Introduction au Machine Learning

Clustering

Théo Lopès-Quintas

BPCE Payment Services,
Université Paris Dauphine

2022 - 2025

Distance

Définition

Définition 1

Une métrique pour un ensemble M est une fonction $d : M \times M \to \mathbb{R}_+$ telle que pour tout $x, y, z \in M$:

1. ***Indiscernabilité*** *:* $d(x, y) = 0 \iff x = y$
2. ***Symétrie*** *:* $d(x, y) = d(y, x)$
3. ***Sous-additivité*** *:* $d(x, z) \leqslant d(x, y) + d(y, z)$

Les distances les plus classiques sont de la famille $\mathcal{L}_p$ et sont de la forme $d(x, y) = \left(\sum_{i=1} \|x_i - y_i\|^p \right)^{\frac{1}{p}}$.
La distance Manhattan en est un cas particulier avec $p = 1$ et la distance la plus classique est la distance euclidienne avec $p = 2$.
On doit garder en tête que toutes les métriques de cette famille de la forme souffrent grandement du fléau de la dimension [1]. Le meilleur conseil que l'on puisse donner est donc de rester autant que possible avec relativement peu d'indicateurs par rapport au nombre d'observations que l'on a à disposition.

1. Voir l'annexe sur le sujet

Approche statistique

K-Means

L'algorithme K-Means vise à partitionner l'espace des features en K clusters où chaque observation appartient au cluster avec la distance la moyenne la plus faible.

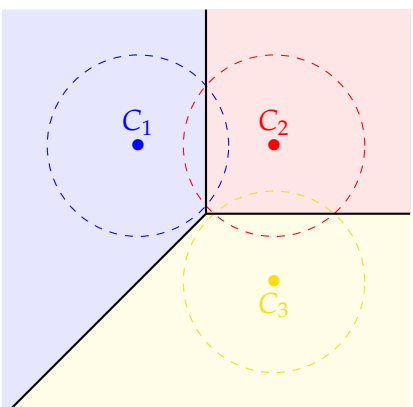


Figure – Exemple d'un clustering avec K-Means pour $K = 3$ clusters

Dans la figure (1) il faut bien comprendre que la partition de l'espace est représentée par les trois espaces colorés. Les cercles ne représentent pas les clusters, mais donnent une idée de la concentration des données autour du centre. L'algorithme K-Means ne renvoie pas des cercles, mais des cellules de Voronoi (les espaces colorés) par construction.

Pour exploiter l'algorithme K-Means, il faut spécifier une distance $d : \mathbb{R}^d \to \mathbb{R}_+$ et un nombre de clusters K que l'on souhaite obtenir. On définit les notations :

- C_k le k-ième cluster de coordonnées $\mu_k \in \mathbb{R}^d$
- $\mu \in \mathcal{M}_{d,K}$ la matrice engendrée par les $(\mu_k)_{k \leqslant K}$
- $z_i^k = \mathbb{1}_{\{x_i \in C_k\}}$ et $z \in \mathcal{M}_{n,K}$ la matrice engendrée par les $(z_i^k)_{i \leqslant n}^{k \leqslant K}$

Nombre d'observations — Nombre de clusters

$$J(\mu, z) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_i^k \|x_i - \mu_k\|^2 \qquad \text{(Distortion)}$$

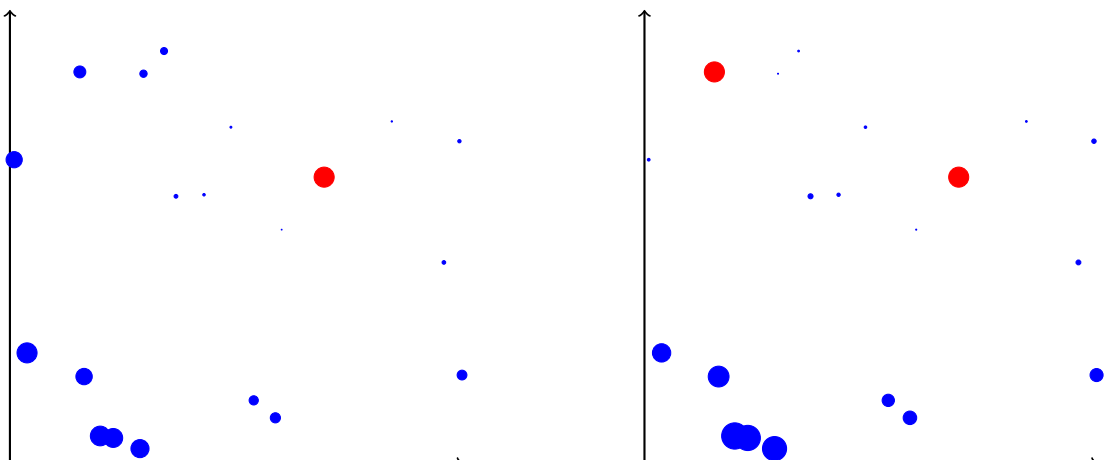
On cherche les meilleurs $(\mu_k)_{k \leqslant K}$ qui permettent de minimiser J :

