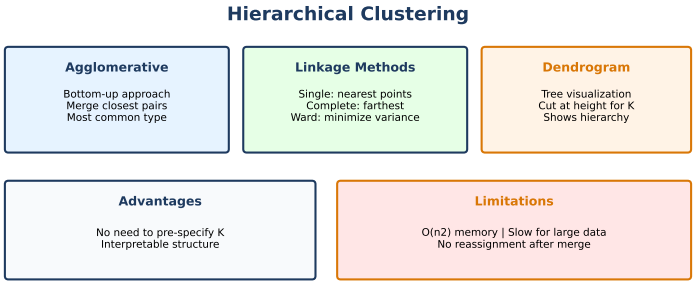
Lesson 30 Summary: Hierarchical Clustering

Data Science with Python – Key Concepts

Data Science Program

Hierarchical Clustering Overview



dendrogram(linkage(X, method="ward")) | AgglomerativeClustering(n_clusters=K)

Hierarchical clustering builds a tree of clusters

Agglomerative Approach

Bottom-up clustering:

- **Start:** Each point is its own cluster
- **Merge:** Combine two closest clusters
- **Repeat:** Until single cluster or threshold

Agglomerative is the most common hierarchical method

Linkage Methods

How to measure cluster distance:

- **Single:** Minimum distance between points
- **Complete:** Maximum distance between points
- **Average:** Mean distance between all pairs
- **Ward:** Minimizes within-cluster variance

Ward linkage usually gives the most balanced clusters

The Dendrogram

Tree visualization:

- **X-axis:** Data points or cluster labels
- **Y-axis:** Distance at which clusters merge
- **Cut:** Horizontal line determines K clusters

Dendrograms reveal hierarchical structure

Choosing Number of Clusters

Methods:

- **Dendrogram:** Cut where gaps are large
- **Silhouette:** Same as K-means
- **Domain knowledge:** Business requirements

Dendrograms let you explore multiple K values visually

Advantages

When to use hierarchical:

- **No pre-specified K:** Explore structure first
- **Interpretable:** Tree shows relationships
- **Small data:** Works well for hundreds of points

Great for exploratory analysis

Limitations

Computational challenges:

- **Memory:** O(n2) distance matrix
- **Speed:** Slower than K-means
- **Irreversible:** Once merged, cannot undo

Not suitable for very large datasets

Python Implementation

Two approaches:

- **scipy:** For dendrogram visualization
- **sklearn:** For prediction and labels

scipy for visualization, sklearn for clustering

Use scipy.cluster.hierarchy for dendrograms

Quick Reference

Essential Commands:

Task	Code
Linkage matrix	`linkage(X, method='ward')`
Dendrogram	`dendrogram(Z)`
Get labels	`fcluster(Z, t=K, criterion='maxclust')`
sklearn	`AgglomerativeClustering(n_clusters=K)`

Hierarchical clustering reveals data structure