THINKFUL

Hierarchical Clustering

DATA SCIENCE

Warm Up

Consider the following questions:

- What might a hierarchy of clusters look like?
- When could a hierarchy of clusters be preferable to a one-to-one assignment of data points to clusters?

Agenda

- Overview of hierarchical clustering
- Walkthrough of agglomerative clustering
  - Comparisons of various linkage types
- Advantages and disadvantages

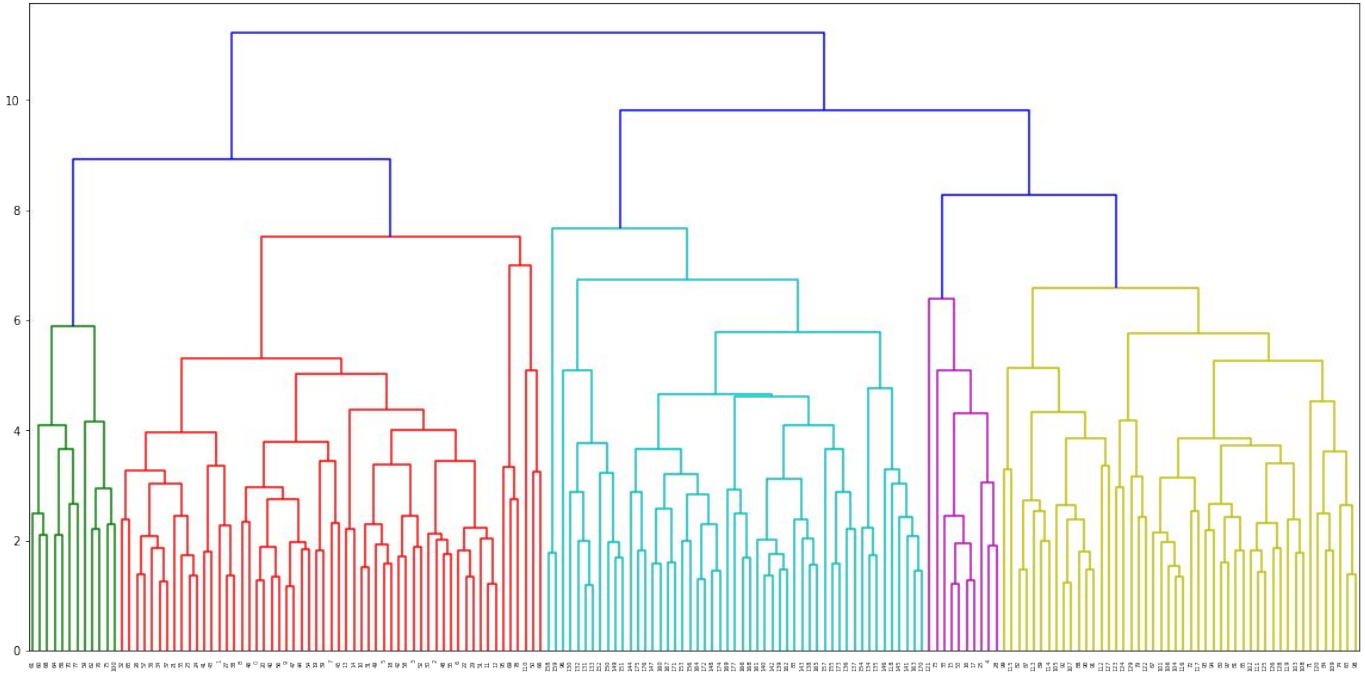
Overview

Hierarchical clustering builds out a hierarchy of clusters (“clusters within clusters”)

- Unlike centroid-based approaches, no fixed number of clusters; after the hierarchy is built, the user can decide on how many clusters are appropriate
- Like medoid-based approaches, can work with any arbitrary dissimilarity measure and only requires a dissimilarity matrix, not the actual data
- Clusters connect observations together at varying distances (as opposed to partitioning them, as in k-means and k-medoids); also known as *connectivity-based clustering*

