Chapitre 1

Introduction

1.1 Qu'est-ce que le data mining ?

En français *data mining* se traduit *fouille de données*. Mais on utilise aussi l'expression *extraction de connaissances à partir de données*. Ainsi le data mining est un ensemble de méthodes qui permettent d'extraire de l'information (de façon semi-automatique) de grandes bases de données. Les informations découvertes sont des règles, des associations, des tendances inconnues (non fixées a priori), des structures particulières. Elles doivent aider à prendre des décisions.

Le terme *data mining* est apparu au départ dans le secteur tertiaire : banque, assurance, vente par correspondance. Dans ce secteur, l'objectif d'une étude de data mining est, en général, de mieux connaître sa clientèle pour la fidéliser ou pour attirer de nouveaux clients. Cependant le data mining est aujourd'hui utilisé dans un grand nombre d'applications de tout type et dans des domaines variés (vente, industrie, médecine, environnement, etc)

Exemples d'applications

1. En bioinformatique, de nombreuses personnes travaillent sur le développement de méthodes permettant de regrouper de gènes en fonction de leurs caractéristiques. De même en écologie, un problème usuel consiste à regrouper des espèces (animales ou végétales) en fonction de caractéristiques communes : type d'habitat, besoin nutritionels, résistence, ...
2. Une banque constate qu'un pourcentage de ses clients, plus important que d'habitude, partent à la concurence. Elle cherche alors à savoir si ces clients ont des caractéristiques particulières. Si oui, elle s'appuira sur les informations découvertes pour leur proposer des avantages les incitant à rester.

Un statisticien distingue généralement deux phases dans l'étude d'un problème (auquel est associé un ensemble d'observations) :

1. Analyse exploratoire, phase dans laquelle on se familiarise avec les données et on privilégie les représentations graphiques.
2. Modélisation, phase dans laquelle on cherche à caractériser un phénomène (physique, biologique, économique ou autre) en développant un modèle probabiliste appris sur des observations.

Un processus complet de data mining comporte, lui, 6 phases :

1. Compréhension métier
2. Compréhension des données
3. Préparation des données
4. Modélisation
5. Évalusation
6. Déploiement

Exemple : plaintes chez Chrysler (D. Larose).

1.2 Quelles sont les différences entre le data mining et la statistique ?

Généralement, une approche statistique consiste à formuler des hypothèses théoriques qui sont confirmées ou infirmées à l'aide de tests statistiques. Si l'étude est correctement menée, on formule les hypothèses, on en déduit un plan d'expérience, on recueille les données, puis on les analyses.

En data mining, la recherche d'information est moins guidée. Cependant, le data mining repose sur un ensemble de méthodes issues des statistiques et de l'intelligence artificielle.

Quelques repères :

Statistique :
- quelques centaines d'individus
- quelques variables recueillies avec un protocole spécial (échantillonnage, plan d'expérience...)
- fortes hypothèses sur les lois statistiques suivies
- les modèles sont issus de la théorie et confrontés aux données
- méthodes probabilistes et statistiques
- utilisation en laboratoire

Analyse des données :
- méthodes essentiellement descriptives
- quelques dizaines de milliers d'individus
- quelques dizaines de variables
- importance du calcul et de la représentation visuelle

Data mining :
- plusieurs millions d'individus
- plusieurs centaines de variables
- nombreuses variables non numériques, parfois textuelles
- données recueillies avant l'étude, et souvent à d'autres fins
- données imparfaites, avec des erreurs de saisie, de codification, des valeurs manquantes, aberrantes
- population constamment évolutive (difficulté d'échantillonner)
- nécessité de calculs rapides, parfois en temps réel

- on ne recherche pas toujours l'optimum mathématique, mais le modèle le plus facile à appréhender par des utilisateurs non statisticiens
- faibles hypothèses sur les lois statistiques suivies
- les modèles sont issus des données et on en tire des éléments théoriques
- méthodes statistiques, d'intelligence artificielle et de théorie de l'apprentissage ("machine learning")
- utilisation en entreprise

1.3 Quels logiciels utilise t'on en data mining ?

1. Les logiciels usuels de statistique : SAS, R, Splus,
2. Des logiciels de calcul numérique : matlab
3. Des parties de logiciels dédiées : SAS/EM, SPSS/Clémentine, Statsoft/Statistica Data Miner, Insight/Insightful Miner
   SPAD
4. Des logiciels libres : weka, tanagra

1.4 Plan du cours

Donner des exemples de pb types ?

2 & 3. Analyses Factorielles : analyse en composantes principales, analyse des correspondances multiples.
4. Positionnement Multidimensionnel.
5. Classification (nuées dynamiques, classification hiérarchique, choix du nombre de classes)
6. Le problème des données manquantes.
7. Régles d'association
8. Arbres de Décision
9. Introduction à la Reg. Logistique
10. Introduction aux Réseaux de neurones
11. Validation et comparaison de modèles

Ouvrages de référence

D. Larose, (2003). *Des données à la connaissance. une introduction au data mining*. Editions Vuibert.

S. Tuffery, (2007). *Data mining et statistique décisionnelle*. Editions Technip.

Han, Kamber, (2006). *Data Mining : Concepts and Techniques, 2nd ed*. The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor.