

Résumé méthodes classification



ADD_Classification.pdf
PDF

donner moi le résumé de ce pdf expliquer bien et avec des tableaux si possible\$



plus



Résumé Détaillé du Document : Méthodes de Classification

1 Objectifs des Méthodes de Classification

- **But principal** : Regrouper des objets ou des variables en classes homogènes.
- **Types de regroupement** :
 - Individus décrits par des variables.
 - Variables observées sur des individus.
- **Résultat attendu** : Interprétation des regroupements pour identifier des structures ou tendances.

2 Organisation des Données

Élément	Description
Individus (n)	Objets observés (ex : I1, I2, ..., In).
Variables (p)	Caractéristiques des objets (ex : X1, X2, ..., Xp).
Tableau des données (X)	Matrice (Individus × Variables) avec les valeurs x_{ij} .

Types de Variables :

- **Quantitatives** : Mesures numériques (ex : taille, poids).
- **Qualitatives** : Catégories (ex : couleur, genre).
- **Binaires** : Présence/absence (0 ou 1).

3 Distances et Dissimilarités

Les distances permettent de mesurer les écarts entre individus.

A. Pour Variables Quantitatives :

Type de Distance	Formule
Euclidienne	$d^2(I_i, I_l) = \sum_{j=1}^p (x_{ij} - x_{lj})^2$
Minkowski	$(d(I_i, I_l))^p = \sum_{j=1}^p x_{ij} - x_{lj} ^p$
L1	$d(I_i, I_l) = \sum_{j=1}^p x_{ij} - x_{lj} $

B. Pour Variables Qualitatives :

- **Distance Khi-Deux (χ^2)** : Utilisée pour les variables catégorielles.

4 Similarités pour Données Binaires

Pour les variables binaires (0/1), on utilise des indices spécifiques :

Indice de Similarité	Formule	Interprétation
Jaccard	$d = \frac{a}{a+b+c}$	Ignorer les absences communes.
Russel et Rao	$d = \frac{a}{a+b+c+d}$	Prend en compte toutes les situations.
Dice	$d = \frac{2a}{2a+b+c}$	Pondère les présences positives.
Ochiaï	$d = \frac{a}{\sqrt{(a+b)(a+c)}}$	Normalise par les fréquences.

Exemple :

Individu	X1	X2	X3	X4
I1	1	1	0	1
I3	1	0	1	1
I5	1	1	0	1

- $a = 2, b = 1, c = 1, d = 0$
- Jaccard : $\frac{2}{2+1+1} = 0.5$
- Russel et Rao : $\frac{2}{2+1+1+0} = 0.5$

5 Méthodes de Classification

A. Classification Hiérarchique Ascendante (CHA)

Principe :

- Itérative et basée sur le regroupement successif des individus les plus proches.
- Utilise une **matrice de distances** entre individus.

B. Stratégies d'Agrégation :

Méthode	Formule / Description
Saut minimal	$d(h, z) = \min\{d(x, z), d(y, z)\}$
Saut maximal	$d(h, z) = \max\{d(x, z), d(y, z)\}$

Méthode	Formule / Description
Distance moyenne	$d(h, z) = \frac{d(x,z)+d(y,z)}{2}$
Centroïdes	$d(h, z) = d(g_h, g_z)$
Ward	$d_w(h, z) = \frac{n_h n_z}{n_h + n_z} d^2(g_h, g_z)$

6 Algorithme de la CHA

1. **Étape 1** : Créer une matrice de distances ($n \times n$).
2. **Étape 2** : Identifier les deux individus les plus proches.
3. **Étape 3** : Fusionner ces deux individus en un groupe.
4. **Étape 4** : Mettre à jour la matrice et répéter jusqu'à l'obtention d'une seule classe.
5. **Résultat final** : Un **dendrogramme** représentant les regroupements successifs.

Propriétés du Dendrogramme :

- **Feuilles terminales** : Les individus initiaux.
- **Nœuds** : Points de fusion entre deux classes.

7 Exemple Pratique : Classification Hiérarchique Ascendante avec Ward

Individu	X1	X2
I1	2	2
I2	7.5	4
I3	3	3
I4	0.5	5
I5	6	4

1. **Étape 1** : Calcul des distances euclidiennes entre chaque paire.
2. **Étape 2** : Regroupement basé sur la méthode de Ward.
3. **Étape 3** : Représentation finale par un dendrogramme.

✓ Conclusion

- Les méthodes de classification permettent d'analyser et d'interpréter efficacement des jeux de données complexes.
- La **classification hiérarchique ascendante** est particulièrement utile pour visualiser les relations hiérarchiques entre les objets.

Souhaitez-vous que je développe une partie spécifique ou ajoute un exemple détaillé avec les calculs ?

