Machine Learning

Clustering

Fabio Vandin

December 22^{nd}, 2023

Choice of number k of clusters

Choosing the number k of clusters (e.g., for k-means) is not easy.

Common approach:

1. run clustering algorithm for various values of k, obtaining a clustering $C^{(k)} = \{C_1^{(k)}, C_2^{(k)}, \ldots, C_k^{(k)}\}$ for each value of k considered;
2. use a *score* S to evaluate each clustering $C^{(k)}$, getting scores $S(C^{(k)})$ for each value of k
3. pick the value of k (and clustering) of maximum score: $C = \arg\max_{C^{(k)}}\{S(C^{(k)})\}$

A very common score based on distances alone: *silhouette*

Silhouette

Given a clustering $C = (C_1, C_2, \ldots, C_k)$ of $\mathcal{X}$ and a point $\mathbf{x} \in \mathcal{X}$, let $C(\mathbf{x})$ be the cluster to which $\mathbf{x}$ is assigned to. Assume $|C_i| \geq 2 \; \forall \; 1 \leq i \leq k$. Define:

$$A(\mathbf{x}) = \frac{\sum_{\mathbf{x}' \neq \mathbf{x}, \mathbf{x}' \in C(\mathbf{x})} d(\mathbf{x}, \mathbf{x}')}{|C(\mathbf{x})| - 1}$$

Given a cluster $C_i \neq C(\mathbf{x})$, let

$$d(\mathbf{x}, C_i) = \frac{\sum_{\mathbf{x}' \in C_i} d(\mathbf{x}, \mathbf{x}')}{|C_i|}$$

and $B(\mathbf{x}) = \min_{C_i \neq C(\mathbf{x})} d(\mathbf{x}, C_i)$.

Then the *silhouette* $s(\mathbf{x})$ of $\mathbf{x}$ is

$$s(\mathbf{x}) = \frac{B(\mathbf{x}) - A(\mathbf{x})}{\max\{A(\mathbf{x}), B(\mathbf{x})\}}$$

Intuition: $s(\mathbf{x})$ measures if $\mathbf{x}$ is closer to points in its "nearest cluster" than to the cluster it is assigned to.

Question: what is the range for $s(\mathbf{x})$? $[-1, 1]$

The silhouette of clustering $C = (C_1, C_2, \ldots, C_k)$ is

$$S(C) = \frac{\sum_{\mathbf{x} \in \mathcal{X}} s(\mathbf{x})}{|\mathbf{X}|}$$

The higher $S(C)$, the better the clustering quality.