Ministère de l'Enseignement supérieur et de la recherche scientifique
Université Abderrahmane Mira de Bejaïa
Faculté des Sciences Exactes
Département de Mathématiques

Module: Data Mining **Année: 2022-2023**
Nature de document: Cours
Niveau : L3STID **N.BERMAD**
Chapitre: 4
Titre: Apprentissage non supervisé

II. Clustering (la classification non supervisée / segmentation):

II.1 Introduction

Pour résoudre certains problèmes complexes, il peut s'avérer utile de commencer par segmenter la population (la diviser en groupes) en espérant que le problème soit alors plus simple à résoudre sur les groupes ainsi constitués. La segmentation est une tâche d'apprentissage *" non supervisée "* car on ne dispose d'aucune autre information préalable que la description des exemples.

II.2 Formulation de problème de segmentation

L'objectif de la segmentation est le suivant : on dispose de données non étiquetées. On souhaite les regrouper par données ressemblantes.
soit un ensemble X de N données décrites chacune par leurs P attributs.

La segmentation consiste à créer une partition ou une décomposition de cet ensemble en groupes telle que :

critère1. les données appartenant au même groupe se ressemblent
critère 2. les données appartenant à deux groupes différents soient peu ressemblantes.

Le problème consiste alors à identifier les nuages denses de points qui sont naturellement présents dans les données.

II.3. Qualité d'un Clustering

- Une bonne méthode de clustering produira des clusters d'excellente qualité avec:
  - ✓ Similarité **intra-classe** importante (forte cohésion interne)

  - ✓ Similarité **inter-classe** faible (faible couplage externe).
- La qualité d'un clustering dépend de :
  - ✓ La mesure de similarité utilisée
  - ✓ L'implémentation de La mesure de similarité.

II.4. Les application de Clustering

- **Marketing:** segmentation du marché en découvrant des groupes de clients à partir de base de données d'achats.
- **Environnement:** identification des zones terrestres similaires (en terme d'utilisation) dans une base de données d'observation de la terre.

- **Assurance:** identification de groupes d'assurés distincts associés à un nombre important de déclarations.
- **Médecine:** la localisation des tumeurs dans le cerveau
  - ✓ Nuage des points du cerveau fournis par le neurologue
  - ✓ Identification des points définissant une tumeur
- **WWW:** la classification de documents

II.5 .Mesure de la similarité

- La similarité entre deux objet i, j est exprimée en terme d'une fonction de distance entre ces deux objets:

$$sim(i,j) = d(i,j)$$

Ministère de l'Enseignement supérieur et de la recherche scientifique
Université Abderrahmane Mira de Bejaïa
Faculté des Sciences Exactes
Département de Mathématiques

Module: Data Mining **Année: 2022-2023**
Nature de document: Cours
Niveau : L3STID **N.BERMAD**
Chapitre: 4
Titre: Apprentissage non supervisé

- La définition de la similarité entre objets dépend de :
  - ✓ Le type de données considérées
  - ✓ Le type de similarité recherchée

II.5.1.Choix de la distance

- Propriétés d'une distance:
  1. $\boldsymbol{d(x,y) \geq 0}$
  2. $\boldsymbol{d(x,y) = 0\ si\ x = y}$
  3. $\boldsymbol{d(x,y) = d(y,x)}$
  4. $\boldsymbol{d(x,z) \leq d(x,y) + d(y,z)}$
- Définir une distance sur chacun de type de données:

II.5.1.1.Distance –Données numériques

Soient $X = (x1, x2, \dots \dots xn)\ et\ Y = (y1, y2, \dots, yn)$ deux objets

- **Distance euclidienne :**

$$\boldsymbol{d(x,y) = \sqrt{\sum_{i=1}^{n} (xi - yi)^2}}$$

- **Distance de Manhattan:**

$$\boldsymbol{d(x,y) = \sum_{i=1}^{n} |xi - yi|}$$

- **Distance Minkoswski:**

$$\sqrt[q]{\sum_{i=1}^{n} |xi - yi|^q}$$

II.5.1.2.Distance –Données binaires

		Object j		
		1	0	sum
Object i	1	a	b	A+b
	0	c	d	C+d
	sum	a+c	b+d	p

Table de contingence (dissimilarité)

- Coefficient de correspondance simple (similarité invariante, si la variable binaire est symétrique):

$$d(i,j) = \frac{b + c}{a + b + c + d}$$

- Coefficient de Jaccard(similarité non invariante, si la variable binaire est asymétrique):

$$d(i,j) = \frac{b + c}{a + b + c}$$

Exemple: dissimilarité entre variables binaires

- Table de patients

Nom	Sexe	Fièvre	Toux	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	p	N	N	N	N