$$\mu = \underset{\mu \in \mathcal{M}_{d,K}}{\arg\min}\, J(\mu, z)$$

Approche statistique

KMeans++ : un meilleur départ

Au départ nous utilisions plusieurs fois l'algorithme avec des vecteurs de départs aléatoires et on conserve la partition qui minimise le plus la distortion. Suivre cette méthode, nous expose à des problèmes théoriques de convergence, qu'on rencontre en pratique.
L'idée est de construire et étendre les centres de proche en proche.



(a) Sélection du deuxième centre

(b) Sélection du troisième centre

Figure – Distribution proportionnelle à la distance au carré des centres pour plusieurs itérations

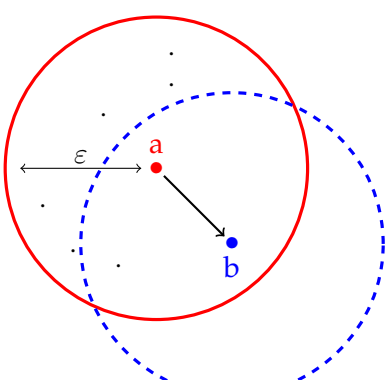
Approche par densité

Objet directement atteignable

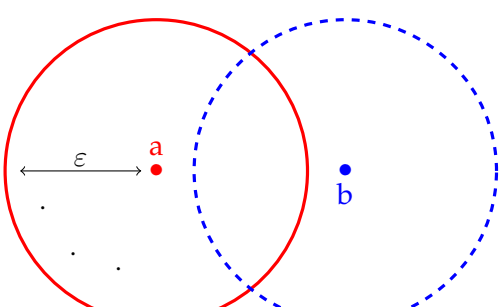
- ε : un seuil, qui sera utilisé pour décider de la proximité entre deux objets
- *MinPts* : le nombre minimum d'objets dans le voisinage d'un point pour qu'il soit considéré comme central
- $N_\varepsilon(x) = \{y \in M | d(x,y) \leqslant \varepsilon\}$: le voisinage d'un objet $x \in M$

Définition 2

Un objet b est directement atteignable depuis un objet a dans un ensemble d'objets D si $b \in N_\varepsilon(a)$ *et* $\#N_\varepsilon(a) \geqslant MinPts$



(a) *b* est directement atteignable depuis *a*



(b) *b* n'est pas directement atteignable depuis *a*

Définition 3

Un objet b est atteignable par densité depuis un objet a dans un ensemble d'objets D s'il existe une chaîne d'objets $o_0, \ldots, o_{n-1}$ telle que $o_0 = a$ et $o_{n-1} = b$ et que pour tout $i \leqslant n-1, o_i \in D$ et o_{i+1} est directement atteignable depuis o_i

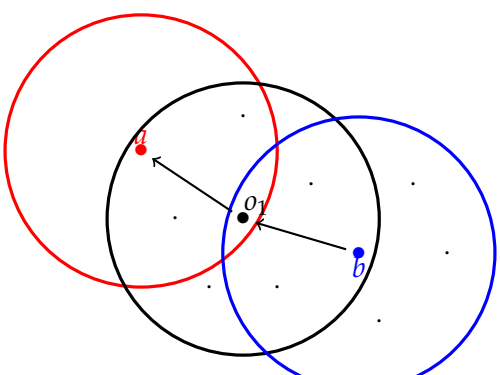


Figure – Exemple de point atteignable par densité

Définition 4

Un objet a est connecté par densité à un objet b dans un ensemble D s'il existe un objet $o \in D$ tel que a et b soit tous les deux atteignables par densité depuis o.

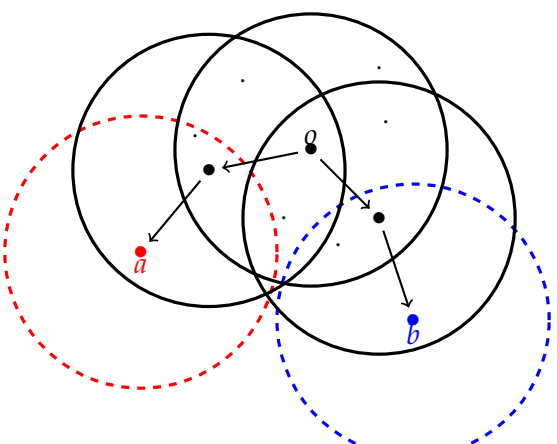


Figure – Exemple d'objet connecté par densité

Définition 5

Soit D un ensemble d'objets. Un cluster C est un sous-ensemble non vide de D qui vérifie :

