Méthodes de Classification

AFC

• Le but des méthodes de classification est de construire une partition d'un ensemble d'objets dont on connaît les distances deux à deux. Les classes formées doivent être le plus homogène possible.

• Les méthodes de classification sont utilisées pour regrouper les individus décrits par un ensemble de variables, ou pour regrouper les variables observées sur des individus et d'interpréter les regroupements obtenus.

Classification

Les données:

Les données de départ sont souvent organisées dans un tableau de données X de type (Individus $\times$ Variables) :

Suppose qu'on a p variables $X_1, X_2, ..., X_p$ observées sur n individus $I_1, I_2, ..., I_n$.

	X_1	$\dots$	X_j	...	X_p
I_1	x_{11}	$\dots$	x_{1j}	$\dots$	x_{1p}
.	.		.		.
.	.		.		.
.	.		.		.
I_i	x_{i1}	...	x_{ij}	...	x_{ip}
.	.		.		.
.	.		.		.
.	.		.		.
I_n	x_{n1}	$\dots$	x_{nj}	$\dots$	x_{np}

Classification

- x_{ij} est la valeur de la variable X_j pour l'individu I_i
- n représente le nombre d'individus
- p représente le nombre des variables

L'ensemble des variables peuvent être:

-Quantitatives

-Qualitatives

-Binaires

Distances et dissimilarités

Pour calculer les distances, les données peuvent se présenter sous différentes formes; elles concernent n individus:

• Cas 1: Un tableau de distances entre les n individus pris deux à deux (c-à-d un tableau de n lignes et n colonnes).

• Cas 2: Les observations de p variables quantitatives sur ces n individus.

• Cas 3: Les observations, toujours sur ces n individus, de variables qualitatives (ou binaires).

D'une façon ou d'une autre, il s'agit, dans chaque cas, de se ramener au tableau des distances deux à deux entre les individus (c-à-d au cas 1).

Classification

• Lorsque les données se présentent sous forme d'un tableau X de p variables quantitatives et n individus, on utilise souvent les distances suivantes:

Distance euclidienne:

$$d^2(I_i, I_l) = \sum_{j=1}^{p} (x_{ij} - x_{lj})^2$$

Distance de Minkowsky : dépend d'un paramètre $\lambda > 0$

$$d(I_i, I_l) = (\sum_{j=1}^{p} |x_{ij} - x_{lj}|^{\lambda})^{\frac{1}{\lambda}}$$

Distance L_1:

$$d(I_i, I_l) = \sum_{j=1}^{p} |x_{ij} - x_{lj}|$$

• Lorsque les variables sont qualitatives on utilise la distance de khi-deux χ^2 (voir le cours de l'AFC).

Similarité entre des objets à structure binaire:

Ce cas concerne des données du type suivant: n individus sont décrits par la présence ou l'absence de p variables binaires (c-à-d $X_j \in \{0, 1\}$ pour $j = 1, ..., p$). De nombreux indices de similarité ont été proposés qui combinent de diverses manières les quatre nombres suivants associés à un couple d'individus (I_i, I_l):

- $a = \sum_{j=1}^{p} \mathbb{1}_{(x_{ij}=x_{lj}=1)}$

  c-à-d a = le nombre de fois où $x_{ij} = x_{lj} = 1$
- $b = \sum_{j=1}^{p} \mathbb{1}_{(x_{ij}=0, x_{lj}=1)}$

  c-à-d b = le nombre de fois où $x_{ij} = 0$ et $x_{lj} = 1$
- $c = \sum_{j=1}^{p} \mathbb{1}_{(x_{ij}=1, x_{lj}=0)}$

  c-à-d c = le nombre de fois où $x_{ij} = 1$ et $x_{lj} = 0$
- $d = \sum_{j=1}^{p} \mathbb{1}_{(x_{ij}=x_{lj}=0)}$

  c-à-d d = le nombre de fois où $x_{ij} = x_{lj} = 0$

Classification

Similarité entre des objets à structure binaire:

