Clustering

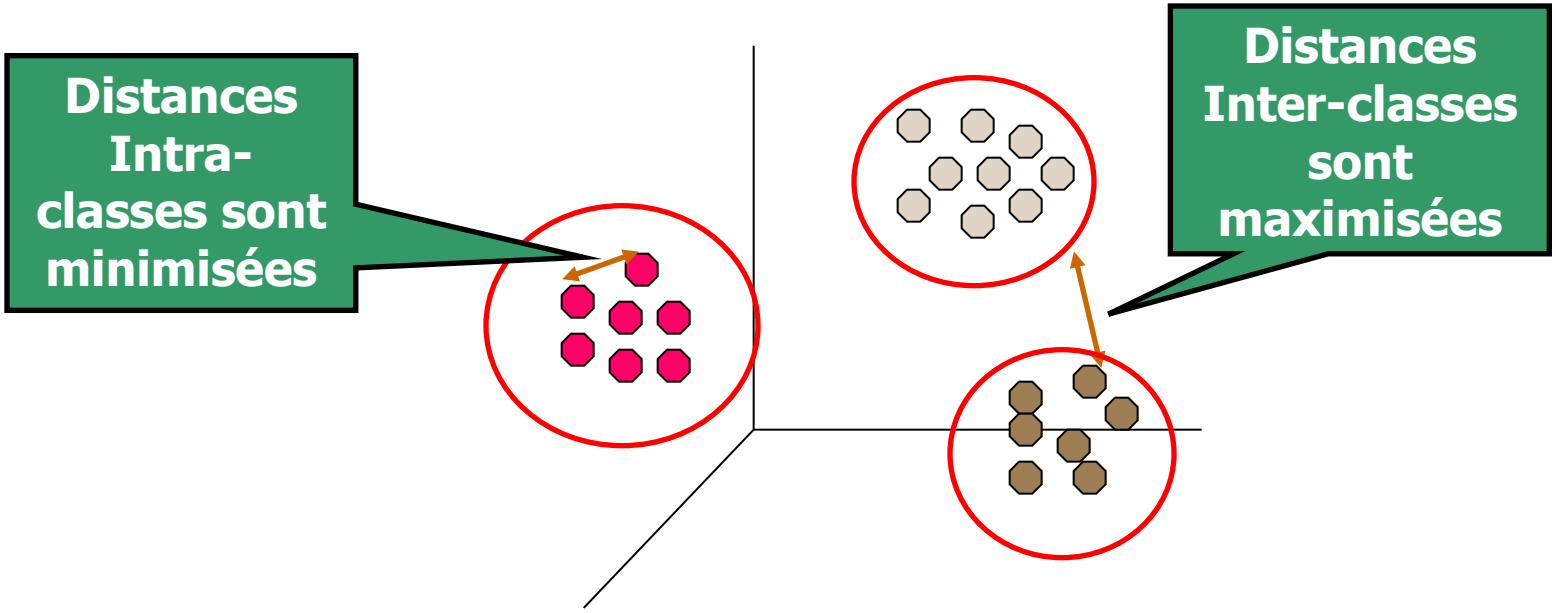
- **La Classification est un apprentissage supervisé. La supervision est faite en nommant les classes des instances d'apprentissage.**

- **Le Clustering est un apprentissage non supervisé. Il n'y a pas une connaissance apriori des classes, ni un ensemble d'apprentissage.**

- **L'algorithme de clustering nécessite une affectation de chaque instance à un groupe ou classe (cluster) de telle façon que tous les objets d'un même groupe sont plus semblables que les autres.**

