Classification Automatique.

Avant propos.

Ce cours est adressé aux étudiants de la 2ème année STID.

Trois notions essentielles seront données dans ce cours

- Partition
- Hierarchie
- Discrimination.

Nous donnons des méthode simples de construction d'une partition, une hierarchie et fonction discriminante.

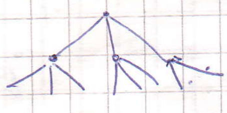
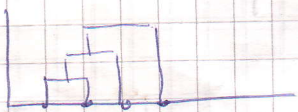
Pour ce but, le premier chapitre presente les notions de base et definitions necessaires.

Classification automatique

I / Introduction.

La classification automatique est un axe des statistiques qui regroupe un ensemble de methodes permettant de rassembler ds une même classe les elts (objets ou individus) qui se ressemble.

Citons : classification automatique par :

- partition  ^{E} : Partition = class disj dont $\cup = E$
- Arbre  grand pere — fils — petits fils — —
- Hierarchie  departement — faculté — universite — Acadie
- Classes empietants : Les classes peuvent se couper.

Nous allons nous interessé à deux entre elles partition basé sur la ressemblance : si deux inds se ressembles ils seront ds la m̄ class sinon ils seront dont de classes ≠. Et la hierarchisation.

Les données sont souvent resumées ds un tabl. ind × variable : n inds caracterisés par p variables.

On resume les etapes comme suit :

étape 1 : Choix des donnés : la population à étudier

étape 2 : Construction d'un tabl de ressemblance, en calculant les ressemblance ou dissimilarités entre touts les inds pris 2 à 2.

étape 3 : Choisi un algorithme de classification : phase de regroupement d'inds.

étape 4 : interpretation des résultats.

II Déf de Distance et dissimilarité.

def 1 : Soit E une population de n inds ou objets.

On appelle indice de dissimilarité d, toute application

$$d : E \times E \longrightarrow \mathbb{R}_+$$
$$(x, y) \longmapsto d(x, y) \quad \text{tq}$$

. $\forall (x,y) \in E\times E$ $\quad d(x,y) = d(y,x)$ symetrie.

. $\forall x \in E$ $\quad d(x,x) = 0$.

si de plus d verifie la inegalité triangulaire

$\forall x, y$ et $z \in E^3$ $\quad d(x,z) \leq d(x,y) + d(y,z)$

alors d est une distance.

Rp : Une distance est donc un indice de dissimilarité, l'invers est faux.

def 2 : On appelle indice de similarité s, toute application

$$s : E\times E \longrightarrow \mathbb{R}_+$$
$$(x,y) \longmapsto s(x,y) \quad tq$$

. $d(x,y) = d(y,x)$ symetrie

. $d(x,x) \geq d(x,y)$.

II 2) Types de distances et dissimilarité

Soit un tab X regroupant n inds operés par k variables quanti

(a) Distance (euclidienne).

A chaque inds $x \in E$, on associe un k-uplet $^t(x_1, \ldots, x_k) \in \mathbb{R}^k$ où x_j represente sa mes par la j^{eme} variable.

- La distance euclidienne entre deux inds x, y est donnée par

$$d(x,y) = \sqrt{\sum_{j=1}^{k} (x_j - y_j)^2}$$

On note par $\sigma_j = \sqrt{Var(x^j)}$ ecart type de la j^{eme} vars.

- La distance euclidienne normé est donnée par.

$$d_N(x,y) = \sqrt{\sum_{j=1}^{k} \frac{1}{\sigma_j^2} (x_j - y_j)^2}$$

- La distance L_1 :

$$d_{L_1}(x,y) = \sum_{j=1}^{k} |x_j - y_j|$$

- La distance Minkowski L_q

$$d_{L_q}(x,y) = \left(\sum_{j=1}^{k} |x_j - y_j|^q \right)^{1/q}$$

$q = 2$ L_2 distance euclidienne

- La distance de Mahalobis.

$$d(x,y) = {}^t(x-y)\, V^{-1}\,(x-y)$$

où V est la matrice des vars_covars de k vars. V de dim k×k

- La distance de khi_2.

$$d^2(x,y) = \sum_{j=1}^{k} \frac{n}{n_{\cdot j}} \left(\frac{n_{xj}}{n_{x\cdot}} - \frac{n_{yj}}{n_{y\cdot}} \right)^2$$

utilisé ds le cas des tabs de contingence, où

n_{xj} est la fréquence de (x,j).

$$n_{x\cdot} = \sum_{j=1}^{k} n_{xj}$$

b) Similarité

Soit une population de n inds à classifier et décris par ~~k~~ variables qualitatives ~~ou~~ k ~~modalités~~ (caractéristiques)

On note a 1 si l'ind possède la caractéristique et par 0 sinon.

