TD
MODELISATION NON-SUPERVISEE
CLASSIFICATIONS AUTOMATIQUES

4 : Modélisation non-supervisée - 1 : la classification 2

Généralités sur la classification.....2

Principes 2
Exemple marketing 2
Vocabulaire 2
Technique 2
Les méthodes de classifications 6

La méthode des K-moyennes et ses variantes.....7

Présentation 7
Vocabulaire et historique 7
Caractéristiques de l'algorithme des K-moyennes (paramètres en entrée) 7
Algorithme 7
Simulation de l'algorithme des K moyennes 8
Avantages et inconvénients 11
Les variantes 11

4 : MODELISATION NON-SUPERVISEE
- 1 : LA CLASSIFICATION

Généralités sur la classification

Principes

La classification est **la plus répandue des techniques descriptives**. Il existe de très nombreux algorithmes de classification.

L'objectif d'une classification est de **distinguer des sous-ensembles (ou classes) distincts** dans la population de départ.

Rappelons que **la classification se distingue du classement** par le fait que les critères de classification ne sont pas connus *a priori* (avant étude de la population). C'est la population qui détermine les critères.

La classification est le plus souvent un **préalable à d'autres opérations de data mining**.

La classification permet de **limiter le nombre de variables par sous-ensemble**. Les variables très discriminantes ou trop peu discriminantes peuvent être éliminées.

La classification permet de **rechercher des corrélations propres à chaque classe** et donc plus précises.

Attention : il n'existe pas une solution unique au problème de la classification. Autrement dit, **il n'y a pas « LA » bonne classification**, mais plusieurs classifications possibles.

Exemple marketing

L'intérêt de la classification en marketing est de définir les profils de client. Chaque classe « résume » une clientèle ce qui permet une **communication spécifique, « one-to-one »**.

Les classes permettent de se **constituer un panel représentatif** (un panel est échantillon permanent de personnes ou de classes de personnes que l'on interroge régulièrement sur différents sujets).

Vocabulaire

En marketing, on parle de **segmentation**, de **typologie**, d'**analyse typologique** ou de **clustering** (en anglais) à la place de classification.

On parle de **classe**, de **segment** ou de **cluster** pour parler tant de l'extension (les individus) que de l'intension (les variables et leurs valeurs possibles) des sous-ensembles définis par la classification.

On parle de **typologie** ou de **type** pour parler de l'intension (les variables et leurs valeurs possibles).

Technique

L'objectif de la présentation technique est de comprendre les grandes lignes du fonctionnement des techniques de classification et les paramétrages possibles.

- ***Nombre de classes possibles***

Le nombre de sous-ensembles possible d'un ensemble de n éléments est appelé « **nombre de Bell** » et est donné par la formule suivante :

$$B(n+1) = \text{Somme}(k=0 ; n)(Cnk * B(k))$$

Avec B(0)=1 et Cnk = n ! / (k ! * (n-k) !)

Les premières valeurs de la suite sont :

B(1)=1, B(2)=2, B(3)=5, b(4)=15, B(5)=52, B(6)=203, …

Assez vite, on arrive à des nombres énormes :

$$B(30)=84,7.10^{22}$$

On ne pourra donc pas tester tous les sous-ensembles possibles.

- ***Relations entre les sous-ensembles obtenus***

En général, on obtient des **sous-ensembles disjoints**. C'est le résultat de la plupart des techniques.

On peut aussi envisager le cas de sous-ensembles disjoints avec **certains sous-ensembles en incluant d'autres**. C'est le cas de certaines **techniques dites « mixtes ».**

Enfin, on peut aussi envisager le cas de **sous-ensembles non disjoints mais sans inclusion** : on parle alors d'**analyse floue (fluzzy)**. On n'abordera pas ce cas.

- ***Distance entre les individus***

La distance entre les individus sera donnée soit par une **métrique** particulière (par exemple la métrique euclidienne, la métrique de Manhattan qui prend les valeurs absolues plutôt que les carrés, la métrique économétrique, etc.), soit par une matrice des similarités (par exemple la matrice des corrélations). Cf. Analyses factorielles pour plus de détails.

- ***Préparation des données***

On a intérêt à **se séparer des individus hors norme**.

