FOUILLE DE DONNEES
-
Fondements

Cours FMIN311 – M2

E. KERGOSIEN

eric.kergosien@lirmm.fr

UM2 MONTPELLIER

Laboratoire d'Informatique de Robotique et de Microélectronique de Montpellier

LIRMM

cnrs

irstea

UNIVERSITÉ MONTPELLIER 2
SCIENCES ET TECHNIQUES

Planning

- Cours de M. Roche : Fouille de Textes
- Ce matin : cours (FD – introduction & algorithmes)
- Jeudi 14/03 : cours (FD – Algorithmes)
- Jeudi 21-28/03 : TP Weka
- Jeudi 4/04 : cours M. Teisseire (Motifs)
- Jeudi 11/04 : TP Weka

Références

- Cours de Anne Laurent, Université Montpellier 2
- Cours de Christelle Scharff, IFI
- Cours de Marius Fieschi, Université de Marseille
- Cours de Fabrice Rossi, Télécom ParisTech
- Cours de Karine Zeitouni, Université de Versailles
- Introduction au Data Mining. M.Jambu. Eyrolles. 1998.
- Data Mining: Concepts and Techniques. J. Han and M. Kamber, The Morgan Kaufmann Series in Data Management Systems, 2000.
- Sites Web : KD Nuggets & UCI machine learning repository

Organisation du cours

- Introduction
  - Pourquoi la fouille de données? Et qu'est ce que la fouille de données?
  - Des données aux connaissances
  - Exemples concrets
- Point sur le projet
- Le data Mining et les autres disciplines
  - Liens entre DM et d'autres disciplines
  - Les deux types d'approches en DM
  - Variables numériques et catégorielles
- Les six grands types de techniques du Data Mining
  - La description, la classification, l'association, l'estimation , la segmentation et la prévision
- Algorithmes de ce cours pour la classification
  - Approches de classification supervisée et non supervisée
  - Règles d'association et motifs séquentiels
  - Évaluation des méthodes
- Conclusion
  - Quelques produits et zoom sur Weka

I. Introduction

a. Pourquoi la Fouille de donnée ?

b. Métaphore

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

d. Des données aux connaissances

e. Exemples d'applications concrètes

UNIVERSITÉ MONTPELLIER 2
SCIENCES ET TECHNIQUES

Pourquoi la fouille de données ?

- L'explosion des données

  Les outils de collecte automatique des données et les bases de données conduisent à d'énormes masses de données stockées dans des entrepôts

  - Entrepôts du Web : ex. Google,
  - Réseaux sociaux et hébergement de documents : ex. Facebook, gmail…
  - e-commerce Achats dans les supermarchés,
  - Transactions de cartes bancaires

Pourquoi la fouille de données ?

- L'explosion des données

  Les outils de collecte automatique des données et les bases de données conduisent à d'énormes masses de données stockées dans des entrepôts

  - Entrepôts du Web : ex. Google,
  - Réseaux sociaux et hébergement de documents : ex. Facebook, gmail…
  - e-commerce Achats dans les supermarchés,
  - Transactions de cartes bancaires
- Les données sont collectées et stockées rapidement (GB/heures)
  - Capteurs : RFID, supervision de procédé,
  - Télescopes,
  - Puces à ADN générant des expressions de gènes,
  - Simulations générant des téra-octets de données.

Pourquoi la fouille de données ?

- L'explosion des données

  Les outils de collecte automatique des données et les bases de données conduisent à d'énormes masses de données stockées dans des entrepôts

  - Entrepôts du Web : ex. Google,
  - Réseaux sociaux et hébergement de documents : ex. Facebook, gmail…
  - e-commerce Achats dans les supermarchés,
  - Transactions de cartes bancaires
- Les données sont collectées et stockées rapidement (GB/heures)
  - Capteurs : RFID, supervision de procédé,
  - Télescopes,
  - Puces à ADN générant des expressions de gènes,
  - Simulations générant de téra-octets de données.
- Submergés par les données, manque de connaissance !

Pourquoi la fouille de données ?

- Pourquoi maintenant ?
  - Limites de l'approche humaine & Techniques traditionnelles ne sont pas adaptées

Requêtes traditionnelles (SQL) impossibles : « Rechercher tous les enregistrements indiquant une fraude »

  - Solutions et compétences en Fouille récentes disponibles

Fournir de meilleurs services, s'adapter aux clients (e.g. dans les CRM)

  - Les données sont produites électroniquement et archivées
  - Le contexte est ultra-concurrentiel : Industriels, médicaux, marketing, etc.
  - Plateformes de calculs disponibles à bas prix

Pourquoi la fouille de données ?

- Pourquoi maintenant ?
  - Limites de l'approche humaine & Techniques traditionnelles ne sont pas adaptées

Requêtes traditionnelles (SQL) impossibles : « Rechercher tous les enregistrements indiquant une fraude »

  - Solutions et compétences en Fouille récentes disponibles

Fournir de meilleurs services, s'adapter aux clients (e.g. dans les CRM)

  - Les données sont produites électroniquement et archivées
  - Le contexte est ultra-concurrentiel : Industriels, médicaux, marketing, etc.
  - Plateformes de calculs disponibles à bas prix

- Un nouveau marché
  - Nouveau concept : Information as a product
  - Toute société ou organisme qui collecte des données valorisables est potentiellement un broker d'information, qu'il peut vendre ou en exploiter commercialement les modèles pour des utilisations essentiellement marketing.

