Clustering

It is basically a type of unsupervised learning method.

Clustering is the task of dividing the data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. The aim is to segregate groups with similar traits and assign them into clusters.

Clustering is used to find meaningful structure, explanatory underlying processes, generative features and groupings inherent in a set of examples.

Silhouette Analysis

The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation).

The silhouette ranges from −1 to +1 where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters.

The Silhouette Coefficient is calculated using the mean intra-cluster distance (a) and the mean nearest-cluster distance (b) for each sample. Let’s define Silhouette for each of the sample in the data set.

Let’s define a(i) to be the mean distance of point (i) w.r.t to all the other points in the cluster its assigned (A). We can interpret a(i) as how well the point is assigned to the cluster. Smaller the value better the assignment. Similarly let’s define b(i) to be the mean distance of point(i) w.r.t. to other points to its closet neighboring cluster (B). The cluster (B) is the cluster to which point (i) is not assigned to but its distance is closest amongst all other cluster.

The Silhouette Coefficient is defined as

S(i) = (b(i) – a(i)) / (max (a(i), b(i))

where

a(i): the average distance between 'i' and all other data within the same cluster

b(i): the lowest average distance of 'i' to all points in any other cluster of which 'i' is not a member

If silhouette value is close to 1, sample is well-clustered and already assigned to a very appropriate cluster. A value of +1 indicates that the sample is far away from its neighboring cluster and very close to the cluster its assigned

If silhouette value is about to 0, sample lies equally far away from both the clusters. A value of 0 means it’s at the boundary of the distance between the two cluster.

If silhouette value is close to -1, sample is misclassified and the sample is close to its neighboring cluster than to the cluster its assigned.

The Silhouette validation technique calculates the silhouette index for each sample, average silhouette index for each cluster and overall average silhouette index for a dataset. Using the approach each cluster could be represented by Silhouette index which is based on the comparison of its tightness and separation. Mean Silhouette score represents the Silhouette score of the entire cluster. This gives us one value representing the Silhouette score of the entire cluster.