Cluster validation
Cluster validation is the process of evaluating the quality and performance of a clustering algorithm on a given data set. It can help you choose the optimal number of clusters, compare different clustering methods, and assess the stability and robustness of the clusters.

clustering portion divided into three categories

- **internal cluster validation** uses internal information of the clustering process, e.g., the within-cluster sum of squares.
- **external cluster validation** compares results to externally known results, e.g., provided labels.
- **relative cluster validation** varies parameters of the clustering method, e.g., number of clusters

Internal Clustering Validation

Internal clustering validation, which use the internal information of the clustering process to evaluate the goodness of a clustering structure. **It can be also used for estimating the number of clusters and the appropriate clustering algorithm.**

The internal measures included in **clValid** package are:

1. **Connectivity** - what extent items are placed in the same cluster as their nearest neighbors in the data space. It has a value between 0 and infinity and should be **minimized**.
2. **Average Silhouette width** - It lies between -1 (poorly clustered observations) to 1 (well clustered observations). It should be **maximized**.
3. **Dunn index** - It is the ratio between the smallest distance between observations not in the same cluster to the largest intra-cluster distance. It has a value between 0 and infinity and should be **maximized**.

External Clustering Validation

External cluster validation uses ground truth information. That is, the user has an idea how the data should be grouped. This could be a know class label not provided to the clustering algorithm. Since we know the "true" cluster number in advance, **this approach is mainly used for selecting the right clustering algorithm for a specific dataset.**

The external cluster validation measures includes:

1. Corrected Rand Index
2. Variation of Information (VI)

The **Corrected Rand Index** provides a measure for assessing the similarity between two partitions, adjusted for chance. Its range is -1 (no agreement) to 1 (perfect agreement). **It should be maximized.**

The **Variation of Information** is a measure of the distance between two clustering (partitions of elements). It is closely related to mutual information. **It should be minimized.**

Relative Clustering Validation

Relative clustering validation, which evaluates the clustering structure by varying different parameter values for the same algorithm (e.g.,: varying the number of clusters k). **It's generally used for determining the optimal number of clusters.**

Silhouette Score

The Silhouette Score represents the within-cluster and between-cluster variation. For this, the Silhouette Score compares the mean distance *a* between a data point and all other data points in the same cluster with the mean distance *b* between a data point and all other data points in the next closest cluster. Then the ratio for each data point is averaged.

$$s = \frac{\sum_{i=1}^{N} \frac{b_i - a_i}{\max(a_i, b_i)}}{N}$$

The Silhouette Score can vary between -1 (incorrect clustering) and 1 (dense clustering). A score of 0 indicates overlapping clusters. Moreover, the Silhouette Score is higher for clusters which are dense and well separated.

Within-groups sum of squares This is also an internal measure but only measures compactness. It sums the squared Euclidean distance between each instance and the centroid of its cluster. From Equation (5.4) we know that the squared Euclidean distance between two instances *p* and *q* with *m* attributes each is given by:

$$sed(p, q) = \sum_{k=1}^{m} |p_k - q_k|^2$$

The within groups sum of squares is given by:

$$s = \sum_{i=1}^{K} \sum_{j=1}^{J_i} sed(p_j, C_i)$$

where K is the number of clusters and J_i is the number of instances of cluster i, and C_i is the centroid of cluster i.

Jaccard external measure This is a variation of a similar measure used in classi- fication tasks. It evaluates how uniform the distribution of the objects in each cluster is with respect to the class label. It uses the following equation:

$$J = (M_{11})/(M_{01} + M_{10} + M_{11})$$

where:

- M_{01} is the number of objects in other clusters but with the same label
- M_{10} is the number of objects in the same cluster, but with different labels
- M_{00} is the number of objects in other clusters with different labels
- M_{11} is the number of objects in the same cluster with the same label.