



Série de TD N°4

Apprentissage & Classification

Non-Supervisés

M. Taffar

Dpt Informatique. Université de Jijel

Cours pour Master 1 –IA, nov.-21

Exercice 1

1. Dites sur quel principe se base un algorithme de classification automatique non-supervisée.
2. Quel est l'inconvénient de l'algorithme **K-means** ?
3. Quel est l'inconvénient majeur de l'algorithme **de regroupement hiérarchique** ?
4. Quelle est la différence majeure entre l'algorithme **K-means** et le **regroupement hiérarchique ascendant** ?
5. Dans un tableau, comparer les classifieurs **K-means** et **groupement hiérarchique** en donnant leurs avantages et leurs inconvénients ?
6. Dans un tableau, donner **3 différences majeures** entre les **classifieurs K-nn et K-means**.

Exercice 2

- Soient 8 objets de la nature suivants, décrits chacun par 2 paramètres : $A_1 = (2, 10)$; $A_2 = (2, 5)$; $A_3 = (8, 4)$; $A_4 = (5, 8)$; $A_5 = (7, 5)$; $A_6 = (6, 4)$; $A_7 = (1, 2)$ et $A_8 = (4, 9)$.
- La matrice des distances Euclidienne entre les différents objets mutuels est donnée ci-après :

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8
A_1	0	$\sqrt{25}$	$\sqrt{36}$	$\sqrt{13}$	$\sqrt{50}$	$\sqrt{52}$	$\sqrt{65}$	$\sqrt{5}$
A_2		0	$\sqrt{37}$	$\sqrt{18}$	$\sqrt{25}$	$\sqrt{17}$	$\sqrt{10}$	$\sqrt{20}$
A_3			0	$\sqrt{25}$	$\sqrt{2}$	$\sqrt{2}$	$\sqrt{53}$	$\sqrt{41}$
A_4				0	$\sqrt{13}$	$\sqrt{17}$	$\sqrt{52}$	$\sqrt{2}$
A_5					0	$\sqrt{2}$	$\sqrt{45}$	$\sqrt{25}$
A_6						0	$\sqrt{29}$	$\sqrt{29}$
A_7							0	$\sqrt{58}$
A_8								0

Exercice 2 (suite)

1. Utilisez **l'algorithme *k-means*** et la **distance Euclidienne** pour regrouper en **3 groupes** les 8 exemples. Au début les centres de chaque groupe sont A_1 , A_4 et A_7 . Exécuter l'algorithme *k-means* pour un nombre suffisant de période (itération).
 - A la fin de chaque période montrez :
 - a) Les nouveaux groupes d'exemples.
 - b) Les nouveaux centres.
 - c) Dessinez les 8 points (exemples) dans un espace 10x10 puis montrer les groupes et leurs nouveaux centroïdes.
 - d) Mettez à jour les distances des objets par rapport aux nouveaux centroïdes.
 - e) Dites combien d'itérations sont nécessaires pour converger ?

Exercice 2 (suite)

2. Appliquez l'algorithme du regroupement hiérarchique pour réaliser la classification des 8 éléments et tracez à chaque étape le dendrogramme correspondant, selon les critères suivants :
- a) à *saut minimal*
 - b) à *diamètre maximal*
 - c) le **critère de Ward** (inertie)
 - d) le **Lien moyen** (consiste à calculer la moyenne des distances entre les individus des groupes)

Choisissez convenablement le critère d'arrêt, soit, par exemple, un seuil de 6 ou nombre d'itérations égal à 3, pour obtenir le meilleur clustering possible. **Comparez.**

Exercice 3

- Ecrire un algorithme qui met à jour une table de distances entre objets et/ou clusters obtenus après **groupement hiérarchique**, selon le **critère de diamètre maximal**, pour le cas suivant :

d	1	2	3	4	5
1	0	5	4	5	6,5
2		0	17	2	20,5
3			0	13	0,5
4				0	14,5
5					0

Table 1 (Avant clustering hiérarchique)

d	1	2	4	c_6
1	0	5	5	6,5
2		0	2	20,5
4			0	14,5
c_6				0

Table 2 (Après clustering, table 1 mise à jour)

- Groupement hiérarchique ascendant pour un indice d'agrégation $f(c_6) = d(3, 5) = 0,5$, avec un nouveau cluster $c_6 = \{3, 5\}$.

Exercice 3 (suite)

- L'algorithme doit comprendre 3 fonctions :
 - fonction qui calcul la distance minimale (niveau d'agrégation $f(c_6) = 0,5$) entre éléments (objets ou clusters), exemple, $d = 0,5$ dans la table 1 ;
 - fonction qui détermine le nouveau groupe (ex., $c_6 = \{3, 5\}$) ;
 - Une fonction qui met à jour la table selon le **diamètre maximal** (c'est-à-dire, la fonction qui détermine la distance du nouveau groupe c_6 aux autres objets restants), ex., table 2.