Métaphore

- Trop de données...
  - Paradoxe : trop données mais pas assez d'informations

Métaphore

- Trop de données...
  - Paradoxe : trop données mais pas assez d'informations

- Difficulté d'accès à l'information…
  - Trop de données tue …l'information

Métaphore

- Trop de données...
  - Paradoxe : trop données mais pas assez d'informations

- Difficulté d'accès à l'information…
  - Trop de données tue …l'information

- Trop de pistes à explorer

Métaphore

- Trop de données...
  - Paradoxe : trop données mais pas assez d'informations

- Difficulté d'accès à l'information…
  - Trop de données tue …l'information

- Trop de pistes à explorer

- …Pas d'accès facile à l'information

Métaphore

- Ce dont on a besoin...
  - Automatisation

Métaphore

- Ce dont on a besoin...
  - Automatisation

  - Extraction des connaissances des bases de données

Métaphore

- Ce dont on a besoin...
  - Automatisation

  - Extraction des connaissances des bases de données

  - Génération d'hypothèses

Solution : Fouille de données

- Objectifs

  - Par analogie à la recherche des pépites d'or dans un gisement, la fouille de données vise :

    ✓ 1. à extraire des informations cachées par analyse globale ;

    ✓ 2. à découvrir des modèles ("patterns") difficiles à percevoir car :
      - le volume de données est très grand
      - le nombre de variables à considérer est important
      - ces "patterns" sont imprévisibles (même à titre d'hypothèse à vérifier)

Evolution de la technologie des bases de données

•1970… : Bases de données relationnelles (RDBMS)

•1980… : RDBMS, modèles de données avancés (extension du relationnel, OO, ...) et DBMS orientés application (spatial, scientifique, …)

• 1990 – 2000 : Fouilles de données et entrepôts de données, BDD multimédia, bases de données Web

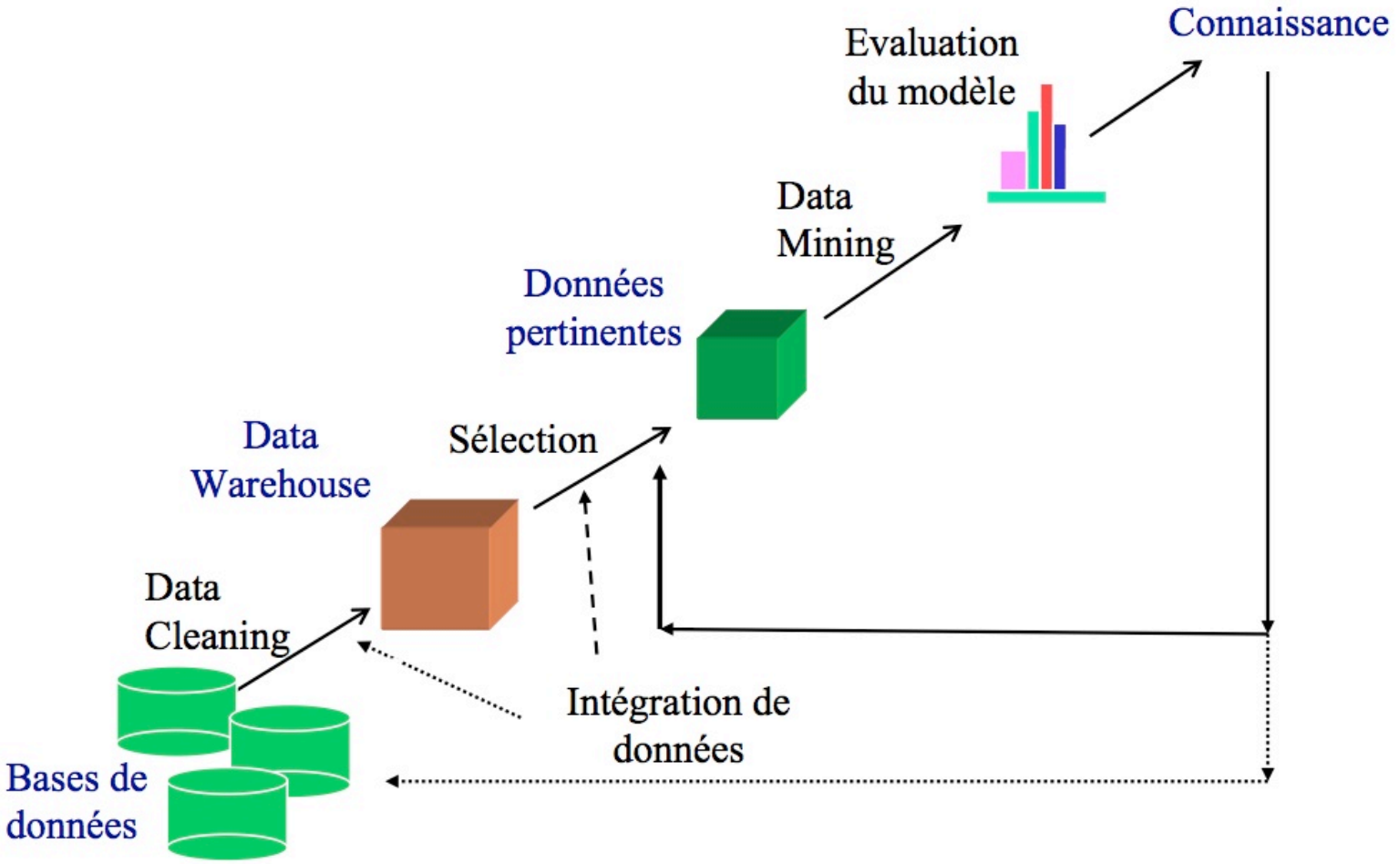
Définitions : KDD vs Fouille de Données

- Extraction de connaissances à partir des données (Knowledge Discovery in Databases – KDD) :
  - *Fayyad (1996) Knowledge Discovery in Databases :"the non-trivial process of identifying valid, potentially useful and ultimately understandable patterns in data"*
  - cycle de découverte d'information regroupant la conception de grandes bases de données ou entrepôts de données (Data Warehouse)
  - tous les traitements à effectuer pour extraire de l'information des données
  - L'un de ces traitements est la Fouille de données (Data Mining)

