Unsupervised learning
Part 2: Hierarchical clustering

By: Noureddin Sadawi, PhD
University of London

Hierarchical clustering

- ‘Hierarchical clustering is an alternative to K-means that can yield very different clusters.’
- ‘Hierarchical clustering allows the user to visualize the effect of specifying different numbers of clusters.’
- ‘…more sensitive in discovering outlying or aberrant groups or records.‘
- ‘…lends itself to an intuitive graphical display, leading to easier interpretation of the clusters.’

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

Hierarchical clustering

- ‘Hierarchical clustering’s flexibility comes with a cost, and it does not scale well to large data sets with millions of records.’
- ‘For even modest-sized data with just tens of thousands of records, it can require intensive computing resources.’
- ‘Indeed, most of the applications of hierarchical clustering are focused on relatively small data sets.’

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

Key terms for hierarchical clustering

‘Dendrogram

A visual representation of the records and the hierarchy of clusters to which they belong.

Distance

A measure of how close one *record* is to another.

Dissimilarity

A measure of how close one *cluster* is to another.’

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

How hierarchical clustering works

It works on a data set with *n* records and *p* variables and is based on two basic building blocks:

- ‘A distance metric $d_{i,\ j}$ to measure the distance between two records *i* and *j*.’
- ‘A dissimilarity metric $D_{A,\ B}$ to measure the difference between two clusters *A* and *B* based on the distances $d_{i,\ j}$ between the members of each cluster.’

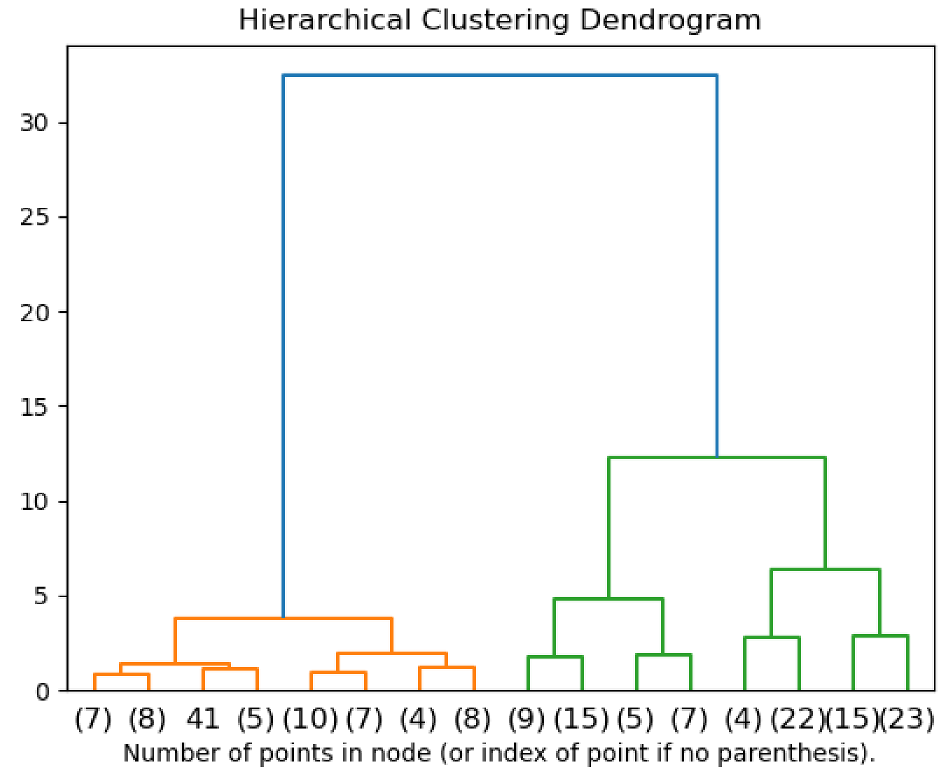
Dissimilarity metric is very important!

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

The dendrogram

- ‘Hierarchical clustering starts by setting each record as its own cluster and iterates to combine the least dissimilar clusters.’
- ‘Hierarchical clustering lends itself to a natural graphical display as a tree, referred to as a *dendrogram*.’

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

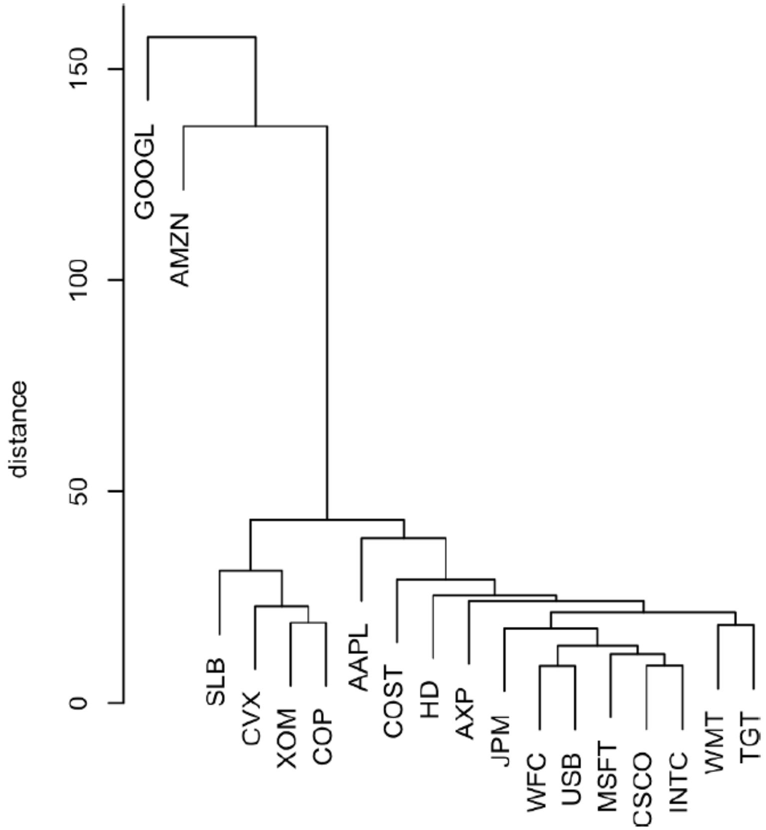


https://scikit-learn.org/stable/modules/clustering.html#hierarchical-clustering

Hierarchical vs K-means

- ‘In contrast to K-means, it is not necessary to pre-specify the number of clusters.’
- ‘Graphically, you can identify different numbers of clusters with a horizontal line that slides up or down.’
- ‘…a cluster is defined wherever the horizontal line intersects with the vertical lines.’
- ‘…you can see that Google and Amazon each belong to their own cluster.’
- ‘The oil stocks… all belong to another cluster.’
- ‘The remaining stocks are in the fourth cluster.’

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).



A dendrogram of stocks.

The agglomerative algorithm

‘1. Create an initial set of clusters with each cluster consisting of a single record for all records in the data
2. Compute the dissimilarity $D(C_x, C_y)$ between all pairs of clusters x, y
3. Merge the two clusters C_x and C_y that are least dissimilar as measured by $D(C_x, C_y)$
4. If we have more than one cluster remaining, return to step 2. Otherwise, we are done!’

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

Measure of dissimilarity

There are four common measures of dissimilarity: *complete linkage*: *single linkage*, *average linkage*, and *minimum variance*

The single linkage method is the minimum distance between the records:

$$D(A, B) = min\, d\left(a_i, b_j\right) for\ all\ pairs\ i, j$$

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).