Exercice 4

- Un algorithme de *regroupement hiérarchique* comprend essentiellement 4 procédures :
 - ✓ Initialisation.
 - ✓ Déterminer la distance minimale.
 - ✓ Comparaison des distances et regroupement des éléments ou Arrêt (avec seuil).
 - ✓ Mise à jour de la matrice des distances.
- En prenant comme éléments à classer les **N** points $a_i = (x_i, y_i)$ d'un espace 2D et en utilisant la distance Euclidienne entre les points. **Faites convenablement les déclarations nécessaires et écrivez un algorithme qui permet de réaliser un *regroupement hiérarchique avec saut minimal*.**

Exercice 5

- Soient 6 objets de la nature représentés par 2 valeurs (x_i, y_i) chacun : $A_1 = (2, 10)$; $A_2 = (2, 5)$; $A_3 = (8, 4)$; $A_4 = (5, 8)$; $A_5 = (7, 5)$ et $A_6 = (1, 2)$.
 1. Calculer la matrice des **distances Manhattan** entre les différents objets mutuels.
 2. Utilisez l'algorithme ***k-means*** et la **distance Manhattan** pour regrouper en 3 groupes les 6 exemples. Au début les centres de chaque groupe sont A_1 , A_4 et A_6 .
 - A chaque itération :
 - Préciser les nouveaux groupes et leurs nouveaux centres.
 - Dessinez dans un espace 10x10, les 6 points (exemples) puis montrer les groupes et leurs nouveaux centroïdes.
 - Mettez à jour les distances des objets par rapport aux nouveaux centroïdes.
 - Dites s'il y a convergence ou pas ?

Exercice 5 (suite)

3. Appliquez l'algorithme de **groupement hiérarchique** à critère **saut minimal**, selon les deux cas ci-dessous, pour déterminer le nombre de classes du problème, grouper les 6 éléments et reconnaître leurs classes. Tracez à chaque étape le dendrogramme correspondant.
 - Nombre d'itérations = 3.
 - Seuil d'agrégation = 9,5.
4. Comparer les deux méthodes : **k-means** et **regroupement hiérarchique**. Que peut-on dire ?

Exercice 6

- Soient 8 objets de la nature définis chacun par 2 valeurs (x_i, y_i) : $A_1 = (2, 10)$; $A_2 = (2, 5)$; $A_3 = (8, 4)$; $A_4 = (5, 8)$; $A_5 = (7, 5)$; $A_6 = (1, 2)$; $A_7 = (1, 12)$ et $A_8 = (8, 10)$.

La distance de *Manhattan* entre le point $A = (a_1, a_2)$ et $B = (b_1, b_2)$ est donnée par :
 $1 - dist(A, B) = \sum_{i=1}^n |a_i - b_i|$

1. Calculer la matrice des distances de Manhattan entre les différents objets mutuels.
2. Utilisez ***k-means*** et la **distance Manhattan** pour regrouper en 3 *groupes* les exemples. Au début les centroïdes de chaque groupe sont : A_1 , A_4 et A_6 . Exécuter l'algorithme *k-means* pour un nombre suffisant d'itérations.

A chaque itération :

- Précisez dans un tableau, les *nouveaux groupes* et (recalculer) leurs *nouveaux centres* respectifs.
- Dessinez, dans un espace 12x12, les 8 points (objets) et montrez les groupes et leurs centres.
- Mettez à jour la table des distances des objets par rapport aux nouveaux centroïdes.
- Dites combien d'itérations sont nécessaires pour converger ?

Exercice 6 (suite)

3. En utilisant **3 itérations**, appliquez l'algorithme du regroupement hiérarchique à critère :

- *saut minimal, puis*
- *Critère de Ward (variance)*

pour grouper (classifier) des 8 objets et tracez à chaque itération le dendrogramme correspondant.

4. Comparez les deux méthodes. Que peut-on dire ?

Exercice 7

1. Décrivez de deux manières (probabiliste et géométrique) l'algorithme de clustering *Mean-Shift*.
2. Donner un exemple simple pour expliquer son fonctionnement.
3. Programmer l'algorithme *Mean-shift* sur un ensemble de n points (formant l'ensemble d'apprentissage) du plan 2D euclidien orthonormé (O, i, j) .

Exercice 8

- Soit l'ensemble d'apprentissage composé des mots échantillon d'un langage L restreint.

$L = \{"lune", "mer", "maladie", "bateau", "avion", "pafum"\}$

- En se basant sur
 1. la **distance de Levenshtein** et le **critère du diamètre maximal**,
 2. la **distance de Gower** (le vecteur descripteur est la chaîne de caractères, et la comparaison se fait caractère par caractère)

utiliser la méthode du **groupement hiérarchique ascendant**, en **4 itérations**, pour déterminer les clusters de mots ; puis tracer le dendrogramme de ce *Clustering*.