Plan of Attack

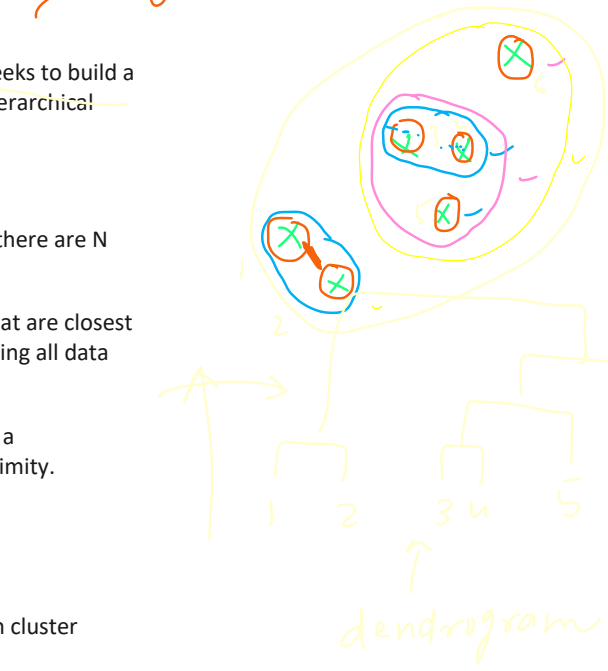
22 December 2023 11:41

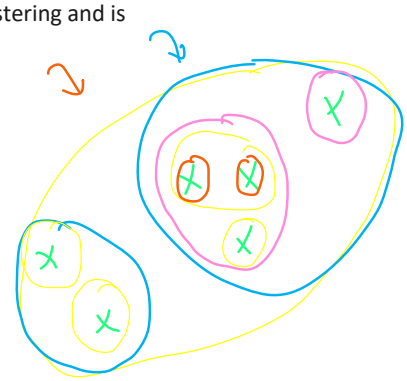
Hierarchical Clustering

22 December 2023 11:41

Hierarchical clustering is a method of cluster analysis used in data mining. It seeks to build a hierarchy of clusters in a step-by-step manner. There are two main types of hierarchical clustering:

1. Agglomerative (Bottom-Up Approach):
   - Initial Step: Starts by treating each data point as a separate cluster. So, if there are N data points, you begin with N clusters.
   - Clustering Process: In each step, the algorithm merges the two clusters that are closest to each other until all the clusters are merged into one big cluster containing all data points.
   - Dendrogram: The result can be represented in a tree-like structure called a dendrogram, which shows the arrangement of the clusters and their proximity.
2. Divisive (Top-Down Approach):
   - Initial Step: Begins with all data points in a single cluster.
   - Clustering Process: At each step, the algorithm splits the cluster until each cluster contains only one data point.
   - Top-Down Splitting: This is less common compared to agglomerative clustering and is computationally more intensive.

