Classification automatique

Ghislain WANDJI

Polytech Lille

GIS2A4

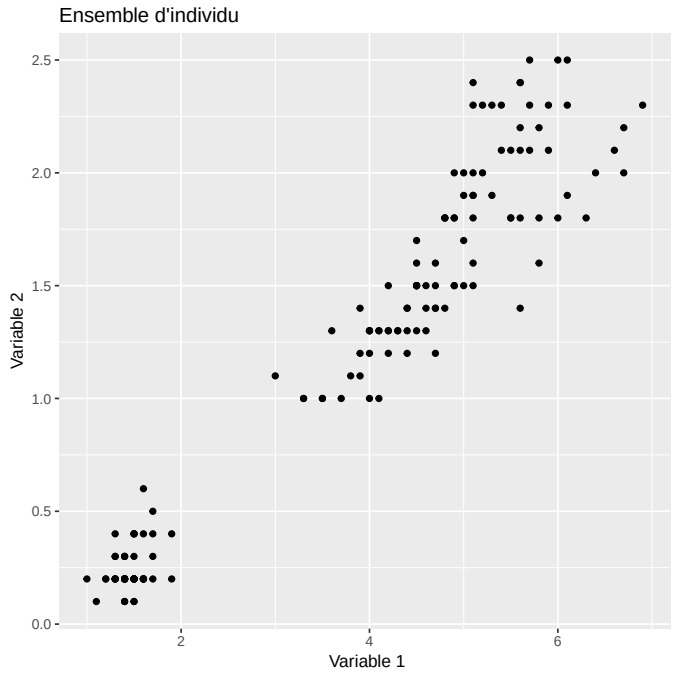
Sommaire

Sommaire

La classification automatique est la plus répandue des techniques d'analyses de données.

Elle permet de définir des sous-ensembles homogènes d'individus (ou de variables) à partir d'un volume important de données.

Elle n'a pas un but prédictif, on ne dispose ici d'aucune variable à expliquer.



Domaines d'application

Différentes expressions sont utilisées pour la désigner suivant le contexte :

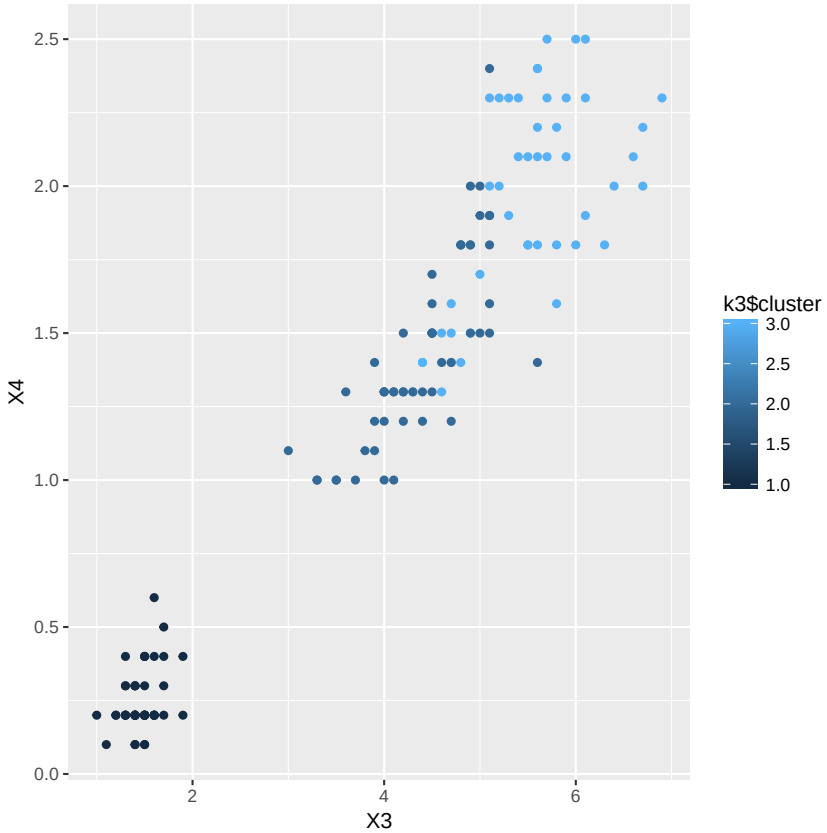
- Marketing : on parlera très souvent de segmentation de la clientèle
- Médécine : Nosologie (Etudie les principes généraux de classification des maladies)
- De nombreuses applications dans les domaines de data-mining et machine learning

Attention

Classification en anglais désigne la classification supervisée
Pour la classification automatique (non supervisée), on utilise très souvent le terme de clustering

Opération statistique consistant à regrouper les objets d'un ensemble de telle sorte que :

- Les objets de chacun de sous-ensemble présentent des caractéristiques similaires (homogénéité interne)
- les différents sous-ensembles présentent des caractéristiques les plus différents possibles (hétérogénéité externe)



Les méthodes de classifications peuvent être divisées en deux grands groupes :

- **La classification par partitionnement** : Elle consiste à partir d'un nombre de classes fixé au départ (k), à diviser les données d'un ensemble en k classes disjointes.
- **La classification hiérarchique** : Elle produit une séquence de partitions emboitées, de la plus fine à la plus grossière, conduisant à des resultats sous forme de dendogramme.

Pour chacun de ces deux groupes, les méthodes peuvent êtres divisées en trois catégories d'appoches fondées sur :

Une distance : Utilisation d'une notion de distance à partir de la quelle on essaye de regrouper entre eux les individus les proches.

Un modèle : Les données de chaque classe sont supposées suivre une distribution statistique spécifique, l'ensemble formant un mélange de distribution.

La densité : Chaque classe est considérée comme une région dense, que l'on compare à des régions plus clairesemées.

On s'intéresse ici aux approches géométriques fondées sur les notions de distances.

Sommaire

On considère ici un ensemble de données constitué de n observations, sur lesquelles sont mesurées p caractéristiques

Notations :

- $\Gamma = \{\omega_1, \ldots, \omega_n\}$ représente l'ensemble de n individus

- Pour chacun des $\omega_i, i = \{1, \ldots, n\}$ individus, on dispose de p valeurs des p caractères $X_1, \ldots, X_p$.

- L'ensemble des données est représenté sous la forme matricielle

$$M = \begin{pmatrix} x_{1,1} & \cdots & x_{1,p} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,p} \end{pmatrix}, \text{ où } \forall (i,j) \in (1, \ldots, n) \times (1, \ldots, p)$$

$x_{i,j}$ représente l'observation de la variable X_j pour l'individu i.

