Analyse et Fouille de Données

Data mining

Saoussen MATHLOUTHI

Saoussen.mathlouthi@ensi-uma.tn

Informatique de gestion: 2LIG

ISG Bizerte

A-U: 2021-2022

Objectifs du cours

Ce cours introduit **l'analyse et la fouille de données (FDD)** appelé aussi **Data Mining**.

Il présente les concepts de base, les principes, les méthodes, les techniques de mise en œuvre et les applications de Data Mining, en mettant l'accent sur deux fonctions principales :

(1) Les **outils de base** d'exploration de données.

(2) La **découverte de modèles.**

du cours

Chapitre I

Introduction à la fouille de données

Chapitre II

Approches de fouille de données

Chapitre III

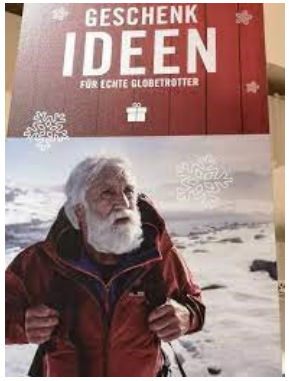
Outils de base de fouille de données

Chapitre IV

La découverte de patrons

Questions

- Est-ce une bonne idée de faire de la publicité de musique rap dans des magazines pour personnes âgées?



- Est-ce que vous saviez que les compagnies de cartes de crédit peuvent suspecter un vol de carte, même si le propriétaire de la carte n'est pas encore au courant?

- Est-ce que vous savez pourquoi les entreprises liées à la bourse sponsorisent les tournois de golfs télévisés?

Plan

- Introduction à la fouille de données
- Exemple d'application à la fouille de données
- Type des données exploitées
- Quelques logiciels de fouille de données

Introduction à la fouille de données

01

Qu'est que c'est la fouille de données?

Fouille de données ou data mining:

- Ensemble d'approches statistiques permettant d'extraire de l'information de grands jeux de données dans une perspectives d'aide à la décision.
- Ensemble des techniques d'exploration de données permettant d'extraire des connaissances sous la forme de **modèles de description** afin de :
  - **Décrire** le comportement actuel des données.
  - Et/ou **Prédire** le comportement futur des données.

Données et information

- Les entreprises produisent des quantités énormes de données
  - Sources: affaires, science, géographie, environnement…
- Les données sont des ressources de valeur
- Besoin de techniques pour analyser les données et extraire des informations / connaissances automatiquement
  - Données = faits
  - Information = modèle / motif (pattern) dans les données

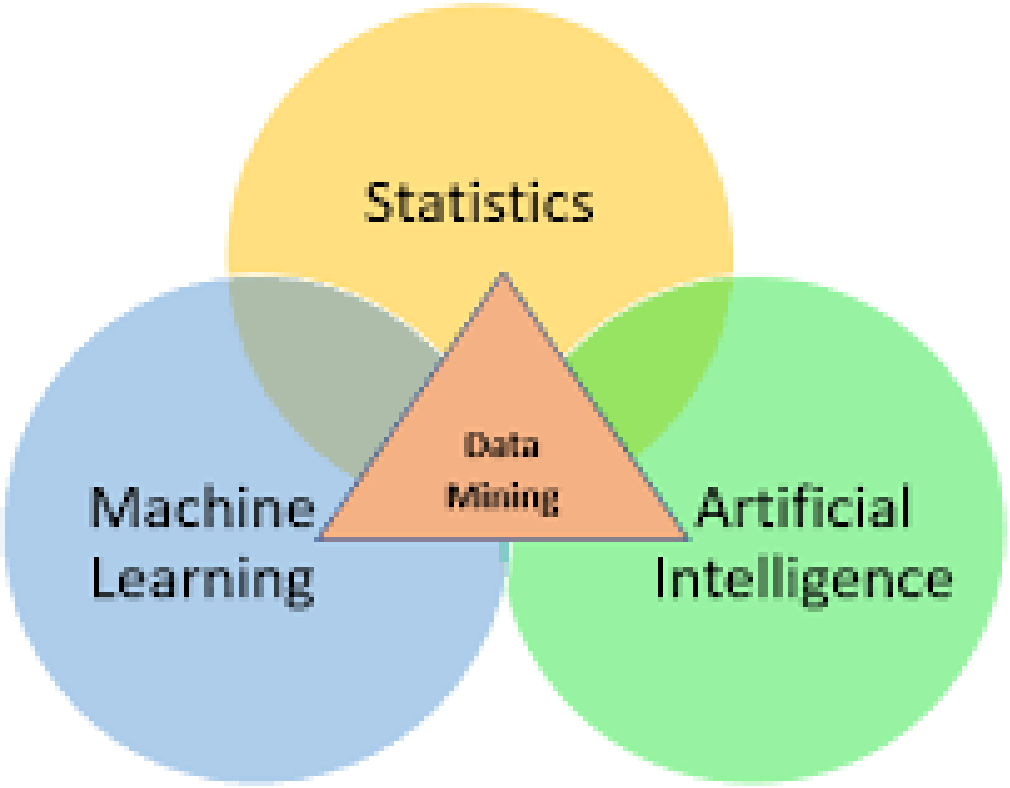
Fouille de données (FDD)

