De la donnée à la décision

Sofian MAABOUT
LaBRI. Université Bordeaux 1

Décider c'est

- choisir, parmi plusieurs actes possibles, celui qui apparaît comme le plus pertinent pour atteindre un résultat envisagé, dans un délai jugé souhaitable et possible, en utilisant au mieux les informations et les ressources disponibles.

- ➔ Extraire de l'information

Information intéressante ?

- Exemple: CRM
  - La richesse d'une entreprise est, entre autres, sa clientèle
  - Objectifs de l'entreprise
    - Augmenter la rentabilité et la fidélité de ses clients
    - En maitrisant les risques
    - En utilisant les bons canaux au bon moment pour vendre le bon produit
  - Un moyen d'y parvenir
    - Gestion de la relation client (GRC)
    - Customer Relationship Management (CRM)
  - Sur quoi se bases-t-on ?
    - Les données sur les clients

De plus en plus de données

- L'accroissement des expertises et de la technicité
  - … font perdre l'approche globale
  - … obligent à stocker de plus en plus de données pour les besoins opérationnels de la gestion quotidienne

- Mais « trop de données tue la donnée » ➔ on connaît de moins en moins les clients

Fouiller les données

- ensemble des algorithmes et méthodes
  - Destinés à l'exploration et l'analyse
  - De grandes quantités de données
  - Sans a priori
  - En vue de détecter des règles, des tendances inconnues ou cachées, restituant de façon concise l'essentiel de l'information utile
  - … pour l'aide à la décision

Téléphonie

- Deux événements
  - Fin du monopole de France télécom
  - Arrivée à saturation du marché
- Sujets dominants
  - Score d'attrition (churn=changement d'opérateur)
  - Text mining (analyse des lettres de réclamation)
  - Optimisation des campagnes marketing
- Problème du churn:
  - Coût d'acquisition d'un nouveau client:: 300 euros
  - + d'un million d'utilisateurs changent chaque année d'opérateur

Commerce

- La vente par correspondance (VPC)
  - Utilise depuis longtemps des scores d'appétence
  - Optimiser les cibles pour réduire les coûts
  - La Redoute envoie (il y a 5 ans) 250 millions de documents à sa clientèle
- E-commerce
  - Personnalisation des pages du site en fonction du profil de l'internaute (Amazon le fait)
- Distribution
  - Détermination des profils de consommateurs, « le panier de la ménagère », l'effet des soldes ou de la publicité
  - Détermination des meilleures implantations (géomarketing)

Les 2 grandes familles d'outils

- Techniques descriptives
- Techniques prédictives

Description

- Il s'agit de **mettre en évidence des informations présentes mais cachées par le volume des données**

- Réduit, résume et synthétise les données

Techniques descriptives

- Regroupement (ou segmentation, ou clustering)
- Recherche d'associations, de corrélations
- Recherche de séquences similaires

Prédiction

- Vise à **extrapoler de nouvelles informations** à partir d'informations déjà présentes
- Explique les données
- Il y a une variable cible à prédire

Techniques prédictives

- Classification
  - Arbres de décision
  - Classification bayésienne
  - Réseaux neuronaux
  - Méthodes SVM (support vector machine)
  - Régression
  - …
- Certaines techniques ne s'appliquent qu'à un type de variable cible (quantitative ou qualitative)

Associations (1)

