11. Clustering Evaluation

11.1 Introduction

Clustering evaluation refers to the task of figuring out how well the generated clusters are.

Assume a set of *N* samples, $\mathbf{x}_1$, $\mathbf{x}_2$, $\cdots$, $\mathbf{x}_N$, have been partitioned by a clustering algorithm into *K* disjoint subsets (clusters)

$$D_1, D_2, \cdots, D_K$$

with $n_1, n_2, \cdots, n_K$ samples, respectively, and

$$n_1 + n_2 + \cdots + n_K = N.$$

The following are some commonly used evaluation metrics for clustering:

(1) Silhouette coefficient
(2) Dunn index
(3) Davies-Bouldin index
(4) Calinski-Harabasz index
(5) Rand index

Details of each evaluation matric are introduced next.

11.2 Silhouette coefficient

The Silhouette coefficient is a metric that measures how well each sample fits into its assigned cluster. It combines information about both

(1) **Cohesion**: how close a sample is to other samples in its own cluster and the

(2) **Separation**: how far a sample is from samples in other clusters.

The Silhouette coefficient for a particular sample $\mathbf{x}_i$ is defined as follows:

$$SC_i = \frac{b_i - a_i}{max\{a_i, b_i\}}$$

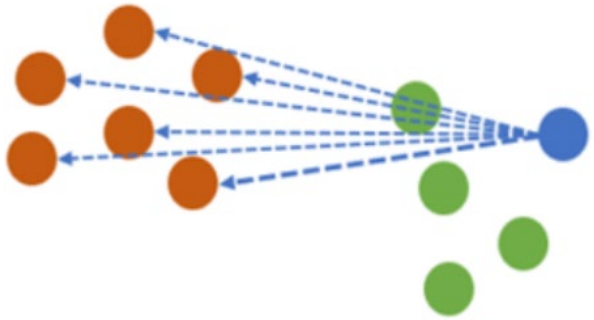
Where a_i is the average distance between $\mathbf{x}_i$ and all the other samples within the same cluster

$$a_i = \frac{1}{n_p - 1} \sum_{\mathbf{x}_i, \mathbf{x}_j \in D_p, i \neq j} d(i,j)$$

Where $d(i,j)$ denotes the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$. In Matlab, the default distance is squared Euclidean distance.

b_i is the minimum distance between $\mathbf{x}_i$ and all other samples belonging to other clusters

$$b_i = \min_{\mathbf{x}_i \in D_p, \mathbf{x}_j \in D_q, p \neq q} d(i,j)$$



Then the overall Silhouette coefficient is obtained as follows

$$SC = \frac{1}{N}\sum_{i=1}^{N} SC_i$$

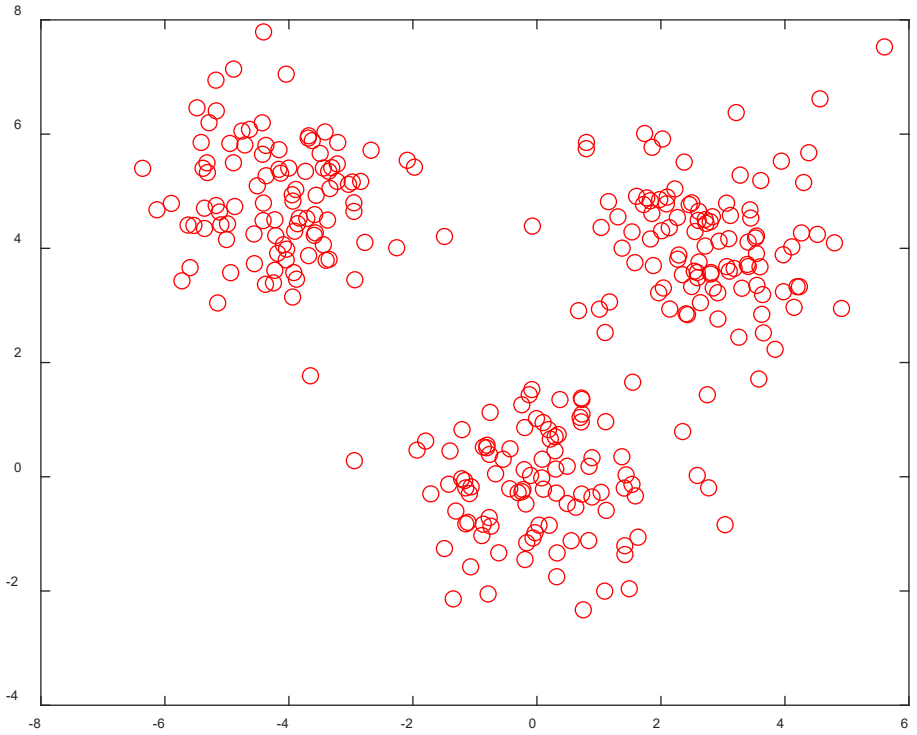
The Silhouette coefficient ranges from -1 to 1.