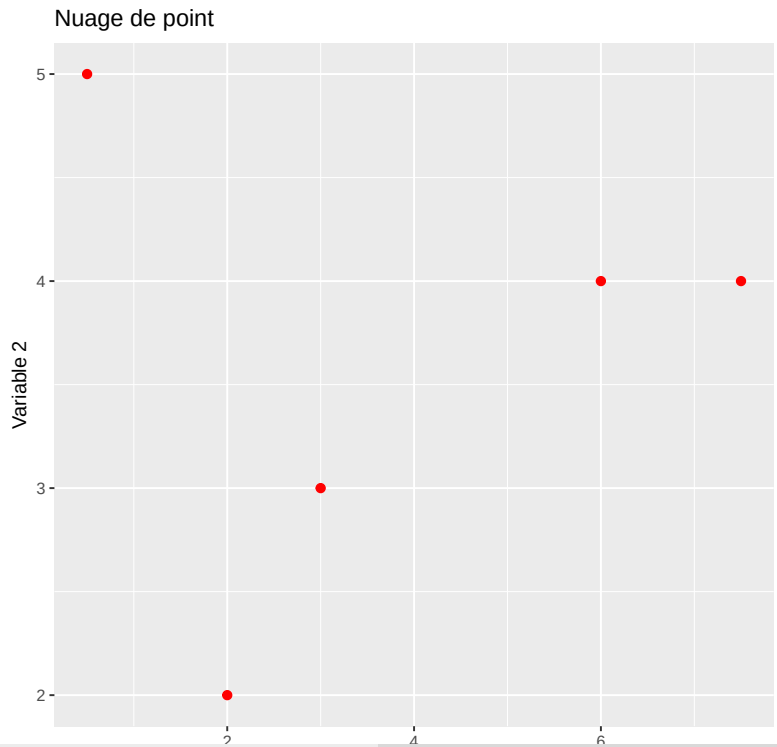
Il peut s'agir d'un tableau de données numériques, d'un tableau de contingence, ou plus généralement d'un tableau de données mixtes.

La representation graphique dans $\mathbb{R}^p$ des $\omega_i, i = \{1, \ldots, n\}$ est appelée **Nuage de points**.

Chaque individu ω_i étant représenté par le point de coordonnées $(x_{i,1}, \ldots, x_{i,p}) \in \mathbb{R}^p$

Exemple :

$$X = \begin{pmatrix} 2 & 2 \\ 7.5 & 4 \\ 3 & 3 \\ 0.5 & 5 \\ 6 & 4 \end{pmatrix}$$



Considérons maintenant l'ensemble illustratif de 6 individus dans un espace de dimmension 4 suivant :

	probabilites	Statistique	genie.logiciel	structure.de.donnes
Pierre	19.00	17.00	2.00	8.00
Paul	9.00	11.00	13.00	12.00
Yoann	20.00	19.00	11.00	13.00
Annabelle	1.00	6.00	18.00	17.00
Cindy	10.00	11.00	12.00	12.00
Stacy	20.00	12.00	18.00	18.00

Une représentation graphique dans $\mathbb{R}^4$ est impossible.

Une des solutions serait de considerer le plan factoriel de l'ACP.

Question : Comment procède-t-on et sur quel(s) critère(s) nous basons-nous pour définir les groupes d'individus ?

On appelle distance sur un ensemble E, toute application $d : E \to \mathbb{R}_+$ satisfaisant les axiomes suivants :

1. $\forall (x, y) \in E^2, d(x, y) > 0 \Rightarrow x \neq y$,
2. $\forall (x, y) \in E^2, d(x, y) = 0 \Leftrightarrow x = y$
3. $\forall (x, y) \in E^2, d(x, y) = d(y, x)$
4. $\forall (x, y, z) \in E^3, d(x, z) \leqslant d(x, y) + d(y, z)$

On appelle similarité, toute application $s : E^2 \to \mathbb{R}_+$ telle que :

1. $\forall (x, y) \in E^2, s(x, y) = s(y, x) \geqslant 0$
2. $\forall (x, y) \in E^2, s(x, x) = s_{max} > s(x, y)$, où s_{max} est la plus grande similarité possible.

Plus les individus se ressemblent, plus la valeur prise par la fonction est élevée.

On appelle dissimilartié, toute application $dis : E^2 \rightarrow \mathbb{R}_+$ telle que :

1. $\forall (x, y) \in E^2, dis(x, y) = 0 \Leftrightarrow x = y$
2. $\forall (x, y) \in E^2, dis(x, y) = dis(y, x)$

Moins deux éléments se ressemblent, plus la valeur de la fonction est élevée

La dissimilarité est définie à partir de l'indice de similarité par $dis(x, y) = s_{max} - s(x, y), \forall (x, y) \in E^2$

Les différentes méthodes de classification utilisent différentes distances et indices de similarités afin d'apprécier la ressemblance entre des individus.

L'évaluation de cette ressemblance dépend le plus souvent de la nature des données étudiées.

Mesure de ressemblance entre individus

Sur des données numériques

On utilise la notion de distance afin d'évaluer La ressemblance entre les individus.

La distance la plus générale utilisée dans $\mathbb{R}^p$ à partir de données numériques est la distance de `Minkowski` définie par :

Sachant $k \in \mathbb{N}, x = (x_1, \ldots, x_k) \in \mathbb{R}^k, y = (y_1, \ldots, y_k) \in \mathbb{R}^k$:

- $d(x, y) = (\sum_{i=1}^{k} |x_i - y_i|^q)^{\frac{1}{q}}), \forall q \geqslant 1.$

Cas particuliers :

- $q = 1$: distance de city-block ou manhattan $d(x, y) = \sum_{i=1}^{k} |x_i - y_i|$
- $q = 1$: distance de euclidienne $d(x, y) = \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2}$

Mesure de ressemblance entre individus

Sur des données numériques

On utilise en général :

- la distance euclidienne simple losrque toutes les variables ont la même échelle de mesure.
- La distance euclidienne normalisée sinon.

$$d(x,y) = \sqrt{\sum_{i=1}^{k} \frac{1}{\sigma_j^2}(x_i - y_i)^2}$$

remarque : Cela revient à calculer la distance euclidienne simple sur des données standardisées.

Mesure de ressemblance entre individus

Sur des données numériques

La distance entre deux groupes d'individus de taille respectives n_1 et n_2 et dont les caractéristiques de dispersion sont $\sum_1$ et $\sum_2$ est obtenue à partir de la distance `Mahalanobis` par :

$$- D = [(g_1 - g_2)' \sum^{-1} (g_1 - g_2)]^{\frac{1}{2}},$$

où g_1 et g_2 représentent les moyennes des deux groupes d'individus,

et $\sum = \frac{n_1 \sum_1 + n_2 \sum_2}{n_1 + n_2 - 2}$ est la matrice de dispersion moyenne estimée.

remarque : Si $\sum$ est la matrice diagonale on obtient la distance euclidienne normalisée.

