Chapitre II : Clustering (Regroupement) non supervisé : Méthode K-means

Rappel : Distance Euclidienne

I. Rappel sur la Distance Euclidienne

Définition : Dans un espace à deux dimensions, la distance euclidienne entre deux points (x_1, y_1) et (x_2, y_2) est :

$$d\big((x_1, y_1), (x_2, y_2)\big) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}.$$

Interprétation

- Plus la distance est faible, plus les deux points sont proches.
- Cette distance est au cœur de nombreux algorithmes de **clustering**, dont **K-Means**.

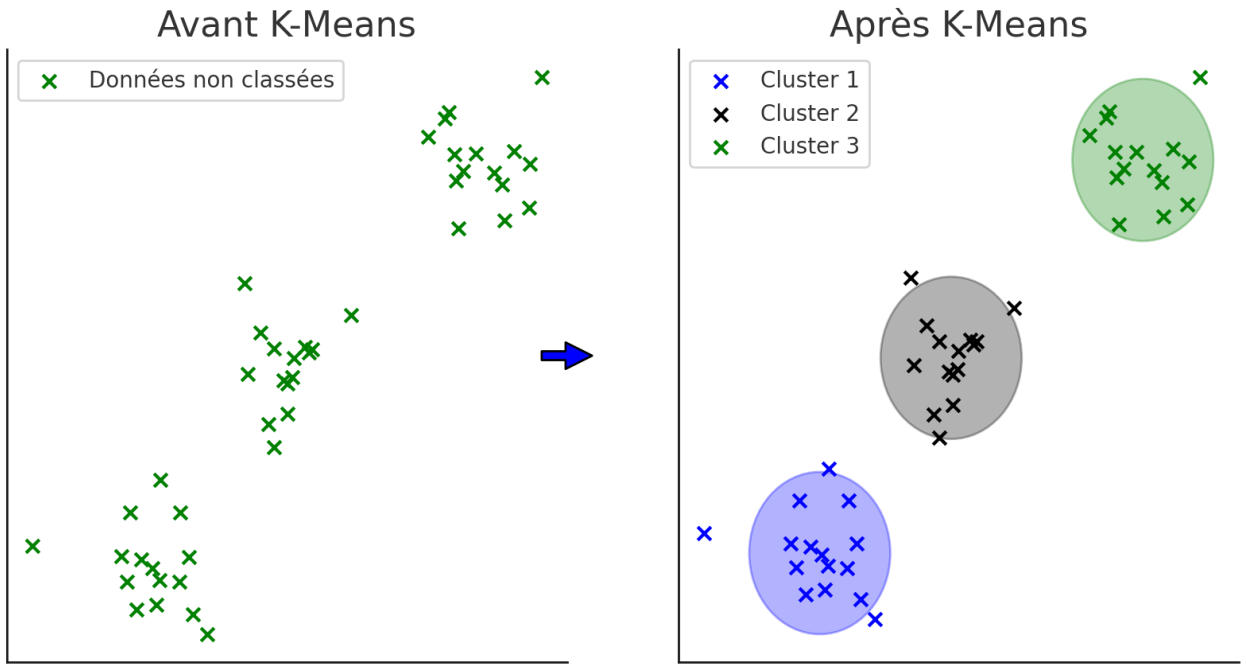
Clustering(regroupement) non supervisé : Contexte

1. Clustering non supervisé

Idée :

- Contrairement à la *classification supervisée* (où l'on connaît des étiquettes *a priori*), le **clustering** regroupe les données selon leurs similarités *sans* information préalable sur les classes.
- L'objectif est de **segmenter** les données en groupes (ou *clusters*) tels que les points d'un même cluster soient **similaires** entre eux et **différents** des points des autres clusters.

Illustration du Clustering



Intérêt du *clustering* :

- **Découverte de patrons cachés** dans les données.
- **Synthèse** et **réduction de la complexité** d'un ensemble de données volumineux.

Présentation de K-Means

2. Algorithme K-Means : Principes

Objectif : Partitionner n observations en K clusters (K étant fixé) de sorte que les points d'un même cluster soient **proches** entre eux.

Idée générale

1. Choisir aléatoirement K centres de clusters (ou **centroïdes**).
2. Affecter chaque point au cluster dont le centroïde est le plus proche (distance euclidienne).
3. Mettre à jour les centroïdes (moyenne des points affectés à chaque cluster).
4. Répéter l'affectation et la mise à jour jusqu'à **convergence**.

Étapes de l'Algorithme K-Means

3 - Algorithme

1. **Initialisation :**
   - Choisir K, le nombre de clusters.
   - Choisir K points initiaux (ou aléatoires) comme centroïdes.
2. **Affectation :** Pour chaque point x_i, calculer la distance à chaque centroïde et l'affecter au plus proche.
3. **Mise à jour :** Recalculer chaque centroïde comme la **moyenne** des points qui lui sont affectés.
4. **Itération :** Répéter affectation/mise à jour jusqu'à stabilisation (plus de changement ou changement minime).

Condition d'arrêt

L'algorithme s'arrête quand les centroïdes ne bougent plus.

