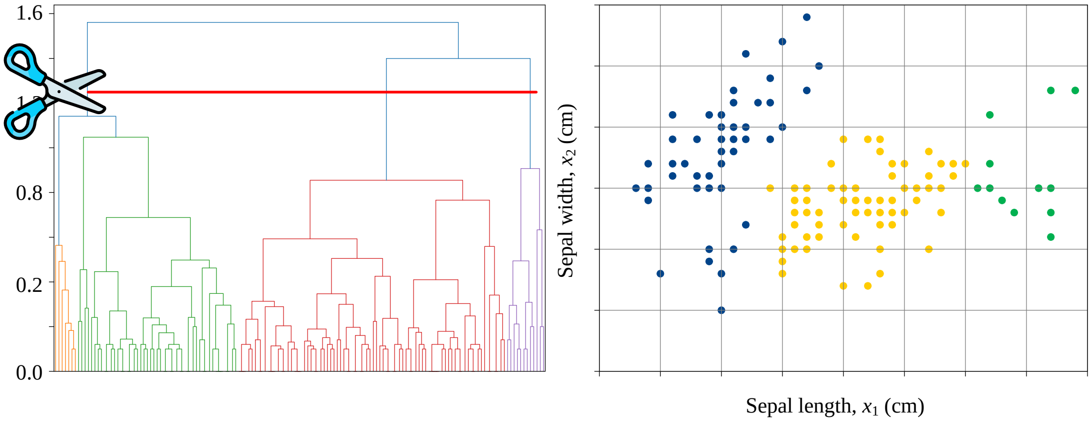


Figure 18. Hierarchical clustering on the Iris dataset using average linkage. Figures generated by Ch31_02_hierarchical_clustering.ipynb.

31.4 Conclusion

Hierarchical clustering builds a tree of nested groups, allowing us to understand data structure at multiple levels. In the agglomerative (bottom-up) approach, each point starts as its own cluster, and the algorithm repeatedly merges the closest clusters until only one remains. The result is visualized with a dendrogram, where the merge height represents the distance at which clusters join. By cutting the dendrogram at different levels, we can obtain different numbers of clusters.

To construct the tree, we first compute a pairwise distance matrix, often using Euclidean distance. Clusters are then merged step by step based on a chosen linkage rule, which defines how to measure the distance between clusters. Common linkage methods include single (nearest pair), complete (farthest pair), and average (mean pairwise distance). Different linkages produce very different dendrogram shapes and clustering results, ranging from chain-like structures to compact, well-separated groups.

Finally, hierarchical clustering is non-parametric and does not produce a fixed predictive model, but it offers a rich and intuitive view of cluster structure. It is especially useful for exploratory data analysis, where understanding relationships and hierarchy is more important than assigning labels to new samples.