Silhouettes

The *silhouette method* is an unsupervised method for evaluation of clusterings that computes certain coefficients for each object. The set of these coefficients allows an evaluation of the quality of the clustering.

Let $O = \{o_1, \ldots, o_n\}$ be a collection of objects, $d : O \times O \longrightarrow \mathbb{R}_+$ a dissimilarity on O, and let $f : O \longrightarrow \{C_1, \ldots, C_k\}$ be a clustering function, that is a function such that $f(o) = C$ if o is in the cluster C. Since the clusters mutually disjoint and exhaustive, f is well-defined.

Suppose that $f(o_i) = C_\ell$. The (f, d)-*average dissimilarity* is the function $a_{f,d} : O \longrightarrow \mathbb{R}$ given by

$$a_{f,d}(o_i) = \frac{\sum\{d(o_i, u) \mid f(u) = f(o_i) \text{ and } u \neq o_i\}}{|f(o_i)|},$$

that is, the average dissimilarity of o_i to all objects of $f(o_i)$, the cluster to which o_i is assigned.

For a cluster C and an object o_i let

$$d(o_i, C) = \frac{\sum\{d(o_i, u) \mid f(u) = C\}}{|C|},$$

be the average dissimilarity between o_i and the objects of the cluster C.

Let $f : O \longrightarrow \{C_1, \ldots, C_k\}$ be a clustering function A *neighbor of* o_i is a cluster $C \neq f(o_i)$ for which $d(o_i, C)$ is minimal.

In other words, a neighbor of an object o_i is "the second best choice" for a cluster for o_i. Let $b : O \longrightarrow \mathbb{R}$ be the function defined by

$$b_{f,d}(o_i) = \min\{d(o_i, C) \mid C \neq f(o_i)\}.$$

If f and d are clear from the context, we shall simply write $a(o_i)$ and $b(o_i)$ instead of $a_{f,d}(o_i)$ and $b_{f,d}(o_i)$, respectively.

The *silhouette* of the object o_i for which $|f(o_i)| \geq 2$ is the number $\mathsf{sil}(o_i)$ given by

$$\mathsf{sil}(o_i) = \frac{b(o_i) - a(o_i)}{\max\{a(o_i), b(o_i)\}}$$

for $o_i \in O$.

It is easy to see that

$$\mathsf{sil}(o_i) = \begin{cases} 1 - \frac{a(o_i)}{b(o_i)} & \text{if } a(o_i) < b(o_i) \\ 0 & \text{if } a(o_i) = b(o_i) \\ \frac{b(o_i)}{a(o_i)} - 1 & \text{if } a(o_i) > b(o_i). \end{cases}$$

If $f(o_i) = 1$, then $s(o_i) = 0$.

Observe that $-1 \leq \mathsf{sil}(o_i) \leq 1$. When $\mathsf{sil}(o_i)$ is close to 1, this means that $a(o_i)$ is much smaller than $b(o_i)$ and we may conclude that o_i is well-classified. When $\mathsf{sil}(o_i)$ is near 0, it is not clear which is the best cluster for o_i. Finally, if $\mathsf{sil}(o_i)$ is close to -1, the average distance from u to its neighbor(s) is much smaller than the average distance between o_i and other objects that belong to the same cluster $f(o_i)$. In this case, it is clear that o_i is poorly classified.

The *average silhouette width of a cluster* C is

$$\mathsf{sil}(C) = \frac{\sum\{\mathsf{sil}(o) \mid o \in C\}}{|C|}.$$

The *average silhouette width of a clustering* κ is

$$\mathsf{sil}(\kappa) = \frac{\sum\{\mathsf{sil}(o) \mid o \in O\}}{|O|}.$$

The silhouette of a clustering can be used for determining the "optimal" number of clusters. If the average silhouette of the clustering is above 0.7, we have a strong clustering.