Analyse des données et liens avec les systèmes décisionnels

Ahlame Douzal
Ahlame.Douzal@imag.fr
(2016)

1. INTRODUCTION AU DATA MINING

Qu'est-ce que le data mining
Émergence du data mining
Problématiques du data mining et applications
Les grandes étapes d'un projet en data mining

2. LA PREPARATION DES DONNEES EN DATA MINING

Les différents types de données
Transformation des données
Mesures de similarités et codages

3. LES PRINCIPALES TECHNIQUES DE DATA MINING

Description : techniques descriptives et exploratoires
Structuration : techniques de classification et de classement
Explication : techniques prédictives (arbres de décision et segmentation)
Association : techniques pour l'extraction de règles d'association
(Analyse du panier de la ménagère)

4. ANALYSES DESCRIPTIVES ET EXPLORATOIRES

Indicateurs de résumé et de synthèse
Analyse factorielle

5. CLASSIFICATION AUTOMATIQUE

Classification par partitionnement
Classification hiérarchique

6. SEGMENTATION PAR ARBRE

Construction d'un arbre de décision
Critères de sélection du meilleur sous arbre
Les règles d'affectation
Estimation du risque d'erreur

Introduction au data mining (DM)

Objectif : Exploration et analyse de données volumineuses afin d'extraire des connaissances cachées pour prédire et agir.

Connaissance : liens, règles, objets similaires, groupes, association,…

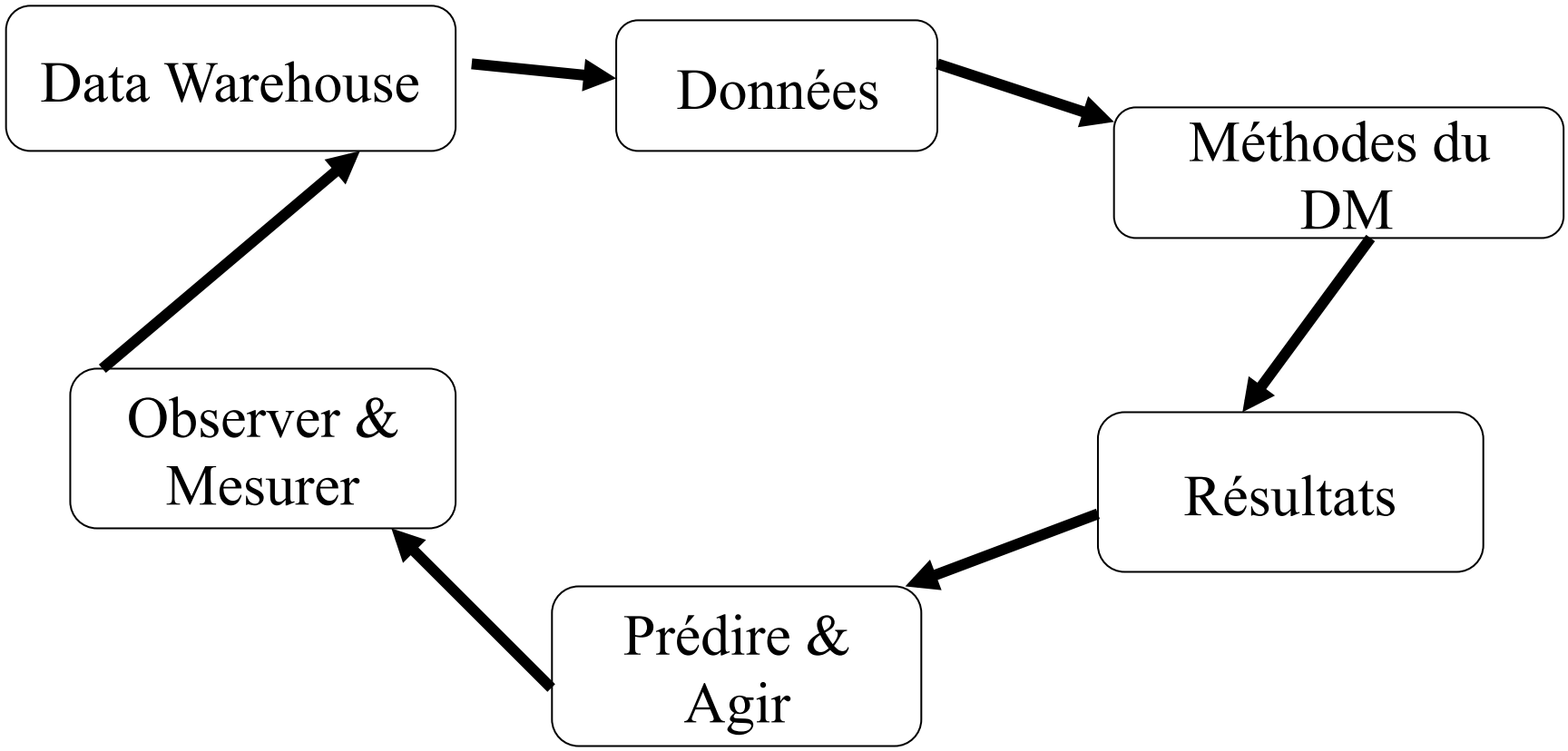
Facteurs d'émergence du DM :

- la production massive des données.
- de grandes capacités de stockage.
- de puissants processeurs.
- un contexte très concurrentiel.
- la disponibilité de logiciels de DM.

Domaines d'application et problématiques

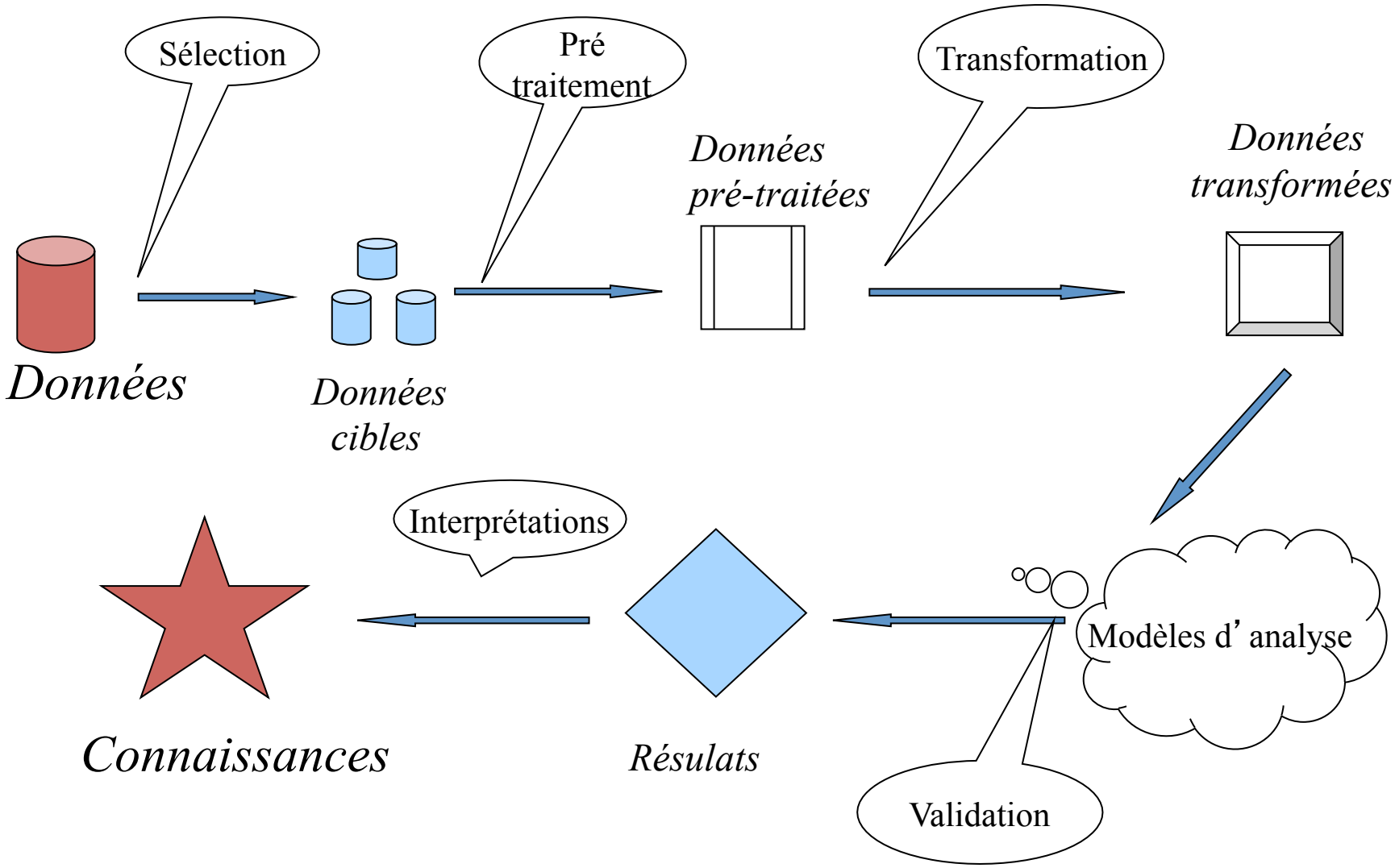
Secteur Industriel	Problématique décisionnelle
Grande Distribution	• Analyse des comportements des consommateurs. • Recherche des similarités des consommateurs en fonction des critères géographiques. • Prédiction des taux de réponse en Marketing direct
Laboratoires pharmaceutiques	• Identification des meilleures thérapies pour différentes maladies • Optimisation des plans d'action des visiteurs médicaux pour le lancement de nouveaux produits
Banques	• Recherche de formes d'utilisation de cartes caractéristiques d'une fraude. • Modélisations prédictives des clients partants.
Assurances	• Analyse des sinistres • Recherche des critères explicatifs du risque ou de la fraude
Aéronautique, automobile	• Prévision des ventes • Dépouillement d'enquête de satisfaction
Télécommunication, eau et énergie	• Détection des formes de consommation frauduleuses • Classification des clients selon la forme d'utilisation des services • Prévision du départ des clients

