Silhouette Coefficient

Silhouette Coefficient (SC) is an evaluation metric for clustering when ground truth labels are unknown. It refers to how similar an object is to its own cluster (cohesion) compared to other clusters (separation).

Silhouette Coefficient for object n is defined as

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

where $a(n)$ and $b(n)$ are in turn defined as follows:

- $a(n)$: the mean distance between object n and all other objects in the same cluster (i.e. the distance to the cluster mean).
- $b(n)$: The mean distance between a object n and all other objects in the next nearest cluster (i.e. the minimum distance to other cluster means).

If $s(n)$ is close to 1, we have $a(n) << b(n)$. This means a small intra-cluster distance and a large nearest-cluster distance, corresponding to a proper cluster. Vice versa, if $s(n)$ close to -1, we have $a(n) >> b(n)$. This means a large intra-cluster distance and a small nearest-cluster distance, corresponding to a poor cluster.

Normally, the Silhouette Coefficient for a clustering is calculated as by averaging the coefficients for each of the objects:

$$SC = \frac{1}{N} \sum_{n=1}^{N} s(n)$$

References:

- Rousseeuw P J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis[J]. Journal of computational and applied mathematics, 1987, 20: 53-65.
- `https://en.wikipedia.org/wiki/Silhouette_(clustering)`
- `http://scikit-learn.org/stable/modules/clustering.html`