A higher Silhouette coefficient indicates that the samples are well-clustered, with clear separation between clusters and tight cohesion within each cluster.

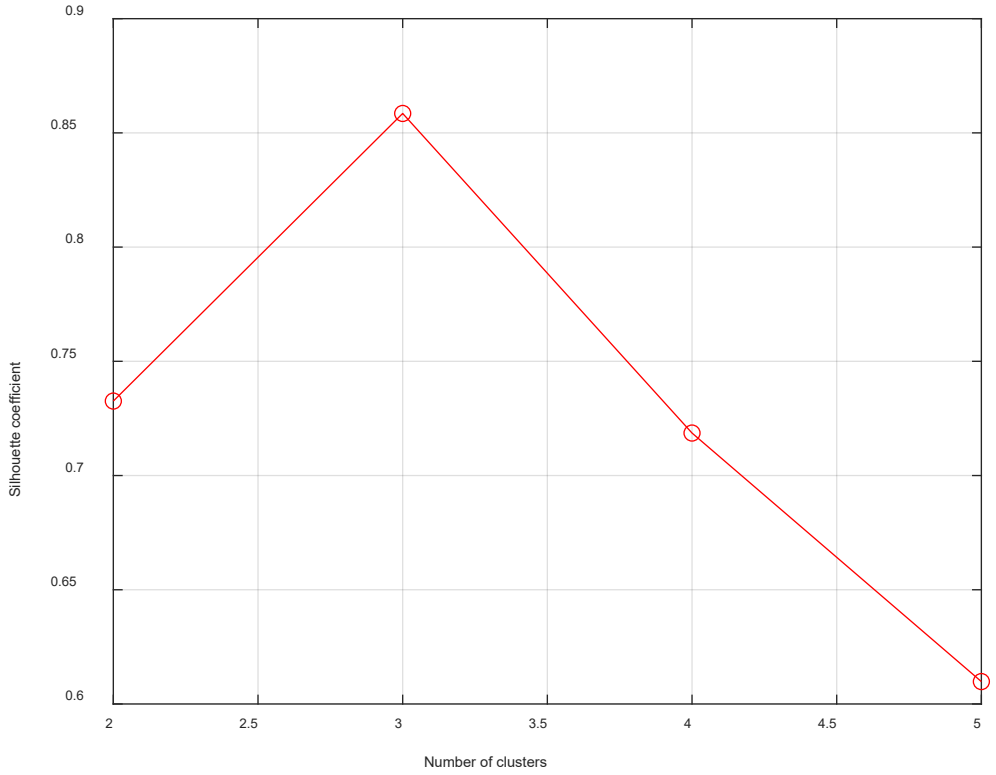
Conversely, a lower Silhouette coefficient suggests that the clustering may be less accurate, with overlapping clusters or samples that are not well-assigned to their respective clusters.