Positionnement du DW et du DM



Les grandes étapes d'un projet en data mining

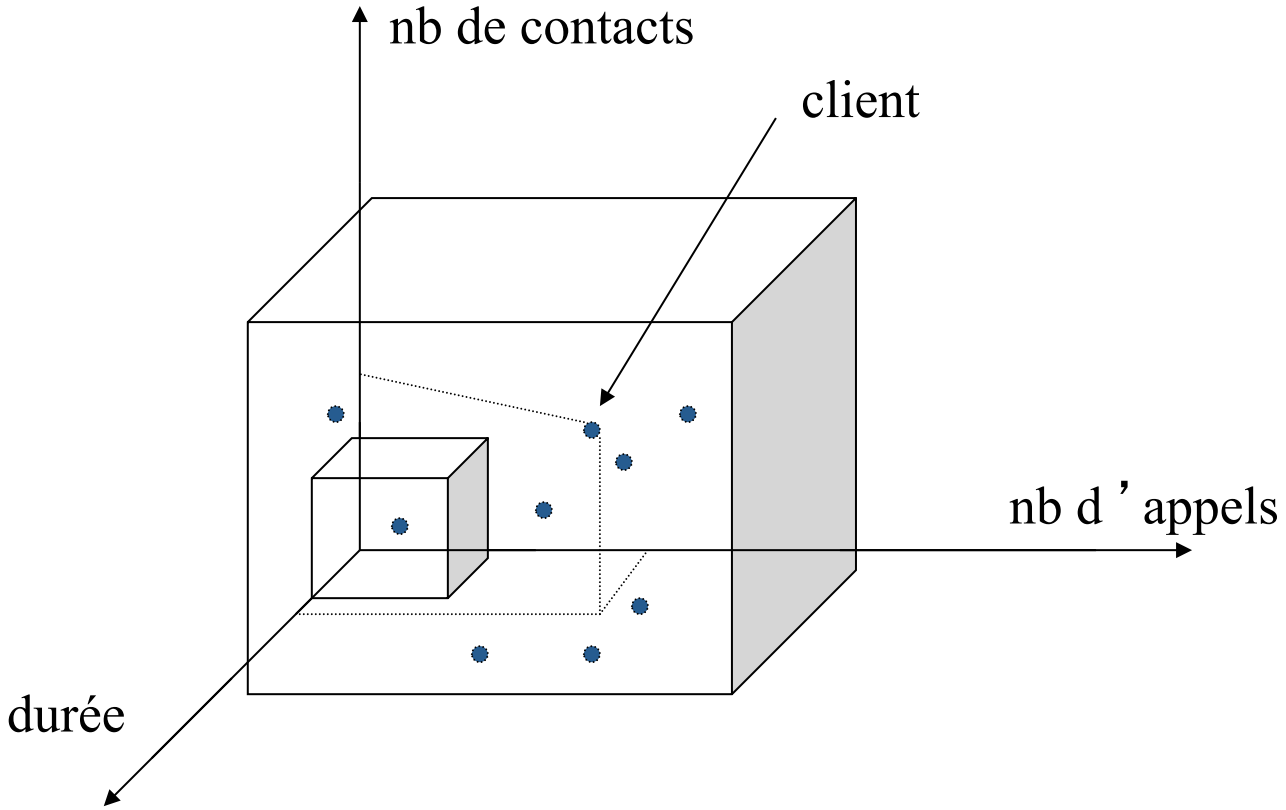
- Analyse des objectifs
- La construction de la base d'analyse
- La préparation des données
  - La normalisation, le codage, la gestion des données aberrantes et manquantes, …
- Le choix des modèles d'analyse
- L'analyse et l'interprétation des résultats
- La validation des résultats
- La prédiction


Sélection
Pré traitement
Transformation
Données pré-traitées
Données transformées
Données
Données cibles
Interprétations
Modèles d'analyse
Connaissances
Résulats
Validation

Systèmes décisionnels

- Les moteurs des bases de données (Oracle, Informix, SqlServer, Ingres,…) pour le stockage et la structuration
- Les outils de requêtes (Business Object, Brio query, GQL, etc.) pour le reporting et l'interrogation des données.
- Les outils OLAP (SAS MDDB, Oracle Express, Pilot de Compshare, Cognos,…) pour l'analyse multidimensionnelle
- Les outils du data mining pour l'extraction de connaissances cachées dans les données.

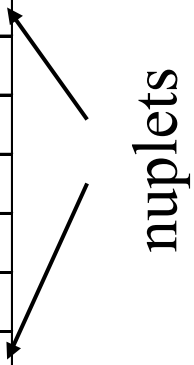
Exemple de données multidimensionnelles en télécommunication



Description de l'activité mensuelle des clients

Attributs / Dimensions

Clients	Nb Appel	Durée Com	Nb Appel Entrant	Courrier Vocal	Nb Contact
C1	35	500	41	Non	16
C2	9	170	25	Non	13
C3	7	210	45	Oui	3
C4	12	220	5	Non	17
C5	31	580	39	Non	19
C6	11	180	30	Oui	5
C7	11	110	10	Oui	20
C8	40	600	50	Oui	12



Quelques problématiques …

- **Constituer des groupes de profils de consommation similaires**
  Classification , Analyse Factorielle

- **Extraire les attributs caractérisant au mieux ces groupes**
  Analyse des corrélations, analyse factorielle

- **Analyser les liens entre attributs (variables)**
  Analyse des corrélations, Analyse d'associations

- **Identifier le groupe d'appartenance d'un nouveau client**
  Classement, Classification

- **Extraire des règles de décision portant sur le bon ou mauvais potentiel d'un client**
  Segmentation, Arbre de décision

- **Prédire le comportement d'un client**
  Réseaux neuronaux, régression, …

La préparation des données en data mining

- Les différents types de données

- Transformation des données
  - La normalisation :
    - Moyenne, variance covariance et corrélations
    - …
- Mesures de similarités et codages des données

Les différents type de données

Structure	Continu	Dénombrable	Cardinal
=, #		CSP	Nominal
<= , >=	Age Température	Rang ressemblance	Ordinal
<=, >=, +, *	Revenu		Mesurable
	Quantitatif	Qualitatif	Attribut

Tableau quantitatif (W , X)

