2- Model evaluation 3 - Clustering Models

By: eng. Esraa Madhi

In Supervised Learning, the labels are known and evaluation can be done by calculating the degree of correctness by comparing the predicted values against the labels. However, in Unsupervised Learning, the labels are not known, which makes it hard to evaluate the degree of correctness as there is **no ground truth**.

That being said, it is still consistent that a *good* clustering algorithm has clusters that have **small within-cluster variance** (data points in a cluster are similar to each other) and **large between-cluster variance** (clusters are dissimilar to other clusters).

Silhouette Coefficient

Silhouette Coefficient measures the between-cluster distance against within-cluster distance.

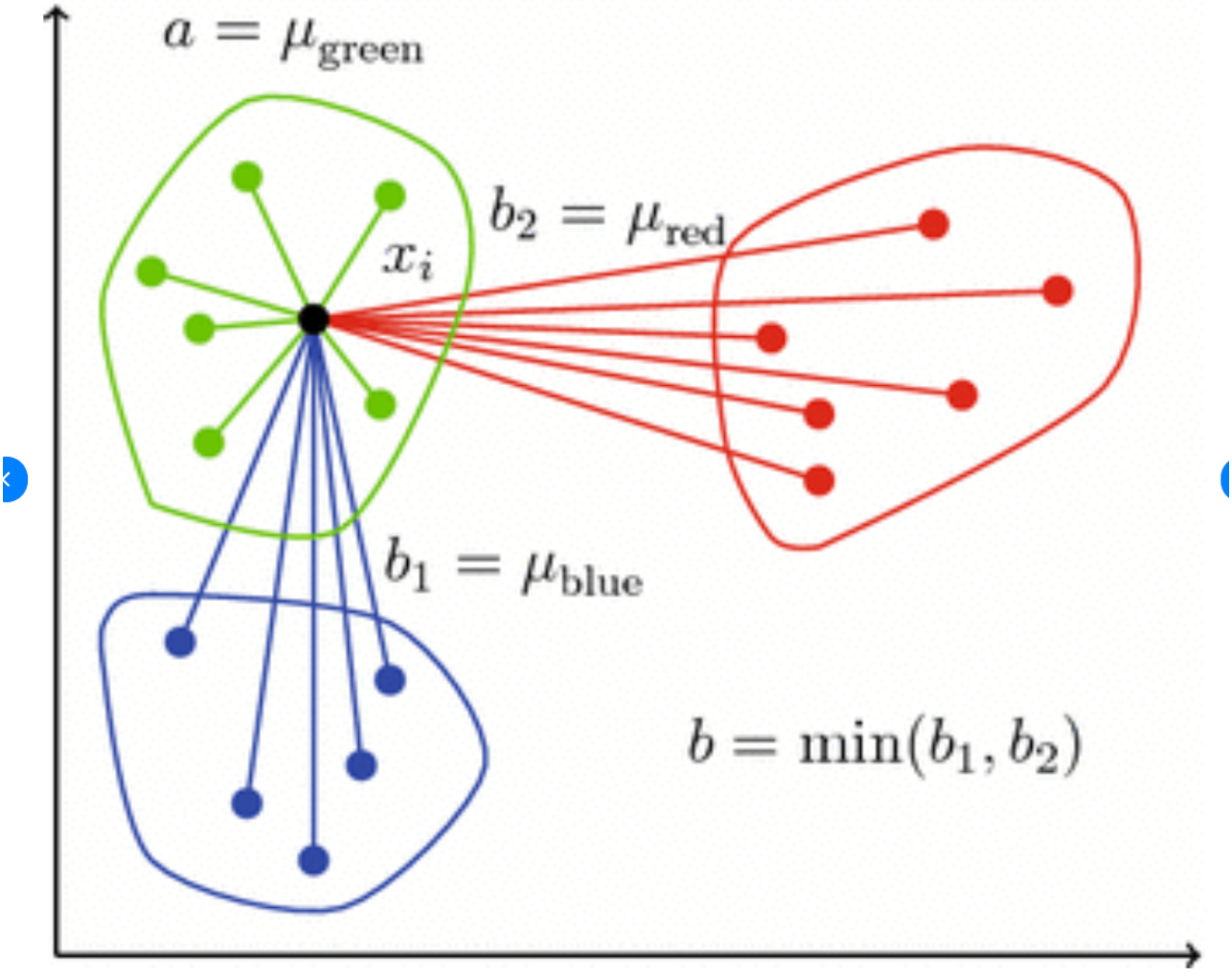
A higher score signifies better-defined clusters, which range from -1 to 1. An object is said to be well-matched to its own cluster and poorly-matched to nearby clusters if its score is close to 1. A score of about -1, on the other hand, suggests that the object might be in the incorrect cluster.

