CLUSTER VALIDITY INDICES

WHY TO EVALUATE THE "GOODNESS" OF THE RESULTING CLUSTERS?

- To avoid finding patterns in noise
- To compare clustering algorithms
- To compare two sets of clusters
- To compare two clusters

DIFFERENT ASPECTS OF CLUSTER VALIDATION

1. Determining the clustering tendency of a set of data, i.e., distinguishing whether non-random structure actually exists in the data.
2. Comparing the results of a cluster analysis to externally known results, e.g., to externally given class labels.
3. Evaluating how well the results of a cluster analysis fit the data *without* reference to external information.
   - Use only the data
4. Comparing the results of two different sets of cluster analyses to determine which is better.
5. Determining the 'correct' number of clusters.

For 2, 3, and 4, we can further distinguish whether we want to evaluate the entire clustering or just individual clusters.

FRAMEWORK FOR CLUSTER VALIDITY

- Need a framework to interpret any measure.
  - For example, if our measure of evaluation has the value, 10, is that good, fair, or poor?
- Statistics provide a framework for cluster validity
  - The more "atypical" a clustering result is, the more likely it represents valid structure in the data
  - Can compare the values of an index that result from random data or clusterings to those of a clustering result.
    - If the value of the index is unlikely, then the cluster results are valid
  - These approaches are more complicated and harder to understand.
- For comparing the results of two different sets of cluster analyses, a framework is less necessary.
  - However, there is the question of whether the difference between two index values is significant

MEASURES OF CLUSTER VALIDITY

- Numerical measures that are applied to judge various aspects of cluster validity, are classified into the following three types.
  - External Index: Used to measure the extent to which cluster labels match externally supplied class labels.
    - Entropy
  - Internal Index: Used to measure the goodness of a clustering structure *without* respect to external information.
    - Sum of Squared Error (SSE)
  - Relative Index: Used to compare two different clusterings or clusters.
    - Often an external or internal index is used for this function, e.g., SSE or entropy
- Sometimes these are referred to as criteria instead of indices
  - However, sometimes criterion is the general strategy and index is the numerical measure that implements the criterion.

EXTERNAL MEASURES

- The correct or ground truth clustering is known priori.
- Given a clustering partition C and ground truth partitioning T, we redefine TP, TN, FP, FN in the context of clustering.
- Given the number of pairs N

$$N=TP+FP+FN+TN$$

EXTERNAL MEASURES …

- True Positives (TP): X_i and X_j are a true positive pair if they belong to the same partition in T, and they are also in the same cluster in C. TP is defined as the number of true positive pairs.
- False Negatives (FN): X_i and X_j are a false negative pair if they belong to the same partition in T, but they do not belong to the same cluster in C. FN is defined as the number of false negative pairs.
- • False Positives (FP): X_i and X_j are a false positive pair if the do not belong to the same partition in T, but belong to the same cluster in C. FP is the number of false positive pairs.
- True Negatives (TN): X_i and X_j are a false negative pair if they do not belong to the same partition in T, nor to the same cluster in C. TN is the number of true negative pairs.

JACCARD COEFFICIENT

- Measures the fraction of true positive point pairs, but after ignoring the true negatives as,

  $$\text{Jaccard} = TP/(TP+FP+FN)$$

- For a perfect clustering C, the coefficient is one, that is, there are no false positives nor false negatives.

- Note that the Jaccard coefficient is asymmetric in that it ignores the true negatives

RAND STATISTIC

- Measures the fraction of true positives and true negatives over all pairs as

$$\text{Rand} = (TP + TN)/ N$$

- The Rand statistic measures the fraction of point pairs where both the clustering C and the ground truth T agree.
- A perfect clustering has a value of 1 for the statistic.
- The adjusted rand index is the extension of the rand statistic corrected for chance.

FOWLKES-MALLOWS MEASURE

- Define precision and recall analogously to what done for classification,

  Prec = TP/ (TP+FP) and Recall = TP / (TP+FN)

- The Fowlkes–Mallows (FM) measure is defined as the geometric mean of the pairwise precision and recall

$$FM = \sqrt{(precision \cdot recall)}$$

- FM is also asymmetric in terms of the true positives and negatives because it ignores the true negatives. Its highest value is also 1, achieved when there are no false positives or negatives.