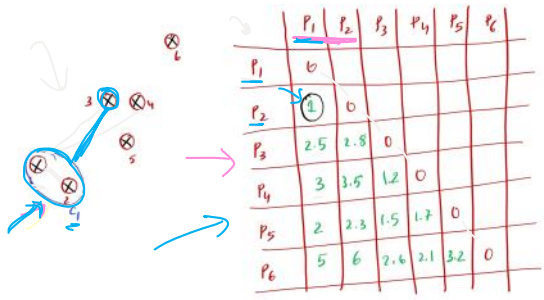


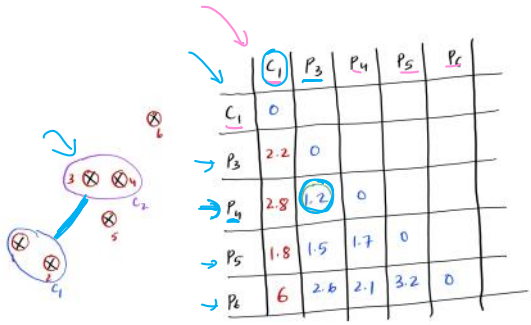


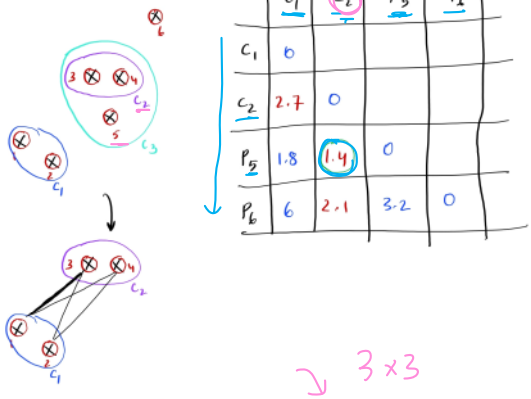
Algorithm

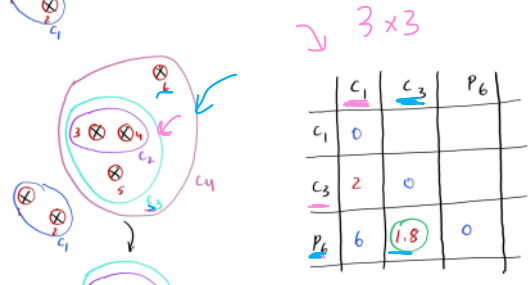
22 December 2023 08:57

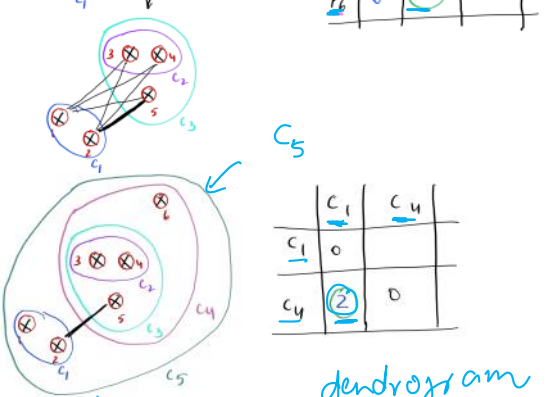
1. Initialization:
   - Treat each data point as a separate cluster. Thus, if you have N data points, you start with N clusters, each containing just one data point.
2. Compute Distance Matrix:
   - Calculate the distance between each pair of clusters. Common distance metrics include Euclidean, Manhattan, and Cosine distances. The choice of distance metric can significantly affect the outcome of the clustering.
   - This results in an N×N distance matrix, where the distance between a cluster and itself is zero.
3. Find the Closest Clusters:
   - Identify the two clusters that are closest to each other based on the distance matrix.
4. Merge Clusters:
   - Combine the two closest clusters into a single cluster.
   - This step reduces the total number of clusters by one.
5. Update Distance Matrix:
   - Recalculate the distances between the new cluster and all the existing clusters.
   - The method of recalculating the distance depends on the linkage criterion used. Common linkage criteria include:
     - Single Linkage: Distance between two clusters is defined as the shortest distance between any two points in the clusters.
     - Complete Linkage: Distance is the longest distance between any two points in the clusters.
     - Average Linkage: Distance is the average distance between all pairs of points in the clusters.
     - Ward's Method: Distance is calculated as the increase in the total within-cluster variance after merging the clusters.
6. Repeat:
   - Repeat steps 3 to 5 until all data points are merged into a single cluster.

