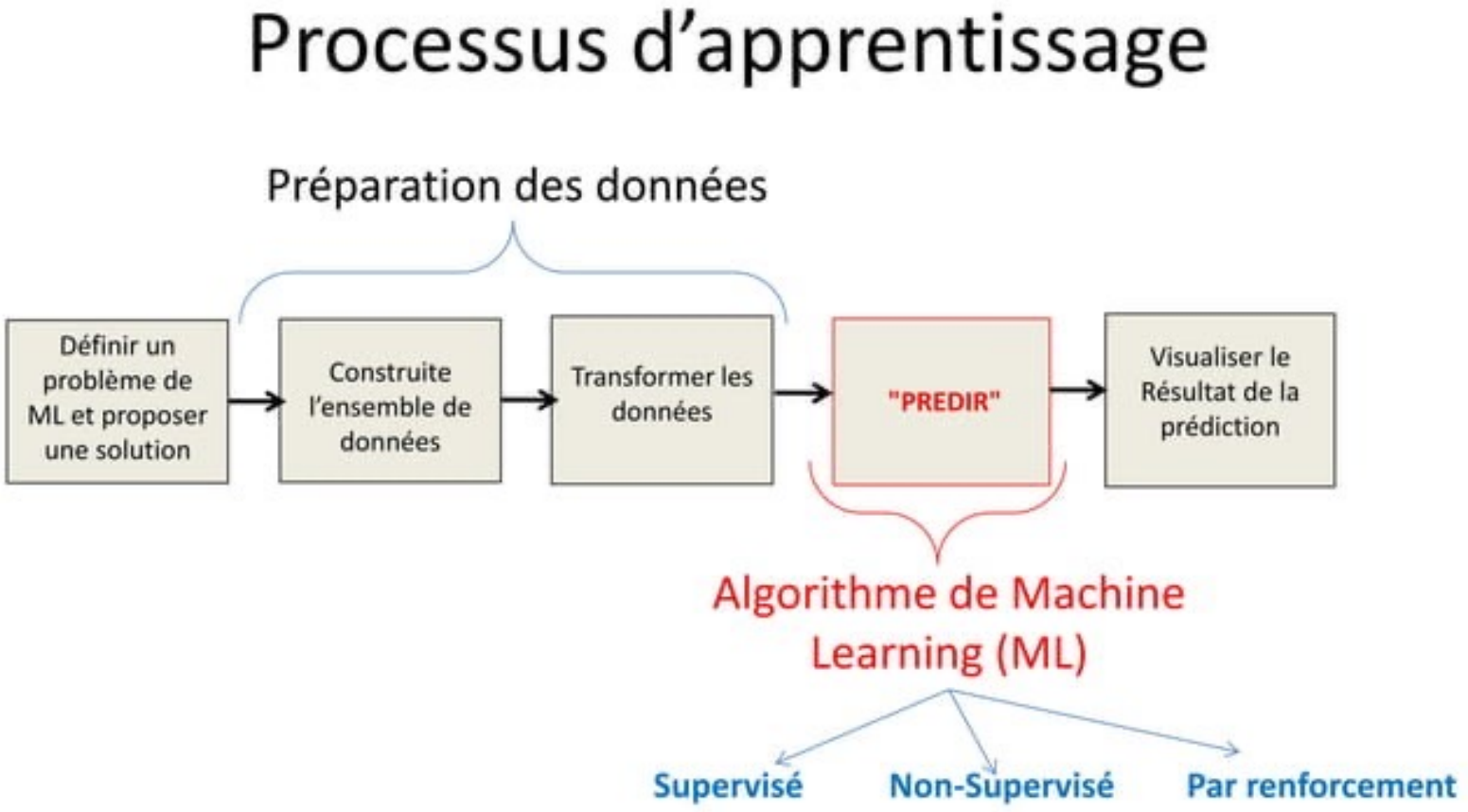
Chapitre 4:
Apprentissage non supervisé
K-means

Pr Amadou Dahirou GUEYE
Université Amadou Mahtar Mbow
Master SISD, Ecole Polytech
Novembre 2024

Plan

- Introduction à l’apprentissage non supervisé et au k-means
- Qu'est-ce que le clustering ?
- Les types de clustering
- Qu'est-ce que K-means ?
- Notion de similarité
- Choisir K : Le nombre de clusters
- Fonctionnement de l'algorithme K-means
- Avantages et limites
- Applications de K-means
- Cas Pratique
- Conclusion