Mesure de ressemblance entre individus

Sur des données ordinales

On remplace les $X_{i,j}$ par leur rang $r_{i,j}$, $r_{i,j} = 1, 2, \ldots, k_j$, où k_j est le nombre de modalité de la variables i

On Calcule pour chaque individu i, $Z_{i,j} = \frac{r_{i,j}-1}{k_j-1}$; $Z_{i,j} \in [0, 1]$

Les distances entre les individus sont ensuite calculées à partir des $Z_{i,j}$, en les considérant comme des données numériques.

Mesure de ressemblance entre individus

Sur des données binaires (1/2)

Les valeurs possibles pour les p variables sont 0 ou 1. La ressemblance entes deux individus x, y se calcule à partir du `tableau de contingence` suivant :

	0	1	
0	$p_{0,0}$	$p_{0,1}$	$p_{0,.}$
1	$p_{1,0}$	$p_{1,1}$	$p_{1,.}$
	$p_{.,0}$	$p_{.,1}$	p

Sont calculables à partir de ce tableau :
- Le nombre de concordance $p_c = p_{0,0} + p_{1,1}$
- Le nombre de discordance $p_d = p_{0,1} + p_{1,0}$

Parmi les indices de similarités entre deux individus, les plus connus sont :

Jaccard : $\frac{p_{0,0}}{p_{0,0}+p_{0,1}+p_{1,0}}$ Russel-Rao : $\frac{p_{0,0}}{p}$

Mesure de ressemblance entre individus

Sur des données de fréquences

La distance entre deux lignes d'un tableau de fréquence (n, p), de terme général $f_{i,j}$ se calcule à partir des composants $X_{i,j} = \frac{f_{i,j}}{f_i}$
- f_i représentee les termes généraux des profils lignes.

$$d(i, i') = \sum_{j=1}^{p} \frac{1}{f_{.,j}} \left(\frac{f_{i,j}}{f_i} - \frac{f_{i',j}}{f_{i'}} \right)$$

Cette distance est connue sous le nom de distance du khi-deux.

Mesure de ressemblance entre individus

Sur des données nominales

Il existe deux méthodes couramment uitilisées pour évaluer la distances à partir de données nominales :

Méthode 1 : - Chaque variable nominale est transformée en autant de variables binaires que de modalités qu'elle présente.
- On se ramène ensuite à un calcul de distance en cas de données binaires.

Méthode 2 : Si deux individus (x,y) présentent la même modalité pour les m variables (m étant appelé nombre d'appartenance), alors $d(x, y) = \frac{p-m}{p}$, où p est le nombre de variables nominales.

Exemple :

	Sexe	CSP	RGM
A_1	H	R	C
A_2	H	R	C
A_3	F	S	M
A_4	F	R	K
A_5	F	S	M
A_6	H	S	K

1. En considérant l'indice de Jaccard, calculer $s(A_6, A_5)$ et $s(A_1, A_4)$.
2. L'individu A_6 est-il plus proche de A_5 que A_1 de A_4 ?

Mesure de similarités entres variables

Sur des données numériques

La similarité entre deux variables j et j' est donnée par la corrélarion

$$r_{j,j'} = \frac{\sum_{i=1}^{n}(X_{i,j}-\bar{X}_j)(X_{i,j'}-\bar{X}_{j'})}{[\sum_{i=1}^{n}(X_{i,j}-\bar{X}_j)^2 \sum_{i=1}^{n}(X_{i,j'}-\bar{X}_{j'})^2]^{\frac{1}{2}}}$$

n : est le nombre d'individus de l'ensemble à classer
$\bar{X}_j$: est la moyenne de la variable j.

A noter qu'une similarité est une distance ne vérifiant pas nécessairement l'inégalité triangulaire.

Mesure de similarités entres variables

Sur des données binaires

	0	1	
0	$n_{0,0}$	$n_{0,1}$	$n_{0,.}$
1	$n_{1,0}$	$n_{1,1}$	$n_{1,.}$
	$n_{.,0}$	$n_{.,1}$	n

L'indice de similarité le plus courant est le Φ^2 de Pearson, qui prend des valeurs comprises entre 0 et 1 et est obtenu à partir du khi-deux de contingence :

$$\chi^2_{j,j\prime} = \frac{n(n_{0,0}n_{1,1} - n_{0,1}n_{1,0})^2}{n_{.,0}n_{.,1}n_{0,.}n_{1,.}} \text{ ; par } \Phi^2_{j,j\prime} = \frac{\chi^2_{j,j\prime}}{n}$$

Mesure de similarités entres variables

Sur des données nominales

L'indice de similarité entre deux variables j et j' est calculé à partir du tableau de contingence $\mathcal{T}_{q,r}$ croisant leurs modalités respectives q et r.

Leur indice de similartié est compris entre 0 et 1, il s'agit du coefficient de Cramer obtenu à partir de $\Phi^2_{j,j'}$:

$$\mathcal{C}_{j,j'} = \frac{\Phi^2_{j,j'}}{\min(r-1,q-1)}$$

Définition

Soit $\mathcal{P}(\Gamma)$ l'ensemble des parties de Γ. On appelle écart, toute application $e : \mathcal{P}(\Gamma) \to \mathbb{R}_+$ définie à partir d'une distance et évaluant la proximité entre deux groupes d'individus.

Les écarts usuels entres groupes d'individus sont :

Ecart simple ou simple linkage

Il s'agit de la méthode du plus proche voisin.
L'écart entre deux groupes correspond à la plus faible distance entre deux points de chacun des groupes.

$$\forall (A, B) \subset \mathcal{P}(\Gamma)^2, e(A, B) = \min_{\omega \in A, \omega^* \in B} d(\omega, \omega^*)$$

Ecart Complet ou complete linkage

Il s'agit de la méthode du voisin le plus éloigné.
L'écart entre deux groupes correspond à la plus forte distance entre deux points de chacun des groupes.

$$\forall (A, B) \subset \mathcal{P}(\Gamma)^2, e(A, B) = \max_{\omega \in A, \omega^* \in B} d(\omega, \omega^*)$$

Ecart moyen ou average linkage

C'est la distance moyenne entre tous les points de A et B.

$$\forall (A, B) \subset \mathcal{P}(\Gamma)^2, e(A, B) = \frac{1}{n_A n_B} \sum_{\omega \in A} \sum_{\omega^* \in B} d(\omega, \omega^*)$$

Ecart de Ward

Soient $(A, B) \subset \mathcal{P}(\Gamma)^2$. G_A et G_B leur centre de gravité respectif et d la distance euclidienne.

$$e(A, B) = \frac{n_A n_B}{n_A + n_B} d^2(G_A, G_B)$$

Elle prend en compte la dispersion à l'interieur mais aussi à l'extérieur d'un groupe