- Huit attributs , avec
  - ✓ Sexe un attribut symétrique, et
  - ✓ Les attributs restants sont asymétriques(test HIV,...)
- Les valeurs Y et P sont initialisées à 1, et la valeur N à 0
- Calculer la distance entre patients, basée sur le coefficient de Jaccard.

$$d(jack, mary) = \frac{0+1}{2+0+1} = 0.33$$
$$d(jack, jim) = \frac{1+1}{1+1+1} = 0.67$$
$$d(jim, mary) = \frac{1+2}{1+1+2} = 0.75$$

II.5.1.3.Distance –données énumératives

- Généralisation des variables binaires, avec plus de deux états , e.g., rouge, jaune , bleu, vert
- Méthode **: correspondance simple**
  - ✓ **M**: le nombre de correspondance, **P:**le nombre totale de variables.

$$\mathbf{d(i,j) = \frac{P - M}{P}}$$

II.5.1.4.Distance –données mixes

Exemple:(âge, propriétaire résidence principale, montant des mensualités en cours .

- x=(30,1,1000), y=(40,0,2200), z=(45,1,4000)
- $d(x,y) = \sqrt{\left(\left(\frac{10}{15}\right)^2 + 1^2 + \left(\frac{1200}{3000}\right)^2\right)} = 1.27$
- $d(x,z) = \sqrt{\left(\left(\frac{15}{15}\right)^2 + 0^2 + \left(\frac{3000}{3000}\right)^2\right)} = 1.41$
- $d(y,z) = \sqrt{\left(\left(\frac{05}{15}\right)^2 + 0^2 + \left(\frac{1800}{3000}\right)^2\right)} = 1.21$
- Plus proche voisin de x=y
- Sommation: d(x,y)=d1(x1,y1)+...+dn(xn,yn)

II.6.Méthodes & algorithmes de clustering

Un algorithme de regroupement vise à produire des d'éléments (classes) à partir d'un ensemble d'éléments en regroupant

Ministère de l'Enseignement supérieur et de la recherche scientifique
Université Abderrahmane Mira de Bejaïa
Faculté des Sciences Exactes
Département de Mathématiques

Module: Data Mining **Année: 2022-2023**
Nature de document: Cours
Niveau : L3STID **N.BERMAD**
Chapitre: 4
Titre: Apprentissage non supervisé

les éléments qui possèdent des caractéristiques communes. Il existe différent techniques de clustering en fonction de la nature de la structure résultante:

- Algorithmes de partitionnements (K-moyennes)
- Algorithmes hiérarchiques
- algorithmes par voisinage dense

II.6.1. Algorithmes de partitionnement

- Partitionner une base de données D de N objets à un ensemble de k clusters

- **Problème:**
  - ✓ Sachant *K*, chercher un partitionnement de *K clusters* qui optimise le critère de segmentation choisis

II.6.1. 1.Algorithme de K moyennes(K-Means)

- **Entrée**: un échantillon de m enregistrements $X_1,...X_m$, le nombre de groupes à constituer , $K \in N$

  1.Choisir K centre initiaux c_1, ...c_k

  2.Répartir chacun des m enregistrements dans le groupe i dans le centre ci est le plus proche

3. Si aucun élément ne change de groupe alors arrêt et sortir les groupes

4. Calculer les nouveaux centre mobiles : pour tous i , ci est la moyenne des éléments du groupe i

5. Aller en 2

Exemple

- Huit points $A,.....,H$ de l'espace euclidien $2D$, $k=2$ (deux groupes)
- Tirer aléatoirement 2 centre: B et D choisis.

Points	Centre D(2,4),B(2,2)	Centre D(2,4),I(27/7,17/7)	Centre J(5/3,10/3),K(24/5,11/5)
A(1,3)	B	D	J
B(2,2)	B	I	J
C(2,3)	B	D	J
D(2,4)	D	D	J
E(4,2)	B	I	K
F(5,2)	B	I	K
G(6,2)	B	I	K
H(7,3)	B	I	K

- **<u>Avantages & inconvénients de K-moyennes</u>**
  - ⊕ Relativement extensible
  - ⊕ Relativement efficace
  - ⊖ Applicable seulement dans le cas où la moyenne des objets est définie
  - ⊖ Besoin de spécifier K
- **Les variantes de K-moyennes**
  - ✓ Sélection des centres initiaux
  - ✓ Calcul des centres (k-medoids)
  - ✓ GMM: variantes de K-moyennes basées sur les probabilités
  - ✓ K-modes: données catégorielles
  - ✓ K-prototype: données mixtes(numériques et catégoriques)