

Fouille de Données et Apprentissage

C5 - Fouille de données complexes

Lina Soualmia

Université de Rouen

LITIS - Équipe TIBS-CISMeF

lina.soualmia@chu-rouen.fr

29 janvier 2016

Crédits

- P.Poncelet (Montpellier)
- S.Ullman (Stanford)
- R.Rakotomalala (Lyon)
- N.Pasquier (Nice)
- parmi d'autres ...

Plan

- Algorithme Close
- Motifs séquentiels
- Fouille du web, fouille de textes
- Retour sur la préparation des données

Algorithme Close

| R | a | b | c | d | e |
|-------|-----|-----|-----|-----|-----|
| x_1 | 1 | 0 | 1 | 1 | 0 |
| x_2 | 0 | 1 | 1 | 0 | 1 |
| x_3 | 1 | 1 | 1 | 0 | 1 |
| x_4 | 0 | 1 | 0 | 0 | 1 |
| x_5 | 1 | 1 | 1 | 0 | 1 |
| x_6 | 0 | 1 | 1 | 0 | 1 |

- $O = \{x_1, x_2, x_3, \dots, x_6\}$
- $P = \{a, b, c, d, e\}$
- $x R p$ si et seulement si la ligne de x et de p se croisent sur un 1 : $x_1 R a$

| R | <i>items</i> |
|-------|--------------|
| x_1 | acd |
| x_2 | bce |
| x_3 | abce |
| x_4 | be |
| x_5 | abce |
| x_6 | bce |

Table avec attribut multivalué

Connexion de Galois

- La connexion de Galois associée à une base de données formelle (O, P, R) est le couple de fonctions (f, g) définies par :
 - $f : 2^P \longrightarrow 2^O$

$$m \longmapsto f(m) = \{x \in O \mid x \text{ possède } m\}$$
 - $g : 2^O \longrightarrow 2^P$

$$X \longmapsto g(X) = \{p \in P \mid \forall x \in X, x R p\}$$
 - g est dit dual de f et f dual de g .
 - $f(m)$ est l'image du motif m .

- Beaucoup de règles : problème d'intelligibilité pour l'analyste
- ① Une stratégie d'élagage : définir de nouvelles mesures de plausibilité (autres que le support et la confiance) pour limiter cet ensemble de règles.
- ② Une deuxième stratégie : élaguer les règles déjà contenues dans la base de connaissances.
 - ▶ Si, par exemple, on dispose de l'ontologie du cours 1, et qu'on a extrait la règle tortue \rightarrow quadrupède, reptile, cette règle peut être supprimée de l'ensemble des règles valides puisqu'elle n'apporte pas de nouvelles connaissances.
- ③ Une troisième stratégie : constater que certaines règles sont moins informatives que d'autres.
 - ▶ Si on a les règles $R_1 = ab \rightarrow c$ et $R_2 = a \rightarrow cd$ avec le même support et la même confiance, alors, on pourra élaguer R_1 , qui est une conséquence de R_2 .

Quelles mesures utiliser ?

- Propriété des mesures

| Mesure | corrélation | null-invariant | fréq.Consequence | directionnelle |
|------------|-------------|----------------|------------------|----------------|
| confiance | N | O | N | O |
| lift | O | N | O | N |
| conviction | O | N | O | N |
| cosine | O | O | N | N |

Quelles mesures utiliser ?

- Chaque propriété précise l'interprétation
 - ▶ Directionnelle : distingue les liens $A \rightarrow C$ et $C \rightarrow A$
 - ▶ Corrélacion : validité statistique
 - ▶ Null-invariant : indépendamment des autres lignes
- Support nécessaire : taille de la population concernée
- Quelles mesures ?
 - ▶ Optimal : une mesure pour chaque propriété
 - ▶ Minimum : support, confiance, lift (ou conviction)

Éviter la redondance

- Ne garder que les règles les plus *informatives*
- Pour un motif donné, les règles valides les plus intéressantes sont à conclusions maximales (autrement dit : à prémisses minimales)
- Si on a une règle $R = p_1 \rightarrow p_2 \setminus p_1$ valide et p'_1 tel que $p_1 \subseteq p'_1 \subseteq p_2$, alors, avec $R' = p'_1 \rightarrow p_2 \setminus p'_1$, on a :
 - ▶ $support(R') = support(R) = support(p_2)$
 - ▶ $confiance(R') = \frac{support(p_2)}{support(p'_1)} \geq \frac{support(p_2)}{support(p_1)} = support(R)$
 - ▶ donc si R est valide, R' l'est aussi.