- Fouille / découverte de connaissances dans les données ou création d'un modèle des données
- Processus qui emploie des techniques d'apprentissage automatiques et intelligentes pour analyser et extraire des connaissances, de grandes quantités de données

Pourquoi la FDD maintenant?

- Des machines plus puissantes
- Existence d'algorithmes de fouille de données
- Collections et sauvegardes des données améliorées
- Domaine à la confluence de différents domaines: base de données, statistiques, intelligence artificielle, visualisation, parallélisme…
- Domaine pluridisciplinaire

fouille de données (Data Mining) à l'intersection de plusieurs disciplines



La fouille de données à la rencontre de plusieurs disciplines

Statistiques

Théorie de l'estimation/ tests
Econométrie

Analyse de données

Description factorielle
Discrimination

Apprentissage symbolique
Reconnaissance de formes

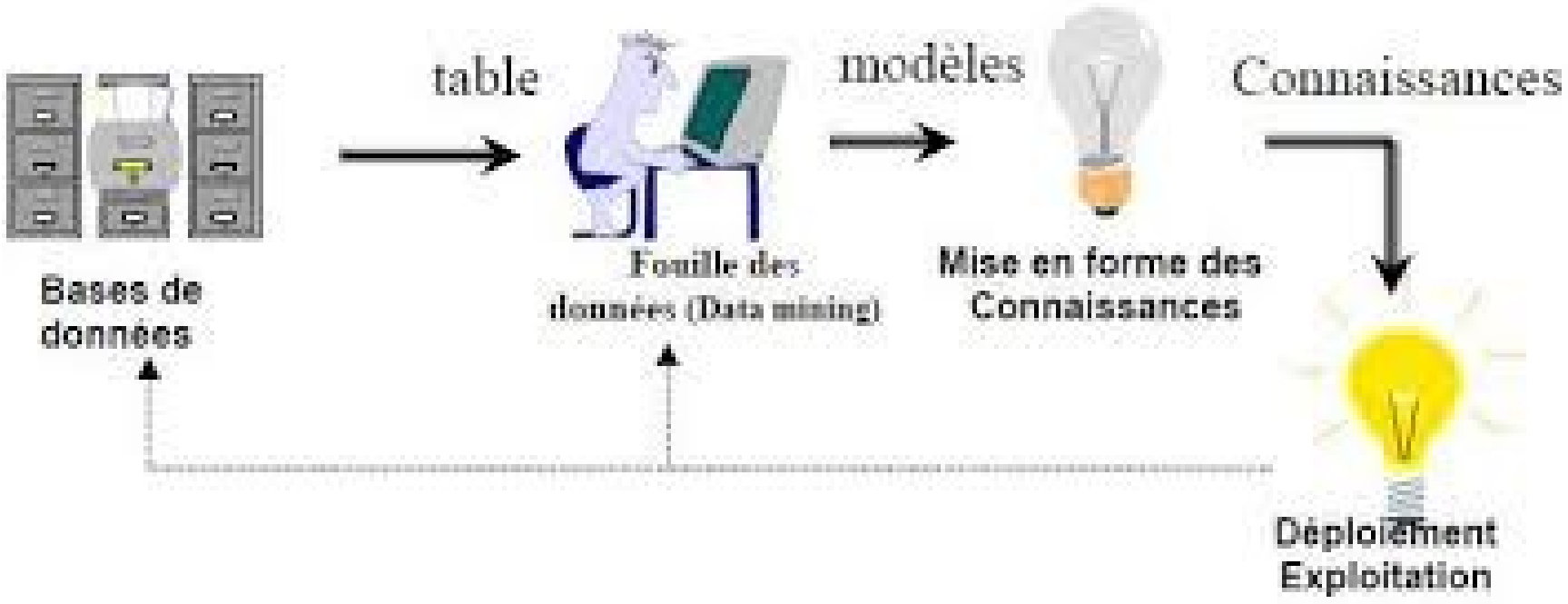
Informatique
Base de données

Exploration de base de données

Informatique
Intélligences
Artificielles

Processus d'extraction de connaissances à partir de données

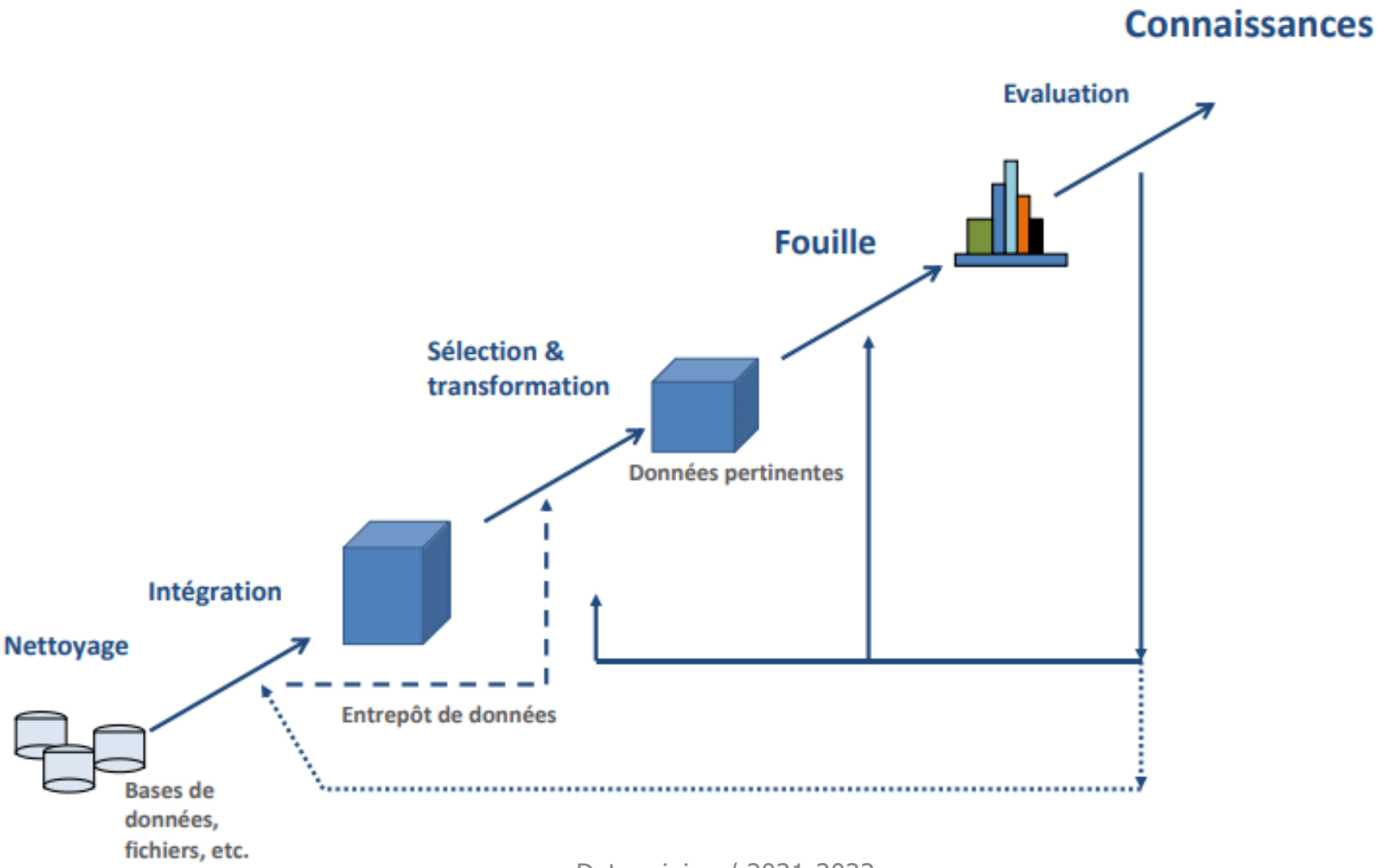
Processus ECD (Extraction de connaissances à partir de données) ou KDD (Knowledge Discovery in Databases) :



