

DATA MINING

TP2: Algorithme K-Médoïdes

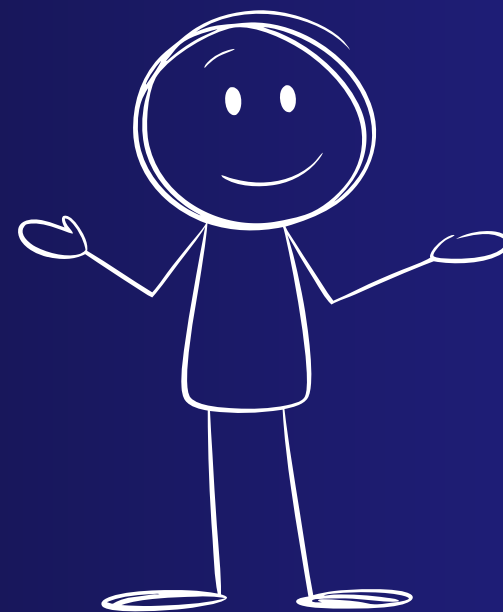


Introduction

La méthode du coude (Elbow methode) est couramment utilisée pour déterminer le nombre optimal de clusters (k) pour les algorithmes de clustering.

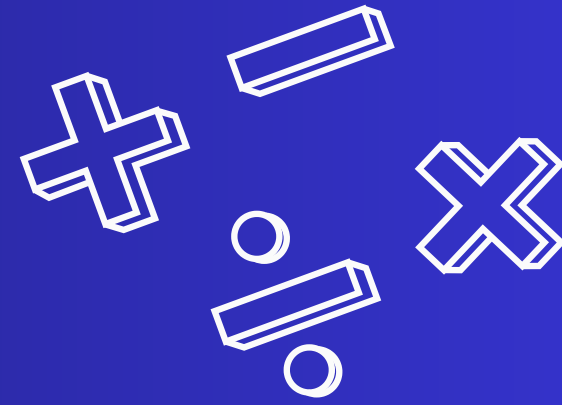
Cependant, cette méthode présente plusieurs limitations :

- Le point de courbure n'est pas toujours évident
- Plusieurs interprétations possibles
- Ambiguïté dans la détermination précise du point de coude
- Ne considère que la distance intra-cluster
- Ignore d'autres métriques importantes



**Le score de silhouette
(Silhouette score)**

Silhouette score

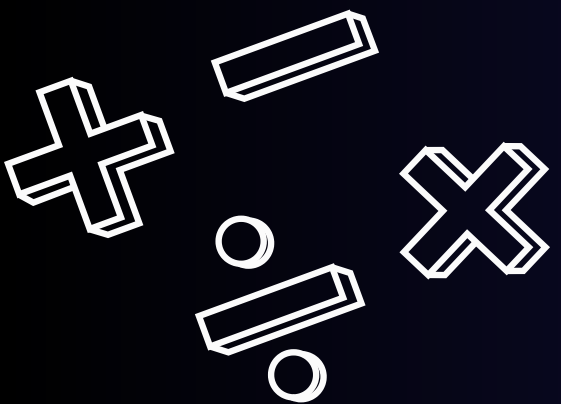


Le score de silhouette (Silhouette score) offre une alternative pour évaluer la qualité du clustering.

La valeur de silhouette est une mesure de la similarité d'un objet avec son propre cluster (cohésion) comparée à d'autres clusters (séparation).

La silhouette varie de **-1 à +1** où une valeur élevée indique que l'objet est bien apparié à son propre cluster et mal apparié aux clusters voisins .