Éviter la redondance

Exemple :

- la règle $e \rightarrow abc$ est valide.
- On peut en déduire que les règles suivantes sont également valides :
 - ▶ $ea \rightarrow bc$, $eb \rightarrow ac$, $ec \rightarrow ab$, $abe \rightarrow c$, $ebc \rightarrow a$...etc.

Éviter la redondance

Algorithme Close

- repose sur l'extraction de **générateurs** d'itemsets **fermés** fréquents
- le nombre d'itemsets fermés fréquents est généralement bien inférieur au nombre d'itemsets fréquents

Connexion de Galois

- La connexion de Galois associée à une base de données formelle (O, P, R) est le couple de fonctions (f, g) définies par :
 - $f : 2^P \longrightarrow 2^O$

$$m \longmapsto f(m) = \{x \in O \mid x \text{ possède } m\}$$
 - $g : 2^O \longrightarrow 2^P$

$$X \longmapsto g(X) = \{p \in P \mid \forall x \in X, x R p\}$$
 - g est dit dual de f et f dual de g .
 - $f(m)$ est l'image du motif m .

Opérateur de fermeture

- La fermeture d'un itemset A est un itemset B tel que B apparaît dans les mêmes objets que A .
- Pour la calculer on utilise les deux fonctions :
 - ▶ f : qui associe à un itemset les objets qui le contiennent
 - ▶ g : qui associe à un ensemble d'objets les itemsets qu'ils ont en commun
- Soit A un itemset : $fermeture(A) = g \circ f(A)$

Algorithme Close

- 1 Initialiser l'ensemble des générateurs avec l'ensemble des singletons formés par les items
- 2 Calculer la fermeture des générateurs de niveau k et de leur support
- 3 Ajouter les fermetures des générateurs à l'ensemble des itemsets fermés fréquents
- 4 Construire des générateurs de niveau $k + 1$

Algorithme Close

- Les générateurs de niveau $k + 1$ sont obtenus de la même manière que dans l'algorithme Apriori, mais ceux appartenant à la fermeture d'un générateur de niveau k sont supprimés.

Génération des règles d'association par Close [02]

- L'ensemble des générateurs et de leurs fermés permettent de déduire une base générique de règles exactes ($\text{conf}=1$) :
 - ▶ par exemple : si le générateur est $\{a\}$ et que son fermé est $\{abc\}$ la règle exacte extraite est $a \rightarrow bc$
- L'ensemble des générateurs, des fermés et de leurs sur-ensembles fermés permettent de déduire une base de règles approximatives ($\text{conf}<1$) :
 - ▶ si le générateur est $\{a\}$, son fermé $\{abc\}$ et le sur-ensemble fermé $\{abcd\}$ la règle approximative extraite est $a \rightarrow bcd$

Motifs séquentiels

Pourquoi la recherche de séquence ?

- nombreuses applications :
 - ▶ Analyse des achats des clients
 - ▶ Analyse de puces ADN
 - ▶ Processus
 - ▶ Conséquences de catastrophes naturelles
 - ▶ Web usage mining
 - ▶ Détection de tendances dans des données textuelles

Motifs séquentiels

- Les séquences temporelles
- Les patterns séquentiels
 - ▶ séquence d 'items ordonnés (pas ensemble)
 - ▶ similaire aux règles associatives mais l'ordre importe
 - exemple : " achat de graines, puis de terreau, puis de gants "
 - consultation de pages web (pageA, pageC, pageF)
- Les règles cycliques
 - ▶ règles vérifiées périodiquement
 - ▶ tous les matins, café \rightarrow sucre, gâteaux
 - ▶ $X \rightarrow Y$ cycle (l,o) signifie que $X \rightarrow Y$ toutes les l unités de temps en commençant au temps o.

