**Measures of Cluster Validity**

* Numerical measures that are applied to judge various aspects of cluster validity, are classified into the following three types.
  + **External Index**: Used to measure the extent to which cluster labels match externally supplied class labels.
    - Entropy
  + **Internal Index**: Used to measure the goodness of a clustering structure *without* respect to external information.
    - Sum of Squared Error (SSE)
    - [**Davies–Bouldin index**](https://en.wikipedia.org/wiki/Davies%E2%80%93Bouldin_index)
    - [**Dunn index**](https://en.wikipedia.org/wiki/Dunn_index)
    - [Silhouette coefficient](https://en.wikipedia.org/wiki/Silhouette_%28clustering%29)
  + **Relative Index**: Used to compare two different clusterings or clusters.
    - Often an external or internal index is used for this function, e.g., SSE or entropy
* Sometimes these are referred to as criteria instead of indices
  + However, sometimes criterion is the general strategy and index is the numerical measure that implements the criterion.

**Measuring Cluster Validity Via Correlation**

**Corr = -0.9235**

**Corr = -0.5810**

**Corr = -0.9235**

**Corr = -0.5810**

* Correlation of incidence and proximity matrices for the K-means clusterings of the following two data sets.
* The **Euclidean distance** between points **p** and **q** is the length of the [line segment](https://en.wikipedia.org/wiki/Line_segment) connecting them (\overline{\mathbf{p}\mathbf{q}}).
* In [Cartesian coordinates](https://en.wikipedia.org/wiki/Cartesian_coordinates), if **p** = (*p*1, *p*2,..., *pn*) and **q** = (*q*1, *q*2,..., *qn*) are two points in [Euclidean *n*-space](https://en.wikipedia.org/wiki/Euclidean_space), then the distance (d) from **p** to **q**, or from **q** to **p** is given by the [Pythagorean formula](https://en.wikipedia.org/wiki/Pythagorean_theorem):

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| \begin{align}\mathrm{d}(\mathbf{p},\mathbf{q}) = \mathrm{d}(\mathbf{q},\mathbf{p}) & = \sqrt{(q_1-p_1)^2 + (q_2-p_2)^2 + \cdots + (q_n-p_n)^2} \\[8pt] & = \sqrt{\sum_{i=1}^n (q_i-p_i)^2}.\end{align} | |  |  | | --- | --- | |  |  | |  | |

----------------------------------------------------------------------------------------------------------

The following methods can be used to assess the quality of clustering algorithms based on internal criterion:

* [**Davies–Bouldin index**](https://en.wikipedia.org/wiki/Davies%E2%80%93Bouldin_index)

The [Davies–Bouldin index](https://en.wikipedia.org/wiki/Davies%E2%80%93Bouldin_index) can be calculated by the following formula:


DB = \frac {1} {n} \sum_{i=1}^{n} \max_{j\neq i}\left(\frac{\sigma_i + \sigma_j} {d(c_i,c_j)}\right)


where n is the number of clusters, c_xis the [centroid](https://en.wikipedia.org/wiki/Centroid) of cluster x, \sigma_xis the average distance of all elements in cluster xto centroid c_x, and d(c_i,c_j)is the distance between centroids c_iand c_j. Since algorithms that produce clusters with low intra-cluster distances (high intra-cluster similarity) and high inter-cluster distances (low inter-cluster similarity) will have a low Davies–Bouldin index, the clustering algorithm that produces a collection of clusters with the smallest [Davies–Bouldin index](https://en.wikipedia.org/wiki/Davies%E2%80%93Bouldin_index) is considered the best algorithm based on this criterion.

Details:

Let *Ci* be a cluster of vectors. Let *Xj* be an n dimensional feature vector assigned to cluster *Ci*.

