K-Means : de la théorie mathématique à la pratique

1 Introduction

Le partitionnement en **k-moyennes** (ou *k-means*) est une méthode de clustering largement utilisée en apprentissage non supervisé. Son objectif est de regrouper un ensemble de données en k clusters homogènes et compacts.

2 Formulation mathématique du problème

Soit un ensemble de données :

$$X = \{x_1, x_2, \ldots, x_n\}, \quad x_i \in \mathbb{R}^d$$

On cherche à partitionner ces points en k clusters :

$$C = \{C_1, C_2, \ldots, C_k\}$$

Chaque cluster est représenté par un centroïde :

$$\mu_j \in \mathbb{R}^d$$

2.1 Fonction objectif (inertie intra-cluster)

L'algorithme K-Means cherche à minimiser la somme des distances intra-cluster :

$$\boxed{J = \sum_{j=1}^{k} \sum_{x_i \in C_j} \|x_i - \mu_j\|^2}$$

3 Intuition géométrique

Chaque cluster est représenté par son centre de gravité. Les points sont assignés au centroïde le plus proche, ce qui induit une partition de type **diagramme de Voronoï**.

4 Algorithme K-Means

4.1 Étape 1 : Initialisation

Choisir k centroïdes initiaux :

$$\mu_1^{(0)}, \mu_2^{(0)}, \ldots, \mu_k^{(0)}$$

4.2 Étape 2 : Calcul des distances

Pour chaque point x_i et chaque centroïde μ_j, on calcule une distance $d(x_i, \mu_j)$.

5 Fonctions de distance

5.1 Distance euclidienne (L2)

$$\boxed{d(x,y) = \sqrt{\sum_{l=1}^{d}(x_l - y_l)^2}}$$

Version au carré (souvent utilisée) :

$$d^2(x,y) = \sum_{l=1}^{d}(x_l - y_l)^2$$

5.2 Distance de Manhattan (L1)

$$d(x,y) = \sum_{l=1}^{d}|x_l - y_l|$$

5.3 Distance de Minkowski

$$d(x,y) = \left(\sum_{l=1}^{d}|x_l - y_l|^p\right)^{1/p}$$

5.4 Distance de Chebyshev

$$d(x,y) = \max_{l}|x_l - y_l|$$

5.5 Distance cosinus

$$d(x,y) = 1 - \frac{x \cdot y}{\|x\|\|y\|}$$

6 Étape 3 : Assignation des points

Chaque point est affecté au cluster dont le centroïde est le plus proche :

$$\boxed{C(x_i) = \arg\min_{j}\|x_i - \mu_j\|^2}$$

7 Étape 4 : Mise à jour des centroïdes

Le centroïde est recalculé comme la moyenne des points du cluster :

$$\boxed{\mu_j = \frac{1}{|C_j|}\sum_{x_i \in C_j} x_i}$$

7.1 Justification mathématique

Cette moyenne minimise la somme des distances au carré :

$$\frac{\partial}{\partial \mu_j} \sum_{x_i \in C_j} \|x_i - \mu_j\|^2 = 0$$

8 Étape 5 : Critère d'arrêt

L'algorithme s'arrête si :

— les centroïdes ne changent plus :

$$\|\mu_j^{(t)} - \mu_j^{(t-1)}\| < \varepsilon$$

— les clusters restent inchangés
— le nombre maximal d'itérations est atteint

9 Convergence

À chaque itération, la fonction objectif J décroît. L'algorithme converge donc toujours, mais pas nécessairement vers le minimum global.

10 Complexité algorithmique

$$\boxed{O(n \cdot k \cdot d \cdot i)}$$

où :

— n : nombre de points
— k : nombre de clusters
— d : dimension des données
— i : nombre d'itérations

11 Hypothèses et limites

K-Means suppose :

— des clusters sphériques
— des variances similaires
— des données numériques
— peu d'outliers

12 Conclusion

K-Means est un algorithme simple, rapide et efficace, reposant sur une formulation mathématique élégante. Il reste une méthode de référence en clustering non supervisé.