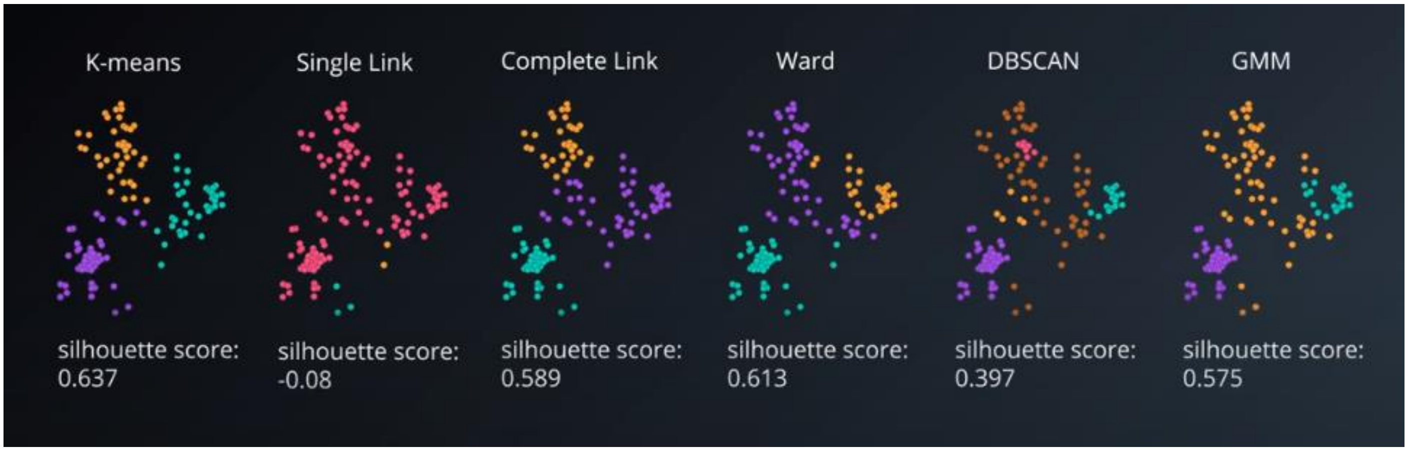
How to Evaluate the Performance of Clustering Algorithms Using Silhouette Coefficient

Silhouette Coefficient:

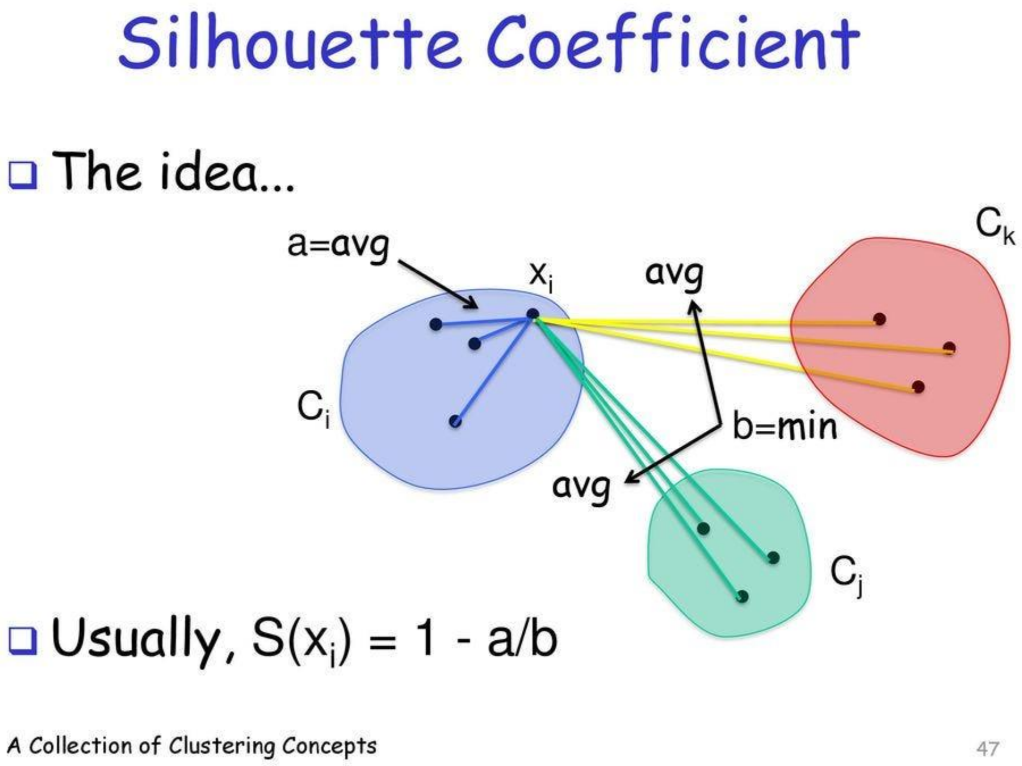
The silhouette value is a measure of how similar an object is to its own cluster (**cohesion**) compared to other clusters (**separation**). The Silhouette coefficient is a value between -1 and 1, where higher values indicate a better clustering. This index is especially useful for high-dimensional datasets where visualizing the clustering's is not possible. We can also calculate the silhouette coefficient for each point, values for individual points are calculated by averaging across clusters or an entire dataset.

