



Silhouette Clustering

Silhouette Clustering involves using the Silhouette Score to evaluate the quality of clusters produced by a clustering algorithm. The Silhouette Score measures how well-separated clusters are and indicates the quality of clustering by quantifying the compactness and separation of the clusters.

Silhouette Score Formula:

The Silhouette Score for a single data point i is calculated as:

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

Where:

- $a(i)$ is the average distance from the i th data point to other data points in the same cluster. It measures the cohesion of the data point within its own cluster.
- $b(i)$ is the smallest average distance from the i th data point to data points in a different cluster, minimized over clusters. It measures the separation of the data point from other clusters.

Interpretation:

- The Silhouette Score ranges from -1 to 1.
- A score close to +1 indicates that the data point is well-clustered and lies far from neighboring clusters.
- A score close to 0 indicates that the data point is close to the decision boundary between two clusters.
- A score close to -1 indicates that the data point may have been assigned to the wrong cluster.

Example:

Let's say we have a dataset of customers with features like age, income, and spending score. We apply a clustering algorithm, let's say K-means, to cluster these customers into three segments.

After clustering, we calculate the Silhouette Score for each data point. For instance, let's consider a specific customer, and we find that:

- $a(i)=0.4$ (average distance to other points in the same cluster)
- $b(i)=0.6$ (smallest average distance to points in other clusters)

Using the formula, we compute the Silhouette Score:

$$s(i) = 0.6 - 0.4 \max\{0.4, 0.6\} = 0.2$$

So, this particular customer has a Silhouette Score of 0.2, indicating moderate clustering quality.

Conclusion:

Silhouette Clustering provides a way to assess the quality of clustering results objectively. Higher Silhouette Scores indicate better-defined clusters, helping in the selection of the optimal number of clusters and the evaluation of clustering algorithms' performance.