Introduction sur l'apprentissage non supervisé



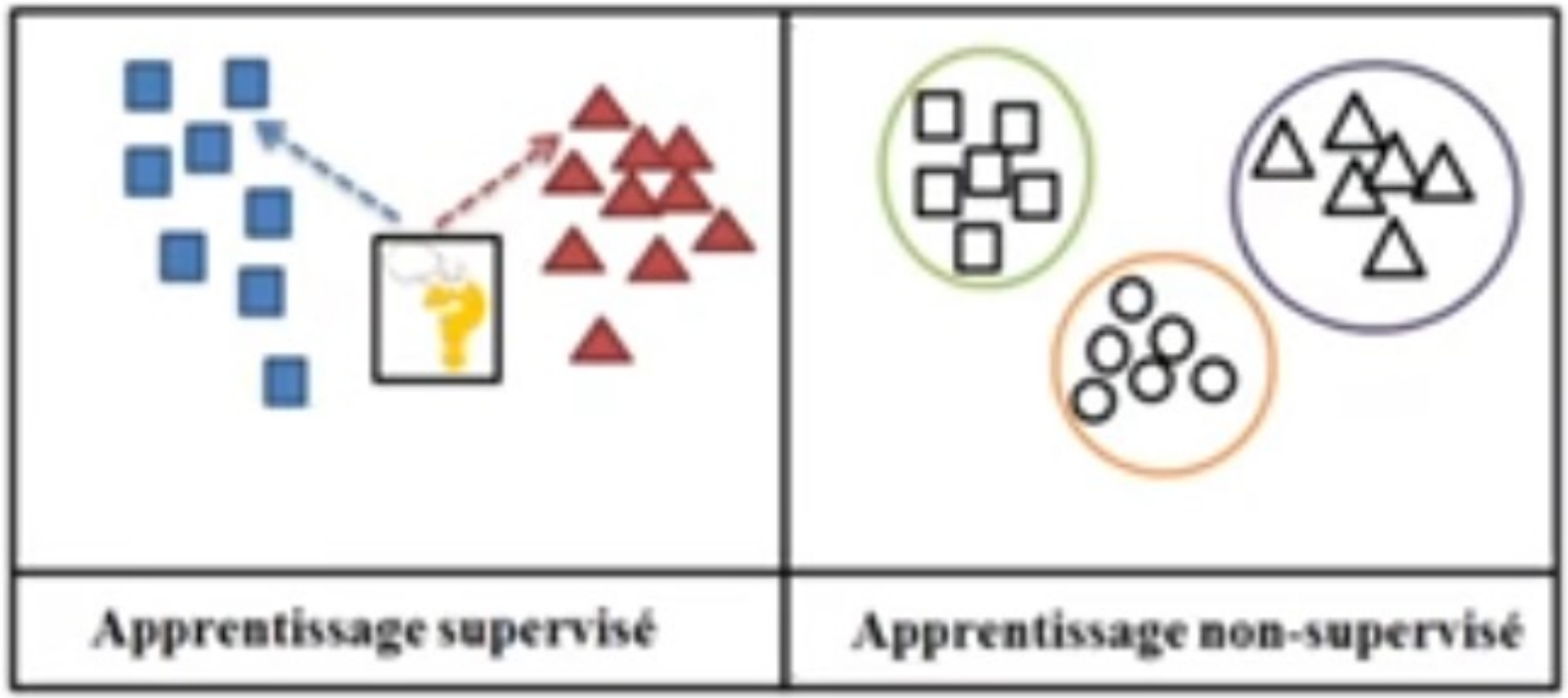
Introduction sur l'apprentissage non supervisé

Différents contextes d'apprentissage

- Les algorithmes de Machine Learning (ML)
  - L'apprentissage non-supervisé
    - Aucun expert n'est disponible. L'algorithme doit découvrir par lui-même la structure des données.
  - L'apprentissage supervisé
    - un expert est employé pour étiqueter correctement des exemples (instances).
  - L'apprentissage par renforcement
    - l'algorithme apprend un comportement.