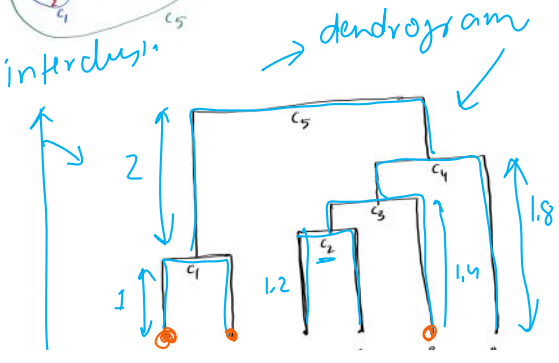




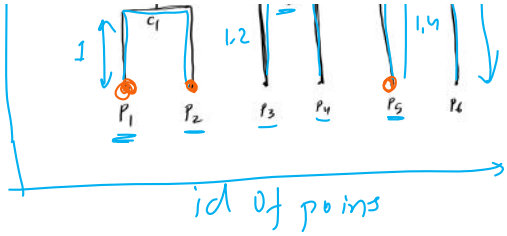


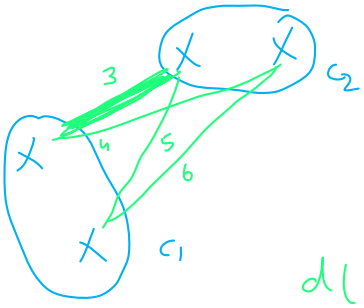






Session 5 Agglomerative Clustering Page 3





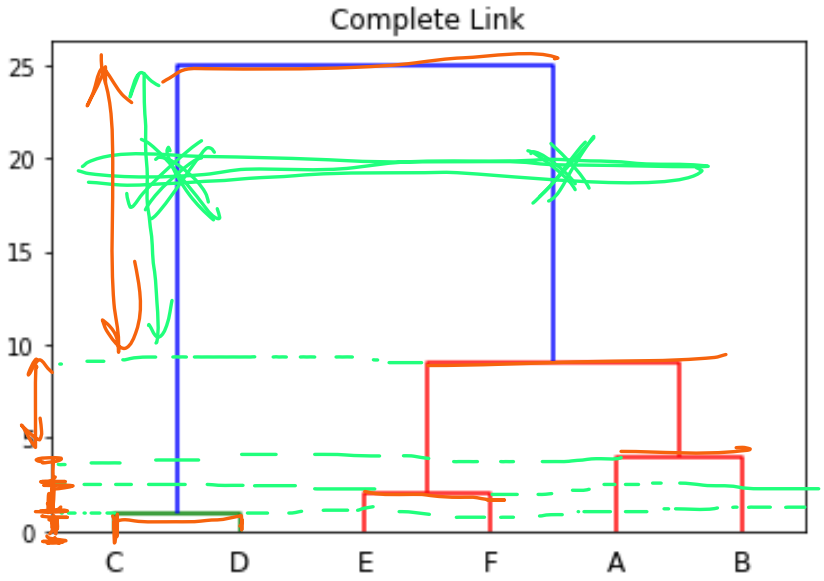
linkage
↓
min

$d(c_1, c_2) = 3$

Finding n_clusters

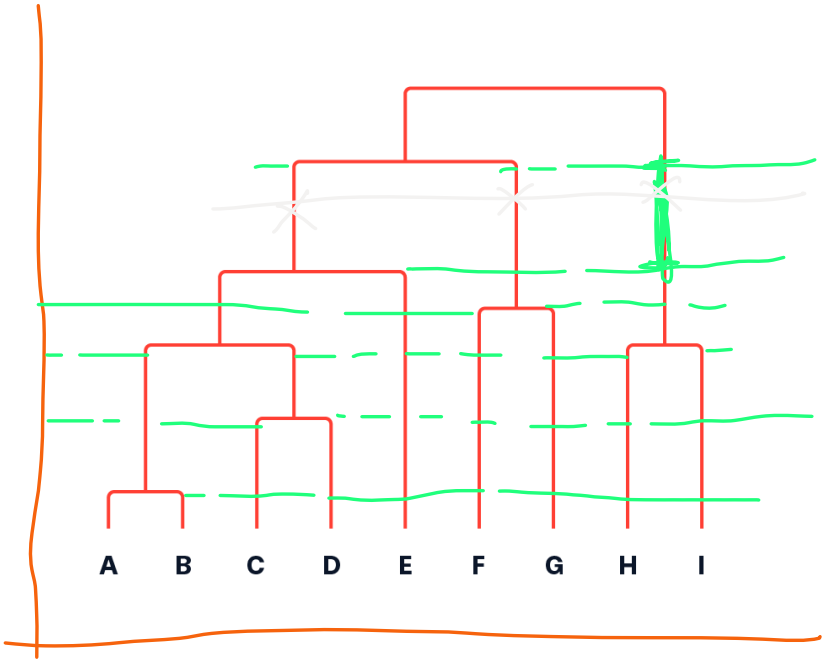
22 December 2023 08:57

dendrogram.