The Silhouette Coefficient of a sample measures the average distance of a sample with all other points in the next nearest cluster against all other points in its cluster. A higher ratio signifies the cluster is far away from its nearest cluster and that the cluster is more well-defined.

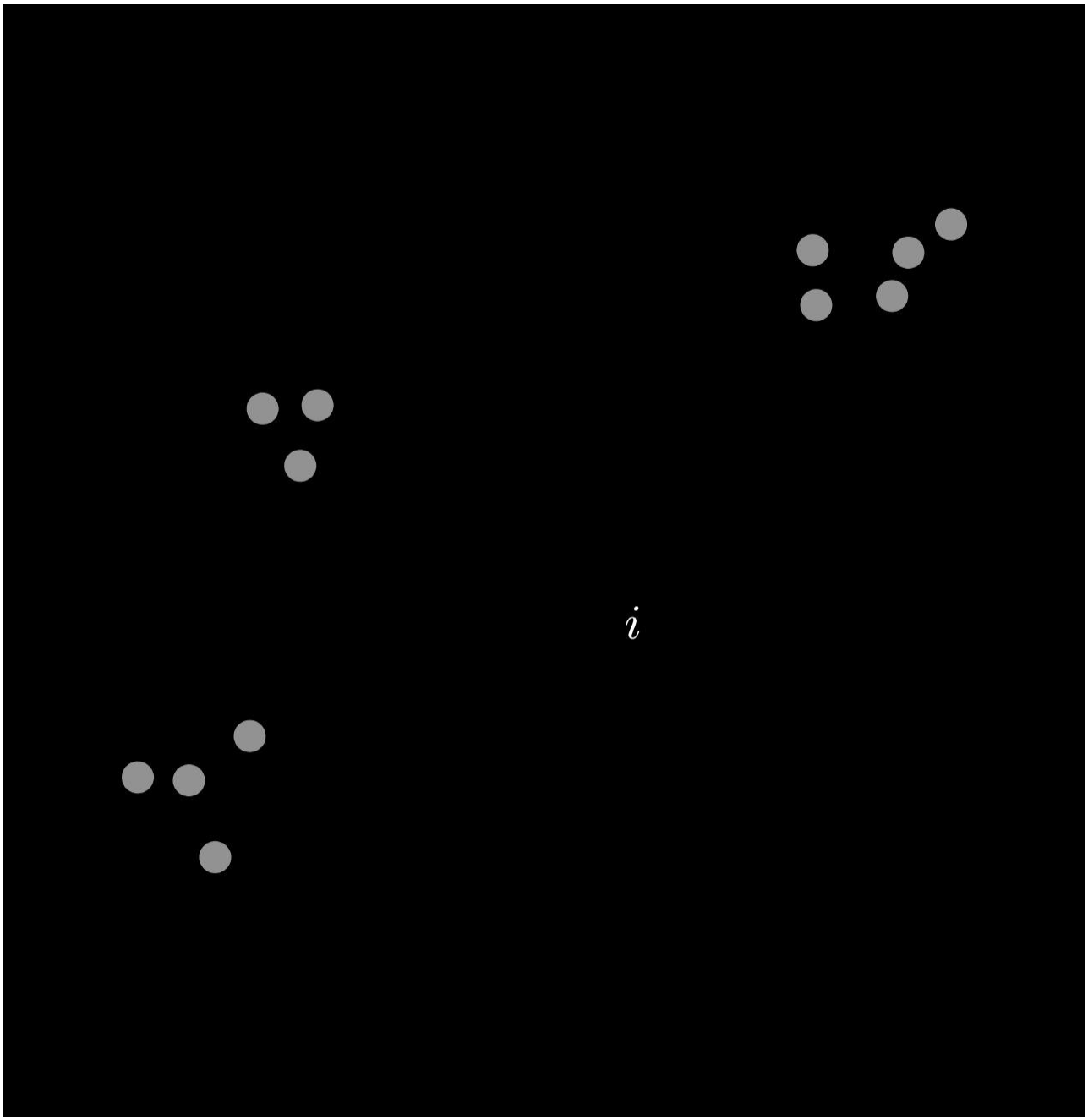
$$s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}}$$

Where:

- $(s(i))$: The Silhouette Coefficient for the $(i) - th$ data point.
- $(a(i))$: The average distance between the $(i) - th$ data point and all other points in the same cluster. This measures how well the data point fits into its own cluster.
- $(b(i))$: The smallest average distance of the $(i) - th$ data point to all points in any other cluster, of which (i) is not a member. This measures how poorly the data point fits into neighboring clusters.