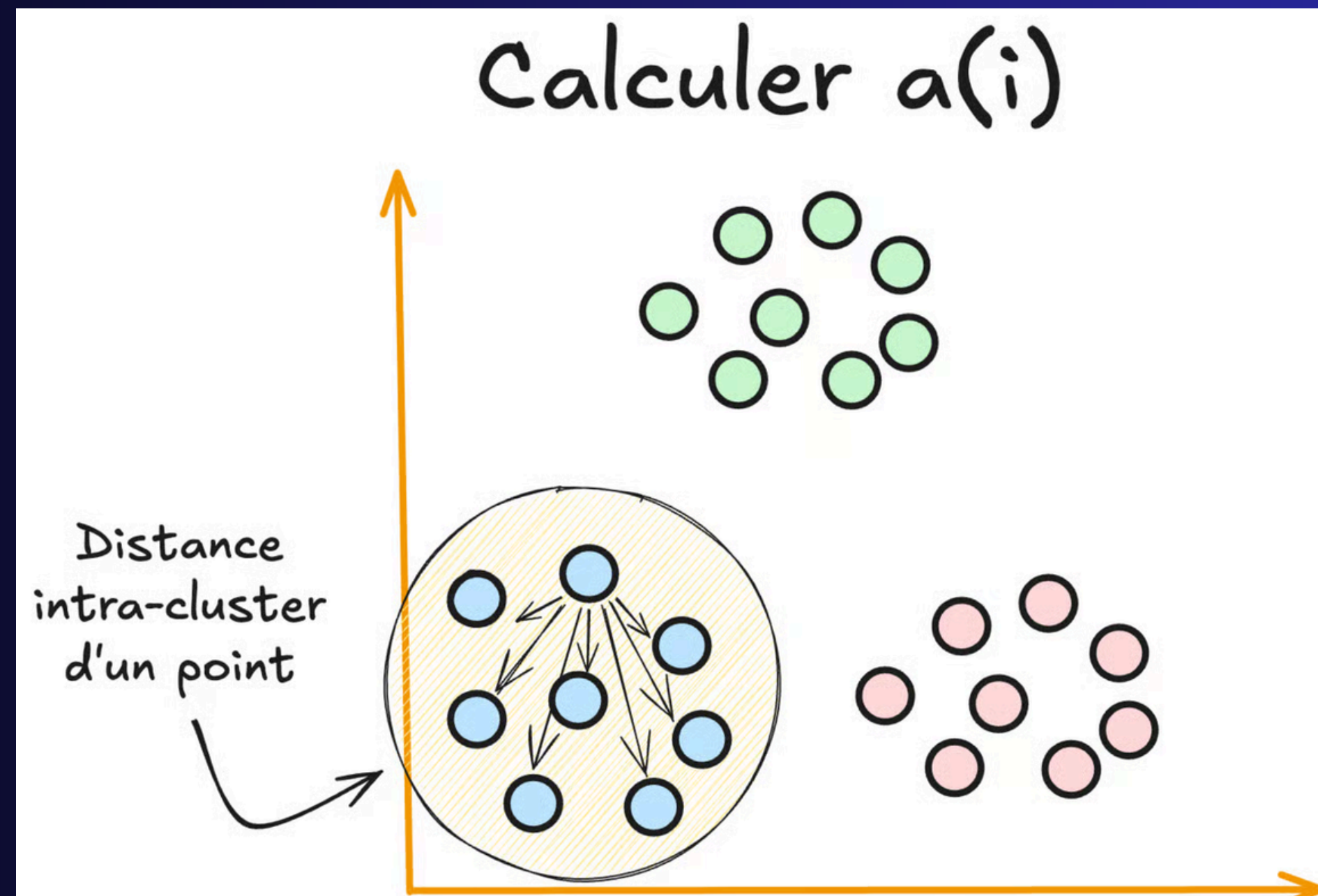
La silhouette peut être calculée avec n'importe quelle métrique de distance, telle que la distance Euclidienne ou la distance de Manhattan.