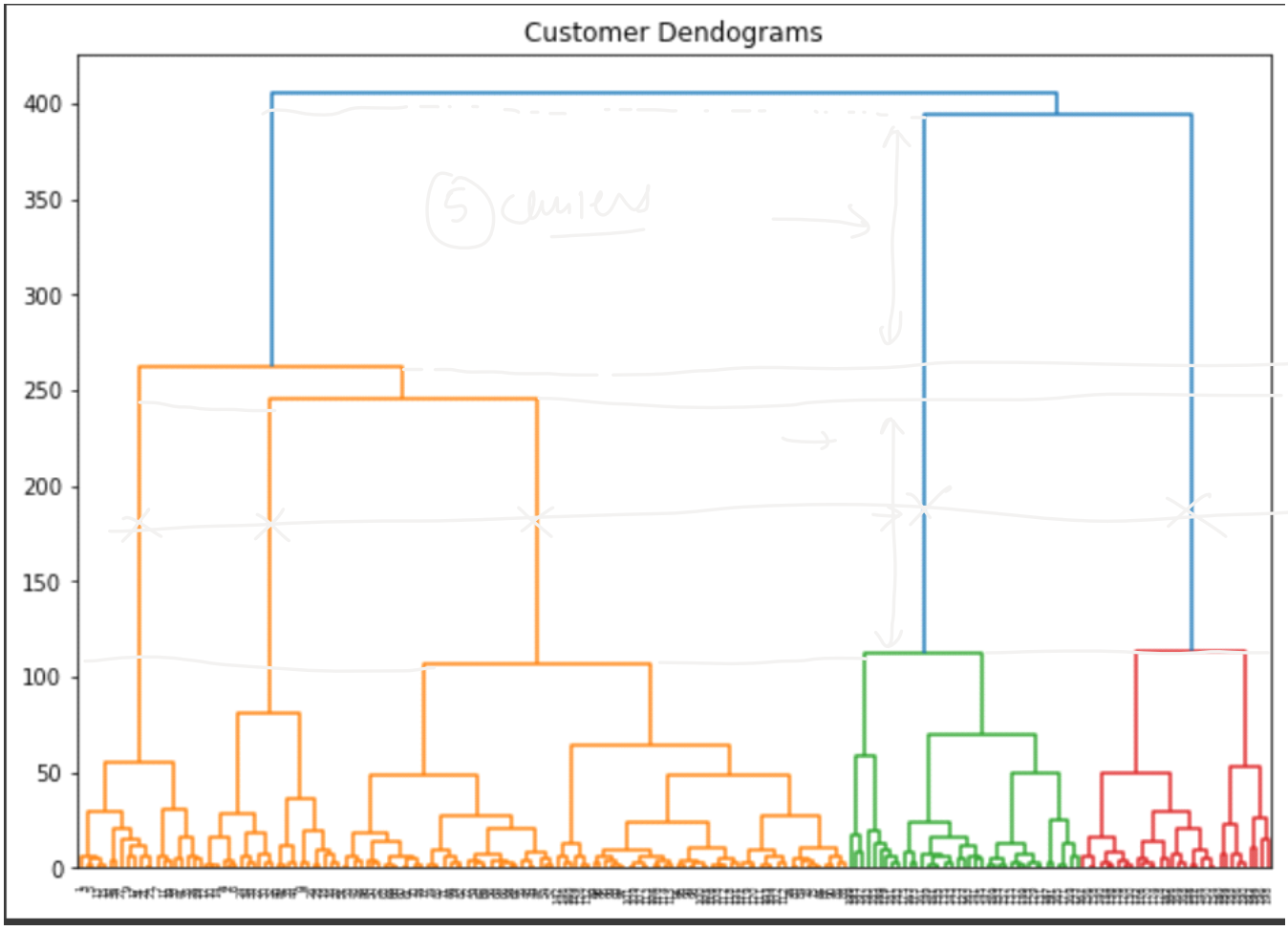


→ ② clusters

inter cluster



cluster - 3


Customer Dendograms
400
350
300
250
200
150
100
50
0

Python Code

22 December 2023 08:58

Linkage

22 December 2023 11:32

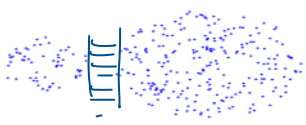
In hierarchical clustering, linkage is the criterion that determines the distance between sets of observations as a function of the pairwise distances between observations. It's essentially the algorithm used to decide the proximity of clusters. There are several linkage methods, each defining the distance between clusters differently:

1. Single Linkage (Nearest Point Algorithm):
   - The distance between two clusters is defined as the shortest distance from any member of one cluster to any member of the other cluster.
   - Capable of detecting non-elliptical shapes in the data.
   - Works well for datasets where the clusters are well-separated.
   - May not perform well when clusters are close together or overlap as it is sensitive to outliers.

2. Complete Linkage (Farthest Point Algorithm):
   - The distance between two clusters is defined as the longest distance from any member of one cluster to any member of the other cluster.
   - Less susceptible to noise and outliers compared to single linkage.
   - Can struggle with elongated clusters or non-convex shapes.

