Apprentissage automatique

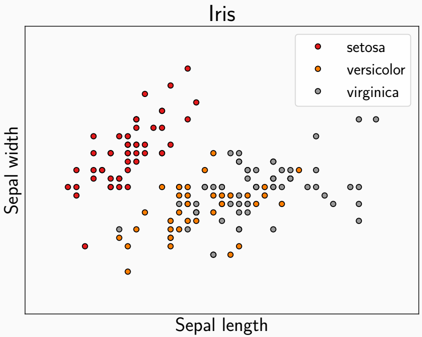
Clustering

Brice Chardin
ISAE-ENSMA

2023–2024

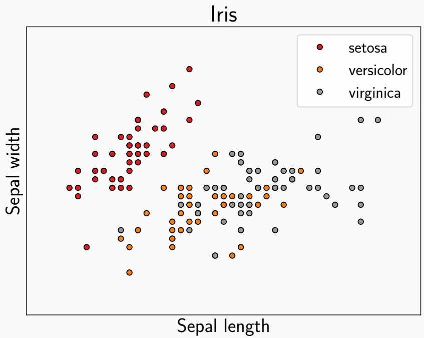
Clustering

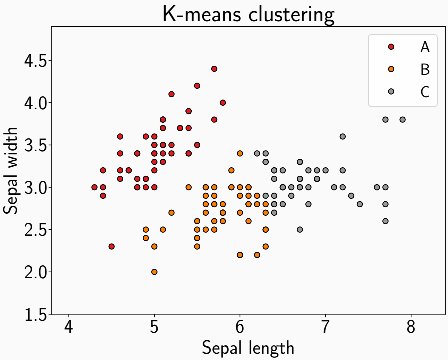
Catégoriser un jeu de données sans connaître la vérité de terrain
(i.e. apprentissage non supervisé)



Clustering

Catégoriser un jeu de données sans connaître la vérité de terrain (i.e. apprentissage non supervisé)





Définitions

Partition : ensemble d'ensembles $\{P_1, \ldots, P_k\}$ non nuls, disjoints et dont l'union est la population X

- $\forall i \in \{1, .., k\},\ P_i \neq \emptyset$
- $\forall i, j \in \{1, .., k\},\ i \neq j \implies P_i \cap P_j = \emptyset$
- $\bigcup_{i=1}^{k} P_i = X$

Clustering : partitionnement de la population en ensembles *homogènes* et *bien séparés*
Homogénéité et séparation dépendent des besoins

Objectif (pour certains algorithmes) : critère quantifiable à minimiser ou maximiser

Contraintes usuelles

Clustering sous contraintes, paramètres d'entrée du modèle

Contraintes sur les partitions :

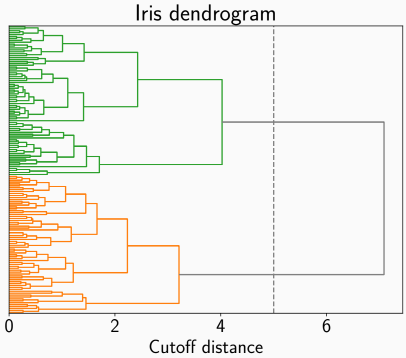
- nombre maximum de clusters
- dimensions maximales (e.g. diamètre)
- nombre minimal ou maximal d'éléments par cluster
- etc.

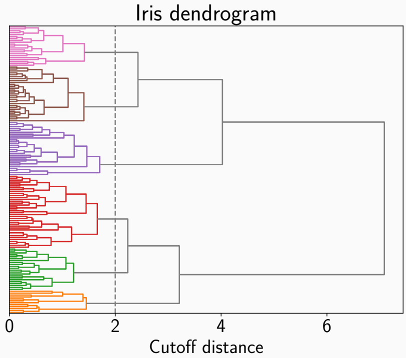
Contraintes sur les instances (apprentissage semi-supervisé)

- paires d'instances incompatibles (*cannot-link*)
- paires d'instances similaires (*must-link*)

Classification des algorithmes (1)

- Clustering hiérarchique → hiérarchie de partitions
- Clustering de partitionnement → partition unique en résultat





Classification des algorithmes (2)

- Méthodes de division $\rightarrow$ état initial : un seul cluster contenant la population
- Méthodes d'agglomération $\rightarrow$ état initial : un cluster par élément

Critère applicable à toute méthode hiérarchique et à certaines méthodes de partitionnement