Les étapes du data mining

1. Nettoyage des données (erreurs, données manquantes, outliers)
2. Transformation éventuelle des données (normalisation, linéarisation...)
3. Explicitation de l'objectif de l'analyse en terme statistique (régression, classification, clustering...)
4. Choix de la méthode et mise en oeuvre informatique
5. Test (validation de la qualité des résultats)
6. Exploitation

Les étapes du data mining



02

Exemple d'application à la fouille de données

La fouille de données : à quoi cela sert ?

- Publicité ciblée sur internet
- Identification des prospects les plus susceptibles de devenir clients
- Reconnaissance faciale dans une image
- Calcul de la rentabilité des clients
- Evaluer le risque d'un client (credit scoring)
- Détection de fraudes bancaires
- Analyse automatique de contenus textuels (text mining)
- Reconnaissance de la parole
- Calcul de score de réachat
- Prévision de consommation d'électricité
- Prévision de traffic routier
- Tester l'efficacité d'un traitement médical
- …

Concepts / Modèles

- Les concepts sont les résultats de la fouille de données
- Ils montrent les relations dans les données ou ils groupent des éléments fondés sur leur ressemblance
- Structures:
  - Arbres de décision
  - Règles
  - Réseaux de neurones
  - Équations mathématiques…
- Parfois, pas de réelles structures
  - Exemples / Instances
  - Regroupements (clusters)

Exemples de modèles

- Vue classique:
  - Si salaire annuel >= 30,000 et années de service >= 5 et propriétaire = vrai

    alors risque de défaut de paiement = faux
  - Age(X, “jeune”) et Salaire(X,”élevé”) ⇒ Classe(X,”A”)
- Vue probabilistique:
  - La plupart des personnes qui ont un bon crédit sont propriétaires

Exemples de modèles

- Vue exemplaire:
  - Un exemple est déterminé comme une instance d'un concept
  - Les exemples suivants sont considérés comme ayant un bon crédit:
    - Salaire = 32,000, années = 6, propriétaire
    - Salaire = 52,000, années = 16, locataire

03

Type des données exploitées

FDD: Sur quelles données? (1)

- Base de données relationnelles (SGDBR)
- Base de données transactionnelles
  - Exemple: Données de ventes des supermarchés
- Base de données orientées objets (SGBDO)
- Base de données relationnelles objets (SGBDRO)
- Base de données temporelles
  - Exemple: Bourse

FDD: Sur quelles données? (2)

- Base de données spatiales
  - Exemples: Images provenant de satellites, cartes géographiques
- Base de données hétérogènes
- Base de données “héréditaires”
- Entrepôts de données
- Base de données textuelles
  - Les données sont décrites par des mots
  - Exemples: La toile, le courrier électronique, les pages html/xml

FDD: Sur quelles données? (3)

- Base de données multimédia
  - Des techniques de recherche et de stockage avancées sont nécessaires
- La toile / Internet
- Besoins de techniques particulières à chaque type de données pour la fouille de données

On ne peut pas tout faire… et tout n'est pas de la FDD

- La FDD offre des outils et des méthodologies qui peuvent aider à comprendre les données et faire des prédictions
- Mais:
  - Il n'y a pas de solution générale et complètement automatique
  - Les techniques doivent être adaptées au problème considéré
  - La FDD doit être comprise
  - La FDD ne remplace pas les experts, mais les assiste

04 Les fonctions de la FDD

Classification des fonctions de la FDD

Les fonctions de la FDD sont classées en 2 catégories:

Description des propriétés des données

Inférence sur les données pour faire des **prédictions**

Les fonctions de la FDD: Classification et Prédiction

- Classification et prédiction
  - Diviser / grouper les instances dans des classes spécifiques pour des prédictions futures
  - Prédire des valeurs (classes) inconnues ou manquantes
- Exemples:
  - Les clients loyaux / les clients non-loyaux
  - Les transactions frauduleuses / les transactions générales
  - Prédire les taux de réponses à une offre
- Algorithmes:
  - Arbres de décision, règles de classification, classification Bayésienne, algorithmes génétiques, algorithme des k plus proches voisins, l'approche Rough Sets, régression linéaire et non linéaire

Les fonctions de la FDD: Regroupement (Cluster)

- Regroupement d'éléments de proche en proche fondé sur leur ressemblance
- Les classes sont inconnues, et sont donc créées
- Exemple:
  - Segmentation du marché
- Algorithmes:
  - K-moyennes
  - Réseaux neuronaux

Les fonctions de la FDD: Associations

- Mise en relation des données
- Exemples:
  - Le panier de la ménagère:
    - Les personnes qui achètent des chips achètent aussi du coca-cola
  - Age(X, “20…29”) et Salaire(X, “20000…29000”) ⇒ Achete(X,”PC”) [support = 2%, confidence = 60%]
- L’algorithme A Priori pour obtenir des règles d’association

Les fonctions de la FDD: Caractérisation et Discrimination

- Caractérisation: Généralise les données d'une classe
- Exemple:
  - Général profil des régions pluvieuses
- Discrimination: Compare les données d'une classe avec une ou plusieurs classes (de comparaison)
  - Gestion des anomalies
- Exemple:
  - Les ventes de logiciels qui ont augmenté de 10% par rapport à celles qui ont augmente de 30%

Évaluation des modèles (1)

- Génération d'un grand nombre de modèles
- Le modèle est-il intéressant?
- Mesures d'intérêt d'un modèle:
  - Nouveau
  - Facile à comprendre
  - Valide sur de nouvelles données (avec une certaine mesure de certitude)
  - Utile
  - Confirme (ou infirme) les hypothèses d'un expert

Évaluation des modèles (2)

- Évaluation d'un modèle
  - subjective (expert)
  - objective (statistiques et structure des modèles)
- Peut-on trouver tous les modèles? (Complétude)
- Peut-on générer seulement les modèles intéressants? (Optimisation)
  - Génération de tous les modèles et filtrage suivant certaines mesures et caractéristiques
    - Non réaliste
  - Générer seulement les modèles vérifiant une condition particulière

Mesure pour l'évaluation des modèles

- Exemple des règles d'association $X \Rightarrow Y$
- Support $(X \Rightarrow Y) = P(X \text{ et } Y)$
- Confidence $(X \Rightarrow Y) = P(Y \mid X) = P(X \text{ et } Y) / P(X)$
- Exemple:

Age(X, "20…29") et Salaire(X, "20000…29000") ⇒ Achète(X,"PC") [support = 2%, confidence = 60%]

Construction et évaluation d'un modèle

- Les données sont séparées en 3 ensembles:
  - Ensemble d'entraiment
  - Ensemble de validation
  - Ensemble test

05

Quelques logiciels de fouille de données

Logiciels

- Logiciels sophistiqués
  - Weka
  - Dbminer
  - Xlminer
  - SQL Server 2000
- Logiciels de programmation avec python (package Python Scikit-learn)
  - Jupyter Notebook
  - Spyder …

Analyse et fouille de données

Data mining

Saoussen Mathlouthi

Saoussen.mathlouthi@ensi-uma.tn

Informatique de gestion: 2LIG

ISG Bizerte

A-U: 2021-2022

Plan

- Un exemple de découverte d'information et de connaissance
- Types de connaissances extraites
- Typologie des méthodes de fouilles de données
- Familles d'approches

01

Un exemple de découverte d'information et de connaissance

Un exemple de découverte d'information et de connaissance

Exemple issu du livre de P. Adriaans et D. Zantige [Adriaans & Zantige 96]

- Un éditeur vend 5 sortes de magazines : sport, voiture, maison, musique et BD
- Il souhaite mieux étudier ses clients pour découvrir de nouveaux marchés ou vendre plus de magazines à ses clients habituels

Quelques questions qu'il peut se poser :

Q1 : Combien de personnes ont pris un abonnement à un magazine de sport cette année ?