1/2

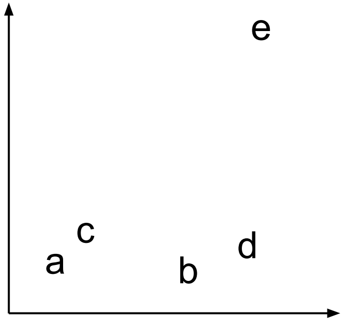
Overview



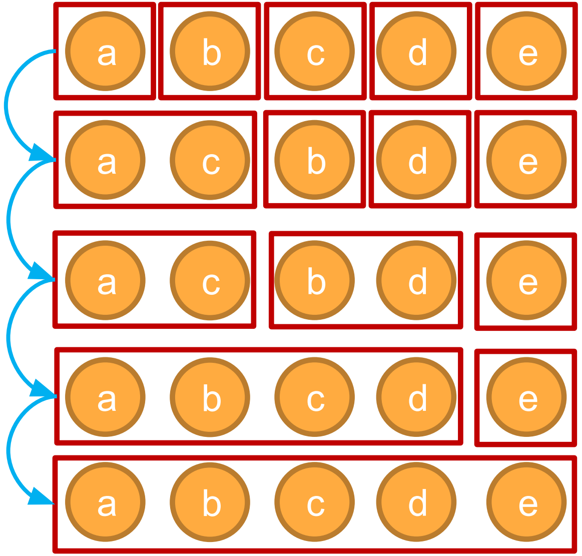
Agglomerative vs. Divisive

Two primary approaches to building clusters: agglomerative (bottom-up) and divisive (top-down)

Original Data



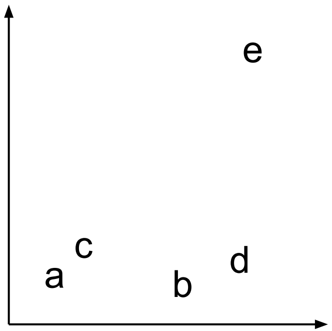
Agglomerative: each observation starts in its own cluster; clusters are merged together until one big cluster is reached



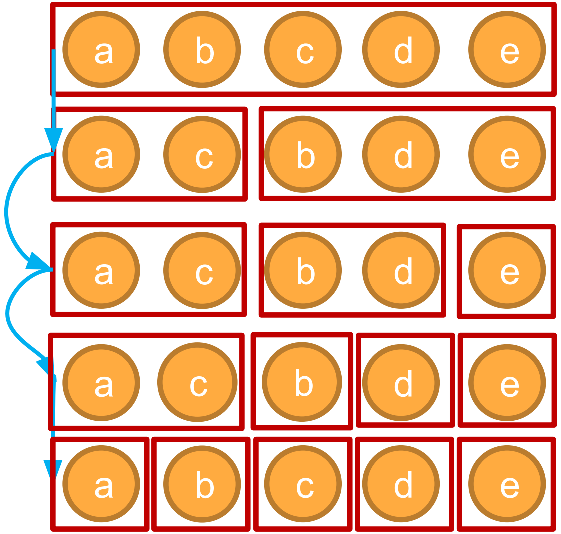
Agglomerative vs. Divisive

Two primary approaches to building clusters: agglomerative (bottom-up) and divisive (top-down)

Original Data



Divisive: all observations start in the same clusters; cuts are made to each cluster until each observation is in its own cluster

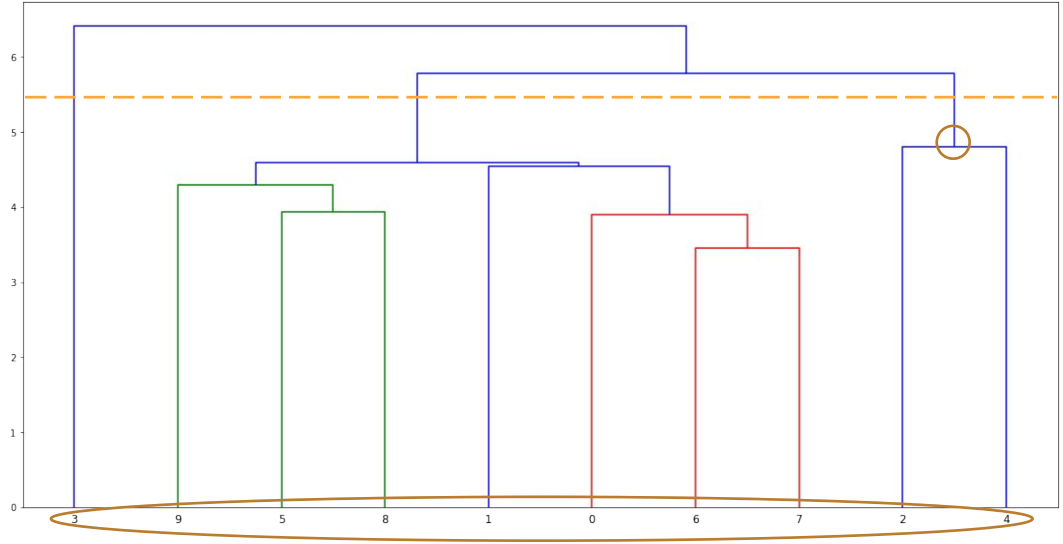


Visualization: Dendrograms

The results of hierarchical clustering are commonly visualized using tree diagrams called dendrograms

We can cut the dendrogram at any height to obtain a concrete number of clusters

The number of times the cut line intersects the dendrogram corresponds to the number of clusters



The split points of the tree correspond to the merging of the two child clusters

The height of the split point corresponds to the distance or dissimilarity between the merged clusters

