Hierarchical Agglomerative Clustering (HAC) vs. K-Means Clustering

Your Name

March 11, 2025

Introduction

In this document, we will discuss the differences between **Hierarchical Agglomerative Clustering (HAC)** and **K-Means Clustering**. We will also walk through an example of HAC using the **single linkage** method and demonstrate how to compute the Euclidean distance matrix, merge clusters, and draw dendrograms.

Differences Between HAC and K-Means

- **K-Means Clustering**:
  - Starts with predefined centroids.
  - Assigns data points to the nearest centroid.
  - Updates centroids iteratively until convergence.
- **Hierarchical Agglomerative Clustering (HAC)**:
  - Starts with each data point as its own cluster.
  - Iteratively merges the closest pair of clusters.
  - Continues until all points belong to a single cluster.

Example: Single Linkage Clustering

Let's walk through an example of HAC using the **single linkage** method. We are given six points with their x and y coordinates. The steps are as follows:

Step 1: Compute the Euclidean Distance Matrix

The Euclidean distance between two points $(x1, y1)$ and $(x2, y2)$ is:

$$\text{Distance} = \sqrt{(x1 - x2)^2 + (y1 - y2)^2}.$$

Compute the distance between all pairs of points to create a distance matrix. The diagonal elements are 0 because the distance between a point and itself is 0.

Step 2: Find the Minimum Distance

Identify the smallest distance in the matrix. For example, suppose the smallest distance is 0.11 between $P3$ and $P6$. These two points form the first cluster.

Step 3: Merge Clusters and Update the Distance Matrix

Merge $P3$ and $P6$ into a single cluster. Update the distance matrix using the single linkage method:

$$\text{Distance}(C1, C2) = \min(\text{dist}(x, y)) \quad \text{for all } x \in C1, y \in C2.$$

For example:

- The distance between $P1$ and the new cluster $P36$ is the minimum of $\text{dist}(P1, P3)$ and $\text{dist}(P1, P6)$.
- Repeat this for all points to update the distance matrix.

Step 4: Repeat the Process

Continue finding the minimum distance in the updated matrix, merging the closest clusters, and updating the distance matrix until all points belong to a single cluster.

Step 5: Draw the Dendrogram

A dendrogram is a tree-like diagram that represents the hierarchical clustering process. The steps to draw it are:

- Start with the first merged cluster (e.g., $P3$ and $P6$) and plot it at height 0.11.
- Add the next merged cluster (e.g., $P2$ and $P5$) at height 0.14.
- Continue adding clusters at their respective heights until all points are merged.

Example Walkthrough

Let's go through the steps with an example:

Initial Distance Matrix

	$P1$	$P2$	$P3$	$P4$	$P5$	$P6$
$P1$	0	0.23	0.22	0.37	0.34	0.23
$P2$	0.23	0	0.15	0.20	0.14	0.25
$P3$	0.22	0.15	0	0.15	0.28	0.11
$P4$	0.37	0.20	0.15	0	0.29	0.22
$P5$	0.34	0.14	0.28	0.29	0	0.39
$P6$	0.23	0.25	0.11	0.22	0.39	0

Table 1: Initial Euclidean Distance Matrix

Step 1: Find the Minimum Distance

The smallest distance is 0.11 between $P3$ and $P6$. Merge $P3$ and $P6$ into a single cluster.

Step 2: Update the Distance Matrix

Update the distance matrix using the single linkage method. For example:

$$\text{Distance}(P1, P36) = \min(0.22, 0.23) = 0.22.$$

Repeat this for all points to get the updated matrix.

Step 3: Repeat Until All Points Are Merged

Continue finding the minimum distance, merging clusters, and updating the matrix until all points belong to a single cluster.

Step 4: Draw the Dendrogram

- Merge $P3$ and $P6$ at height 0.11.
- Merge $P2$ and $P5$ at height 0.14.
- Merge $P36$ and $P25$ at height 0.15.
- Continue until all points are merged.

Conclusion

- **HAC** is a bottom-up clustering method that merges the closest clusters iteratively.
- **Single linkage** uses the minimum distance between clusters.
- Dendrograms provide a visual representation of the clustering process.