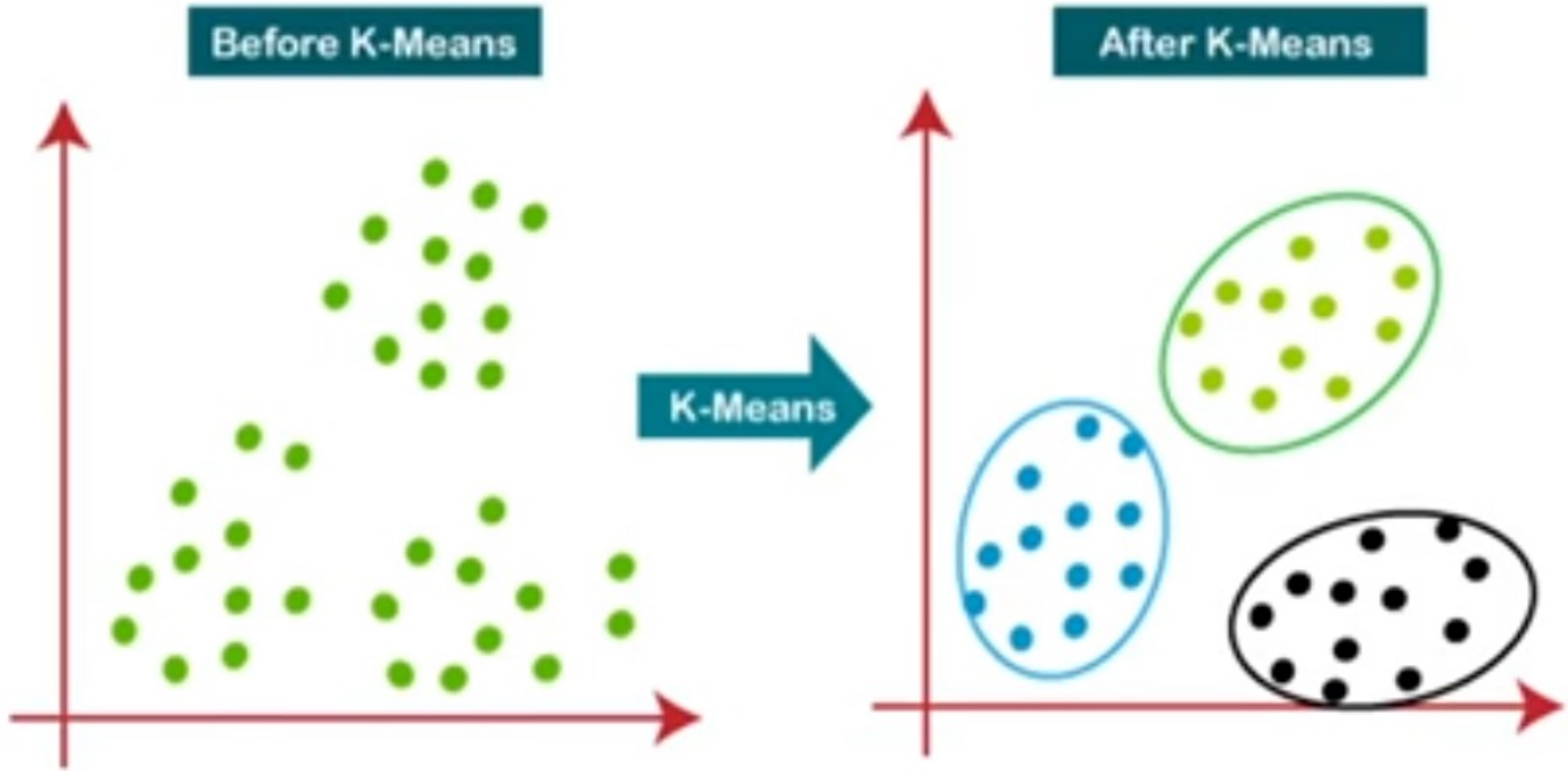
On s'intéresse, dans ce chapitre, au algorithmes d'apprentissage non supervisé.

3


?
Apprentissage supervisé
Apprentissage non-supervisé

Introduction au K-means

K-means est un algorithme non supervisé populaire en Machine Learning. Il est utilisé pour regrouper des données en clusters similaires.


Before K-Means
K-Means
After K-Means

Qu'est-ce que le clustering ?

Le clustering est une méthode d'apprentissage non supervisé (unsupervised learning). Ainsi, on n'essaie pas d'apprendre une relation de corrélation entre un ensemble de features X d'une observation et une valeur à prédire Y.

L'apprentissage non supervisé va plutôt trouver des patterns dans les données. Notamment, en regroupant les choses qui se ressemblent.

Qu'est-ce que le clustering ?