The original observations are leaves at the bottom of the dendrogram, each starting out as its own cluster in agglomerative clustering

UP NEXT

Linkage Criteria

Linkage Overview

One of the decisions we must make when updating clusters for hierarchical clustering is how to calculate the distance between two clusters A and B. The function that determines these distances is referred to as the “linkage criterion” and is often a function of the pairwise dissimilarities between the observations in both clusters. Three of the most popular linkage criteria are:

- D: distance
- M and N: clusters to calculate distance between
- i and j: observations assigned to clusters M and N, respectively
- n_M and n_N: number of observations in clusters M and N, respectively

Minimum or Single-Linkage

$$D_{single}(M,N) = \min_{i\,\in M, j\in N} d_{ij}$$

Maximum or Complete-Linkage

$$D_{complete}(M,N) = \max_{i\,\in M, j\in N} d_{ij}$$

Average Linkage

$$D_{average}(M,N) = \frac{1}{n_M n_N} \sum_{i\,\in M, j\in N} d_{ij}$$

Single Linkage

Single-linkage considers the **minimum** pairwise dissimilarity between elements of two clusters to be the distance between the clusters

- Interpretation of dendrogram cut at height h:

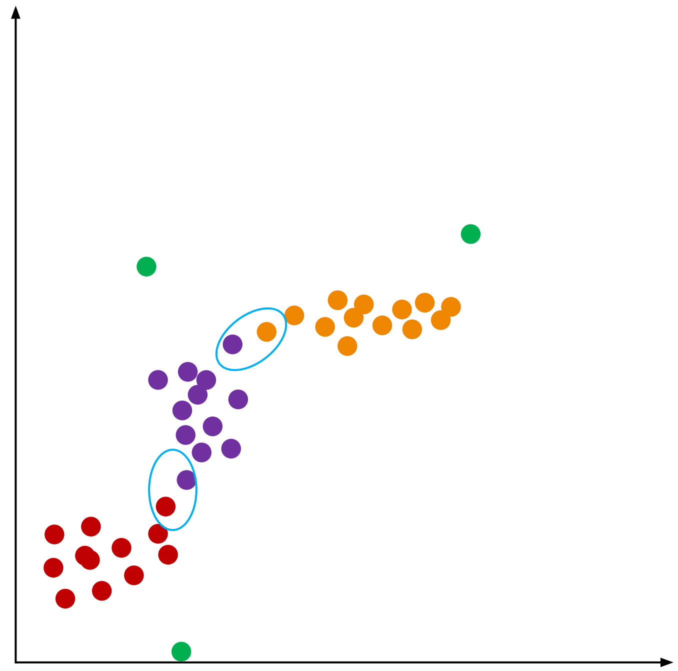
  For each point in a given cluster, there is another point in the same cluster with dissimilarity less than or equal to h
- Suffers from “chaining” phenomenon: even if two clusters are generally far apart, they only need to have one pair of similar points to be considered “close”
- Often leads to long, loose clusters with last few merges only merging in individual points

1/2

Single Linkage

Although the red, purple, and orange clusters are generally far apart, they happen to have pairs of points that are close to each other. Under single-linkage, the distance between the red and purple clusters will be equal to the distance between the pair of circled red and purple observations.

Consequently, they will be merged before the green points, resulting in a long, loose cluster that is not compact.



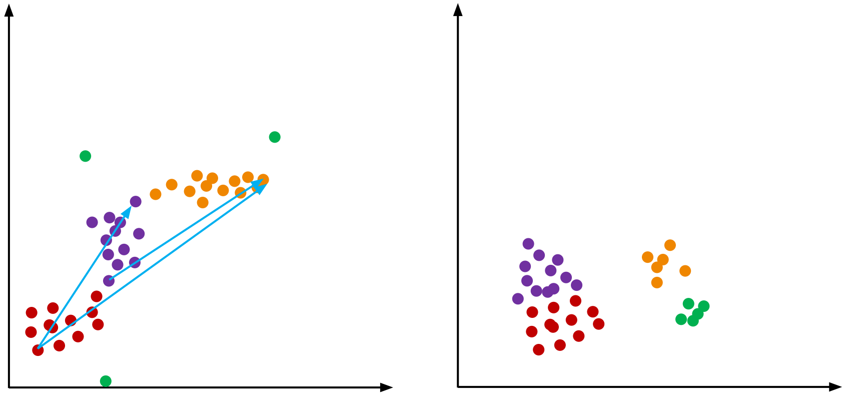
Complete Linkage

Complete-linkage considers the **maximum** pairwise dissimilarity between elements of two clusters to be the distance between the clusters

- Interpretation of dendrogram cut at height h:
  For each point in a given cluster, all other points within the same cluster have a dissimilarity of h or less
- Avoids chaining problem associated with single linkage, since distance is dependent on worst-case scenario instead of best
- Tends to produce more balanced, compact clusters; however, can suffer from “crowding” problem in which clusters are compact but not very far apart

Complete Linkage

Under complete-linkage, the distances between the red, purple, and orange clusters depend on the maximum dissimilarity, indicated by the blue arrows in the left diagram. As a result, they are considered much further apart and are unlikely to be merged until the very end, avoiding the chaining problem from single-linkage.