 S_i =  \frac{1}{T_i} \sum_{j=1}^{T_i} {\left|\left| X_j-A_i\right|\right|_p}

Here A_iis the [centroid](https://en.wikipedia.org/wiki/Centroid) of *Ci* and *Ti* is the size of the cluster *i*. *Si* is a measure of scatter within the cluster. Usually the value of *p* is 2, which makes this a [Euclidean distance](https://en.wikipedia.org/wiki/Euclidean_distance) function between the centroid of the cluster, and the individual feature vectors. Many other distance metrics can be used, in the case of [manifolds](https://en.wikipedia.org/wiki/Manifolds) and higher dimensional data, where the euclidean distance may not be the best measure for determining the clusters. It is important to note that this distance metric has to match with the metric used in the clustering scheme itself for meaningful results.

 M_{i,j} = \left|\left|A_i-A_j\right|\right|_p = \Bigl(\displaystyle\sum_{k=1}^{n}\left|a_{k,i}-a_{k,j}\right|^p\Bigr)^{\frac 1 p} 

 M_{i,j}is a measure of separation between cluster C_iand cluster C_j.

 a_{k,i} is the *k*th element of A_i, and there are n such elements in *A* for it is an n dimensional centroid.

Here *k* indexes the features of the data, and this is essentially the [Euclidean distance](https://en.wikipedia.org/wiki/Euclidean_distance) between the centers of clusters *i* and *j* when *p* equals 2.

Let *Ri,j* be a measure of how good the clustering scheme is. This measure, by definition has to account for *Mi,j* the separation between the *ith* and the *jth* cluster, which ideally has to be as large as possible, and *Si*, the within cluster scatter for cluster i, which has to be as low as possible. Hence the Davies–Bouldin index is defined as the ratio of *Si* and *Mi,j* such that these properties are conserved:

1.  R_{i,j} \geqslant 0 .
2.  R_{i,j} = R_{j,i} .
3. When  S_j \geqslant S_k and  M_{i,j} = M_{i,k} then  R_{i,j} > R_{i,k} .
4. When  S_j = S_k and  M_{i,j} \leqslant M_{i,k} then  R_{i,j} > R_{i,k} .

With this formulation, the lower the value, the better the separation of the clusters and the 'tightness' inside the clusters.

A solution that satisfies these properties is:

 R_{i,j} = \frac{S_i + S_j}{M_{i,j}} 

This is used to define *Di*:

 D_i \equiv \max_{j \neq i} R_{i,j}

If N is the number of clusters:

 \mathit{DB} \equiv \frac{1}{N}\displaystyle\sum_{i=1}^N D_i

*DB* is called the Davies–Bouldin index. This is dependent both on the data as well as the algorithm. *Di* chooses the worst-case scenario, and this value is equal to *Ri,j* for the most similar cluster to cluster *i*. There could be many variations to this formulation, like choosing the average of the cluster similarity, weighted average and so on

* [**Dunn index**](https://en.wikipedia.org/wiki/Dunn_index)

The Dunn index aims to identify dense and well-separated clusters. It is defined as the ratio between the minimal inter-cluster distance to maximal intra-cluster distance. For each cluster partition, the Dunn index can be calculated by the following formula:[[32]](https://en.wikipedia.org/wiki/Cluster_analysis" \l "cite_note-32)


D = \frac{\min_{1 \leq i < j \leq n} d(i,j)}{\max_{1 \leq k \leq n} d^{\prime}(k)} \,,

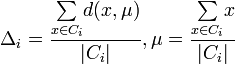

where *d*(*i*,*j*) represents the distance between clusters *i* and *j*, and *d* '(*k*) measures the intra-cluster distance of cluster *k*. The inter-cluster distance *d*(*i*,*j*) between two clusters may be any number of distance measures, such as the distance between the [centroids](https://en.wikipedia.org/wiki/Centroids) of the clusters. Similarly, the intra-cluster distance *d* '(*k*) may be measured in a variety ways, such as the maximal distance between any pair of elements in cluster *k*. Since internal criterion seek clusters with high intra-cluster similarity and low inter-cluster similarity, algorithms that produce clusters with high Dunn index are more desirable.