Example

There are 300 unlabelled data as shown below



If we set the number of cluster to 2, 3, 4, 5, respectively, and use K-mean clustering algorithm to cluster the 300 samples. We then calculate the Silhouette coefficient, and the result is shown below:



11.3 Dunn Index

The *Dunn index* aims at quantifying the compactness and variance of the clustering.

A cluster is considered *compact* if there is small variance between samples of the cluster. This can be quantified using

$$\delta_p = \max_{\mathbf{x}_i, \mathbf{x}_j \in D_p, i \neq j} d(i,j)$$

Where $d(i,j)$ denotes the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$.

Two clusters can be considered as *well separated* if the clusters are far-apart. This can be quantified using:

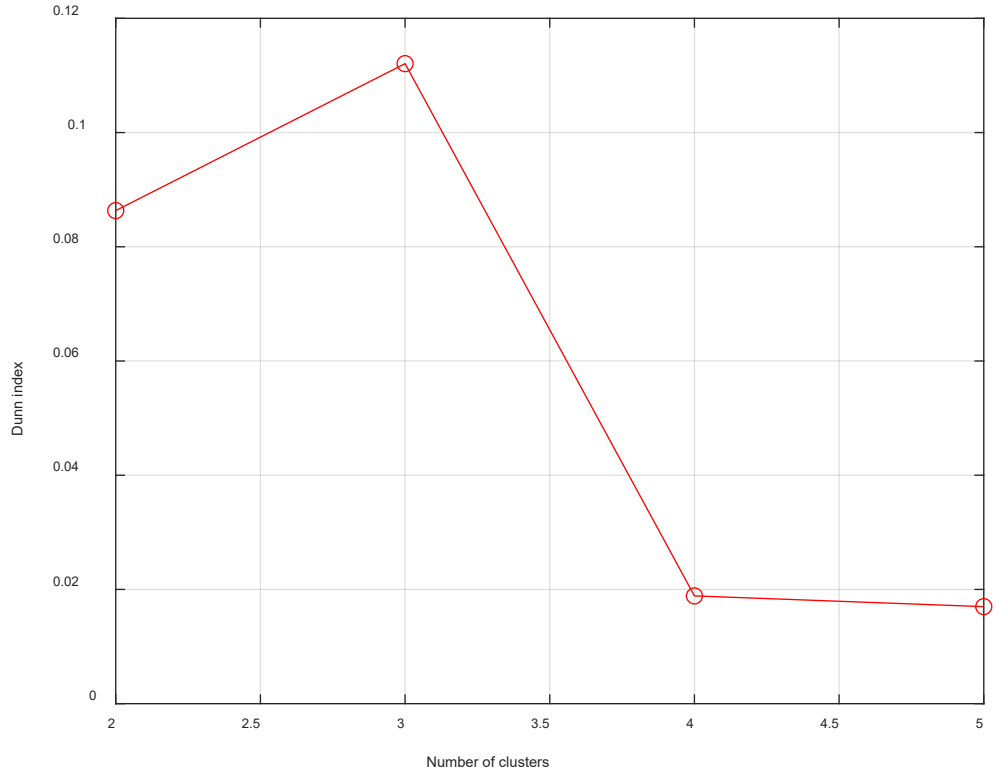
$$d_{p,q} = \min_{\mathbf{x}_i \in D_p, \mathbf{x}_j \in D_q, p \neq q} d(i,j)$$

Given these quantities, the *Dunn index* is then defined as:

$$DI = \frac{\min\limits_{p,q \in K} d_{pq}}{\max\limits_{p \in K} \delta_p}$$

A higher *Dunn Index* indicates compact, well-separated clusters, while a lower index indicates less compact or less well-separated clusters.

For the example used in Silhouette coefficient illustration, the Dunn index is shown below.



11.4 Davies-Bouldin index

Davies-Bouldin index also evaluates the goodness of clustering based on within-cluster dispersion and between-cluster separation.