However, the right diagram illustrates the "crowding" issue that can be seen with complete linkage. A common manifestation of this issue occurs when large clusters are split into smaller, compact clusters that are very close to each other, as the red and purple are. Complete-linkage may very well merge the orange and green before the red and purple.

Average Linkage

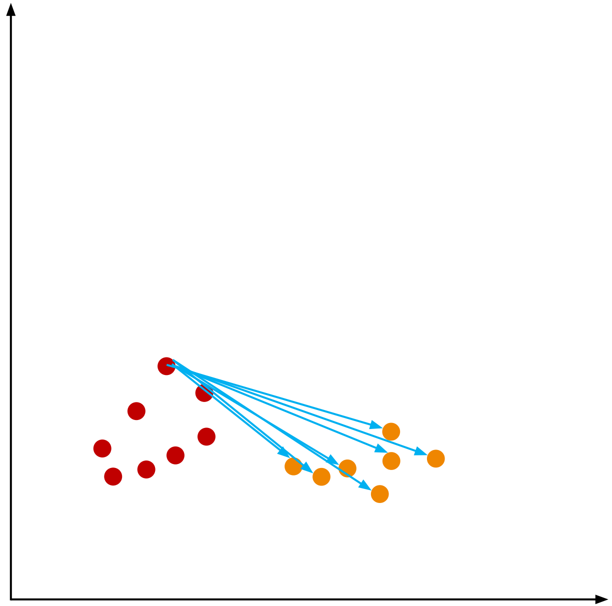
Average-linkage considers the **average** pairwise dissimilarity between elements of two clusters to be the distance between the clusters

- Interpretation of dendrogram cut at height h:
  No easy interpretation
- Balances the downsides of chaining (from single-linkage) and crowding (from complete-linkage)
- Tends to produce clusters that demonstrate a balance of compactness and separation
- However, no clear interpretation; also, monotone transformations of the distance function can produce different clustering results

Average Linkage

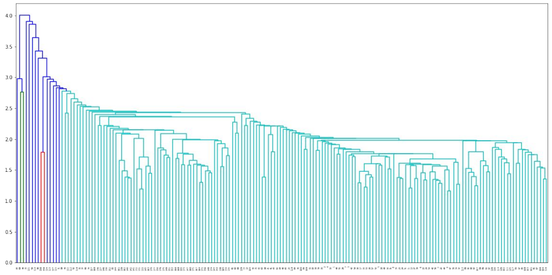
The above diagram shows all of the pairwise distances for one of the points in the red cluster.

For average-linkage, we would repeat this process for all of the points in the red cluster (i.e. draw lines from every red point to every orange point) and then average the resulting 8 * 7, or 56 distances to determine the distance between the two clusters.

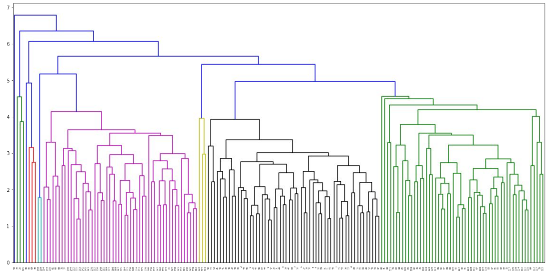


Comparison Of Linkage Methods

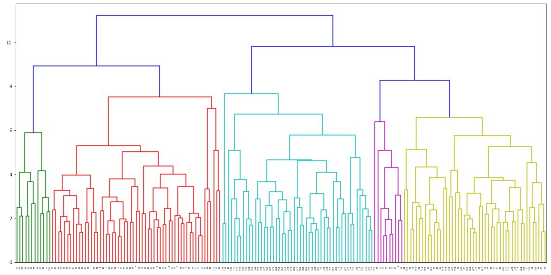
The following dendrograms visualize the results of hierarchical clustering on scikit-learn’s *wine* dataset with the aforementioned linkages



Single



Average



Complete

Advantages & Disadvantages

Primary Advantage: Flexibility

- Doesn’t require any pre-specified number of clusters
  - Hierarchy can be cut at any level to obtain any desired number of clusters
- Like k-medoids, works with arbitrary dissimilarity measures and categorical data
- Many real-world contexts naturally lend themselves to a hierarchical framework
  - For example, products that a grocery store sells are easy to think about hierarchically – perishable vs. non-perishable, then within perishable, packaged vs. non-packaged, etc.

