Data-Mining

I- Introduction au Data-Mining
II- Text mining
III- Web mining
IV- KNIME

Dr Ghazi Aziz

I.Introduction au

Data-Mining

Data-Mining : introduction

Data-mining ≡ Fouille de données

Regroupe un ensemble de techniques et d'outils de la Statistique, l'Informatique et la Science de l'information

A évolué vers la science des donnéesMachine Learning, Data-Mining Big Data (explosion des données) Formalismes de stockage et de traitement distribués des données(NoSQL, Hadoop, MapReduce, Spark ...)

Historique des techniques de statistique et de data mining

1875 Régression linéaire de Francis Galton.
1896 Formule du coefficient de corrélation de Karl Pearson2.
1900 Distribution du X^2 de Karl Pearson.
1936 Analyse discriminante de Fischer et Mahalanobis
1941 Analyse factorielle des correspondances de Guttman
1943 Réseaux de neurones de Mac Culloch et Pitts
1944 Régression logistique de Joseph Berkson
1958 Perceptron de Rosenblatt
1962 Analyse des correspondances de J.-P. Benzécri
1962 Régression logistique de J. Cornfield
1964 Arbre de décision AID de J.-P. Sonquist et J.-A. Morgan
1965 Méthode des centres mobiles de E. W. Forgy
1967 Méthode des k means (k moyennes) de Mac Queen
1971 Méthode des nuées dynamiques de Diday
1972 Modèle linéaire généralisé de Nelder et Wedderburn
1975 Algorithme génétique de Holland
1977 Méthode de classement DISQUAL de Gilbert Saporta
1980 Arbre de décision CHAID de KASS
1983 Régression PLS de Herman et Svante Wold
1984 Arbre CART de Breichman, Friedman, Olshen, Stone
1986 Perceptron multicouches de Rumelhart et Mac Clelland
1989 Réseaux de T. Kohonen (cartes auto-adaptatives)
1990 Apparition du concept de Data Mining
1993 Arbre C4.5 de J. Ross Quinlan
1996 Bagging (Breiman) et boosting (Freund-Shapire)
1998 Support vector machine de Vladimir Vapnik
2001 Régression logistique PLS de Tenenhaus

QUESTIONS : Exemples

- Combien de clients ont acheté tel produit pendant telle période ? ⟹ SQL
- A-t-on vendu plus d'un tel produit cette année que l'année dernière? ⟹ OLAP
- Quel est le profil des clients ?
- Quels autres produits les intéresseront ?
- Quand seront-ils intéressés ?

⟹ Data Mining

Data-Mining : définition

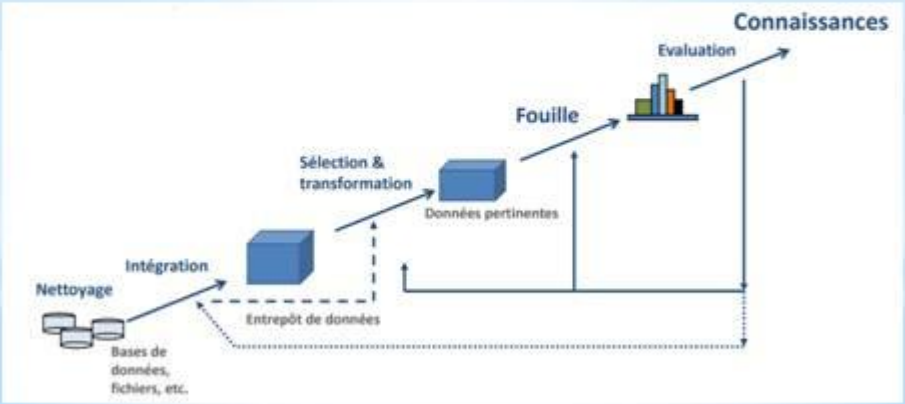
Définition 1

Le data-mining est un processus de découverte de règle, relations, corrélations et/ou dépendances à travers une grande quantité de données, grâce à des méthodes statistiques, mathématiques et de reconnaissances de formes.

Définition 2

Le data-mining est un processus d'extractions automatique d'informationspredictives à partir de grandes bases de données.

Définition : Processus ou méthode qui extrait des connaissances « intéressantes » ou des motifs (patterns) à partir d'une grande quantité de données



Data-Mining : les raisons du développement

Données

Big Data : augmentation sans cesse de donnéesgénérées

- **Twitter : 50M de tweets /jour (=7 téraoctets)**
- **Facebook : 10 téraoctets /jour**
- **Youtube : 50h de vidéos uploadées /minute**
- **2.9 million de mail /seconde**

Puissance de calcul

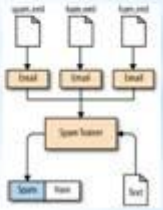
- **Loi de Moore**
- **Calcul massivement distribué**

Création de valeur ajoutée

- **Intérêt : du produit aux clients.**
- **Extraction de connaissances des big data**

Exemples d'applications

- **Entreprise et Relation Clients** : création de profils clients, ciblage de clients potentiels et nouveaux marchés
- **Finances** : minimisation de risques financiers
- **Bioinformatique** : analyse du génome, mise au point de médicaments
- **Internet** : spam, e-commerce, détection d'intrusion, recherche d'informations ...
- **Sécurité**
- ...



Exemples d'applications : E-commerce

Targeting (ciblage)

Stocker les séquences de clicks des visiteurs, analyser les caractéristiques des acheteurs

Faire du "targeting" lors de la visite d'un client potentiel

Systèmes de recommandation

Opportunité : les clients notent les produits ! Comment tirer profit de ces données pour proposer des produits à un autre client ?

Solutions : technique dit de filtrage collaboratif pour regrouper les clients ayant les mêmes "goûts".

Exemples d'applications : Analyse des risques

Détection de fraudes pour les assurances

- Analyse des déclarations des assurés par un expert afin d'identifier les cas de fraudes.
- Applications de méthodes statistiques pour identifier les déclarations fortement corrélées à la fraude.

Prêt Bancaire

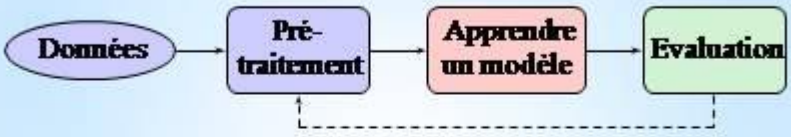
- Objectif des banques : réduire le risque des prêts bancaires.
- Créer un modèle à partir de caractérisques des clients pour discriminer les clients à risque des autres.

Exemples d'applications : Commerce

Opinion mining

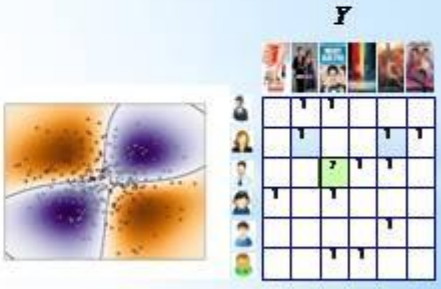
Exemple : analyser l'opinion des usagers sur les produits d'une entreprise à travers les commentaires sur les réseaux sociaux et les blogs

Mise en œuvre d'un projet de DM



Principales étapes

1. Collecte de données
2. Pré-traitement
3. Analyse statistique
4. Identifier le problème de DM
5. Apprendre le modèle mathématique
6. Évaluer ses capacités



Ensemble de données

Données si un problème de DM

Les informations sont des exemples avec des attributs

On dispose généralement d'un ensemble de N données

Attributs

Un attribut est un descripteur d'une entité. On l'appelle également variable, ou caractéristique

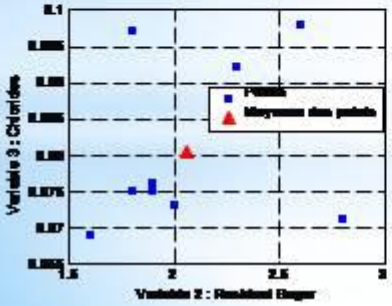
Exemple

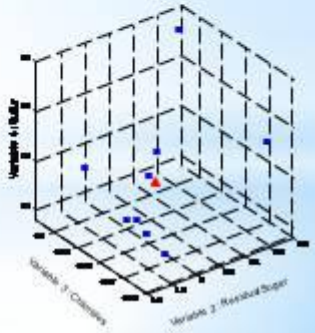
C'est une entité caractérisant un objet ; il est constitué d'attributs.

Synonymes : point, vecteur (souvent dans R^d)

Données: illustration

Points x / Variables	citric acid	residual sugar	chlorides	sulfur dioxide
1	0	1.9	[illegible]	11
2	0	2.6	[illegible]	25
3	0.04	2.3	[illegible]	15
Point x $\in \mathbb{R}^4$	0.56	1.9	[illegible]	17
5	0	1.9	[illegible]	11
6	0	1.8	[illegible]	13
7	0.06	1.6	[illegible]	15
8	0.02	2	[illegible]	9
9	0.36	2.8	[illegible]	17
10	0.08	1.8	[illegible]	15





11/58

Type de données

Capteurs → variables quantitatives, qualitatives, ordinales
Texte → Chaîne de caractères
Parole → Séries temporelles
Images → données2D
Videos → données2D + temps
Réseaux → Graphes
Flux → Logs, coupons…
Etiquettes → information d'évaluation

Big Data (volume, vélocité, variété)

Flot "continu" de données
→ Pre-traitement des données(nettoyage, normalisation, codage …)
→ Représentation : des donnéesaux vecteurs

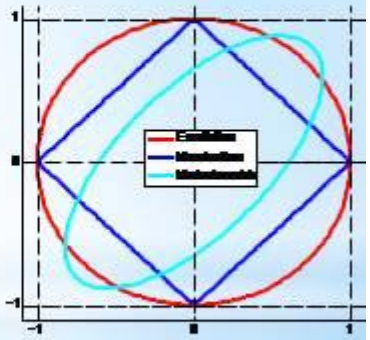
12/58

Donnéeset Métriques

Les algorithmes nécessitent une notion de similarité dans l'espace X des données. La similarité est traduite par la notion de distance.

