Data Mining

Analyse et Fouille de données

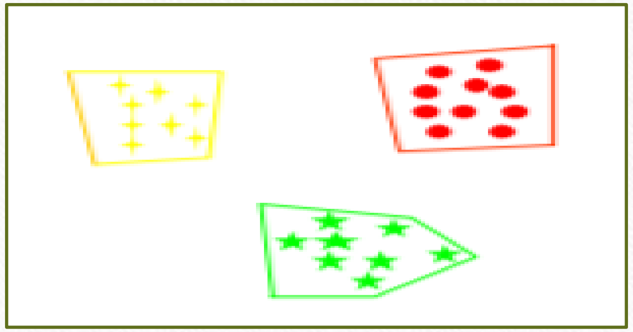
Dr. Sana Hamdi

sana.hamdi@fst.utm.tn

Introduction

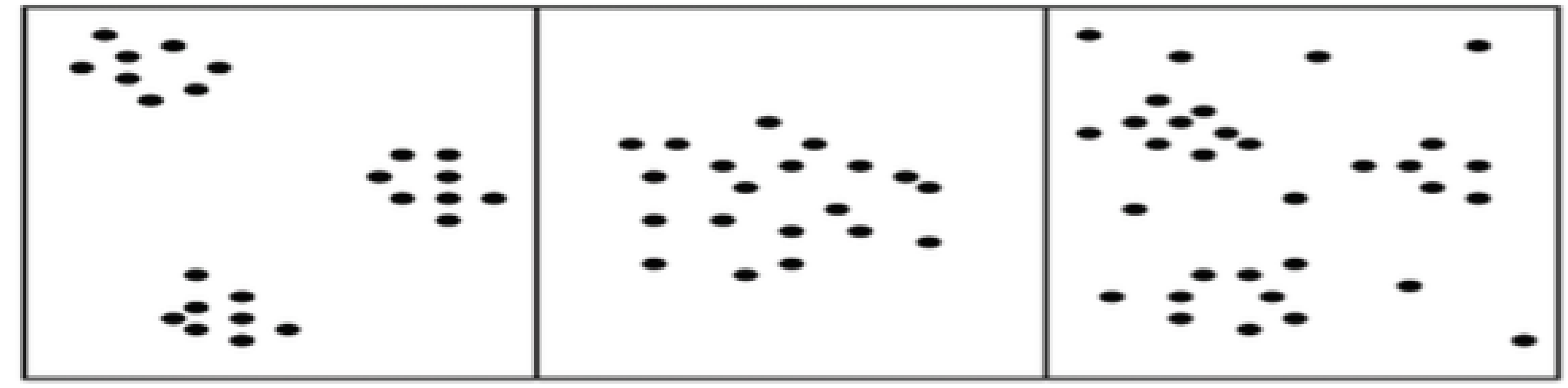
Le Clustering

- C'est une approche qui partitionne un ensemble de données en sous-classes (clusters) ayant du sens.
- Il s'agit de regrouper les points proches/similaires en « paquets ou groupes ou classes »
- Optimiser le regroupement: Une bonne méthode va produire des clusters tout en:
  - ✓ Maximisant la similarité intra-classe
  - ✓ Minimisant la similarité inter-classes
- La qualité d'un clustering dépend de la mesure de similarité



Le Clustering

- MAIS les groupes peuvent être assez bien définis et séparés, ou au contraire imbriqués/sans frontières claires, et de formes quelconques



Applications

Domaine	Forme des données	Clusters
Text mining	Textes, Mails	Textes proches, Dossiers automatiques
Web mining	Textes et images	Pages Web proches
Bio-informatique	Gènes	Gènes ressemblants
Marketing	Infos Clients, produits achetés	Segmentation de la clientèle
Segmentation d'images	Images	Zones homogènes dans l'image
Web log Analysis	ClickStream	Profils utilisateurs, groupes d'accès similaires