1/2

Advantages & Disadvantages

Primary Disadvantage: Poor scalability

- Naïve implementation involves recomputing all pairwise distances at every iteration/merge
- Naïve time complexity of O(n^3) and naïve memory scaling of O(n^2)
  - Unsuitable for anything beyond small datasets
- Well-optimized single-linkage and complete-linkage algorithms can get down to O(n^2) time complexity
  - Still cannot handle large datasets

Summary

- Hierarchical clustering is a fundamentally different approach from k-means, k-medoids, etc., that relies on connectivity rather than portioning
- The output of hierarchical clustering is heavily influenced by the linkage criterion chosen; commonly used ones include single, complete, and average
- Hierarchical clustering is very flexible and can produce any number of desired clusters
- Hierarchical clustering also works with categorical and mixed data
- However, it scales very poorly and is difficult to use with large datasets

EXERCISE

1. Jupyter notebook
2. Data set: Starbucks locations in the U.S.
3. Choose a reasonably sized subset of the locations in the U.S.
4. Calculate the geographical distance matrix using haversine distance
5. Build a hierarchical clustering model with average linkage and a fixed number of clusters
6. Plot the dendrogram resulting from hierarchical clustering
7. Plot the resulting clusters of locations on a map using plotly.express

THINKFUL

Thank You

Hierarchical Clustering

Warm Up

Consider the following questions:

- What might a hierarchy of clusters look like?
- When could a hierarchy of clusters be preferable to a one-to-one assignment of data points to clusters?

High Level Agenda

- Overview of hierarchical clustering
- Walkthrough of agglomerative clustering
    - Comparisons of various linkage types
- Advantages and disadvantages

Overview

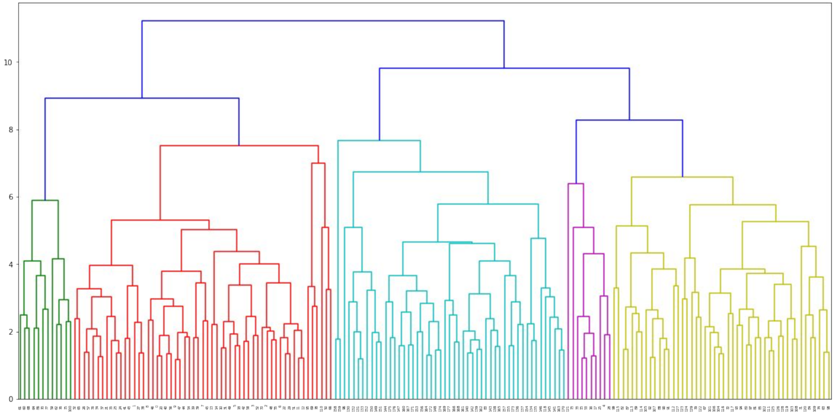
Overview

Hierarchical clustering builds out a hierarchy of clusters ("clusters within clusters")

- Unlike centroid-based approaches, no fixed number of clusters; after the hierarchy is built, the user can decide on how many clusters are appropriate

- Like medoid-based approaches, can work with any arbitrary dissimilarity measure and only requires a dissimilarity matrix, not the actual data

- Clusters connect observations together at varying distances (as opposed to partitioning them, as in k-means and k-medoids); also known as *connectivity-based clustering*

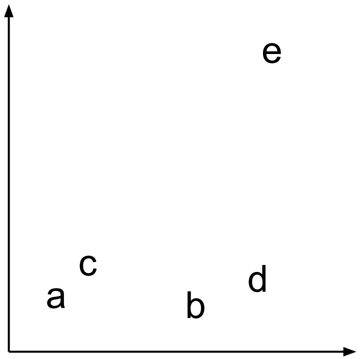


THINKFUL

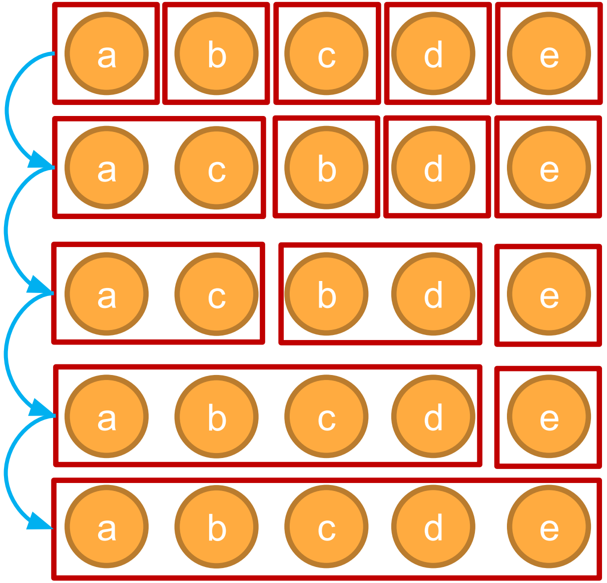
Agglomerative vs. Divisive

Two primary approaches to building clusters: agglomerative (bottom-up) and divisive (top-down)

