How do you measure the accuracy of a Clustering Algorithm?

Supervised ML → Accuracy, Precision, recall, Roc-Auc, F-Beta Score

Clustering ML → Silhouette Score

Silhouette Score

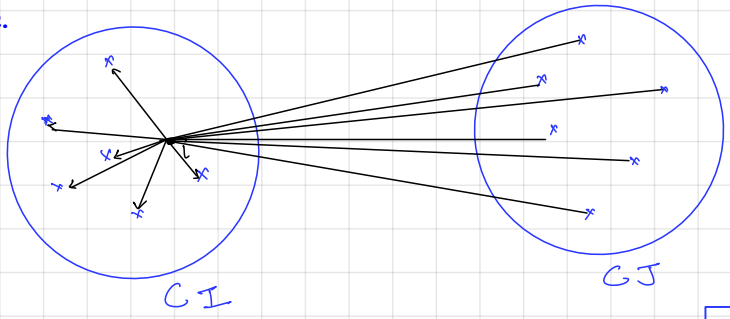
Silhouette refers to a method of interpretation and validation of consistency within clusters of data.

The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation).

The silhouette ranges from −1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters.

If most objects have a high value, then the clustering configuration is appropriate. If many points have a low or negative value, then the clustering configuration may have too many or too few clusters

The silhouette can be calculated with any distance metric, such as the Euclidean distance or the Manhattan distance.



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

$a(i)$ → mean dist b/w i and all other data pt in same cluster

$|C_I| - 1$ → no. of datapt in cluster C_I

→ measure how well i is assigned to its cluster (smaller value better assignment)

$$b(i) = \min_{J \neq I} \frac{1}{|C_J|} \sum_{j \in C_J} d(i,j) \quad \text{for each point } i,\ i \in C_I$$

→ Smallest mean distance of i to all point in any other cluster

Silhouette Score for one datapt (i).

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

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

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

$\therefore \quad -1 \leq S(i) \leq 1$