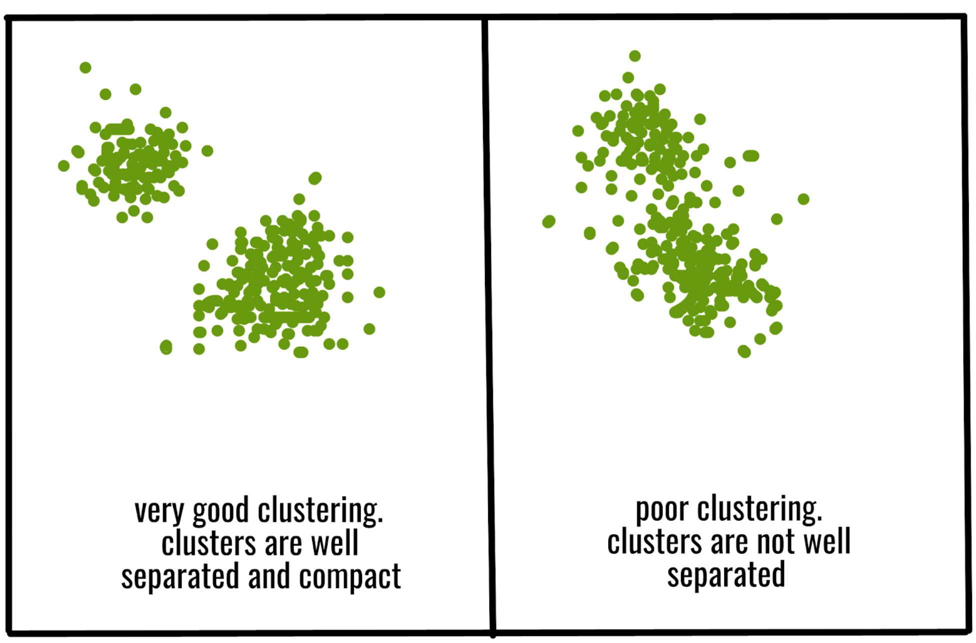
Davies Bouldin Index for Evaluation of Clusters

- DB is a **metric** for the evaluation of the CLUSTERING ALGORITHMS.
- The intuition behind it is the basic properties of clusters that are-
  1. Two different clusters must be as different as possible
  2. The data points in a particular cluster must be as similar as possible.
- DB index looks to minimize the **ratio** of the **intracluster distances to the inter cluster distances** for **each pair** of distinct clusters, on **average.**



- **Let us see how it is defined-**

Given n dimensional points, let C_i be a cluster of data points. Let X_j be an n-dimensional feature vector assigned to cluster C_i.

$$S_i = \left(\frac{1}{T_i}\sum_{j=1}^{T_i}|X_j - A_i|^p\right)^{1/p}$$

Here A_i is the centroid of C_i and T_i is the size of the cluster i. S_i is a measure of scatter within the cluster. Usu

- The above formula is just simply saying that **Si** is the intracluster distance for each cluster **Ci.**
- **P** is used as **2** which means mostly euclidean distance is used.

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

$M_{i,j}$ is a measure of separation between cluster C_i and cluster C_j.

- Again, **Mi,j** is just the **inter cluster distance** for a cluster pair i,j where i!=j.

Let $R_{i,j}$ be a measure of how good the clustering scheme is. This measure, by definition has to account for $M_{i,j}$ the separation between the i^{th} and the j^{th} cluster, which ideally has to be as large as possible, and S_i, 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 S_i and $M_{i,j}$ 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:

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

- Again, the idea that the properties and formulae reflect is the fact that **two different clusters must be as different as possible** and **the data points in a particular cluster must be as similar as possible.**
- **R** denotes the **measure of goodness** of the clustering.
- **Sj >= Sk line means** if there is a cluster j and a cluster k which are **equidistant from i** and **j is more scattered than k** then **Ri,j > Ri,k which means i,k is a better cluster pair than i,j.**
- **Similarly ,** the next line says that if the intracluster distance is the same between a cluster **i and clusters j and k** but the **intercluster is different** then the pair with the **smaller inter cluster distance would have** larger R as it would be a **worse option than the other one. This would be because it is nearer to i.**

This is used to define D_i:

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

If N is the number of clusters:

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

DB is called the Davies-Bouldin index.

- **Note that Di** chooses the **worst case scenario for the cluster i.**
- **From the definition of DB index , we observe that MINIMISING THE DAVIES-BOULDIN INDEX LEADS TO BETTER CLUSTERS.**

- **This index is thus defined as an average over all the *i* clusters, and hence a good measure of deciding how many clusters actually exists in the data is to plot it against the number of clusters it is calculated over. The number *i* for which this value is the lowest is a good measure of the number of clusters the data could be ideally classified into.**