Une dissimilarité est obtenue en combinant les similarités presentées par les nombres a, b, c et d suivants:

a = n^{bre} de caractéristiques commune à deux inds i_1 et i_2.

b = " " possédées par i_1 et non par i_2.

c = " " " par i_2 et non par i_1.

d = " " qui ne sont possédés ni par i_1 ni par i_2.

Plusieurs ds similarité ont été proposées, utilisant toutes les n^{bre} a, b, c et d.

- La similarité de Jaccard :

$$s(i_1, i_2) = \frac{a}{a+b+c}$$

- La similarité de Dice ou de Krakanowski

$$s(i_1, i_2) = \frac{2a}{2a+b+c}$$

- La similarité d'Ochiai

$$s(i_1, i_2) = \frac{a}{\sqrt{(a+b)(a+c)}}$$

Yull

$$\frac{ad-bc}{ad+bc}$$

Pearson $\frac{|ad-bc|}{[(a+b)(c+d)(a+c)(b+}$

- La similarité de Rogers et Tanimoto

$$s(i_1, i_2) = \frac{a+d}{a+d+2(b+c)}$$

- La similarité de Russel et Rao

$$s(i_1, i_2) = \frac{a}{a+b+c+d}$$

- Sokal et ~~Sn~~eath (Sneath)

$$\frac{a}{a+2(b+c)}$$

- Sokal et Michener

$$\frac{a+b}{a+b+c+d}$$

- Kulzinsky $\frac{a}{a+b} + \frac{a}{a+c}$

$$= \frac{a}{b}$$

Rq : toutes ses similarités sont comprises entre 0 et 1.

~~On~~ A chaque ~~in~~dice de similarité correspond un indice de dissimilarité, defini par $d = 1 - s \geq 0$.

Exemples

1/ $E = \{0, 1, 2, 3\}$.

$d : E \times E \longrightarrow \mathbb{R}^+$

. $d(1,0) = d(1,0) = 1$. $\quad d(0,2) = d(2,0) = 2$.

. $d(1,3) = d(3,1) = 4$ $\quad d(0,3) = d(3,0) = 3$.

. $d(1,2) = d(2,1) = 3$

. $d(2,3) = d(3,2) = 6$

$d(0,0) = d(1,1) = d(2,2) = d(3,3) = 0$

d est un indice de dissimilarité mais pas une distance.

$d(2,3) = 6$

$d(2,0) + d(0,3) = 2+3 = 5 < 6$

l'inegalité triangulaire n'est pas verifiée.

On defini le tab de dissimilarité suivant.

E	0	1	2	3
0	0	1	2	3
1	1	0	3	4
2	2	3	0	6
3	3	4	6	0

symetrique de diagonale nulle.

exo 2/ $E = \{0, 1, 2, 3\}$.

- $d(0,0) = d(1,1) = d(2,2) = d(3,3) = 0$.
- $d(0,1) = 1 = d(1,0)$. $d(1,2) = d(2,1) = 1$. $d(2,3) = d(3,2) = 1$
- $d(0,2) = d(2,0) = 2$. . $d(1,3) = d(3,1) = 2$
- $d(0,3) = d(3,0) = 3$

Le tableau de distances est.

	0	1	2	3
0	0	1	2	3
1	1	0	1	2
2	2	1	0	1
3	3	2	1	0

= distance entre 1 et 2.

exo3/ Soit 4 inds à qui on a posé 4 questions A, B, C et D auxquelles ils doivent répondre par oui ou non.

On note par 1 si la réponse est oui et par zéro si la réponse est non et soit le tabl de données suivant:

inds \ Quests	Q_1	Q_2	Q_3	Q_4
i_1	non 0	oui 1	oui 1	oui 1
i_2	0	0	1	1
i_3	1	0	0	1
i_4	1	1	0	1

Calculons les quantités a, b, c et d entre i_1 et i_2. $Q_3^{i_1} = Q_3^{i_2} = 1$ $Q_4^{i_1} = Q_4^{i_2} = 1$

$a = 2$ 2 réponse oui communes.

$b = 1$ $i_1 : Q_1 = 1$ et $Q_2 = 0$.

$c = 0$

$d = 1$ $Q_1^{i_1} = Q_2^{i_2} = 0$

l'indices de similarité ~~de~~ entre i_1 et i_2.

- Jaccard $s(i_1, i_2) = \frac{2}{2+1+0} = \frac{2}{3}$.
- Dice ou Czekanowski $s(i_1, i_2) = \frac{4}{4+1+0} = \frac{4}{5}$
- Ochiai $s(i_1, i_2) = \frac{2}{\sqrt{(2+1)\cdot(2+0)}} = \frac{2}{\sqrt{6}}$
- Rogers et Tanimoto $s(i_1, i_2) = \frac{2+1}{2+1+2(1+0)} = \frac{3}{5}$