Distance intra-cluster

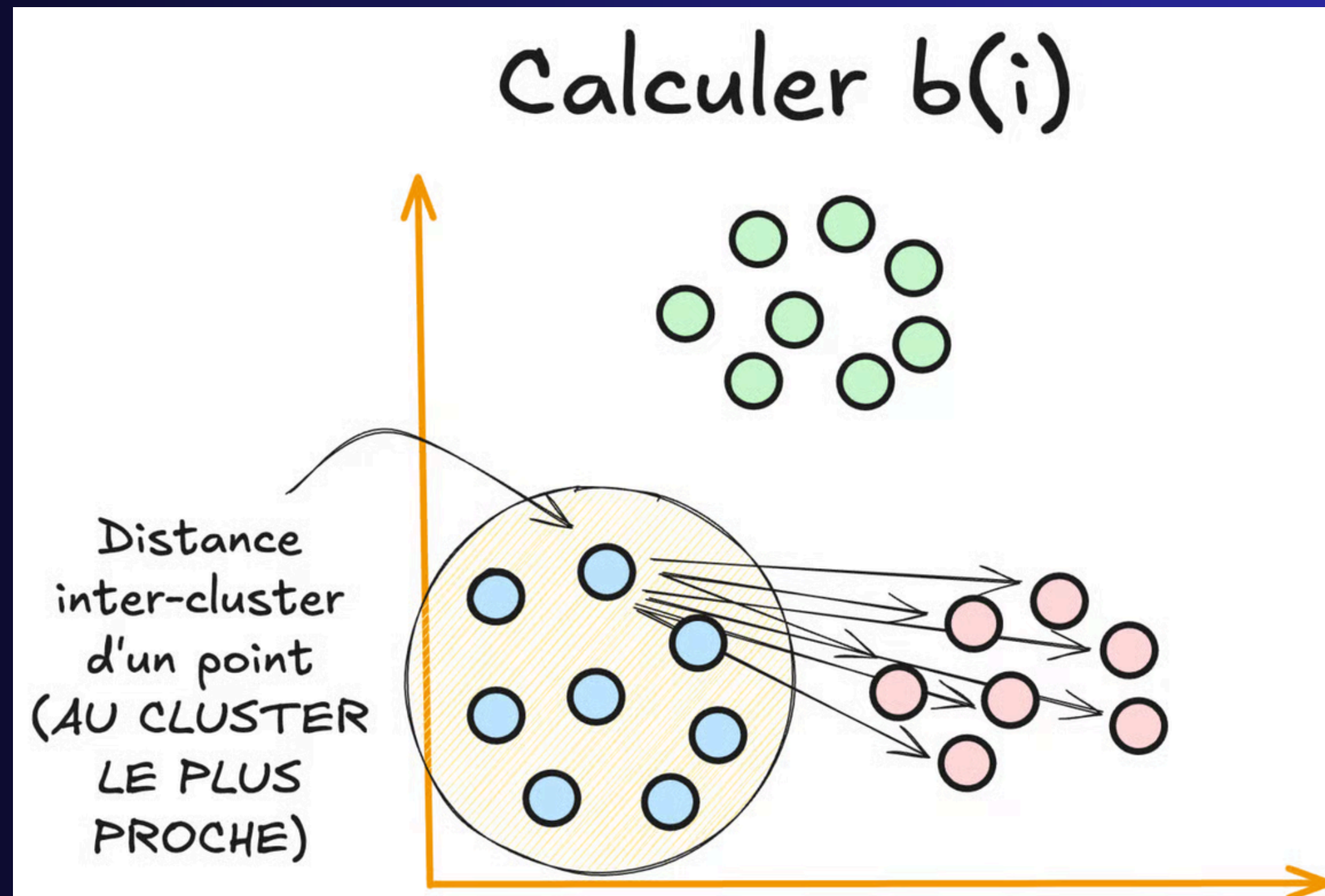
Pour chaque point de données (i), on détermine :

$a(i)$: la distance moyenne avec tous les autres points au sein du même cluster



Distance intr-cluster

b(i) : Distance moyenne avec les points du cluster voisin le plus proche



Calcul du score de silhouette

Pour un point de données spécifique (i), le score s'obtient par :