∀j :{1..p} Xj quantitatif

Tableau qualitatif (W , X)

j :{1..p} Xj qualitatif

Tableau contingence (W , X)

xij est la fréquence d'apparition de la modalité xj
pour l'individu wi

Tableau de préférence (W , X)

xij exprime le degré de préférence de la modalité Xj par wi

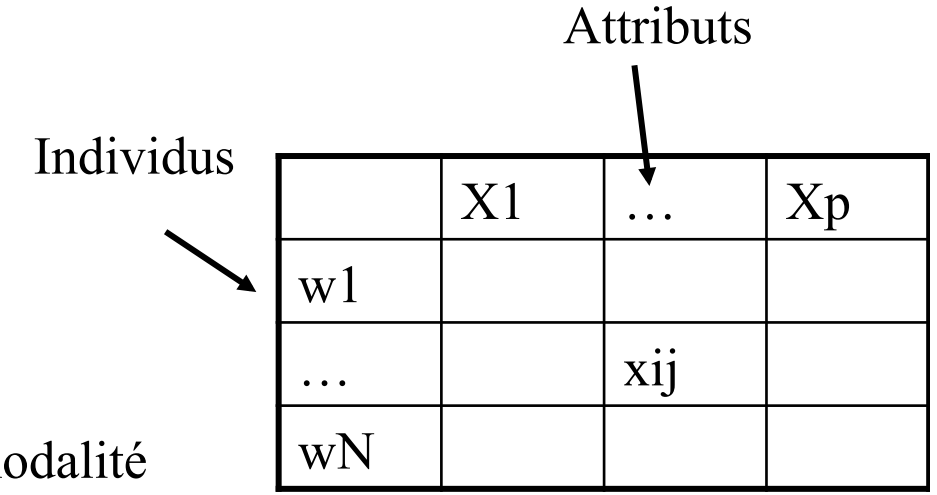
Tableau binaire (W , X)

xij :{0 ,1} exprime la présence ou pas de la modalité Xj pour wi

Tableau de proximité (W, W)

xij exprime une mesure de similarité ou de dissimilarité entre deux individus wi, wj

Tableau hétérogène



Indicateurs de position et de dispersion d'attribut quantitatif

- Indicateur de Position
  - Moyenne de Xj

$$\overline{Xj} = \frac{1}{N}\sum_{i=1}^{N} xij$$

- Indicateur de dispersion
  - Variance

$$\text{var}(Xj) = \frac{1}{N}\sum_{i=1}^{N}(xij - \overline{Xj})^2$$

  - Covariance

$$\text{cov}(Xj, Xk) = \frac{1}{N}\sum_{i=1}^{N}(xij - \overline{Xj})(xik - \overline{Xk})$$

  - Corrélation

$$cor(Xj, Xk) = \frac{\text{cov}(Xj, Xk)}{\sqrt{\text{var}(Xj) * \text{var}(Xk)}}$$

La normalisation des données quantitatives

Le centrage d'un attribut quantitatif **Xj = (x1j,…,xij,… xNj)**

xij $\longrightarrow$ $xij^c = xij - \overline{Xj}$

La réduction d'un attribut quantitatif **Xj = (x1j,…,xij,… xNj)**

xij $\longrightarrow$ $xij^n = xij / \sqrt{\text{var}(Xj)}$

La normalisation d'un attribut quantitatif Xj = (x1j,…,xij,… xNj)

xij $\longrightarrow$ $xij^* = \dfrac{(xij - \overline{Xj})}{\sqrt{\text{var}(Xj)}}$

Les mesures standards de proximités entre n-uplets

1- Tableaux quantitatifs

$$wi = (xi1, \ldots xip)$$

Distance de Minkowski

$$d(wi, ws) = \left(\sum_{j=1}^{p} |xij - xsj|^{r} \right)^{1/r}$$

r=1 (distance de Manhattan) $$d(wi, ws) = \sum_{j=1}^{p} |xij - xsj|$$

r=2 (distance euclidienne) $$d(wi, ws) = \sqrt{\sum_{j=1}^{p} (xij - xsj)^2}$$

r~> ∞(distance Chebychev) $$d(wi, ws) = \max_{1 \le j \le p} (|xij - xsj|)$$

2- Tableaux binaires

$wi = (xi1,...xip)$ xij=0/1 $ws = (xs1,...xsp)$ xsj=0/1

Codage des données binaires

wi → ws	1	0
1	a	b
0	c	d

- a : nombre d'occurrence où xij=1 et xsj=1
- b : nombre d'occurrence où xij=0 et xsj=1
- c : nombre d'occurrence où xij=1 et xsj=0
- d : nombre d'occurrence où xij=0 et xsj=0

Mesures de dissemblance usuelles

Russel et Rao	$d(wi,ws) = 1 - \frac{a}{a+b+c+d}$
Jaccard	$d(wi,ws) = 1 - \frac{a}{a+b+c}$
Dice	$d(wi,ws) = 1 - \frac{2a}{2a+b+c}$
Sokal & Sneath	$d(wi,ws) = 1 - \frac{a}{a+2(b+c)}$
Roger & Tanimoto	$d(wi,ws) = 1 - \frac{a+d}{a+d+2(b+c)}$
Kulzinsky	$d(wi,ws) = 1 - \frac{a}{b+c}$
Yule	$d(wi,ws) = 1 - \frac{ad-bc}{ad+bc}$

3- Tableau qualitatif nominal et/ou ordinal

- On procède au codage des attributs qualitatifs en attributs binaires
- Applications des mesures de similarités vues précédemment

	Couleur	Forme
wi	Rouge	Ellipsoïde
ws	Jaune	Circulaire

- Codage binaire

	Rouge	Jaune	Bleue	Ellipsoïde	Circulaire
wi	1	0	0	1	0
ws	0	1	0	0	1

Les mesures standards de proximités entre attributs

1- Attributs quantitatifs

$$D(Xj,Xk)=cor(Xj,Xk)$$

2- Attributs qualitatifs nominaux

(voir similarités entre vecteur binaire)

3- Attributs qualitatifs ordinaux

Le coefficient de corrélation des **rangs de Kendall**

…

Le coefficient de corrélation des rangs de Kendall

Xj(x1j, …xNj) Xk(x1k, … , xNk)

On procède au codage des attributs Xj en Yj et Xk en Yk :

$$Yj : \Omega x \Omega \rightarrow \{-1,0,1\}$$

$$\left\{\begin{array}{l} Yj(wi,ws) = -1 \quad si \quad xij \prec xik \\ Yj(wi,ws) = 0 \quad si \quad xij = xik \\ Yj(wi,ws) = 1 \quad si \quad xij \succ xik \end{array}\right\}$$

$$\tau(Xj,Xk) = cor(Yj,Yk)$$

- **Exemple**

	X1	X2
1	a	e
2	c	f
3	b	e
4	a	g

Codage

	Y1	Y2
(1,2)	-1	-1
(1,3)	-1	0
(1,4)	0	-1
(2,3)	+1	1
(2,4)	1	-1
(3,4)	1	-1

a<b<c et e<f<g

$$\tau(X1, X2) = cor(Y1, Y2)$$

Techniques descriptives et exploratoires

L'analyse descriptive

Objectif :

Résumer, synthétiser, structurer un ensemble d'informations en utilisant des représentations graphiques ou indicateurs numériques

Moyens :

Fonctions d'agrégation

(moyenne, moyenne mobile, ratio, cumul,…)

Outils graphiques

Exemples de mesures descriptives

Générer le rapport donnant la fréquence des appels clients, la durée total de communication par mois.

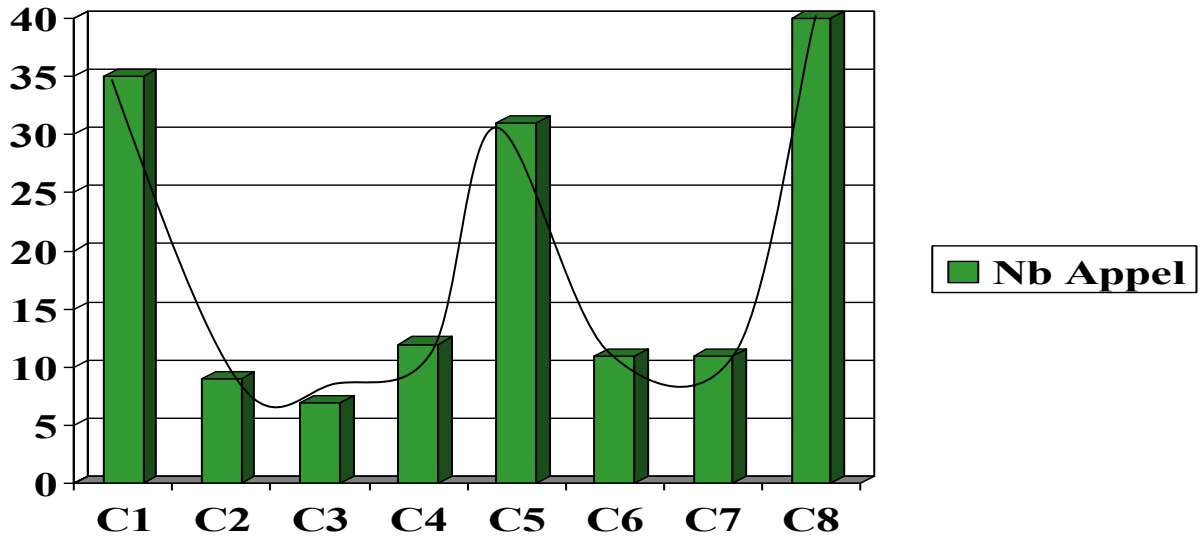
Donner la moyenne mobile des trois derniers mois des nombres d'appels entrants.

