



Intelligence Artificielle

Apprentissage Non Supervisé – Clustering

K-means

Bezzaz Soumia

2025/2026 – Semestre 2

Qu'est-ce que K-means ?

K-means = algorithme de clustering qui groupe des données similaires ensemble

Définition simple

K-means est un algorithme qui prend un ensemble de points et les divise en **K groupes (clusters)** en fonction de leur **similarité**.

Chaque groupe a un **centre (centroïde)** et tous les points du groupe sont plus proches de ce centre que des autres centres.

Le « **K** » dans K-means représente le **nombre de groupes** que vous voulez créer.

Itératif

L'algorithme répète les mêmes étapes jusqu'à ce que les groupes soient stables.

Très rapide

Fonctionne même sur des millions de points. C'est pourquoi il est si populaire.

Simple à comprendre

L'idée est intuitive : rapprocher les points similaires.

Imaginez que vous devez ouvrir 3 supermarchés dans une ville pour servir tous les habitants

Le problème

- Il y a 100 maisons dans la ville (les points)
- Vous voulez ouvrir $K=3$ supermarchés (les centres)
- Objectif : minimiser la distance moyenne que les gens doivent parcourir

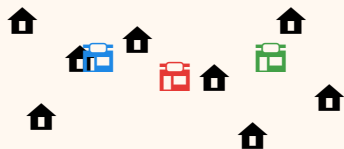
Comment décider où placer les 3 supermarchés ?

La solution K-means

1. Placez 3 supermarchés au hasard dans la ville
2. Chaque maison va au supermarché le plus proche
3. Déplacez chaque supermarché au centre de ses clients
4. Recommencez les étapes 2-3 jusqu'à ce que les supermarchés ne bougent plus

Visualisation :

Étape 1 : Placement initial

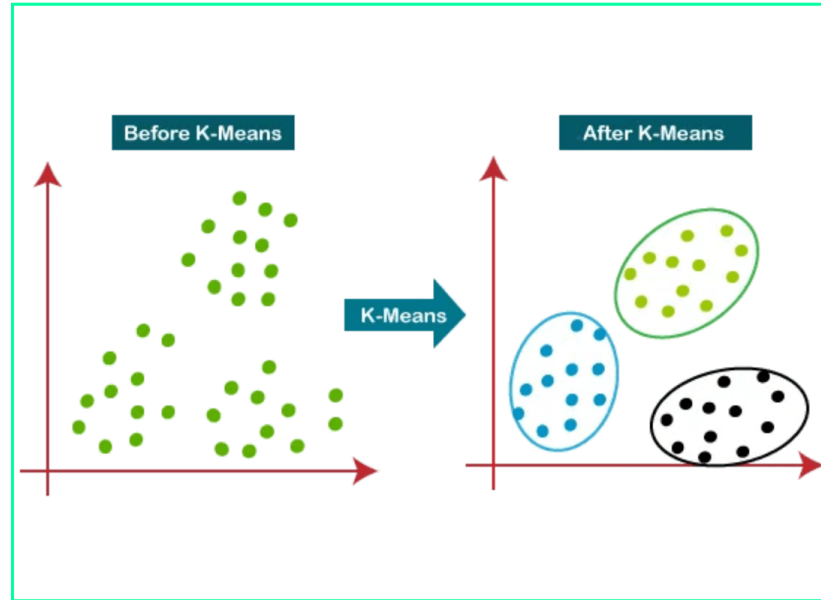


Étape 2 : Chaque maison → son supermarché le plus proche



Étape finale : Supermarchés au centre de leurs zones





Visualisation du regroupement automatique des données

Voici comment l'algorithme fonctionne, étape par étape

1/ INITIALISATION

Placer K centres au hasard

Choisissez K points au hasard dans vos données pour être les premiers centres (centroïdes).

Exemple : Si $K=3$, vous choisissez 3 points au hasard comme centres de départ.

2/ ASSIGNATION

Assigner chaque point au centre le plus proche

Pour chaque point, calculez la distance à tous les centres et assignez-le au centre le plus proche.

Exemple : Si un point est à 5m du centre A, 10m du centre B et 15m du centre C → il rejoint le groupe A.

3/ MISE À JOUR

Recalculer les centres

Pour chaque groupe, calculez le nouveau centre = la moyenne de tous les points du groupe.

4/ RÉPÉTITION

Répéter jusqu'à stabilisation

Répétez les étapes 2 et 3 jusqu'à ce que les centres ne bougent plus.