There are many ways to define the size or diameter of a cluster. It could be the distance between the farthest two points inside a cluster, it could be the mean of all the pairwise distances between data points inside the cluster, or it could as well be the distance of each data point from the cluster centroid. Each of these formulations are mathematically shown below:

Let *Ci* be a cluster of vectors. Let *x* and *y* be any two n dimensional feature vectors assigned to the same cluster *Ci*.

 \Delta_i =   \underset{x , y \in C_i}{\text{max}} d(x,y) , which calculates the maximum distance.

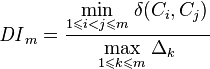
 \Delta_i =   \dfrac{1}{|C_i| |C_i - 1|} \underset{x , y \in C_i, x \neq y}{\sum} d(x,y) , which calculates the mean distance between all pairs.

, calculates distance of all the points from the mean.

This can also be said about the intercluster distance, where similar formulations can be made, using either the closest two data points, one in each cluster, or the farthest two, or the distance between the centroids and so on. The definition of the index includes any such formulation, and the family of indices so formed are called Dunn-like Indices. Let

 \delta(C_i,C_j) be this intercluster distance metric, between clusters *Ci* and *Cj*.

With the above notation, if there are *m* clusters, then the Dunn Index for the set is defined as:

.

* [Silhouette coefficient](https://en.wikipedia.org/wiki/Silhouette_%28clustering%29)

The silhouette coefficient contrasts the average distance to elements in the same cluster with the average distance to elements in other clusters. Objects with a high silhouette value are considered well clustered, objects with a low value may be outliers. This index works well with k-means clustering, and is also used to determine the optimal number of clusters.

Assume the data have been clustered via any technique, such as [k-means](https://en.wikipedia.org/wiki/K-means), into kclusters. For each [datum](https://en.wikipedia.org/wiki/Data) i, let a(i)be the average dissimilarity of iwith all other data within the same cluster. We can interpret a(i)as how well iis assigned to its cluster (the smaller the value, the better the assignment). We then define the average dissimilarity of point ito a cluster cas the average of the distance from ito all points in c.

Let b(i)be the lowest average dissimilarity of ito any other cluster, of which iis not a member. The cluster with this lowest average dissimilarity is said to be the "neighbouring cluster" of ibecause it is the next best fit cluster for point i. We now define a silhouette:

s(i) = \frac{b(i) - a(i)}{\max\{a(i),b(i)\}}


Which can be also written as:

s(i) = \begin{cases}
  1-a(i)/b(i), & \mbox{if } a(i) < b(i) \\
  0,  & \mbox{if } a(i) = b(i) \\
  b(i)/a(i)-1, & \mbox{if } a(i) > b(i) \\
\end{cases}


From the above definition it is clear that

 -1 \le s(i) \le 1


For s(i)to be close to 1 we require a(i) \ll b(i). As a(i)is a measure of how dissimilar iis to its own cluster, a small value means it is well matched. Furthermore, a large b(i)implies that iis badly matched to its neighbouring cluster. Thus an s(i)close to one means that the datum is appropriately clustered. If s(i)is close to negative one, then by the same logic we see that iwould be more appropriate if it was clustered in its neighbouring cluster. An s(i)near zero means that the datum is on the border of two natural clusters.

The average s(i)over all data of a cluster is a measure of how tightly grouped all the data in the cluster are. Thus the average s(i)over all data of the entire dataset is a measure of how appropriately the data has been clustered. If there are too many or too few clusters, as may occur when a poor choice of kis used in the clustering algorithm (e.g.: [k-means](https://en.wikipedia.org/wiki/K-means)), some of the clusters will typically display much narrower silhouettes than the rest. Thus silhouette plots and averages may be used to determine the natural number of clusters within a dataset. One can also increase the likelihood of the silhouette being maximized at the correct number of clusters by re-scaling the data using feature weights that are cluster specific.[[2]](https://en.wikipedia.org/wiki/Silhouette_%28clustering%29#cite_note-2)