Cluster Validation is the process by which clustering's are scored after they are executed. This provides a means of comparing different clustering algorithms and their results on a certain dataset.



Mathematical formulation :

Assume the data have been clustered via any technique, such as k-means, For data point i ∈ Ci (data point i in the cluster Ci), let



Silhouette Coefficient

❑ The idea...

❑ Usually, $S(x_i) = 1 - a/b$

A Collection of Clustering Concepts 47

be the mean distance between i and all other data points in the same cluster, where |Ci| is the number of points belonging to cluster i,

a(i) = avg distance of i to other point same cluster

$$a_i = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i, j)$$

We then define the mean dissimilarity of point i to some cluster Cj as the mean of the distance from i to all points in Cj

b(i) = avg distance to nearest other cluster

$$b_i = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i, j)$$

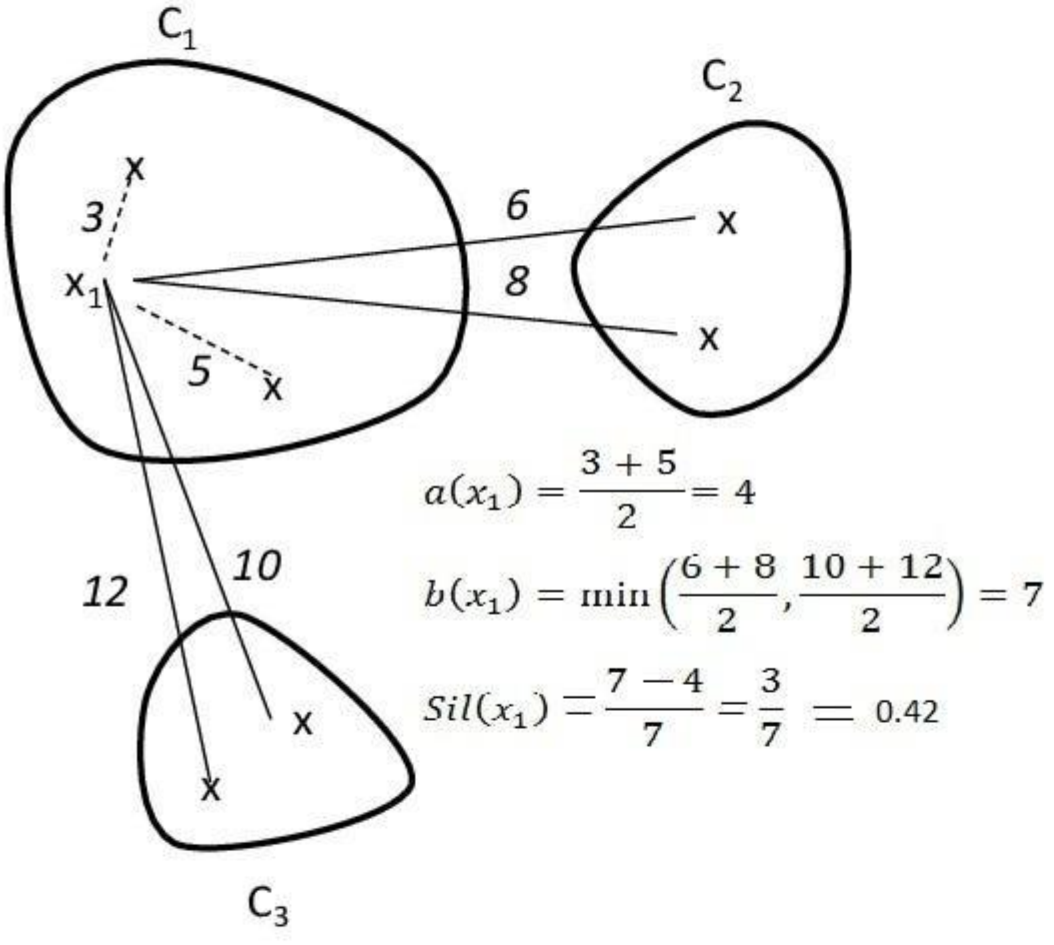
We now define a ***silhouette*** (value) of one data point i

S(i) = b(i) - a(i) / max{b(i), a(i)}

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

Silhouette coefficient

calculate silhouette score for toy dataset



$$a(x_1) = \frac{3+5}{2} = 4$$

$$b(x_1) = \min\left(\frac{6+8}{2}, \frac{10+12}{2}\right) = 7$$

$$Sil(x_1) = \frac{7-4}{7} = \frac{3}{7} = 0.42$$

Overall Silhouette score for the complete dataset can be calculated as the mean of silhouette score for all data points in the dataset. As can be seen from the formula ***silhouette score*** *would always lie between* ***-1 to 1*** *representing better clustering.*

where higher values indicate a better clustering

ideally , **S(i) = 1**

so, **S(i) < 0** indicates outliers

Silhouette Score like many other clustering evaluation metric is susceptible to error. Whenever its being used to quote algorithm performance one must be sure that the distance metric used in the algorithm is able to linearly separate the data.

In cases where the datasets are not linearly separable and the dimensions of the datasets is very high we must be careful while quoting silhouette distance.