- distance euclidienne : $x, z \in \mathbb{R}^d$, on a
  $d(x,z) = \|x - z\|_2 = \sqrt{\sum_{j=1}^{d}(x_j - z_j)^2} = \sqrt{(x-z)^\top (x-z)}$
- distance de manhattan
  $d(x,z) = \|x - z\|_1 = \sum_{j=1}^{d} |(x_j - z_j)|$
- distance de mahalanobis
  $d(x,z) = \sqrt{(x-z)^\top \Sigma^{-1} (x-z)}$

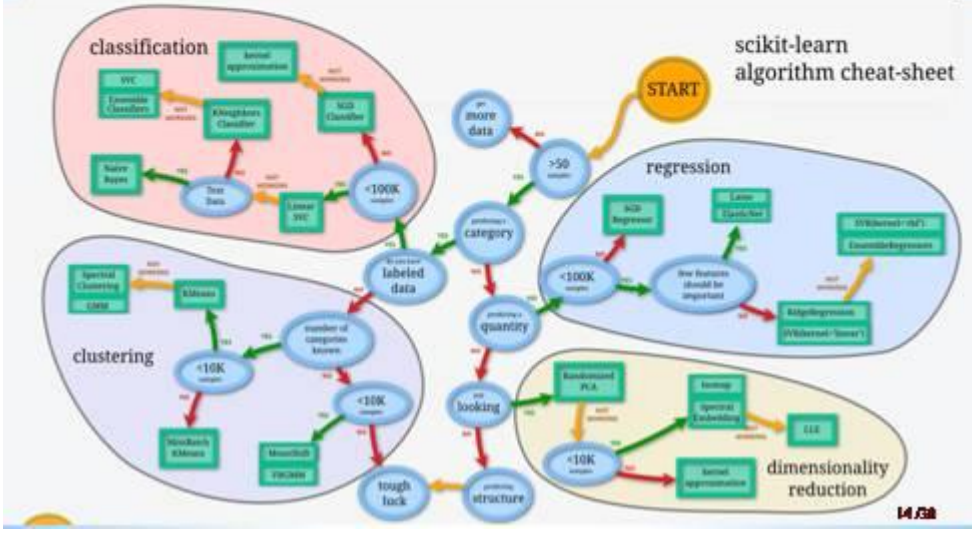
  $\Sigma \in \mathbb{R}^{d \times d}$: matrice carrée définie positive



Caractérisation des méthodes de Data-Mining

Types d'apprentissage:

- **Apprentissage supervisé**
- **Apprentissage non-supervisé**
- **Apprentissage semi-supervisé**



Récapitulatif

1. Donnez une définition au DM
2. Donnez qq applications du DM
3. Donnez le schéma de mise en œuvre d'un projet DM
4. Donnez qq type de donnéesutilisés en DM
5. Donnez la formule de calcul de la distance euclidienne ? De Manhattan ? Et Mahalanobis?
6. Donnez l'objectif du ML supervisé et donnez qq algorithmes
7. Donnez l'objectif du ML non supervisé et donnez qq algorithmes
8. Donnez l'objectif du ML semi supervisé et donnez qq algorithmes

Tâches du datamining

1. Classification
2. Estimation
3. Prédiction
4. Segmentation (clustering)
5. Association
6. Description

Tâches du Data Mining : Exemples

Classification : Déterminer grade en fonction de l'âge, l'ancienneté, le salaire et les affectations.

Estimation : (sur Variables continues)

Estimer le salaire en fonction de l'âge, ancienneté et affectations

Prédiction : Prédire quelle sera la prochaine affectation d'un militaire.

Association: Déterminer des règles de type :

Le militaire qui est sergent entre 25 et 30 ans sera lieutenant colonel entre 45 et 50 ans (fiable à n %).

Segmentation (ou clustering) :

Segmenter les militaires en fonction de leurs parcours (carrière) et affectations.

Description : Indicateurs statistiques traditionnels :

Age moyen, pourcentage de femmes, salaire moyen

Les types de modèles

En plus des différents types de référentiels de donnéeset d'informations sur lesquels l'exploitation peut être effectuée, les types de modèles peuvent être également exploités. Il existe un certain nombre de fonctionnalités d'exploration de données:

- La caractérisation et discrimination
- L'exploration de modèles fréquents, d'associations et de corrélations
- Classification et régression
- Analyse de regroupement
- Analyse des valeurs aberrantes

1. Description du concept : Caractérisation et discrimination

Les entrées de donnéespeuvent être associées à des classes ou à des concepts. De telles descriptions d'une classe ou d'un concept sont appelés des descriptions de classe/concept. Ces descriptions peuvent être dérivées en utilisant :

- La caractérisation des données: en résumant les donnéesde la classe étudiée appelée souvent la classe cible en termes généraux
- La discrimination des données: par comparaison de la classe cible avec une ou un ensemble de classes comparatives appelées souvent classes
- Hybride : à la fois la caractérisation et la discrimination des données.

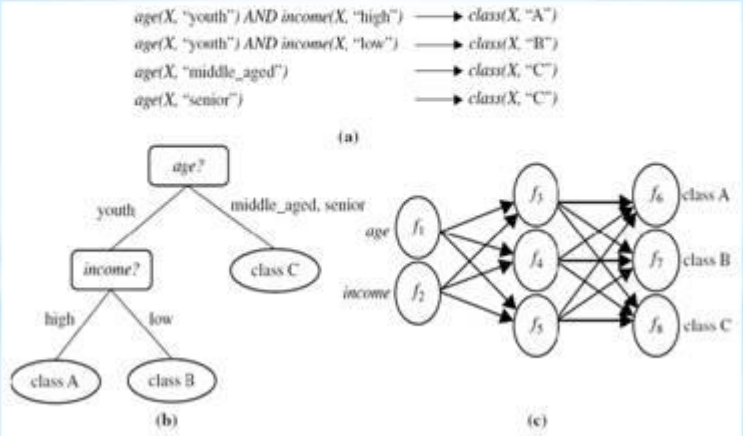
2. Extraction de modèles fréquents, d'associations et de corrélations

Les modèles fréquents sont des modèles qui se produisent fréquemment dans les données. Il existe de nombreux types de modèles fréquents :

- Des ensembles d'éléments fréquents : fait généralement référence à un ensemble d'éléments qui apparaissent souvent ensemble dans un ensemble de donnéestransactionnelles, par exemple, le lait et le pain, qui sont fréquemment achetés ensemble dans les épiceries par de nombreux clients.
- Des sous-séquences fréquentes : également appelées modèles séquentiels telle que le modèle selon lequel les clients ont tendance à acheter d'abord un ordinateur portable, suivi d'un appareil photo numérique, puis d'une carte mémoire, est un modèle séquentiel (fréquent)
- Des sous-structures fréquentes : peut faire référence à différentes formes structurelles (par exemple, des graphiques, des arbres ...) qui peuvent être combinéesavec des ensembles d'éléments ou des sous-séquences. Si une sous-structure se produit fréquemment, on l'appelle un modèle structuré (fréquent).

3. Classification et régression pour l'analyse prédictive

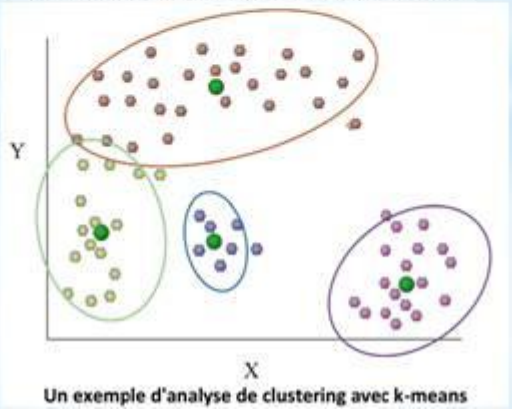
La classification est le processus de recherche d'un modèle (ou d'une fonction) qui décrit et distingue des classes de donnéesou des concepts. Le modèle dérivé peut être représenté sous diverses formes, telles que des règles de classification (c'est-à-dire des règles SI-ALORS), des arbres de décision, des formules mathématiques ou des réseaux neuronaux (Figure). Un test sur une valeur d'attribut de chaque branche représente un résultat du test et les feuilles de l'arbre représentent des classes ou des distributions de classes.



Modèles de classification : (a) Des règles SI-ALORS, (b) Un arbre de décision ou (c) Un réseau de neurones.