Inertie d'un nuage de points

Etant donné un nuage de point $\mathcal{N} = \{m_1, \ldots, m_n\}$ son centre de gravité a pour coordonnées $(\bar{x}_1, \ldots, \bar{x}_p)$ avec $\bar{x}_j = \frac{1}{n} \sum_{i=1}^{p} x_{i,j}, \forall j \in \{1, \ldots, p\}$

Soient $k \in \{1, \ldots, n\}$ et $\mathcal{P} = (\mathcal{N}_r)_{r \in \{1,\ldots,k\}}$ une partition de $\mathcal{N}_r$. L'inertie autour de $\mathcal{N}_r$:

$\mathcal{I}(\mathcal{N}_r) = \frac{1}{n_r} \sum_{i \in \mathcal{N}_r} d^2(\omega_i, g_r)$, où g_r représente le centre de gravité de $\mathcal{N}_r$

$\mathcal{N}_r$ est d'autant plus homogène que son inertie est faible

Inertie interclasse

L'inertie interclasse est la moyenne des carrées des distances des barycentres des classes au barycentre global.

$$\mathcal{I}_{inter}(\mathcal{P}) = \frac{1}{n}\sum_{r=1}^{k} n_r d^2(g_r, g)$$

g_r représente le centre de gravité de la classe $\mathcal{C}_r$ et g celui de l'ensemble.

Elle mesure la séparation entre les sous-nuages, plus elle est élevée plus les classes sont séparées les unes des autres. Ce qui indique une bonne classification.

Inertie intraclasse

L'inertie intraclasse mesure l'hétérogénéité de l'ensemble des sous-nuages de $\mathcal{N}$. C'est la somme des interties totales de chaque classe.

$$\mathcal{I}_{intra}(\mathcal{P}) = \frac{1}{n}\sum_{r=1}^{k} n_r \mathcal{I}(\mathcal{N}_r)$$

Une bonne partition a une inertie intraclasse faible et une inertie interclasse élevée.

Remarques :

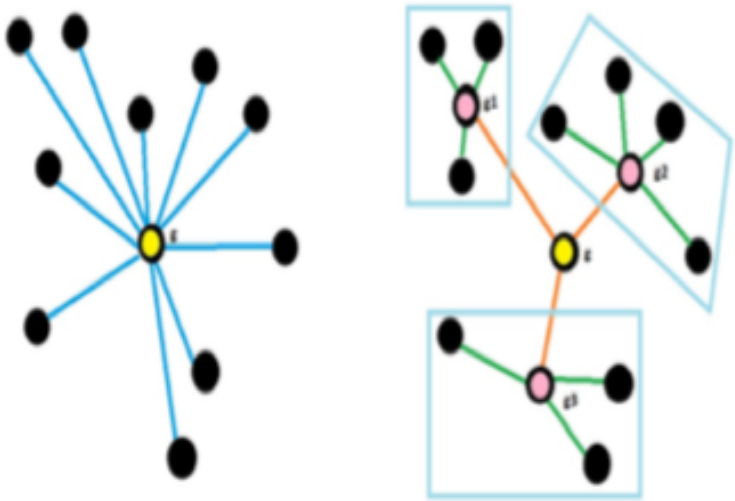
l'inertie totale, intra-classe et inter-classe est aussi appellée

- variance intra et inter-classe lorsque $n_i = \frac{1}{n}$
- somme des carrés intra et inter-classe losrque $n_i = 1$.

Décomposition de Huygens

Pour toute partition $\mathcal{P}$ de $\mathcal{N}$, l'inertie totale de $\mathcal{P}$ est obtenue par :

$$\mathcal{I}_{totale}(\mathcal{P}) = \mathcal{I}_{intra}(\mathcal{P}) + \mathcal{I}_{inter}(\mathcal{P})$$



Inertie Expliquée

Pour une partition $\mathcal{P}$ donnée, son pourcentage d'inertie expliquée est donné par la formule :

$$100(1 - \frac{\mathcal{I}_{intra}(\mathcal{P})}{\mathcal{I}_{totale}(\mathcal{P})})$$

Sa valeur croît lorsque le nombre de classes augmente, il permet de comparer deux partitions ayant le même nombre de classes.

Exemple

Exercice

calculer l'inertie totale,l'inertie intra-classe et l'inertie inter-classe de la partition $\mathcal{P} = \{\{\omega_1, \omega_3\}, \{\omega_4\}, \{\omega_2, \omega_5\}, \{\omega_6\}\})$

On donne

	X_1	X_2
ω_1	2	2
ω_2	7.5	4
ω_3	3	3
ω_4	0.5	5
ω_5	6	4
ω_6	15	10

En déduire le pourcentage d'inertie expliquée.

Partitions

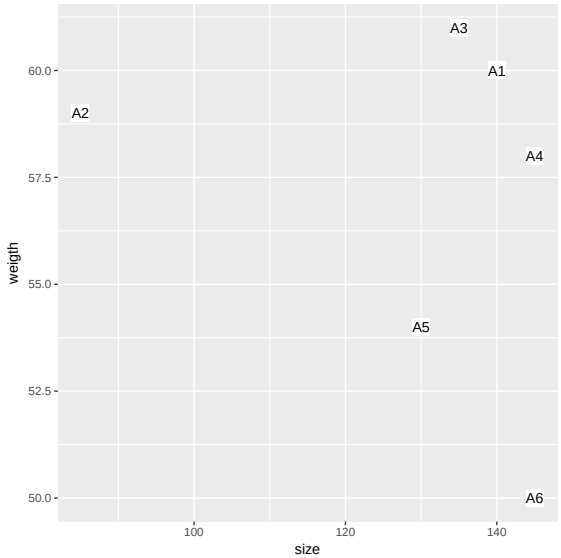
Soient $\mathcal{P}_1, \ldots, \mathcal{P}_n$, n éléments de Γ

$(\mathcal{P}_1, \ldots, \mathcal{P}_n)$ Constitue une partition de Γ si et seulement si :

- $\forall k \in \{1, \ldots, n\}, \mathcal{P}_k \neq \emptyset$
- $\forall p \neq q, \mathcal{P}_p \cap \mathcal{P}_q = \emptyset$
- $\cup_{k=1}^{n} \mathcal{P}_k = \Gamma$

Exemple : $\{\{1, 2\}, \{3\}, \{4, 5, 6\}\}$ forme une partition de $\{1, 2, 3, 4, 5, 6\}$

Qu'en est-il de $\{\{1, 2, 3\}, \{3\}, \{4, 5, 6\}\}$?



Hierarchie

Une famille H_Γ de $\mathcal{P}(\Gamma)$ est une hiérarchie si :

- $\Gamma \in H_\Gamma$
- $\forall \gamma \in \Gamma, \{\gamma\} \in H_\Gamma$
- $\forall A, B \in H_\Gamma, A \cap B \in \{\emptyset, A, B\}$

A toute hiérarchie correspond un arbre de classification.