Mesurer la variabilité de la durée de connexion des clients sur l'année

Exemple de mesures descriptives

Clients	C1	C2	C3	C4	C5	C6	C7	C8	Tot
Nb Appel	35	9	7	12	31	11	11	40	156
Fréquence	0.22	0.05	0.04	0.07	0.19	0.07	0.07	0.25	

Fréquence :
Moyenne :
Variance :



L'analyse exploratoire

- ***Objectif :***

  Extraire des propriétés à un ensemble de nuplets.

  Étendre (inférer) ces propriétés à la base de données.

  Valider ou infirmer ces propriétés à l'aide de tests d'hypothèses.

Moyens :

Mesure de similarité, de distance entre nuplets.

Mesure de corrélations, de liens entre descripteurs

Outils graphiques

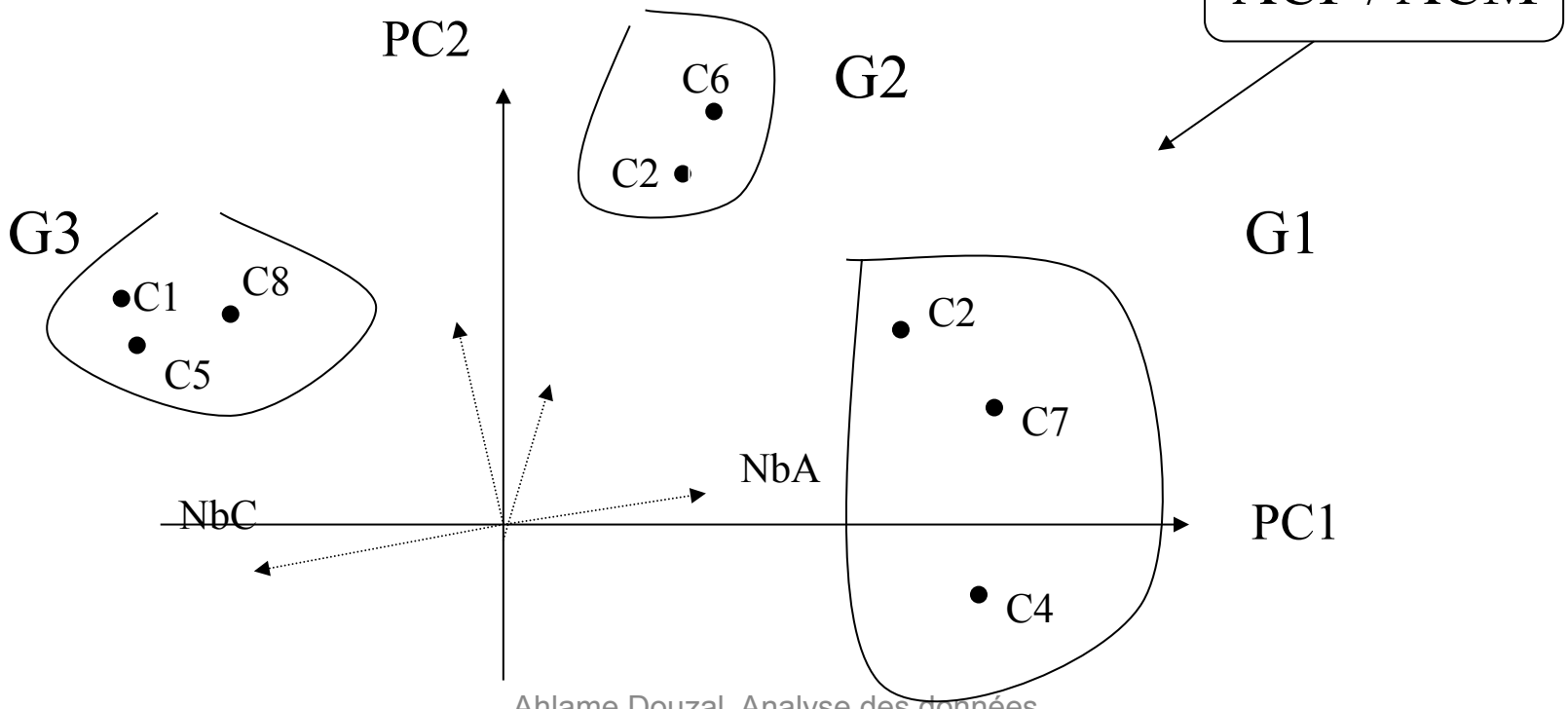
Quelques méthodes de l'analyse exploratoire

- Analyse factorielle
  - Objectif :
    - extrait les attributs pertinents
    - fournit des représentations graphiques des individus
    - mesure les liens entre descripteurs
    - prépare les données à une éventuelle classification.
  - Méthodes :
    - Analyse en composantes principales (ACP) : données continues
    - Analyse des correspondances multiples (ACM) : données continues et nominales.
    - …

L'analyse factorielle

Clients	Nb Appel	Durée Com	Nb Appel Entrant	Courrier Vocal	Nb Contact
C1	35	500	41	Non	16
C2	9	170	25	Non	13
C3	7	210	45	Oui	3
C4	12	220	5	Non	17
C5	31	580	39	Non	19
C6	11	180	30	Oui	5
C7	11	110	10	Oui	20
C8	40	600	50	Oui	12

ACP / ACM



Techniques classificatoires

Objectif :

Fournit une représentation graphique

Extrait des groupes d'individus similaires

Méthodes

Par partitionnement (Centres-mobiles) : les groupes forment

une partition

Hiérarchique : les groupes peuvent se recouvrir par

inclusion

Méthode des Centres-mobiles

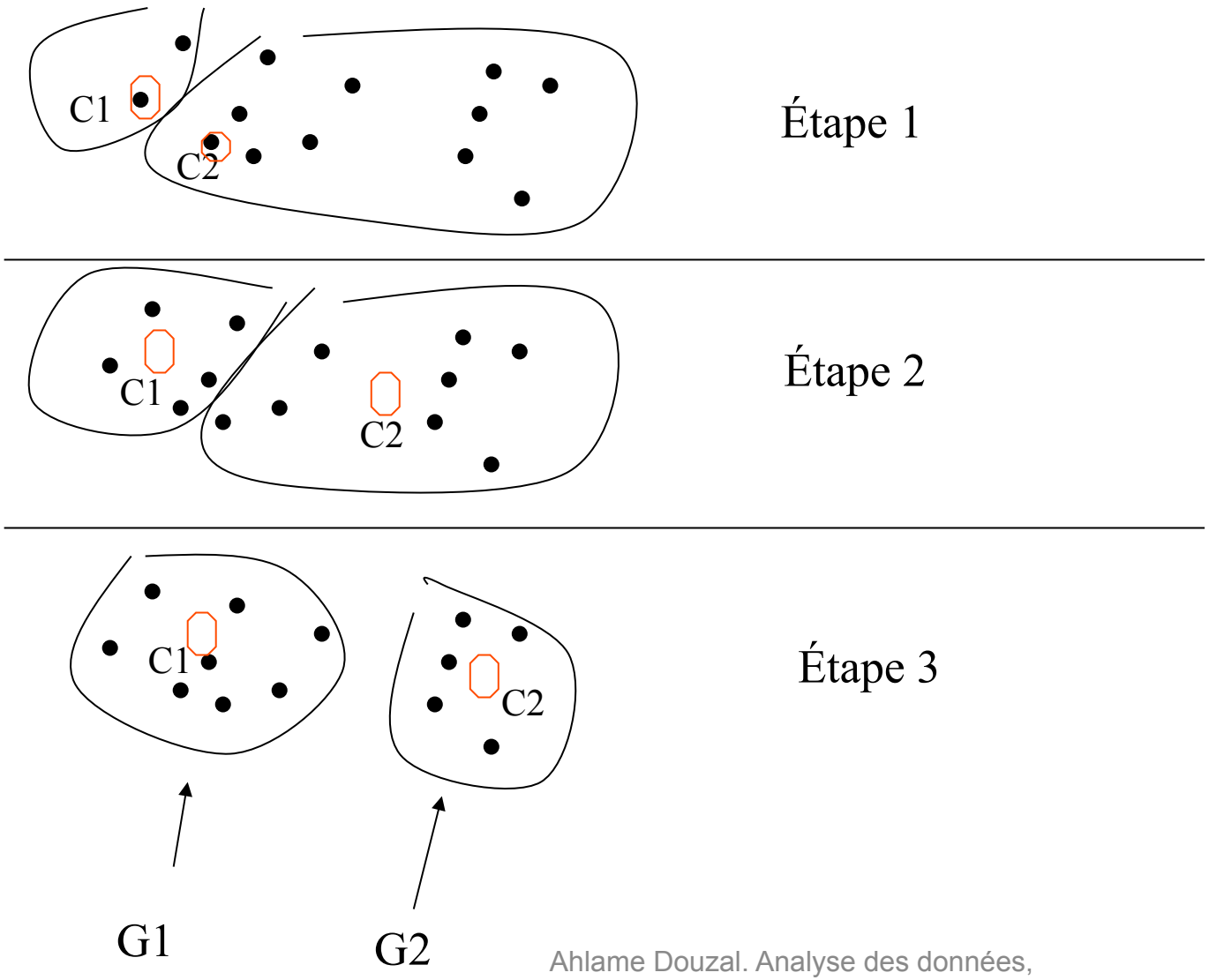
Objectif :