Q2 : A-t-on vendu plus d'abonnements de magazines de sport cette année que l'année dernière ?

Q3 : Est-ce que les acheteurs de magazines de BD sont aussi amateurs de sport ?

Q4 : Quelles sont les caractéristiques principales de mes lecteurs de magazines de voiture ?

Q5 : Peut-on prévoir les pertes de clients et prévoir des mesures pour les diminuer ?

Questions de natures différentes mettant en jeu des processus différents

Un exemple de découverte d'information et de connaissance

Q1 : Combien de personnes ont pris un abonnement à un magazine de sport cette année ?

- réalisable en SQL à partir des données opérationnelles sous réserve d'indexations suffisantes des tables concernées
- seule difficulté : ne pas pénaliser le serveur transactionnel par des requêtes trop longues

Q2 : A-t-on vendu plus d'abonnements de magazines de sport cette année que l'année dernière?

- nécessite de conserver toutes les dates de souscription même pour les abonnements résiliés
- l'utilisateur devrait poser de nombreuses requêtes de ce type
- elles devraient être résolues par des requêtes multidimensionnelles de type OLAP.

➢ Q1 et Q2 :

- Réponse par simples requêtes SQL : les données recherchées sont que le résultat d'un calcul simple sur un ou des groupes d'enregistrements
- ce qui distingue Q1 et Q2, c'est la notion de temps

Un exemple de découverte d'information et de connaissance

Q3 : Est-ce que les acheteurs de magazines de BD sont aussi amateurs de sport ?

- exemple simplifié de problème où l'on demande si les données vérifient une règle
- réponse formulée par une valeur estimant la probabilité que la règle soit vraie
- en général des outils statistiques sont utilisés
- cette question peut être généralisée, on pourrait ainsi :
  - chercher des associations fréquentes entre acheteurs de magazine pour effectuer des actions promotionnelles
  - introduire une composante temporelle pour chercher si le fait d'être lecteur d'un magazine implique d'être, plus tard, lecteur d'un autre magazine

Un exemple de découverte d'information et de connaissance

Q4 : Quelles sont les caractéristiques principales de mes lecteurs de magazines de voiture ?

question beaucoup plus ouverte : il s'agit de trouver une règle et non plus de la vérifier ou de l'utiliser

Q5 : Peut-on prévoir les pertes de clients et prévoir des mesures pour les diminuer ?

question ouverte : Il faut disposer d'indicateurs comme : durées d'abonnement, délais de paiement, ...

question (classique dans le bancaire) avec une forte composante temporelle et nécessite des données historiques

c'est pour ce type de question que sont mis en œuvre des outils de fouille de données

02

Types de connaissances extraites

Les données ?

Les données peuvent être vues comme une collection d'objets (enregistrements) et leurs attributs.

- Un attribut est une propriété et/ou une caractéristique de l'objet.
- Un ensemble d'attributs décrit un objet.

Attributs

Objets

	petal_length	petal_width	species
0	1.4	0.2	0
1	1.4	0.2	0
2	1.3	0.2	0
3	1.5	0.2	0
4	1.4	0.2	0
5	1.7	0.4	0
6	1.4	0.3	0
7	1.5	0.2	0
8	1.4	0.2	0
9	1.5	0.1	0
10	1.5	0.2	0

Les données

Attribut - valeur

- La valeur d'un attribut est un nombre ou un symbole.
- Ne pas confondre attribut et valeur

Types

- **Quantitative** (numérique, exprime une quantité)
  - Discrète (ex : nombre d'étudiants dans un cours) ou continue (ex : longueur)
  - Echelle proportionnelle (chiffre d'affaires, taille), ou échelle d'intervalle (température)
- **Qualitative**
  - Variable ordinale (classement à un concours, échelle de satisfaction client)
  - Variable nominale (couleur de yeux, diplôme obtenu, sexe)
- Les **modalités** d'une variable sont l'ensemble des valeurs qu'elle prend dans les données

ex : les modalités de notes sont {0; 1; 2; ; 20} les modalités de couleur sont {bleu,vert,noir,...}

Types de connaissances extraites

Connaissances extraites = connaissances sous la forme de **modèles de description** permettant de :