- ***Choix des variables de la classification***

Pour **distinguer les bonnes classes**, il est souvent nécessaire de faire plusieurs « passes », en supprimant ou en ajoutant des variables à la méthode de classification.

Cela concerne les variables trop peu discriminantes.

- ***Nombre de classes***

Les **techniques sans *a priori*** laissent l'algorithme déterminer le nombre optimum de classes.

Les **techniques avec *a priori*** obligent à fixer *a priori* le nombre de classes attendues.

On peut commencer par utiliser des techniques sans *a priori* puis utiliser les résultats dans les techniques avec *a priori*.

On peut toujours imposer moins de classes qu'il n'y en a sans *a priori*. C'est utile d'un point de vue pratique si le nombre de classes trouvé sans *a priori* est trop élevé pour être opérationnel en pratique (un service commercial peut vouloir travailler sur 5 segments, et pas sur les 10 proposés par la classification sans *a priori*).

Par contre, imposer un **nombre de classes plus grand qu'il n'y en a sans *a priori*** risque de conduire à des **résultats arbitraires**.

- ***Choix d'une classification : techniques empiriques***

Pour valider une classification on peut utiliser plusieurs méthodes complémentaires :

- **L'analyse sémantique des caractéristiques des variables dans chaque classe.** On peut ainsi, intuitivement, trouver une signification à chaque classe. Cette analyse sémantique rejoint la question du nombre de classes.
- **La projection des individus dans le plan factoriel d'une ACP** : on vérifiera ainsi que les individus de chaque classe sont visuellement regroupés ensemble et séparés des autres classes. ACP : Analyse en composantes principales. On abordera ce point avec la technique des composantes principales.
- **L'utilisation de la technique prédictive de l'arbre de décision** en prenant le numéro de la classe comme variable cible. On abordera ce point avec la technique des arbres de décision.

- ***Choix d'une classification : mesure des inerties inter- et intra- classes***

L'inertie mesure l'écart entre les individus d'une classe.

Il existe 3 mesures d'inertie pour une population :

- **L'inertie totale, I** est une mesure indépendante de toute division de la population en classes.
- **L'inertie inter-classes, Ir** et **l'inertie intra-classes, Ia** sont deux mesures fonctions d'une division de la population en classes :

➢ ***L'inertie totale : I***

C'est la moyenne des carrés des distances des individus au centre de gravité (barycentre).

$$I = \left(\sum_{i=1}^{n} [\, d(x_i, g)^2 \,] \right) / n$$

Avec :

n : nombre d'individus de la population

x : valeur d'un individu

g : valeur du centre de gravité de la population

L'intertie totale I tend vers 0 quand n tend vers 1.

- ***L'inertie intra-classe : Ia***

C'est la somme des inerties totales de chaque classe :

$$\mathbf{Ia = Somme(i=1 \text{ à } k)[\ Ii\ ]}$$

Avec :

k : nombre de classes de la population

I : inertie totale d'une classe

Ia tend vers I quand k tend vers 1. S'il n'y a qu'une classe, l'inertie intra-classe, c'est l'inertie totale.

Ia tend vers 0 quand k tend vers n. Puisqu'il y a un individu par classe, il n'y a aucun décalage entre deux individus d'une même classe.

- ***L'inertie inter-classe : Ir***

C'est la « moyenne pondérée » des carrés des distances des centre de gravité de chaque classe au centre de gravité de la population totale.

$$\mathbf{Ir = Somme(i=1 \text{ à } k)[\ pi * d\ (gi,\ g)^2\ ]}$$

Avec :

k : nombre de classes de la population initiale.

g : valeur du centre de gravité de la population initiale.

p : poids de la classe (égale au nombre d'individus de la classe divisé par le nombre d'individus de la population totale).

Ir tend vers I quand k tend vers n. S'il n'y a qu'un individu par classe, l'inertie inter-classe, c'est l'inertie totale.

Ir tend vers 0 quand k tend vers 1. S'il n'y a qu'une seule classe, il n'y a aucun décalage entre la population totale et la classe.

- ***Formule de Huygens***

$$\mathbf{I = Ia + Ir}$$

Ia tend vers 0 et Ir tend vers I quand le nombre de classes tend vers le nombre d'individus, autrement dit quand chaque individu devient une classe.

Ia tend vers I et Ir tend vers 0 quand le nombre de classe tend vers 1, autrement dit quand il n'y a plus de classes.