4. Analyse de cluster

Contrairement à la classification et à la régression, qui analysent des ensembles de données(d'apprentissage) étiquetés par classe, le clustering analyse les objets de donnéessans consulter les étiquettes de classe. Dans de nombreux cas, les donnéesétiquetées par classe peuvent tout simplement ne pas exister au début. Le clustering peut être utilisé pour générer des étiquettes de classe pour un groupe de données. Les objets sont regroupés selon le principe de maximisation de la similarité intraclasse et de minimisation de la similarité interclasse.



Un exemple d'analyse de clustering avec k-means

5. Analyse des valeurs aberrantes

Un ensemble de donnéespeut contenir des objets qui ne sont pas conformes au comportement général ou au modèle des données. Ces objets de donnéessont des valeurs aberrantes.
De nombreuses méthodes d'exploration de donnéesrejettent les valeurs aberrantes comme du bruit ou des exceptions.
Cependant, dans certaines applications (par exemple, la détection de fraude), les événements rares peuvent être plus intéressants que ceux qui se produisent plus régulièrement.
L'analyse des donnéesaberrantes est appelée analyse des valeurs aberrantes ou extraction d'anomalies.
Les valeurs aberrantes peuvent être détectées à l'aide de tests statistiques qui supposent un modèle de distribution ou de probabilité pour les données, ou à l'aide de mesures de distance où les objets éloignés de tout autre cluster sont considérés comme des valeurs aberrantes

Exercice 1 : K-Means

Soit l'ensemble D des entiers suivants : D= { 2, 5, 8, 10, 11, 18, 20 } On veut répartir les donnéesde D en trois (3) clusters, en utilisant l'algorithme Kmeans. La distance d entre deux nombres a et b est calculée ainsi : d(a , b) = |a - b| (la valeur absolue de a moins b) Travail à faire :
1/ Appliquez Kmeans en choisissant comme centres initiaux des 3 clusters respectivement : 8, 10 et 11. Montrez toutes les étapes de calcul. -Premier clusters : C1={ 2, 5, 8} C2={10} C3={11, 18, 20} -
2/ Donnez le résultat final et précisez le nombre d'itérations qui ont été nécessaires.

Correction Exercice 1 :

1)

- **Itération 1 :**

Mise à jour des clusters : C1={ 2, 5, 8} C2={10} C3={11, 18, 20}
R-estimation des centres de gravité : μ1= (2+5+8)/3 μ2=10/1 μ3=(11+18+20)/3
=> μ1=5 μ2=10 μ3=16.33

- **Itération 2 :**

Mise à jour des clusters : C1={ 2, 5} C2={8, 10, 11} C3={18, 20}
R-estimation des centres de gravité : μ1= (2+5)/2 μ2=(8+10+11)/3 μ3=(18+20)/2
=>μ1=3.5 μ2=9.66 μ3=19

- **Itération 3**

Mise à jour des clusters : C1={ 2, 5} C2={8, 10, 11} C3={18, 20}
R- estimation des centres de gravité : μ1= (2+5)/2 μ2=(8+10+11)/3 μ3=(18+20)/2
μ1=3.5 μ2=9.66 μ3=19 Stabilité : Les centres de gravité n'ont pas changé. L'algorithme s'arrête

2) Les clusters résultats : C1={ 2, 5} C2={8, 10, 11} C3={18, 20}
Nombre d'itérations = 3

Exercice 2 : Table de décision

Le tableau suivant contient des données sur les résultats obtenus par des étudiants de Tronc Commun (première année à l'Université). Chaque étudiant est décrit par 3 attributs : Est-il doublant ou non, la série du Baccalauréat obtenu et la mention. Les étudiants sont répartis en deux classes : Admis et Non Admis. On veut construire un arbre de décision à partir des données du tableau, pour rendre compte des éléments qui influent sur les résultats des étudiants en Tronc Commun. Les lignes de 1 à 12 sont utilisées comme données d'apprentissage. Les lignes restantes (de 13 à 16) sont utilisées comme données de tests.

1/ Utiliser les données des lignes de 1 à 12 pour construire l'arbre en utilisant l'algorithme ID3. Montrez toutes les étapes de calcul. Dessinez l'arbre final.

	Doublant	Série	Mention	Classe
1	Non	Maths	ABien	Admis
2	Non	Techniques	ABien	Admis
3	Oui	Sciences	ABien	Non Admis
4	Oui	Sciences	Bien	Admis
5	Non	Maths	Bien	Admis
6	Non	Techniques	Bien	Admis
7	Oui	Sciences	Passable	Non Admis
8	Oui	Maths	Passable	Non Admis
9	Oui	Techniques	Passable	Non Admis
10	Oui	Maths	TBien	Admis
11	Oui	Techniques	TBien	Admis
12	Non	Sciences	TBien	Admis
13	Oui	Maths	Bien	Admis
14	Non	Sciences	ABien	Non Admis
15	Non	Maths	TBien	Admis
16	Non	Maths	Passable	Non Admis

Entropie et gain d'information

- S contient s tuples de classe C pour i = {1, m} S contient s_i tuples de classe Ci pour i = {1..m}
- L'information mesure la quantité d'information nécessaire pour classer un objet classer un objet :

$$I(s_1,s_2,\ldots,s_m) = -\sum_{i=1}^{m} \frac{s_i}{s} \log_2 \frac{s_i}{s}$$

- Entropie d'un attribut A ayant pour valeurs $\{a_1, \ldots, a_v\}$:

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j}+\ldots+s_{mj}}{s} I(s_{1j},\ldots,s_{mj})$$

- Gain d'information en utilisant l'attribut A :

$$Gain(A) = I(s_1,s_2,\ldots,s_m) - E(A)$$

Correction :

1) On remarque que sur les 12 lignes des données d'apprentissage, 8 correspondent à la classe "Admis" et 4 à la classe "Non admis". L'entropie de l'ensemble S (à la racine de l'arbre) est donc égale à :

Entropie (S)= -(8/12) * Log2 (8/12) - (4/12) * Log2 (4/12) = 0.92

Pour connaître quel attribut on doit choisir comme test au niveau de la racine de l'arbre, il faut calculer le gain d'entropie sur chacun des attributs : "Condition", "Série" et "Mention".

Calcul du gain d'entropie sur l'attribut "Condition" :

Gain(S, Condition) = Entropie(S) - 7/12 * Entropie(SOui) - 5/12 * Entropie(SNon)

avec Entropie(SOui)= -5/7*Log2(5/7) - 4/7*Log2(4/7)

et Entropie(SNon)= -0/5*log2(0/5) Gain(S, Condition) =0.34

Calcul du gain d'entropie sur l'attribut "Série" :

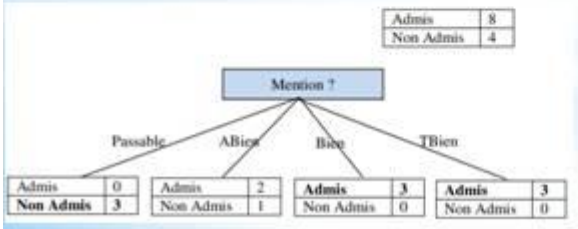
Gain(S, Série) = Entropie(S) - 4/12 * Entropie(SMaths) - 4/12 * Entropie(STechniques) - 4/12 * Entropie(SSciences) Gain(S, Série) = 0.01

Calcul du gain d'entropie sur l'attribut "Mention" :

Gain(S, Série) = Entropie(S) - 3/12 * Entropie(SPassable) - 3/12 * Entropie(SABien) - 3/12 * Entropie(SBien) - 3/12 * Entropie(STBien)

Gain(S, Mention) = 0.69

On constate que le plus grand gain d'entropie est obtenu sur l'attribut "Mention". C'est donc cet attribut qui est choisi comme test à la racine de l'arbre. Nous obtenons l'arbre partiel suivant :



Caractérisation des méthodes : apprentissage supervisé

Objectif

A partir des données $\{(x_i, y_i) \in X \times Y, i = \dots, N\}$, estimer les dépendances entre X et Y.

On parle d'**apprentissage supervisé** car les y_i permettent de guider le processus d'estimation.

Exemples

Estimer les liens entre habitudes alimentaires et risque d'infarctus. x_i : attributs concernant le régime d'un patient, y_i sa catégorie (risque, pas risque).

Applications : détection de fraude, diagnostic médical ...

Techniques

k-plus proches voisins, SVM, régression logistique, arbre de décision ...

Caractérisation des méthodes : Apprentissage non-supervisé

Objectifs

Seules les données $\{x_i \in X, i = \cdots, N\}$ sont disponibles. On cherche à décrire comment les données sont organisées et en extraire des sous-ensemble homogènes.

Exemples

Catégoriser les clients d'un supermarché. x_i représente un individu (adresse, âge, habitudes de courses ...)

Applications : identification de segments de marchés, catégorisation de documents similaires, segmentation d'images biomédicales ...

Techniques

Classification hiérarchique, Carte de Kohonen, K-means, extractions de règles ...

Caractérisation des méthodes : apprentissage semi-supervisé

Objectifs

Objectifs : parmi les données, seulement un petit nombre ont un label i.e $\{(x_1, y_1), \cdots, (x_n, y_n), x_{n+1}, \cdots, N\}$. L'objectif est le même que pour l'apprentissage supervisé mais on aimerait tirer profit des données sans étiquette.

