

Prof. Mohamed Aly LOULY

Data Science et Machine Learning

1

Classification automatique non-supervisée

Catégorisation ou Clustering

CLASSIFICATION AUTOMATIQUE NON-SUPERVISÉE

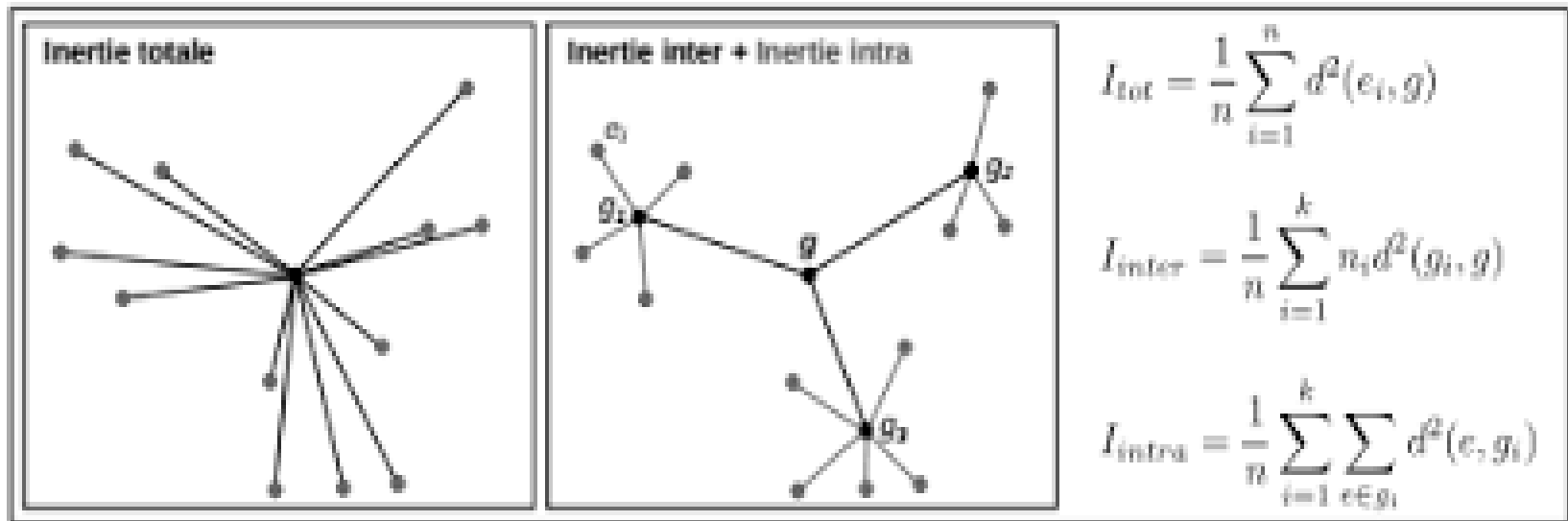
L'objectif des méthodes de la classification automatique non-supervisée est de construire des catégories (**clusters**) d'individus, à partir de leurs **similarités**, pour faciliter l'**exploration statistique** des données multidimensionnelles.

Cette catégorisation (**clustering**) opère un **partitionnement** ou une **classification hiérarchique** d'un ensemble d'individus à partir de l'observation de plusieurs variables.

Les méthodes de **partitionnement** supposent connu le nombre de classes à trouver alors que la **classification hiérarchique** propose un regroupement progressif de tous les individus.

THÉORÈME DE HUYGENS

L'inertie totale peut être décomposée en inertie intra-classes et inertie inter-classes.



LA MÉTHODE K-MEANS

La méthode des k-moyennes est une méthode de partitionnement donc l'objectif est de grouper n individus dans m classes homogènes à partir de l'observation de p variables.

Les observations sont données par une matrice X de taille (n, p) . Les lignes de cette matrice représentent les individus et les colonnes représentent les variables. Le coefficient x_{ij} désigne l'observation de la variable j sur le l'individu i .

La méthode se base sur un algorithme de minimisation locale de l'intértie intra-classes pour ce problème NP-complet.