En apprentissage non supervisé, les données sont représentées comme suit :

$$X = \begin{pmatrix} x_{(1,1)} & x_{(1,2)} & x_{(1,...)} & x_{(1,n)} \\ x_{(2,1)} & x_{(2,2)} & x_{(2,...)} & x_{(2,n)} \\ \dots & \dots & \dots & \dots \\ x_{(m,1)} & x_{(m,2)} & x_{(m,...)} & x_{(m,n)} \end{pmatrix}$$

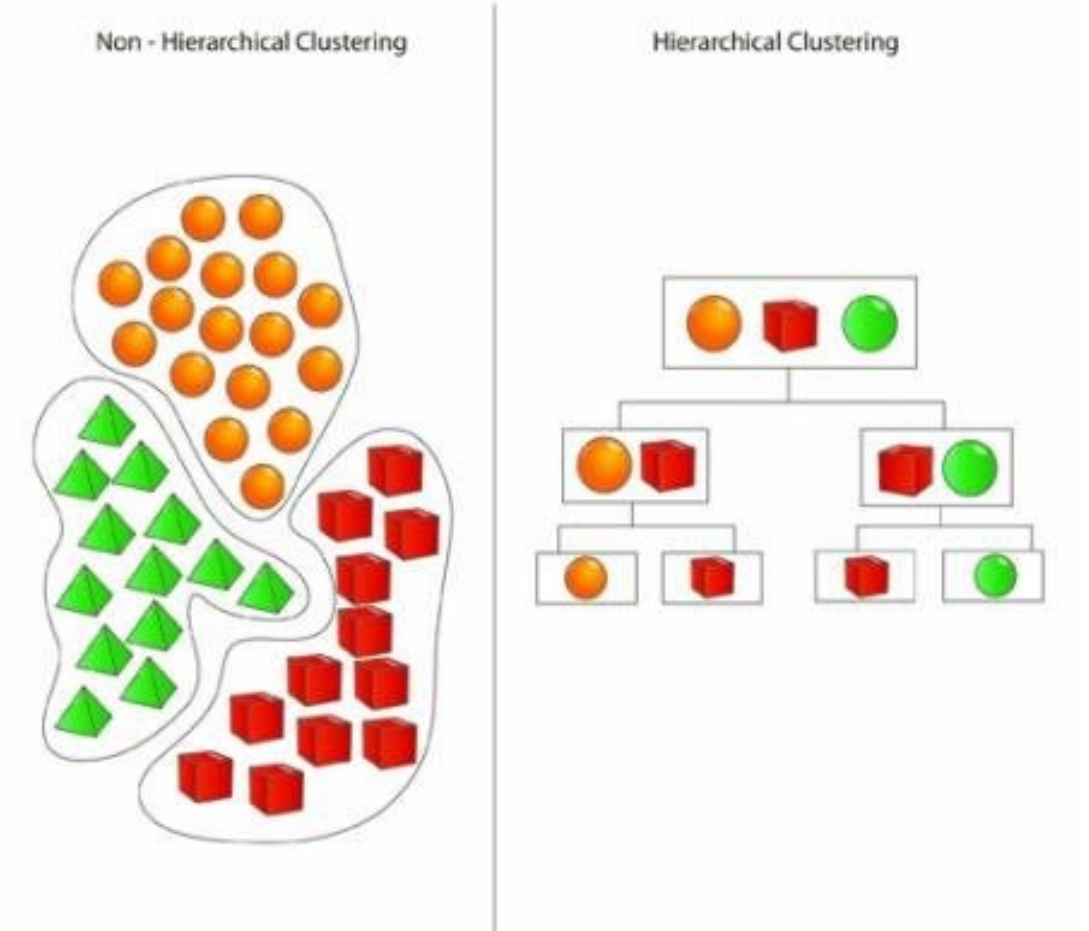
Qu'est-ce que le clustering ?

Il existe deux types de clustering :

- Le clustering hiérarchique
- Le clustering non-hiérarchique (partitionnement)

Qu'est-ce que le clustering ?

Il existe deux types de clustering :



Qu'est-ce que K-means ?

K-means est un algorithme non supervisé de clustering non hiérarchique. Il permet de regrouper en K clusters distincts les observations du dataset. Ainsi les données similaires se retrouveront dans un même cluster.

Notion de similarité

Pour pouvoir regrouper un jeu de données en K cluster distincts, l'algorithme K-Means a besoin d'un moyen de comparer le degré de similarité entre les différentes observations.