INTERNAL MEASURES: COHESION AND SEPARATION

- Cluster Cohesion (Compactness): Measures how closely related are objects in a cluster.
- Cluster Separation (Separation): Measure how distinct or well-separated a cluster is from other clusters.
- Example: Squared Error
  - Cohesion is measured by the within cluster sum of squares (SSE)

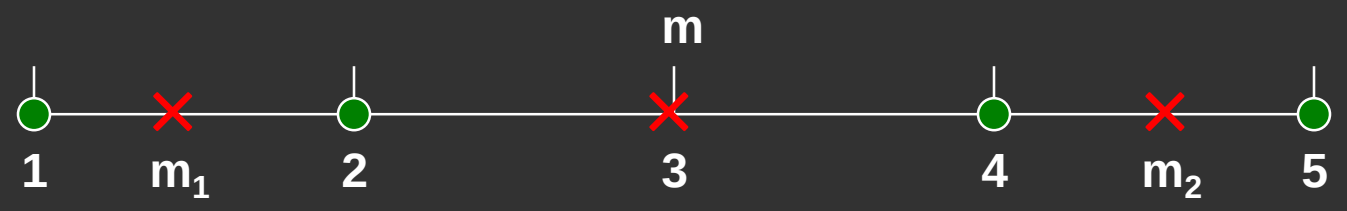
$$WSS = \sum_{i} \sum_{x \in C_i} (x - m_i)^2$$

  - Separation is measured by the between cluster sum of squares

$$BSS = \sum_{i} |C_i| (m - m_i)^2$$

    - Where $|C_i|$ is the size of cluster i

INTERNAL MEASURES: COHESION AND SEPARATION



K=1 cluster:

$$WSS = (1-3)^2 + (2-3)^2 + (4-3)^2 + (5-3)^2 = 10$$

$$BSS = 4 \times (3-3)^2 = 0$$

$$Total = 10 + 0 = 10$$

K=2 clusters:

$$WSS = (1-1.5)^2 + (2-1.5)^2 + (4-4.5)^2 + (5-4.5)^2 = 1$$

$$BSS = 2 \times (3-1.5)^2 + 2 \times (4.5-3)^2 = 9$$

$$Total = 1 + 9 = 10$$

INTERNAL MEASURES: COHESION AND SEPARATION

- Generally most of the indices used for internal clustering validation combine compactness and separation measures as follow:

- $\text{index} = \frac{(\alpha \times \text{Separation})}{(\beta \times \text{Compactness})}$

- Where α and β are weights.

CALINSKI-HARABAZ INDEX:

- Given k clusters, the Calinski-Harabaz score s is given by the ratio of the between-cluster dispersion mean and the within-cluster dispersion

$$CH = \frac{BSS(C)/(k-1)}{WSS(C)/(N-k)}$$

- The score is higher when clusters are dense and well separated, which relates to a standard concept of a cluster

SILHOUETTE COEFFICIENT

- A measure of how close each point in one cluster is to points in the neighboring clusters.

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

For each observation *i*, calculate the average dissimilarity $a(i)$ between *i* and all other points of the cluster to which *i* belongs.

For all other clusters C, to which *i* does not belong, calculate the average dissimilarity *d(i,C)* of *i* to all observations of C. The smallest of these *d(i,C)* is defined as $b(i) = min_C\ d(i, C)$. The value of $b(i)$ can be seen as the dissimilarity between *i* and its "neighbor" cluster, i.e., the nearest one to which it does not belong.

- Observations with a large Si (almost 1) are very well clustered.
- A small Si (around 0) means that the observation lies between two clusters.
- Observations with a negative Si are probably placed in the wrong cluster.

