Lesson 30: Hierarchical Clustering

Data Science with Python – BSc Course

45 Minutes

Learning Objectives

The Problem: K-Means requires choosing K upfront. What if we want to see the full hierarchy of cluster relationships at all levels?

After this lesson, you will be able to:

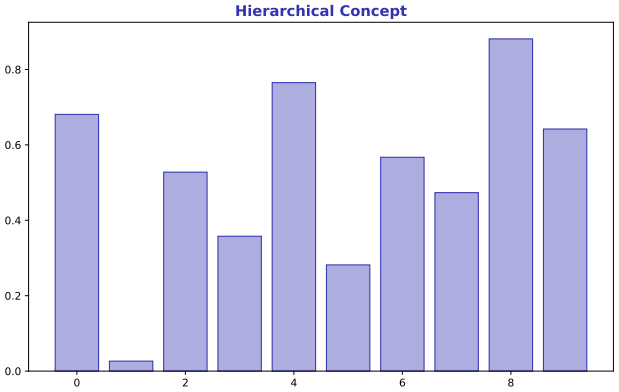
- Build and interpret dendrograms
- Choose between linkage methods (single, complete, ward)
- Cut dendrograms to obtain flat clusters
- Apply hierarchical clustering to portfolio construction

Finance Application: Hierarchical Risk Parity (HRP) portfolio optimization

Hierarchical Concept

Building a Tree of Clusters

- Agglomerative: start with N clusters, merge closest pairs
- Result: nested hierarchy showing relationships at all scales

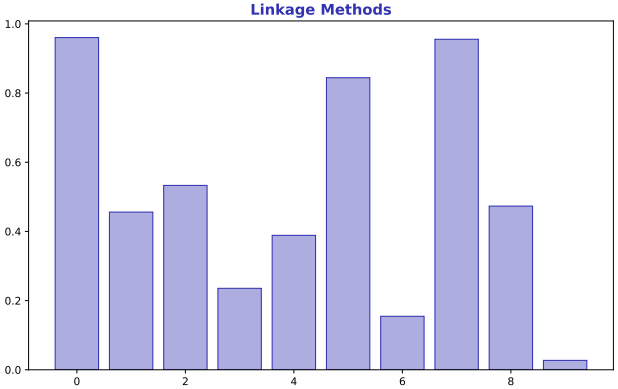


Advantage over K-Means: no need to specify K in advance

Linkage Methods

How to Measure Cluster Distance?

- Single: min distance between clusters (chains)
- Complete: max distance (compact clusters)
- Ward: minimize variance increase (balanced)

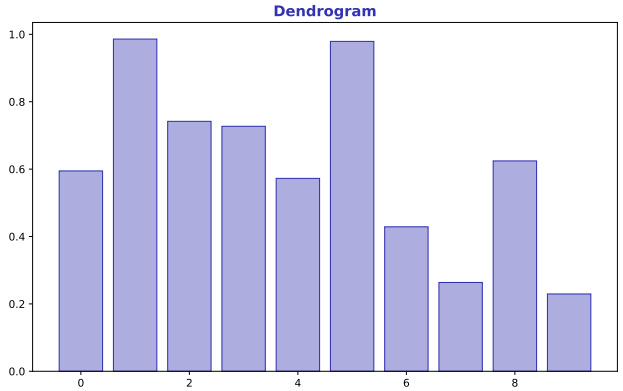


Default recommendation: Ward linkage for most applications

Dendrogram

Visualizing the Hierarchy

- Y-axis: distance at which clusters merge
- X-axis: individual observations (leaves)

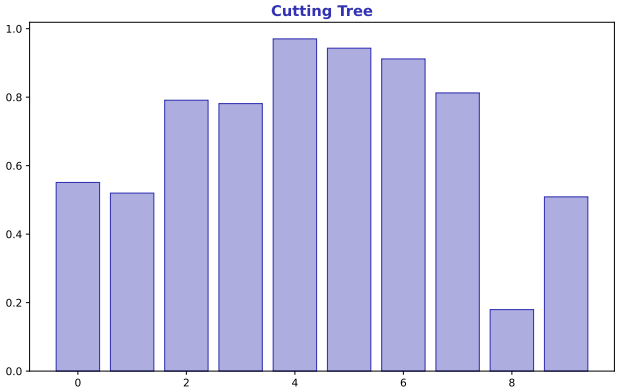


Read dendrograms bottom-up: similar items merge early, different items merge late

Cutting the Tree

From Hierarchy to Flat Clusters

- Draw horizontal line at chosen height – clusters below are groups
- Higher cut = fewer clusters, lower cut = more clusters