Exemples

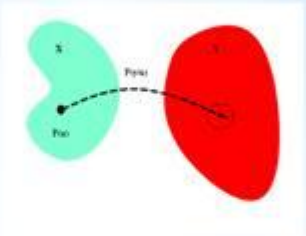
Exemple : pour la discrimination de pages Web, le nombre d'exemples peut être très grand mais leur associer un label (ou étiquette) est coûteux.

Techniques

Méthodes bayésiennes, Self training, Co-Training, SVM...

Apprentissage supervisé : les concepts

Soit deux ensembles X et Y munis d'une loi de probabilité jointe p(X, Y).



Objectifs : On cherche une fonction $f : X \rightarrow Y$ qui à X associe f(X) qui permet d'estimer la valeur y associée à x. f appartient à un espace H appelé **espace d'hypothèses**.

Exemple de H : ensemble des fonctions polynomiales

Apprentissage supervisé : les concepts

On introduit une notion de **coût** L(Y, f(X)) qui permet d'évaluer la pertinence de la prédiction de f, et de pénaliser les erreurs.

L'objectif est donc de choisir la fonction f qui minimise

$$R(f) = E_{X,Y}[L(Y, f(X))] \quad \text{(Espérance)}$$

où R est appelé le **risque moyen** ou **erreur de généralisation**. Il est également noté EPE(f) pour **expected prediction error**

Apprentissage supervisé : les concepts

Exemples de fonction coût et de risque moyen associé.

Coût quadratique (moindres carrés)

$$L(Y, f(X)) = (Y - f(X))^2$$

$$R(f) = E[(Y - f(X))^2] = \int (y - f(x))^2 p(x, y) dxdy$$

Coût (moindres valeurs absolues)

$$L(Y, f(X)) = |Y - f(X)|$$

$$R(f) = E[|Y - f(X)|] = \int |y - f(x)| p(x, y) dxdy$$

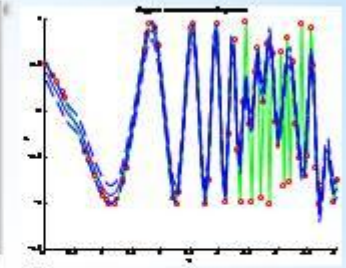
Apprentissage supervisé : les concepts

Régression

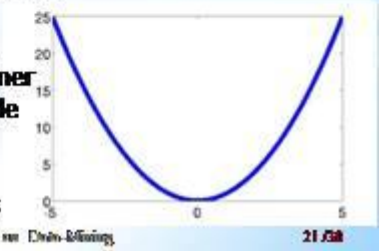
On parle de régression quand Y est un sous-espace de $\mathbb{R}^d$.

Fonction de coût typique :

quadratique $(y - f(x))^2$



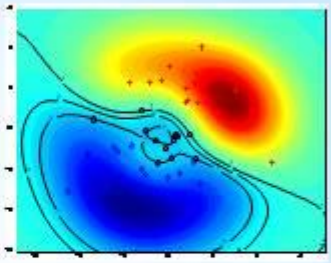
NB : La régression en apprentissage supervisé est une méthode qui permet de prédire une valeur numérique à partir de donnéesd'entrée. Par exemple, on peut utiliser la régression pour estimer le prix d'une maison en fonction de sa surface, de son nombre de pièces, de son quartier, etc. La régression cherche à trouver une fonction qui minimise l'écart entre les valeurs prédites et les valeurs réelles



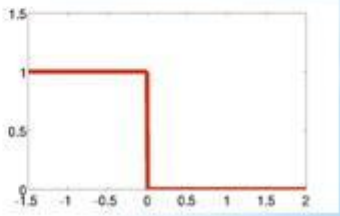
Apprentissage supervisé : les concepts

Discrimination (classification)

si Y est un ensemble discret non-ordonné, (par exemple $\{-1, 1\}$), on parle de discrimination ou classification.



La fonction de coût la plus utilisée est : $\Theta(-yf(x))$ où Θ est la fonction échelon.



Apprentissage supervisé : les concepts

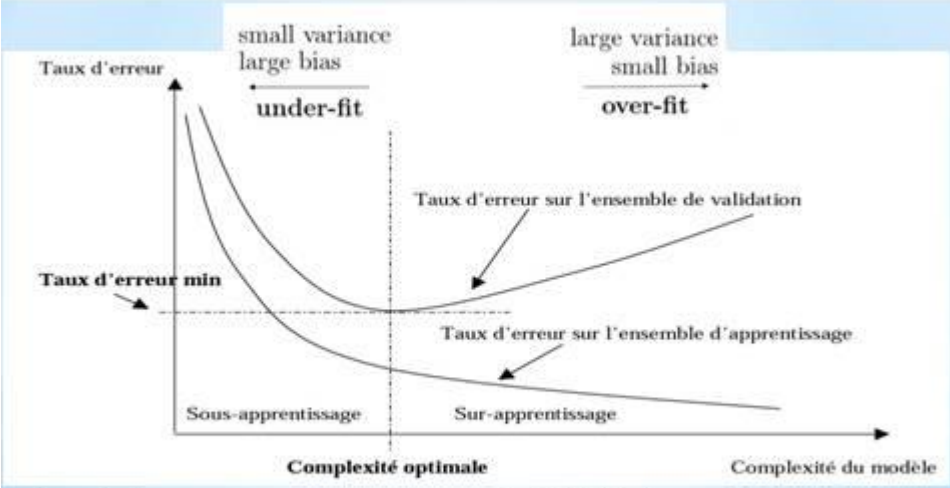
- En pratique, on a un ensemble de données $\{(x_i, y_i) \in X \times Y\}_{i=1}^{N}$ appelé **ensemble d'apprentissage** obtenu par échantillonnage indépendant de $p(X, Y)$ que l'on ne connaît pas.

On cherche une fonction f, appartenant à H qui minimise le **risque empirique** :

$$R_{emp}(f) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, f(x_i))$$

- Le risque empirique ne permet pas d'évaluer la pertinence d'un modèle car il est possible de choisir f de sorte que le risque empirique soit nul mais que l'erreur en généralisation soit élevée. On parle alors de **sur-apprentissage**.

Illustration du sur-apprentissage et sous-apprentissage



Il s'agit de trouver un compromis entre la fiabilité du modèle sur l'ensemble d'apprentissage et la généralisation du modèle : trouver le juste milieu entre sous et sur -apprentissage.

Sélection de modèles

Problématique

On cherche une fonction f qui minimise un risque empirique donné. On suppose que f appartient à une classe de fonctions paramétrées par α. Comment choisir α pour que f minimise le risque empirique et généralise bien ?

Exemple : On cherche un polynôme de degré α qui minimise un risque $R_{emp}(f_\alpha) = \frac{1}{N}\sum_{i=1}^{N}(y_i - f_\alpha(x_i))^2$.

Objectifs :

1. **proposer une méthode d'estimation d'un modèle afin de choisir (approximativement) le meilleur modèle appartenant à l'espace hypothèses.**
2. **une fois le modèle choisi, calculer son erreur de généralisation.**

Sélection de modèles : approche classique

Cas idéal : les données D_N abondent (N est très grand)

1. Découper aléatoirement $D_N = D_{app} \cup D_{val} \cup D_{test}$
2. Apprendre chaque modèle possible sur D_{app}
3. Evaluer sa performance en généralisation sur D_{val}

$$R_{val} = \frac{1}{N_{val}} \sum_{i \in D_{val}} L(y_i, f(x_i))$$

4. Sélectionner le modèle qui donne la meilleure performance sur D_{val}
5. Tester le modèle retenu sur D_{test}

Remarque

D_{test} n'est utilisé qu'une seule fois !

Sélection de modèles : Validation Croisée

Cas moins favorable : les données D_N sont modestes (N est petit)

Estimation de l'erreur de généralisation par rééchantillonnage.

Principe

1. Séparer les N données en K ensembles de part égales.
2. Pour chaque $k = 1, \cdots, K$, apprendre un modèle en utilisant les $K-1$ autres ensemble de données et évaluer le modèle sur la k-ième partie.
3. Moyenner les K estimations de l'erreur obtenues pour avoir l'erreur de validation croisée.

Sélection de modèles : Validation Croisée (2)

Détails :

$$R_{CV} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{N_k}\sum_{i=1}^{N_k} L(y_i^k, f^{-k}(x_i^k))$$

où f^{-k} est le modèle f appris sur l'ensemble des données sauf la k-ième partie.

Propriétés : Si K = N, CV est approximativement un estimateur sans biais de l'erreur en généralisation. L'inconvénient est qu'il faut apprendre N − 1 modèles.

typiquement, on choisit K = 5 ou K = 10 pour un bon compromis entre le biais et la variance de l'estimateur.

Conclusions

Pour bien mener un projet de DM

- Identifier et énoncer clairement les besoins.
- Créer ou obtenir des données représentatives du problème.
- Identifier le contexte de l'apprentissage.
- Analyser et réduire la dimension des données.
- Choisir un algorithme et/ou un espace d'hypothèses.
- Choisir un modèle en appliquant l'algorithme aux données prétraitées.
- Valider les performances de la méthode.

Etapes de Traitement des données:

- Nettoyage des données
- Transformation des données
- Sélection des données
- Réduction des données

Techniques d'Extraction de connaissances

- Techniques descriptives :
- Techniques prédictives :

Nettoyage des données

Objectif :

Supprimer les données bruitées ou non pertinentes.

