USTHB,
FI, Département IASD
M2, Master SII
Module : Data mining

TD 5

Exercice 1

Considérer le dataset suivant contenant 10 instances et 5 attributs nommés A, B, C, D et E.

	A	B	C	D	E
I1	1	4	13	2	3
I2	1	2	12	0	7
I3	1	3	13	2	6
I4	1	4	11	2	7
I5	1	4	14	2	7
I6	0	4	15	2	7

1) Rappeler l'algorithme k-means en fournissant les précisions sur :
    a. Le choix des centroides
    b. Le calcul de la similarité en tenant compte des types des attributs
    c. Le choix du paramètre k
2) Rappeler la complexité de l'algorithme.
3) Appliquer l'algorithme k-means sur le dataset pour k = 2 et en démarrant avec les instances I2 et I4 comme centroides initiaux. Considérer tous les types des attributs comme des entiers.

Exercice 2

Considérer les valeurs suivantes d'un attribut âge :

14, 16, 17, 17, 20, 21, 21, 22, 23, 23, 26, 26, 26, 26, 31, 34, 34, 36, 36, 36, 36, 37, 41, 46, 47, 53, 69.

1) Répartir les données en trois intervalles selon chacune des méthodes suivantes :
    a. Partitionnement à largeur égale.
    b. k-means : prendre respectivement 20, 23 et 37 comme centroides.
    c. Commenter la performance des différents résultats obtenus.

Exercice 3

Considérer un repère euclidien avec les 9 points suivants :
$A_1(2, 10)$; $A_2(4, 9)$; $A_3(5, 8)$; $A_4(5, 9)$; $A_5(8, 5)$; $A_6(8, 4)$; $A_7(7, 5)$; $A_8(6, 4)$; $A_9(7, 4)$ où (x, y) représente les coordonnées du point.
Rappelons que DBSCAN est un algorithme de classification non supervisée basé densité. Il rassemble des régions voisines et forme des clusters sur la base de 2 paramètres qui sont respectivement :

- le voisinage d'un point **p** défini dans un rayon **r** et qui est égal à l'ensemble des points éloignés de **p** à une distance inférieure ou égale à **r**.

- la densité d'un point **p** définie par le nombre de points appartenant à son voisinage, **MinPts** étant le nombre de points minimal exigé pour former un cluster.

1) Ecrire la procédure qui détermine le voisinage d'un point dans un rayon égal à **r** ?
2) Rappeler l'algorithme DBSCAN. Quelle est sa complexité ?
3) Avec **r** = 1 unité, déterminer le voisinage de chacun des points donnés précédemment.
4) Avec **MinPts** = 2, appliquer l'algorithme DBSCAN sur les 9 points.
5) Schématiser l'exécution de l'algorithme sur un plan euclidien en montrant clairement les étapes de l'algorithme.
6) Déterminer le bruit engendré par DBSCAN.

Exercice 4

Parmi les méthodes de classification non supervisée hiérarchiques, nous avons étudié la méthode agglomérative dite AGNES (AGglomerative NESting).

1) Rappeler l'algorithme AGNES. Quelle est sa complexité ?
2) Appliquer l'algorithme sur les 9 points $A_1(2, 10)$; $A_2(4, 9)$; $A_3(5, 8)$; $A_4(5, 9)$; $A_5(8, 5)$; $A_6(8, 4)$; $A_7(7, 5)$; $A_8(6, 4)$; $A_9(7, 4)$.
3) Schématiser le résultat de l'algorithme sur un plan euclidien.
4) Dessiner le dendrogramme des clusters.
5) Spécifier les clusters qui sont à une distance de séparation supérieure à $\sqrt{5}$.
6) Rappeler l'algorithme DIANA.
7) Reprendre les mêmes données que celles de la question 2 et Appliquer l'algorithme DIANA.