

Série 3 en Fouille de données (Clustering)

Exercice 1: On veut regrouper les utilisateurs d'une application Android sur la base des notes qu'ils ont attribué à l'application selon deux critères: fonctionnalités et fiabilité. L'évaluation de chaque critère se fait sur une échelle de 1 à 5. Les données suivantes représentent les évaluations de chaque utilisateur pour chaque critère:

A (1, 1), B (5, 4), C (2, 1), D (4, 4), E (1, 2), F (4, 5), G (1, 3) et H (3, 5)

Travail demandé

1. En utilisant la technique de K-means, regrouper les utilisateurs en deux groupes avec les utilisateurs E et G comme seed initiaux et la distance de Manhattan pour mesurer les distances. Effectuer deux itérations seulement. (La représentation graphique des données est recommandée.)
2. Proposer un label (nom) à chaque groupe selon son contenu.
3. Selon les groupes obtenus à la deuxième itération, de combien d'itérations supplémentaires a-t-on besoin pour stabiliser les clusters? Justifier la réponse.
4. Peut-on améliorer la convergence vers les groupes finaux par un meilleur choix des seeds? Justifier la réponse.

Exercice 2: Soit trois individus ayant subi un test de positivité par rapport à 5 virus. Les tableaux suivants illustrent les résultats des tests

Individu	Test1	Test2	Test3	Test4	Test5
I_1	1	1	0	1	0
I_2	0	0	1	0	0
I_3	1	0	1	1	1

Question: Si nous devons isoler un individu de ces trois sur la base de leur similitude, quel sera l'individu à isoler? Justifier la réponse par les calculs nécessaires.

Exercice 2 : Soit la liste suivante des voitures avec la puissance et leur poids.

Voiture	V1	V2	V3	V4	V5
Poids en kg	1500	1675	1559	1557	1564
Puissance	130	165	150	150	140

1. Donner le tableau des valeurs standardisées des variables.
2. Regrouper les individus avec les valeurs standardisées en utilisant la classification hiérarchique ascendante et en prenant comme mesure de similarité la distance euclidienne, et ce pour chacun des cas suivants : (a) linkage single, linkage complet, comparaison entre centroïdes.