Exemple : Le Web Usage Mining

- Le Weblog contient des données sur la dynamique du Web
 - ▶ Son analyse permet de cibler les utilisateurs (clients, marchés) potentiels
- La recherche de régularités (séquences fréquentes de pages visitées) permet :
 - ▶ D'ajuster la conception des pages et des liens et d'améliorer les performances des sites
 - ▶ Les associations de pages côté client permettent d'optimiser le cache du navigateur, d'effectuer du « prefetching »

Web Usage Mining

- Analyse de l'usage des visiteurs sur un site Web
- Les pages contiennent l'information
- Les liens sont des « routes » (hyperliens)
- Comment les personnes naviguent-elles sur Internet ?
- Principe : intégrer et « fouiller » ces données pour en produire de l'information et de la connaissance.
L'information sur les chemins de navigation disponibles dans des fichiers logs.

Buts :

- La connaissance sur la manière dont les visiteurs utilisent un site Web permet de :
 - ▶ Fournir une aide pour réorganiser le site
 - ▶ Aider le concepteur à positionner l'information importante que les visiteurs recherchent.
 - ▶ Précharger et cacher les pages
 - ▶ Fournir des sites adaptatifs (personnalisation)
 - ▶ Eviter le « zapping »
 - ▶ Utile dans le cas du e-commerce

Exemple : Le panier de la ménagère

- But : mettre en évidence un ordre fréquent dans l'historique de l'achat.
 - ▶ Ex : Ordinateur puis webcam
- La séquence : $\langle (ef) (ab) (df) c b \rangle$ est composée d'ensembles d'articles.
- L'ordre est important pour les séquences mais pas dans les ensembles

Mesure de support

- Support minimal :
 - ▶ nombre minimum d'occurrences d'un motif séquentiel pour être considéré comme fréquent
 - ▶ l'occurrence n'est prise en compte qu'une fois dans la séquence
- Support (b) dans $\langle (a) (bc) (d) (b) \rangle = 1$ (voir exple)

Problématique des motifs séquentiels fréquents

- Soit D une base de données de transactions de clients
- Soit σ une valeur de support minimal
- Rechercher toutes les séquences S
 - ▶ telles que : $\text{support}(S) \geq \sigma$ dans D

| id | séquence |
|----|-----------------------|
| 10 | <a (abc) (ac) d (cf)> |
| 20 | <(ad) c (bc) (ae)> |
| 30 | <(ef) (ab) (df) c b> |
| 40 | <e g (af) c b c> |

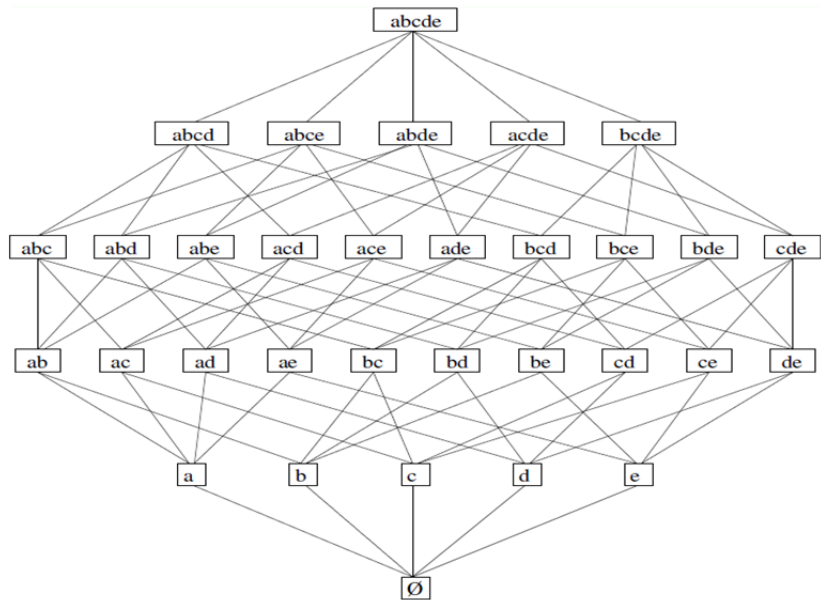
Avec un seuil de support $\text{min_sup} = 2$, $\langle (ab) c \rangle$ est un motif séquentiel fréquent