Questions :

- **Que faire si certaines données sont manquantes?**
  - **Certains clients n'ont pas donné leur adresse.**
- **Toutes les données sont-elles fiables (problèmes d'inconsistance)?**
  - **Un même article appartient à différentes catégories (dans des magasins différents).**
  - **Le prix d'un même article est très supérieur à la normale dans un magasin donné.**
- **Que faire si certaines données sont numériques dans le cas où la technique d'extraction ne peut manipuler que des données symboliques?**

Données manquantes

Solutions :

- Ne pas tenir compte des tuples contenant des données manquantes (valeurs nulles).
- Remplir manuellement les champs non remplis.
- Utiliser les valeurs connues :
  - Remplacer un salaire manquant par le salaire médian des clients.
  - Prédire les valeurs manquantes, en le déduisant d'autres paramètres (salaire à partir de l'âge et de la profession).

Données bruitées

Plusieurs solutions : lissage, segmentation, régression linéaire.

Techniques de lissage (data smoothing) :

1. Trier les différentes valeurs de l'attribut considéré.
   {4, 8, 15, 21, 21, 24, 25, 28, 34}
2. Partitionner l'ensemble résultat.
   {{4, 8, 15}, {21, 21, 24}, {25, 28, 34}}
3. Remplacer les valeurs initiales par de nouvelles valeurs en fonction du partitionnement réalisé :
   - par la valeur moyenne des regroupements réalisés
     {9, 22, 29}
   - par les min et max des regroupements réalisés.
     {{4, 4, 15}, {21, 21, 24}, {25, 25, 34}}

Implique une perte de précision ou d'information.

Données bruitées

Techniques de segmentation (clustering) :

- Les valeurs similaires sont placées dans une même classe.
- On ne tient pas compte des valeurs isolées (dans une classe comportant trop peu d'éléments).

Techniques de régression linéaire :

- Hypothèse : un attribut Y dépend linéairement d'un attribut X.
  - Années d'expérience X et salaire Y.
- Trouver les coefficients a et b tels que $Y = aX + b$.
- Remplacer les valeurs de Y par celles prédites.

Données bruitées : régression linéaire

Données de départ :
Un ensemble de couples (X_i, Y_i).

Détermination des coefficients :

- Soient $\overline{X}$ et $\overline{Y}$ les valeurs moyennes des attributs X et Y.
- $a = \frac{\mathrm{cov}(x, y)}{V(x)}$.
- $b = \overline{Y} - a\overline{X}$.

Données inconsistantes

Données inconsistantes dans une base de données :

- Contraintes d'intégrités ou dépendances fonctionnelles non respectées.
- Exemples :
  - La contrainte I_ID→I_CATEGORY n'est pas respectée au moment de l'intégration des données.
  - Unicité de clés non respectée.

Intégration des données

Objectif :

Regrouper les données provenant de différentes sources.

→ Problématique typique lors de la construction d'entrepôts de données.

Exemple :

Un attribut nommé C_ID dans la BD de Paris peut très bien se nommer CUST_ID dans la BD de Londres.

→ Utilisation de méta-données (XML) pour la mise en correspondance.

Transformation des données

- **Lissage de données** : utilisation de techniques de régression.
- **Normalisation des données** : normaliser certains attributs numériques afin qu'ils varient entre 0 et 1.
  - Pour ne pas privilégier les attributs ayant les plus grands domaines de variation (salaire/âge).
- **Agrégation des données** : opérations OLAP (On-Line Analytical Processing) permettant une analyse multidimensionnelle sur les BD volumineuses afin de mettre en évidence une analyse particulière des données.
  - Calculer les niveaux de ventes réalisées de tel produit par mois plutôt que par jour.
- **Généralisation des données** : remplacer les données finies par des données de plus haut niveau.
  - Remplacer les adresses precises des clients par leur code postal.
  - Remplacer l'âge des clients par « jeune », « adulte », « sénior ».

Discrétisation des connaissances

Répartition des valeurs des attributs :

À chaque étape, on cherche à découper l'intervalle de variation des données en K intervalles comportant le même nombre de valeurs.

On divise C_AGE=[0, 100] en A_1 =[0, 20] et A_2 =[20, 100] si 50 % des clients ont moins de 20 ans.

Entropie et classification à priori des données :

On cherche à caractériser les individus achetant les différents types de lait (entier, demi-écrémé, écrémé).

Facilité à appréhender le découpage obtenu :

On veut obtenir des intervalles du type [−12.5, 0] plutôt que [−12.536, 0.0005].

Discrétisation basée sur l'entropie (1/2)

Entropie d'un ensemble de données S :

Définition :

- **S est découpé en k classes $C_1, \ldots, C_k$.**
- $Ent(S) = \sum p_i . \log(p_i)$ avec $p_i = \frac{|C_i|}{|S|}$.

Propriétés :

- **Ent(S) est maximale (égale à 0) si les données sont réparties dans une seule et même classe.**
- **Ent(S) est minimale si les données sont uniformément réparties dans toutes les classes.**

Discrétisation basée sur l'entropie (2/2)

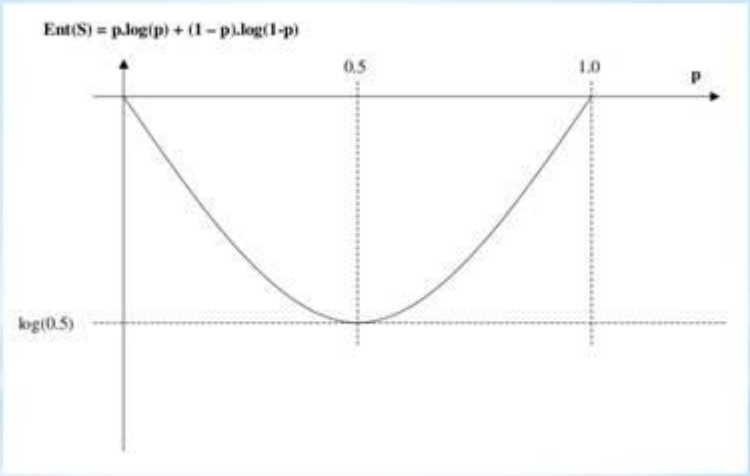
Méthode :

- **Découper S =[a, b] en S_1 =[a, c] et S_2 =[c , b].**
- **Maximiser le gain d'information**

$$I(S, c) = \frac{|S_1|}{|S|} Ent(S_1) + \frac{|S_2|}{|S|} Ent(S_2) - Ent(S).$$

- **Arrêt du découpage si le gain devient insuffisant, quel que soit c.**

Variation de l'entropie



Sélection des données

Objectif :

Garder uniquement les données pertinentes pour l'étude réaliser.

Exemple :

- **Doit-on s'intéresser à toutes les catégories de produits de vente ?**
- **Doit-on s'intéresser aux ventes réalisées il y a plus d'un an ?**

Réduction des données

Réduction en ligne par échantillonnage :

- Pour des raisons de performance.
- Du fait de la complexité importante des algorithmes d'extraction.
- Plusieurs méthodes : échantillonnage aléatoire (avec ou sans remise), échantillonnage par clustering/segmentation.

Réduction en colonne par suppression des attributs redondants :

- Cas triviaux (âge et date de naissance).
- Via une analyse des corélation entre attributs :

$$corr_{A,B} = \frac{P(A \wedge B)}{P(A).P(B)} = \frac{P(B/A)}{P(B)}$$

- Indépendance : $corr_{A,B} = 1$ si $P(B/A) = P(B)$.
- Corrélation positive : $corr_{A,B} > 1$ si $P(B/A) > P(B)$.

Réduction des données
Matrice de contingence

Exemple de matrice de contingence :

	Avec pain	Sans pain	Total
Avec beurre	4.000	3.500	7.500
Sans beurre	2.000	500	2.500
Total	6.000	4.000	10.000

Analyse de corélation :

- $P(Beurre) = \frac{7.500}{10.000} = 0.75$ et $P(Pain) = 0.6$.
- $P(Beurre \wedge Pain) = \frac{4.000}{10.000} = 0.4$.
- $corr_{Pain,Beurre} = \frac{0.4}{0.75 \times 0.6} = 0.89 < 1$
  $\rightarrow$ Indique une corrélation négative.