Lorsqu'il existe une relation de pré-ordre compatible avec une relation d'ordre naturelle, on dit que H_Γ est stratifié.

Une hiérarchie est dite indicée si il existe une application croissante v définie de H_Γ à valeur dans $\mathbb{R}_+$.

Sommaire

introduction

Méthodes visant à constituer au sein d'un ensemble une partition de k classes disjointes non vides.

Processus itératif visant à optimiser un certain critère défini à priori
Les algorithmes de partitionnement sont divisés en deux grands groupes :

- Les algorithmes k-moyennes
- Les algorithmes k-représentants

Pour un ensemble E à n éléments, il existe $B_n = \frac{1}{e}\sum_{k=1}^{\infty}\frac{k^n}{k!}$ Partitions possibles. B_n est appelé **nombre de Bell**

Exemple

$\Gamma = \{a, b, c, d\}$: Combien de partitions possibles ?

Introduction

Méthodes très populaires dans les applications scientifiques

Elles produisent k classes à partir d'un ensemble de données de n objets de telle sorte que la fonction objective soit minimum.

Fonction objective

$$\Phi = \sum_{r=1}^{k} \sum_{\omega_i \in \mathcal{C}_r} (\omega_i - \mathcal{G}_r)$$

où $\mathcal{C}_r$ représente les classes obtenues, et $\mathcal{G}_r$ le barycentre de la classe $\mathcal{C}_r$.

Il existe de nombreuses méthodes de type centres mobiles, qui différent suivant :

- La sélection des centres initiaux.
- Les stratégies de calcul des moyennes des classes.
- La prise en compte de données catégorielles

Description

Construction d'une partition en k classes à partir de k objets considérés comme centres initiaux des classes.

Centres initiaux sont tirés au hasard, et affectation des objets aux classes en fonction de leurs proximités à ces centres.

Les centres des gravités des classes ainsi obtenus constituent des nouveaux centres, qui fournissent une nouvelle partition.

On réitère le processus jusqu'à la stabilité des partitions.

Algorithme

Etape 0 : - Séléction au hasard de k individus définis comme centres initiaux de classes, formant ainsi un ensemble $\mathcal{C}^0 = \{\mathcal{C}_1^0, \ldots, \mathcal{C}_k^0\}$.

- Affectation des individus au centre le plus proche :
  $\forall i \in n$, on determine k^* tel que $k^* = \text{argmin}_{k \in 1,\ldots,K}\, d(\omega_i, \mathcal{C}_k^0)$

Est ainsi obtenu une partition en k classes $\mathcal{P}^0 = \{\mathcal{P}_1^0, \ldots, \mathcal{P}_k^0\}$

Etape 1 :- Pour chacune des $\mathcal{P}_{i,i=1,\ldots,k}^0$ classes, est calculé son barycentre $\mathcal{C}_i^1$, qui sera défini comme centre de la classe. Formant ainsi l'ensemble $\mathcal{C}^1 = \{\mathcal{C}_1^1, \ldots, \mathcal{C}_k^1\}$

- En utilisant le même principe qu'à l'étape précédente, est définit une nouvelle partition $\mathcal{P}^1 = \{\mathcal{P}_1^1, \ldots, \mathcal{P}_k^1\}$

Etape m : Les Centres $\mathcal{C}^m = \{\mathcal{C}_1^m, \ldots, \mathcal{C}_k^m\}$ sont déterminés à partir de la partition

$\mathcal{P}^{m-1} = \{\mathcal{P}_1^{m-1}, \ldots, \mathcal{P}_k^{m-1}\}$ obtenue en $m-1$

Exemple d'application

Soient 6 individus pour lesquels sont observées les variables X_1,X_2

	X_1	X_2
ω_1	3	0
ω_2	-2	3
ω_3	-2	2
ω_4	-2	-1
ω_5	2	2
ω_6	0	-1

1. A partir des centres initiaux $\mathcal{C}_1^0 = \omega_6$ et $\mathcal{C}_2^0 = \omega_2$, construire à l'aide de la distance euclidienne, en utilisant l'aglorithme des centres mobiles une partition des 6 individus.
2. idem avec $\mathcal{C}_1^0 = \omega_4$ et $\mathcal{C}_2^0 = \omega_2$.

Variantes : Méthode k-means

Proposée par Mac Queen (1967) ; il s'agit d'une modification de l'algorithme des centres mobiles.

Les centres des classes,sont recalculés à chaque affectation d'un individu à une classe.

L'algorithme est plus efficace, mais dépend de l'ordre des individus dans le fichier.
Exemple :

	X_1	X_2
A	1	1
B	2	1
C	3	3
D	4	4

En considérant A et B comme centres initiaux, appliquer l'algorithme k-means afin de constituer une partition de l'ensemble des individus.

Variantes :nuées dynamiques

Description

La méthode des nuées dynamique se distingue de celle des centres mobiles par le fait que chaque classe n'est plus représentée par son centre, mais par un sous-ensemble de la classe appelé **noyau**.

Lorsque le noyau est bien constitué, celui-ci est plus représentatif de la classe que le barycentre.

Variantes :nuées dynamiques

Algorithme

Sont choisis au hasard k sous-ensemble tels que :

- $\text{card}(\mathcal{N}_r^0) = p, \forall r = 1, \ldots, k$
- $\mathcal{N}_r^0 \cap \mathcal{N}_s^0 \forall r \neq s = 1, \ldots, k$

L'ensemble $\mathcal{N}^0 = \{\mathcal{N}_1^0, \ldots, \mathcal{N}_k^0\}$ consitue une famille de Γ.

Variantes :nuées dynamiques

Algorithme

Sont définis deux fonctions :

fonction d'affectation f $f : \mathcal{N}_r^0 \to \mathcal{P}_r^0 \subset \Gamma$;

$\mathcal{P}_r^0 = \{x \in \Gamma \, d(x, \mathcal{N}_r^0) \leqslant d(x, \mathcal{N}_s^0), \forall r \neq s = 1, \ldots, k\}$.

L'ensemble $\mathcal{P}^0 = \{\mathcal{P}_1^0, \ldots, \mathcal{P}_k^0\}$ constitue une partition de Γ.

fonction de représentation g $g : \mathcal{P}_r^0 \to \mathcal{N}_r^1; card(\mathcal{N}_r^1) = q, \forall r = 1, \ldots, k$.