Cycle de vie du KDD

Définitions : KDD vs Fouille de Données

- Extraction de connaissances à partir des données (Knowledge Discovery in Databases – KDD) :



Etapes de processus de découverte de connaissance

1. Connaître le domaine d'application
   Connaissance pertinente déjà établie et buts de l'application
2. Sélection des données cibles
3. Data cleaning, prétraitement
4. Réduction de données et transformation
   - Choix des fonctions du data mining
     Synthèse, résumé, classification, régression, association, clustering.
   - Choix des algorithmes de fouille
5. Data mining
   Recherche des modèles intéressants
6. Evaluation des patterns et présentation de la connaissance
   Visualisation, transformation, etc.
7. Utilisation de la connaissance

Etape 1 : Datawarehousing

- Datawarehouse (Entrepôt de données)

–Base de données construite dans un but décisionnel depuis des bases de production souvent multi-sources et archivant des données historisées

  - actualisées soit par interrogation des bases sources (data pull), soit par envoie automatiques des modifications par les serveurs (data push)
  - généralement de grande taille correspondant à l'archivage du résultat des requêtes

–Datamart : magasin de données ciblé sur quelques sujets particuliers à l'échelle d'un département de l'entreprise

Etape 2 : OLAP (sélection)

- On-Line Analytical Processing (OLAP)

  – exploration (lecture) d'un datawarehouse par analyse multi-dimensionnelle et interactive

  – représente les données dans des «Data Cubes» donnant des comptages, totaux, ..., pour chaque variable et pour toute combinaison de variables avec différents niveaux de détail (ex : total annuel, sous-totaux mensuels, par semaine, ...)

Etape 3 : Fouille de données

- Fouille de données (Data Mining) :

  Ensemble de techniques d'exploration des données permettant d'extraire d'une base de données 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

Etape 4 : Reporting

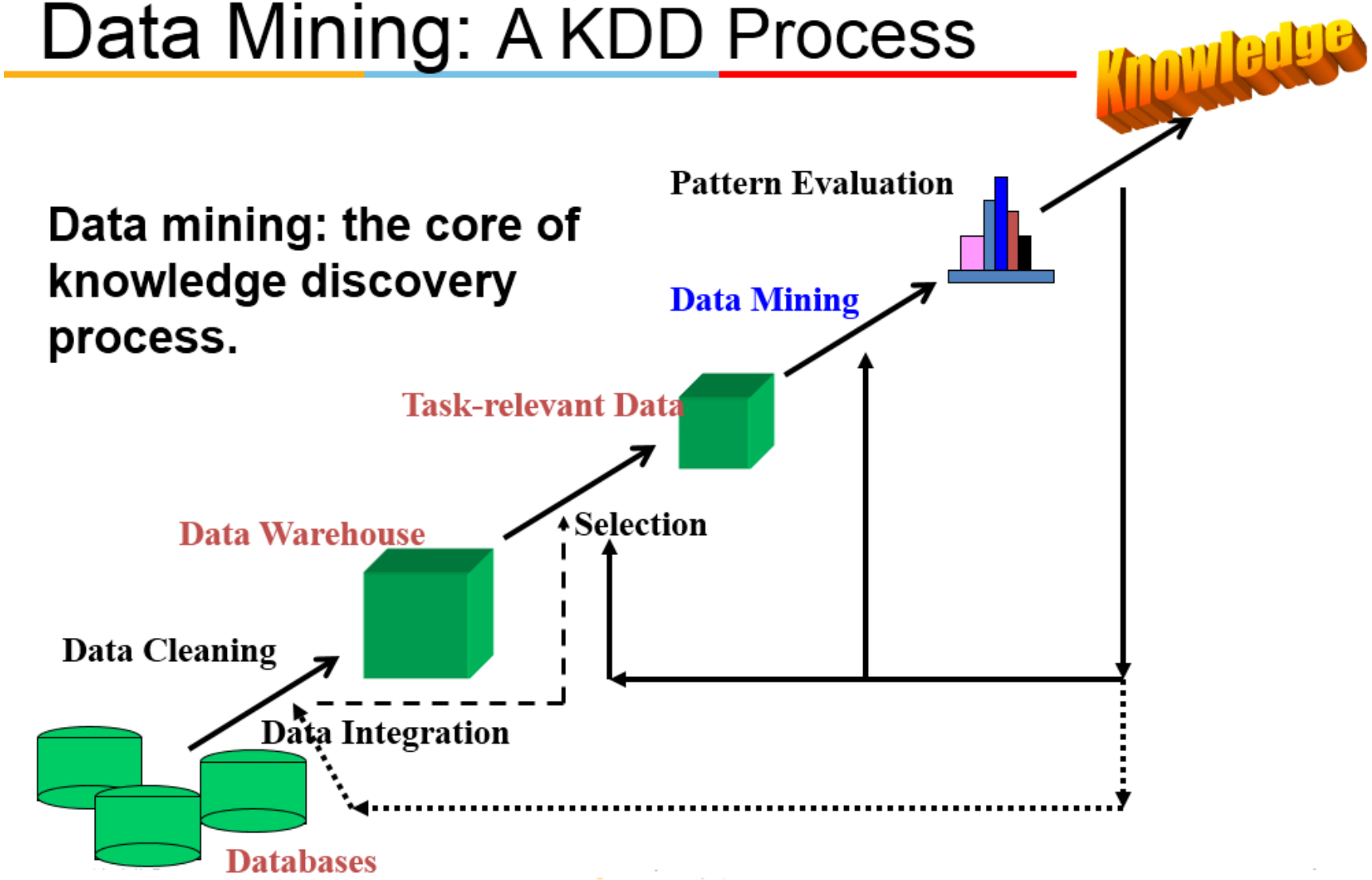
- Formalisation et diffusion

  – Les résultats sont formalisés pour être diffusés. Ils ne seront utiles qu'une fois devenus une connaissance partagée.