Classification des algorithmes (3)

- Hard clustering $\rightarrow$ partitions discrètes (i.e. labels)
- Fuzzy clustering $\rightarrow$ degré (ou probabilité) d'appartenance à une partition

Classification des algorithmes (3)

- Hard clustering $\rightarrow$ partitions discrètes (i.e. labels)
- Fuzzy clustering $\rightarrow$ degré (ou probabilité) d'appartenance à une partition

- Déterministe $\rightarrow$ mêmes entrées $=$ mêmes sorties
- Stochastique $\rightarrow$ intègre un processus aléatoire

Classification des algorithmes (3)

- Hard clustering $\rightarrow$ partitions discrètes (i.e. labels)
- Fuzzy clustering $\rightarrow$ degré (ou probabilité) d'appartenance à une partition

- Déterministe $\rightarrow$ mêmes entrées $=$ mêmes sorties
- Stochastique $\rightarrow$ intègre un processus aléatoire

Si objectif défini :

- Exact $\rightarrow$ garantit une solution optimale
- Approximatif $\rightarrow$ ne garantit pas une solution optimale

Clustering en k-moyennes (*k-means*)

Contrainte : nombre prédéfini k de clusters

Objectif : minimiser l'écart quadratique moyen du clustering

Soient $\{C_1, \ldots, C_k\}$ clusters de centres $\{\mu_1, \ldots, \mu_k\}$

$$\text{écart} = \sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||^2$$

Clustering en k-moyennes (*k-means*)

Contrainte : nombre prédéfini k de clusters

Objectif : minimiser l'écart quadratique moyen du clustering

Soient $\{C_1, \ldots, C_k\}$ clusters de centres $\{\mu_1, \ldots, \mu_k\}$

$$\text{écart} = \sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||^2$$

Solution optimale difficile à calculer

Algorithme des k-moyennes (*k-means*)

Algorithme itératif simple :

1. affecter un point à chaque cluster
2. répéter :
   - calculer le centre de chaque cluster
   - affecter chaque point au cluster dont le centre est le plus proche
3. jusqu'à convergence (i.e. pas de réaffectation)

Algorithme des k-moyennes (*k-means*)

Algorithme itératif simple :

1. affecter un point à chaque cluster
2. répéter :
    - calculer le centre de chaque cluster
    - affecter chaque point au cluster dont le centre est le plus proche
3. jusqu'à convergence (i.e. pas de réaffectation)

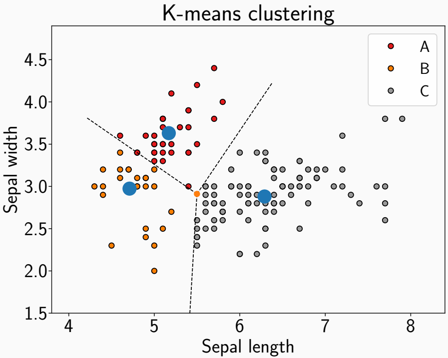
Convergence garantie, mais vers un minimum local (solution approximative)

Résultat (très) dépendant de l'initialisation

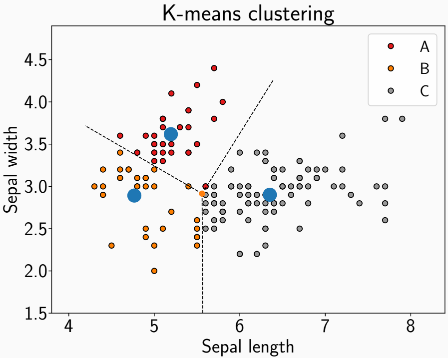
- meilleur résultat de plusieurs exécutions avec initialisations différentes
- heuristiques d'initialisation (e.g. k-means++)