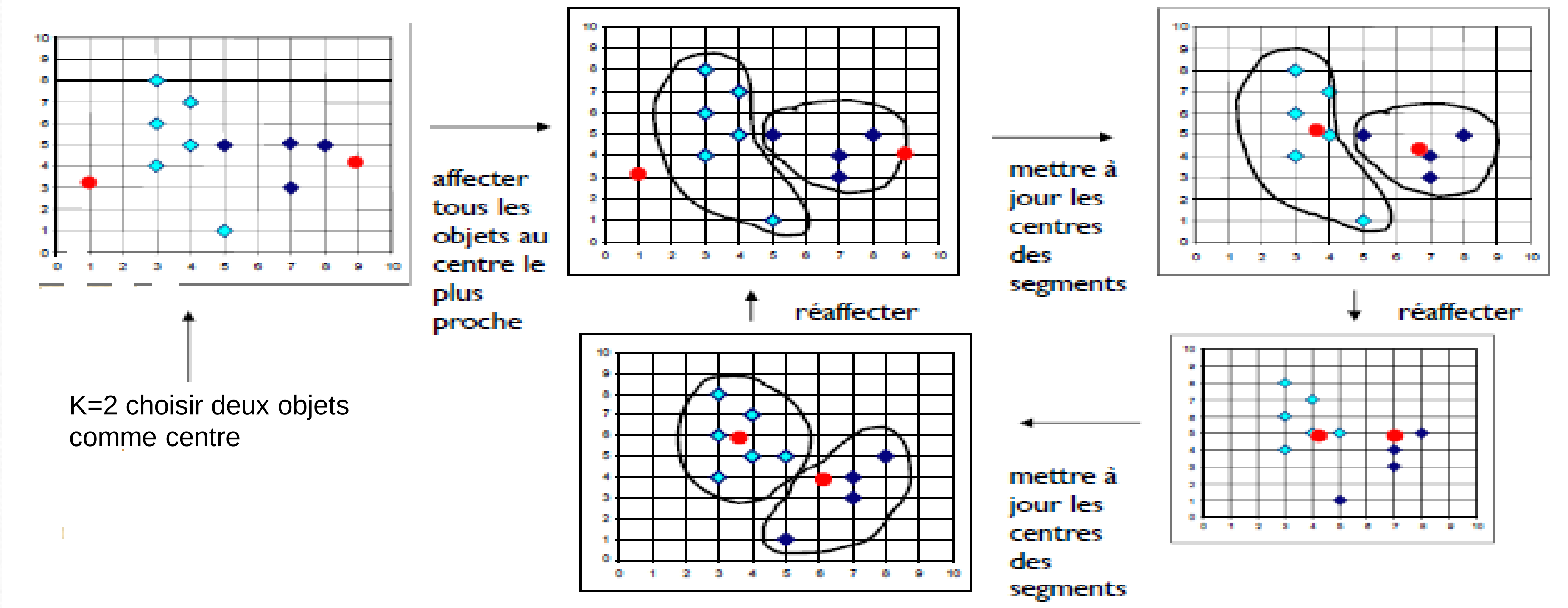
Partionnement

- Construire une partition à ***k*** clusters d'un jeu de données ***D*** de ***n*** objets
- Les ***k*** clusters doivent optimiser le critère de partitionnement choisi (fonction de similarité)
- Critère de performance : l'erreur est définie comme la distance euclidienne entre individus et centres des groupes
- Approches heuristiques :
  - K-means : chaque cluster est représenté par son centre de gravité
  - K-medoïds ou *PAM (partition around medoids)* : chaque cluster est représenté par un objet du cluster