1. ***Maximalité** : pour tout $a, b \in D$, si $b \in C$ et que b is atteignable par densité depuis a, alors $a \in C$*
2. ***Connectivité** : pour tout $a, b \in C$, b est connecté par densité avec a*

Tous les objets de D qui ne sont contenus dans aucun cluster sont regroupés dans un cluster qu'on appelle le ***bruit****.*

L'algorithme DBSCAN exploite ces notions de la manière suivante : il va chercher un point central et définir son voisinage. Puis il va inspecter son voisinage pour étendre avec des points centraux ce cluster. Il le fera tant qu'il ne peut plus ajouter de points à ce cluster. L'algorithme répétera la procédure jusqu'à ce qu'il ne puisse plus créer de cluster, et les points restants seront labellisés comme du bruit.

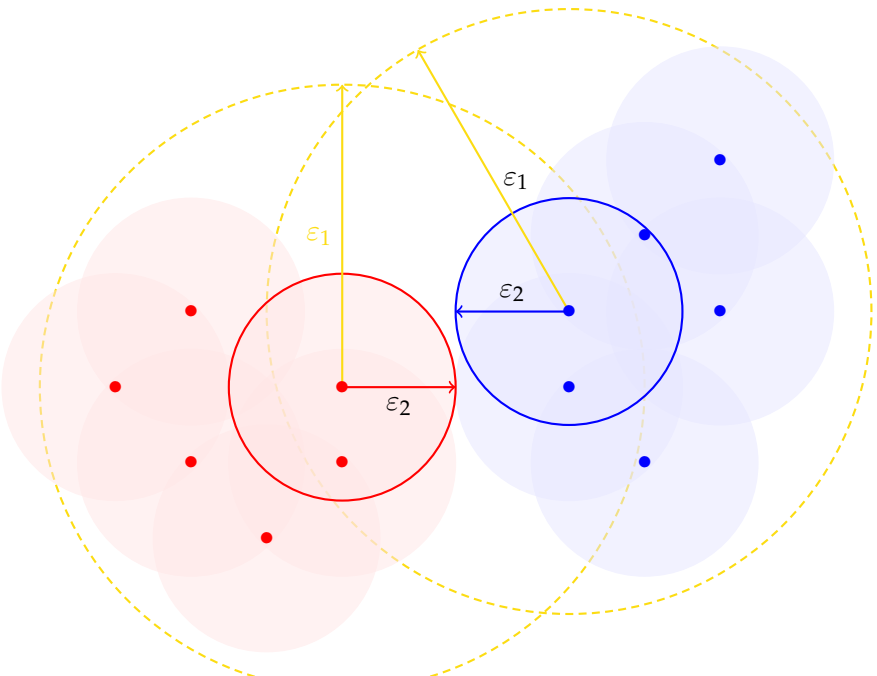


Figure – Deux clustering différents pour deux valeurs de ε

Performance d'un clustering

Silhouette score

Le silhouette score nécessite pour chaque observation x_i le calcul de deux nombres :

- a_i : la distance moyenne entre x_i et les autres points du cluster
- b_i : la distance moyenne entre x_i et les autres points du cluster le plus proche

On calcule, à l'aide des deux précédentes valeurs un score s_i :

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)}$$

Finalement, le silhouette score est définie comme :

$$S = \frac{1}{n} \sum_{i=1}^{n} s_i$$

Exercice 1 (Silhouette score)

On s'intéresse à l'interprétation du silhouette score S comme défini précédemment.

1. *Montrer que* $\forall i \leqslant n,\ s_i \in [-1, 1]$
2. *En déduire que* $S \in [-1, 1]$
3. *Quelle information nous est donnée quand S est proche de 1 ? Même question pour 0 et -1.*

