

TP 5 : k-means / FCM

Le **clustering** est une technique d'apprentissage automatique non supervisée. Les techniques de clustering peuvent être regroupées principalement en deux classes : clustering dur (Hard clustering) et clustering souple ou flou (soft clustering). Dans le clustering dur, chaque observation est mise dans un seul cluster. Dans le clustering flou, chaque observation peut appartenir à plusieurs clusters avec un degré d'appartenance (une probabilité).



Fuzzy C-means (FCM) est une méthode de clustering flou non supervisé très populaire, dérivée de l'algorithme k-means. FCM est fréquemment utilisée dans le traitement d'images, la reconnaissance de formes. Dans cet algorithme, chaque cluster est représenté par son centre de gravité.

<i>K-means</i>	<i>FCM</i>
Initialisation du nombre de cluster k	1. Initialisation du nombre de cluster C Initialisation de ϵ , le degré de flou m ($1,25 < m < 2$)
Initialisation aléatoire des centres	2. Initialisation aléatoire des centres
MAJ des centres $\bar{x}_{kj} = \frac{1}{ C_k } \sum_{i \in C_k} x_{ij}$	3. Mettre à jour la matrice U des degrés d'appartenance : $U = \begin{bmatrix} u_{11} & \dots & u_{1k} & \dots & u_{1n} \\ u_{i1} & \dots & u_{ik} & \dots & u_{in} \\ u_{c1} & \dots & u_{ck} & \dots & u_{cn} \end{bmatrix} \begin{matrix} k = 1, \dots, n \\ i = 1, \dots, c \end{matrix}$ $u_{ik} = \left[\sum_{j=1}^c \left(\frac{d_{ik}}{d_{jk}} \right)^{\frac{2}{m-1}} \right]^{-1}, \quad u_{ik} = \left[\sum_{j=1}^c \left(\frac{\ X_k - V_i\ }{\ X_k - V_j\ } \right)^{\frac{2}{m-1}} \right]^{-1}$
	4. MAJ le vecteur V des centres des clusters $v_i = \frac{\sum_{k=1}^n (u_{ik})^m x_k}{\sum_{k=1}^n (u_{ik})^m}$
Répéter jusqu'à convergence	5. Répéter les étapes 3 et 4 jusqu'à convergence $\ V^t - V^{t+1}\ < \epsilon$ et le nombre d'itération

De grandes valeurs de m brouilleront les classes et tous les éléments ont tendance à appartenir à tous les clusters. Les solutions du problème d'optimisation dépendent du paramètre m . Autrement dit, différentes sélections de m conduisent généralement à des partitions différentes.

Le terme u_{ij} représente le degré d'appartenance de x_j à la classe i . La matrice U est appelée matrice de C-partitions floues si et seulement si elle satisfait aux conditions :

$$\forall i \in \{1..C\}, \forall j \in \{1..N\} \begin{cases} u_{ij} \in [0,1] \\ 0 < \sum_{j=1}^N u_{ij} < N \end{cases} \quad \forall j \in \{1..N\} \sum_{i=1}^C u_{ij} = 1$$

Travail demandé

1. Ecrire une fonction python qui permet d'implémenter k-means. N'utiliser pas la fonction prédéfinie dans cette question.
2. Ecrire une fonction python qui permet d'implémenter l'algorithme FCM. N'utiliser pas la fonction prédéfinie dans cette question.
3. Appliquer les deux algorithmes pour la segmentation d'une image de votre choix.
4. En utilisant les fonctions prédéfinies K-means et FCM, comparer les résultats.
5. Tester l'effet de la valeur de m sur le résultat du clustering (1-10).

Bon courage