Construire une partition de r classes.

Algorithme :

1) Choisir le nombre de classe r
2) Choisir r nuplets comme centres des r classes
3) Affecter chaque nuplet au centre le plus proche.
4) Recalculer le centre de la classe d'affectation.
5) Répéter les étapes 3) et 4) jusqu'à stabilisation.

Centres-mobiles



Exemple

nuplets	A1 Esthétique du Package	A2 Mémorisation Accroche publicitaire
P1	1	1
P2	1	2
P3	4	3
P4	4	5
P5	2	2

Nombre de classe = 2

P1, P5 comme centres des deux classes

Centres-mobiles

Étape 1 :

Classe 1 : P1, P2

Classe 2 : P4, P5, P3

Nouveaux nuplets centres :

Classe 1 : P12 (1, 1.5)

Classe 2 : P12 (3.33, 3.33)

Étape 2 :

Classe 1 : P1, P2, P5

Classe 2 : P4, P3

Classification Hiérarchique

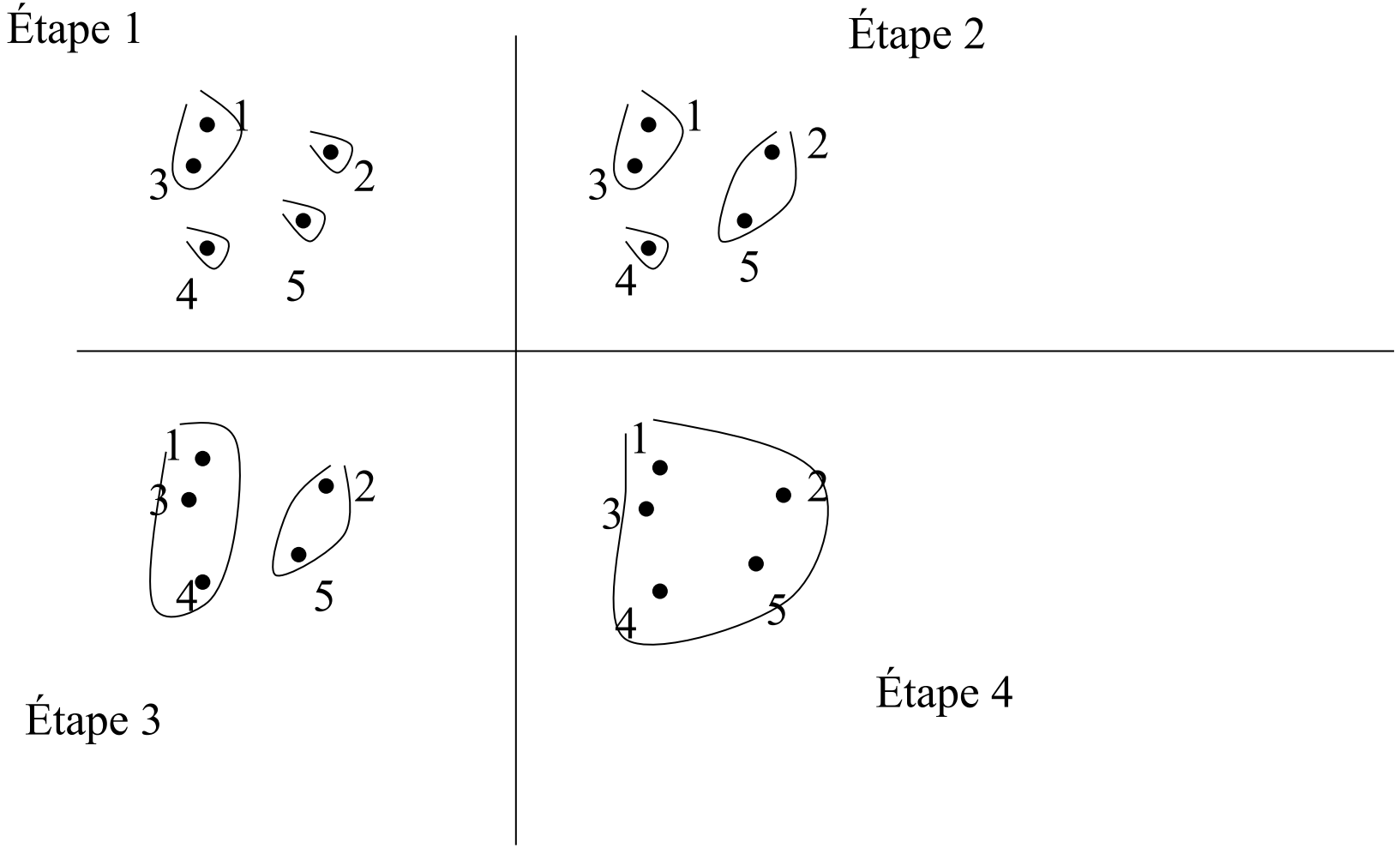
Objectif :

Construire une succession de partitions à p classes, p-1classes,…, 1 classe.

Algorithme :

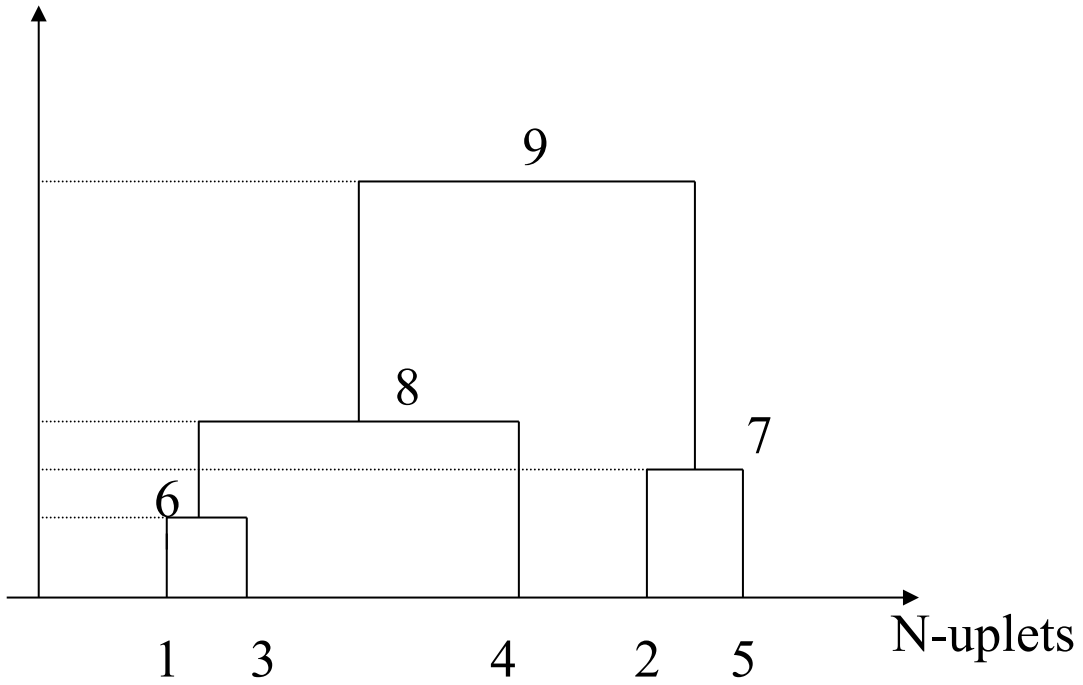
1) Soit p nuplets à classer
2) On construit la matrice des distances entre les p nuplets (p classes).
3) On agrège en un nouvel nuplet les deux nuplets les plus proches (p-1 classes)
4) on réitère les étapes 2 et 3 jusqu'à ce qu'il n' y ait plus qu'une classe.