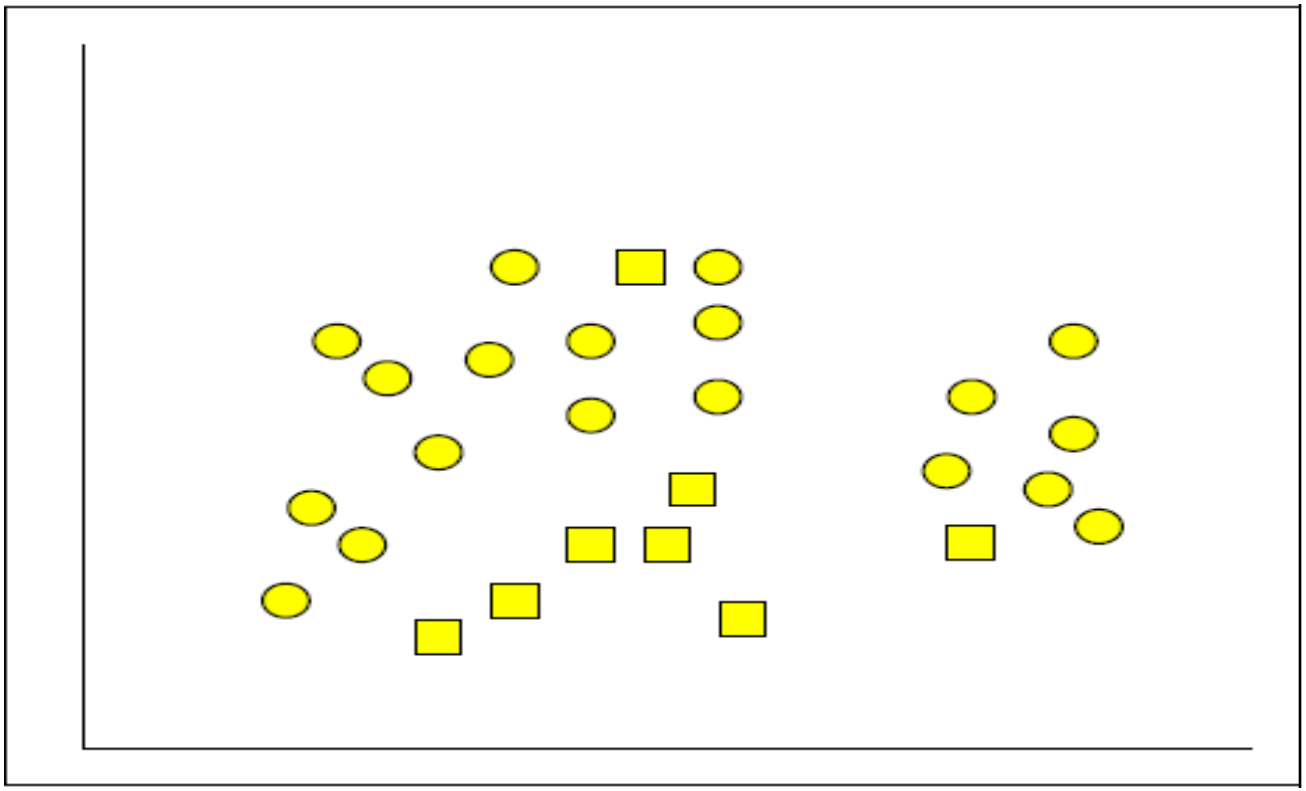
Clustering

- **Trouver des groupes (classes) d'objets tels que chaque objet d'un groupe est similaire qu'un autre objet du même groupe et différent des autres objets des autres groupes**
- L'objectif est de trouver un groupement le plus naturel possible des instances.
  - A l'intérieur d'un groupe: Maximiser la similarité entre instances.
  - Entre les groupes: Minimiser la similarité entre les instances.



