CLUSTER VALIDATION

→ Tendency of a data point to become or form cluster.

→ Inshort we check randomness or non-randomness.

Unsupervised measure
- → Cluster cohension (compactness)/(tightness)
- → Cluster seperation (isolation)
  How well two datapoints are seperated from each other.

In Kmeans we have silhouette Coefficient to measure two points.

So Normally we do 3 types of validation.

1) Inter cluster validation → It uses the internal information of the clustering process to evaluate the goodness of a clustering structure without reference to external information. It can be also used for estimating the number of clusters and the appropiate clustering algorithim without any external data.

- Normally we observe the following
  - i) The objects in the same cluster are similar as much as possible.
  - ii) The objects in different cluster are highly distinct
- In short, we want the average distance within the cluster to be small as possible and the average distance between the cluster to be as large as possible.
- Methods → i) Silhouette coefficient ii) Dunn index.

2) External cluster validation → It consist in comparing the result of a cluster analysis to an externally known result, such as externally provided class labels. It measures the extent to which cluster labels match externally supplied class labels. Since we know the "true" cluster number in advance, this approach is mainly used for selecting the right clustering algorithim for a specific data set.

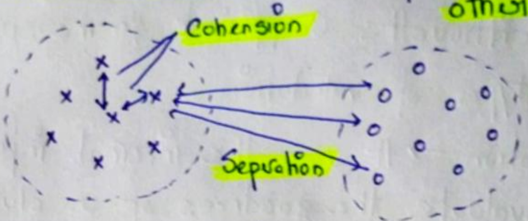
- Aim is to compare the identified clusters (by Kmeans/hierarchical/or any) clustering to an external reference.
- Method — Rand index varies from -1 (no agreement) to 1 (perfect agreement)

3) Connectivity → It corresponds to what extent items are placed in the same cluster as their nearest neighbour in the data space. The connectivity has a value between 0 and infinity and should be minimized.

Silhouette Coefficient

→

Cohension	Seperation
Measure the distance from one point to all other point in same cluster.	Measure of one point in one cluster to measure with all point in other cluster.



→ silhouette coefficient value ranges from [-1, 1], -1 → clustering is wrong

0 → Indifferent, clusters are same

1 → both clusters are faraway

Step 1 → create a distance matrix, i.e, euclidean $|x_i - y_i|^2$.

Step 2 → For each data point, x. calculate.

a) Cohension
Intra class dist
Same cluster.

b) Seperation
other Cluster.

Step 3 → $\boxed{\text{Silhouette coefficient} = \text{Seperation} - \text{Cohension}.}$

Many time we get -ve value, so we normalize $= \dfrac{\text{Seperation} - \text{Cohension}}{\text{Max (Seperation, cohension)}}$

or other method, $a < b$, $1 - \left(\frac{a}{b}\right)$

$a > b$, $\left(\frac{b}{a}\right) - 1$.

$a = b$, 0

- A value of Silhouette coefficient(s_i) close to 1 indicates that the objects are well clustered. In other word object i is similar to other object in the group.
- A value of s_i close to -1 indicates that the object is poorly clustered, the assignment to some other cluster would probably improve the result.