Algorithme des k-moyennes

1. Choisir k objets les représentants initiaux de k clusters
2. (Ré)affecter chaque objet O au cluster C_i de centre M_i tel que $d(O, M_i)$ est minimal
3. Recalculer M_i de chaque cluster (le barycentre)
4. Aller à l'étape 2 et 3 jusqu'à ce qu'il n'y ait plus (ou peu) de changement dans les clusters

k-moyennes: Trace d'exécution



k-moyennes: Exemple

- $A=\{1,2,3,6,7,8,13,15,17\}$. Créer 3 clusters à partir de A
- On prend 3 objets au hasard. Supposons que c'est 1, 2 et 3. Ca donne:
  - $C_1=\{1\}$, $M_1=1$,
  - $C_2=\{2\}$, $M_2=2$,
  - $C_3=\{3\}$ et $M_3=3$
- Chaque objet O est affecté au cluster dont le centre est le plus proche.
- 6 est affecté à C_3 car $d(M_3,6)<d(M_2,6)$ et $d(M_3,6)<d(M_1,6)$
- On a: $C_1=\{1\}$, $M_1=1$,

  $C_2=\{2\}$, $M_2=2$

  $C_3=\{3, 6,7,8,13,15,17\}$, $M3=69/7=9.86$

k-moyennes: Exemple

- d(3,M2)<d(3,M3) ➔ 3 passe dans C_2. Tous les autres objets ne bougent pas.

C_1={1}, M_1=1, C_2={2,3}, M_2=2.5, C_3={6,7,8,13,15,17} et M_3= 66/6=11

- d(6,M_2)<d(6,M_3) ➔ 6 passe dans C2. Tous les autres objets ne bougent pas.

C_1={1}, M_1=1, C_2={2,3,6}, M_2=11/3=3.67, C3={7,8,13,15,17}, M_3= 12

- d(2,M1)<d(2,M2) ➔ 2 passe en C1. d(7,M2)<d(7,M3) ➔ 7 passe en C2. Les autres ne bougent pas.

C1={1,2}, M1=1.5, C2={3,6,7}, M2=5.34, C3= {8,13,15,17}, M3=13.25

- d(3,M1)<d(3,M2) ➔ 3 passe en 1. d(8,M2)<d(8,M3) ➔ 8 passe en 2

C1={1,2,3}, M1=2, C2={6,7,8}, M2=7, C3={13,15,17}, M3=15

Plus rien ne bouge

Indice de validité: Silhouette

- L'erreur n'est pas un bon indice de qualité/validité : n'est basée que sur la dispersion intra-groupes
- Le score de silhouette se définit d'abord sur un point i dont le groupe est k
- Il se base sur:

- la distance moyenne du point à son groupe: $a(i) = \frac{1}{|I_k|-1} \sum_{j \in I_k, j \neq i} d(x^i, x^j)$
- la distance moyenne du point à son groupe voisin: $b(i) = \min_{k' \neq k} \frac{1}{|I_{k'}|} \sum_{i' \in I_{k'}} d(x^i, x^{i'})$.
- Avec I_k est l'ensemble des points appartenant à un groupe k