Autre exemple

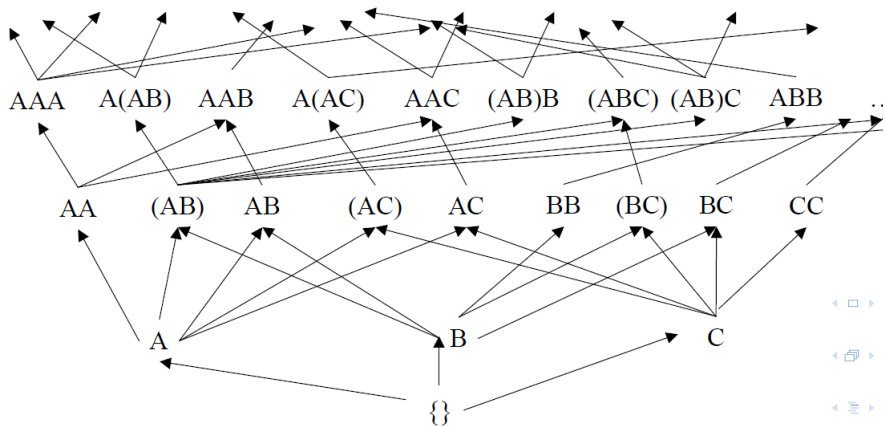
| Clients | Date 1 | Date 2 | Date 3 | Date 4 |
|---------|----------|----------|----------|----------|
| C1 | 10 20 30 | 20 40 50 | 10 20 60 | 10 40 |
| C2 | 10 20 30 | 10 20 30 | | 20 30 60 |
| C3 | 20 30 50 | | 10 40 60 | 10 20 30 |
| C4 | 10 30 60 | 20 40 | 10 20 60 | 50 |

min_support= 3 : $\langle (10\ 30)\ (20)\ (20\ 60) \rangle$ est un motif séquentiel fréquent

Itemsets : Espace de recherche



Motifs Séquentiels : l'espace de recherche



La propriété d'antimonotonie est conservée dans cet espace

- Si une séquence n'est pas fréquente, aucune des super-séquences de S n'est fréquente
 - ▶ $\text{Sup}(< (10) (20\ 30) >) < \text{minsupp}$
 - ▶ $\text{Sup}(< (10) (20\ 30) (40) >) < \text{Sup}(< (10) (20\ 30) >) < \text{minsupp}$

Principaux algorithmes

- introduction du concept initial et algorithme Apriori-like 1996 (voir exple)
- GSP - Generalized Sequential Patterns : 1996
- Pattern-growth : FreeSpan and PrefixSpan : 2001

Algo Apriori-like pour les séquences [96]

- 1 recherche des séquences de taille 1
- 2 calcul des supports, élimination des séquences non fréquentes
- 3 génération des candidats de taille 2 à partir des fréquents de taille 1
- 4 ...etc.

Génération des séquences candidates par jointure :

- S-Extension : ajout d'une séquence
- I-Extension : ajout d'un itemset

Amélioration de l'algorithme

- PSP (Prefix Tree for Sequential Patterns) [01]
- Même principe que l'algorithme FP-Tree et FP-Growth (cf Cours 3)

Text Mining

Autre application de la fouille de données : Text Mining (depuis les années 2000 vs 94 pour le datamining)

- Objectifs
- Documents électroniques
 - ▶ Structurés (10%) et non-structurés (90%)
 - ▶ Beaucoup d'outils limités au structuré (BDR)
 - ▶ Grand volume, croissance exponentielle
- Les BD textuelles sont omniprésentes
 - ▶ Bases de données de bibliothèques,
 - ▶ bases de données de documents, mails, www ...
 - ▶ exemple Medline : 22 millions d'articles scientifiques en médecine (de 1960 à 2013)
- Problèmes
 - ▶ Recherche plein texte (IR)
 - ▶ Extraction de connaissances (catégorie, mots-clés, ...)
 - ▶ Structuration (XML, Tables, RDF LinkedData)

Text Mining

- L'extraction de connaissance à partir de données textuelles (découvertes de tendances, classification/ organisation,)
- Procédé consistant à synthétiser (classer, structurer, résumer, . . .) les textes en analysant les relations, les patterns, et les règles entre unités textuelles (mots, groupes, phrases, documents)
- Techniques :
 - ▶ Classification
 - ▶ Apprentissage
 - ▶ Recherche d'information
 - ▶ Statistiques
 - ▶ TALN = Traitement automatique du langage naturel