- ***Inertie et valeur d'une classification***

Intuitivement, on est tenté de dire qu'une bonne classification est celle qui réduit l'inertie intra-classe (pour chaque classe, les individus sont bien regroupés) et qui augmentent la distance inter-classe (les classes, prises globalement, sont bien séparées).

Toutefois, la formule de Huygens montre que cette approche intuitive ne résiste pas à l'analyse mathématique : l'approche intuitive conduit finalement à ne plus avoir de classes.

Bien que Ir tende vers I quand le nombre de classes tend vers le nombre d'individus, on constate toutefois une inflexion dans la courbe d'accroissement de Ir. Le point d'inflexion peut être considéré comme correspondant au nombre de classes optimum.

Les techniques mesurant ce nombre de classes à partir de l'inflexion dans l'évolution de Ir parlent de R^2, R squared, RSQ, Cubic Clustering Criterion (CCC), pseudo F, R^2 semi-partiel (SPRSQ) et pseudo t^2 (PST2) pour aborder cette caractéristique.

Les méthodes de classifications

On peut distinguer 5 grands types de méthodes de classifications :

- Par partitionnement
- Hiérarchique
- Mixte
- Neuronale
- Relationnelle

La méthode des K-moyennes et ses variantes

Présentation

L'algorithme des K-moyennes est un algorithme qui permet de trouver des classes dans des données.

C'est un algorithme « non hiérarchique » : les classes qu'il construit n'entretiennent jamais de relations hiérarchiques : une classe n'est jamais incluse dans une autre classe.

C'est un algorithme très utilisé.

Vocabulaire et historique

La méthode est due à Forgy (1965).

Elle devient méthode des K-means de Mac Queen (1967).

La méthode des nuées dynamiques de Diday (1971) est une nouvelle version.

En français, on parle de méthode des K-moyennes.

On parle aussi de méthode des centres mobiles.

Caractéristiques de l'algorithme des K-moyennes (paramètres en entrée)

L'algorithme fonctionne en précisant le nombre de classes attendues.

L'algorithme calcule la distance intra-classe. Il travaille donc sur des variables continues.

Algorithme

Début

K ← Nombre de classes attendues ;

Initialiser la valeur du centre de ces classes avec la valeur de K enregistrements choisis aléatoirement.

IA ← + ∞ // Initialisation de l'inertie intra-classe. On pourrait aussi l'initialiser à la valeur de l'inertie totale.

Répéter

- IAec ← IA ; // On mémorise l'inertie intra-classe « en cours » : IAec
- Pour chaque enregistrement, le mettre dans la classe la plus proche.
- Calculer le nouveau centre de chacune des classes (barycentre).
- Calculer la nouvelle inertie intra-classe : IA

Tant IA < IAec

Fin

On s'arrête quand la nouvelle inertie intra-classe est plus grande ou égale à la précédente.

Simulation de l'algorithme des K moyennes

- ***Itération 1***

On part de 8 points. On choisit deux points : pt7 et pt8 et on en fait les centres. On calcule les distances de chaque point aux centres et on repère la distance la plus petite.

Classe	Axes	Centre 1	Centre 2	pt1	pt2	pt3	pt4	pt5	pt6	pt7	pt8
Totalité	X	1	2	1	3	4	5	1	4	1	2
	Y	1	1	3	3	3	3	2	2	1	1
			distance (pt, C1) :	2,00	2,83	3,61	4,47	1,00	3,16	0,00	1,00
			distance (pt, C2) :	2,24	2,24	2,83	3,61	1,41	2,24	1,00	0,00
			distance min	2,00				1,00		0,00	
			distance min		2,24	2,83	3,61		2,24		0,00

En fonction de la distance la plus petite, on classe les points dans une classe ou dans l'autre. On calcule le centre de gravité dans chaque classe.

