Série de TD N°1 - Corrigé
Clustering 1 (Types des variables)

Exercice 1 : Les variables de données peuvent être de type : numérique, ordinale, nominale (ou catégoriale), asymétrique binaire, intervalle ou vecteurs.

a- Donner un exemple pour chaque type de variables. **(voir cours)**

b- Décrire brièvement (en donnant un exemple) comment mesurer la similarité entre deux éléments de chacune des variables citées plus haut.

1. Numérique : distance Euclidienne (Manhattan, Minkowski)
2. Intervalle : 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*) $$z_{if} = \frac{x_{if} - m_f}{s_f}$$

- Utiliser une distance pour calculer la similarité

Exemple :

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

$M_{age}=60$, $S_{age}=5$

$M_{salaire}=11074$, $S_{salaire}=148$

	Age	Salaire
Personne1	-2	-0.5
Personne2	2	0.175
Personne3	0	0.324
Personne4	0	2

3. Asymétrique : coefficient de Jaccard (Exemple : Voir cours)
4. Nominale : Une généralisation des variables binaires, ex: rouge, vert et bleu

- Utiliser un grand nombre de variables binaires : créer une variable binaire pour chaque modalité. Ensuite utiliser l'indice de Jaccard.
- Exemple : supposons que deux individus i et j ont répondu à deux questions pour lesquelles ils avaient trois choix: *i* répond *b* aux deux questions alors que *j* répond *b* à la première et *c* à la deuxième. Les choix seront codés selon le tableau ci-dessous.

Individu	Q1a	Q1b	Q1c	Q2a	Q2b	Q2c
i	0	1	0	0	1	0
j	0	1	0	0	0	1

5. Ordinales : peuvent être utilisées comme des variables intervalles.
   - Remplacer x_{if} par son rang $r_{if} \in \{1,..,M_f\}$
   - 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

Exemple :

Etudiant	Mention
E1	Très bien
E2	Excellent
E3	Bien

- Remplacement de chaque x_{if} par son rang $r_{if} \in \{1,2,3\}$: *E1 : 2 ; E2 :3, E3 :1*
- Calcul de Z_{if} *:* Z_{1f}*= (2-1)/(3-1)=1/2=0.5,* Z_{2f}*=(3-1)/(3-1)=1,* Z_{3f}*=(1-1)(3-1)=0*
- D'où les nouvelles valeurs *: E1 : 0.5, E2 : 1, E3 : 0*

c- Comment procéder pour mesurer la similarité entre des éléments de types différents ?

On doit tout d'abord recoder toute variable nominale sous forme de variables binaires et toute variable ordinale sous forme de variable numérique, ensuite pour chaque type de variables utiliser une mesure adéquate. Enfin, on utilise une formule pondérée pour faire la combinaison :

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

Où δ est un poids accordé à la variable 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
- *f* est ordinale : (voir réponse 5 question b de cet exercice)

Exercice 2 : Soient les mesures suivantes de la variable âge : 18, 22, 25, 42, 28, 43, 33, 35, 56, 28. Standardiser cette variable par l'écart absolu moyen ensuite par le Z-score.

Voir réponse 2 question b de cet exercice.

M_{age} = (18+ 22+ 25+ 42+ 28+ 43+ 33+ 35+ 56+ 28)/10=33.

Sage = (|18-33|+|22-33|+….+|28-33|)/10

=(15+11+12+9+5+10+0+2+23+5)/10=9.2

Les mesures standardisées seront : (18-33)/9.2=-1.63, (22-33)/9.2=1.19, ….

Exercice 3 : Soient deux objets représentés par les tuples (22, 1, 42, 10) et (20, 0, 36, 8).

1. Calculer la distance euclidienne entre ces deux objets
2. Calculer la distance de Manhattan entre ces deux objets
3. Calculer la distance de Minkowski entre ces deux objets, en utilisant q=3.

- Distance Euclidienne :

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

$$d(a,b)=\sqrt{(|22-20|^2+|1-0|^2+|42-36|^2+|10-8|^2)}$$

- Distance de Manhattan :

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

- Distance de Minkowski pour q=5:

$$d(i,j)=\sqrt[5]{(|x_{i_1}-x_{j_1}|^5+|x_{i_2}-x_{j_2}|^5+...+|x_{i_p}-x_{j_p}|^5)}$$

Exercice 4: L'observation de quatre variables binaires sur quatre individus différents a conduit aux résultats suivants. Construire la matrice des similitudes en utilisant l'indice de Jaccard.

Indice (ou coefficient) de Jaccard : $d(i,j)=\frac{b+c}{a+b+c}$

	A	B	C	D
A	0			
B	2/3	0		
C	3/4	1	0	
D	2/4	1	1/3	0