Étapes de la fouille de textes

- ❶ Sélection du corpus de documents
 - ▶ Documents pré-classés
 - ▶ Documents à classer
- ❷ Extraction des termes
 - ▶ Analyse grammaticale et/ou lemmatisation
 - ▶ Filtrage des termes extraits
- ❸ Transformation
 - ▶ Passage à un espace vectoriel
 - ▶ Réduction des dimensions
- ❹ Classification
 - ▶ Automatique supervisée ou non
 - ▶ Élaboration de taxonomie (classement)
- ❺ Visualisation des résultats et interprétation des résultats

Text Mining vs. Recherche d'information

Recherche d'Information (Information Retrieval)

- Domaine développé en parallèle des bases de données
- L'information est organisée dans (un grand nombre de) documents
- Pb : localiser les documents pertinents en se basant sur l'entrée de l'utilisateur (mots clés ou documents exemples)

Text Mining - Classification automatique

- Classification automatique d'un grand nombre de documents (pages Web, mails, fichiers textuels) basée sur un échantillon de documents pré-classifié (en fonction de thèmes)
- Mise en oeuvre :
 - ▶ Echantillon : des experts génèrent l'échantillon
 - ▶ Classification : l'ordinateur découvre les règles de classification
 - ▶ Application : les règles découvertes peuvent être utilisées pour classer des nouveaux documents et les affecter à la bonne classe

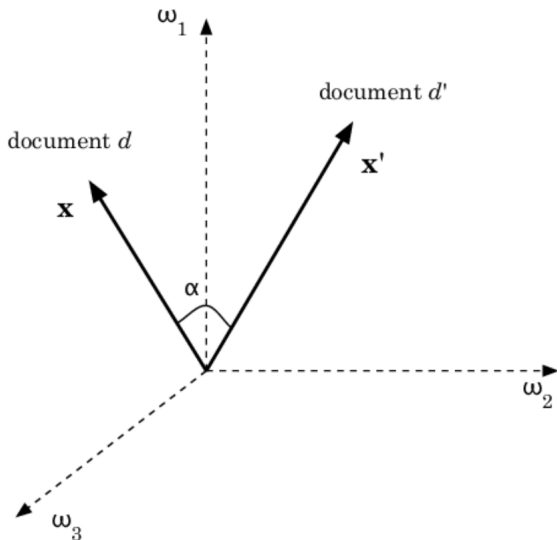
Text Mining - Classification

Quelques problèmes

- Synonymie : Même concept qualifié par termes différents (ventre/abdomen)
 - ▶ un mot peut ne pas apparaître dans un document même si le document lui est très lié
- Polysémie : le même mot peut avoir plusieurs sens
 - ▶ Termes identiques utilisés dans des contextes sémantiques différents (avocat véreux, avocat mûr...)
- Représentation des documents
 - ▶ vecteurs de termes, choix des termes représentatifs, calcul de la distance entre un vecteur représentant le groupe de documents et celui du nouveau document, ...
- Évolution des classes dans le temps

L'espace des vecteurs

- Chaque document est vu comme une séquence de mots
- Le nombre de mots du lexique présents dans les documents du corpus détermine la dimension de l'espace



Représentation des documents

- Vecteurs de documents
- Matrice Terme/Document ou Document/terme
- Nécessité de pondérer : pondération (importance relative)
- Nécessité de réduire l'espace : réduction de dimension

| | d_1 | d_2 | • | • | • | d_d |
|-------|----------|----------|---|---|---|----------|
| t_1 | w_{11} | w_{12} | • | • | • | w_{1d} |
| t_2 | w_{21} | w_{22} | • | • | • | w_{2d} |
| • | • | • | | | | • |
| • | • | • | | | | • |
| • | • | • | | | | • |
| t_t | w_{t1} | w_{t2} | • | • | • | w_{td} |

Term frequency (TF)

- hypothèse : un terme qui apparait plusieurs fois dans un document est plus important qu'un terme qui n' apparaît qu'une seule fois
- TF_{ij} = Fréquence du terme t_i dans le document d_j

Inverse document frequency (IDF)