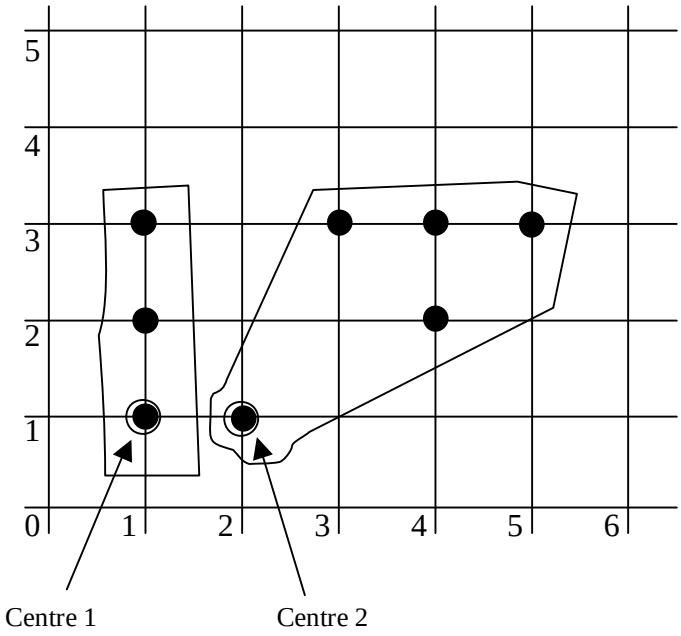
Classe	Axes	Centre 1	Centre 2	Pt1	pt2	pt3	pt4	pt5	pt6	pt7	pt8
Classe 1	X	1		1				1		1	
	Y	2		3				2		1	
Classe 2	X		3,6		3	4	5		4		2
	Y		2,4		3	3	3		2		1

On calcule enfin l'inertie totale pour chaque classe : on calcule d'abord la distance au carré de chaque point au centre de la classe, puis la moyenne de ces distance.

Enfin, on calcule l'inertie intra-classe : somme des inerties totales de chaque classe.

Classe			pt1	pt2	pt3	pt4	pt5	pt6	pt7	pt8
I 1	0,67	distance² (pt, C1) :	1,00				0,00		1,00	
I 2	1,68	distance² (pt, C2) :		0,72	0,52	2,32		0,32		4,52
I A	**2,35**									

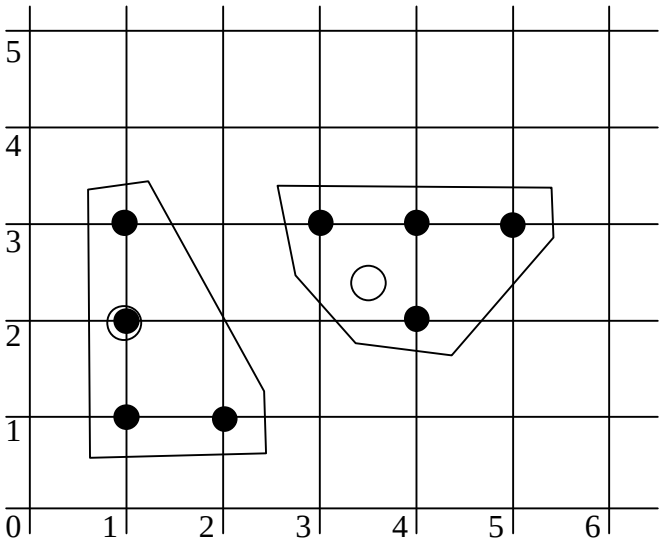
Représentation graphique :



- ***Itération 2***

Itération 2											
Classe	**Axes**	**Centre 1**	**Centre 2**	**pt1**	**pt2**	**pt3**	**pt4**	**pt5**	**pt6**	**pt7**	**pt8**
Totalité	X	1	3,6	1	3	4	5	1	4	1	2
	Y	2	2,4	3	3	3	3	2	2	1	1
			distance (pt, C1) :	1,00	2,24	3,16	4,12	0,00	3,00	1,00	1,41
			distance (pt, C2) :	2,67	0,85	0,72	1,52	2,63	0,57	2,95	2,13
			distance min	1,00				0,00		1,00	1,41
			distance min		0,85	0,72	1,52		0,57		
Classe	**Axes**	**Centre 1**	**Centre 2**	**pt1**	**pt2**	**pt3**	**pt4**	**pt5**	**pt6**	**pt7**	**pt8**
Classe 1	X	1,25		1				1		1	2
	Y	1,75		3				2		1	1
Classe 2	X		4		3	4	5		4		
	Y		2,75		3	3	3		2		
IA 1	0,88	distance² (pt, C1) :		1,63				0,13		0,63	1,13
IA 2	0,69	distance² (pt, C2) :			1,06	0,06	1,06		0,56		
IA	**1,56**										