Original Data



Agglomerative: each observation starts in its own cluster; clusters are merged together until one big cluster is reached

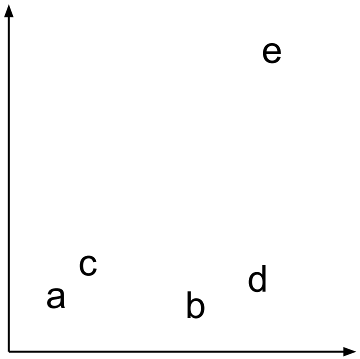


THINKFUL

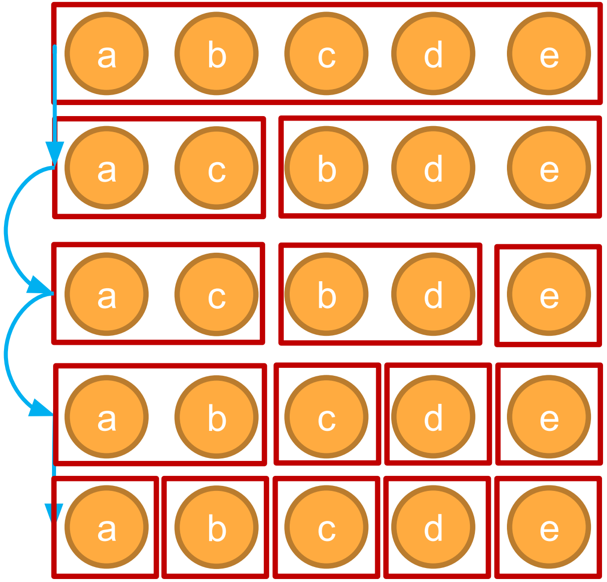
Agglomerative vs. Divisive

Two primary approaches to building clusters: agglomerative (bottom-up) and divisive (top-down)

Original Data



Divisive: all observations start in the same clusters; cuts are made to each cluster until each observation is in its own cluster

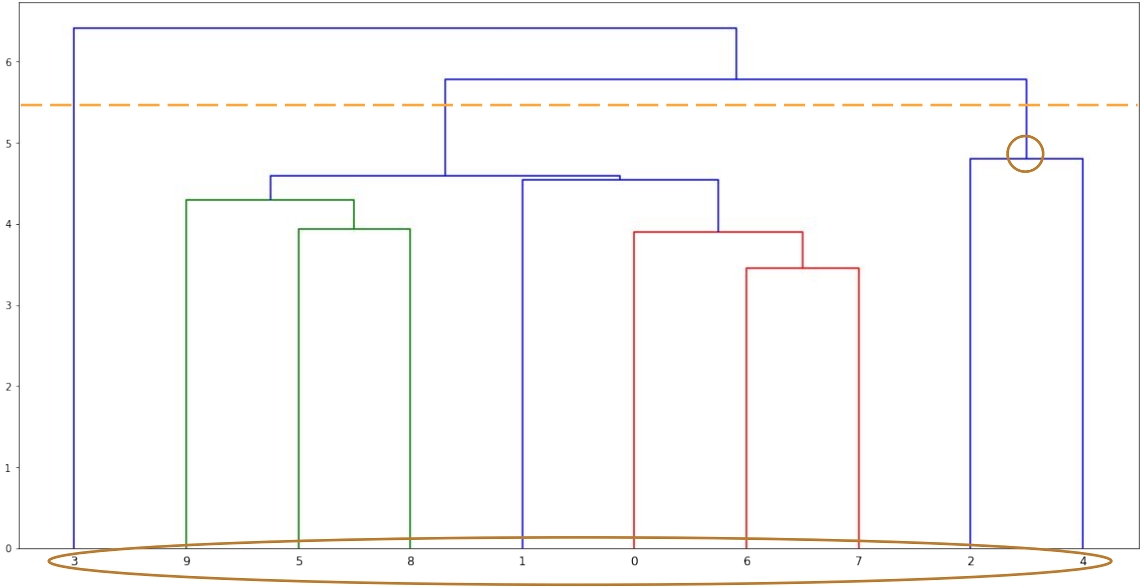


Visualization: Dendrograms

The results of hierarchical clustering are commonly visualized using tree diagrams called dendrograms

We can cut the dendrogram at any height to obtain a concrete number of clusters

The number of times the cut line intersects the dendrogram corresponds to the number of clusters



The split points of the tree correspond to the merging of the two child clusters

The height of the split point corresponds to the distance or dissimilarity between the merged clusters