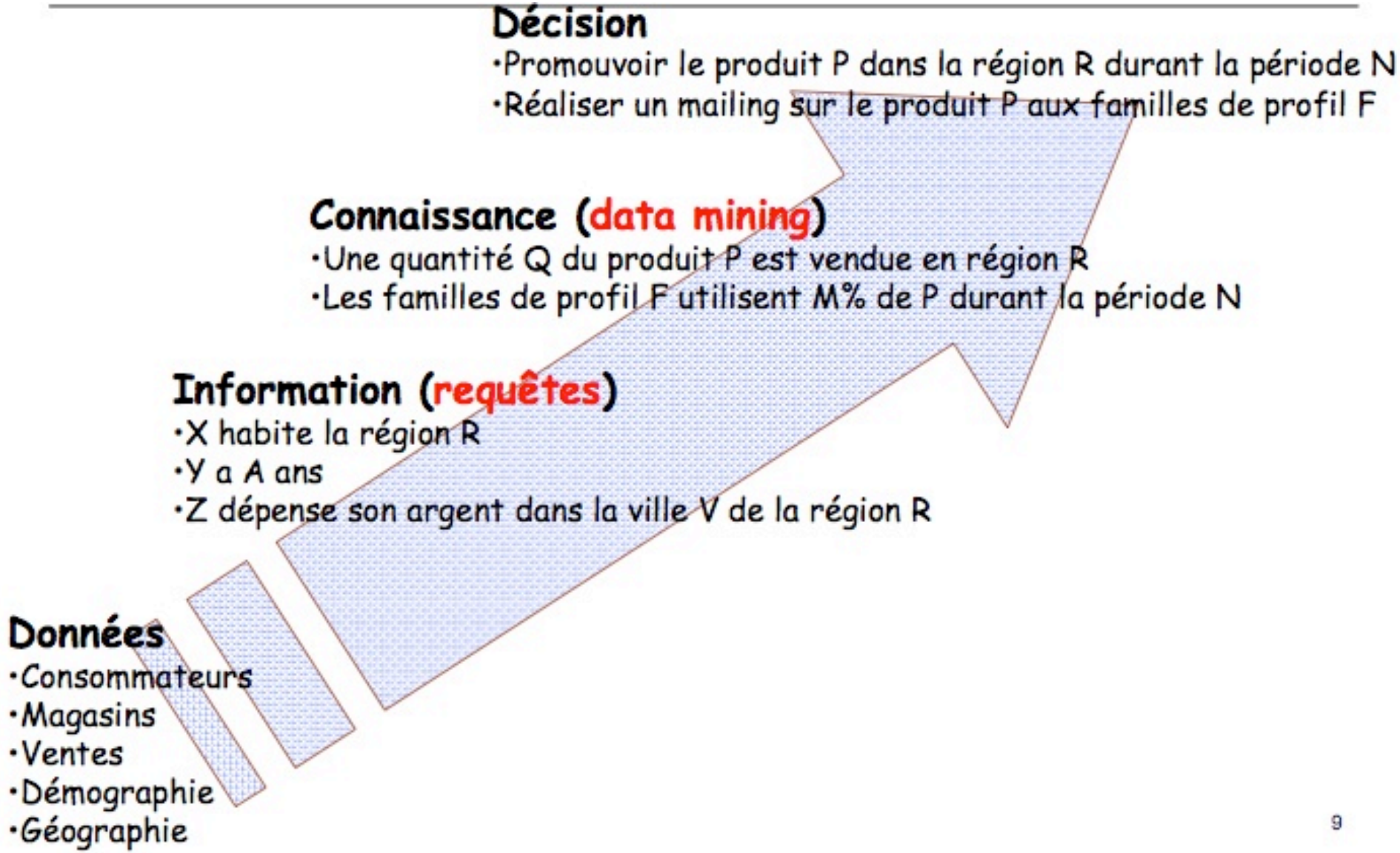
  – Aboutissement de la démarche : de nombreux outils de reporting, tableaux de bord, .. existent