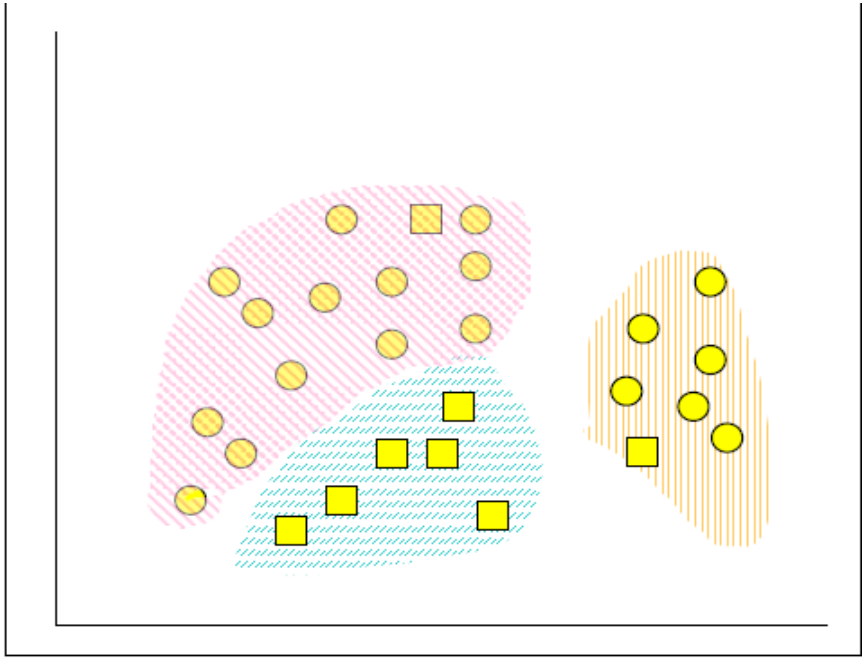
Clustering

- **Par exemple, soit l'ensemble de figures suivant:**



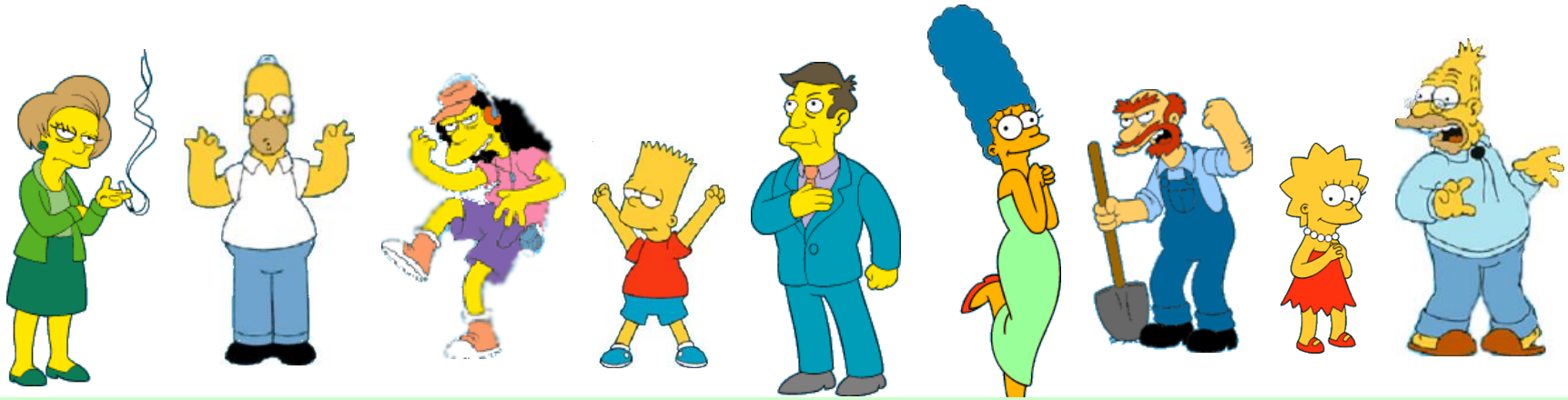
Clustering

- **Un algorithme de clustering peut trouver les clusters suivants:**

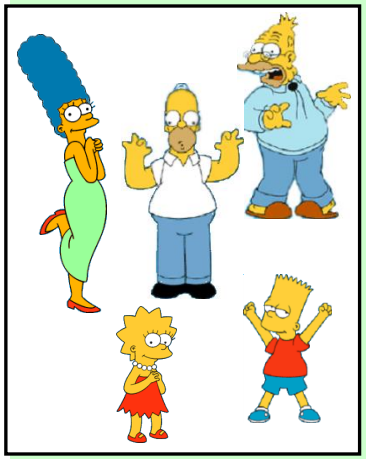


- **Bien que certaines figures différentes coexistent dans un cluster.**

Quel est le groupement naturel de ces objets?

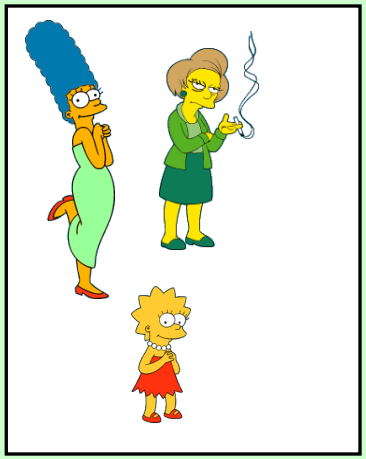


Les algorithmes de Clustering sont subjectifs



Famille deTatkare

Employés de l'école



Feminins

Masculins

Le problème du Clustering

- **Etant donnée une base de données D={t_1,t_2,…,t_n} de tuples et une valeur entière k, le *Clustering est de définir une application* f:D→{1,..,k} où chaque t_i est affecté à un seul cluster (groupe ou classe) K_j, 1<=j<=k.**

- **Un *Cluster*, K_j, contient exactement les tuples qui lui sont affectés.**

- **Contrairement au problème de classification, les clusters ne sont pas connus apriori.**

Applications

- **Marketing:** Découvrir les groupes de clients basés sur leurs habitudes d'achats

- **Plans de cités:** Identifier les groupes de batiments par type, vameur, localisation

