Université de Batna 2 (Mostefa Ben Boulaid)
Faculté de Technologie
Département de Génie Industriel

Apprentissage Automatique
Cours 3 (Clustering)

Pr Hassen BOUZGOU

Différents problèmes de ML

	Apprentissage Supervisé	Apprentissage Non-supervisé
Discrète	Classification / Catégorisation	Clustering
Continue	Régression	Réduction de dimensionnalité

Regroupement

- Regroupement (Clustering): construire une collection d'objets
  - **Similaires** au sein d'un **même** groupe,
  - **Différents** quand ils appartiennent à des groupes **différents**
- Le Clustering est une classification non supervisée: pas de classes prédéfinies

Utilisation:

- Afin de mieux comprendre les données
- Comme prétraitement avant d'autres analyses

Regroupement (1)

- Une bonne méthode de regroupement permet de garantir:
  - Une grande ressemblance ***intra-groupe***
  - Une faible ressemblance ***inter-groupe***
- La capacité d'une méthode de regroupement dépend essentiellement de **la mesure de similarité** utilisée par la méthode
- La qualité d'une méthode peut aussi être mesurée par sa capacité à trouver quelques ou tous les motifs intéressants

Données

Matrice de données

$$\begin{bmatrix} x_{11} & \dots & x_{1f} & \dots & x_{1p} \\ \dots & \dots & \dots & \dots & \dots \\ x_{i1} & \dots & x_{if} & \dots & x_{ip} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & \dots & x_{nf} & \dots & x_{np} \end{bmatrix}$$

Matrice de similarité

d(2,1) distance entre

Instance 2 et instance 1

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \dots & \dots & 0 \end{bmatrix}$$

Mesure de similarité

- Métrique de similarité/dissimilarité: exprimée en termes d'une fonction de distance, typiquement *d(i,j)*
- Fonction de distance dépend du type des données : binaires, nominales, ordinales ou continues
  - ✓ **Continue** sur un intervalle ex: poids, taille
  - ✓ **Binaire**: oui, non ex: malade, sein
  - ✓ **Nominale**: par nom: ex: couleur (bleu, vert,)
  - ✓ **Ordinale**: naturellement ordonnées. Ça peut être le classement à une course, par exemple ou le résultat à questionnaire (1 : pas du tout d'accord, 2 … 5 : Tout à fait d'accord)

Fonction Distance

Valeurs continues sur un intervalle

- Distance de Minkowski :

$$d(i,j)=\sqrt[q]{(|x_{i_1}-x_{j_1}|^q+|x_{i_2}-x_{j_2}|^q+\ldots+|x_{i_p}-x_{j_p}|^q)}$$

avec $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ et $x_j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ deux objets à p dimensions, et q un entier positif

si $q = 1$: distance de Manhattan

$$d(i,j)=|x_{i_1}-x_{j_1}|+|x_{i_2}-x_{j_2}|+\ldots+|x_{i_p}-x_{j_p}|$$

si $q = 2$:distance euclidienne

$$d(i,j)=\sqrt{(|x_{i_1}-x_{j_1}|^2+|x_{i_2}-x_{j_2}|^2+\ldots+|x_{i_p}-x_{j_p}|^2)}$$

Mesure de similarité

Variables Binaires

table de contingence

Objet *i* \ Objet *j*	1	0	*sum*
1	a	b	$a+b$
0	c	d	$c+d$
sum	$a+c$	$b+d$	p

a= nombre de positions où i a 1 et j a 1

- **coefficient simple d'appariement**

Exemple o_i=(1,1,0,1,0) et

o_j=(1,0,0,0,1)

a=1, b=2, c=1, d=1 ⇒ $d(o_i, o_j)$=3/5

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

- **Coefficient de Jaccard**

$d(o_i, o_j)$=3/4

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

Mesure de similarité

Exemple

Nom	Sexe	Fièvre	Toux	Test-1	Test-2	Test-3	Test-4
Amar	M	P	N	P	N	N	N
Mohamed	M	P	N	P	N	P	N
Karima	F	P	P	N	N	N	N

- sexe est symétrique
- les autres sont asymétriques
- soit P = 1, et N = 0

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

Objet *i* \ Objet *j*	1	0	*sum*
1	*a*	*b*	*a+b*
0	*c*	*d*	*c+d*
sum	*a+c*	*b+d*	*p*

$$d(Amar, Mohamed) = \frac{0+1}{2+0+1} = 0.33$$

$$d(Amar, Karima) = \frac{1+1}{1+1+1} = 0.67$$

$$d(Karima, Mohamed) = \frac{1+2}{1+1+2} = 0.75$$

Les plus similaires sont les instances avec la plus petite distance, Amar et Mohamed $\Rightarrow$ atteints du même mal

