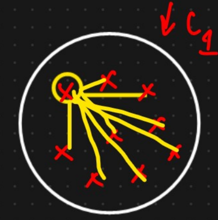
Silhouette (Clustering)

(1)

a(i)



For data point $i \in C_I$ (data point i in the cluster C_I), let

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

be the mean distance between i and all other data points in the same cluster, where $|C_I|$ is the number of points belonging to cluster i, and $d(i, j)$ is the distance between data points i and j in the cluster C_I (we divide by $|C_I| - 1$ because we do not include the distance $d(i, i)$ in the sum). We can interpret $a(i)$ as a measure of how well i is assigned to its cluster (the smaller the value, the better the assignment).

a(i)

(2)

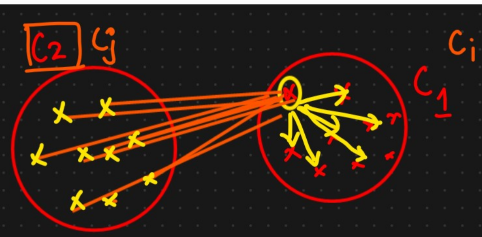
We then define the mean dissimilarity of point i to some cluster C_J as the mean of the distance from i to all points in C_J (where $C_J \neq C_I$).

For each data point $i \in C_I$, we now define

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

$\boxed{a(i) \ll b(i)}$

to be the *smallest* (hence the $\min$ operator in the formula) mean distance of i to all points in any other cluster, of which i is not a member. The cluster with this smallest mean dissimilarity is said to be the "neighboring cluster" of i because it is the next best fit cluster for point i.



a(i)

③ Silhoutte Score

We now define a *silhouette* (value) of one data point i

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \text{ if } |C_I| > 1$$

$\boxed{-1 \text{ to } 1}$

More near to 1 better

Clustering model we have created

and

$$s(i) = 0, \text{ if } |C_I| = 1$$

Which can be also written as:

+ve. ≈ 1

$$s(i) = \begin{cases} 1 - a(i)/b(i), & \text{if } a(i) < b(i) \\ 0, & \text{if } a(i) = b(i) \\ b(i)/a(i) - 1, & \text{if } a(i) > b(i) \end{cases}$$

≈ -1

From the above definition it is clear that

$$\{ -1 \le s(i) \le 1 \}$$