DUNN'S INDEX:

- If the data set contains compact and well-separated clusters, the diameter of the clusters is expected to be small and the distance between the clusters is expected to be large. Thus, Dunn index should be maximized.

$$Dunn's\ Index\ (D) = \frac{\text{min\{inter-cluster separation\}}}{\text{max\{intra-cluster distance or diameter\}}}$$

$$Dunn = \min_{1 \le i \le K} \min_{1 \le j \le K, j \ne i} \frac{\delta(C_i, C_j)}{\max_{1 \le k \le K} \triangle(C_k)}$$

Here,

$$\triangle(S) = \max_{\overline{x}, \overline{y} \in S} d(\overline{x}, \overline{y})$$

max{intra–cluster distance or diameter}

$$\delta(S, T) = \min_{\overline{x} \in S, \overline{y} \in T} d(\overline{x}, \overline{y}),$$

min{inter–cluster separation}

DAVIES–BOULDIN INDEX:

- A function of the ratio of the sum of *within-cluster scatter* to *between-cluster separation.*
- The objective is to minimize the DB index for achieving proper clustering.

$$DB = \frac{1}{K}\sum_{i=1}^{K} R_{i,q,t}$$

Here, $R_{i,q,t} = \max_{j,j\neq i} \frac{S_{i,q} + S_{j,q}}{d_{i,j,t}}.$

$S_{i,q}$ is the *qth* moment of the total points in cluster C_i *w.r.t. their mean* $\bar{c}_i$.

$$S_{i,q} = \left(\frac{\sum_{\overline{x}\in C_i}\{\|\overline{x} - \overline{c}_i\|_2^q\}}{|C_i|}\right)^{\frac{1}{q}}$$

$d_{i,j,t}$ *is the* Minkowski distance of order t between the centroids $\bar{c}_i$ *and* $\bar{c}_j$

$$d_{i,j,t} = \|\overline{c}_i - \overline{c}_j\|_t$$

XIE-BENI INDEX:

- In the definition of XB-index, the numerator indicates the compactness of the obtained cluster while the denominator indicates the strength of the separation between clusters.
- The objective is to minimize the XB-index for achieving proper clustering.

$$XB = \frac{\sum_{i=1}^{K} \sum_{j=1}^{n} \mu_{ij}^2 \|\overline{x}_j - \overline{c}_i\|^2}{n\left(\min_{i \neq k} \|\overline{c}_i - \overline{c}_k\|^2\right)}.$$

PS INDEX:

$$PS = \frac{1}{K}\sum_{i=1}^{K}[\frac{1}{n_i}\sum_{j\in S_i}\frac{d_c(\bar{x}_j,\bar{c}_i)}{min_{m,n=1,\ldots,K,m\neq n}(d_c(\bar{c}_m,\bar{c}_n))}]$$

Given partition of the data set X = xj : j = 1, 2, . . . , n and the center of each cluster ci(i = 1, . . . , K).

Here, Si is the set whose elements are the data points assigned to the ith cluster, ni is the number of elements in Si

- The most desirable partition is obtained by minimizing PS-index

I-INDEX:

- $$I = \left(\frac{1}{K} \times \frac{E_1}{E_k} \times D_k\right)^2$$

Here, E_1 is constant for given dataset.

$$E_k = \sum_{k=1}^{K} \sum_{i=1}^{n_k} \|\bar{x}_i - \bar{c}_k\|^2$$

$$D_k = max_{i,j=1}^{K} \left\|\bar{c}_i - \bar{c}_j\right\|^2$$

n_i is the number of points in cluster C_i

$\bar{c}_k$ is the centre of kth cluster.

The best partitioning occurs at the maximum value of the I-index.

CS-INDEX:

- Used for tackling clusters of different densities and/or sizes.

$$CS = \frac{\sum_{i=1} K[\frac{1}{N_i} \sum_{\overline{X}_i \in C_i} \max_{\overline{X}_q \in C_i} d(\overline{X}_i, \overline{X}_q)]}{\sum_{i=1}^{K} [\min_{j \in K, j \neq i} d(\overline{m}_i, \overline{m}_q)]}$$

where mi , i = 1, . . . K are the cluster centers.