

Série N°3 en Fouille de données (Clustering)

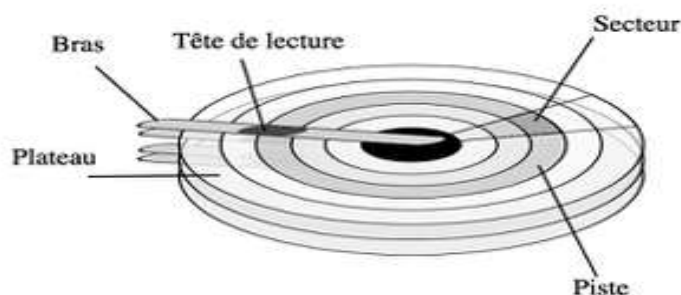
Exercice n°1 : Soit 10 individus caractérisés par deux variables comme suit :

A(1, 0), B (1, 1), C(2, 1), D (3, 1), E (2, 2), F(5, 3), G (4, 3), H(4, 4), I (4, 5), J (3, 4)

1. Donner la représentation graphique des individus sous forme de nuage de points.
2. En utilisant la technique des K-means, regrouper les individus en **deux** clusters avec la distance de **Manhattan** et les points **B** et **C** comme centroïdes initiaux avec **deux** itérations seulement.
3. En vous basant sur la représentation graphique des points, quel est le nombre de choix de centroïdes initiaux qui permettent
 - a. de converger rapidement vers les clusters finaux.
 - b. de converger lentement vers les clusters finaux.

Exercice n°2 : Soit 07 fichiers stockés dans des blocs de données sur un disque de stockage à raison d'un fichier par bloc. Les coordonnées relatives des blocs sont les suivantes : A (6, 1), B (8, 1), C (6, 3), D (8, 3), E (1, 3), F (3, 3) et G (2, 6).

1. Donner la représentation graphique des blocs par rapport à l'origine (0, 0).
2. En utilisant le clustering par K-means et la distance de Manhattan, regrouper les blocs en deux groupes en prenant les blocs B et G comme seeds initiaux.
3. Dans la réalité, les coordonnées réelles des blocs correspondent aux couples « piste / secteur » du disque. On veut lire les fichiers en passant d'un fichier au fichier le plus proche. Pour atteindre un fichier, on fait tourner le disque et on lit le fichier par une tête de lecture (voir figure).



Question : Comment se fait le calcul de distance entre les fichiers si

- Le bras possède une tête de lecture par piste ?
- Le bras possède une seule tête de lecture ?