The Silhouette Coefficient: Evaluating Clustering Quality

Jerome Chou

May 15, 2025

What is the Silhouette Coefficient?

- A metric to evaluate the quality of clustering results.
- Measures how well each data point fits its assigned cluster compared to other clusters.
- Range: $[-1, 1]$
  - $+1$: Well-clustered, far from other clusters.
  - 0: Near cluster boundaries, ambiguous assignment.
  - -1: Likely misclustered, closer to another cluster.

Intuition: Balances intra-cluster cohesion and inter-cluster separation.

Mathematical Definition

For a data point i, the Silhouette Coefficient is:

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

where:

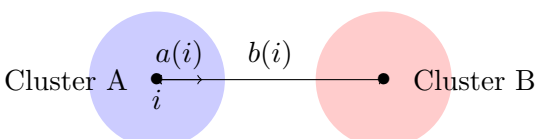
- $a(i)$: Mean distance to all other points in the same cluster (cohesion).
- $b(i)$: Mean distance to all points in the nearest neighboring cluster (separation).

Intuition: $s(i)$ compares how close i is to its own cluster versus the next closest cluster.

Computation Steps

1. Compute $a(i)$: Average distance from point i to all points in its cluster.
2. Compute $b(i)$: Average distance from point i to all points in the nearest other cluster.
3. Calculate $s(i) = \frac{b(i)-a(i)}{\max(a(i),b(i))}$ for each point.
4. Average $s(i)$ across all points for the overall coefficient.

Visual Aid:



Applications

- **Cluster Validation**: Assesses how cohesive and separated clusters are.
- **Optimal Cluster Number**: Choose k that maximizes the average Silhouette Coefficient.
- **Algorithm Comparison**: Compare different clustering methods or parameters.

Why it matters: Provides an intuitive, standardized metric for clustering quality.

Logical Comparison with Other Metrics

- **Versus Within-Cluster Sum of Squares (WCSS)**:
  - WCSS only measures intra-cluster cohesion.
  - Silhouette also considers inter-cluster separation.
- **Versus Davies-Bouldin Index**:
  - Davies-Bouldin focuses on cluster centroids and scatter.
  - Silhouette evaluates individual point assignments.

Advantage: Silhouette's per-point analysis offers fine-grained insights.

Takeaways

- The Silhouette Coefficient measures clustering quality via cohesion (a) and separation (b).
- Formula: $s(i) = \frac{b(i)-a(i)}{\max(a(i),b(i))}$, averaged over all points.
- Values near $+1$ indicate good clustering; near -1 suggest misclustering.
- Useful for validating clusters, choosing k, and comparing algorithms.
- Intuitive and standardized, but sensitive to distance metrics.