Aglomerative Clustering and Clustering Evaluation

David Mareček

December 09, 2021

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

EUROPEAN UNION
European Structural and Investment Fund
Operational Programme Research,
Development and Education

Hierarchical clustering

Each observation starts in its own cluster, and clusters are successively merged together.

The linkage criteria:

- **Mimimum (Single linkage)** minimizes the minimum distance between pairs of clusters.

$$d_{single} = \min\{d(x_i, x_j),\ i \in A, j \in B\}$$

- **Maximum (Complete linkage)** minimizes the maximum distance between pairs of clusters.

$$d_{complete} = \max\{d(x_i, x_j),\ i \in A, j \in B\}$$

- **Average linkage** minimizes the average of the distances between all observations of pairs of clusters.

$$d_{average} = \frac{1}{|A| \cdot |B|} \sum_{i \in A} \sum_{j \in B} d(x_i, x_j)$$

- **Cenroid linkage** minimizes the distance between centers of clusters.

$$d_{centroid} = d\left(\frac{1}{|A|} \sum_{i \in A} x_i, \frac{1}{|B|} \sum_{i \in B} x_i\right)$$

Hierarchical clustering

- **Ward linkage** is a variance minimizing approach. The distance between two clusters A and B is how much the sum of saquares will increase when we merge them. It is similar to the k-means objective function but tackled with an agglomerative hierarchical approach.

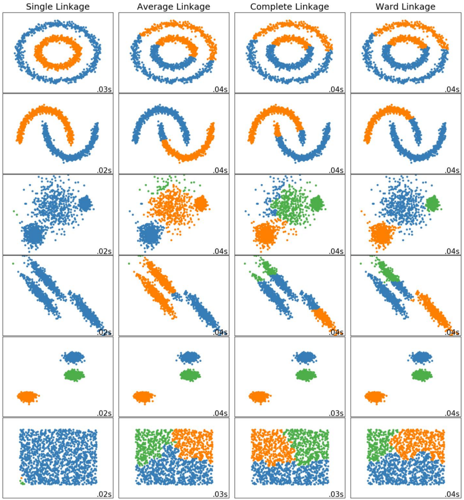
  $$d_{Ward}(A,B) = \sum_{i \in A \cup B} ||x_i - m_{A \cup B}||^2 - \sum_{i \in A} ||x_i - m_A||^2 - \sum_{i \in B} ||x_i - m_B||^2,$$

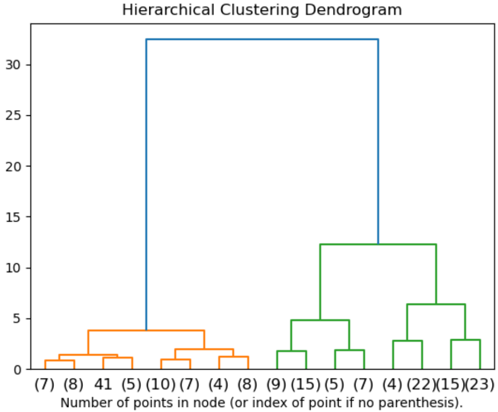
  where m_X is the mean (center) of cluster X. It also corresponds to the squared distance between the centers of the clusters

  $$d_{Ward}(A,B) = \frac{n_A n_B}{n_A + n_B} ||m_A - m_B||^2,$$

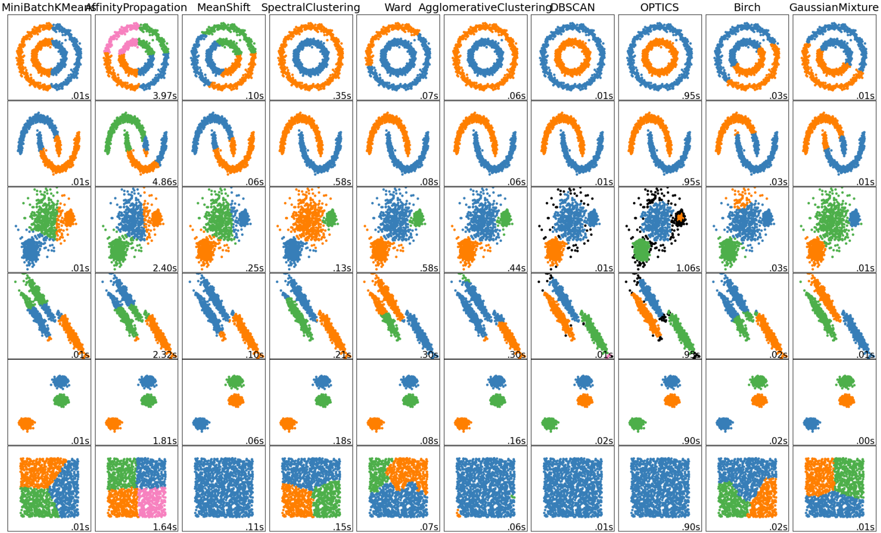
  where n_A and n_B are number of points in clusters A and B, respectively.