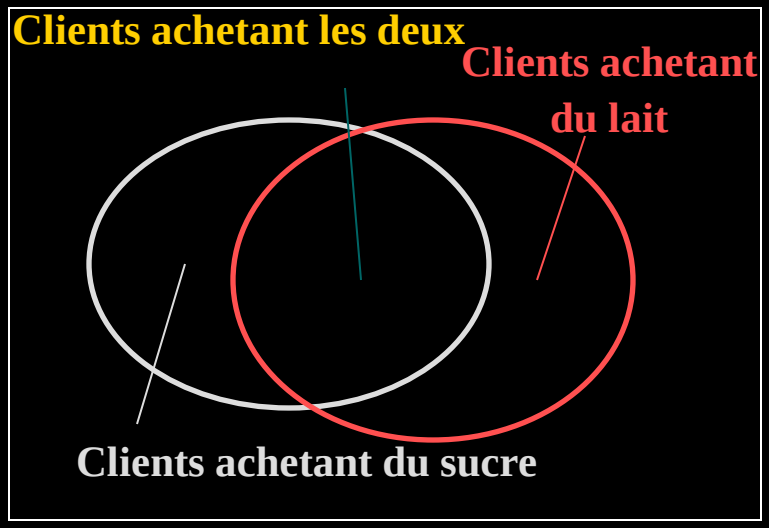
- Les enseignes de grands magasins proposent régulièrement des promotions sur divers produits
- Une promotion représente un manque à gagner pour le magasin
- Dilemme : Comment proposer des promotions intéressantes pour les clients tout en réduisant le manque à gagner ?
- Regarder les habitudes d'achats des clients : si en général, les clients qui achètent du lait achètent aussi du sucre, alors il n'est pas intéressant de faire des promotions sur les 2 produits en même temps

Associations (2)

- Règles d'association :
  - motifs de la forme : Corps ➔ Tête
  - Exemple : Lait ➔ sucre
- Etant donnés: (1) une base de transactions, (2) chaque transaction est décrite par un identifiant et une liste d'items
  - Trouver: toutes les règles qui expriment une association entre la présence d'un item avec la présence d'un ensemble d'items
    - Ex., *98% des personnes qui achètent du lait achètent du sucre*

Associations: Support et Confiance (3)

Trouver les règles *X & Y ⇒ Z avec un support > s et une confiance >c*



- support *s*, probabilité qu'une transaction contienne {X, Y, Z}
- confiance *c*, probabilité conditionnelle qu'une transaction qui contient {X, Y} contienne aussi *Z*

Confiance=support(X,Y,Z)/support(X,Y)

ID Transaction	Items
2000	A,B,C
1000	A,C
4000	A,D
5000	B,E,F

Soit support minimum 50%, et confiance minimum 50%,

$A \Rightarrow C$ (50%, 66.6%)

$C \Rightarrow A$ (50%, 100%)

Problème algorithmique

- Si on a 10^6 produits, on a $2^{10^{\wedge}6}$ ensembles de produits à vérifier ! (2^30= 1 milliard)
- Pourtant, en procédant intelligemment, on y arrive
- Idée: Exploiter la propriété de non monotonicité :
  - Si {A,B,C} n'est pas fréquent, alors {A,B,C,D} ne peut pas l'être

Prévision (1)

- Les établissements financiers accordent des crédits à leurs clients
- L'attribution d'un crédit dépend de certains critères que le client doit satisfaire
- Dilemme :
  - Si on ne prête qu'aux très riches, on n'aura pas de problèmes de remboursement mais on perd les autres clients (pas de risque).
  - Si on prête aux moins riches, on ne va pas perdre les clients mais on est exposé aux non remboursements (trop de risque)
- Idée : se baser sur l'historique des clients pour dresser des profils de bons clients, clients moyens, et mauvais payeurs

Prévision (2)