Parmi les algorithmes de partitionnement les plus rapides

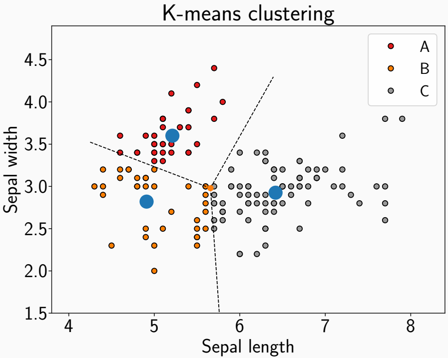
Itérations de k-means



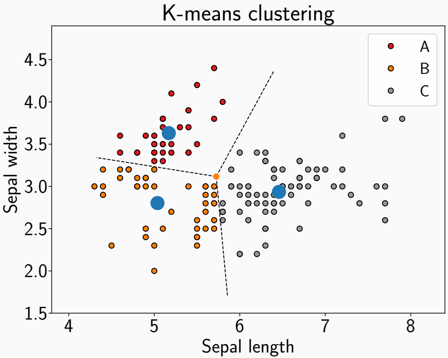
Itérations de k-means



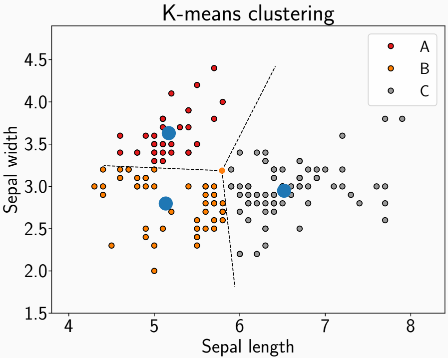
Itérations de k-means



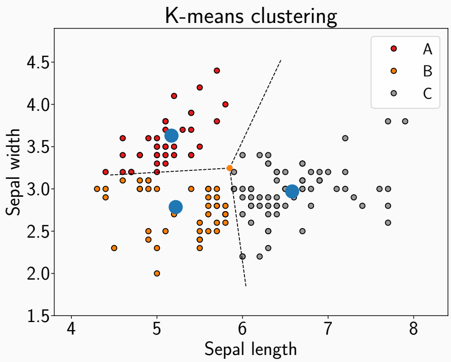
Itérations de k-means



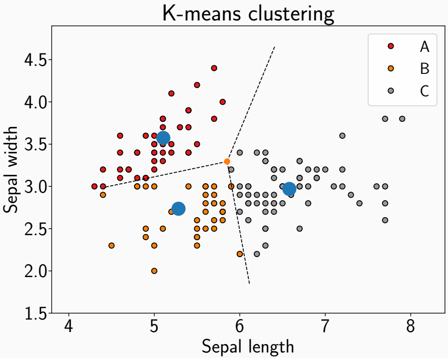
Itérations de k-means



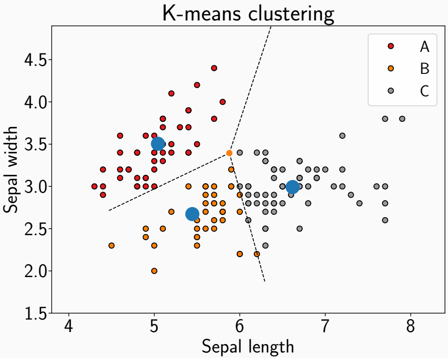
Itérations de k-means



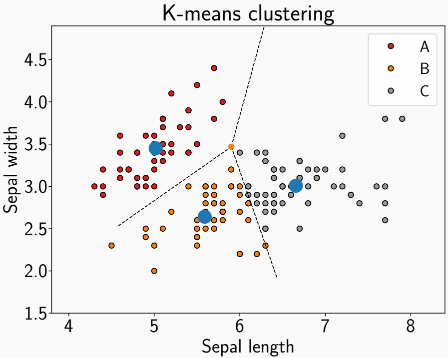
Itérations de k-means



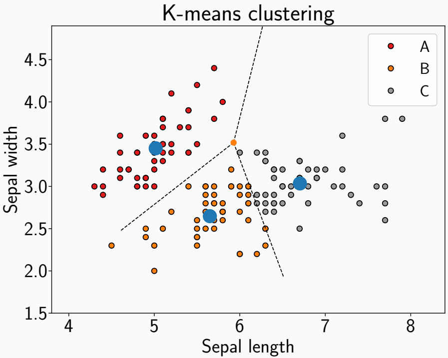
Itérations de k-means



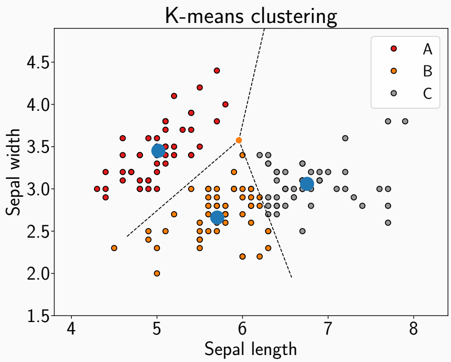
Itérations de k-means



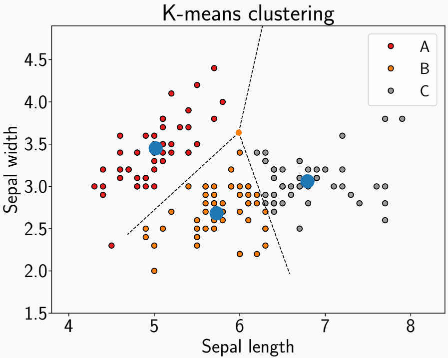
Itérations de k-means



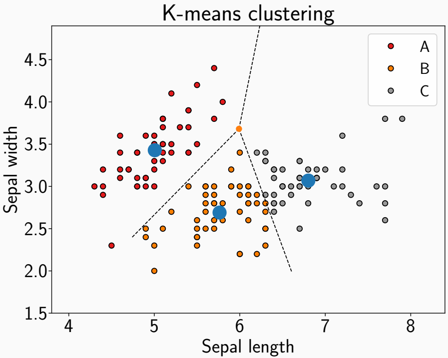
Itérations de k-means



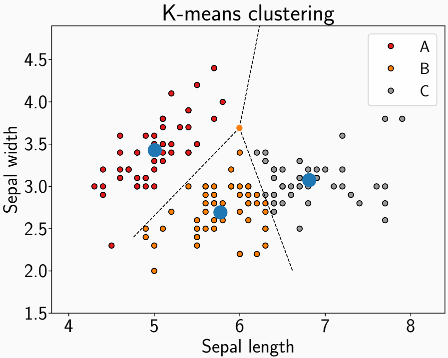
Itérations de k-means



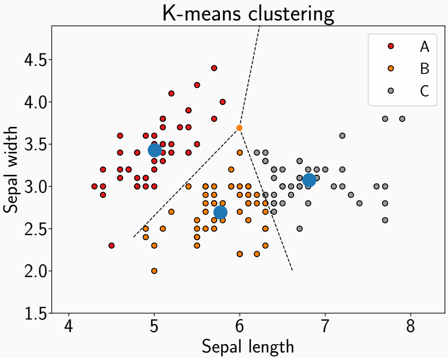
Itérations de k-means



Itérations de k-means



Itérations de k-means



→ Convergence

k-médoïdes

Même principe que k-means en remplaçant les centres par des médoïdes

Utile lorsqu'une moyenne n'est pas calculable

Données catégorielles