  – Difficulté d'interprétation et de généralisation & peu d'automatisation

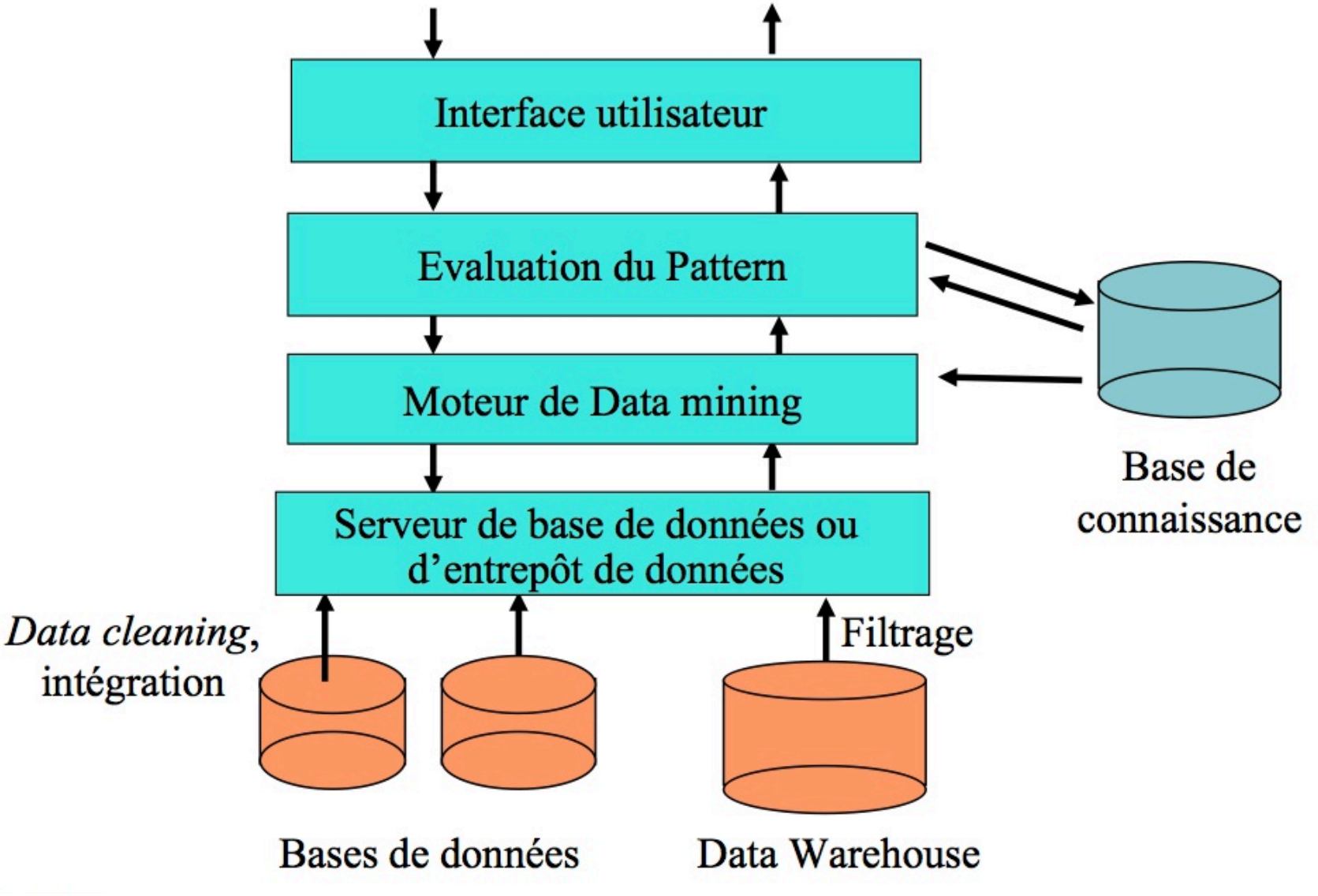
Définitions : KDD vs Fouille de Données



Data Mining : des données aux connaissances



Architecture d'un système type de Data Mining



Exercice : Data Mining ou non?

	Oui	Non
Rechercher le salaire d'un employé		
Les supporters achètent de la bière le samedi et de l'aspirine le dimanche		
Interroger un moteur de recherche Web pour avoir des informations sur le Data Mining		
Regrouper ensemble des documents retournés par un moteur de recherche en fonction de leur contenu		

Exercice : Data Mining ou non?

	Oui	Non
Rechercher le salaire d'un employé		X
Les supporters achètent de la bière le samedi et de l'aspirine le dimanche		
Interroger un moteur de recherche Web pour avoir des informations sur le Data Mining		
Regrouper ensemble des documents retournés par un moteur de recherche en fonction de leur contenu		

Exercice : Data Mining ou non?

	Oui	Non
Rechercher le salaire d'un employé		X
Les supporters achètent de la bière le samedi et de l'aspirine le dimanche	X	
Interroger un moteur de recherche Web pour avoir des informations sur le Data Mining		
Regrouper ensemble des documents retournés par un moteur de recherche en fonction de leur contenu		

Exercice : Data Mining ou non?

	Oui	Non
Rechercher le salaire d'un employé		X
Les supporters achètent de la bière le samedi et de l'aspirine le dimanche	X	
Interroger un moteur de recherche Web pour avoir des informations sur le Data Mining		X
Regrouper ensemble des documents retournés par un moteur de recherche en fonction de leur contenu		

Exercice : Data Mining ou non?

	Oui	Non
Rechercher le salaire d'un employé		X
Les supporters achètent de la bière le samedi et de l'aspirine le dimanche	X	
Interroger un moteur de recherche Web pour avoir des informations sur le Data Mining		X
Regrouper ensemble des documents retournés par un moteur de recherche en fonction de leur contenu	X	

Applications par domaine

Services financiers
- Attrition (churn)
- Détection de fraudes
- Identification opportunités de ventes

Marketing
- Gestion de la relation client (CRM)
- Optimisation de campagnes marketing
- Ventes croisées

Télécommunications
- Fidélisation (anti-churn)
- Ventes croisées
- Incidentologie

Assurances, Secteur public
- Indiquer les anomalies des comptes
- Réduire le coût d'investissement d'activité suspecte
- Détection de la fraudes

Grande Distribution
- Fidélisation
- Ventes croisées
- Analyses de panier
- Détection de fraudes

Sciences de la vie
- Trouver les facteurs de diagnostic typiques d'une maladie
- Alignement gênes & protéïnes
- Identifier les capacités d'interaction de médicaments

Internet
- Personnalisation des pub affichées
- Optimisation des sites web
- Profilage et Recommendation

Autre
- Rech. d'info (web ou document)
- Recherche par similarité (images…)
- Analyse spatiale…

Qu'est ce que la fouille de données?

En résumé : c'est quoi le Data Mining ??

Qu'est ce que la fouille de données?

En résumé

Découverte d'informations intéressantes dans un paquet de données

Qu'est ce que la fouille de données?

En résumé

Découverte d'informations intéressantes dans un paquet de données

- qu'est-ce qu'un paquet de données ?
- qu'est-ce qu'une information intéressante ?
- qu'entend-on par découvrir ?