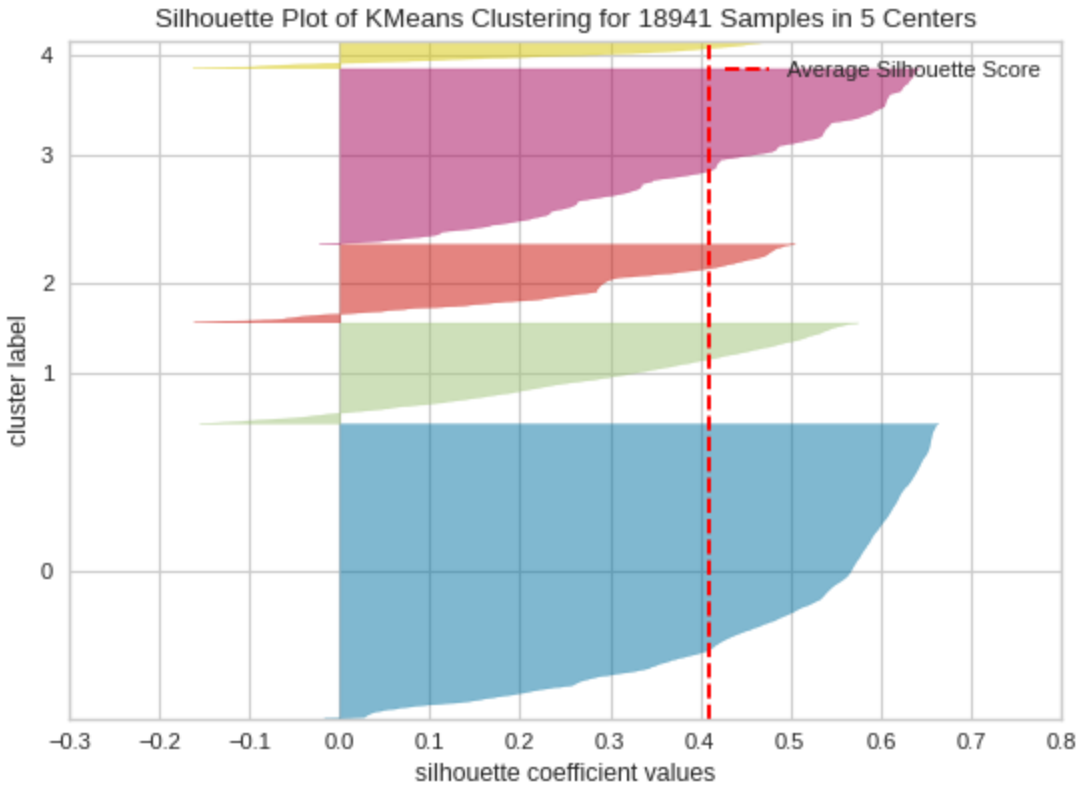
Silhouette coefficient example



Range of (s(i)): The value of $(s(i))$lies between (-1) and (1).

- When $(s(i) = 1)$, the sample is well matched to its own cluster and poorly matched to neighboring clusters.
- If $(s(i) = 0)$, the sample is on or very close to the decision boundary between two neighboring clusters.

- A value of $(s(i) = -1)$ indicates that the sample has been assigned to the wrong cluster.



The Silhouette Coefficient for a set of samples takes the average Silhouette Coefficient for each sample.

$$\text{Average Silhouette Score} = \frac{1}{N} \sum_{i=1}^{N} s(i)$$

This measure is powerful for assessing the effectiveness of clustering because it not only incorporates how compact the clusters are but also how well-separated they are from each other, which are key indicators of good clustering results.

Resources:

- https://www.geeksforgeeks.org/clustering-metrics/
- https://blog.paperspace.com/ml-evaluation-metrics-part-2/