- d(pomme, pomme) $= 0$
- d(pomme, poire) $= 1$
- avg(pomme, poire) $= ?$

Définition (médoïde)

$$\text{médoïde}(C) = \underset{c \in C}{\operatorname{argmin}} \sum_{e \in C} d(e, c)$$

I.e. point du cluster dont la distance moyenne aux autres membres est minimale

Exemple d'algorithme approximatif : PAM (*Partitioning Around Medoids*)

Méthodes par densité : DBSCAN et OPTICS

Méthodes par densité, définies par deux paramètres : ϵ et *minpts*

Zone dense : zone contenant au moins *minpts* points dans un rayon de ϵ

Propagation de la notion de densité pour obtenir les clusters

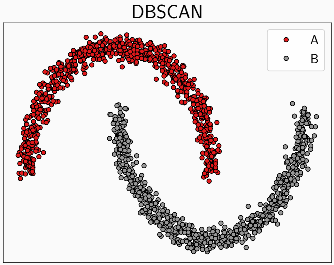
Méthodes par densité : DBSCAN et OPTICS

Méthodes par densité, définies par deux paramètres : ϵ et *minpts*

Zone dense : zone contenant au moins *minpts* points dans un rayon de ϵ

Propagation de la notion de densité pour obtenir les clusters

Permet d'obtenir des formes non convexes



Classification ascendante hiérarchique

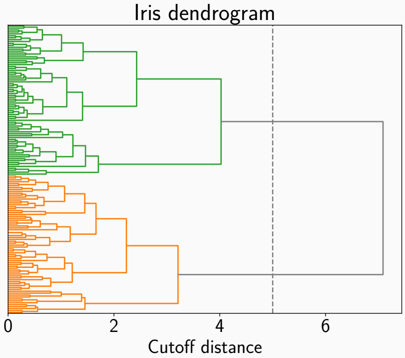
1. affecter chaque point à un cluster distinct
2. tant qu'il existe plusieurs clusters
   - fusionner les deux clusters les plus *proches*