$$s_i = (b_i - a_i) / \max(a_i, b_i)$$

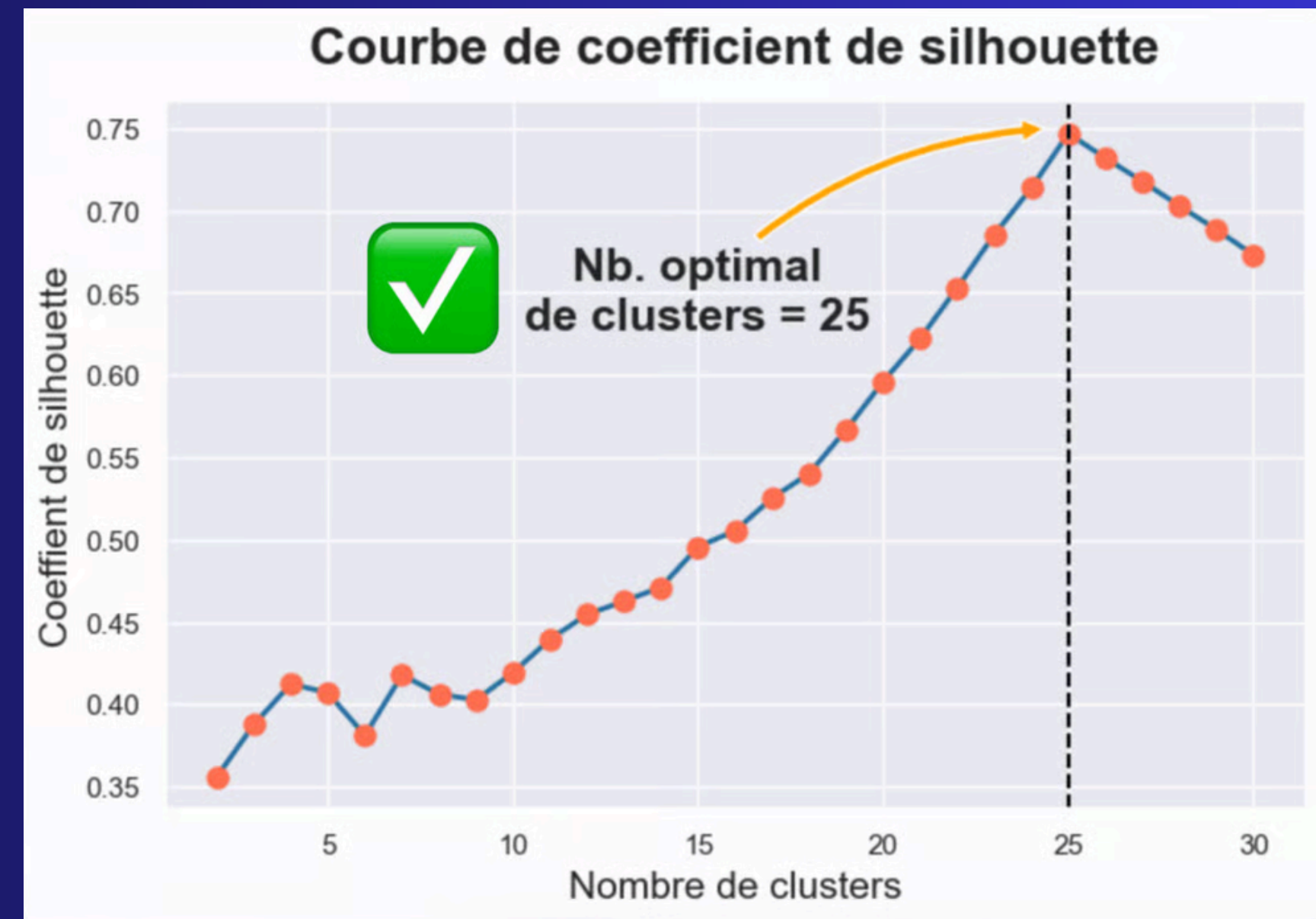
Où :

a_i = distance intra-cluster (cohésion)

b_i = distance inter-cluster (séparation)

Le score global du clustering est la moyenne :