Techniques de Clustering

- Construire une partition à ***k*** clusters d'une base ***D*** de ***n*** objets
- Les ***k*** *clusters doivent* optimiser le critère choisi
  - Global optimal: Considérer toutes les k-partitions
  - Méthodes Heuristiques
    - ***k-means*** (MacQueen'67): Chaque cluster est représenté par son *centre*
    - ***k-medoids*** or PAM (Partition around medoids) (Kaufman & Rousseeuw'87): Chaque cluster est représenté par *un de ses objets*

Une **heuristique** est une méthode de calcul qui fournit rapidement une solution réalisable, pas nécessairement optimale ou exacte, pour un problème d'optimisation difficile

K-Means (1)

- Le partitionnement en ***k*-moyennes** (ou **k-*means*** en Anglais) est une méthode de partitionnement de données et un problème d'optimisation combinatoire.
- Étant donné des points et un entier ***k***, le problème est de diviser les points en ***k*** groupes, souvent appelés ***clusters***, de façon à minimiser une certaine fonction.
- On considère la distance d'un point à la moyenne des points de son cluster ; la fonction à minimiser est la somme des carrés de ces distances.

un problème d'**optimisation combinatoire** (on dit aussi d'**optimisation discrète**) consiste à trouver dans un ensemble discret un parmi les *meilleurs* sous-ensembles (ou solutions) réalisables, la notion de *meilleure solution* étant définie par une fonction objectif

K-Means (2)

4 étapes

1. Partitionne les objets en *k* ensembles non vides

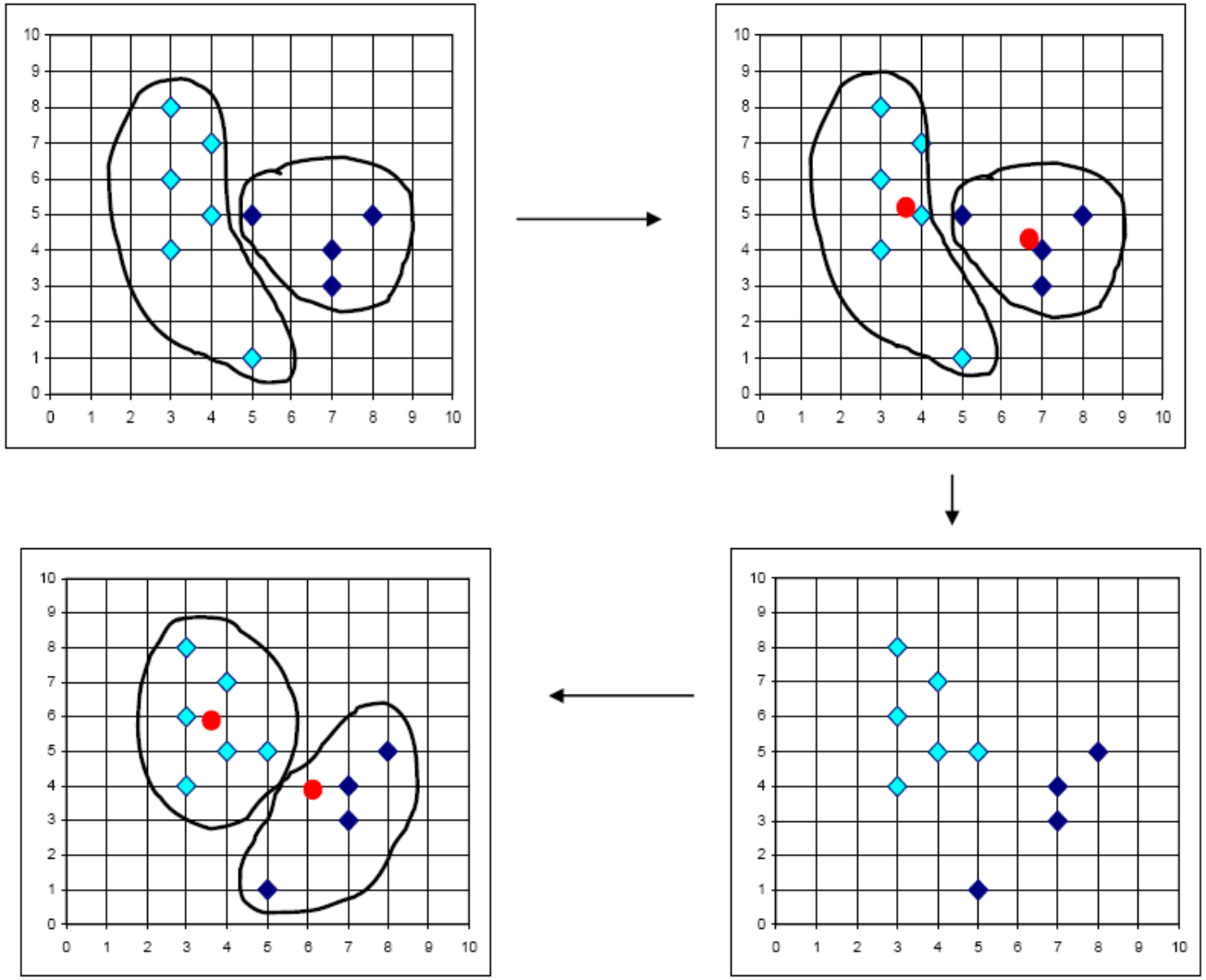
2. Calcule le centroïde de chaque partition/cluster