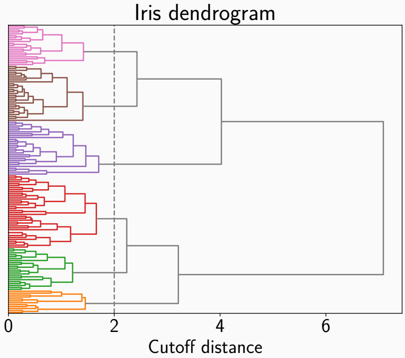
Critères de proximité entre deux clusters A et B :

- distance maximale entre deux points (complete-link) : $\max\{d(a,b) : a \in A, b \in B\}$
- distance minimale entre deux points (single-link) : $\min\{d(a,b) : a \in A, b \in B\}$
- distance entre les médoïdes de A et B
- distance moyenne : $\dfrac{\displaystyle\sum_{a \in A} \sum_{b \in B} d(a,b)}{|A| \cdot |B|}$
- etc.

Classification ascendante hiérarchique

Critère de coupure sur la hiérarchie résultat : distance ou nombre de clusters





Mesures de qualité

Taux de bonne classification

Si vérité de terrain connue : $\text{Tx} = \dfrac{\text{nb éléments bien classés}}{\text{nb éléments total}}$

Matrice de contingence

	Clusters				
Classes	**C_1**	**C_2**	**C_3**	**C_4**	Attendu
A	15	7	10	0	
B	12	2	5	0	
C	3	20	3	11	
D	0	4	12	10	

Reel

Taux de bonne classification

Si vérité de terrain connue : $\text{Tx} = \dfrac{\text{nb éléments bien classés}}{\text{nb éléments total}}$

Matrice de contingence

	Clusters			
Classes	C_1	C_2	C_3	C_4
A	15	7	10	0
B	12	2	5	0
C	3	20	3	11
D	0	4	12	10

Considérer le cas le plus favorable

Taux de bonne classification

Si vérité de terrain connue : $\text{Tx} = \dfrac{\text{nb éléments bien classés}}{\text{nb éléments total}}$

Matrice de contingence

	Clusters			
Classes	**C_1**	**C_2**	**C_3**	**C_4**
A	15	7	10	0
B	12	2	5	0
C	3	20	3	11
D	0	4	12	10

Considérer le cas le plus favorable

$$\text{Tx} = \frac{10 + 12 + 20 + 10}{114} = 45.6\%$$

Indice de Rand

Proportion de paires d'éléments égaux ou différents dans les deux cas

Indice de Rand

Proportion de paires d'éléments égaux ou différents dans les deux cas

Indice de Rand non ajusté :

Classes	**Clusters**
A	C_1
B	C_2
A	C_2
B	C_2
C	C_1

Concordances : 0

Total : 10

Indice de Rand

Proportion de paires d'éléments égaux ou différents dans les deux cas

Indice de Rand non ajusté :

Classes	Clusters
A	C_1
B	C_2
A	C_2
B	C_2
C	C_1

Concordances : 1

Total : 10

Indice de Rand

Proportion de paires d'éléments égaux ou différents dans les deux cas

Indice de Rand non ajusté :

	Classes	**Clusters**	
	A	C_1	
$=$	B	C_2	$\neq$
	A	C_2	
	B	C_2	
	C	C_1	

Concordances : 1

Total : 10

Indice de Rand

Proportion de paires d'éléments égaux ou différents dans les deux cas

Indice de Rand non ajusté :

$$\begin{array}{cccc}
 & \textbf{Classes} & \textbf{Clusters} & \\
 & \mathrm{A} & \mathrm{C}_1 & \\
 & \mathrm{B} & \mathrm{C}_2 & \\
\neq & \mathrm{A} & \mathrm{C}_2 & \neq \\
 & \mathrm{B} & \mathrm{C}_2 & \\
 & \mathrm{C} & \mathrm{C}_1 &
\end{array}$$

Concordances : 2

Total : 10

Indice de Rand

Proportion de paires d'éléments égaux ou différents dans les deux cas

Indice de Rand non ajusté :

$$
\neq \quad
\begin{array}{cc}
\textbf{Classes} & \textbf{Clusters} \\
\mathrm{A} & \mathrm{C}_1 \\
\mathrm{B} & \mathrm{C}_2 \\
\mathrm{A} & \mathrm{C}_2 \\
\mathrm{B} & \mathrm{C}_2 \\
\mathrm{C} & \mathrm{C}_1
\end{array}
\quad =
$$