ALGORITHME DE LA MÉTHODE K-MEANS

- Générer aléatoirement des centres initiaux, $g_1^0, \dots, g_m^0$.
- Tant que l'inertie intra-classes diminue :
 - Réaliser une partition P_t en affectant chaque individu i à la classe C_k^t dont le centre g_k^{t-1} est le centre le plus proche de i .
 - Trouver les centres de gravités, $g_1^t, \dots, g_m^t$, des classes $C_1^t, \dots, C_m^t$ et calculer l'inertie intra-classes correspondante :

$$g_k^t = \frac{1}{n_k} \sum_{i \in C_k^t} Z_i \quad \text{et} \quad I_W^t = \frac{1}{n} \sum_{k=1}^m \sum_{i \in C_k^t} d^2(i, g_k^t)$$

CONVERGENCE DE L'ALGORITHME

On montre facilement que l'inertie intra-classes est décroissante dans l'algorithme de K-means : $I_W^t \geq I_W^{t+1}$.

En effet, la partition P_{t+1} est constituée en affectant chaque individu i à la classe générée par le centre g_k^t le plus proche. Nous avons donc bien :

$$I_W^t = \frac{1}{n} \sum_{k=1}^m \sum_{i \in C_k^t} d^2(i, g_k^t) \geq \frac{1}{n} \sum_{k=1}^m \sum_{i \in C_k^{t+1}} d^2(i, g_k^t)$$

CONVERGENCE DE L'ALGORITHME

Le théorème de Huygens montre ensuite facilement le reste :

$$\begin{aligned}
 \frac{1}{n} \sum_{k=1}^m \sum_{i \in C_k^{t+1}} d^2(i, g_k^t) &= \frac{1}{n} \sum_{k=1}^m \sum_{i \in C_k^{t+1}} \|Z_i - g_k^t\|^2 \\
 &= \frac{1}{n} \sum_{k=1}^m \sum_{i \in C_k^{t+1}} \|Z_i - g_k^{t+1} + g_k^{t+1} - g_k^t\|^2 \\
 &= I_W^{t+1} + \frac{1}{n} \sum_{k=1}^m n_k^{t+1} \|g_k^{t+1} - g_k^t\|^2 \geq I_W^{t+1}
 \end{aligned}$$

CONVERGENCE DE L'ALGORITHME

La **convergence** est donc bien assurée car l'inertie intra-classes est décroissante et l'ensemble des partitions possibles est fini. L'observation empirique montre que cette convergence est généralement **rapide**.

La qualité du résultat de cet algorithme de **recherche locale** dépend toutefois des centres initiaux, aléatoirement générés. Il est donc raisonnable de **lancer plusieurs fois l'algorithme** avec différents points initiaux.

CLASSIFICATION ASCENDANTE HIÉRARCHIQUE

Cette méthode de classification non-supervisée réalise une suite imbriquée de partitions en **agrégeant graduellement tous les individus**.

La **partition initiale** est celle de n classes, composées, chacune, d'un seul individu.

La **partition finale** est composée d'une seule classe regroupant tous les individus.

Chaque itération de l'algorithme consiste à grouper les deux classes les plus proches.

CLASSIFICATION ASCENDANTE HIÉRARCHIQUE

L'objectif est de grouper progressivement n individus sur lesquels nous disposons de l'observation de p variables.

La fusion à chaque itération des deux classes les plus proches nécessite l'introduction d'une **distance** ou une **dissimilarité** entre les classes.

L'algorithme commence avec une partition en n **singletons** et prend $n-1$ itérations pour arriver à une partition composée d'une **seule classe complète**.

CRITÈRES D'AGRÉGATION

L'agrégation de deux classes A et B augmente l'inertie intra-classes de la quantité :

$$\text{Ind}(A, B) = \frac{n_A n_B}{n_A + n_B} \|g_A - g_B\|^2$$

Le critère de Ward consiste donc à grouper les deux classes pour lesquelles cette quantité est minimale afin de minimiser l'augmentation de l'inertie intra-classes à chaque itération.