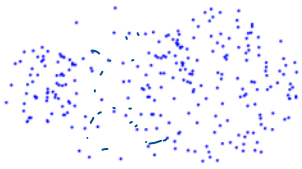
3. Average Linkage (Average Link Algorithm):
   - The distance between two clusters is defined as the average distance between each member of one cluster to every member of the other cluster.

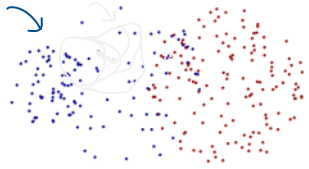
4. Ward's Method:
   - Objective: The main goal of Ward's method is to find the pair of clusters that, when merged, will increase the total within-cluster variance as little as possible. This is like trying to keep the clusters as compact as possible.
   - Within-Cluster Variance: This is a measure of how spread out the points are within a cluster. A lower within-cluster variance means the points are closer to each other, and therefore, the cluster is more compact.
   - How it Works: At each step of the algorithm, Ward's method looks at all possible pairs of clusters and calculates how much the within-cluster variance would increase if those two clusters were merged. It then merges the two clusters that result in the smallest increase in variance.
   - Resulting Clusters: Because Ward's method tries to keep the within-cluster variance low, it tends to create clusters that are compact and roughly spherical in shape. This can be particularly effective if the natural groups in your data are also compact and spherical.

