In the bottom-up or agglomerative method, each data point is initially designated as a cluster. Clusters are iteratively combined based on homogeneity that is determined by some distance measure. On the other hand, the divisive or top-down approach starts with a cluster and subsequently splits into homogeneous sub-groups.

Hierarchical clustering creates a tree-like representation of the partitioning called a dendrogram. A dendrogram is drawn somewhat similar to a binary tree with the root at the top and the leaves at the bottom. The leaf on the dendrogram represents a data sample. The dendrogram is constructed by iteratively combining the leaves based on homogeneity to form clusters moving up the tree. An illustration of hierarchical clustering is shown in Figure 25-5.

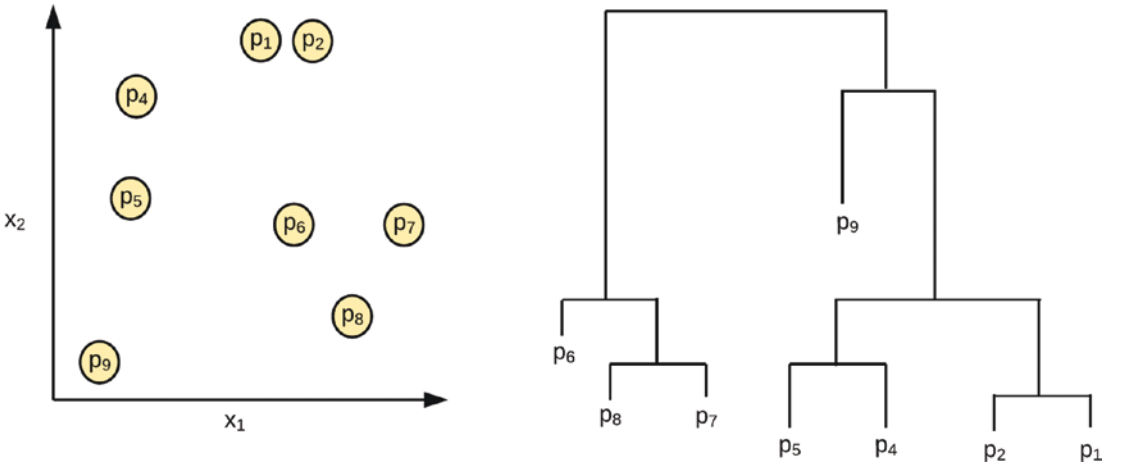

Figure 25-5. *An illustration of hierarchical clustering of data points in a 2-D feature space. Left: The spatial representation of points in 2-D space. Right: A hierarchical cluster of points represented by a dendrogram.*

How Are Clusters Formed

Clusters are formed by computing the nearness between each pair of data points. The notion of nearness is most popularly calculated using the Euclidean distance measure. Beginning at the leaves of the dendrogram, we iteratively combine those data points that are closer to one another in the multi-dimensional vector space until all the homogeneous points are placed into a single group or cluster.

The Euclidean distance is used to compute the nearness between *n* data points. After each pair of data points has combined to form a cluster, the new cluster pairs are then pulled into groups going up the tree, with the tree branch or dendrogram height reflecting the dissimilarity between the clusters.

Dissimilarity computes how different each cluster of data is from one another. The notion of dissimilarity between two clusters or groups is described in terms of *linkage*. Four types of linkage exist for grouping clusters in hierarchical clustering. They are centroid, complete, average, and single.

The centroid linkage computes the dissimilarity between two clusters using the geometric centroid of the clusters. The complete linkage uses the two farthest data points between the two clusters to compute the dissimilarity (see Figure 25-6).

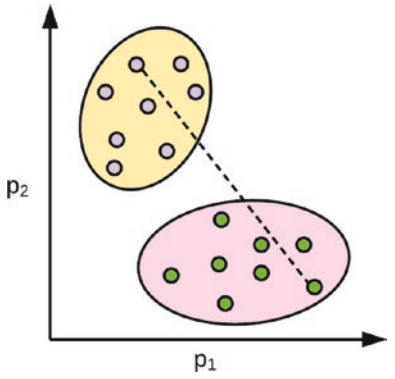

Figure 25-6. *Complete linkage*

The average linkage finds the means of points within the pair of clusters and uses that new artificial point to calculate the dissimilarity (see Figure 25-7).

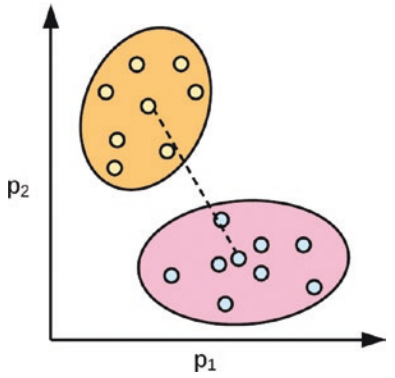

Figure 25-7. *Average linkage*

The single linkage uses the closest data point between the cluster pairs to compute the dissimilarity measure (see Figure 25-8).

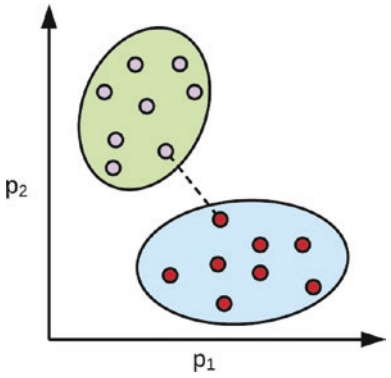

Figure 25-8. *Single linkage*

Empirically, the complete and average linkages are preferred in practice because they yield more balanced dendrograms. Other dissimilarity measures exist for evaluating the nearness or homogeneity of data points. One of such is the Manhattan distance, another distance-based measure, or the correlation-based distance which groups pairs of data samples with highly correlated features. A correlated-based dissimilarity measure may be more useful in datasets where proximity in multi-dimensional spaces is not as useful a metric for homogeneity as compared to the correlation of their features in the space. A choice of calculating dissimilarity has a significant impact on the ensuring dendrogram.

After running the algorithm, the dendrogram is cut at a particular height, and the number of distinct lines or branches after the cut is circumscribed as the number of clusters in the dataset. An illustration of cutting the dendrogram is shown in Figure 25-9.

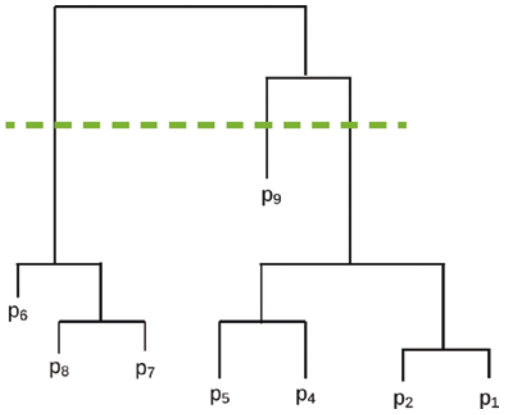

Figure 25-9. *Dendrogram cut*

Hierarchical Clustering with the SciPy Package

This example implements hierarchical or agglomerative clustering with SciPy. The 'scipy.cluster.hierarchy' package has simple methods for performing hierarchical clustering and plotting dendrograms. This example uses the 'complete' linkage method. The plot of the dendrogram is shown in Figure 25-10.

```
# import packages
import matplotlib.pyplot as plt
from scipy.cluster.hierarchy import dendrogram
from scipy.cluster import hierarchy

Z = hierarchy.linkage(X, method='complete')

plt.figure()
dn = hierarchy.dendrogram(Z, truncate_mode='lastp')
```