Performance d'un clustering

Silhouette score

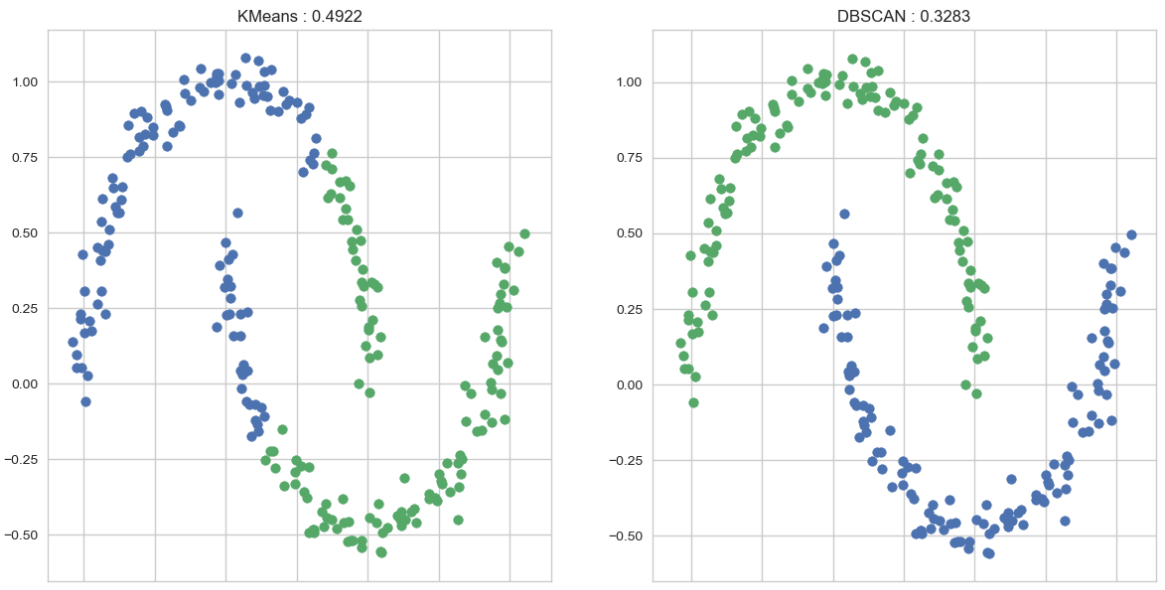


Figure – Comparaison de la valeur du silhouette score entre deux méthodes de clustering

Performance d'un clustering

Index de Calinski-Harabasz

On appelle W_k la matrice de dispersion intra-cluster et B_k la matrice de dispersion inter-cluster, définies par :

$$\begin{aligned} W_k &= \sum_{q=1}^{k} \sum_{x \in \mathcal{C}(q)} (x - C_q)(x - C_q)^T \\ B_k &= \sum_{q=1}^{k} n_q (C_q - C_{\mathcal{D}})(C_q - C_{\mathcal{D}})^T \end{aligned}$$

Finalement on définit le score comme :

$$S = \frac{\mathrm{tr}\,(B_k)}{\mathrm{tr}\,(W_k)} \frac{n_{\mathcal{D}} - k}{k - 1}$$

Performance d'un clustering

Index de Calinski-Harabasz

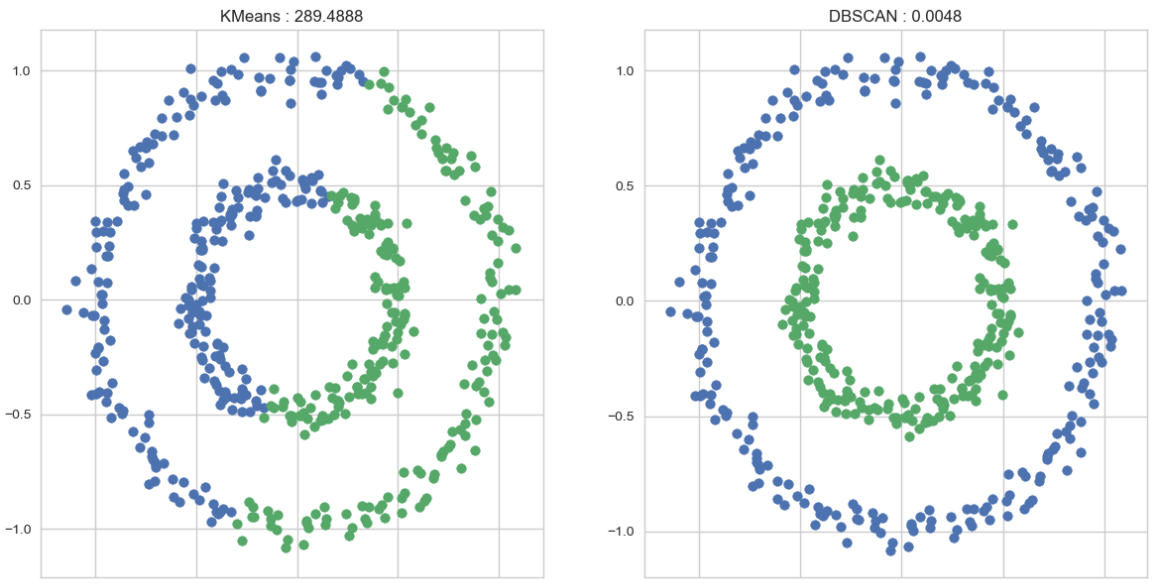


Figure – Comparaison de l'index Calinski-Harabasz pour deux algorithmes de clustering