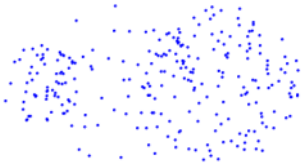


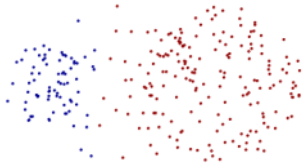
Original Points

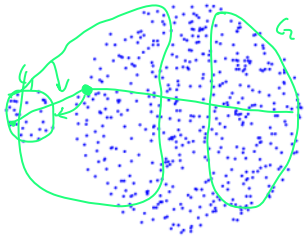
Two Clusters

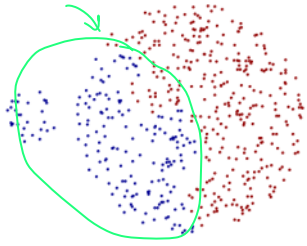


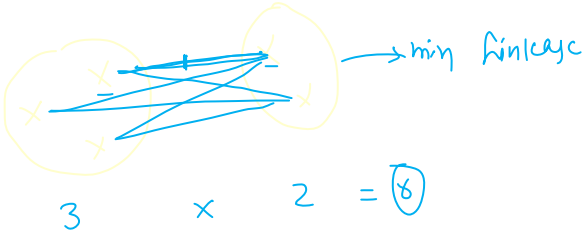


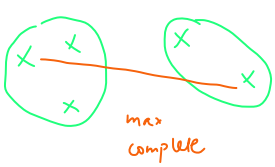


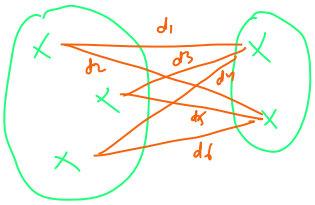












$$\frac{d_1 + d_2 + \ldots + d_6}{6}$$

↓ proximity

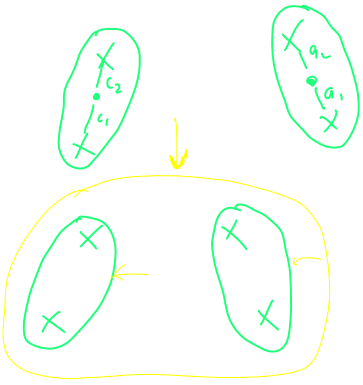
↓ sim → dist

ward min

min

$$\left[d_1^2 + d_2^2 + d_3^2 + d_4^2 - c_1^2 - c_2^2 - a_1^2 - a_2^2 \right]$$

distance



$[WCSS - WCSS_1 - WCSS_2]$

↓ gain in variance

Time Complexity

22 December 2023 08:59

$n \times n$ small 6×6 6×5 $n \times (n-1)$

Space complexity big dataset medium

$n^2 \rightarrow$ millions

$10^6 \rightarrow 10^6 \times 10^6$

8 GB / 16 GB

10^2

GBs RAMs

Time complexity

Space complexity → $O(n^2)$

Time complexity → $O(n \log n)$ to $O(n^3)$

↓

drawbacks

Advantages and Disadvantages

22 December 2023 08:58

Advantages

1. **Discovery of Hierarchical Structure**: The algorithm reveals the hierarchy and nested structure within the data, which can be informative for understanding complex relationships.
2. **Useful for Any Distance Measure**: The method can be used with any distance measure, which is beneficial for different types of data, such as genomic data or mixed data types.
3. **Does Not Assume Clusters as Spherical**: Unlike K-means, agglomerative clustering does not assume that clusters are spherical in shape, which can result in more natural cluster shapes.
4. **Easy to Implement and Understand**: The algorithm is conceptually simple and can be easily implemented, making it accessible for users with varying levels of expertise.
5. **Robust to Noise and Outliers**: With the appropriate choice of linkage criteria (such as Ward's method), hierarchical clustering can be relatively robust to noise and outliers, as these will typically be merged into clusters at later stages of the process.

Disadvantages

1. Computational Complexity: One of the biggest drawbacks is its computational cost. The algorithm has a time complexity of O(n^3) and space complexity of O(n^2) for the simplest implementations, making it impractical for large datasets.
2. Sensitivity to Noise and Outliers: Certain linkage criteria, such as single linkage, can be highly sensitive to noise and outliers, which can lead to misleading results. Outliers can cause clusters to merge prematurely, distorting the true structure of the data.
3. Difficulty in Identifying the Number of Clusters: While the dendrogram can provide insights into the potential number of clusters, there is often subjectivity involved in interpreting where to 'cut' the dendrogram to define the clusters.
4. Arbitrary Decisions in Linkage Criteria: The choice of linkage criteria (single, complete, average, Ward's, etc.) can significantly affect the results, and there is no definitive rule for choosing the best method, which can make the process somewhat arbitrary.
5. No Global Objective Function: Unlike K-means, which minimizes within-cluster variance, there's no clear global objective in hierarchical clustering, which can make it difficult to assess the quality of the resulting clusters.

Applications

22 December 2023 08:59