Les similarités suivantes ont été proposées par différents auteurs:

Jaccard: $d_{il} = \frac{a}{a+b+c}$

Russel et Rao: $d_{il} = \frac{a}{a+b+c+d}$

Dice: $d_{il} = \frac{2a}{2a+b+c}$

Ochiaï: $d_{il} = \frac{a}{(a+b)(a+c)}$

Exemple

On considère le tableau suivant:

	X_1	X_2	X_3	X_4
I_1	1	1	0	1
I_2	1	1	1	1
I_3	1	0	1	1
I_4	0	0	1	0
I_5	1	1	0	1
I_6	0	1	0	0

On cherche à déterminer la similarité entre individus I_3 et I_5.

Classification

Dans ce cas, on a: $a = 2$, $b = 1$, $c = 1$ et $d = 0$

$d_{35} = \frac{a}{a+b+c} = \frac{2}{2+1+1} = \frac{1}{2}$ (Jaccard)

$d_{35} = \frac{a}{a+b+c+d} = \frac{2}{2+1+1+0} = \frac{1}{2}$ (Russel et Rao)

$d_{35} = \frac{2\times 2}{2\times 2+1+1} = \frac{1}{3}$ (Dice)

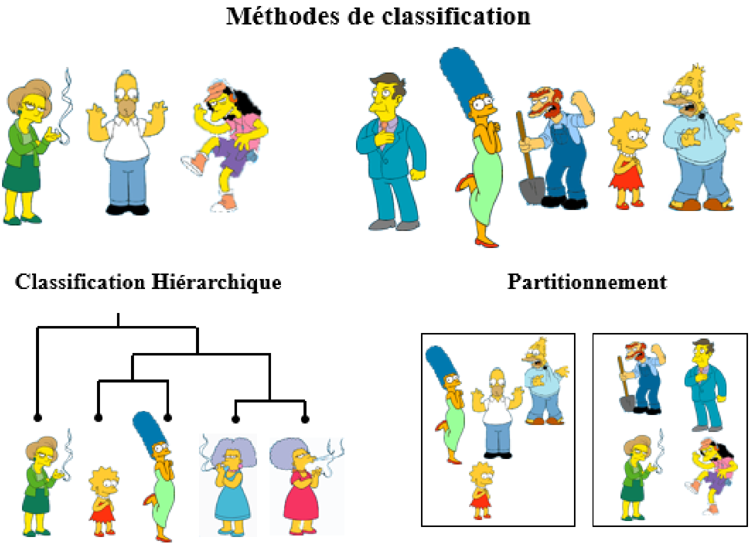
$d_{35} = \frac{a}{(a+b)(a+c)} = \frac{2}{(2+1)(2+1)} = \frac{2}{9}$ (Ochiaï)

Méthodes de classification:

1) Classification hiérarchique ascendante

2) Méthode des centres mobiles

Classification



1) Classification hiérarchique ascendante:

La classification hiérarchique ascendante est une méthode itérative qui consiste, à chaque étape, à regrouper les classes les plus proches. C-à-d à chaque étape, on cherche à créer une partition en agrégeant deux à deux les individus les plus proches.

Le nuage des individus N_I qu'on cherche à classer est supposé muni d'une distance (ou similarité ou dissimilarité) d.

La façon de regrouper des individus ou des groupes d'individus repose sur des **critères d'agrégation**.

Stratégie d'agrégation:

• Première étape:

Si d est une dissimilarité, on choisit I_i et $I_{i'}$ tel que $d(I_i, I_{i'})$ est minimale $\Rightarrow$ $G_1 = \{I_i, I_{i'}\}$

• Deuxième étape:

Nouveau tableau de dissimilarités $(n-1) \times (n-1)$ $\Rightarrow$ nécessite de définir une **méthode d'agrégation** entre un individu et un groupe d'individus ou entre deux groupes d'individus.