Qu'est ce que la fouille de données?

En résumé

Découverte d'informations intéressantes dans un paquet de données

- qu'est-ce qu'un paquet de données ?
- qu'est-ce qu'une information intéressante ?
- qu'entend-on par découvrir ?

En anglais : Data Mining

Fortement lié à l'apprentissage automatique (machine learning)

Les données

- Un tableau de données
  - N lignes : les individus, les objets d'étude
  - P colonnes : les variables, les caractéristiques des objets
- Une base de données relationnelle
  - des tables des tableaux
  - des liens entre les tables : un client (dans la table des clients) a acheté des produits (dans la table des produits)
- Un entrepôt de données (data warehouse)
  - un entrepôt de données (data warehouse) : mise en commun de bases de données
  - agrégation de valeurs : nombre de commandes par enseigne et par mois d'un produit

Les données

- Un tableau de données
  - N lignes : les individus, les objets d'étude
  - P colonnes : les variables, les caractéristiques des objets
- Une base de données relationnelle
  - des tables des tableaux
  - des liens entre les tables : un client (dans la table des clients) a acheté des produits (dans la table des produits)
- Un entrepôt de données (data warehouse)
  - un entrepôt de données (data warehouse) : mise en commun de bases de données
  - agrégation de valeurs : nombre de commandes par enseigne et par mois d'un produit

→ Difficultés

Données complexes, hétérogènes, évolutives et volumineuses

Les données : exemples concrets

- Sciences de la vie
  - médecine : patients et maladies, essais cliniques
  - génomique : gènes, patients, tissus
- Marketing
  - fichiers clients
  - traces d'usage (site web, communication mobile)
  - Achats
- Industrie
  - senseurs : température, vibration
  - Images
  - analyse physico-chimique

Informations intéressantes

- Liens entre variables
  - Corrélation
  - dépendance non linéaire
  - capacité de prédiction
- Liens entre individus
  - interactions significatives
  - groupes homogènes
- Liens entre évènements
  - co-occurrence
  - dépendance logico-temporelle

Informations intéressantes : exemples concrets

- Sciences de la vie
  - lien entre tabagisme et maladies cardio-vasculaires
  - lien entre tabagisme et cancer du poumon
  - maladies génétiques : mutation → gène détérioré → protéine non produite → maladie
- Marketing
  - évaluation du risque de défaillance pour un crédit
  - typologie des clients
  - recommandation de produits
- Industrie
  - identification de modes de fonctionnement normaux d'un matériel
  - lien entre un mode vibratoire et une défaillance future
  - qualité subjective d'un produit à partir de mesures objectives

Découverte

- Exploration
  - l'analyste fait tout
  - rapports
  - outils visuels
- Semi-automatique
  - l'analyste guide le processus
  - algorithmes d'apprentissage : inférence à partir d'exemples de résultats voulus
- Automatique
  - intervention minimale de l'analyste : choix d'une méthode et analyse des résultats
  - parfois proche de l'exploration
  - souvent presque impossible mais souhaitable

Découverte : exemples concrets

- Exploration
  - statistiques classiques : moyenne, médiane, coefficient de corrélation
  - version visuelle : histogrammes, diagramme à bâtons
- Semi-automatique
  - segmentation d'un ensemble de clients
  - construction d'un modèle en vue d'une exploitation automatique
- Automatique
  - reconnaissance d'empreintes digitales
  - recherche de cooccurrences fréquentes
  - recommandations

Quelques applications concrètes

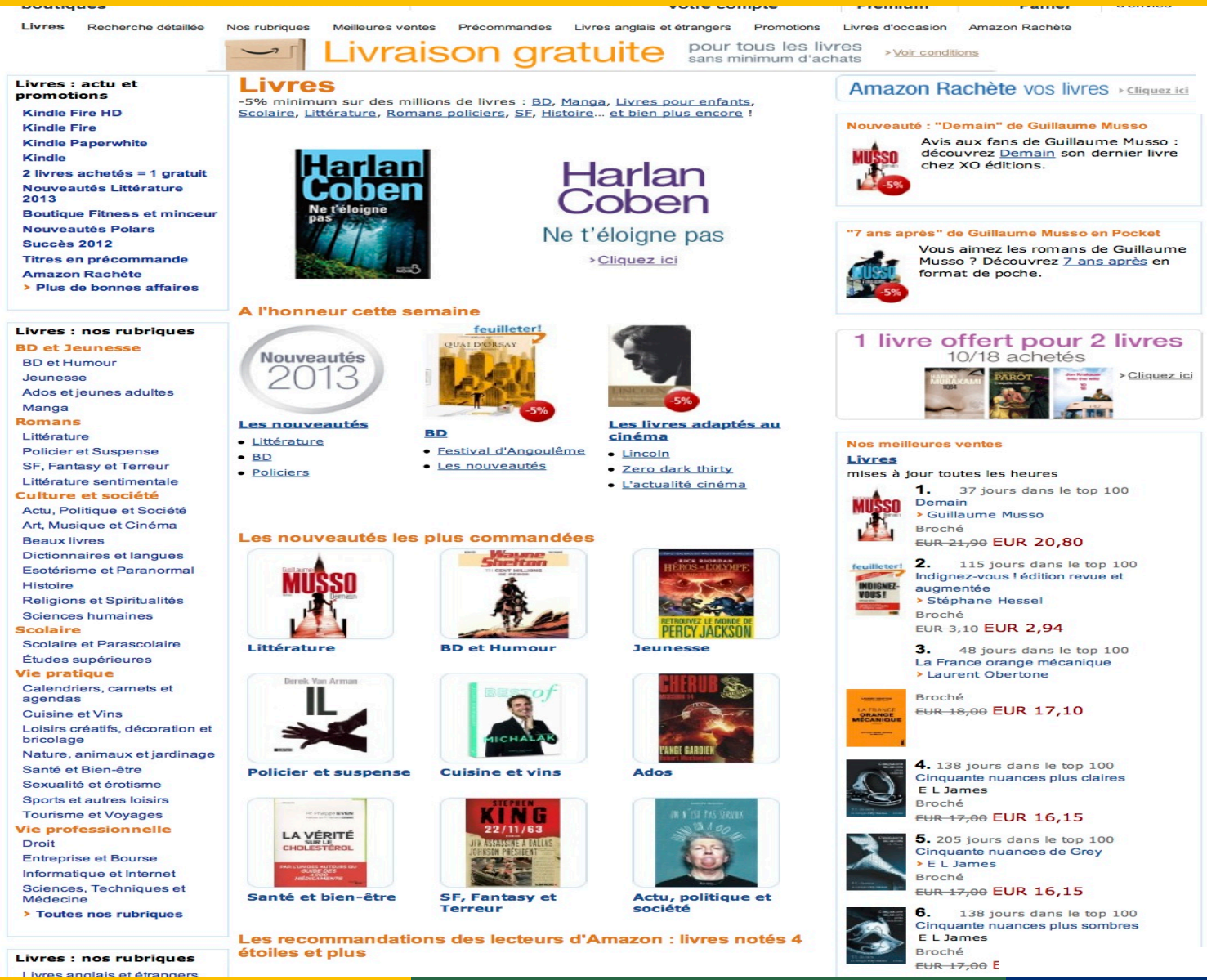
- Visualisation de l'information : liens entre profils facebook

