Silhouette **Coefficient**

Introduction

In the field of data analysis and machine learning, clustering plays a crucial role in identifying patterns and grouping similar data points together. While there are several clustering algorithms available, evaluating the quality of clustering results is equally important. One popular method for measuring the quality of clusters is the **silhouette coefficient**.

What is the Silhouette Coefficient?

The **silhouette coefficient** is a metric that measures how well each data point fits into its assigned cluster. It combines information about both the cohesion (how close a data point is to other points in its own cluster) and the separation (how far a data point is from points in other clusters) of the data point.

The coefficient ranges from -1 to 1.

Calculating the Silhouette Coefficient: Step-by-Step

1. For each data point, calculate two values:

— Average distance to all other data points within the same cluster (cohesion).

— Average distance to all data points in the nearest neighboring cluster (separation).

2. Compute the silhouette coefficient for each data point using the formula:

silhouette coefficient = (separation — cohesion) / max(separation, cohesion)

3. Calculate the average silhouette coefficient across all data points to obtain the overall silhouette score for the clustering result.

Example

To illustrate the calculation of the silhouette coefficient, let's consider a scenario with 3 clusters, where each cluster has 3 (2-dimensional points).

Cluster 1:

Point A1: (2, 5)
Point A2: (3, 4)
Point A3: (4, 6)

Cluster 2:

Point B1: (8, 3)
Point B2: (9, 2)
Point B3: (10, 5)

Cluster 3:

Point C1: (6, 10)
Point C2: (7, 8)
Point C3: (8, 9)