- décrire le comportement actuel des données et/ou
- prédire le comportement futur des données

- **des analyses** :
  - exemple: la distribution du trafic routier en fonction de l'heure
- **des règles** :
  - exemple : si un client n'a pas payé une facture > 1 000 € alors il est dans 70% des cas en faillite
- **l'attribution de scores de qualité** :
  - par exemple, score de fidélité aux clients
- **la classification d'entités** :
  - par exemple, les mauvais payeurs

03

Typologie des méthodes de fouilles de données

Typologie des méthodes de fouilles de données (1)

Typologie selon l'objectif:

- **Classification** : examiner les caractéristiques d'un objet et lui attribuer une classe.

  e.g. diagnostic ou décision d'attribution de prêt à un client.

- **Prédiction** : prédire la valeur future d'un attribut en fonction d'autres attributs.

  e.g. prédire la qualité d'un client .

- **Association** : déterminer les attributs qui sont corrélés.

  e.g. analyse du panier de la ménagère

- **Segmentation** : former des groupes homogènes à l'intérieur d'une population.

Typologie des méthodes de fouilles de données (2)

Typologie selon le type de modèle obtenu:

- **Modèles prédictifs.**
  - Utilisent les données existantes et des résultats connus sur ces données pour développer des modèles capables de prédire les valeurs d'autres données.
  - e.g. Prédire les clients qui ne rembourseront pas leur crédit.
  - Utilisés principalement en classification et prédiction.
- **Modèles descriptifs.**
  - Proposent des descriptions de données pour aider à la prise de décision.
  - Souvent en amont de la construction de modèles prédictifs.
  - Utilisés principalement en segmentation et association.

Typologie des méthodes de fouilles de données (3)

Typologie selon le type d'apprentissage utilisé:

- **Apprentissage supervisé** : fouille supervisée
  - Processus qui prend en entrée des exemples d'apprentissage contenant à la fois des données d'entrée et de sortie.
  - Les exemples d'apprentissage sont fournis avec leur classe.
  - But : classer correctement un nouvel exemple.
  - Utilisés principalement en classification et prédiction.
- **Apprentissage non supervisé** : fouille non supervisée
  - Processus qui prend en entrée des exemples d'apprentissage contenant que des données d'entrée
  - Pas de notion de classe
  - But : regrouper les exemples en paquets (clusters) d'exemples similaires.
  - Utilisés principalement en segmentation et association.

Analyse et fouille de données

Data mining

Saoussen Mathlouthi

Saoussen.mathlouthi@ensi-uma.tn

Informatique de gestion

ISG Bizerte

A-U: 2021-2022

04

Familles d'approches

- Classification
- Segmentation
- Association

04

Familles d'approches

- **Classification**
- **Segmentation**
- **Association**

Classification

Examiner les caractéristiques d'un objet et lui attribuer une classe (un champ particulier à valeurs discrètes).

- Etant donnée une collection d'enregistrements (ensemble d'apprentissage).
  - Chaque enregistrement contient un ensemble d'attributs et un de ces attributs est sa classe.
- Trouver un modèle pour l'attribut classe comme une fonction de la valeurs des autres attributs
- But : permettre d'assigner une classe à des enregistrements inconnus de manière aussi précise que possible.
- Un ensemble de test est utilisé pour déterminer la précision du modèle.

