3. put an Edge between all core points that are within Eps
4. Make Each group of Connected Core points to Separate Cluster.
5. Assign Each Border point to one of Cluster of its Associated Core points.

* Strength & Weakness :- DBSCAN uses a density-based defn of a cluster
* It is relatively resistant to Noise & can handle Clusters of arbitary Shapes & Sizes
* DBSCAN has trouble when the Clusters have widely Varying densities.
* It also has trouble with high-dimensional data because density is more difficult to define for such data.
* DBSCAN can be Expensive when the computation of Nearest Neighbours requires computing all pairwise proxi-mities.

* Cluster Evaluation :- (Cluster Validation)

The Evaluation of the resulting classific-ation Model is an Integral part of the process of developing a classification Model & there are well-Separated Measures & procedures Ex "Accuracy & Cross-validation" Etc.

→ "Cluster Evaluation" Should be a part of any cluster Analysis, A Key Motivation is that almost Every clustering Algo will find clusters in a data set, Even if that data Set has no Natural Cluster Structure.

→ The Cluster Evaluation / Validity are traditionally classified into the following three types

* Unsupervised cluster Evaluation
* Supervised cluster Evaluation
* Relative cluster Evaluation

* Unsupervised cluster Evaluation :- Measure the goodness of a clustering structure without considering External info. Ex:- Sum of Squared Error (SSE)

→ Unsupervised Measures of cluster validity are often further divided into two classes

* Measure of cluster cohesion (compactness)
* Measure of cluster Separation (Isolation)

* Measure of cluster cohesion & Separation → "cluster cohesion" determines how closely related the objects in a cluster.

"cluster Separation" determines how distinct or well separated a cluster is from other clusters.

→ Cohesion & Separation for a cluster can be Expressed using the following Equations

$$\text{Cohesion } (C_i) = \sum_{\substack{x \in C_i \\ y \in C_i}} \text{proximity } (x, y)$$

$$\text{Separation } (C_i, C_j) = \sum_{\substack{x \in C_i \\ y \in C_j}} \text{proximity } (x, y)$$

→ In general, we can Express overall cluster validity for a set of K-clusters as a weighted sum of validity of individual clusters

$$\text{overall validity} = \sum_{i=1}^{k} w_i \text{ validity } (C_i)$$

The "validity" function can be cohesion, separation or some combination of these quantities.

* "Silhouette Co-efficient" → is a popular method which combines both cohesion & separation.
The following steps show how to compute it:
1) For ith-object, Calculate its Average distance to all other objects, call this value "a_i"
2) For ith object, Calculate its Average distance to all other objects & find the min such value, call this value "b_i"
3) Compute Silhouette Co-Efficient. $S_i = (b_i - a_i) / \text{Max}(a_i, b_i)$

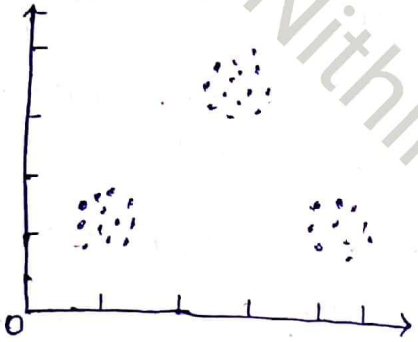
The value of the Silhouette Co-efficient can vary between -1 & 1.

* Measuring Cluster Validity Via Correlation → If we are given Similarity Matrix for a data set & the cluster labels from a cluster Analysis of data set, then we can Evaluate "goodness" of the clustering by looking at the correlation between Similarity Matrix & an ideal version of Similarity Matrix based on cluster labels.

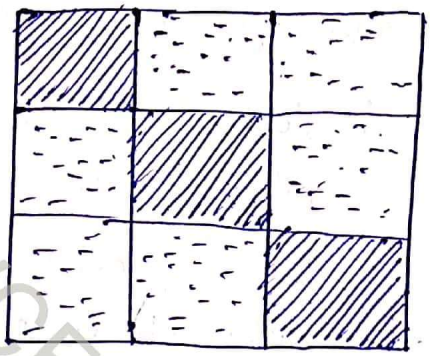
"1" to all points within the cluster & Similarity of "0" to all points in other clusters.
Thus if we sort the rows & columns of Similarity Matrix so that all objects belonging to same class together, then an Ideal Similarity Matrix has "block diagnol" structure

→ High Correlation b/w Ideal & Actual Similarity Matrices indicates that the points that belong to same cluster are close to each other
while low-Correlation indicates the opposite.

→ Consider the foll Example



(well-Separated clusters)



(Similarity Matrix Sorted)

If we have well-Separated clusters, then the Similarity Matrix should be "roughly block-diagnol"

→ This Approach may seem Hopelessly Expensive for large data sets, since the Computation of proximity Matrix takes $O(m^2)$ time.

* Supervised Measures of cluster Evaluation :- These Approaches Measure the Extent to which two objects that are in the same class are in the same cluster & Vice versa.

→ There are two types of Supervised Measures.
  * classification-Oriented
  * Similarity-Oriented

* Classification-oriented Measures → are a set of Measures such as "Entropy", "purity", "precision", "Recall" & "F-Measure" that are commonly used to Evaluate the performance of a classification Model.

→ "Entropy":- The degree to which Each cluster consists of objects of a single class.
The total Entropy for a set of clusters is calculated as the sum of Entropies of Each cluster weighted by the size of Each cluster i.e. $e = \sum_{i=1}^{k} \frac{m_i}{m} e_i$

→ "purity" :- Another Measure of the Extent to which a cluster contains objects of a single class
The overall purity of a clustering is $\text{purity} = \sum_{i=1}^{k} \frac{m_i}{m} p_i$

→ "precision" :- The fraction of a cluster that consists of objects of a specified class
The precision of cluster i w.r.t class j is $\text{precision}(i,j) = p_{ij}$

→ "Recall" :- The Extent to which a cluster contains all objects of a specified class.
The recall of cluster i w.r.t class j is $\text{recall}(i,j) = m_{ij}/m_j$

→ "F-Measure" :- A combination of both precision & Recall that measures the Extent to which a cluster contains only objects of a particular class & all objects of that class.
The F-Measure of cluster i w.r.t class j is

$F(i,j) = (2 \times \text{precision}(i,j) \times \text{recall}(i,j))/(\text{precision}(i,j) +$

* Similarity-oriented Measures → In this Approach, the Cluster Evaluation will be done using the Comparison of two Matrices

  * Ideal cluster Similarity Matrix
  * Ideal class Similarity Matrix.

"Ideal cluster Similarity Matrix", which has a "1" in the ijth Entry if two objects are in Same Cluster, otherwise "0".

"Ideal class Similarity Matrix", which has a "1" in the ijth Entry if two objects are in Same Class, otherwise "0".

* Relative Methods of cluster Evaluation :- Compares different Clusterings or Clusters. A relative cluster Evaluation is a Supervised or Unsupervised Evaluation Measure that is used for the purpose of Comparison.

→ Relative Measures are not actually a Separate type of Cluster Evaluation Measure, but are instead a Specific use of Such Measures.

Ex:- Two K-Means Clusterings can be Compared using Either "SSE" or "Entropy".

* Density-Based clustering :- Density Based Clustering Algorithm has played a vital role in finding non linear Shapes Structures based on the density.

→ "Density-Based Spatial Clustering of Application with Noise