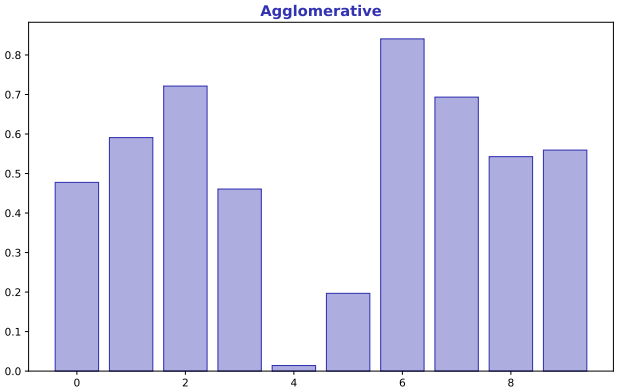


Use `fcluster(Z, t=height, criterion='distance')` **to cut in scipy**

Agglomerative Clustering

sklearn Implementation

- `from sklearn.cluster import AgglomerativeClustering`
- Specify `n_clusters` or `distance_threshold`

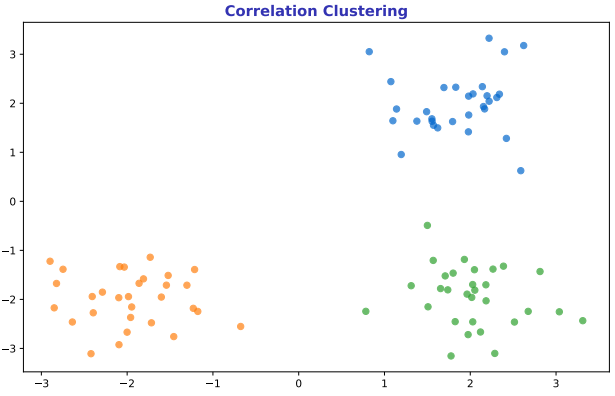


For dendrogram plotting, use scipy.cluster.hierarchy instead

Correlation-Based Clustering

Using Correlation as Similarity

- Distance $= 1 - \text{correlation}$ (or $\sqrt{2(1-\rho)}$)
- Cluster assets by return correlation patterns

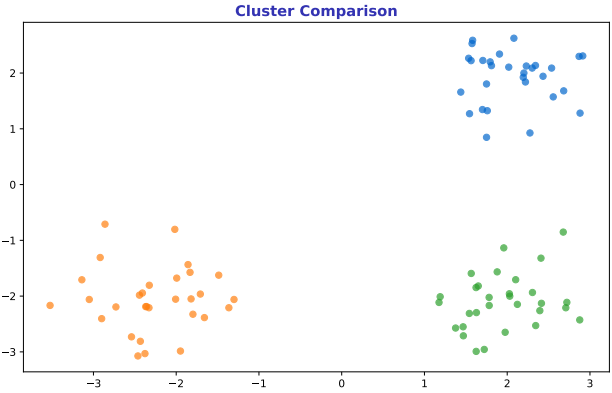


Finance standard: cluster correlation matrix to find asset groups

Comparison with K-Means

When to Use Which?

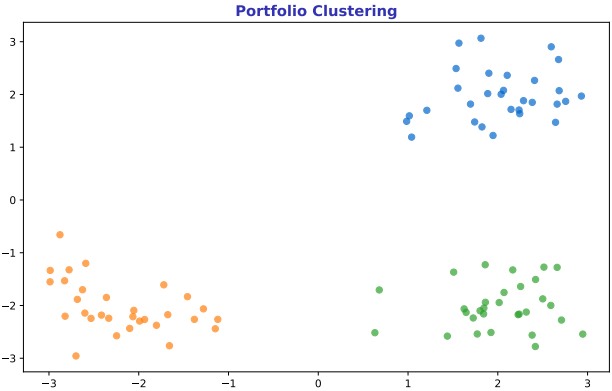
- K-Means: faster, need K, spherical clusters
- Hierarchical: slower, visual hierarchy, any cluster shape



Use hierarchical for exploratory analysis, K-Means for production

Hierarchical Risk Parity (HRP)

- Cluster assets by correlation, then allocate within/across clusters
- More robust than mean-variance optimization



HRP: Modern portfolio construction using hierarchical clustering

Hands-On Exercise (25 min)

Task: Build Asset Hierarchy

1. Calculate correlation matrix for 20 stocks (1 year daily returns)
2. Convert to distance matrix: $d = \sqrt{2(1-\rho)}$
3. Build dendrogram with Ward linkage
4. Cut at 2-3 different heights – compare resulting clusters
5. Label clusters by dominant sector

Deliverable: Dendrogram with cluster cut lines annotated.

Extension: Implement simple HRP allocation based on your clusters

Lesson Summary

Problem Solved: We can now discover hierarchical relationships and create clusters at any granularity.

Key Takeaways:

- Hierarchical clustering builds a tree (dendrogram)
- Linkage method matters: Ward for balanced clusters
- Cut dendrogram at desired height to get flat clusters
- Finance: correlation-based clustering for portfolio construction

Next Lesson: PCA (L31) – reducing dimensions while preserving information

Memory: Dendrogram = tree. Cut horizontally to get clusters. Ward = balanced.