CRITÈRES D'AGRÉGATION

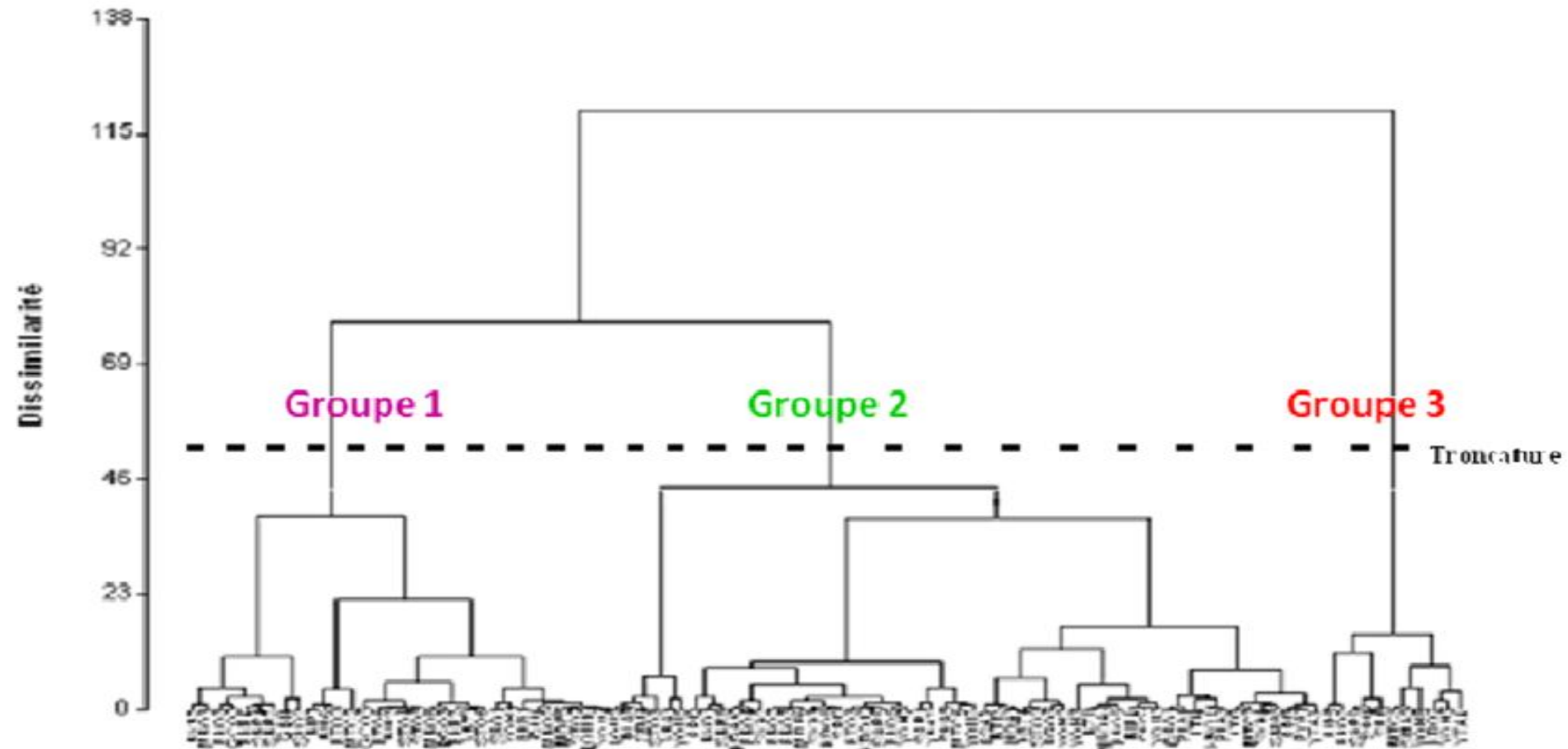
Le centre de la classe C qui résulte du groupement de A et B est :