Classification

Méthodes d'agrégation:

Soit x, y et z trois classes. Si les classes x et y sont regroupées en une seule classe h, plusieurs critères d'agrégation sont possibles :

- distance du saut minimal : $d(h, z) = min\{d(x, z); d(y, z\}$
- distance du saut maximal : $d(h, z) = max\{d(x, z); d(y, z\}$
- distance moyenne : $d(h, z) = \frac{d(x,z)+d(y,z)}{2}$

Classification

- Méthode des centroïdes: $d(h, z) = d(g_h, g_z)$
- Méthode de la variance (Ward): $d_w(h, z) = \frac{n_h n_z}{n_h + n_z} d^2(g_h, g_z)$

Avec g_h et g_z sont des centres de gravité des classes h et z. n_h et n_z sont des effectifs des classes h et z.

Le saut de Ward joue un rôle particulier et est la stratégie d'agrégation la plus courante.

L'idée de la méthode de Ward est d'agréger les individus en minimisant l'inertie (la variance) intraclasse et en maximisant l'inertie interclasse.

Remarque:

La distance de Ward entre G_1 et G_2, notée $d_w(G_1, G_2)$, est une mesure de la perte d'inertie interclasse lors du regroupement de deux classes G_1 et G_2.

C-à-d, la perte d'inertie inter-classe lors du regroupement de G_1 et G_2 est égale à $\frac{d_w(G_1,G_2)}{n}$

Algorithme de la classification hiérarchique ascendante

Étape 1: Le nuage des individus N_I est une partition P_n de n éléments.

Étape 2: Calculons la matrice des distances $n \times n$ entre les individus. Ensuite, nous recherchons les deux éléments à agréger, c-à-d les deux individus les plus proches en terme de distance.

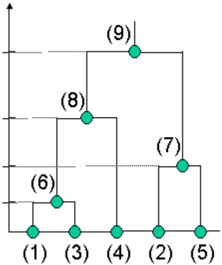
$\Rightarrow$ L'agrégation des deux individus fournit une partition P_{n-1} à $n-1$ individus.

Étape 3: Nous construisons la nouvelle matrice $(n-1) \times (n-1)$ des distances, puis nous recherchons les deux nouveaux éléments à agréger en utilisant une méthode d'agrégation.

$\Rightarrow$ L'agrégation des deux éléments fournit une partition P_{n-2} à $n-2$ individus.

Étape m: Calculons la matrice $(n-(m-1)) \times (n-(m-1))$ des distances, puis nous cherchons à agréger deux éléments jusqu'à l'obtention de la dernière partition P_1.

Les regroupement successifs sont représentés sous la forme d'un arbre ou dendrogramme.



• Les éléments terminaux de dendrogramme représentent les individus.
• Les nœuds de l'arbre correspondent aux regroupements de deux éléments.
Dans le dendrogramme précédent, les éléments terminaux sont les individus (1), (2), (3), (4) et (5). Les nœuds sont (6), (7), (8) et (9). Avec l'effectif de nœud (6) est 2, de nœud (7) est 2, de nœud (8) est 3 et de nœud (9) est 5.

Remarque:

On sait que:

$$I_{totale} = I_{inter} + I_{intra}$$

• Dans l'**Étape 1**, on a $I_{totale} = I_{inter}$ et $I_{intra} = 0$

• Dans l'**Étape 2**, la quantité $\frac{d_w}{n}$ représente la perte d'inertie interclasse lors du premier regroupement, avec d_w est la distance de Ward associé au premier regroupement (agrégation).

• Dans la **dernière étape**, on a $I_{totale} = I_{intra}$ et $I_{inter} = 0$

Exemple: On considère le tableau X de données suivant:

	X_1	X_2
I_1	2	2
I_2	7.5	4
I_3	3	3
I_4	0.5	5
I_5	6	4

On cherche à faire une classification hiérarchique ascendante en utilisant la distance euclidienne et la méthode d'agrégation de Ward.

On note $N_I = \{I_1, I_2, I_3, I_4, I_5\}$ le nuage des individus à classer.

- **Étape 1:** $P_5 = \{I_1, I_2, I_3, I_4, I_5\}$