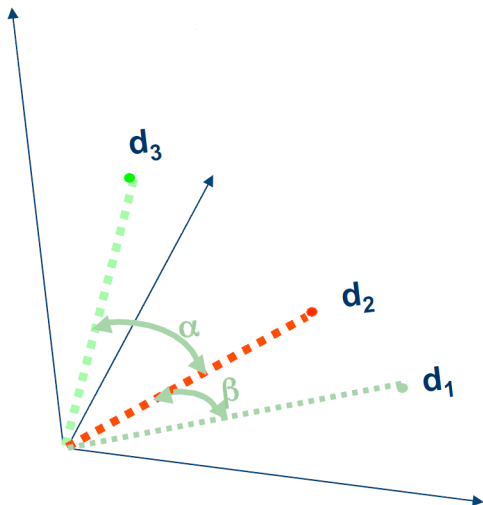
- Un terme qui apparaît dans peu de documents est un meilleur discriminant qu'un terme qui apparaît dans tous les documents
- df_i = nombre de documents contenant le terme t_i
- d = nombre de documents du corpus
- Inverse document frequency $IDF_i = \log\left(\frac{d}{df_i}\right)$

Pondération TF-IDF

- TF-IDF signifie Term Frequency x Inverse Document Frequency [89]
- mesure l'importance d'un terme dans un document relativement à l'ensemble des documents.
- $W_{ij} = tf_{ij} \times \log(\frac{d}{df_i})$
 - ▶ tf_{ij} fréquence du terme i dans le document j
 - ▶ df_i nombre de documents contenant le terme i
 - ▶ d nombre de documents du corpus

Similarité entre documents

- Permet de ranger les documents par pertinence
- Le cosinus de l'angle entre 2 documents est souvent utilisé
- $\alpha > \beta \implies \cos(\alpha) < \cos(\beta)$; d_2 est plus proche de d_1 que de d_3



Classification de documents

- Trois algorithmes de classification supervisée souvent considérés
 - ▶ KNN (K Nearest Neighbor) : Un document est classé dans la catégorie dominante parmi ses k plus proches voisins
 - ▶ Centroïd : Sélection de la catégorie de plus proche centroïde
 - ▶ Naïve Bayes : Sélectionner la catégorie la plus probable (voir cours apprentissage)

Phase 1 : Apprentissage

- ensemble de documents “exemple” pré-affectés
- pré-traitement et sélection des termes
- représentation des documents
- estimation des paramètres du classifieur
- classifieur

Phase 2 : Classement

- document d à classer
- représenter d
- Utiliser le classifieur de la phase 1 et calculer le score (C_i , d)
- Affecter d à C_i
- document d avec la ou les catégories affectées

Algorithme KNN

- Calcul de similarité
 - ▶ Entre le nouveau doc. et les exemples pré-classés
 - ▶ $\text{Similarité}(d1,d2) = \cos(d1,d2)$
 - ▶ Trouve les k exemples les plus proches
- Recherche des catégories candidates
 - ▶ Vote majoritaire des k exemples
 - ▶ Somme des similarités > seuil
- Sélection d'une ou plusieurs catégories
 - ▶ Plus grand nombre de votes
 - ▶ Score supérieur à un seuil

Text Mining - Corrélations

- Analyse d'associations basée sur des mots clés
- Rechercher des associations/corrélations parmi des mots clés ou des phrases
- Mise en oeuvre
 - ▶ Pré-traitement des données : parser, supprimer les mots inutiles (le, la, ...) et prise en compte d'une analyse morpho-syntaxique (e.g. lemmatiseur)
 - ▶ un document est représenté par : (id, {mots clés})
 - ▶ Appliquer des algorithmes de recherche de règles d'association

Text Mining - Corrélations

- Quelques problèmes
 - ▶ Ceux du traitement de la langue naturelle
- Lemme : forme canonique
 - ▶ book, books [book]
 - ▶ mange, mangera, mangeaient, mangeant, [manger]
 - ▶ Nécessite une grammaire
 - ▶ Généralement entrée de référence en dictionnaire
- Stemming : racine + dérivation [préfixe/suffixe]
 - ▶ produire, production, productivité [produc]
 - ▶ Calculer par un algorithme (Stemmer) (Porter)
- Les mots inutiles (ordinateur ? Utile ?) (Exple des 200 mots d'Oracle text)
- Réduction de l'espace de recherche : les associations de mots, phrase, paragraphe, ...

