Section 12: Hierarchical clustering

STA 35C – Statistical Data Science III

Instructor: Akira Horiguchi

Fall Quarter 2025 (Sep 24 – Dec 12)
MWF, 12:10 PM – 1:00 PM, Olson 158
University of California, Davis

Overview

Based on Chapter 12 of ISL book James et al. (2021).

- For more R code examples, see R Markdown files in `https://www.statlearning.com/resources-second-edition`

Motivation

K-means clustering requires you to prespecify the number of clusters K.

- This can be an issue.
- *Hierarchical clustering* is an alternative that does not require this.

Suppose we have n observations $x_1, \ldots, x_n \in \mathbb{R}^p$. (Example below: $n = 9$ and $p = 2$.)

Algorithm:

1. *Treat each observation as a cluster.* I.e., create n singleton clusters.
2. *Keep merging together similar clusters until all observations have been merged into a single cluster.* For $i = \boxed{n}, \boxed{n-1}, \ldots, 2$:
    (a) For each of the $\binom{i}{2}$ cluster pairs, compute the pair's dissimilarity. (Dissimilarity measure is often Euclidean distance; will discuss more later).
    (b) Identify the least dissimilar (i.e. most similar) pair of clusters. Merge these two clusters.

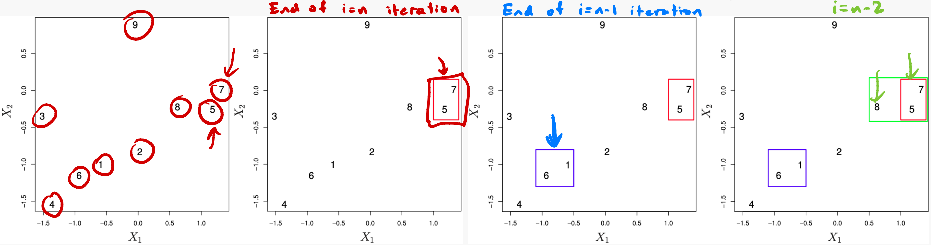

Figure 1: From James et al. (2021). First few steps of the hierarchical clustering algorithm with complete linkage and Euclidean distance.

Dendrogram view

H-clust process can be visualized using a tree-based illustration called a *dendrogram*.

- Each leaf of dendrogram represents an observation. (Step 1 of algorithm)
- As we move up the tree, some leaves begin to fuse into branches. Then branches begin to fuse. Each fusion corresponds to an iteration of Step 2 of algorithm.

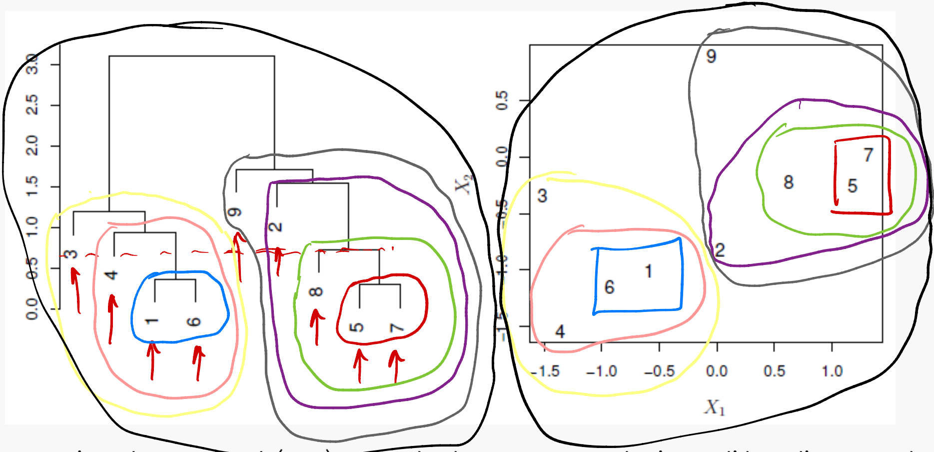

Figure 2: Figure by James et al. (2021). Left: A dendrogram generated using Euclidean distance and complete linkage. Right: The raw data used to generate the dendrogram.

Interpreting a dendrogram

More comments

- For any two observations, height of fusion indicates how different the observations are. (Ignore horizontal proximity.)
- To identify clusters, make a horizontal cut across dendrogram.
- Height of cut controls number of clusters obtained.
- A single dendrogram can be used to get any number of clusters. Eyeball.
- H-clust algorithm is deterministic (i.e. non-random).

A larger example:

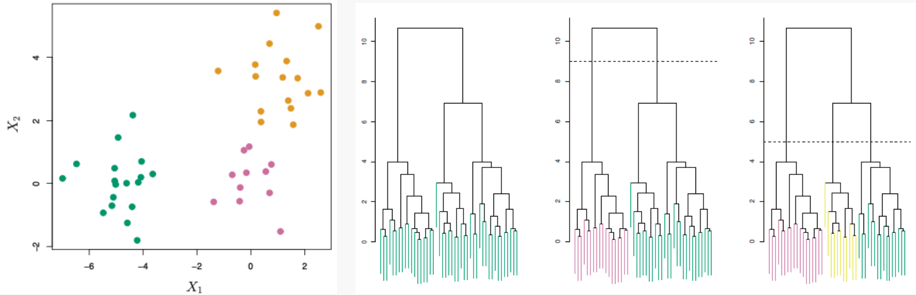

Figure 3: From James et al. (2021). 45 observations.

Some comments

Hierarchical clustering sometimes produces worse results than K-means clustering.

- Suppose we record various body measurements (e.g., height, weight, nose length) of 60 raccoons.
  - 20 from NYC, 20 from Tokyo, 20 from Cairo.
  - 30 males and 30 females.
- K-means clustering with $K = 2$ might group raccoons by sex, and with $K = 3$ by city.
- These two partitions are not nested, so they cannot be achieved by the same dendrogram from a hierarchical clustering.

Dissimilarity between two clusters

How to define dissimilarity between e.g., cluster $\{5, 7\}$ and cluster $\{8\}$?

- Need to extend dissimilarity to two groups of observations.
- *Linkages* define the dissimilarity between two groups of observations.
  1. *Complete*: computes all dissimilarities between an observation in cluster A and an observation in cluster B, and record largest of these $n_A n_B$ dissimiliarities.
  2. *Single*: same, except record smallest of these $n_A n_B$ dissimiliarities.
  3. *Average*: same, except record mean of these $n_A n_B$ dissimiliarities.
  4. *Centroid*: dissimilarity between two cluster centroids.

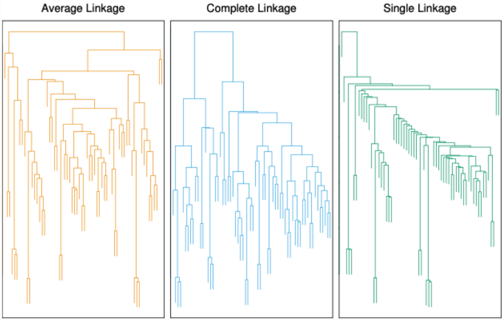

Figure 4: From James et al. (2021). Average, complete, and single linkage applied to an example data set. Average and complete linkage tend to yield more balanced clusters.

Issues in hierarchical clustering

- What dissimilarity measure should be used?
- What type of linkage should be used?
- Where shall the dendogram be cut (i.e., how many clusters do we need/want)?