For cluster D_i, the centroid (cluster centre) is as

$$\mathbf{m}_i = \frac{1}{n_i} \sum_{\mathbf{x} \in D_i} \mathbf{x}$$

Then the within-cluster (intra-cluster) dispersion is defined as:

$$\delta_i = \sqrt{\frac{1}{n_i} \sum_{\mathbf{x} \in D_i} \|\mathbf{x} - \mathbf{m}_i\|^2}$$

The separation between clusters D_i and D_j are defined as the distance between two centroids $\mathbf{m}_i$ and $\mathbf{m}_j$, and is denoted by $d(i,j)$.

Then the similarity between clusters D_i and D_j is defined as:

$$S_{ij} = \frac{\delta_i + \delta_j}{d(i,j)}$$

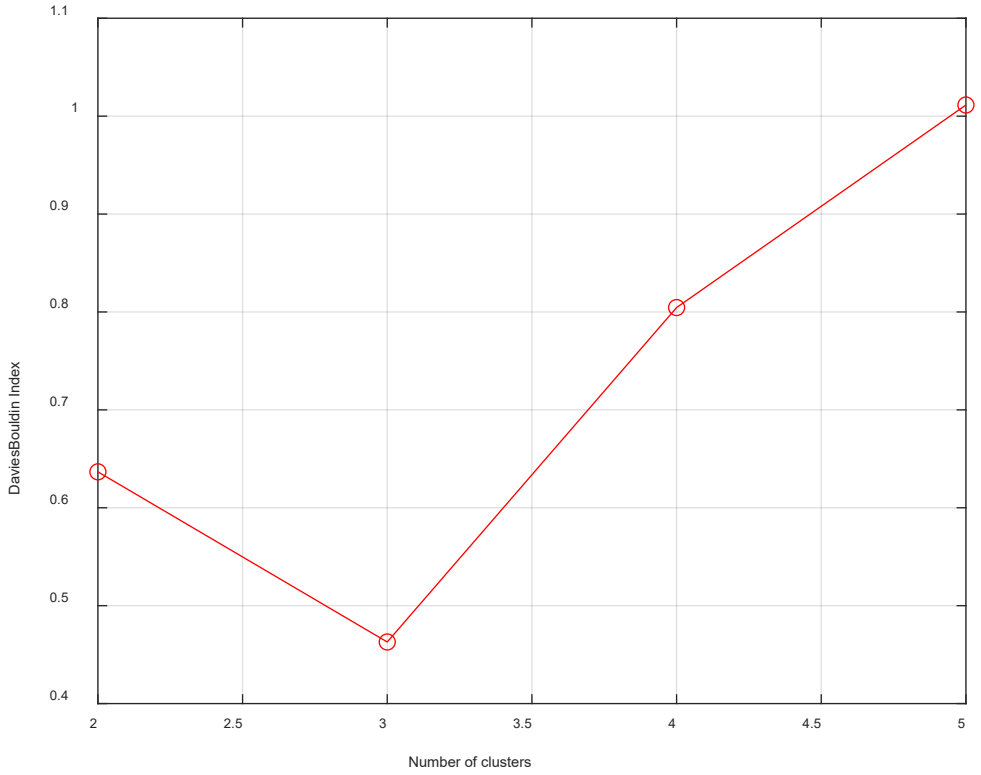
The similarity of the most similar cluster to cluster D_i is:

$$S_i = \max_{i \neq j} S_{ij}$$

The Davies-Bouldin index is defined as:

$$DB = \frac{1}{K} \sum_{i=1}^{K} S_i$$

The best choice for clusters is where the average similarity is minimized, therefore a smaller DB represents better defined clusters. For the previous example, the DB index is



11.5 Calinski-Harabasz Index

The Calinski-Harabasz (CH) Index is a measure of how similar a sample is to its own cluster (cohesion) compared to other clusters (separation).