Applications de K-means dans le monde réel

K-means est utilisé dans des dizaines de domaines différents :

Géosciences & Cartographie

- Segmentation d'images satellites (forêt, eau, urbain)
- Classification de types de sols
- Zonage climatique
- Délimitation de bassins versants

💡 *Sentinel-2 : classer chaque pixel en 5 catégories d'occupation des sols*

Marketing & Business

- Segmentation de clients (comportement d'achat)
- Recommandations de produits
- Détection de fraude
- Analyse de marché

💡 *Amazon regroupe ses clients en 8 segments pour personnaliser les recommandations*

Biologie & Médecine

- Classification de types de cellules
- Analyse génétique
- Segmentation d'images médicales (IRM, scanner)
- Identification de maladies

💡 *Grouper des tumeurs selon leurs profils génétiques pour adapter les traitements*

Technologie & IA

- Compression d'images (réduire les couleurs)
- Détection d'anomalies réseau
- Organisation de documents
- Reconnaissance de patterns

💡 *Réduire une image de 16 millions de couleurs à 64 couleurs principales*

Cas pratique : Segmentation d'image satellite

Exemple concret : classer automatiquement l'occupation des sols d'une région

Le défi

Vous avez une image satellite Sentinel-2 d'une région de 100 km².
L'image contient 10 millions de pixels.

Chaque pixel a 13 bandes spectrales (couleurs invisibles : rouge, vert, bleu, proche-infrarouge, infrarouge moyen...).

Objectif : classer automatiquement chaque pixel en 5 catégories :
Forêt, Eau, Culture, Urbain, Sol nu.

Avant K-means

*Image satellite brute
(13 bandes spectrales)*

Image complexe
avec millions de nuances

Solution avec K-means

1. Fixer K=5 (nos 5 catégories)
2. Chaque pixel = un point à 13 dimensions (les 13 bandes)
3. K-means trouve 5 centres dans l'espace à 13 dimensions
4. Chaque pixel est assigné au centre le plus proche
5. Résultat : une carte colorée avec 5 classes !

Temps de calcul : ~2 minutes pour 10 millions de pixels

Après K-means

*Carte segmentée
(5 classes distinctes)*



Exercice 1 – Comprendre K-means avec un exemple simple

Contexte :

Un géographe étudie 9 villes et mesure leur température moyenne annuelle (en °C). Il veut les regrouper en K=3 zones climatiques (Froide, Tempérée, Chaude) en utilisant K-means.

Voici les températures des 9 villes :

Oslo

5°C

Stockholm

7°C

Helsinki

6°C

Paris

12°C

Berlin

10°C

Madrid

15°C

Athènes

19°C

Le Caire

22°C

Dubaï

27°C

Questions (répondez sans calculs, juste par logique) :

1. Q1. Si K-means place les 3 centres initiaux à 6°C, 12°C et 22°C, dans quel groupe ira chaque ville ?
2. Q2. Après la première itération, où seront les nouveaux centres pour chaque groupe ?
3. Q3. Les 3 zones climatiques finales vous semblent-elles logiques ? Justifiez.

Solution

Q1 – Premier assignment (chaque ville → centre le plus proche)

Groupe 1 (centre 6°C): Oslo, Stockholm, Helsinki

Groupe 2 (centre 12°C): Paris, Berlin, Madrid

Groupe 3 (centre 22°C): Athènes, Le Caire, Dubaï

Q2 – Nouveaux centres (moyenne de chaque groupe)

Groupe 1 : $(5+7+6)/3 = 6^{\circ}\text{C}$ (inchangé ✓)

Groupe 2 : $(12+10+15)/3 = 12.3^{\circ}\text{C}$

Groupe 3 : $(19+22+27)/3 = 22.7^{\circ}\text{C}$

Q3 – Interprétation des zones climatiques

Zone Froide (6°C) : Oslo, Stockholm, Helsinki → Villes nordiques - climat continental froid

Zone Tempérée (12°C) : Paris, Berlin, Madrid → Europe centrale/ouest - climat océanique/continental tempéré

Zone Chaude (23°C) : Athènes, Le Caire, Dubaï → Méditerranée/Moyen-Orient - climat chaud méditerranéen/désertique

Conclusion : K-means a trouvé 3 zones climatiques cohérentes qui correspondent bien à la réalité géographique !

Exercice 2

1. K-means nécessite de fixer K (nombre de groupes) à l'avance.



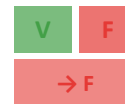
3. L'initialisation aléatoire des centres peut donner des résultats différents.



5. L'algorithme s'arrête lorsque les centres ne se déplacent plus.



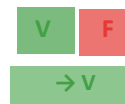
2. Le K-means est un algorithme d'apprentissage supervisé.



4. K-means est lent et ne marche pas sur de grandes données.



6. K-means peut avoir plusieurs itérations avant de converger.



Algorithme de clustering

1

Divise les données en K groupes basés sur la similarité (distance).

4 étapes simples

2

Init → Assign → Update → Repeat. Converge rapidement en général.

K doit être fixé

3

Utiliser la méthode du coude, silhouette ou expertise domaine.

Rapide et scalable

4

Fonctionne même sur des millions de points. Le plus utilisé en pratique.

Applications partout

5

Géosciences, marketing, biologie, tech. Segmentation d'images, classification...