3. Assigner à chaque objet le cluster dont le centroïde est le plus proche

4. boucle en 2, jusqu'à ce les clusters soient stables.

K-Means (3)

Exemple:



K-Means (2)

Avantages

- Relativement efficace : *O(*tkn*)*, avec *n* le nombre d'objets, *t* le nombre d'itérations et en général t et k << n (complexité linéaire)
- Termine souvent sur un optimum local. L'optimum global peut être atteint en utilisant des techniques telles que les algorithmes génétiques

• Faiblesses

- Utilisable seulement lorsque la moyenne est définie. Que faire dans le cas de données nominales ?
- Besoin de spécifier k à l'avance
- Ne gère pas le bruit et les valeurs aberrantes (outliers)

La méthode K-Medoids

- En Anglais **(Partitioning Around Medoids « PAM »)**
- Trouver des objets représentatifs (medoïdes) dans les clusters au lieu de la moyenne (les centres)
- **Principe**
  - Commencer avec un ensemble de medoïdes puis itérativement remplacer un par un autre si ça permet de réduire la distance globale
  - Efficace pour des données de petite taille

K-Medoids (1)

Choisir arbitrairement ***k*** objets représentatifs

- Pour toute paire (h,j) d'objets tels que h est choisi et j non, calculer le coût $\boldsymbol{TC_{jh}}$ du remplacement de j par h
  - Si $TC_{jh} < 0$, ***j*** est remplacé par ***h***
  - Puis affecter chaque objet non sélectionné au medoïde qui lui est le plus proche
- Répéter jusqu'à ne plus avoir de changements

K-Medoids (2)

- TC_{jh} représente le gain en distance globale que l'on va avoir en remplaçant *j* par *h*

- Si TC_{jh} est négatif alors on va perdre en distance. Ca veut dire que les clusters seront plus compacts.

K-Medoids (Exemple)

- Soit A={1,3,4,5,8,9}, k=2 et M={1,8} ensemble des medoides
  ➔C1={1,3,4} et C2={5,8,9}
  $E_{\{1,8\}}$=dist(3,1)^2+dist(4,1)^2+dist(5,8)^2+dist(9,8)^2= 4+ 9+ 9+1 =23

- Comparons 1 et 3➔M={3,8}➔C1={1,3,4,5} et C2={8,9}
  $E_{\{3,8\}}$ =dist(1,3)^2+dist(4,3)^2+dist(5,3)^2+dist(9,8)^2= 4+ 1+ 4+1 =10
  $E_{\{3,8\}} - E_{\{1,8\}}$= -12 <0 donc le remplacement est fait.

- Comparons 3 et 4➔ M={4,8}➔ C1 et C2 inchangés ➔C1={1,3,4,5} et C2={8,9}
- $E_{\{4,8\}}$=dist(1,4)^2+dist(3,4)^2+dist(5,4)^2+dist(8,9)^2= 12➔ 3 n'est pas remplacé par 4
- Comparons 3 et 5➔M={5,8}➔ C1 et C2 inchangés et E{5,8}>E{3,8}