CLUSTER EVALUATION METRICS

- Clustering is an unsupervised learning technique used to group similar data points together.
- Evaluating the performance of clustering algorithms can be challenging due to the absence of true labels.
- However, several metrics have been developed to assess clustering quality, both with and without ground truth labels.

1. Silhouette Score

- The Silhouette Score measures how similar an object is to its own cluster compared to other clusters.
- It ranges from -1 to 1, where a higher score indicates better-defined clusters.

Formula

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

a(i) is the average distance between the i-th data point and all other points in the same cluster.
b(i) is the average distance between the i-th data point and all points in the nearest cluster.

Advantages:

- Easy to interpret and understand.
- Does not require ground truth labels.
- Provides insight into cluster tightness and separation.

Disadvantages:

- Computationally expensive for large datasets.
- May not perform well with clusters of varying densities and shapes.
- Example:
- A Silhouette Score close to 1 indicates well-separated clusters, whereas a score close to -1 indicates overlapping clusters.

2. Davies-Bouldin Index

- The Davies-Bouldin Index (DBI) evaluates the average similarity ratio of each cluster with its most similar cluster.
- A lower DBI indicates better clustering.

Formula

$$DBI = \frac{1}{k}\sum_{i=1}^{k} \max_{j \neq i}\left(\frac{s_i + s_j}{d_{ij}}\right)$$

- s_i and s_j are the cluster dispersions for clusters i and j.
- d_{ij} is the distance between the centroids of clusters i and j.

Advantages:

- Considers both intra-cluster and inter-cluster distances.
- Lower values indicate better clustering.

Disadvantages:

- Sensitive to the number of clusters and their dispersion.
- May not perform well with non-spherical clusters.
- Example:
- A DBI of 0.5 indicates better-defined clusters compared to a DBI of 1.5.

3. Adjusted Rand Index (ARI)

- The Adjusted Rand Index (ARI) measures the similarity between the true cluster assignments and the clustering algorithm's assignments.
- It adjusts for chance grouping, making it more robust.

Formula

$$\text{ARI} = \frac{\text{RI} - \text{Expected RI}}{\text{Max RI} - \text{Expected RI}}$$

Advantages:

- Accounts for random chance.
- Useful for comparing clustering algorithms with ground truth labels.

Disadvantages:

- Requires ground truth labels.
- Computationally intensive for large datasets.

Example:
- An ARI of 0.85 indicates a high similarity between the predicted and true clusters.

4. Normalized Mutual Information (NMI)

- Normalized Mutual Information (NMI) measures the amount of shared information between the predicted clusters and the true clusters. It ranges from 0 to 1, where higher values indicate better clustering.

Formula

$$\mathrm{NMI}(U,V) = \frac{2 \times I(U;V)}{H(U)+H(V)}$$

- $I(U;V)$ is the mutual information between the clusters U and V.
- $H(U)$ and $H(V)$ are the entropies of the clusters U and V.

Advantages:
- Normalized, making it easier to compare different clustering results.
- Symmetric measure, not affected by the direction of comparison.

Disadvantages:
- Requires ground truth labels.
- Can be less intuitive to interpret.

Example:
- An NMI of 0.9 indicates a high degree of agreement between the predicted and true clusters.

5. Dunn Index

- The Dunn Index evaluates the ratio of the minimum inter-cluster distance to the maximum intra-cluster distance. Higher values indicate better clustering.

Formula

$$\text{Dunn Index} = \frac{\min_{1 \le i < j \le k} d(c_i, c_j)}{\max_{1 \le l \le k} d(c_l)}$$

- $d(c_i, c_j)$ is the distance between clusters i and j.
- $d(c_l)$ is the intra-cluster distance for cluster l.

Advantages:

- Useful for identifying compact and well-separated clusters.
- Simple and intuitive interpretation.

Disadvantages:

- Sensitive to noise and outliers.
- Computationally expensive for large datasets.

Example:

- A Dunn Index of 3.0 indicates well-separated and compact clusters compared to a Dunn Index of 1.0.

6. Inertia (Within-Cluster Sum of Squares)

- Inertia measures the sum of squared distances between data points and their respective cluster centroids. Lower inertia indicates better clustering.

Formula

$$\text{Inertia} = \sum_{i=1}^{k} \sum_{x_j \in C_i} \|x_j - \mu_i\|^2$$

- C_i is the i-th cluster.
- μ_i is the centroid of the i-th cluster.

Advantages:

- Simple to compute and understand.
- Useful for determining the optimal number of clusters using the elbow method.

Disadvantages:

- Sensitive to the shape and size of clusters.
- May not perform well with clusters of varying densities and shapes.

Example:

- Inertia can be used in the elbow method to identify the optimal number of clusters by finding the point where adding more clusters does not significantly reduce inertia