$$g_C = \frac{n_A}{n_A + n_B} g_A + \frac{n_B}{n_A + n_B} g_B$$

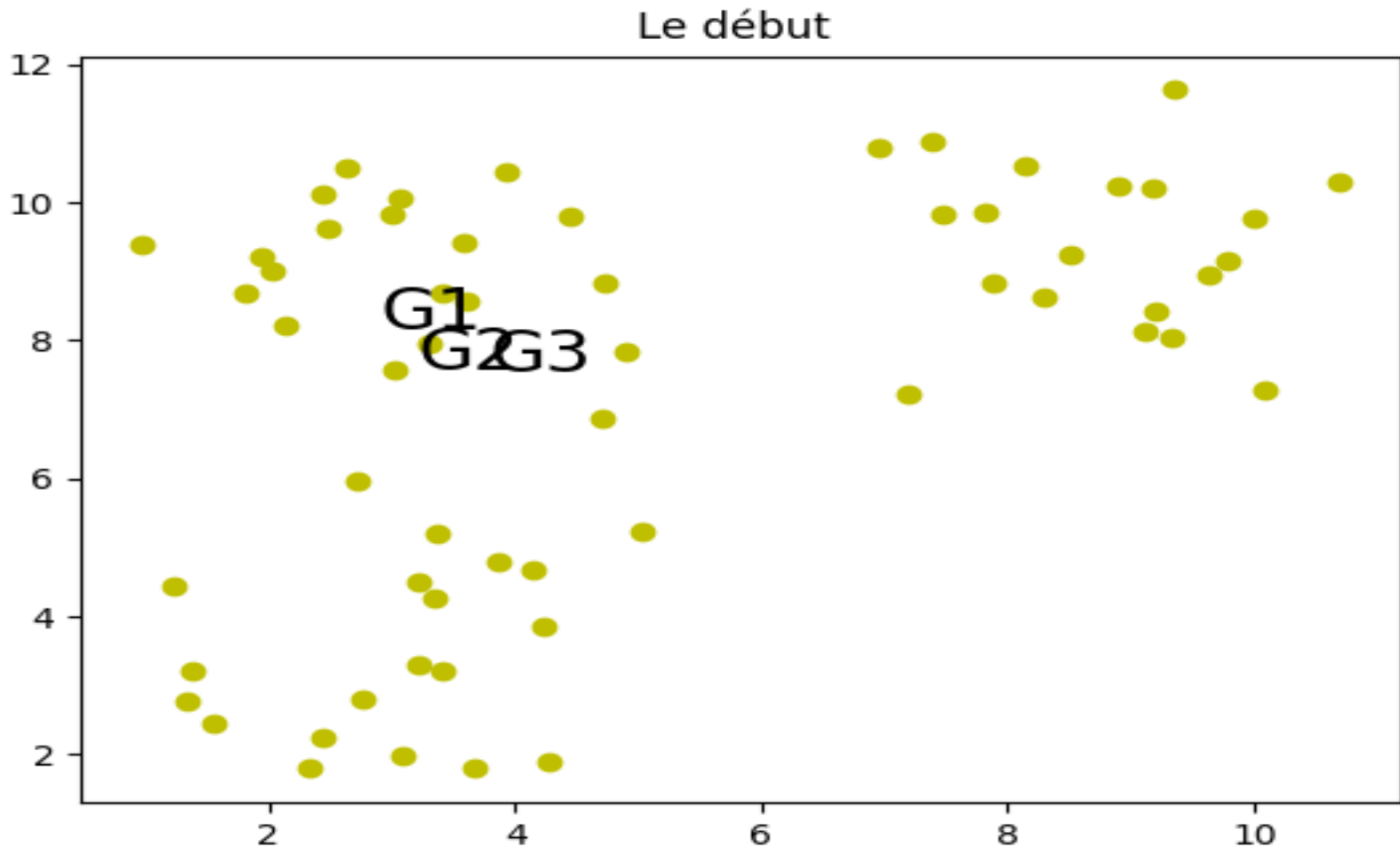
La distance entre un point i et le centre g_C de la classe C peut être déduit des distances de i avec les centres de A et B :

$$\|i - g_C\|^2 = \frac{1}{n_A + n_B} (n_A \|i - g_A\|^2 + n_B \|i - g_B\|^2 - \text{Ind}(A, B))$$