$$s = (1/n) \sum s_i$$



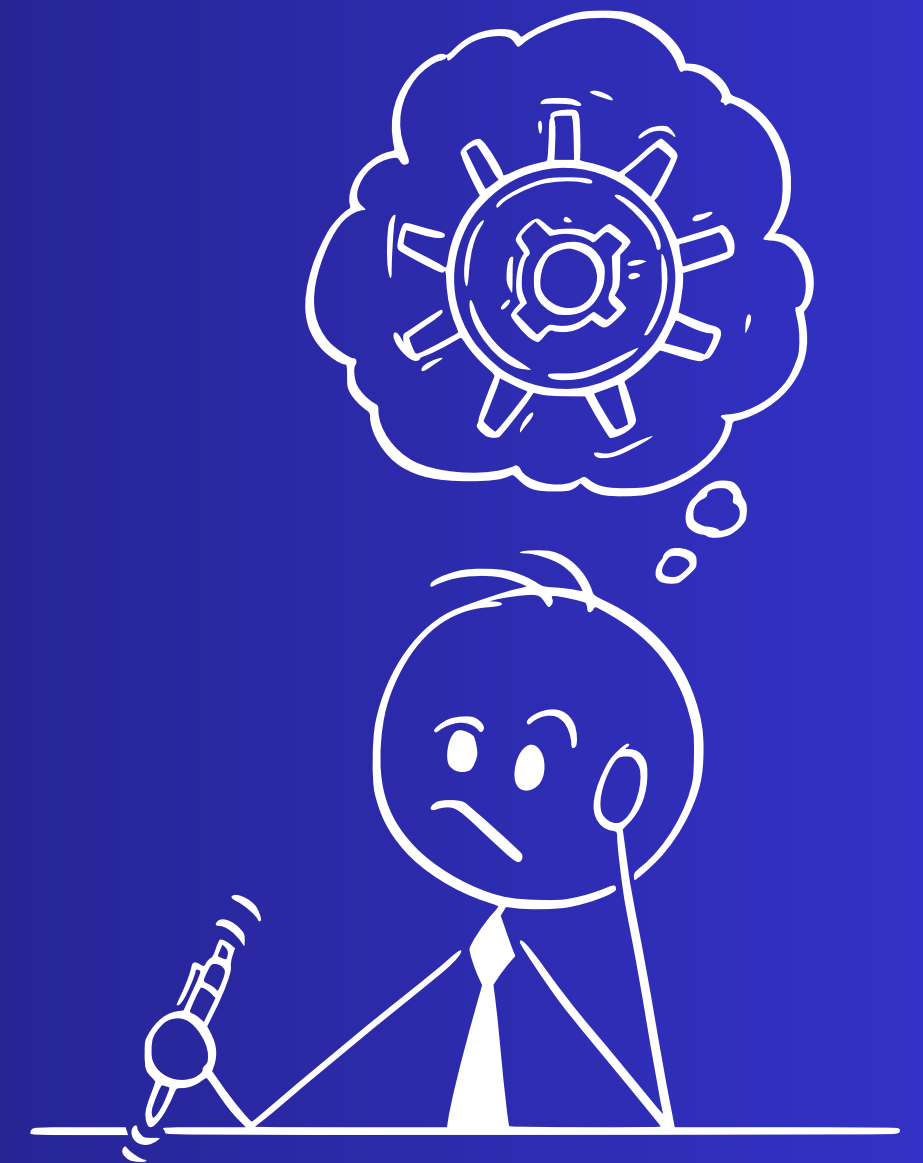
Propriétés

Plage de valeurs : $[-1, 1]$

- 1 : clustering optimal
- 0 : chevauchement des clusters
- -1 : affectation incorrecte

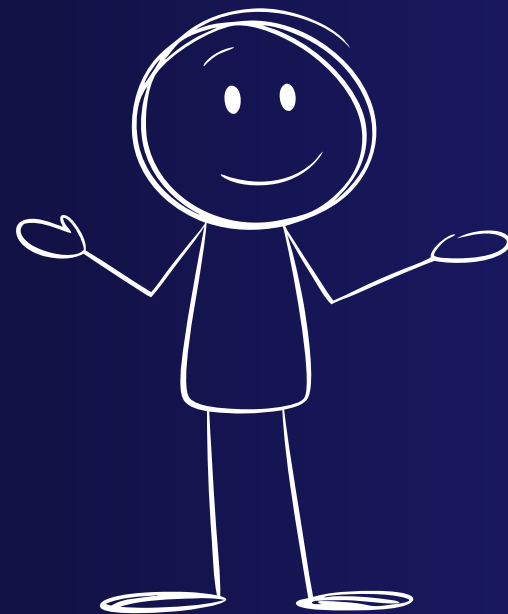
Interprétation :

Un score élevé indique un meilleur clustering



Le problème de l'algorithme K-means

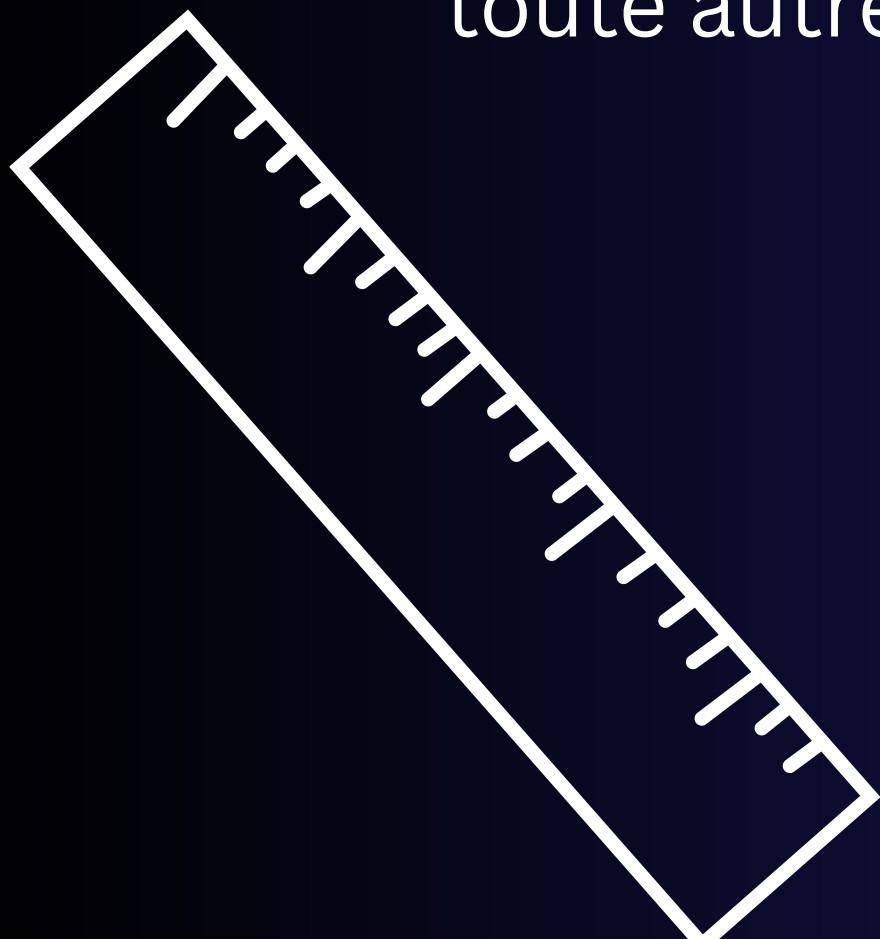
Lorsque des outliers sont assignés à un cluster, ils peuvent fausser considérablement la valeur moyenne de ce cluster. Cela affecte involontairement l'assignation des autres instances aux clusters.