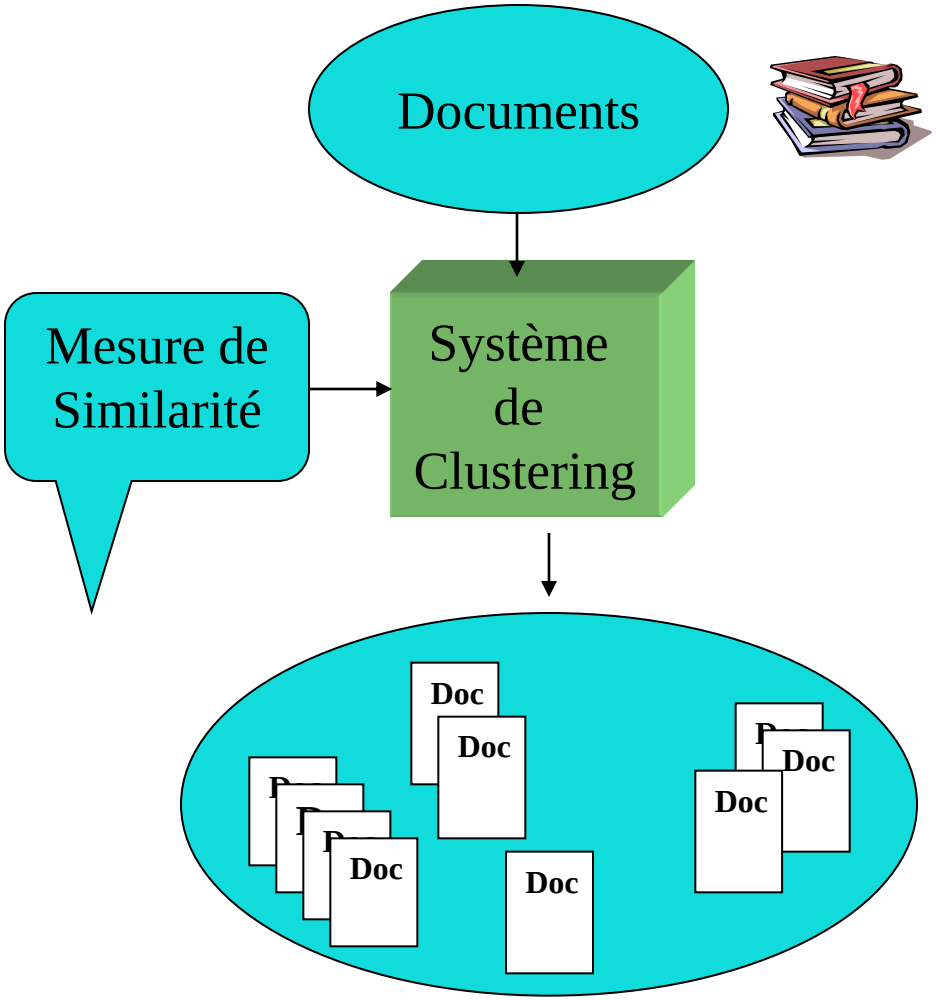
Applications

- **Traitement d'image:** Identifier les clusters d'images similaires (eg chevaux)

- **Biologie:** Découvrir les groupes de plantes/animaux ayant des propriétés similaires

Applications

- Etant donnés
  - Un ensemble de documents de texte
  - Une mesure de Similarité
    - e.g., Combien de mots communs dans tous les documents
- Trouver:
  - Les clusters (groupes) de documents pertinents



Qu’est ce qu’un bon regroupement?

- Une bonne méthode de regroupement permet de garantir
  - Une grande similarité intra-groupe
  - Une faible similarité inter-groupe
- La qualité d’un regroupement dépend donc de la mesure de similarité utilisée par la méthode et de son implémentation

Structures de données

- Matrice de données

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

- Matrice de similarité

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

Mesurer la qualité d'un clustering

- Métrique pour la similarité: La similarité est exprimée par le biais d'une **mesure de distance**
- Une autre fonction est utilisée pour la mesure de la qualité
- Les définitions de distance sont très différentes que les variables soient des intervalles **(continues), catégories, booléennes ou ordinales**
- En pratique, on utilise souvent une pondération des variables

Types des variables

- Intervalles:
- Binaires:
- catégories, ordinales, ratio:
- Différents types:

Intervalle (discrètes)

- Standardiser les données
  - Calculer l'écart absolu moyen:

$$s_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + ... + |x_{nf} - m_f|)$$

où $$m_f = \frac{1}{n}(x_{1f} + x_{2f} + ... + x_{nf}).$$

  - Calculer la mesure standardisée (*z-score*)