Représentation graphique :



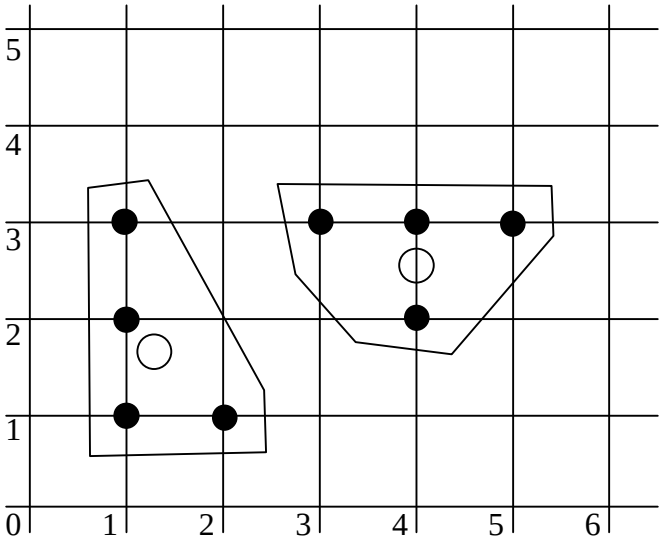
- ***Itération 3***

Itération 3											
Classe	**Axes**	**Centre 1**	**Centre 2**	**pt1**	**pt2**	**pt3**	**pt4**	**pt5**	**pt6**	**pt7**	**pt8**
Totalité	X	1,25	4	1	3	4	5	1	4	1	2
	Y	1,75	2,75	3	3	3	3	2	2	1	1
			distance (pt, C1) :	1,27	2,15	3,02	3,95	0,35	2,76	0,79	1,06
			distance (pt, C2) :	3,01	1,03	0,25	1,03	3,09	0,75	3,47	2,66
			distance min	1,27				0,35		0,79	1,06
			distance min		1,03	0,25	1,03		0,75		
Classe	**Axes**	**Centre 1**	**Centre 2**	**pt1**	**pt2**	**pt3**	**pt4**	**pt5**	**pt6**	**pt7**	**pt8**
Classe 1	X	1,25		1				1		1	2
	Y	1,75		3				2		1	1
Classe 2	X		4		3	4	5		4		
	Y		2,75		3	3	3		2		
IA 1	0,88	distance² (pt, C1) :		1,63				0,13		0,63	1,13
IA 2	0,69	distance² (pt, C2) :			1,06	0,06	1,06		0,56		
IA	**1,56**										

Fichier K moyennes.xls

La dernière itération aboutit à une inertie intra-classes identique à la précédent (1,56).

Représentation graphique :



Avantages et inconvénients

- ***Remaraues sur les individus isolés : hors nome. outliers***

Les individus isolés (hors norme, outliers) vont toujours constituer une classe. La méthode permet donc de mettre au jour ces individus.

- ***Il faut ensuite les éliminer et relancer la méthode pour mettre au iour les classes.Avantaaes***

Le temps de calcul est rapide : il est fonction de N (nombre d'individus dans la population de départ).

- ***Inconvénients***

Le nombre de classes est un paramètre fourni en entrée.

La répartition finale des classes est fonction des premiers centres choisis.

La méthode est surtout bien adaptée aux classes sphériques.

Les variantes

Les variantes sont de deux types :

- ***Optimisation***

Ce sont des variantes algorithmiques qui permettent une accélération des traitements.

- ***Traitement des variables catéaorielles***

L'algorithme de la méthode des K-moyennes ne traite que des variables continues (du fait du calcul des centres de gravité, de la distance et de l'inertie).

Des variantes de cette méthode permettent de prendre en compte les variables catégorielles :

- Méthode de k-medoids (PAM - 1990, CLARA - 1990, CLARANS – 1994).
- Méthode des k-modes (Huang, 1998)

L'algorithme des k-medoids recourt aux médianes (médoïdes) plutôt qu'aux moyennes (centroïdes), d'où la possibilité de travailler sur des variables catégorielles.

Ces algorithmes donnent des résultats de meilleure qualité que les précédents, mais ils sont aussi beaucoup plus coûteux en temps d'exécution.