Classification pour la préparation des données

Préparation des données

- Données réelles imparfaites/endommagées
 - ▶ Incomplètes
 - ▶ Bruitées
 - ▶ Incohérentes
- Nécessité de préparer les données
 - ▶ Nettoyage
 - ▶ Intégration et transformation
 - ▶ Réduction
 - ▶ Discrétisation

Nettoyage - Données bruitées

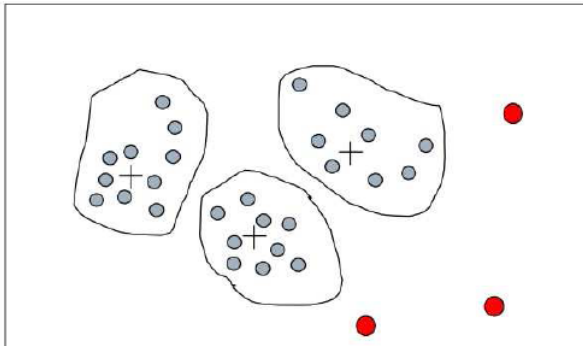
On peut :

- Trier et partitionner (discrétiser)
- Classifier (exceptions)
- Appliquer un modèle de prédiction (ex : une fonction de régression)

Nettoyage - Supprimer les déviations

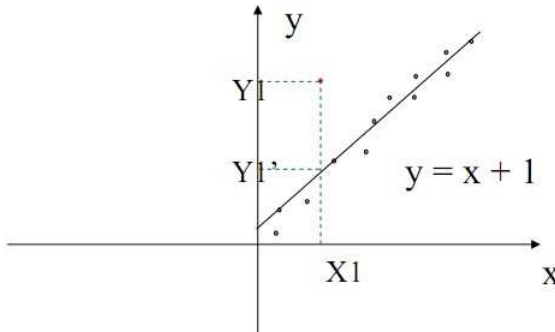
Classifier

- Les valeurs similaires sont organisées en classes
- Les valeurs hors-classes sont considérées comme des déviations



Lissage par régression

- Les données sont lissées de manière à approcher une fonction
 - ▶ Régression linéaire
 - ▶ Régression linéaire multiple



Classification

- Classification : regrouper les instances en groupes ayant une ou des propriétés communes
- Les groupes sont les « classes » distinguées
- Classification non-supervisée :
 - ▶ Clustering (anglo-saxons) ou cluster analysis
 - ▶ Les classes ne sont pas connues à l'avance ;
Apprentissage non-supervisé
- Classification supervisée :
 - ▶ Classification (anglo-saxons) ou classement
 - ▶ Les classes sont connues à l'avance
 - ▶ Apprentissage supervisé

Classement

- Classement : Classification supervisée (classement)
- Tâche de prédiction : Prédit des variables catégorielles
- ① Construire un modèle de classement (classifieur) des données en se basant sur un ensemble appelé ensemble d'apprentissage (EA) (training set)
- ② Utiliser le modèle pour classer de nouvelles données
- Exemple :
 - ▶ On dispose de données sur des patients atteint d'un cancer et d'autres sains
 - ▶ On déduit de leurs caractéristiques un modèle
 - ▶ Ce modèle est utilisé pour déterminer le risque pour d'autres patients de développer un cancer

(cf cours F.Nicart)


Classement par arbre de décision

- Arbre de classes :
 - ▶ Division en sous-classes correspond à un test sur une variable
 - ▶ Règles dont la prémisse exprime une condition sur des variables (dites explicatives) et la conclusion est la variable à expliquer (classe label)
- Ex : Si $\text{age} < 30$ ans et $\text{Etudiant} = \text{où}$ alors $\text{achete_PC} = \text{où}$
- Avantage : Modèle lisible et compréhensible par l'utilisateur