Classification hiérarchique



Classification hiérarchique

- Distance



Exemple

nuplets	A1 Esthétique du Package	A2 Mémorisation Accroche publicitaire
P1	1	1
P2	1	2
P3	4	3
P4	4	5
P5	2	2

Étape 1

P1 et P2 agrégés en P12

Distances	P1	P2	P3	P4	P5
P1	**0**	**1**	**3.6**	**5**	**1.41**
P2		**0**	**3.16**	**4.24**	**1**
P3			**0**	**2**	**2.23**
P4				**0**	**3.6**
P5					**0**

Étape 2

nuplets	A1 Esthétique du Package	A2 Mémorisation Accroche publicitaire
P12	1	1.5
P3	4	3
P4	4	5
P5	2	2

P12 et P5 agrégés en P125

Distances	P12	P3	P4	P5
P12		**3.35**	**4.6**	**1.11**
P3		**0**	**2**	**2.23**
P4			**0**	**3.6**
P5				**0**

Étape 3

nuplets	A1 Esthétique du Package	A2 Mémorisation Accroche publicitaire
P125	1.5	1.75
P3	4	3
P4	4	5

P3 et P4 agrégés en P34

Distances	P125	P3	P4
P125		**2.74**	**4.1**
P3		**0**	**2**
P4			**0**

Arbres de décision

$$R(A1,\ldots,Ap,B)$$

A1,…An : attributs explicatifs

B : attribut à expliquer

$$Ai \Rightarrow_p B$$

si je connais Ai alors je connais B avec une probabilité p ($p \in [0,1]$)

Arbres de décision

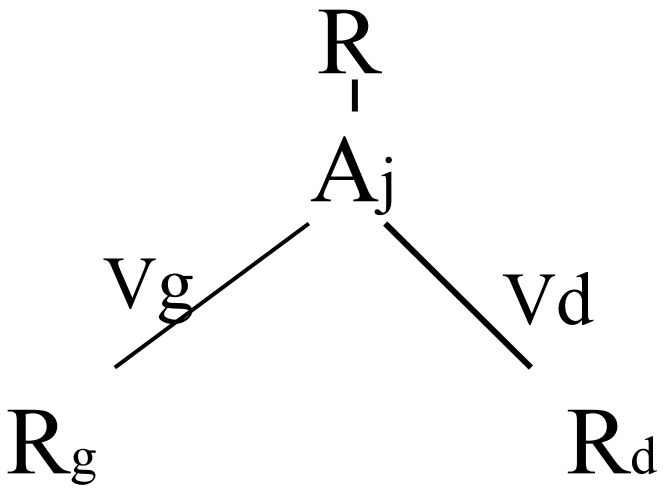
- Objectif

Extraire les attributs les plus discriminants

($A_i \Rightarrow_p B$ avec p fort)

Extraire des règles de décision (Identification)

$A_i \wedge A_j \wedge A_k \Rightarrow_p B$



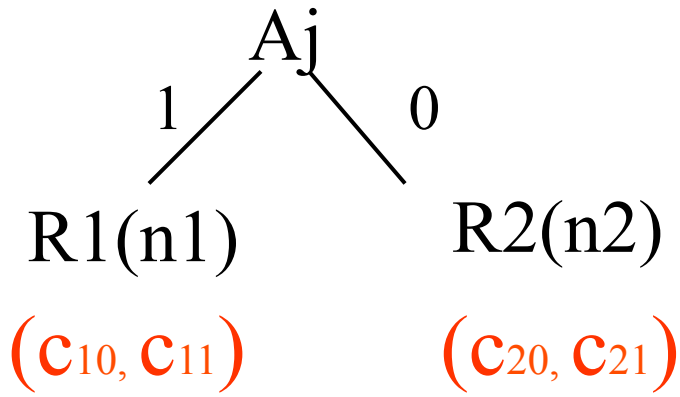
Partitionnement binaire

Type Attribut	Vg	Vd
Qualitatif binaire (a1,a2)	Aj =a1	Aj =a2
Qualitatif ordonné (a1, a2, a3) a1 ≤ a2 ≤ a3	Aj = a1	Aj ≥ a2
	Aj ≤ a2	Aj = a3
Qualitatif non ordonné (a1, a2, a3)	Aj = a1	Aj = a2 ou Aj = a3
	Aj = a1 ou Aj = a2	Aj = a3
	Aj = a1 ou Aj = a3	Aj = a2
Quantitatif (a)	Aj ≤ a	Aj > a

Mesure d'impureté

Soit Aj et B deux attributs binaires

ni : nb de nuplets dans Ri



c_{10} : proportion des nuplets *n1* prenant la modalité *0* de B
c_{11} : proportion des nuplets *n1* prenant la modalité *1* de B

$$I(Aj) = \sum_i c_{i0} * c_{i1}$$

Algorithme de construction de l'arbre de décision

Algorithme

a) Pour chaque attribut explicatif, on partitionne l'ensemble des nuplets puis on calcule le degré d'impureté associé à cette partition.

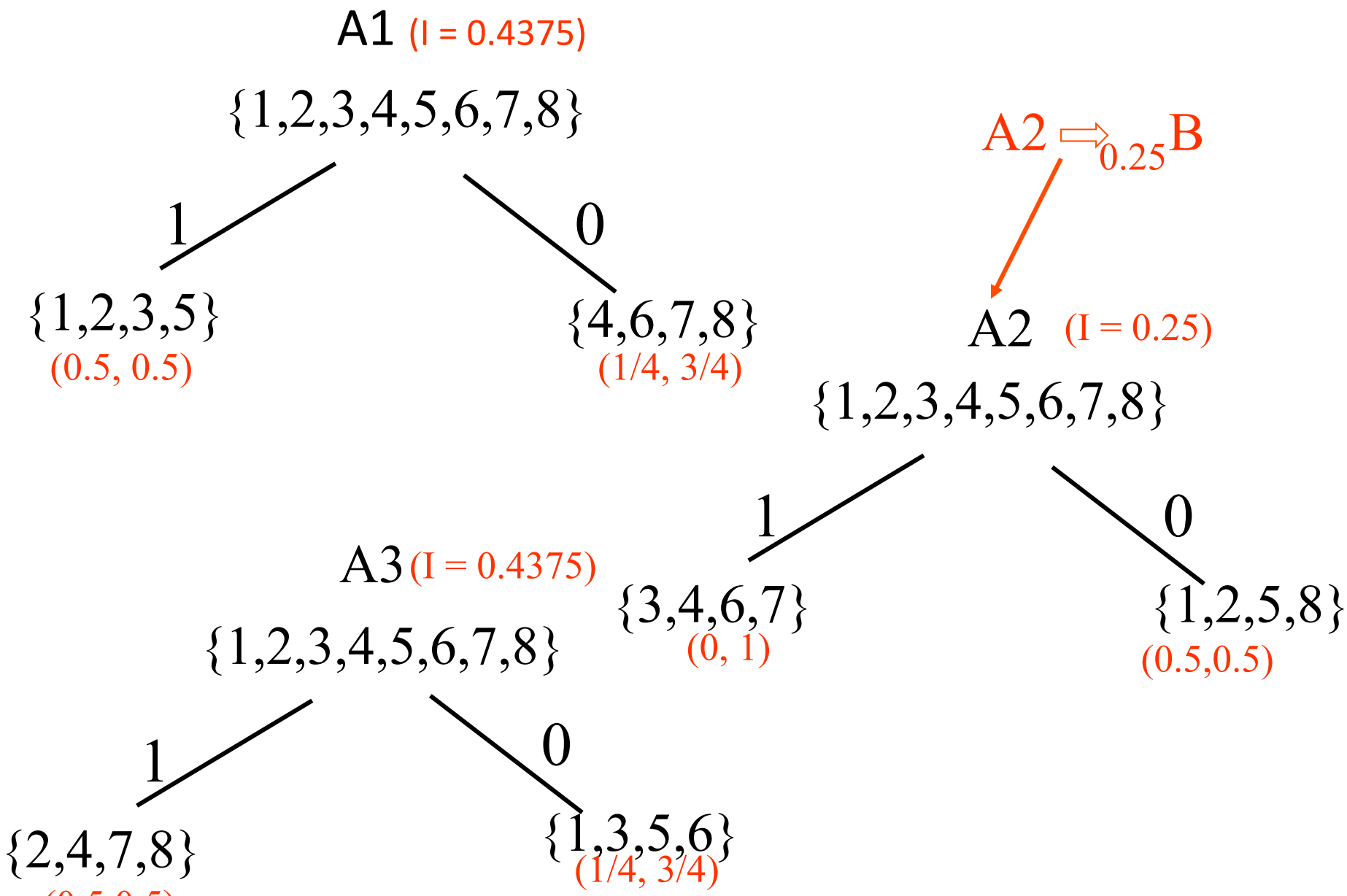
b) On choisit comme premier attribut de partitionnement celui donnant le degré d'impureté le moins élevé.