The original observations are leaves at the bottom of the dendrogram, each starting out as its own cluster in agglomerative clustering

Linkage Criteria

Linkage Overview

One of the decisions we must make when updating clusters for hierarchical clustering is how to calculate the distance between two clusters A and B. The function that determines these distances is referred to as the “linkage criterion” and is often a function of the pairwise dissimilarities between the observations in both clusters. Three of the most popular linkage criteria are:

Minimum or
Single-Linkage

$$D_{single}(M,N) = \min_{i \in M, j \in N} d_{ij}$$

Maximum or
Complete-Linkage

$$D_{complete}(M,N) = \max_{i \in M, j \in N} d_{ij}$$

Average
Linkage

$$D_{average}(M,N) = \frac{1}{n_M n_N} \sum_{i \in M, j \in N} d_{ij}$$

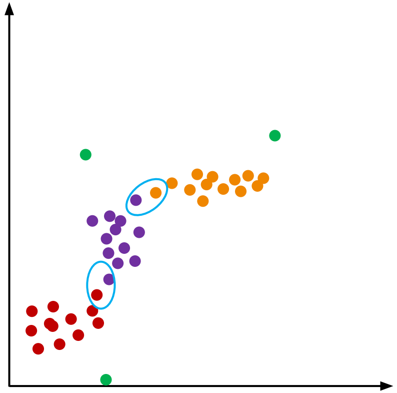
D: distance
M and N: clusters to calculate distance between
i and j: observations assigned to clusters M and N, respectively
n_M and n_N: number of observations in clusters M and N, respectively

THINKFUL

Single Linkage

Single-linkage considers the **minimum** pairwise dissimilarity between elements of two clusters to be the distance between the clusters

- Interpretation of dendrogram cut at height h: For each point in a given cluster, there is another point in the same cluster with dissimilarity less than or equal to h
- Suffers from “chaining” phenomenon: even if two clusters are generally far apart, they only need to have one pair of similar points to be considered “close”
- Often leads to long, loose clusters with last few merges only merging in individual points



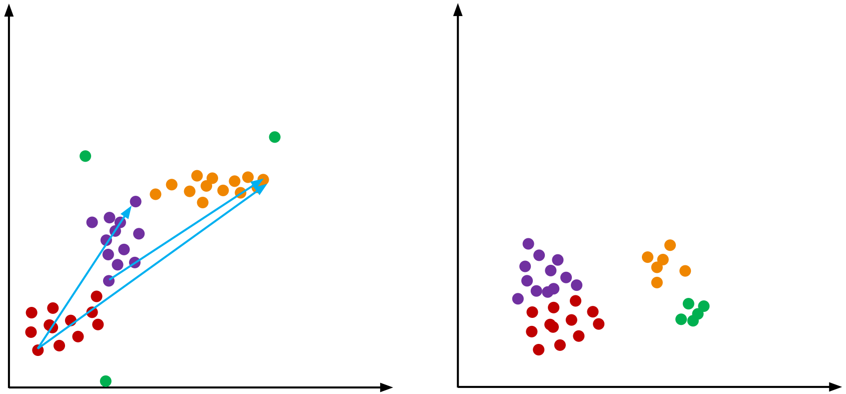
Although the red, purple, and orange clusters are generally far apart, they happen to have pairs of points that are close to each other. Under single-linkage, the distance between the red and purple clusters will be equal to the distance between the pair of circled red and purple observations.

Consequently, they will be merged before the green points, resulting in a long, loose cluster that is not compact.

Complete Linkage

Complete-linkage considers the **maximum** pairwise dissimilarity between elements of two clusters to be the distance between the clusters

- Interpretation of dendrogram cut at height h: For each point in a given cluster, all other points within the same cluster have a dissimilarity of h or less
- Avoids chaining problem associated with single linkage, since distance is dependent on worst-case scenario instead of best
- Tends to produce more balanced, compact clusters; however, can suffer from "crowding" problem in which clusters are compact but not very far apart



Under complete-linkage, the distances between the red, purple, and orange clusters depend on the maximum dissimilarity, indicated by the blue arrows in the left diagram. As a result, they are considered much further apart and are unlikely to be merged until the very end, avoiding the chaining problem from single-linkage.

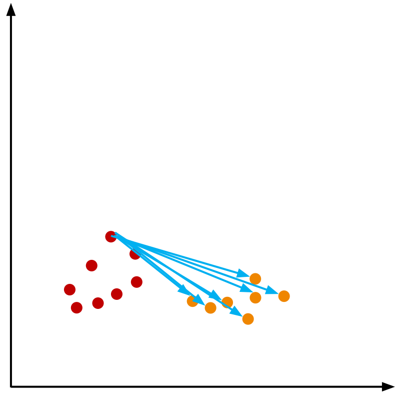
However, the right diagram illustrates the "crowding" issue that can be seen with complete linkage. A common manifestation of this issue occurs when large clusters are split into smaller, compact clusters that are very close to each other, as the red and purple are. Complete-linkage may very well merge the orange and green before the red and purple.