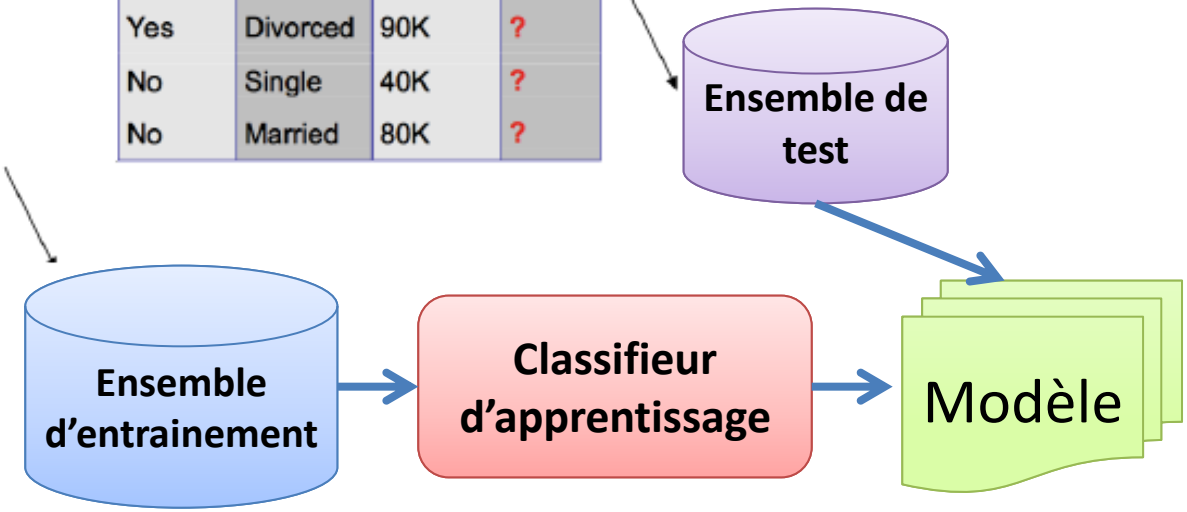
Classification

Exemple:

Les attributs de chaque individu sont: le remboursement (refund), l'état civil (marital status), le revenu imposable (texable income) et la classe fraude (cheat)

	categorical	categorical	continuous	class
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?



Classification

Classification : exemples d'applications

Marketing direct

- But : réduire le coût du mailing en ciblant un ensemble de consommateurs qui achèteront vraisemblablement un nouveau téléphone portable.
- Approche :
  - Utiliser des données pour un produit similaire.
  - On sait quels consommateurs ont acheté. La décision (Achat - Pas achat) est l'attribut classe.
  - Collecter diverses informations sur ce type de consommateurs.
  - Cette information représente les entrées du classifieur.

04

Familles d'approches

- Classification
- **Segmentation**
- Association

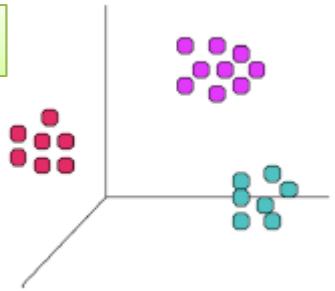
Segmentation

Former des groupes homogènes à l'intérieur d'une population

- Etant donné un ensemble de points, chacun ayant un ensemble d'attributs, et une mesure de similarité définie sur eux, trouver des groupes tels que :
  - Les points à l'intérieur d'un même groupe sont très similaires entre eux.
  - Les points appartenant à des groupes différents sont très dissimilaires.
- Le choix de la mesure de similarité est important.

Regroupement (clustering) basé sur la distance Euclidienne

Distances intra-groupe minimisées



Distances inter-groupe maximisées

Segmentation

Segmentation: exemples d'application

- Segmentation de marchés .
- Segmentation de documents.
- ...

04

Familles d'approches

- Classification
- Segmentation
- **Association**

Association

Entrée : Un ensemble de tickets de caisse

- Une observation = un caddie, un ticket de caisse.
- Non prise en compte de la fréquence des produits.
- Un grand nombre de produits, un grand nombre de caddies (petit sous ensemble de l'ensemble de produits).

TID	Elements
1	Pain, coca, lait
2	Jus, pain
3	Jus, coca, fromage, lait
4	Jus, pain, fromage, lait
5	Coca, fromage, lait

Sortie : Des règles

règles extraites: {lait} → {coca}

{fromage, lait} → {jus}

Association

Association : exemples d'application

- Marketing et promotions sur des produits.
- Gestion du supermarchés : rayonnage.
- Inventaire.
- ...

Exercices

05

Exercice (1)

- Problème: Améliorer le processus d'acceptation et de rejet des crédits pour diminuer les pertes dues aux défauts de paiements
- Les étapes
  - Définir les données
  - Définir les objectifs
  - Traduire le problème en un problème de fouille de données

Exercice (2)

- Prédire si un client va être en défaut de paiement ou pas
- Prédire les profits escomptés résultant de l'ajustement des défauts de paiement
- Sont à prendre en compte:
  - Le domaine d'application
  - Les données considérées
  - Le(s) modèle(s) que l'on veut générer
  - Les techniques que l'on veut/doit utiliser