Les distances les plus connues pour le cas de clustering sont :

- La distance euclidienne
- La distance de Manhattan

Notion de similarité

La distance Euclidienne :

Soit une matrice X à n variables quantitatives. Dans l'espace vectoriel E^n. La distance euclidienne d entre deux observations x_1 et x_2 se calcule comme suit :

$$d(x_1, x_2) = \sqrt{\sum_{j=1}^{n} (x_{1n} - x_{2n})^2}$$

Notion de similarité

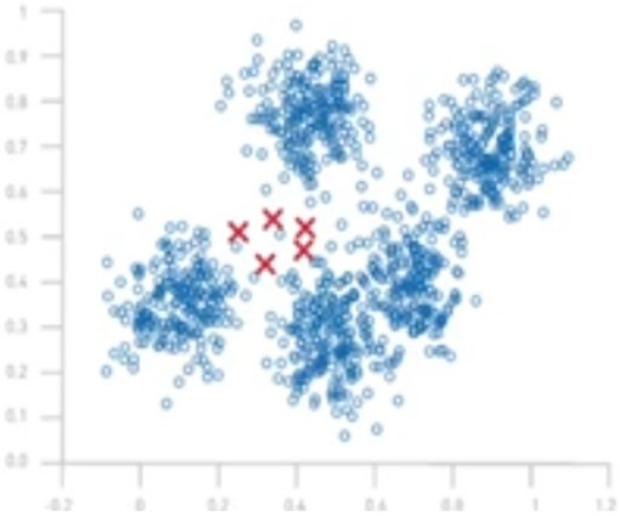
La distance de Manhattan : est la distance entre deux points parcourue.

la distance de Manhattan est donnée par :

$$distance\ Manhttan = \sum_{i=1}^{n} |(x_i - y_i)|$$

Choisir K : Le nombre de clusters

Choisir un nombre de cluster K n'est pas forcément intuitif. Un nombre K grand peut conduire à un partitionnement trop fragmenté des données. Par contre, un nombre de clusters trop petit, conduira à avoir, potentiellement, des cluster trop généralistes contenant beaucoup de données. Dans ce cas, on n'aura pas de patterns à découvrir.



Choisir K : Le nombre de clusters

La difficulté réside sur le fait de trouver un K qui met en lumière des patterns intéressants entre les données.

Il n’existe pas de méthode automatisé pour calculer le nombre de K.

Choisir K : Le nombre de clusters

La méthode qui est souvent utilisée est de choisir un nombre de cluster et de lancer le K-Means avec différentes valeurs de K puis calculer la variance entre les différents clusters.

La variance des clusters se calcule comme suit :