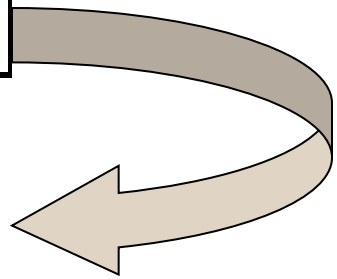
$$z_{if} = \frac{x_{if} - m_f}{s_f}$$

Exemple

	Age	Salaire
Personne1	50	11000
Personne2	70	11100
Personne3	60	11122
Personne4	60	11074

$M_{Age} = 60 \quad S_{Age} = 5$

$M_{salaire} = 11074 \quad S_{salaire} = 148$



	Age	Salaire
Personne1	-2	-0,5
Personne2	2	0,175
Personne3	0	0,324
Personne4	0	2

Similarité entre objets

- Les distances expriment une similarité
- Ex: la *distance de Minkowski* :

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

où $i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ et $j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ sont deux objets p-dimensionnels et q un entier positif

- Si $q = 1$, d est la distance de Manhattan

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

Similarité entre objets(I)

- *Si q = 2, d est la distance Euclidienne* :

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

  – Propriétés
  - $d(i,j) \geq 0$
  - $d(i,i) = 0$
  - $d(i,j) = d(j,i)$
  - $d(i,j) \leq d(i,k) + d(k,j)$

Exemple: distance de Manhattan

	Age	Salaire
Personne1	50	11000
Personne2	70	11100
Personne3	60	11122
Personne4	60	11074

d(p1,p2)=120

d(p1,p3)=132

Conclusion: p1 ressemble plus à p2 qu'à p3 ☹

	Age	Salaire
Personne1	-2	-0,5
Personne2	2	0,175
Personne3	0	0,324
Personne4	0	0

d(p1,p2)=4,675

d(p1,p3)=2,824

Conclusion: p1 ressemble plus à p3 qu'à p2 ☺

Variables binaires

- Une table de contingence pour données binaires

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

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

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

a=1, b=2, c=1, d=1

Mesures de distances

- Coefficient d'appariement (matching) simple (invariant pour variables symétriques):

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

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

$d(o_i, o_j)$=3/5

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

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

Variables binaires (I)

- Variable symétrique: Ex. le sexe d'une personne, i.e coder masculin par 1 et féminin par 0 c'est pareil que le codage inverse

- Variable asymétrique: Ex. Test HIV. Le test peut être positif ou négatif (0 ou 1) mais il y a une valeur qui sera plus présente que l'autre. Généralement, on code par 1 la modalité la moins fréquente
  - 2 personnes ayant la valeur 1 pour le test sont *plus similaires* que 2 personnes ayant 0 pour le test

Variables binaires(II)

- Exemple

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

- Sexe est un attribut symétrique
- Les autres attributs sont asymétriques
- Y et P ≡ 1, N ≡ 0, la distance n'est mesurée que sur les asymétriques

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

$$d(jack, jim) = \frac{1 + 1}{1 + 1 + 1} = 0.67$$

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

Les plus similaires sont Jack et Mary⇒atteints du même mal

Variables Nominales

- Une généralisation des variables binaires, ex: rouge, vert et bleu
- Méthode 1: Matching simple
  - *m*: # d'appariements, *p*: # total de variables

$$d(i, j) = \frac{p - m}{p}$$

- Méthode 2: utiliser un grand nombre de variables binaires
  - Créer une variable binaire pour chaque modalité (ex: variable rouge qui prend les valeurs vrai ou faux)

Variables Ordinales

- Une variable ordinale peut être discrète ou continue
- L'ordre peut être important, ex: classement
- Peuvent être traitées comme les variables intervalles
  - remplacer x_{if} par son rang $r_{if} \in \{1, \ldots, M_f\}$
  - Remplacer le rang de chaque variable par une valeur dans [0, 1] en remplaçant la variable *f* dans l'objet *i* par

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

  - Utiliser une distance pour calculer la similarité

En Présence de Variables de différents Types

- Pour chaque type de variables utiliser une mesure adéquate. Problèmes: les clusters obtenus peuvent être différents
- On utilise une formule pondérée pour faire la combinaison

$$d(i,j) = \frac{\sum_{f=1}^{P} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{P} \delta_{ij}^{(f)}}$$

  - *f* est binaire ou nominale:

    $d_{ij}^{(f)} = 0$ si $x_{if} = x_{jf}$, sinon $d_{ij}^{(f)} = 1$
  - *f* est de type intervalle: utiliser une distance normalisée
  - *f* est ordinale
    - calculer les rangs r_{if} et $z_{if} = \frac{r_{if} - 1}{M_f - 1}$
    - Ensuite traiter z_{if} comme une variable de type intervalle