**Cluster Validation Assignment**

The term cluster validation is used to design the procedure of evaluating the goodness of clustering algorithm results. This is important to avoid finding patterns in a random data, as well as, in the situation where you want to compare two clustering algorithms.

**Generally, clustering validation statistics can be categorized into 3 classes**:

Internal cluster validation, which 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 appropriate clustering algorithm without any external data.

1. External cluster validation, which consists in comparing the results 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 algorithm for a specific data set.
2. Relative cluster validation, which evaluates the clustering structure by varying different parameter values for the same algorithm (example: varying the number of clusters k). It’s generally used for determining the optimal number of clusters.

## **Internal measures for cluster validation:**

In this section, we describe the most widely used clustering validation indices. Recall that the goal of partitioning clustering algorithms is to split the data set into clusters of objects, such that:

* the objects in the same cluster are similar as much as possible,
* and the objects in different clusters are highly distinct

Internal validation measures reflect often the compactness, the connectedness and the separation of the cluster partitions.

1. Compactness or cluster cohesion: Measures how close are the objects within the same cluster. A lower within-cluster variation is an indicator of a good compactness (i.e., a good clustering). The different indices for evaluating the compactness of clusters are base on distance measures such as the cluster-wise within average/median distances between observations.
2. Separation: Measures how well-separated a cluster is from other clusters. The indices used as separation measures include:
   * distances between cluster centers
   * the pairwise minimum distances between objects in different clusters
3. Connectivity: corresponds to what extent items are placed in the same cluster as their nearest neighbors in the data space. The connectivity has a value between 0 and infinity and should be minimized.

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

***Index = (α×Separation) / (β×Compactness)***

Where α and β are weights

### 

### **Silhouette coefficient:**

The silhouette analysis measures how well an observation is clustered and it estimates the average distance between clusters. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters.

Finally, the silhouette width of the observation is defined by the formula:

***Si = (bi−ai)/max(ai,bi).***

**Dunn index:**

### The Dunn index is another internal clustering validation measure which can be computed as follow:

1. For each cluster, compute the distance between each of the objects in the cluster and the objects in the other clusters
2. Use the minimum of this pairwise distance as the inter-cluster separation (min.separation)
3. For each cluster, compute the distance between the objects in the same cluster.
4. Use the maximal intra-cluster distance (i.e maximum diameter) as the intra-cluster compactness
5. Calculate the Dunn index (D) as follow:

***D=min.separation / max.diameter***

## External measures for clustering validation

The aim is to compare the identified clusters (by k-means, pam or hierarchical clustering) to an external reference.