On suppose que l'ensemble $\mathcal{N}_r^1$ de q éléments de Γ qui minimise $\sum_{x \in \mathcal{N}_r^1} d(x, \mathcal{P}_r^0)$ existe et est unique.

$$\sum_{x \in \mathcal{N}_r^1} d(x, \mathcal{P}_r^0) = \min(\sum_{x \in A} d(x, \mathcal{P}_r^0); A \subset \Gamma, \text{card}(A) = q)$$

Variantes :nuées dynamiques

Algorithme

On construit ainsi de manière itérative à l'aide de f et g la chaîne :

$$\mathcal{N}_r^0 \xrightarrow{f} \mathcal{P}_r^0 \xrightarrow{g} \mathcal{N}_r^1 \xrightarrow{f} \mathcal{P}_r^1 \xrightarrow{g} \ldots \xrightarrow{g} \mathcal{N}_r^i \xrightarrow{f} \mathcal{P}_r^i \xrightarrow{g} \ldots$$

- En utilisant le critère d'optimisation $U_i = \mathcal{W}(\mathcal{N}_r^i, \mathcal{P}_r^i) = \sum_{r=1}^{k} d(\mathcal{N}_r^i, \mathcal{P}_r^i)$ à l'itération i.
- Le processus converge vers une partition optimale en un nombre fini d'itérations.

k-modes

Algorithme orienté vers des données qualitatives

L'appoche est similiaire à celle de méthodes de type centre mobile

La seule différence réside sur le choix de la mesure de dissimilarité utilisée pour evaluer la proximité entre deux individus.

k-modes

Les moyennes sont remplacées par les modes, et une méthode basée sur les fréquences est utilisée pour mettre à jour les modes.

$$d(X_i, X_{i'}) = \sum_{j=1}^{p} \frac{n_{X_{i,j}} + n_{X_{i,j'}}}{n_{X_{i,j}} n_{X_{i,j'}}} \mathbb{1}_{\{X_{i,j} \neq X_{i',j}\}}$$

$n_{X_{i,j}}$ correspond au nombre d'objets de l'échantillon dont les valeurs respectives sont $X_{i,j}$

Chaque classe $\mathcal{C}_r$ a un mode défini par un vecteur $V^r = (X_1^r, \ldots, X_p^r)$.
On cherche tout au long du processus l'ensemble de vecteur V^c qui rend minimum $Q = \sum_1^k \sum_{X \in \mathcal{C}_r} d(X, V^r)$

Méthodes relationelles

Les individus sont décrits par p variables qualitatives à $m_1, \ldots, m_p$ modalités respectivement.

Les individus sont représentés sous forme de relation d'équivalence.

Une classification est une relation d'équivalence $\mathcal{R}$ où $i\mathcal{R}j$, si i et j sont dans la même classe.

On associe $\mathcal{R}$ à une matrice $n \times n$ définie par $m_{i,j} = \mathbb{1}_{\{i\mathcal{R}j\}}$

Les 3 propriétés de la relation d'équivalence se traduisent par :

1. $m_{i,j} = m_{j,i}$
2. $m_{i,i} = 1$
3. $m_{i,j} + m_{j,k} - m_{i,k} \leqslant 1$

Méthodes relationelles

La recherche d'une classification revient alors à chercher une matrice $\mathcal{M} = (m_{i,j})$ satisfaisant les trois propriétés précédentes.

On cherche une classification fournissant un bon compromis entre les p classifications initiales.

On pose :
$m_{i,j}$= Nombre de fois où i et j ont été mis dans la même classe.
$\mathcal{M}' = (m'_{i,j}) = 2m_{i,j} - p$

- $m_{i,j} \geqslant 0$ si i et j sont dans la même classe (coïncident) pour une majorité de variables
- $m_{i,j} \leqslant 0$ si i et j sont dans des classes différentes pour une majorité de variables.

Méthodes relationelles

Un critère naturel pour former une partition centrale consiste à mettre i et j dans la même classe si si $m'_{i,j} \geqslant 0$ et les séparer sinon.

Ce critère naturel ne fournit cependant pas toujours une partition : il pourrait y avoir non transitivité de la règle majoritaire (c'est le paradoxe de Condorcet).

On est ramené à un problème de programmation linéaire

Soient :

$\mathcal{C}(A, S) = \sum_{B_i \in S}(A, B_i)$ le critère de Condorcet d'un individu A avec un ensemble S.

$\mathcal{C}(A, B) = m(A, B) - d(A, B)$, le critère de Condorcet pour deux individus A et B

- $m(A, B)$: nombre de variables ayant la même valeur pour A et B
- $d(A, B)$: nombre de variables ayant des valeurs différentes pour A et B

On commence la constitution des classes en affectant chaque individu à la classe S pour laquelle $\mathcal{C}(A, S)$ est maximum et supérieur à 0.

On réalise plusieurs itérations jusqu'à ce que :

- Le nombre d'itérations maximal soit atteint.

- Le critère de Condorcet ne s'améliore plus suffisamment d'une itération à la suivante.

Conclusion

Avantages

- Complexité des méthodes linéaires, le temps d'exécution étant proportionnel au nombre d'individu.
- Les algorithmes de réallocation améliorent continuement la qualité des classes.

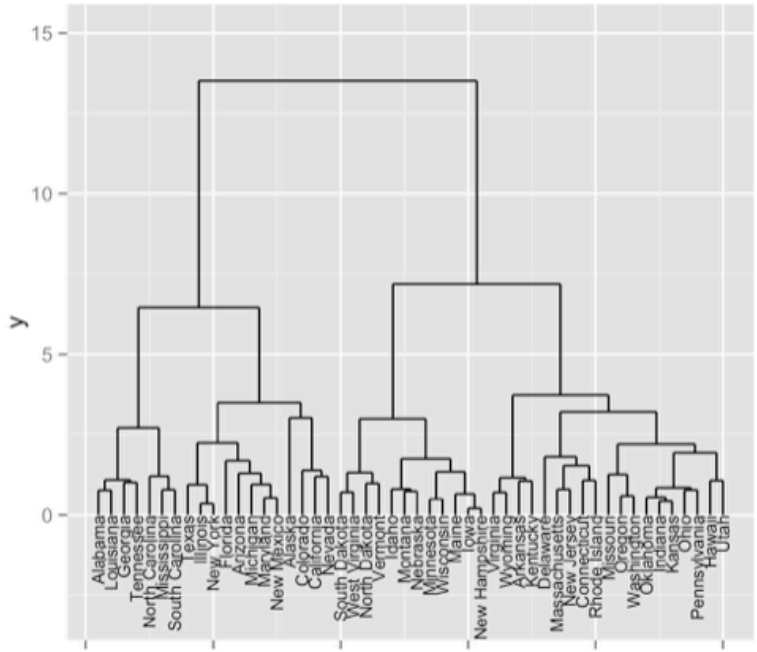
Inconvenients

- La partition finale dépend énormément de l'initialisation des centres.
- Le nombre de classes est fixé au départ

Sommaire

introduction

Les techniques de classifications hiérarchiques produisent des classes emboitées que l'on visualise graphiquement sous forme d'arbre hiérarchique indicé.