NB: En logique classique « ¬ » (non), « ∨ » (ou), et ∧ (et)

Réduction des données

Qualité de la corrélation

Coefficient de corrélation :

$$r_{A,B} = \frac{\sum(A_i - \bar{A})(B_i - \bar{B})}{\sigma_A \cdot \sigma_B}$$

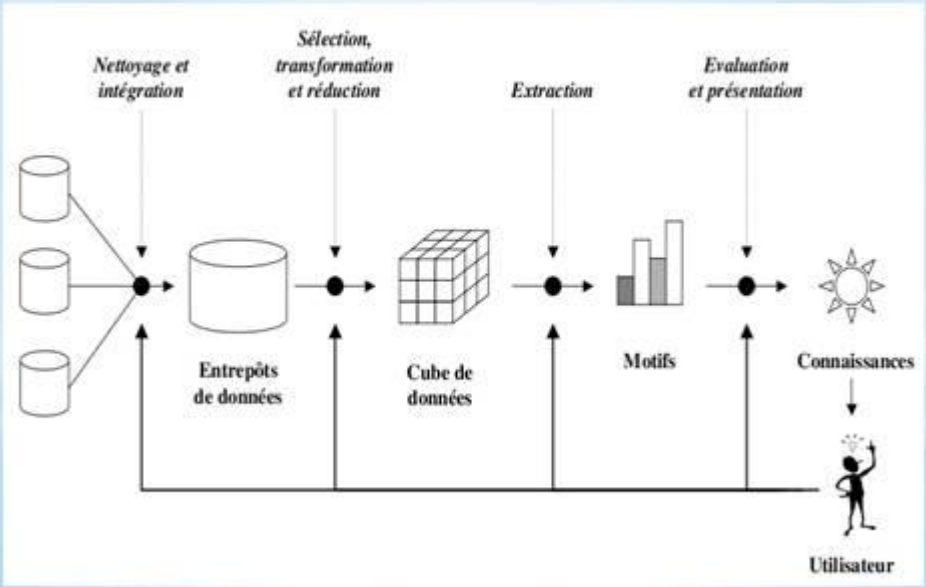
avec $\sigma_X = \sqrt{\sum(X_i - \bar{X})^2}$.

Signification :

Plus $r_{A,B}$ s'éloigne de zéro, meilleure est la corrélation :

- $r_{A,B} = +1$: corrélation positive parfaite.
- $r_{A,B} = -1$: corrélation négative parfaite.
- $r_{A,B} = 0$: absence totale de corrélation.

Extraction de connaissances



Extraction de connaissances (1/2)

Techniques descriptives :

- Visent à mettre en évidence des informations présentes, mais cachées dans les gros volumes de données.
  Cas de la segmentation de la clientèle, de la recherche d'association de produits sur les tickets de caisse.
- Permettent de réduire, de résumer et de synthétiser les données.
- Pas de variable « cible » à prédire.

Exemples :

- Techniques de segmentation/clustering : nuées dynamiques, segmentation hiérarchique, réseaux de neurones.
- Extraction de règles d'association.

Extraction de connaissances (2/2)

Techniques prédictives :

- Visent à extrapoler de nouvelles informations à partir des informations présentes.
  Cas général du scoring (impayés, attrition, crédit).
- Permettent d'« expliquer » les données.
- Il existe une variable « cible » à prédire.

Exemples :

- Classification/discrimination (variable cible qualitative) :
  - analyse discriminante ;
  - arbres de classification ;
  - réseaux neuronaux multi-couches.
- Prédiction (variable cible quantitative) :
  - régression linéaire (simple et multiple) ;
  - arbres de régression.

Présentation des connaissances

Problème :
Comment représenter/visualiser les connaissances extraites?

Formules logiques :

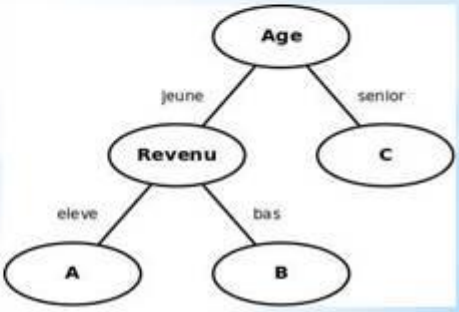
- Age(X, 'jeune'), Revenu(X, 'eleve') → class(X, 'A') [1.402].
- Age(X, 'jeune'), Revenu(X, 'bas') → class(X, 'B') [1.038].
- Age(X, 'senior') → class(X, 'C') [2.160].

Présentation des connaissances

Tableau :

Age	Revenu	Class	Count
jeune	eleve	A	1.402
jeune	bas	B	1.038
senior	eleve	C	786
senior	bas	C	1.374

Arbre de décision :



Les logiciels de data mining

Il existe de nombreux logiciels de statistiques et de data mining.

- Parmi les gros logiciels, on peut citer :
  - Clementine de SPSS. Clementine est la solution de data mining la plus vendue dans le monde.
  - Entreprise Miner de SAS.
  - Statistica Data Miner de StatSoft
  - XL Miner (data mining sous excel)
  - ORACLE, comme d'autres SGBD, fournit des outils de data mining
- Parmi les logiciels gratuits, on peut citer :
  - TANAGRA, logiciel de data mining gratuit pour l'enseignement et la recherche.
  - ORANGE, logiciel libre d'apprentissage et de data mining.
  - WEKA, logiciel libre d'apprentissage et de data mining.

OUTILS de Data Mining

TANAGRA

Logiciel gratuit, destiné à l'enseignement et à la recherche. Open source.

rapidminer

Open source à la fois gratuit et commercial.

WEKA
The University of Waikato

Écrit en Java et développé à l'université de Waikato, Nouvelle-Zélande.
Logiciel libre disponible sous la Licence publique générale GNU.

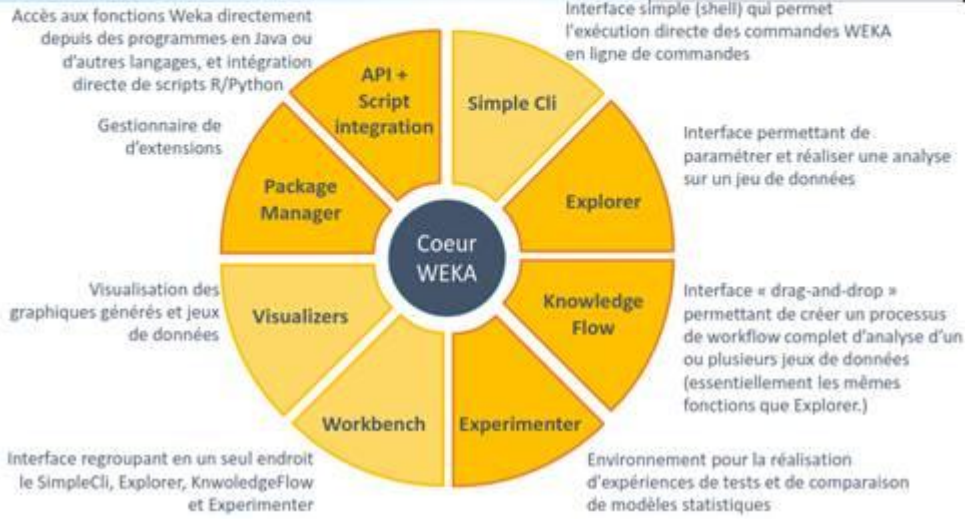
WEKA

- Weka 3: Data Software Mining en Java
- Weka est une collection d'algorithmes d'apprentissage machine pour les tâches d'exploration de données. Les algorithmes peuvent être soit appliquées directement à un ensemble de donnéesou appelés à partir de votre propre code Java.
- Weka contient des outils pour les donnéesde pré-traitement, la classification, la régression, le regroupement, les règles d'association, et la visualisation.
- Il est également bien adapté pour le développement de nouveaux programmes d'apprentissage machine.
- Site web: http://www.cs.waikato.ac.nz/ml/weka/

TP WEKA : pour le Dataset IRIS.CSV déterminer l'algorithme d'apprentissage qui a un bon score d'évaluation

Fonctionnalités et mode opératoire

Interfaces utilisateur Weka



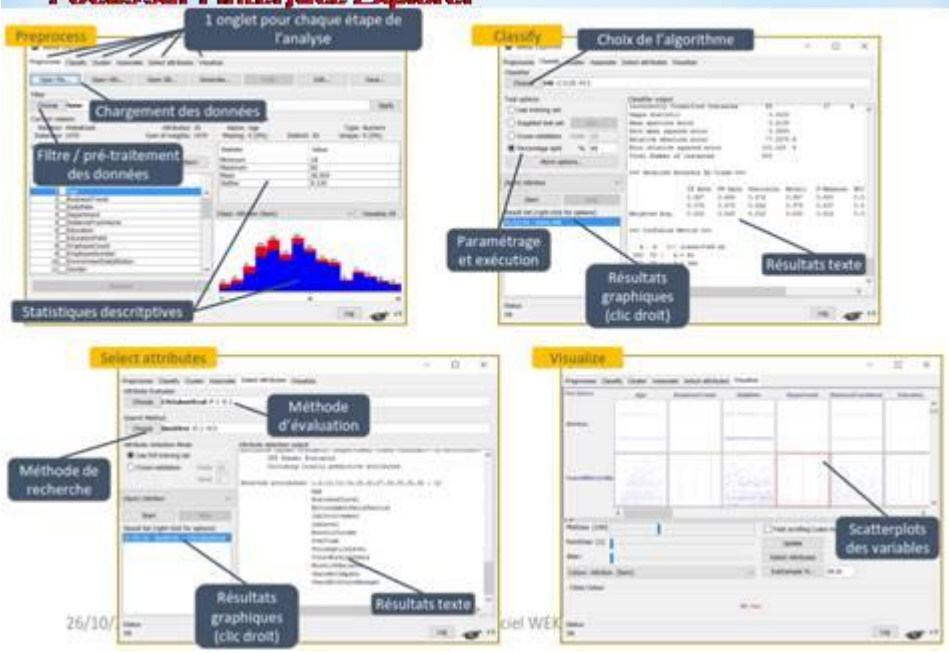
Fonctionnalités et mode opératoire

Principales fonctionnalités de traitement des données

Preprocessing	Import, inspection et préparation/filtre des données
Classification	Mise en œuvre des différents algorithmes de classification
Clustering	Accès aux techniques de clustering comme l'algorithme de k-means
Association	Accès aux apprentissages par règles d'association qui essaient d'identifier toutes les relations importantes entre les variables
Feature (attribute) selection	Choix des variables les plus pertinentes et prometteuses
Visualization	Affichage graphique scatterplot, arbres