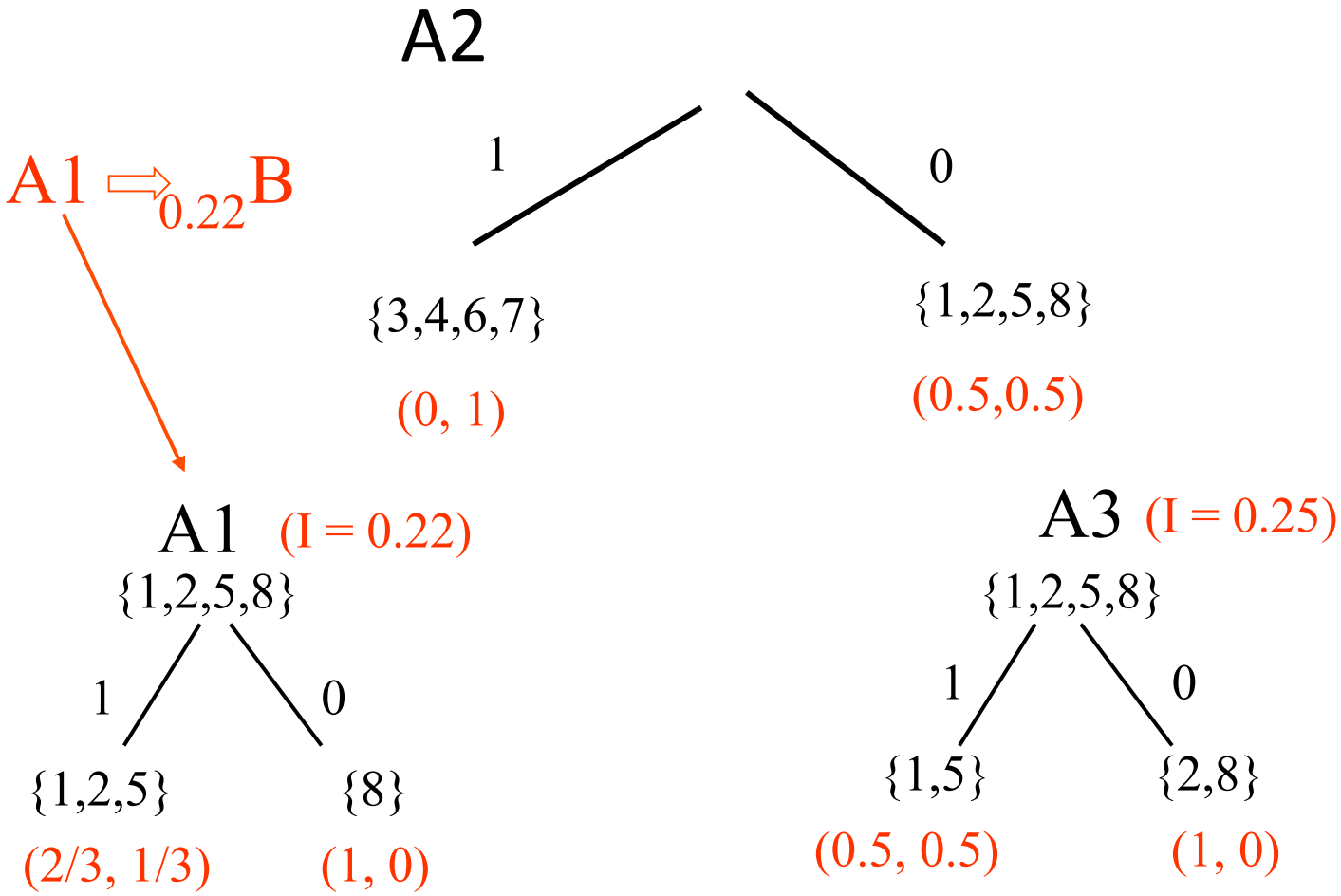
c) Pour chaque attribut explicatif restant, on réitère a) et b) pour segmenter chacune des parties obtenues. On s'arrête quand la partie contient un nuplet ou qu'on a atteint le degré d'impureté 0.

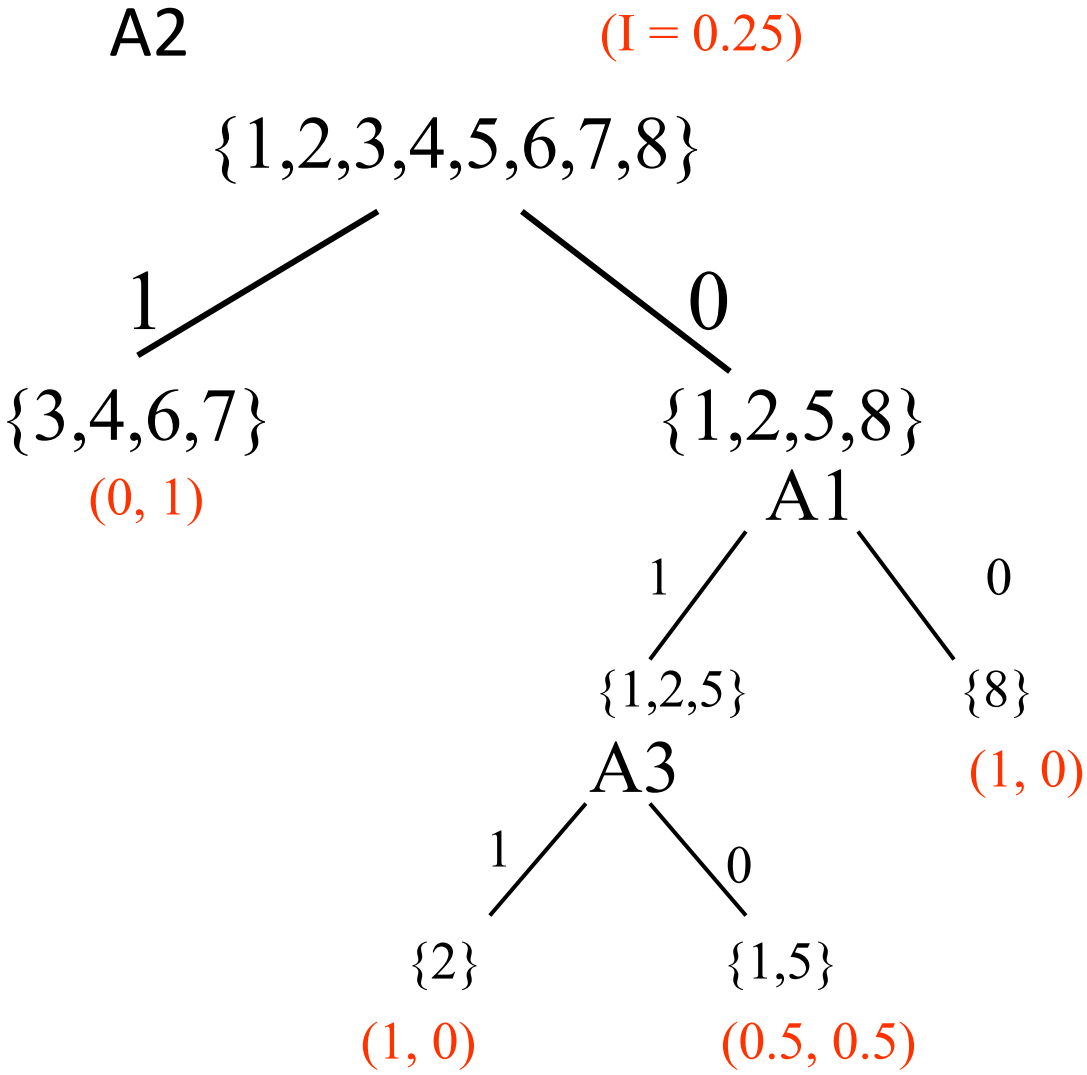
Exemple de construction d'un arbre binaire

explicatifs à expliquer

nuplets	A1 A eu un stagiaire	A2 A embauché un étudiant	A3 Connaît l'école	A4 Rendez-vous
1	1	0	0	0
2	1	0	1	0
3	1	1	0	1
4	0	1	1	1
5	1	0	0	1
6	0	1	0	1
7	0	1	1	1
8	0	0	1	0