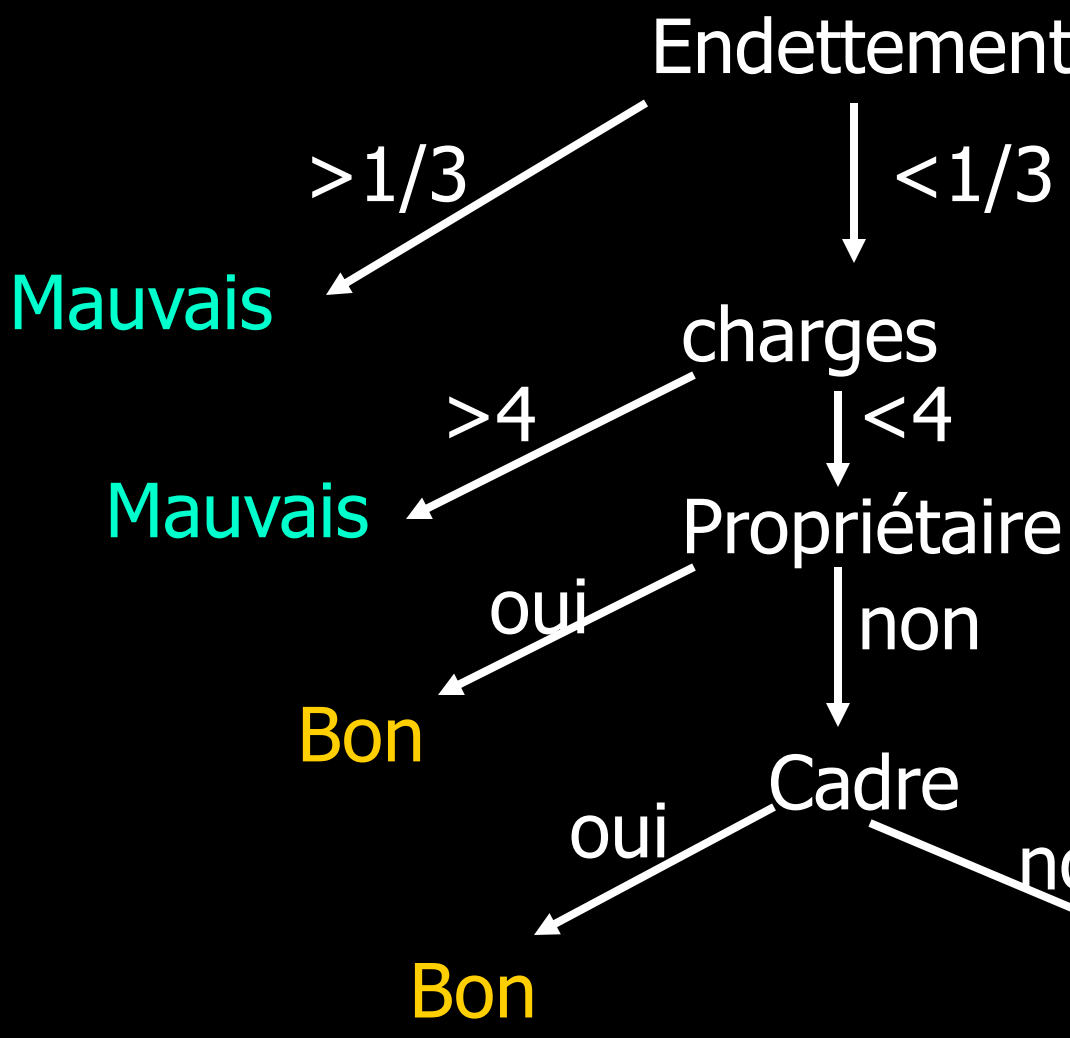
- L'organisme dispose d'un fichier décrivant ses différents clients à qui il a attribué un crédit
- Chaque client est décrit par un certain nombre d'attributs : Salaire, situation marital, emploi, locataire/propriétaire, personnes à charge, montant crédit, …
- A chaque client, on ajoute un attribut particulier qui est le nom de la classe et qui est égal à bon, mauvais ou moyen
- Le but consiste à extraire à partir de ce fichier un ensemble de règles qu'on va utiliser lorsqu'un nouveau client demande un crédit pour savoir si l'on peut le lui attribuer ou pas

Prévision (3)

- Exemples de règles de production:
  - Si crédit > 1/3 salaire → mauvais
  - Si crédit <1/3 salaire & charges >4 → mauvais
  - Si crédit <1/3 salaire & charges <4 & propriétaire = oui → bon
  - Si crédit <1/3 salaire & charges <4 & propriétaire=non & cadre=oui → bon
  - Si crédit <1/3 salaire & charges <4 & propriétaire=non & cadre = non → moyen
  - …

- Ces règles peuvent être représentées par un arbre de décision

Prévision (4)



En pratique, les systèmes construisent d'abord les arbres d'où ils dérivent les règles

Regroupement (1)