THINKFUL

Average Linkage

Average-linkage considers the **average** pairwise dissimilarity between elements of two clusters to be the distance between the clusters

- Interpretation of dendrogram cut at height h:
  No easy interpretation
- Balances the downsides of chaining (from single-linkage) and crowding (from complete-linkage)
- Tends to produce clusters that demonstrate a balance of compactness and separation
- However, no clear interpretation; also, monotone transformations of the distance function can produce different clustering results

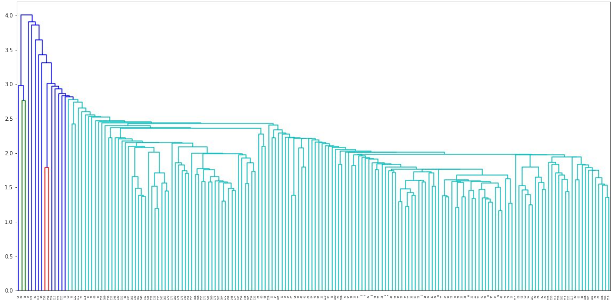


The above diagram shows all of the pairwise distances for one of the points in the red cluster.

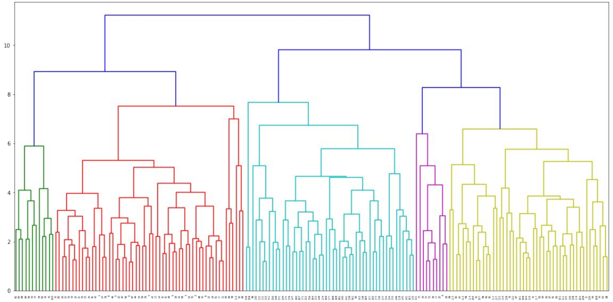
For average-linkage, we would repeat this process for all of the points in the red cluster (i.e. draw lines from every red point to every orange point) and then average the resulting 8 * 7, or 56 distances to determine the distance between the two clusters.

Comparison of Linkage Methods

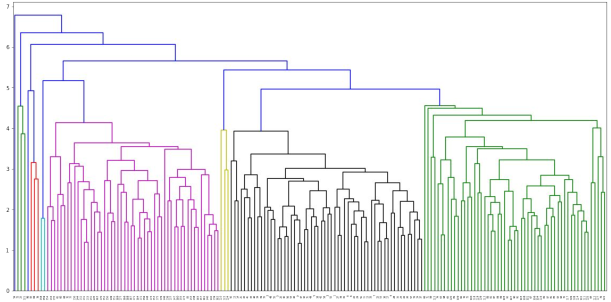
The following dendrograms visualize the results of hierarchical clustering on scikit-learn's *wine* dataset with the aforementioned linkages



Single



Complete



Average

Advantages and Disadvantages

Advantages and Disadvantages

Primary Advantage: Flexibility

- Doesn’t require any pre-specified number of clusters
  - Hierarchy can be cut at any level to obtain any desired number of clusters
- Like k-medoids, works with arbitrary dissimilarity measures and categorical data
- Many real-world contexts naturally lend themselves to a hierarchical framework
  - For example, products that a grocery store sells are easy to think about hierarchically – perishable vs. non-perishable, then within perishable, packaged vs. non-packaged, etc.

Primary Disadvantage: Poor scalability

- Naïve implementation involves recomputing all pairwise distances at every iteration/merge
- Naïve time complexity of $O(n^3)$ and naïve memory scaling of $O(n^2)$
  - Unsuitable for anything beyond small datasets
- Well-optimized single-linkage and complete-linkage algorithms can get down to $O(n^2)$ time complexity
  - Still cannot handle large datasets

Recap

- Hierarchical clustering is a fundamentally different approach from k-means, k-medoids, etc., that relies on connectivity rather than portioning
- The output of hierarchical clustering is heavily influenced by the linkage criterion chosen; commonly used ones include single, complete, and average
- Hierarchical clustering is very flexible and can produce any number of desired clusters
- Hierarchical clustering also works with categorical and mixed data
- However, it scales very poorly and is difficult to use with large datasets

Exercise:
Hierarchical Clustering

- Jupyter notebook
- Data set: Starbucks locations in the U.S.
- Choose a reasonably sized subset of the locations in the U.S.
- Calculate the geographical distance matrix using haversine distance
- Build a hierarchical clustering model with average linkage and a fixed number of clusters
- Plot the dendrogram resulting from hierarchical clustering
- Plot the resulting clusters of locations on a map using plotly.express