Here the cohesion is estimated based on the distances from the samples in a cluster to the cluster centroid:

$$C_i = \sum_{\mathbf{x} \in D_i} \|\mathbf{x} - \mathbf{m}_i\|^2$$

The separation is estimated based on the distance of the cluster centroid from the global centroid:

$$S_i = \|\mathbf{m}_i - \mathbf{m}\|^2$$

Where $\mathbf{m}$ is the global centroid:

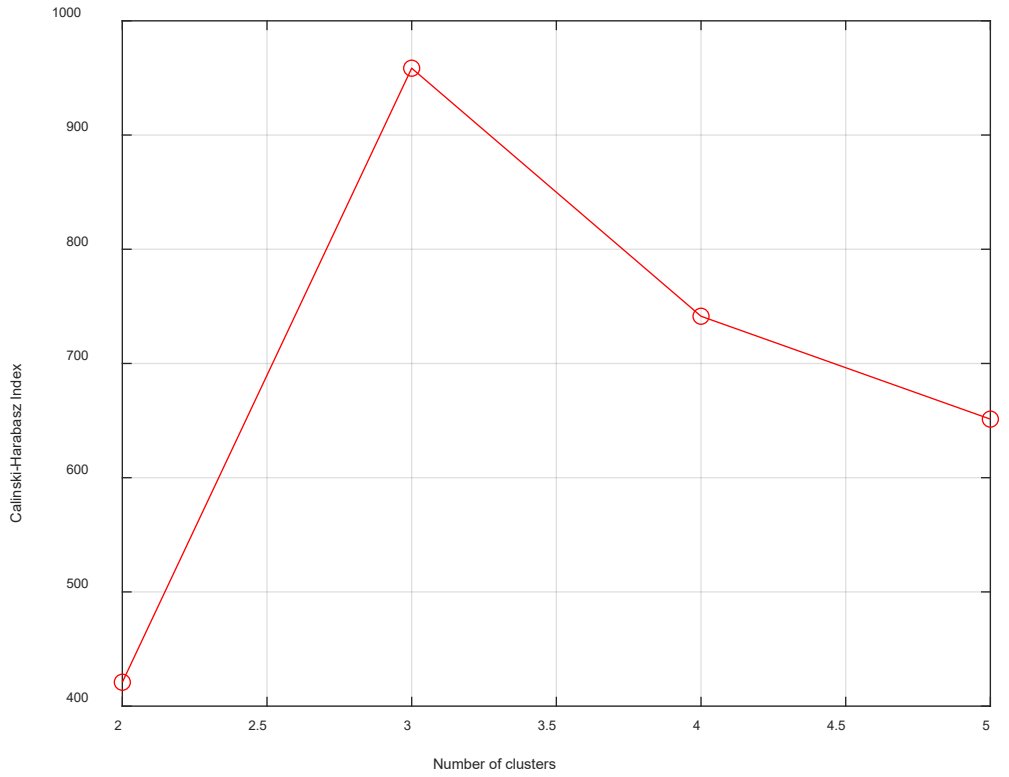
$$\mathbf{m} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i$$

Then the CH index is defined as:

$$CH = \frac{\sum_{i=1}^{K} n_i S_i}{\sum_{i=1}^{K} C_i}$$

Higher value of CH index means the clusters are dense and well separated

For the previous example, the Calinski-Harabasz (CH) Index is shown below



11.6 Rand Index

Another commonly used metric is the Rand Index. It computes a similarity measure between two clustering methods by considering all pairs of samples and counting pairs that are assigned in the same or different clusters across two different methods.

The formula of the Rand Index is:

$$RI = \frac{N_s + N_d}{c(N, 2)}$$

Where

N_s---the number of pairs of samples assigned to the same cluster across two clustering methods

N_d---the number of pairs of samples assigned to two different clusters across two clustering methods

$c(N, 2)$ ---the number of all possible pairs of samples

$$C(N, 2) = \frac{N!}{2!(N-2)!}$$

The Rand index always takes a value between 0 and 1.

- 0 indicates that two clustering methods do not agree on the clustering of any pair of data.
- 1 indicates that two clustering methods perfectly agree on the clustering of every pair of data