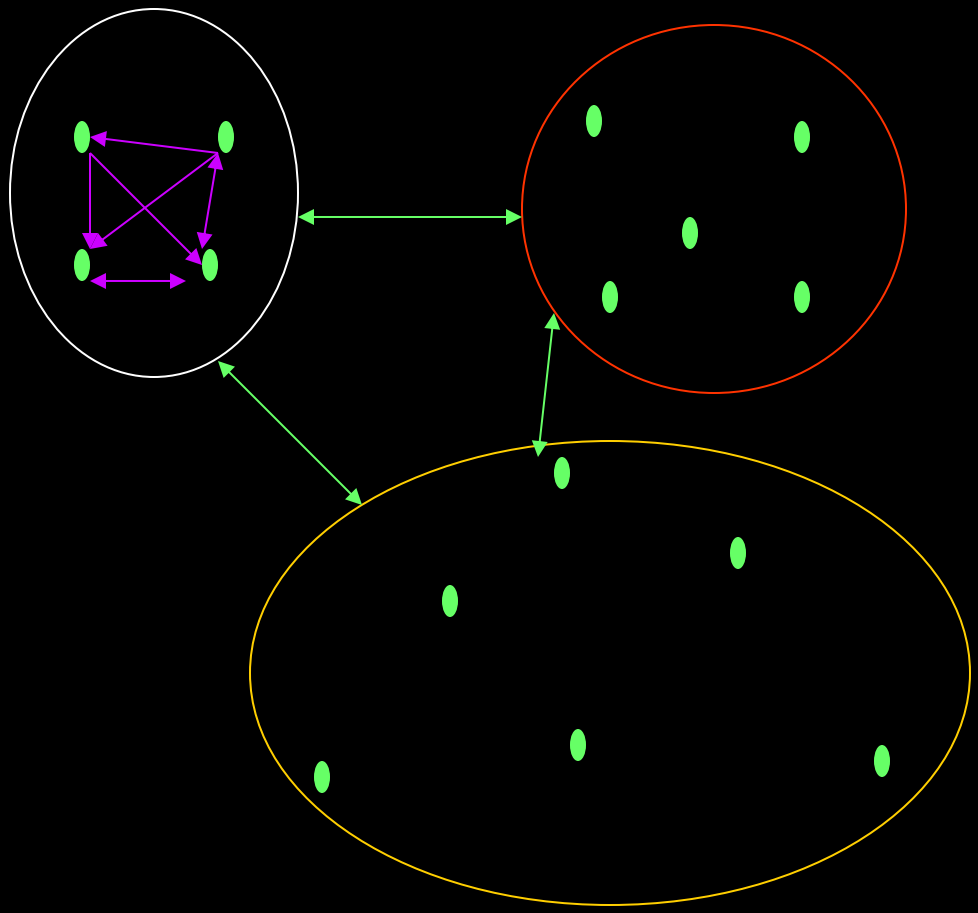
- Considérons une entreprise de vente par correspondance qui veut envoyer des prospectus publicitaires à ses clients
- L'entreprise a un fichier de 100.000 clients. Le coût de la campagne est estimé à 0,5 € ce qui fait un coût global de 50.000 €
- D'où l'intérêt de cibler les envois : un client qui a l'habitude d'acheter du matériel de pêche n'a que faire d'une pub qui porte sur les vêtements pour le golf (en général …)
- Dilemme : ne pas envoyer de prospectus versus en envoyer mais en ciblant les clients
- Idée : construire des groupes de clients. Chaque groupe sera soit destinataire d'un prospectus ciblé soit on ne lui envoie pas du tout.

Regroupement (2)

- Les groupes (ou clusters) sont construits de sorte à
  - Maximiser la similarité entre éléments d'un même groupe
  - Maximiser la dissimilarité entre groupes

- Les questions auxquelles le décideur est confronté :
  - Si chaque individu forme à lui seul un groupe, alors la similarité intra-groupe est maximale mais la dissimilarité inter-groupes peut ne pas l'être
  - Si on ne forme qu'un seul groupe, la dissimilarité intergroupes est maximale, mais la similarité intra-groupe peut ne pas l'être
  - → des techniques qui permettent à l'utilisateur de fixer le nombre k de groupes qu'il veut construire

Regroupement



Conclusion

- Utiliser un système de datamining est intéressant quand on sait
  - Quelles actions nous voulons entreprendre
  - Quelles types d'information nous devons rechercher
- Pour chaque type d'information, il existe plusieurs techniques qui ne sont dans la plupart des cas, pas équivalentes mais complémentaires
- Pour bien exploiter les informations extraites, il est important de comprendre les techniques sous jacentes