- Amazon, lastfm, netflix
  - recommandations par co-achats
  - recommandations personnalisées

Visualisation : Facebook

Liens entre profils :
http://www.facebook.com/notes/facebook-engineering/ visualizing-friendships/ 469716398919

Amazon (1/2)


Livres
Recherche détaillée
Nos rubriques
Meilleures ventes
Précommandes
Livres anglais et étrangers
Promotions
Livres d'occasion
Amazon Rachète
Livraison gratuite
pour tous les livres
sans minimum d'achats
Voir conditions
Livres : actu et promotions
Kindle Fire HD
Kindle Fire
Kindle Paperwhite
Kindle
2 livres achetés = 1 gratuit
Nouveautés Littérature 2013
Boutique Fitness et minceur
Nouveautés Polars
Succès 2012
Titres en précommande
Amazon Rachète
Plus de bonnes affaires
Livres : nos rubriques
BD et Jeunesse
BD et Humour
Jeunesse
Ados et jeunes adultes
Manga
Romans
Littérature
Policier et Suspense
SF, Fantasy et Terreur
Littérature sentimentale
Culture et société
Actu, Politique et Société
Art, Musique et Cinéma
Beaux livres
Dictionnaires et langues
Esotérisme et Paranormal
Histoire
Religions et Spiritualités
Sciences humaines
Scolaire
Scolaire et Parascolaire
Études supérieures
Vie pratique
Calendriers, carnets et agendas
Cuisine et Vins
Loisirs créatifs, décoration et bricolage
Nature, animaux et jardinage
Santé et Bien-être
Sexualité et érotisme
Sports et autres loisirs
Tourisme et Voyages
Vie professionnelle
Droit
Entreprise et Bourse
Informatique et Internet
Sciences, Techniques et Médecine
Toutes nos rubriques
Livres : nos rubriques
Livres anglais et étrangers
Livres
-5% minimum sur des millions de livres : BD, Manga, Livres pour enfants, Scolaire, Littérature, Romans policiers, SF, Histoire... et bien plus encore !
Harlan Coben
Ne t'éloigne pas
Cliquez ici
A l'honneur cette semaine
Nouveautés 2013
Les nouveautés
Littérature
BD
Policiers
feuilleter
-5%
BD
Festival d'Angoulême
Les nouveautés
-5%
Les livres adaptés au cinéma
Lincoln
Zero dark thirty
L'actualité cinéma
Les nouveautés les plus commandées
Littérature
BD et Humour
Jeunesse
Policier et suspense
Cuisine et vins
Ados
Santé et bien-être
SF, Fantasy et Terreur
Actu, politique et société
Les recommandations des lecteurs d'Amazon : livres notés 4 étoiles et plus
Amazon Rachète vos livres
Cliquez ici
Nouveauté : "Demain" de Guillaume Musso
Avis aux fans de Guillaume Musso : découvrez Demain son dernier livre chez XO éditions.
-5%
"7 ans après" de Guillaume Musso en Pocket
Vous aimez les romans de Guillaume Musso ? Découvrez 7 ans après en format de poche.
-5%
1 livre offert pour 2 livres
10/18 achetés
Cliquez ici
Nos meilleures ventes
Livres
mises à jour toutes les heures
1. 37 jours dans le top 100
Demain
Guillaume Musso
Broché
EUR 21,90 EUR 20,80
2. 115 jours dans le top 100
Indignez-vous ! édition revue et augmentée
Stéphane Hessel
Broché
EUR 3,10 EUR 2,94
3. 48 jours dans le top 100
La France orange mécanique
Laurent Obertone
Broché
EUR 18,00 EUR 17,10
4. 138 jours dans le top 100
Cinquante nuances plus claires
E L James
Broché
EUR 17,00 EUR 16,15
5. 205 jours dans le top 100
Cinquante nuances de Grey
E L James
Broché
EUR 17,00 EUR 16,15
6. 138 jours dans le top 100
Cinquante nuances plus sombres
E L James
Broché
EUR 17,00 E

Amazon (2/2)

04/03/13

Amazon.fr : Livres : -5% et livraison gratuite

Nos boutiques éditeurs

l'école des loisirs
Bragelonne
Milady
POCKET
Gallimard 1911-2011 un siècle d'édition

Les livres les plus populaires

- Meilleures ventes
- Dernières nouveautés
- Cadeaux les plus offerts
- Cadeaux les plus demandés

Gagnez de l'argent

Amazon Rachète : reprise de livres
Echangez vos livres contre des chèques-cadeaux

Vendez sur Amazon.fr
C'est facile, rapide, et la mise en vente est gratuite !