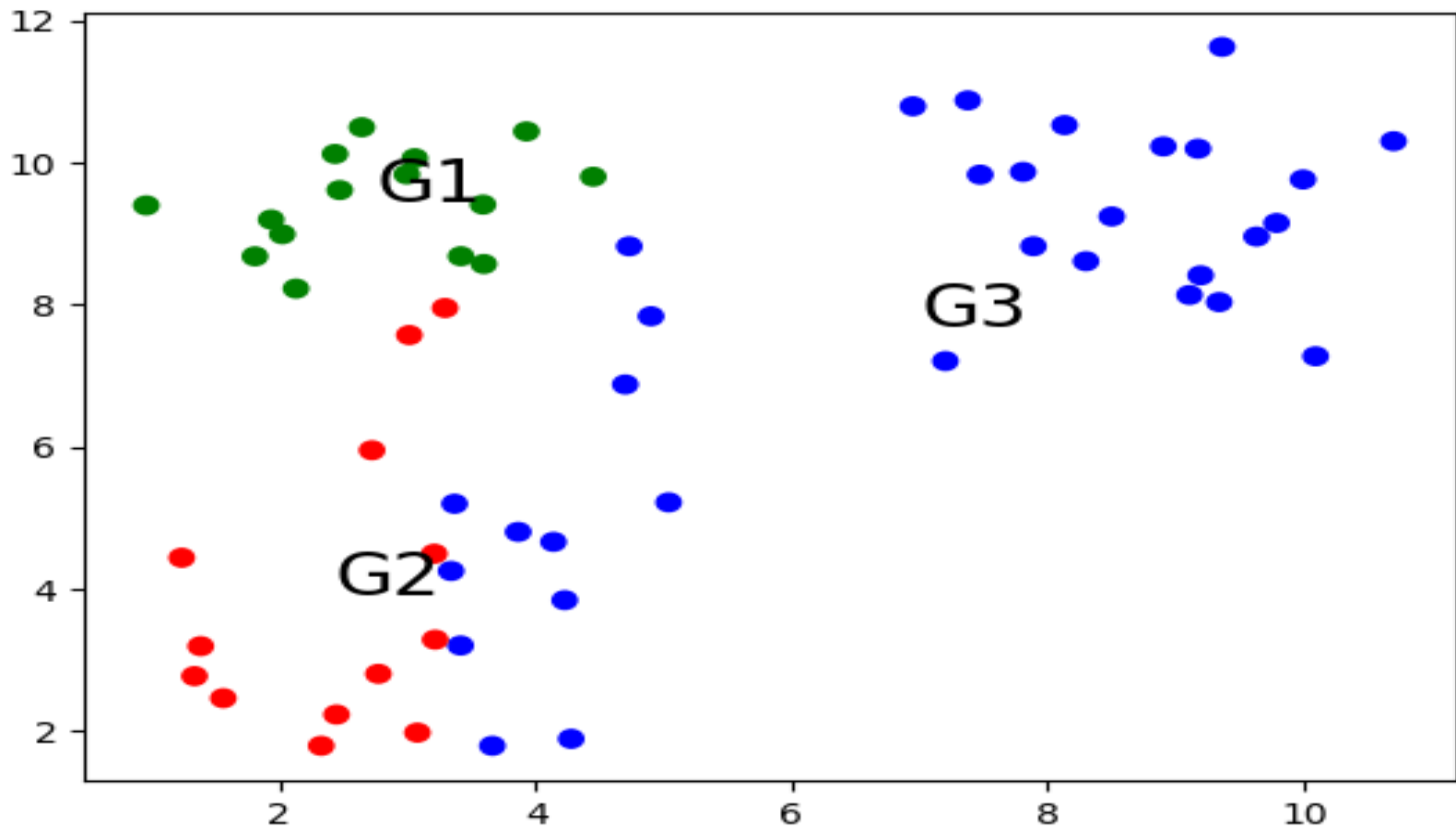
DENDROGRAMME



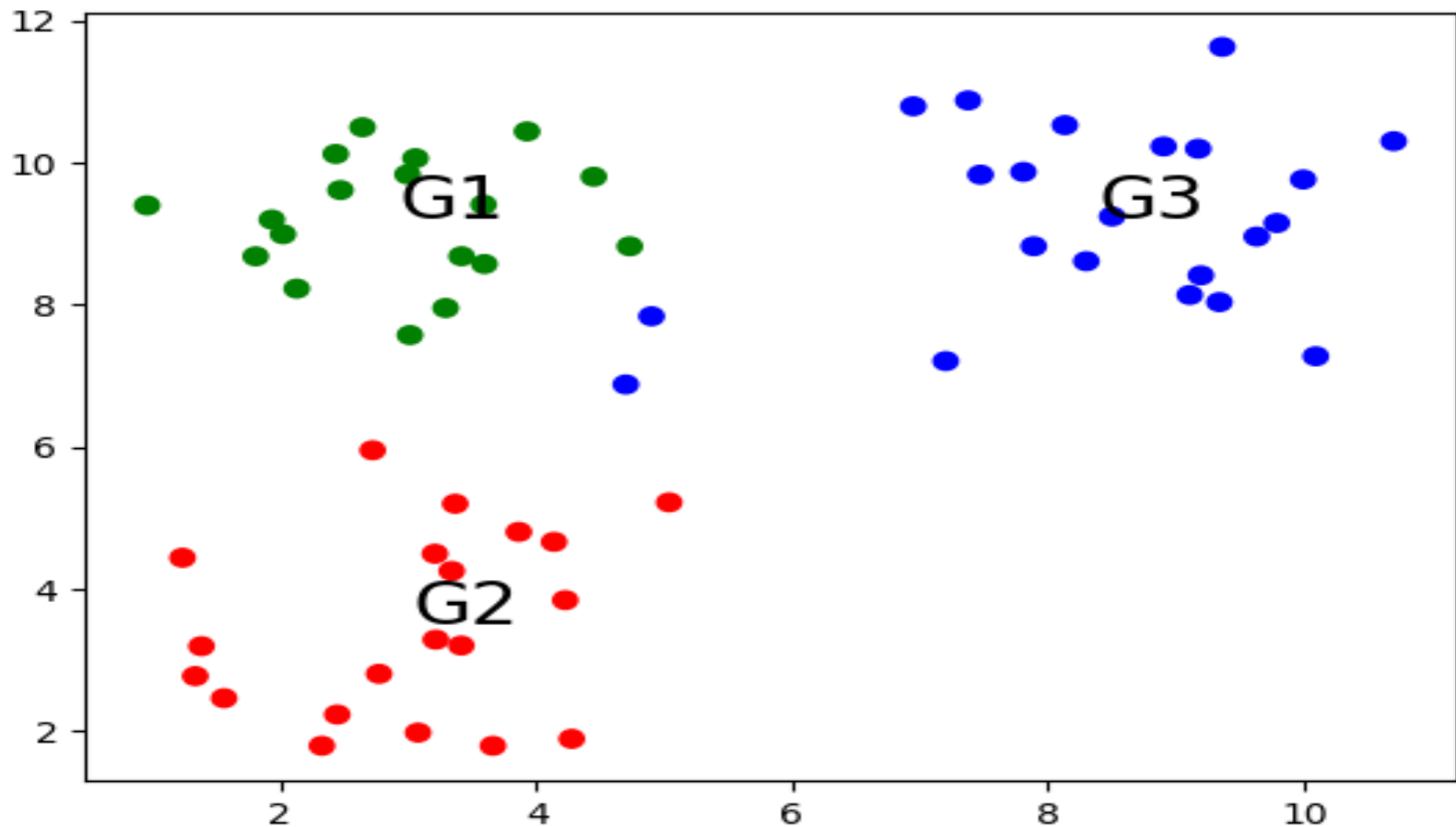
EXEMPLE D'INITIALISATION DE LA MÉTHODE K-MEANS



ALGORITHME K-MEANS APRÈS UNE SEULE ITÉRATION



K-MEANS APRÈS 2 ITÉRATIONS



CONVERGENCE APRÈS 3 ITÉRATIONS