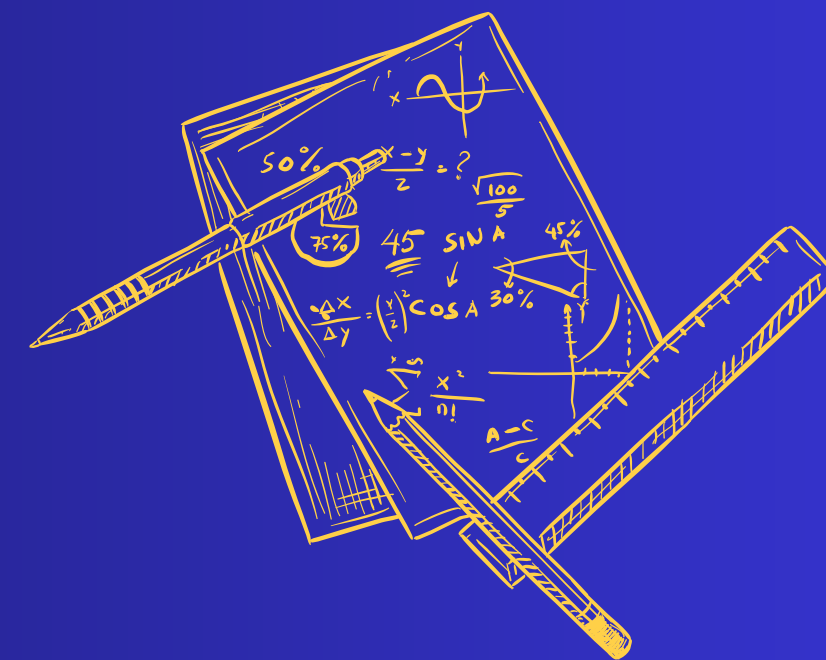
Algorithme K-Médoïdes

Algorithme K-Médoïdes

L'algorithme K-Medoids est un algorithme de clustering non supervisé où les instances, appelés **médoïdes**, servent de centres aux clusters. Un médoïde est un point du cluster dont la somme des distances (ou dissimilarités) à tous les autres instances du cluster est minimale. Cette distance peut être la distance euclidienne, la distance de Manhattan ou toute autre fonction de distance.



$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$



Fonctionnement de l'algorithme des K-médoides

Algorithm: *k-medoids*. PAM, a *k*-medoids algorithm for partitioning based on medoid or central objects.

Input:

- *k*: the number of clusters,
- *D*: a data set containing *n* objects.

Output: A set of *k* clusters.

Method:

- (1) arbitrarily choose *k* objects in *D* as the initial representative objects or seeds;
- (2) **repeat**
- (3) assign each remaining object to the cluster with the nearest representative object;
- (4) randomly select a nonrepresentative object, \mathbf{o}_{random} ;
- (5) compute the total cost, *S*, of swapping representative object, \mathbf{o}_j , with \mathbf{o}_{random} ;
- (6) **if** $S < 0$ **then** swap \mathbf{o}_j with \mathbf{o}_{random} to form the new set of *k* representative objects;
- (7) **until** no change;