Fonctionnalités et mode opératoire

Focus sur l'interface Explorer



Récapitulatif

1. Donnez la signification des fonctions L et R en DM
2. Quelles sont les taches de data mining ?
3. Donnez des exemples de taches : Estimation, prédiction, association
4. Quels sont les types de modèles pervant être exploités par le DM?
5. Quelle est la différence entre régression et classification en apprentissage supervisé ?
6. Quelle est la formule de calcul de l'erreur en validation croisée ?, combien de modèle faut-il apprendre si K=N ?
7. Citez les étapes pour bien mener un projet de DM
8. Quelles sont les étapes de Traitement des données avant étraction de connaissances ?
9. Quelles les techniques utilisées pour l'extraction des connaissances ?
10. Citez qq logiciels propriétaires du DM
11. Citez qq logiciels gratuits du DM
12. Donnez qq caractéristiques de WEKA?
13. Donnez les principales fonctionnalités de traitement de données sur WEKA

III. Web-Mining

Introduction au Web mining

Le Web Mining s'est développé à la fin des années1990. Ce domaine consiste à utiliser l'ensemble des techniques du Data Mining afin de développer des approches et des outils, permettant d'extraire des informations pertinentes à partir d'un large volume de donnéesdu web (documents, traces d'interactions, structure des pages, liens visités, parcours dans le site..)

Basé sur le type de donnéesà étudier dans le domaine du Web Mining, trois grands axes de recherches ont été envisagés portant sur le Web Content Mining (mining using HTML pages), le Web Structure Mining(mining using URL links) et le Web Usage Mining (mining using clicks, visits et logs).

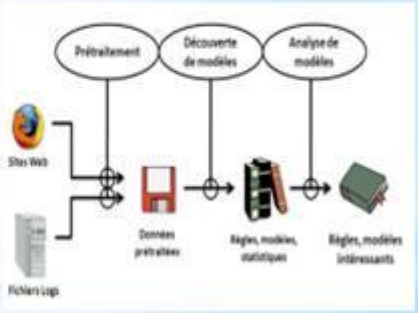
* Processus du Web mining

Le processus du Web Mining se déroule en trois étapes :

1. Collecte des donnéessur l'utilisateur,

2. Utilisation de ces donnéesà des fins de Personnalisation.

3. Présentation à l'utilisateur d'un contenu ciblé.



Processus du Web Mining

29

Les bases de données(BD) comparées à la recherche d'information (RI)

Format des données:

BD : donnéesstructurées. Sémantique claire, basée sur un modèle formel

RI : non-structurées - texte libre

Requêtes :

BD : formelles (par exemple, SQL)

RI : souvent - langues naturelles (recherche de mots-clé)

Résultats :

BD : résultat exact

RI : trouver les informations les plus pertinents pour une requête donnée

Le fonctionnement d'un moteur de recherche. Etapes :

L'exploration ou crawl faite par un robot d'indexation qui parcours récursivement tous les hyperliens qu'il trouve en récupérant les ressources intéressantes. L'exploration est lancée depuis une ressource pivot, comme une page d'annuaire web.

L'indexation des ressources récupérées extraire les mots considérés comme significatifs du corpus de documents `a explorer. Les mots extraits sont enregistrés dans une BD. Un dictionnaire inverse ou l'index TF-IDF permet de retrouver rapidement dans quel document se situe un terme significatif donné.

La recherche un algorithme identifie dans le corpus documentaire (en utilisant l'index), les documents qui correspondent le mieux aux mots contenus dans la requête. Les résultats des recherches sont donnés par ordre de pertinence supposée.

Contenu - GoogleBot

Google calcule pour chaque page un score de crawl => niveaux de crawl de GoogleBot :

couche de base : la plupart des pages du web. Elles sont crawlées avec une fréquence liée à la fréquence de mise à jour du contenu ainsi qu'à leur PageRank.

couche quotidienne : un petit nombre de pages crawlées de façon quotidienne.

couche temps réel : un nombre très petit de pages crawlées avec une fréquence de l'ordre de la minute, l'heure,... (ex. l'actualité)

Les caractéristiques clé de Google :

Utilisation des indexes basés sur le contenu

Indexes peuvent être partitionnés sur plusieurs machines ou bien les données sont dupliquées sur les machines et les requêtes sont distribuées parmi celles-ci

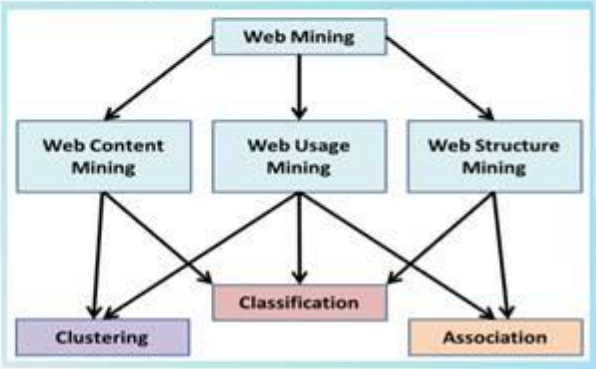
PageRank ordonner les documents trouvés, en utilisant les liens entre pages comme indice de pertinence

Qu'est ce que le Web Mining ?

Le Web Mining a pour objectif l'extraction de connaissances à partir de la structure formée par les liens entre pages/objets/communautés, du contenu des pages Web ou de données d'usages (typiquement des logs de serveurs web).
Bien que cette discipline se base fortement sur le Data Mining, on ne saurait réduire cette discipline à la seule application d'algorithmes existants dans ce contexte nouveau. Ceci est dû aux particularités des donnéesdu Web.
De nouvelles tâches ont été mises au point au cours de la dernière décennie. Une classification des approches du Web Mining fait désormais consensus et est basées sur les types de données manipulées par les approches. Nous décrivons ces 3 catégories, ci-dessous.

1– Web structure mining.

La fouille de structures issues du Web permet l'extraction de connaissances à partir des liens qu'il existe entre les pages Web ou toutes autres entités reliées entre elles. Les liens hypertexts représentent en quelques sortes la structure du Web. Typiquement, l'identification de pages importantes, est très utilisée par les moteurs de recherche. Ces approches permettent également de découvrir des communautés d'utilisateurs qui partager des intérêts communs. Il est important de noter que les approches de Data Mining classiques n'abordent pas cette thématique car les donnéestraitées ne contiennent pas de liens.



2— Web content mining.

La fouille de contenus de pages Web permet l'extraction de motifs à partir des contenus des pages Web. Par exemple, il est possible de regrouper des pages Web selon certaines thématiques communes comme le feraient des techniques de Data Mining plus classiques.
Il existe toutefois des spécificités associées au donnéesdu Web. Par exemple, la fouille des avis d'internautes sur des produits, e.g., musiques, films, voyages, permet de découvrir automatiquement l'opinion des consommateurs face à un produit pour accroître la qualité des systèmes de recommandation

3 — Web usage mining.

La fouille de donnéesd'utilisation du Web fait référence à l'analyse automatique des logs stockés sur les serveurs Web. De nombreux algorithmes peuvent être appliqués sur ce type de donnéescar celles-ci sont structurées, rendant possible l'utilisation de techniques standards. Une des principales difficultés lorsque l'on souhaite considérer de telles donnéesest leur prétraitement. En effet plusieurs problématiques de recherche sont apparues quant à l'identification des sessions utilisateurs ou encore la suppression du bruit.

Web structure mining

Les premiers moteurs de recherche retrouvaient les documents pertinents seulement en fonction de la similarité du contenu avec la requête de l'utilisateur.
Depuis le milieu des années1990, il est devenu de plus en plus évident que cette simple similarité n'était pas suffisante pour au moins deux raisons principales :

- D'abord, l'augmentation rapide du nombre de pages Web a rendu très difficile l'ordonnancement des documents retrouvés. Par exemple, la requête « classification technique » retourne quelques 190000 résultats. Classer ces documents seulement en fonction de la similarité de leur contenu pour ne retenir que 30 ou 40 documents à afficher dans la première page des documents est très difficile.
- Ensuite, il est très facile de « falsifier » le contenu d'une page de telle sorte à la rendre similaire à des très nombreuses requêtes. En effet, un concepteur de pages Web peut facilement ajouter artificiellement pléthore de mots-clés stratégiques pour augmenter la similarité de ses pages avec des requêtes.

Web structure mining

- La prise en compte de liens hypertextes est alors apparue comme la solution. En effet, à la différence des documents textuels « classiques » qui sont souvent considérés comme indépendants, les pages Web sont inter-connectées grâce aux hypertiens qui représentent une source d'information importante. Ces liens peuvent être internes, i.e., permettant la définition de la structure du site Web, ou externes, i.e., et implicitement indique à l'utilisateur que le site ciblé est une source d'information valide. Ces liens externes doivent alors de toute évidence être considérés par les moteurs de recherche lors de la recherche d'information.
- Au cours de la période 1997-1998, deux des plus influents algorithmes de recherche d'information basés sur l'influence des liens ont été proposés : PageRank et HITS. Ces deux algorithmes ont originellement été mis au point pour l'étude des réseaux sociaux. Ils exploitent tous les deux la structure du Web formée par les liens afin de définir « le prestige » ou « l'autorité » d'un site Web.

