Data Mining Machine Learning

Md. Abdullah-Al-Kafi
Lecturer
Daffodil International University

Problem Statement

We are clustering food items based on their sweetness and calorie content using Hierarchical Agglomerative Clustering (HAC) with average linkage.

Data

The data for each food item is as follows:

Food	Sweetness	Calories
Apple	7	52
Banana	9	89
Carrot	3	41
Donut	10	400
Eggplant	2	25

Step 1: Calculate Pairwise Euclidean Distances

The Euclidean distance formula between two points (x_1, y_1) and (x_2, y_2) is:

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

We calculate the pairwise distances between all food items as follows:

Pair	Distance
Apple, Banana	$\sqrt{(9-7)^2 + (89-52)^2} = \sqrt{1373} \approx 37.04$
Apple, Carrot	$\sqrt{(3-7)^2 + (41-52)^2} = \sqrt{137} \approx 11.70$
Apple, Donut	$\sqrt{(10-7)^2 + (400-52)^2} = \sqrt{121113} \approx 348.01$
Apple, Eggplant	$\sqrt{(2-7)^2 + (25-52)^2} = \sqrt{754} \approx 27.46$
Banana, Carrot	$\sqrt{(3-9)^2 + (41-89)^2} = \sqrt{2340} \approx 48.37$
Banana, Donut	$\sqrt{(10-9)^2 + (400-89)^2} = \sqrt{96722} \approx 311.07$
Banana, Eggplant	$\sqrt{(2-9)^2 + (25-89)^2} = \sqrt{4145} \approx 64.38$
Carrot, Donut	$\sqrt{(10-3)^2 + (400-41)^2} = \sqrt{128930} \approx 359.01$
Carrot, Eggplant	$\sqrt{(2-3)^2 + (25-41)^2} = \sqrt{257} \approx 16.03$
Donut, Eggplant	$\sqrt{(2-10)^2 + (25-400)^2} = \sqrt{140689} \approx 375.25$

The resulting distance matrix is:

	Apple	Banana	Carrot	Donut	Eggplant
Apple	0	37.04	11.70	348.01	27.46
Banana	37.04	0	48.37	311.07	64.38
Carrot	11.70	48.37	0	359.01	16.03
Donut	348.01	311.07	359.01	0	375.25
Eggplant	27.46	64.38	16.03	375.25	0

Step 2: Clustering Steps

1. **First Merge**: The smallest distance is 11.70 between **Apple** and **Carrot**. Merge these to form **Cluster A**.

2. **Update Distances for Cluster A**:

$$\begin{aligned}\text{Distance from Cluster A to Banana} &= \frac{37.04 + 48.37}{2} = 42.71\\ \text{Distance from Cluster A to Donut} &= \frac{348.01 + 359.01}{2} = 353.51\\ \text{Distance from Cluster A to Eggplant} &= \frac{27.46 + 16.03}{2} = 21.75\end{aligned}$$

3. **Second Merge**: The smallest distance now is 21.75 between **Cluster A** and **Eggplant**. Merge **Eggplant** into **Cluster A** to form **Cluster B** (Apple, Carrot, Eggplant).

4. **Update Distances for Cluster B**:

$$\begin{aligned}\text{Distance from Cluster B to Banana} &= \frac{37.04 + 48.37 + 64.38}{3} = 49.93\\ \text{Distance from Cluster B to Donut} &= \frac{348.01 + 359.01 + 375.25}{3} = 360.09\end{aligned}$$

5. **Third Merge**: The smallest remaining distance is 49.93 between **Cluster B** and **Banana**. Merge **Banana** into **Cluster B**.

6. **Final Merge**: The only remaining items are **Cluster B** and **Donut**, with a distance of 360.09. Merge these to form the final cluster.

Conclusion

The hierarchical clustering process results in the following clusters: 1. Cluster A: Apple and Carrot 2. Cluster B: Apple, Carrot, Eggplant 3. Further merging with Banana and Donut completes the clustering hierarchy.

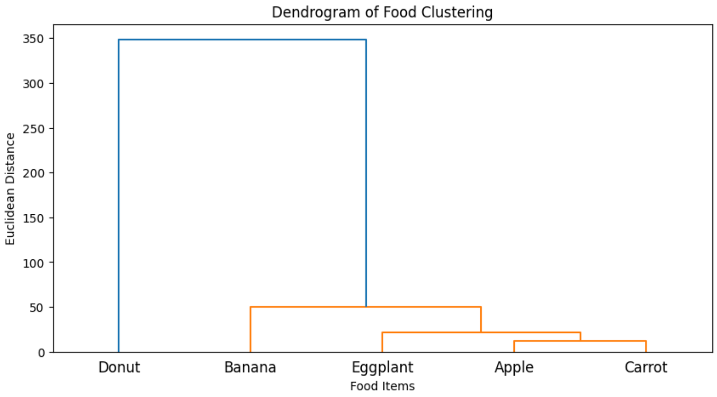

Figure 1: H-Clustering

Fundamental Questions

1. **What is Hierarchical Agglomerative Clustering (HAC)?**
   - Explain the basics of HAC and how it differs from other clustering methods like k-means.

2. **How does the Agglomerative approach work in HAC?**

   - Describe the bottom-up approach of HAC, where each data point starts as its own cluster.

3. **What is a dendrogram, and how is it used in HAC?**
   - Discuss how a dendrogram visually represents the merging of clusters and the hierarchical structure.

4. **What are some common applications of HAC?**
   - Explore practical use cases for HAC in fields like biology, marketing, and document clustering.

Technical Questions

1. **What are the key steps in Hierarchical Agglomerative Clustering?**
   - Outline the major steps, including distance calculation, merging of clusters, and recalculating distances.

2. **How do we calculate distances between clusters in HAC?**
   - Describe distance metrics such as Euclidean distance and different linkage methods (single, complete, average).

3. **What are the differences between single, complete, and average linkage?**
   - Explain each linkage method and how they influence the shape and structure of clusters in HAC.

4. **How is the cut-off threshold determined in a dendrogram?**
   - Discuss the criteria for ”cutting” a dendrogram to form distinct clusters.

Advanced Questions

1. **How does the choice of distance metric affect HAC results?**
   - Analyze how different metrics (e.g., Manhattan, Cosine, or Euclidean distance) impact the clustering outcome.

2. **What are some limitations and challenges of HAC?**
   - Describe limitations such as scalability and sensitivity to noise, and how they affect clustering quality.

3. **How does HAC perform with high-dimensional data?**
   - Examine challenges HAC faces with high-dimensional data and possible dimensionality reduction techniques.

4. **What is the computational complexity of HAC?**
   - Explore the time complexity of HAC and how it impacts clustering large datasets.

Practical and Interpretation Questions

1. **How do we interpret clusters generated by HAC?**
   - Discuss interpreting clusters, what the dendrogram tells us about data relationships, and validating results.

2. **In what scenarios would HAC be preferable to k-means clustering?**
   - Compare scenarios in which HAC's hierarchical approach may outperform partition-based methods like k-means.

3. **How can we use HAC in a practical data science project?**
    - Outline steps to use HAC in a project, including data preprocessing, selecting parameters, and interpreting results.
4. **What are some techniques to optimize HAC for large datasets?**
    - Discuss strategies like dimensionality reduction or sampling to make HAC more scalable for large datasets.