Hierarchical clustering





Clustering Methods Comparison



Clustering Evaluation

If you have a testing data available with annotated gold label:

Rosenberg and Hirschberg (2007) define the following objectives for any cluster assignment:

- **Homogeneity** – each cluster contains only members of a single class
- **Completeness** – all members of a given class are assigned to the same cluster
- **V-measure** – their harmonic mean

If you do not have any labelled data:

- **Silhouette coefficient** – "unsupervised" consistency within clusters of data

Homogeneity

Homogeneity – To what extent each cluster contains only members of a single class?

$$h = 1 - \frac{H(C|K)}{H(C)}$$

$H(C|K)$ is the conditional entropy of the cluster assignments given the classes:

$$H(C|K) = -\sum_{c=1}^{|C|}\sum_{k=1}^{|K|}\frac{n_{c,k}}{n}\log\frac{n_{c,k}}{n_k}$$

$H(C)$ is the entropy of the classes:

$$H(C) = -\sum_{c=1}^{|C|}\frac{n_c}{n}\log\frac{n_c}{n}$$

Completeness

Completeness – To what extent all members of a given class are assigned to the same cluster?

$$c = 1 - \frac{H(K|C)}{H(K)}$$

$H(K|C)$ is the conditional entropy of the classes given the cluster assignments:

$$H(K|C) = -\sum_{c=1}^{|C|} \sum_{k=1}^{|K|} \frac{n_{c,k}}{n} \log \frac{n_{c,k}}{n_c}$$

$H(K)$ is the entropy of the clusters:

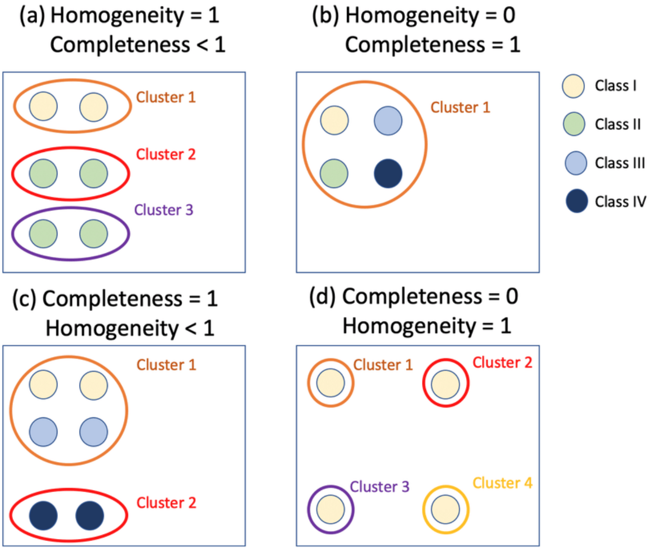
$$H(K) = -\sum_{k=1}^{|K|} \frac{n_k}{n} \log \frac{n_k}{n}$$

V-measure

V-measure – Harmonic mean of homogeneity and comleteness:

$$v = \frac{2 \cdot h \cdot c}{h + c}$$

Homogeneity and Completeness



Silhouette coefficient

How similar an object is to its own cluster (cohesion) compared to other clusters (separation)

Values between -1 and 1

The Silhouette Coefficient s for a single sample is then given as:

$$s_i = \frac{b_i - a_i}{max(a_i, b_i)}, \quad a_i = \frac{1}{C_I - 1} \sum_{j \in C_I, i \neq j} d(i,j), \quad b_i = \min_{J \neq I} \frac{1}{C_J} \sum_{j \in C_J} d(i,j)$$

- a is the mean distance between a sample and all other points in the same cluster
- b is the mean distance between a sample and all other points in the next nearest cluster

The mean over all points of a cluster is a measure of how tightly grouped all the points in the cluster are. Thus the mean over all data of the entire dataset is a measure of how appropriately the data have been clustered.

https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html