A1 (I = 0.4375)
{1,2,3,4,5,6,7,8}
1
0
{1,2,3,5}
(0.5, 0.5)
{4,6,7,8}
(1/4, 3/4)
A2 ⇨0.25 B
A2 (I = 0.25)
{1,2,3,4,5,6,7,8}
1
0
{3,4,6,7}
(0, 1)
{1,2,5,8}
(0.5,0.5)
A3 (I = 0.4375)
{1,2,3,4,5,6,7,8}
1
0
{2,4,7,8}
(0.5,0.5)
{1,3,5,6}
(1/4, 3/4)


A2
1
0
{3,4,6,7}
{1,2,5,8}
(0, 1)
(0.5,0.5)
A1 ⇨0.22 B
A1 (I = 0.22)
{1,2,5,8}
1
0
{1,2,5}
{8}
(2/3, 1/3)
(1, 0)
A3 (I = 0.25)
{1,2,5,8}
1
0
{1,5}
{2,8}
(0.5, 0.5)
(1, 0)


A2
(I = 0.25)
{1,2,3,4,5,6,7,8}
1
0
{3,4,6,7}
(0, 1)
{1,2,5,8}
A1
1
0
{1,2,5}
{8}
A3
(1, 0)
1
0
{2}
{1,5}
(1, 0)
(0.5, 0.5)

Extraction de règles

$$(A_2 = 1) \quad \Rightarrow_1 \quad B = 1$$

Si l'entreprise a embauché un étudiant
alors obtention d'un rendez-vous

$$(A_2 = 0 \text{ et } A_1 = 0) \quad \Rightarrow_1 \quad B = 0$$

$$(A_2 = 0 \text{ et } A_1 = 1 \text{ et } A_3 = 1) \quad \Rightarrow_1 \quad B = 0$$

$$(A_2 = 0 \text{ et } A_1 = 1 \text{ et } A_3 = 0) \quad \Rightarrow_{0.5} \quad B = 1$$

Extraction de règles d'association

	Farine	Sucre	Lait	Œuf	Chocolat	Thé
1	1	1	1	0	0	0
2	0	1	0	1	1	0
3	1	1	0	1	1	0
4	0	0	0	1	1	1

Objectif :

- Extraire les associations du type Ai=1⇒Aj=1 (noté Ai⇒Aj)

  farine ⇒ sucre, chocolat ⇒ thé

- Évaluer la fiabilité des associations extraites

- **Le degré de confiance d'une association**

$$\textbf{Conf}(Ai \Rightarrow Aj) = \frac{Occ(Ai \Rightarrow Aj)}{Occ(Ai)}$$

Conf : le degré de confiance

Occ(Ai) : le nombre d'occurrences dans la table où apparaît la modalité Ai

- **Le degré de support d'une association**

$$\textbf{Sup}(Ai \Rightarrow Aj) = \frac{Occ(Ai \Rightarrow Aj)}{N}$$

N : le nombre de nuplets

Algorithme

a) On calcule le poids de chaque attribut. On élimine les attributs dont le poids est inférieur à un certain seuil de confiance.

b) On calcule le poids de chaque couple d'attribut (Ai=1,Aj=1). On élimine les couples dont le poids est inférieur à un certain seuil de confiance.

c) Sur la base des couples retenus on construit toutes les associations possibles.

d) Pour chaque association on évalue ses degrés de confiance et de support.

Processus du data mining

- Poser le problème
- La recherche des données
- La sélection des données pertinentes
- Le « nettoyage » des données
- Les actions sur les variables
- La recherche d'un modèle
- L'évaluation du résultat
- L'intégration de la connaissance

Poser le problème

- La formulation du problème

  problème de diagnostic de panne

  analyse des défauts de production …

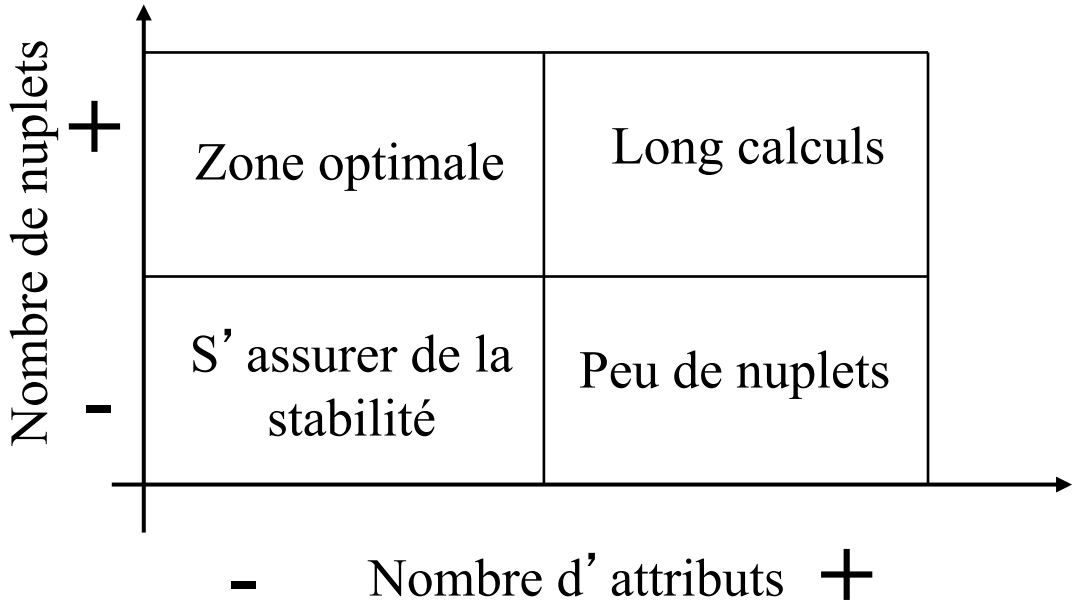
- La typologie du problème

  Exploratoire ? Inférentielle ?

- Les résultats attendus et utilisations

La recherche des données

- Investigation et détermination de la structure générale des données
- La réduction des dimensions (corrélations)



La sélection des données

- Échantillon ou exhaustivité

  choix dépend de l'infrastructure

  détection de tendances générales (échantillon représentatif)

  Exhaustivité : qualité des résultats, coûteux

- mode de création de l'échantillon

  taille : fonction des méthodes à appliquer

  tirage aléatoire à partir des différentes sous-population

Le nettoyage des données

- Gestion des valeurs aberrantes

  isolation des « pics » de distribution statistiques

- Gestion des valeurs manquantes

  exclure les nuplets incomplets

  remplacer les données manquantes

  …

Les actions sur les attributs

- La transformation mono-attribut
  - la normalisation des distributions
  - transformation des dates en durées
  - géocodage (intégrer les contraintes de proximités dans le raisonnement)
- La transformation multi-attribut
  - les ratios, les fréquences,
  - les tendances
  - les combinaison linéaires, ...

La recherche du modèle

- Choix de la base d'apprentissage et de la base de test (70% / 30%)
- Choix des algorithmes de calcul
  - modèle à base d'équations

    (Réseaux de neurones, techniques de régression)
  - analyse logique

    (Arbres de décisions, règles d'association, ensembles flous)
  - techniques de projection (mise en évidence des facteurs principaux d'explication)

    (analyse factorielles, classification, …)

L'évaluation du résultat

- Évaluation qualitative
  - restitution de la connaissance sous forme graphique

- Évaluation quantitative
  - Les intervalles de confiance
  - les tests de validation (étudier la stabilité des résultats dans la base test)

L'intégration de la connaissance

- Dresser un bilan
  - une faible qualité des données conduit à revoir les processus d'alimentation de l'entrepôt
  - la détection du fort pouvoir prédictif d'une donnée pousse à modifier le schéma de la base et le rythme d'alimentation
  - Les agrégats construits se révèlent être des dimensions intéressantes à intégrer dans le tableau de bord existant
  - explication des connaissances contradictoires avec l'existant.