K-moyenne
fiche d'aide

Arts et Metiers

A quoi sert la méthode des K-moyennes ?

Contrairement à la méthode KNN (voir fiche), utilisée pour de la régression ou de la classification, K-moyenne n'est utilisée que pour partitionner les données.

Comment fonctionne la méthode des K-moyennes ?

K-moyenne est une méthode de clustering, c'est à dire permet de partitionner chaque donnée en sous- groupe, de manière non supervisée.
Fonctionnement : On impose le nombre k de classes. En appelant (C1,... ,Ck) ces classes on note : j le barycentre de Cj et mj le moment d'inertie de Cj. L'objectif va être de minimiser la somme des moments d'inertie. On utilise pour cela un algorithme glouton : On choisit aléatoirement k centres (1,... ,k). Chacun des points du nuage est associé au centre j le plus proche; on crée ainsi k classes (C1,...,Ck), puis on calcule les barycentres (1,... ,k) de ces classes, qui remplacent les valeurs précédentes on reprend les étapes précédentes.

Avantages et Inconvenients

Un modele Kmean presente selon les cas des avantages et des inconvenients.

Avantages

1. Simplicité et facilité de prise en main
2. Rapidité

Inconvenients

1. ne permet pas de trouver des groupes ayant des formes complexes.
2. Le choix du paramétre K est difficile à éstimé et peux faire varié de façon significative les résultats
3. On ne peut l'utiliser que lorsque l'on peut définir la valeur moyenne du cluster, ce qui peut ne pas convenir à certaines applications
4. Lorsque l'on veut appliquer l'algorithme K-means, il est d'abord nécessaire de déterminer une partition initiale basée sur le centre de regroupement initial, puis d'optimiser la partition initiale. La sélection de ce centre de clustering initial a un impact plus important sur les résultats du clustering. Si l'on ne sélectionne pas bien la valeur initiale, on risque de ne pas obtenir de résultats de clustering efficaces

Exemple

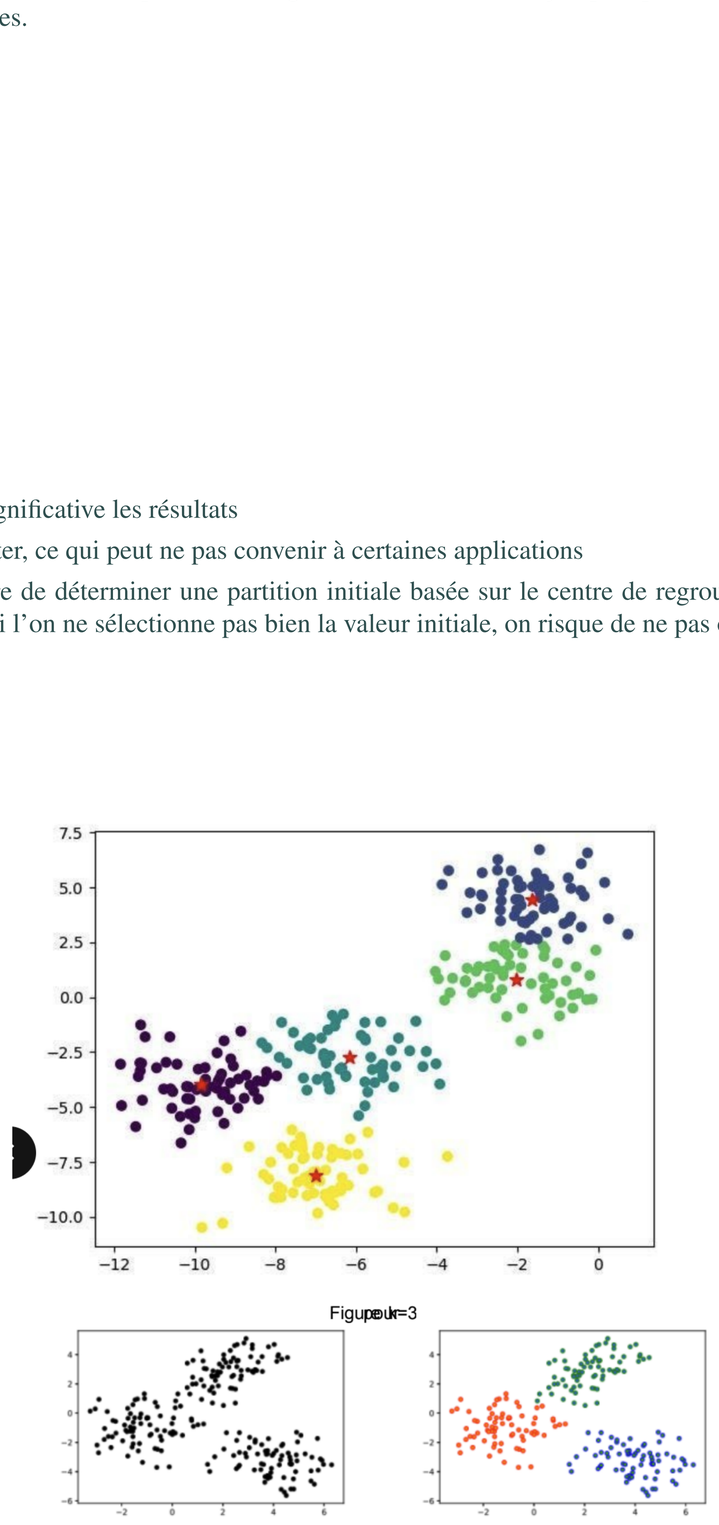

Figure 1: Exemple classification avec sklearn

On peut mettre en place le pseudo-code suivant:

Algorithme des k-moyennes : pseudocode

Algorithme 4 : Algorithme des k moyennes

Data :
- Un ensemble fini de points X d'un espace euclidien
- Un entier k

Result : Une partition de X en parties non vides $X_0, \ldots, X_{k-1}$
Répartir les points de X en sous-ensembles disjoints $X_0, \ldots, X_{k-1}$
while *la situation évolue* **do**
  Calculer $m_0, \ldots, m_{k-1}$ barycentres de $X_0, \ldots, X_{k-1}$;
  for *Chaque point x de X* **do**
    Trouver i tel que $\|x - m_i\|$ est minimal parmi les $\|x - m_j\|$ et placer x dans X_i

Pour aller plus loin

1. Scikit-learn.
2. Wikipedia