Web structure mining

1 PageRank

Développé dans l'université de Stanford par Brin et Page (Page et al., 1998), puis intégré dans le moteur de recherche Google, PageRank est un algorithme d'analyse de structure des liens entre pages très puissant pour mesurer la peupularité d'une page Web (Zhang et al., 2006).

A l'inverse de HITS, PageRank calcul une seule mesure de qualité (on dit aussi popularité ou prestige) pour une page au moment du crawl. Cette mesure et alors combinée avec un autre score similaire à ceux utilisés dans les systèmes de recherche d'information (SRI) lors de l'interrogation. Il est réputé être plus efficace, car la mesure de popularité d'une page est précalculée indépendamment de la requête. Actuellement, il permet de capturer la notion d'importance d'une page en se basant sur les informations des liens.

Ainsi, l'intérêt d'une page A dans le résultat est fonction du nombre de pages qui pointent vers elle.

De plus, et afin d'éviter les liens vains, qui peuvent être de nos jours générés facilement de manière artificielle, cet algorithme étend l'idée des citation, en prenant en compte non seulement le nombre de liens référençant une page mais aussi l'importance et le poids des pages pointant vers elle (Zhang et al., 2006).

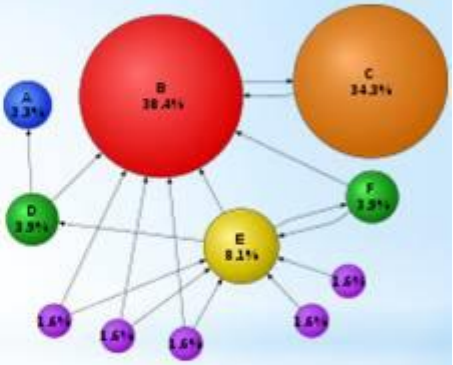
Web structure mining

1 PageRank

L'expression suivante donne la formule de calcul du score r(v), d'une page v utilisé dans PageRank, α étant une constante de normalisation, pa et ch deux fonctions représentant les parents et les fils d'une page (Baldi et al., 2003) :

$$r(v) = \alpha \sum_{w \in \mathrm{pa}[v]} \frac{r(w)}{|\mathrm{ch}[w]|}$$

Remarquons que chaque parent w contribue par une quantité proportionnelle à son score r (w) et inversement proportionnelle à son degré de sortie.



Le *PageRank* d'une page a tendance à être d'autant plus élevé que la somme des *PagesRanks* des pages qui pointent vers elle est élevée.

Web structure mining

2 HITS

L'algorithme HITS (Hyperlink Induced Topic Search) proposé par Kleinberg (Kleinberg, 1999) s'appuie sur un principe simple : toutes les pages n'ont pas la même importance, et ne jouent pas le même rôle. Certaines sont importantes ayant un contenu informatif (authoritative ou de référence), et d'autres sont des pages hubs contenant des liens vers des pages de référence.

Contrairement à PageRank, HITS démarre par un graphe d'une requête qu'il soumis à un SRI standard (les systèmes de recherche d'information), pour collecter un premier ensemble de pages liées au sujet recherché (les nœuds du graphe) appelé ensemble racine (root set) noté R. Ce dernier est utilisé pour obtenir un deuxième ensemble étendu (expanded set) noté E, par l'ajout de tous les nœuds fils, et un nombre fixe de nœuds parents de l'ensemble racine. Vu le nombre important de nœuds pouvant être impliqués, et afin d'alléger l'algorithme, les arrêtes représentant les liens qui relient les nœuds du même serveur seront éliminés. Ils sont seulement considérés comme des liens de navigation.

Web usage mining

- Si le WCM et le WSM utilisent les donnéesWeb primaires ou réelles, le Web Usage Mining (WUM) opère sur les donnéessecondaires de celui-ci, générées par l'interaction des utilisateurs avec les sites Web (Kosala et al., 2000).
- Avec le succès et la popularité du Web, de grandes quantités de donnéessont générées chaque jour par les serveurs Web, en enregistrant dans des fichiers de type log (journaux) les traces de navigation, appelée aussi clickstream (flux continu de clics), (Kimball et al., 2000), englobant les parcours et les actions, effectuées par les utilisateurs lors de leurs visites.
- Le WUM est donc la découverte et l'extraction de connaissances à partir des donnéesd'interaction des utilisateurs avec le Web, sous forme de modèles et de motifs de navigation, en utilisant les techniques du data mining (Cooley et al., 1997).
- De par la nature distribuée du Web, la journalisation (logging) dans le cadre du WUM peut être effectuée dans trois niveaux : les journaux de traces au niveau des serveurs Web, les journaux consignés par les serveurs proxy et les traces de navigation coté client (Cooley, 2000).

Web content mining

le Web contient une masse importante de contenu non structuré, i.e., des donnéestextuelles. Analyser ces textes peut fournir un grand nombre d'informations qui peuvent se révéler utiles pour un décideur. Dans ce cours, nous nous focalisons sur une tâche bien particulière : la fouille d'opinions. Cette tâche n'est pas seulement difficile, car elle implique l'utilisation de techniques issues du traitement automatique de la langue naturelle (TALN)

Des techniques ont été développées pour analyser cette formidable source d'informations. nous nous faciliterons sur 3 tâches particulières :

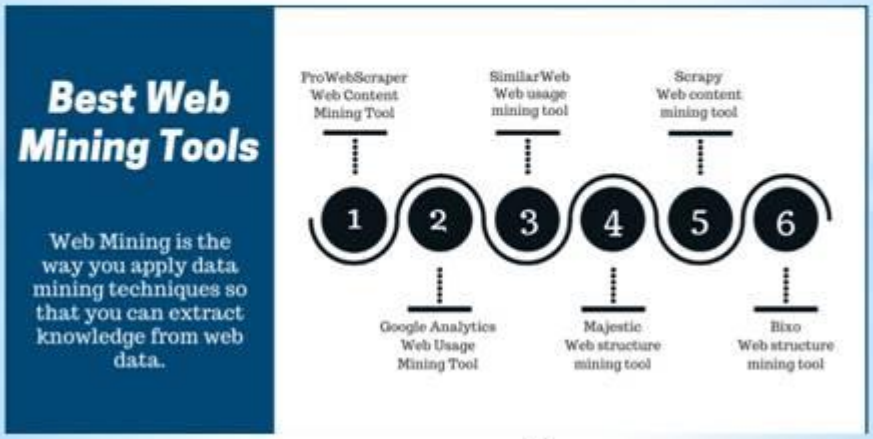
1— Classification de sentiments. La fouille d'opinion est traitée comme un problème de classification. Ces approches fonctionnent généralement au niveau d'un document et permettent de dire si une opinion positive ou négative a été formulée sur ce document

Web content mining

2 – La fouille d'opinions basée sur les caractéristiques. Ces approches fonctionnent à un niveau de granularité un peu plus fin : la phrase généralement. Elles permettent de découvrir les raisons qui poussent un internaute à aimer ou pas un produit ou un service.

3– La détection de spam. Il est très important de pouvoir détecter les faux avis, i.e., des avis malhonnêtes, de ceux qui sont réellement formulés par les internautes. La détection de ces fausses opinions est un problème difficile car les auteurs sont rarement correctement identifiés.

Les meilleures outils de WEB MINING



Google Analytics (Web Usage Mining Tool)

* Google Analytics est considéré comme l'un des meilleurs outils d'analyse des activités. Il peut suivre et signaler le site Web. Il peut suivre et signaler le trafic du site Web.

Caracteristiques:

-Analyse du rendement de la publicité et de la campagne

-Analyse et mise à l'essai du site Web

-Intégration facile avec le produit de Google comme, Adsense, Adwords, Google Display Network, Google Tag Manager, etc

SimilarWeb (Web usage mining tool)

SimilarWeb **est un puissant outil de business intelligence. Il offre des renseignements sur le trafic et le marketing pour n'importe quel site Web.**

Caractéristiques

*** Indicateurs de trafic et d'engagement**

*** Optimisation des moteurs de recherche et mots clés PPC**

*** Intérêts du public**

Exemple d'analyse Similarweb de site web : Menara.ma

*Exemple d'applications

Récapitulatif

1. Donnez les étapes de fonctionnement d'un moteur de recherche
2. Quels sont les niveaux de crawl de GoogleBot pour calculer le score de crawl d'une page web
3. Qu'est ce que le PageRank sur google search
4. Donnez la formule de calcul du score r(v) de PageRank
5. Quel est le principe de l'algorithme HITS
6. Quel est l'objectif principale du web mining
7. Donnez le schéma de la structure web mining
8. Qu'est ce que le Web content mining
9. Qu'est ce que le Web usage mining
10. Quels sont les niveaux de journalisation (logging) dans le cadre du WUM
11. Quelles sont les 3 tâches particulières du TALN dans le WCM ?
12. Donnez les meilleurs applications du Web mining