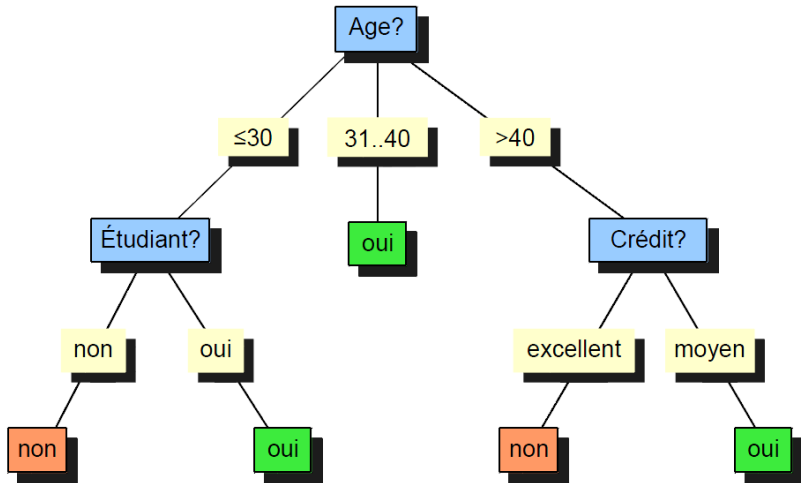
(cf TP3)

Classement : exemple

- Jeu de données D : Clients ayant ou non acheté un PC (classe)



| Attributs prédictifs | | | | Attributs cible |
|----------------------|---------|----------|-----------|-----------------|
| Age | Revenus | Étudiant | Crédit | Achète_PC |
| <=30 | élevés | non | moyen | non |
| <=30 | élevés | non | excellent | non |
| 31...40 | élevés | non | moyen | oui |
| >40 | moyens | non | moyen | oui |
| >40 | faibles | oui | moyen | oui |
| >40 | faibles | oui | excellent | non |
| 31...40 | faibles | oui | excellent | oui |
| <=30 | moyens | non | moyen | non |
| <=30 | faibles | oui | moyen | oui |
| >40 | moyens | oui | moyen | oui |
| <=30 | moyens | oui | excellent | oui |
| 31...40 | moyens | non | excellent | oui |
| 31...40 | élevés | oui | moyen | oui |
| >40 | moyens | non | excellent | non |



Règles de classification

- L'arbre peut être représenté sous forme de règles de classification
- Une règle pour chaque branche allant de la racine à une feuille
- Chaque terme attribut-valeur constitue un opérande de la conjonction en partie gauche
- Chaque feuille correspond à une classe à prédire

Procédure de construction

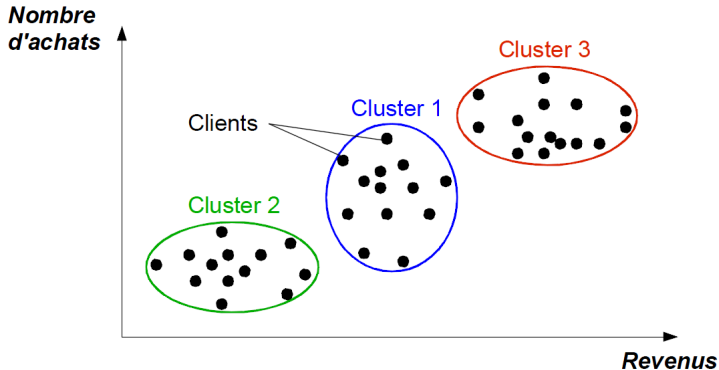
- Principe :
 - ▶ Partitionner les données (nœud de l'arbre) de telle sorte que les parties soient plus homogènes que le nœud parent et qu'elles soient les plus différentes possible entre elles vis à vis de la variable classe
- Problèmes : comment mesurer la bonne partition ?
- PB du passage à l'échelle pour de grands ensembles de données

Clustering

- Clustering (classification non-supervisée)
- Tâche de description
 - ▶ Recherche des groupes (clusters) dans un ensemble de données avec la plus grande similarité possible intra-groupe et la plus grande dissimilarité possible inter-groupes
- Exemple :
 - ▶ On dispose de données sur des clients (age, nombre d'enfants, revenus, nombre d'achats, etc.)
 - ▶ On regroupe en clusters les clients ayant des caractéristiques communes
 - ▶ Pour chaque cluster, on définit une offre commerciale adressée aux clients de ce clusters

(exemple fichier bank-data.csv du TP3)

Exemple



Représentation bi-dimensionnelle des données

Mesures de similarité

- Objets « suffisamment similaires » regroupés en clusters
 - ▶ Définition du seuil de similarité difficile
- Évaluation des clusters
 - ▶ Distance entre objets à l'intérieur du cluster
 - ▶ Distance avec les objets des autres clusters
- Les données bruitées et les déviations (outliers, objets hors-normes) nuisent à la qualité du clustering