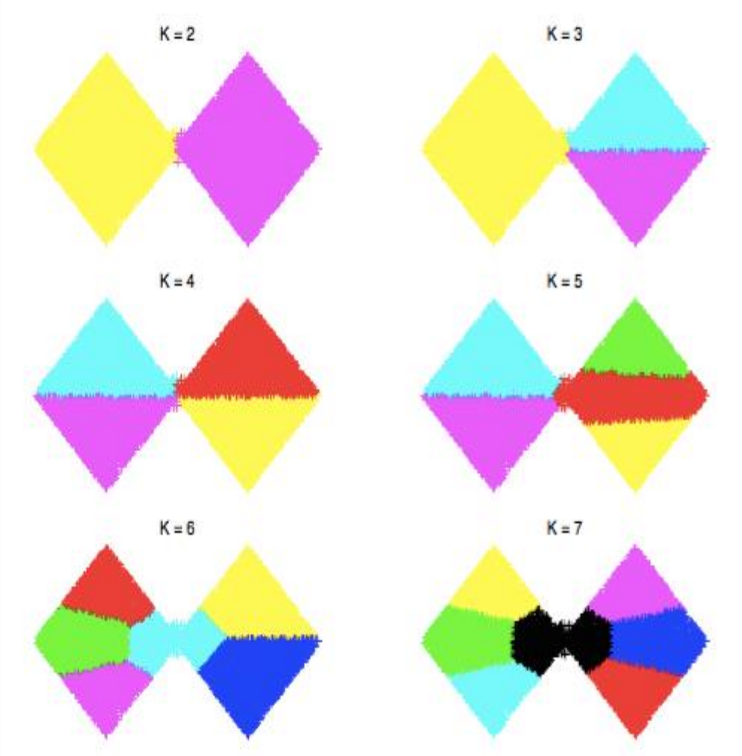
Indice de validité: Silhouette

- Le coefficient de silhouette du point i s’écrit alors: $$s_{sil}(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$
- On peut le moyenner groupe par groupe pour comparer leurs homogénéités : ceux avec les coefficient de silhouette les plus forts sont les plus homogènes. Sur l'ensemble de la classification, il aura pour expression
$$S_{sil} = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{|I_k|} \sum_{i \in I_k} s_{sil}(i)$$
- Le coefficient de silhouette varie entre -1 (pire classification) et 1 (meilleure classification)
- Choisir le K qui maximise le coefficient de Silhouette

Variantes de k-moyennes

Les variantes des K-Means diffèrent dans :

- La sélection des k initiaux
- Calcul de la dissimilarité (distance)
- Stratégies pour calculer la moyenne d’un cluster



k-moyennes: Intérêts

- L’algorithme converge en général pour les mesures typiques de similarité.
- Souvent l’algorithme converge en quelques itérations.
- Relativement efficace: O(t*k*n), où n est nombre objets, k est nombre de clusters, et t est le nombre d’itérations. Normalement, k, t << n.

k-moyennes: limites

- Utilisable seulement lorsque la moyenne est définie. Que faire dans le cas de données nominales ?
- On doit spécifier k en avance (nombre de clusters)
- Les clusters sont construits par rapports à des objets inexistants (les milieux)
- Ne gère pas le bruit et les exceptions

Algorithme des k-médoïdes (PAM)

- Trouve des représentants, appelés médoïdes, dans les clusters
- PAM *(partition around medoids)* :
  - médoïde : l'objet d'un cluster pour lequel la distance moyenne à tous les autres objets du cluster est minimale
  - critère d'erreur : $E = \sum_{i=1}^{k} \sum_{p \in c_i} d(p, mi)^2$

Algorithme des k-médoïdes (PAM)

1. Sélectionner *k* objets arbitrairement
2. Assigner le reste des objets au médoïde le plus proche
3. Sélectionner un objet non médoïde et échanger si le critère d'erreur peut être réduit
4. Répéter 2 et 3 jusqu'à ne plus pouvoir réduire le critère d'erreur

k-médoïdes : exemple

- Soit A={1,3,4,5,8,9}, k=2 et M={1,8} ensemble des medoides ➔ C1={<u>1</u>,3,4} et C2={5,<u>8</u>,9}

  $E_{\{1,8\}}=d(3,1)^2+d(4,1)^2+d(5,8)^2+d(9,8)^2=$**23**
- Comparons 1 et 3 ➔M={3,8} ➔ C1={1,<u>3</u>,4,5} et C2={<u>8</u>,9}

  $E_{\{3,8\}}=d(1,3)^2+d(4,3)^2+d(5,3)^2+d(9,8)^2=$**10**

  $E_{\{3,8\}} - E_{\{1,8\}} = -29 < 0$ donc le remplacement est fait.
- Comparons 3 et 4 ➔ M={4,8}➔ C1 et C2 inchangés et $E_{\{4,8\}}=d(1,4)^2+d(3,4)^2+d(5,4)^2+d(8,9)^2=$ 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}

k-médoïdes : exemple