On part des objets élémentaires, qu'on agrège sur la base d'une notion de distance entres les objets et d'un critère de ressemblance entre les classes.

A chaque étape on cherche une partition de l'ensemble des individus en regroupant ceux les plus proches.

L'hétérogénéité des classes de la CAH augmente avec la taille des classes.

Stratégies et méthodes d'agrégation

- **Méthode du saut minimum** : Méthode très sensible à "l'effet de chaine" : on se retrouve assez souvent avec un groupe démesurément gros et plusieurs petits groupes satellites

- **Méthode de la distance maximale** : Très sensible aux valeurs hors normes, et de ce fait peu utilisée.

- **Méthode de Ward** :
  - L'indice de dissimilarité entre deux classes est la perte d'inertie interclasse résultant de leur regroupement. Il s'agit de l'écart de Ward
  - Sont agrègés les individus faisant le moins varier l'inertie intraclasse.
  - Soient deux groupes d'individus A et B d'effectif respectif n_A et n_B et de centre de gravité respectif g_A et g_B. Le centre de gravité du regroupement est $g_{AB} = \frac{n_A g_A + n_B g_B}{n_A + n_B}$

Algorithme

On munit l'ensemble des éléments d'une mesure de ressemblance, on construit une matrice des écarts entre les éléments pris deux à deux $\mathcal{M}_d^n$

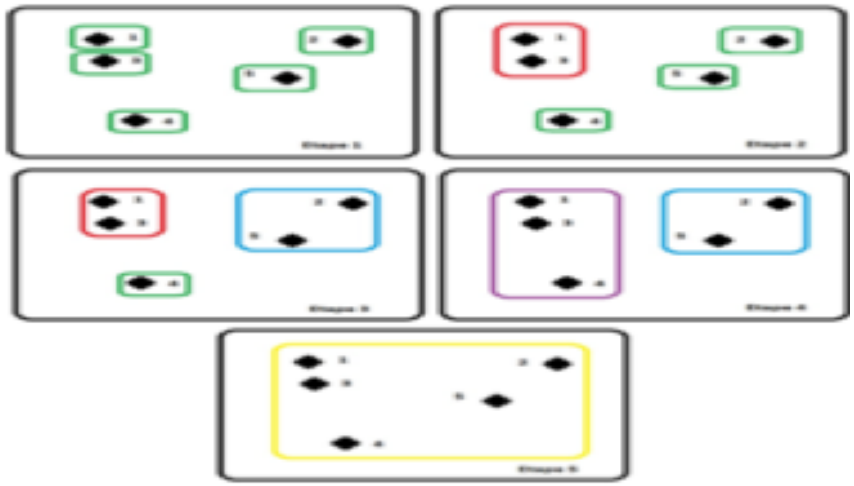
On agrège en un nouvel ensemble les deux éléments les plus proches.

On met à jour la nouvelle matrice de ressemblance entre le nouvel élément formé et les $n-2$ restants $\mathcal{M}_d^{n-1}$

Algorithme

On recherche à nouveau les deux éléments les plus proches que l'on agrège.

On procède ainsi de façon itérative jusqu'à l'obtention d'une classe unique.



Avantages et inconvenients

avantages

- Pas de fixation du nombre de classes à priori
- Permet de classer des individus, des variables et des centres de classes
- Pas de dépendances aux centres initiaux

inconvenients

- On obtient différents résultats suivant les choix des paramètres (distances, choix d'agrégation).
- Les calculs sont lourds dès lors qu'on a un nombre important de données.

Exemple

Construire une CAH des individus suivants, en utilisant la distance euclidienne et la stratégie d'agrégation de Ward.

	X_1	X_2
ω_1	2	2
ω_2	7.5	4
ω_3	3	3
ω_4	0.5	5
ω_5	6	4
ω_6	15	10

Description

Combine efficacement les avantages des méthodes par partitionnement et de le CAH.

Elle consiste à effectuer une première classification sur les n observations par les k-means, en fixant le nombre de classes de manière à limiter le risque de fusion des classes naturelles.

Puis on effectue une CAH sur les centres de ces classes.

La CAH est suivie d'une optimisation en effectuant une classification des centres mobiles sur les centres des classes de la CAH.

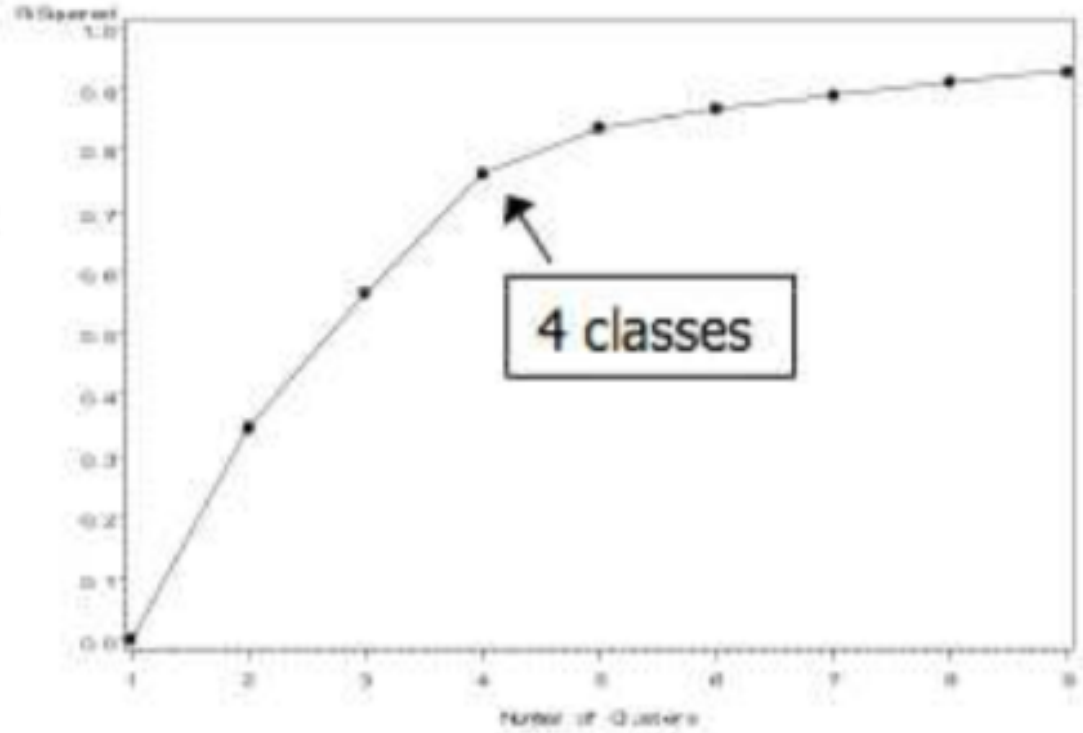
Sommaire

Choix du nombre de classes

R^2 (RSQ)

C'est la proportion d'inertie expliquée par la classification, plus elle est proche de 1, meilleure est la classification.