Exemple : Application de K-Means

4. Exemple de base : Itérations de K-Means

Données : On considère 6 points 2D :

$$\{(2;2), (2.5;2), (3;2.2), (7.5;7), (8;8)\}$$

et on souhaite les regrouper en $K = 2$ clusters.
Initialisation (Itération 0) : Choix (arbitraire) de deux Centroïdes initiaux :

$$C_1^{(0)} = (2;2), \quad C_2^{(0)} = (3;2.2)$$

Itération 1 : Affectation des points

Point	**Distance à $C_1^{(0)}$**	**Distance à $C_2^{(0)}$**	**Cluster**
(2,2)	0.0	1.02	1
(2.5,2)	0.5	0.54	1
(3,2.2)	1.02	0.0	2
(7.5,7)	7.43	6.58	2
(8,8)	8.49	7.65	2

Table – Affectation des points aux clusters après Itération 1

$$\text{Clusters : } \begin{cases} C_1 = \{(2,2), (2.5,2)\} \\ C_2 = \{(3,2.2), (7.5,7), (8,8)\} \end{cases}$$

Itération 1 : Mise à jour des centroïdes

-Recalculer chaque centroïde comme la **moyenne** des points qui lui sont affectés.

$$C_1^{(1)} = \left(\frac{2+2.5}{2}, \frac{2+2}{2}\right) = (2.25, 2)$$

$$C_2^{(1)} = \left(\frac{3+7.5+8}{3}, \frac{2.2+7+8}{3}\right) = (6.17, 5.73)$$

Nouveaux centroïdes :

$$C_1^{(1)} = (2.25, 2), \quad C_2^{(1)} = (6.17, 5.73)$$

Itération 2 : Nouvelle affectation

Point	**Dist. à $C_1^{(1)}$**	**Dist. à $C_2^{(1)}$**	**Nouveau cluster**
(2,2)	0.25	5.60	1
(2.5,2)	0.25	5.23	1
(3,2.2)	0.78	4.75	**1**
(7.5,7)	7.25	1.84	2
(8,8)	8.31	2.92	2

Table – Nouvelle affectation après Itération 2

Changement clé : $(3, 2.2)$ passe de **Cluster 2** $\rightarrow$ **Cluster 1**.

Itération 2 : Mise à jour des centroïdes

-Recalculer chaque centroïde comme la **moyenne** des points qui lui sont affectés.

$$C_1^{(2)} = \left(\frac{2+2.5+3}{3}, \frac{2+2+2.2}{3}\right) = (2.5, 2.07)$$

$$C_2^{(2)} = \left(\frac{7.5+8}{2}, \frac{7+8}{2}\right) = (7.75, 7.5)$$

Nouveaux centroïdes :

$$C_1^{(2)} = (2.5, 2.07), \quad C_2^{(2)} = (7.75, 7.5)$$

Itération 3 : Vérification des distances (Convergence)

Dernière itération : On recalcule les distances pour vérifier qu'aucun point ne change de cluster.

$$C_1^{(2)} = (2.5, 2.07), \quad C_2^{(2)} = (7.75, 7.5).$$

Point	**dist**$(C_1^{(2)})$	**dist**$(C_2^{(2)})$	**Cluster final**
(2,2)	≈ 0.50	≈ 7.96	1
(2.5,2)	≈ 0.07	≈ 7.60	1
(3,2.2)	≈ 0.52	≈ 7.12	1
(7.5,7)	≈ 7.02	≈ 0.56	2
(8,8)	≈ 8.09	≈ 0.56	2

Table – Distances finale à l'itération 3

Aucun point ne change de cluster : Convergence atteinte.

Itération 3 : Convergence

Affectation finale des points :

$$\begin{cases} C_1 = \{(2,2), (2.5,2), (3,2.2)\} \quad \text{(Centroïde final : } (2.5, 2.07)) \\ C_2 = \{(7.5,7), (8,8)\} \quad \text{(Centroïde final : } (7.75, 7.5)) \end{cases}$$

Conclusion

L'algorithme K-Means converge en **3 itérations**, après que le point $(3, 2.2)$ ait changé de cluster.

Enoncé de l'Exercice

Exercice d'Application : Application de K-Means

Objectif : Regrouper les points suivants en $K = 2$ **clusters** en utilisant l'algorithme K-Means :

$$\{(1,1),(2,1),(4,3),(5,4),(6,4),(6,6)\}$$

Centroïdes initiaux :

$$C_1^{(0)} = (1,1), \quad C_2^{(0)} = (5,4)$$

Questions

En utilisant l'algorithme K-Means :

1. Attribuez chaque point au **cluster le plus proche**.
2. Recalculez les **nouveaux centroïdes** après mise à jour.
3. Répétez l'opération jusqu'à convergence.
4. Quels sont les **clusters finaux** et leurs centroïdes ?
5. Comment un **choix différent des centroïdes initiaux** pourrait-il influencer le résultat ?

Conclusion sur K-Means

Bilan :

- **K-Means** est une méthode itérative qui converge généralement rapidement.
- Il est sensible à l'initialisation et suppose le nombre de clusters K connu.
- Malgré ses limites, il est très populaire pour la **segmentation** de données (marketing, imagerie, etc.).