$$V = \sum_j \sum_{x_i \to c_j} D(c_j, x_i)^2$$

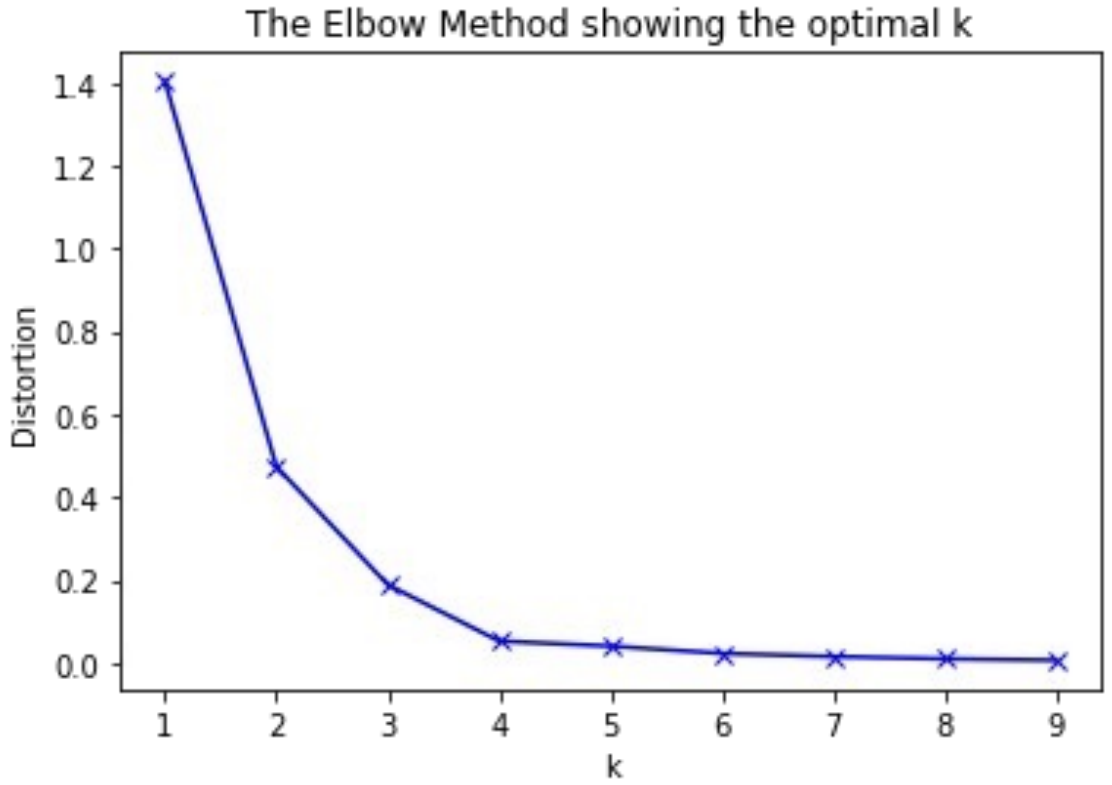
Choisir K : Le nombre de clusters

Avec :

- c_j : Le centre du cluster (le centroïd)
- x_i : la ième observation dans le cluster ayant pour centroïd
- $D(c_j, x_i)$: La distance (euclidienne ou autre) entre le centre du cluster et le point x_i

Choisir K : Le nombre de clusters

En mettant dans un graphique le nombre de cluster K en fonction de la variance on obtient la courbe suivante:

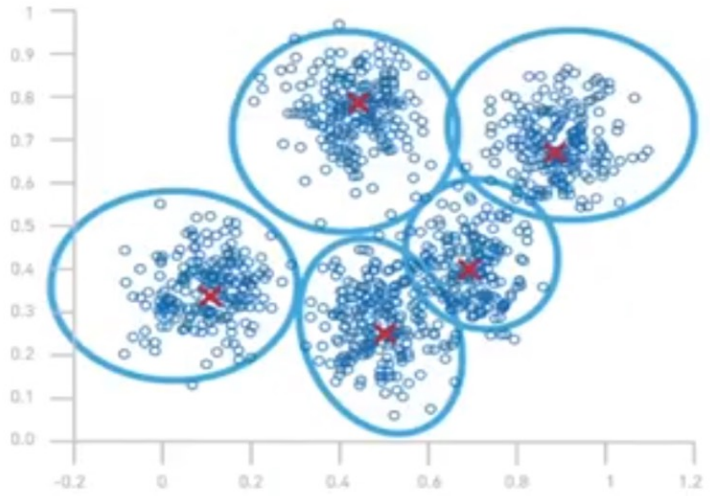


Fonctionnement de K-means

k-means est un algorithme itératif qui minimise la somme des distances entre chaque individu et le centroïd. Le choix initial des centroïdes conditionne le résultat final.

Fonctionnement de K-means

1. Initialisation aléatoire des centroids.
2. Affectation des points au centroid le plus proche.
3. Mise à jour des centroids.
4. Répétition jusqu'à convergence ou stabilisation.



Fonctionnement de K-means

Principe algorithmique

Algorithme K-means

Entrée :

K le nombre de cluster à former

Le Training Set (matrice de données)

Fonctionnement de K-means

DEBUT

Choisir aléatoirement K points (une ligne de la matrice de données).
Ces points sont les centres des clusters (nommé centroïd).

Fonctionnement de K-means

REPETER

Affecter chaque point (élément de la matrice de donnée) au groupe dont il est le plus proche à son centre

Recalculer le centre de chaque cluster et modifier le centroide

JUSQU'A CONVERGENCE OU (stabilisation de l'inertie totale de la population)

FIN ALGORITHME

Avantages et limites

- Avantages : Simple, rapide, adapté aux clusters bien séparés.
- Limites : Sensible à l'initialisation, optimums locaux, nécessite de connaître K.

Applications de K-means

- Segmentation client.
- Clustering de documents (ex. Google Actualités).
- Exploration de données pour détection de similarités.
- Recommandation sur youtube et autres actualités.

Cas Pratique

Un TP sur le K-Means

Réaliser un modèle de K-Means pouvant trouver des similarités sur des données médicales de patients

(Voir dataset et fichier word)

Conclusion

K-means est un algorithme clé en apprentissage non supervisé pour regrouper des données similaires. Bien choisir K est essentiel pour des résultats pertinents.