Silhouette coefficient

Definition and purpose

Silhouette refers to a method of interpretation and validation of consistency within clusters of data. The technique provides a succinct graphical representation of how well each object has been classified.

Assume the data has been clustered into k groups. For data point $i \in C_I$, let

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

Which referred the **intracluster distance**. And then we define the mean dissimilarity (**intercluster distance**)

$$b(i) = \min_{J \neq I} \frac{1}{|C_J|} \sum_{j \in C_J} d(i,j)$$

And

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

For $s(i)$ to be close to 1 we require $a(i) \ll b(i)$. **As $\boldsymbol{a(i)}$ is a measure of how dissimilar $\boldsymbol{i}$ is to its own cluster**, a small value means it is well matched. Furthermore, a large $b(i)$ implies that i is badly matched to its neighbouring cluster. Thus **an $\boldsymbol{s(i)}$ close to 1 means that the data is appropriately clustered**. **If $\boldsymbol{s(i)}$ is close to -1, then by the same logic we see that $\boldsymbol{i}$ would be more appropriate if it was clustered in its neighboring cluster**. **An $\boldsymbol{s(i)}$ near zero means that the datum is on the border of two natural clusters**.