⊛ Evaluating of clustering:-

Cluster validity:-

The validation of cluster created is a troublesome task.

→ The problem here is "cluster are in the eyes of the beholder".

A good cluster will have:-

- High Inter class similarity
- Low Inter class similarity.

Aspects of cluster validation:-

- **External:-** compare your cluster to ground truth.
- **Internal:-** Evaluating the cluster without reference to the external data.
- **Reliability:-** the clusters are not formed by chance (randomly) —
  Some statistical framework can be used

⊙External measures

→ No. of objects in the data P :- $P_1, P_2, \ldots P_m$

the set of ground truth clusters C :- $C_1, C_2 \ldots C_n$

the set of clusters formed by the algorithm.

the Incidence matrix - N * N matrix.

→ $P_{ij} = 1$, if the two points O_i & O_j belong to the same cluster in the ground truth
else, $P_{ij} = 0$.

→ $C_{ij} = 1$, if the two points O_i & O_j belong to the same cluster in the ground truth
else, $C_{ij} = 0$

Now there can be following scenario

1) $C_{ij} = P_{ij} = 1$ ⟶ Both the points belong to the same cluster for both our algorithm & ground truth (Agree) — SD

2) $C_{ij} = P_{ij} = 0$ ⟶ Both the points don't belong to same cluster for both our algorithm & ground truth (Agree) — DD

3) $C_{ij} = 1$ but $P_{ij} = 0$ → the points belong in the same cluster for our algorithm but different cluster in ground truth (Disagree) — SD

4) $C_{ij} = 0$ but $P_{ij} = 1$ → the points don't belong in same cluster for our Algo but same cluster in ground truth (disagree) — DS

↑ Just like accuracy score,

$$\text{Rand Index} = \frac{\text{total Agree}}{\text{total Disagree}} = \frac{(SS+DD)}{(SS+DD+DS+SD)}$$

→ the disadvantage of this is it can be dominated by DD.

$$\text{Jaccard coefficient} = \frac{SS}{(SS+DS+SD)}$$

∴ A higher value of Rand Index & Jaccard coefficient mean that the clusters generated by our algorithm mostly agree to the ground truth.

confusion matrix

n = no. of points
m_i = Points in cluster i
C_j = Points in class j
n_{ij} = Points in cluster i coming from class j.

$P_{ij} = \frac{n_{ij}}{m_i}$

(Probability of element from cluster i to be assigned to class j

	class 1	class 2	class 3	
cluster 1	n_{11}/P_{11}	n_{12}/P_{12}	n_{13}/P_{13}	m_1
cluster 2	n_{21}/P_{21}	n_{22}/P_{22}	n_{23}/P_{23}	m_2
cluster 3	n_{31}/P_{31}	n_{32}/P_{32}	n_{33}/P_{33}	m_3
	C_1	C_2	C_3	n

→ Entropy of cluster P_i,

$$e_i = -\sum P_{ij} (\log P_{ij})$$

→ For entering clustering algorithm, entropy can be given as

$$e = \sum \frac{m_i}{n} e_i$$

→ Purity of cluster i, $p_i = \max(p_{ij})$

And for entire cluster it is $P(C) = \sum \frac{m_i}{n} \cdot p_i$

→ Purity, is the total percentage of data points "clustered correctly".

→ A high value of purity score means that our clustering algorithm performs well against the ground truth.

Note:-

→ In calculating External measures, most of time we don't have ground truth,

---

⊙ External measures:-

These are the methods we do measure the quality of cluster without external reference. there are two aspects of it.

Cohesion:-

How closely the objects in the same cluster are related to each other. It is the within-cluster sum of squared distances. It is the same metric that we used to calculate for K-means Algorithm.

$$WCSS = \sum \sum (x - m_i)^2$$

Separation:-

How different objects in different clusters are and how distinct a well separated cluster is from other clusters.

It is the between cluster sum of squared distance

$$BSS = \sum c_i (m - m_i)^2$$

where, c is the size of individual cluster & m is centroid of all data points

Note:-

- BSS + WSS is always a constant

→ the silhouette can be calculated as:-

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

where,

$a(x)$ is avg. distance of x from all other point in same cluster.

$b(x)$ is avg. distance of x from all other points in other clusters.

And silhouette coefficient is,

$$SC = \frac{1}{N} \sum s(x)$$

Higher $s(x)$ means that the inter cluster similarity is less and the intra cluster dissimilarity is more (Good clustering)