Concordances : 2

Total : 10

Indice de Rand

Proportion de paires d'éléments égaux ou différents dans les deux cas

Indice de Rand non ajusté :

	Classes	**Clusters**	
	A	C_1	
	B	C_2	
$\neq$	A	C_2	$=$
	B	C_2	
	C	C_1	

Concordances : 2

Total : 10

Indice de Rand

Proportion de paires d'éléments égaux ou différents dans les deux cas

Indice de Rand non ajusté :

Classes	**Clusters**
A	C_1
B	C_2
A	C_2
B	C_2
C	C_1

Concordances : 3

Total : 10

Indice de Rand

Proportion de paires d'éléments égaux ou différents dans les deux cas

Indice de Rand non ajusté :

Classes	**Clusters**
A	C_1
B	C_2
A	C_2
B	C_2
C	C_1

$\neq$ $\neq$

Concordances : 4

Total : 10

Indice de Rand

Proportion de paires d'éléments égaux ou différents dans les deux cas

Indice de Rand non ajusté :

	Classes	**Clusters**	
	A	C_1	
	B	C_2	
	A	C_2	
$\neq$	B	C_2	$=$
	C	C_1	

Concordances : 4

Total : 10

Indice de Rand

Proportion de paires d'éléments égaux ou différents dans les deux cas

Indice de Rand non ajusté :

Classes	**Clusters**
A	C_1
B	C_2
A	C_2
B	C_2
C	C_1

$\neq$ $\neq$

Concordances : 5

Total : 10

Indice de Rand

Proportion de paires d'éléments égaux ou différents dans les deux cas

Indice de Rand non ajusté :

Classes	**Clusters**
A	C_1
B	C_2
A	C_2
B	C_2
C	C_1

$\neq$ $\neq$

Concordances : 6

Total : 10

Indice de Rand : $\frac{6}{10} = 0.6$

Autres mesures

Indice de Rand non ajusté

- Score parfait de 1 (concordance totale)
- Pire score de 0 (discordance totale) difficile à obtenir

Autres mesures

Indice de Rand non ajusté

- Score parfait de 1 (concordance totale)
- Pire score de 0 (discordance totale) difficile à obtenir

→ Les concordances sur les différences donnent des scores proches de 1 même pour des clusters très dissimilaires

Autres mesures

Indice de Rand non ajusté

- Score parfait de 1 (concordance totale)
- Pire score de 0 (discordance totale) difficile à obtenir

$\rightarrow$ Les concordances sur les différences donnent des scores proches de 1 même pour des clusters très dissimilaires

Indice de Rand ajusté : score entre -1 et 1, affectation aléatoire $\rightarrow$ 0

Mesures d'homogénéité

Autres mesure si la vérité de terrain n'est pas connue

Inertie intra-cluster (mesure d'homogénéité des clusters)

$$\mathrm{I}_{\text{intra}} = \frac{1}{|X|} \sum_{i=1}^{k} \sum_{x \in C_i} d^2(x, \mu_i) \quad \leftarrow \text{fonction objectif de } \textit{k-means}$$

Score de silhouette : $s(x_i) = \dfrac{b(x_i) - a(x_i)}{\max(a(x_i), b(x_i))}$

$a(x_i)$: distance moyenne entre $x_i \in C_m$ et les points appartenant au même cluster

$b(x_i)$: minimum de la distance moyenne entre x_i et les points d'un autre cluster

Score de silhouette

Score de silhouette : $s(x_i) = \frac{b(x_i) - a(x_i)}{\max(a(x_i), b(x_i))}$

$$a(x_i) = \frac{1}{|C_m| - 1} \sum_{x_j \in C_m, j \neq i} d(x_i, x_j) \qquad b(x_i) = \min_{l \neq m} \left(\frac{1}{|C_l|} \sum_{x_j \in C_l} d(x_i, x_j) \right)$$

Si $s(x_i) < 0$, alors $b(x_i) < a(x_i)$: x_i pourrait (ou devrait) changer de cluster

Score de silhouette global correspond au score de silhouette moyen

$$s(X) = \frac{1}{|X|} \sum_{x_i \in X} s(x_i)$$

Un score de silhouette parfait (i.e. $a(x_i) = 0$, généralement inatteignable) vaut 1