Le nombre de classes à retenir correspond au dernier saut le plus important observé lors d'agrégation de deux classes.



Choix du nombre de classes

Pseudo F

Mesure statistique évaluant la séparation entre toutes les classes

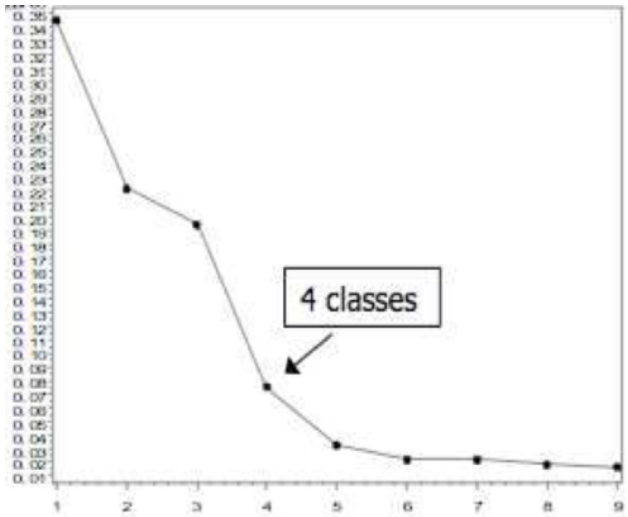
Permet de comparer l'homogénéité entre une parition en k classes et une partition en $k-1$ classes.

$$Pseudo - F = \frac{\frac{R^2}{k-1}}{\frac{1-R^2}{n-k}}$$

Choix du nombre de classes

Semi-Partial R-Square (SPRSQ)

Mesure la perte d'homogénéité lors de l'agrégation de deux classes.
Le nombre de classes correspondant au point où est observé une forte baisse de SPRSQ.



Choix du nombre de classes

Semi-Partial R-Square (SPRSQ)

Le nombre de classes finalement retenu sera choisi de manière à observer une valeur élevée de RSQ, suivi d'un saut important de SPRSQ à l'agrégation suivante.

Choix du nombre de classes

Silhouette

Pour tout individu $\omega_i \in \Gamma$, on note A sa classe d'appartenance.

On définit :

- $a(\omega_i) = \frac{1}{card(A)-1} \sum_{\omega_i \in A, i \neq i'} dis(\omega_i, \omega'_i)$: La moyenne des dissimilitudes entre ω_i et les autres individus du groupe

- $b(\omega_i) = \min_{\mathcal{C} \neq A} \left(\frac{1}{card(\mathcal{C})} \sum_{\omega'_i \in \mathcal{C}} dis(\omega_i, \omega'_i) \right)$: La moyenne des dissimilitudes entre ω_i et les individus du groupe le plus proches

Choix du nombre de classes

Silhouette

On appelle **largeur de la silhouette** associée à ω_i, la quantité

$$s(\omega_i) = \frac{b(\omega_i) - a(\omega_i)}{\max(b(\omega_i), a(\omega_i))},\ s(\omega_i) \in [-1, 1]$$

- $s(\omega_i) = 1$: ω_i est bien classé
- $s(\omega_i) \simeq 0$: ω_i se situe entre son groupe d'affectation, et le groupe le plus proche identifié
- $s(\omega_i) = -1$: ω_i est mal classé

Choix du nombre de classes

Silhouette

La largeur de la silhouette d'une classe correspond à la moyenne des largeurs des silhouettes des individus qui la compose(notée S_r).

On appelle indice de qualité d'une partition à k classes, la moyene globale des largeurs des silhouettes des différentes classes $\mathcal{C}_r, r = 1, \ldots, k$:

$$S(k) = \frac{1}{n} \sum_{r=1}^{k} n_r . S_r$$

On retient le nombre de classe k qui maximise l'indice de qualité.

Choix du nombre de classes

Comparaison de deux partitions

Indice de Rand

Elle mesure la concordance entre deux partitions

Soient deux partitions $\mathcal{P}_1, \mathcal{P}_2$

On définit :

- a : Le nombre d'individus se trouvant dans la même pour les deux partitions
- b : Le nombre d'individus se trouvant dans une même classe de $\mathcal{P}_1$, maus dans deux classes différentes de de $\mathcal{P}_2$
- c : Le nombre d'individus se trouvant dans une même classe de $\mathcal{P}_2$, maus dans deux classes différentes de de $\mathcal{P}_1$
- d : Le nombre d'individus se trouvant dans deux classes pour les deux partitions

$$R(\mathcal{P}_1, \mathcal{P}_2) = \frac{a+d}{a+b+c+d}$$

interprétation

Analyse unidimensionnelle

Sont essentiellement utilisées les variables illustratives,celles-ci pouvant être nominales ou continues.

Pour une variable X donnée, on compare sa moyenne(fréquence) dans la classe avec celle de l'ensemble des individus de la population.

On fait l'hypothèse :

H_0 : Les individus constituant la classes sont tirés au hasard et sans remise de la population totale.

interprétation

Analyse unidimensionnelle

Dans le cas de Variables nominales :

Soit N la variable aléatoire correspondant au le nombre d'individus de la classe, présentant la modalité j de la variable X.
Sous H_0 : $N \backsim$ Hypergéométrique :

- de moyenne $E - j(N) = n_k \frac{n_j}{n}$
- de variance $S_k^2 = n_k * \frac{n-n_k}{n-1} \frac{n_j}{n}(1 - \frac{n_j}{n})$

Lorsque les effectifs sont élevés, on construit la statistique :
$t_k(N) = \frac{N-E_k(N)}{S_k(N)} \backsim \mathcal{N}(0,1)$

On a $p_k(j) = \mathbb{P}(|Z| > t_k(N))$

Plus $p_k(j)$ est faible, plus on rejette H_0

interprétation

Analyse unidimensionnelle

Dans le cas de Variables continues : la valeur-test

Il s'agit d'une statistique permettant de classer les variables lors d'une caractérisation des classes.

On se place sous les mêmes hypothèses que ci-dessus :

On definit la valeur-test $V_k(X) = \frac{\bar{X}_k - X_k}{\sigma_k(X)}$

Au plus elle sera grande en valeur absolue, au plus on rejette H_0, la variable caractérise la classe.

interprétation

Analyse multidimentionelle

On determine les variables qui caractérisent le mieux les classes constituées.

Sont utilisées à la fois les variables actives et illustratives.

Différentes approches sont envisageables :

- Réaliser une Analyse factorielle représentant les classes obtenues et les variables initiales.
- Construire un arbre de décision avec comme variable cible la classe de risque obtenue.
- Effectuer une classification des variables initiales en y incluant les indicatrices des classes obtenues