Devenez Partenaires
Proposez des produits depuis votre site web et touchez jusqu'à 10 % de rémunération !

Avantage
Un programme simple et efficace permettant aux éditeurs de développer leurs ventes.

App-Shop Amazon pour Android
Téléchargez chaque jour une application offerte pour votre smartphone et tablette Android.

Webservices
Les développeurs de logiciels peuvent utiliser des protocoles basés sur XML pour intégrer des fonctions et du contenu Amazon.fr dans d'autres sites Web. *(Page en anglais)*

› Tous nos programmes

Cinquante nuances plus claires
E L James
Broché
(98)
EUR 16,15

Cinquante nuances plus sombres
E L James
Broché
(239)
EUR 16,15

Demain
Guillaume Musso
Broché
(9)
EUR 20,80

La France orange mécanique
Laurent Obertone
Broché
(100)
EUR 17,10

› Plus de livres bien notés

Retrouvez les mots-clés les plus recherchés en Livres

twilight| millenium| stephenie meyer| harry potter| livres| marc levy| biographie| millenium poche| douglas kennedy| musso| nora roberts| pour les nuls| stephen king| harlan coben| books| maxime chattam| dan brown| guillaume musso| eragon| naruto| star wars| fascination| cuisine| les chevaliers d'emeraude| stieg larsson| sophie kinsella| anna gavalda| twilight tome 1| tintin| bd| hugh laurie| fred vargas| mary higgins clark| dictionnaire| millenium stieg larsson| meg cabot| revue technique automobile| pocket| joomla| vampire| manga| tara duncan| chattam| poker| paris| guide du routard| anita blake| robin hobb| philosophie| musso guillaume

Envie d'un bon roman ?

Toute occasion est bonne pour avoir envie d'un bon roman : un départ en vacances, un cadeau, l'utilisation des transports en commun, l'envie de partager un bon moment avec un proche, une envie de dépaysement... Vive les romans ! Le roman, ce genre littéraire apparu dès l'Antiquité avec les récits historiques et la tragédie, se caractérise par la narration fictionnelle. Le roman se destine plus à une lecture individuelle que collective (à la différence du conte ou de l'épopée), et ses lecteurs y puisent souvent un intéret pour les personnages, une intrigue, des péripéties, et l'art d'écrire. Le roman est aujourd'hui la forme de littérature dominante et multiplie ses genres : du roman pour enfants au roman policier, chacun peut de nos jours trouver son bonheur dans le roman. Le roman multiplie également ses formats : du roman épistolaire au roman intéractif, chaque écrivain est libre de choisir la composition la plus adaptée. la pluralité des genres et la lecture individuelle rendent l'appréciation d'un roman très subjective ! Bernard Werber écrira que "tout bon roman doit pouvoir se résumer à une blague"... d'autres débatteront de longues heures sur les blogs et autres forums spécialisés de la toile des qualités dudit roman. Et vous ? Quel sera votre prochain roman ? Trouvez le roman élu sur Amazon.fr, à -5% et livré gratuitement. Bonne lecture !

8. 37 jours dans le top 100
Un sentiment plus fort que la peur
Marc Levy
Broché
~~EUR 21,00~~ EUR 19,95

9. 17 jours dans le top 100
Les lumières de l'invisible.
› Patricia Darré
Broché
~~EUR 17,95~~ EUR 17,05

10. 30 jours dans le top 100
7 ans après...
› Guillaume Musso
Poche
~~EUR 7,80~~ EUR 7,41

› Voir toutes les meilleures ventes Livres

Nos meilleures ventes

Livres : Poches
mises à jour toutes les heures

1. 1364 jours dans le top 100
Cannibale
› Didier Daeninckx
Poche
~~EUR 4,20~~ EUR 3,99

2. 40 jours dans le top 100
Ô ma mémoire : La poésie, ma nécessité
› Stéphane Hessel
Poche
~~EUR 7,90~~ EUR 7,50

3. 670 jours dans le top 100
Je résiste aux personnalités toxiques (et autres casse-pieds...
Christophe André ...
Broché
~~EUR 8,50~~ EUR 8,07

4. 18 jours dans le top 100
Back up
› Paul Colize
Broché
~~EUR 7,50~~ EUR 7,12

5. 1304 jours dans le top 100
Cent ans de solitude
› Gabriel Garcia Marquez ...
Poche
~~EUR 8,00~~ EUR 7,60

› Voir toutes les meilleures ventes Poches

II. Point sur le projet

UNIVERSITÉ MONTPELLIER 2
SCIENCES ET TECHNIQUES

Etapes de processus de découverte de connaissance

1. Quel domaine d'application?
   Connaissance pertinente déjà établie et buts de l'application
2. Sélection des données cibles
3. Data cleaning, prétraitement :

   Préciser les entrées, les sorties au format attendu et les actions de prétraitement à réaliser

1. Réduction de données et transformation : vous en êtes ici au niveau du projet !!
   - Choix des fonctions du data mining
   - Choix des algorithmes de fouille
2. Data mining

   Recherche des modèles intéressants

6. Evaluation des pattern et présentation de la connaissance

   Visualisation, transformation, etc.

7. Utilisation de la connaissance

II. Le data Mining et les autres disciplines

a. Liens entre DM et d'autres disciplines

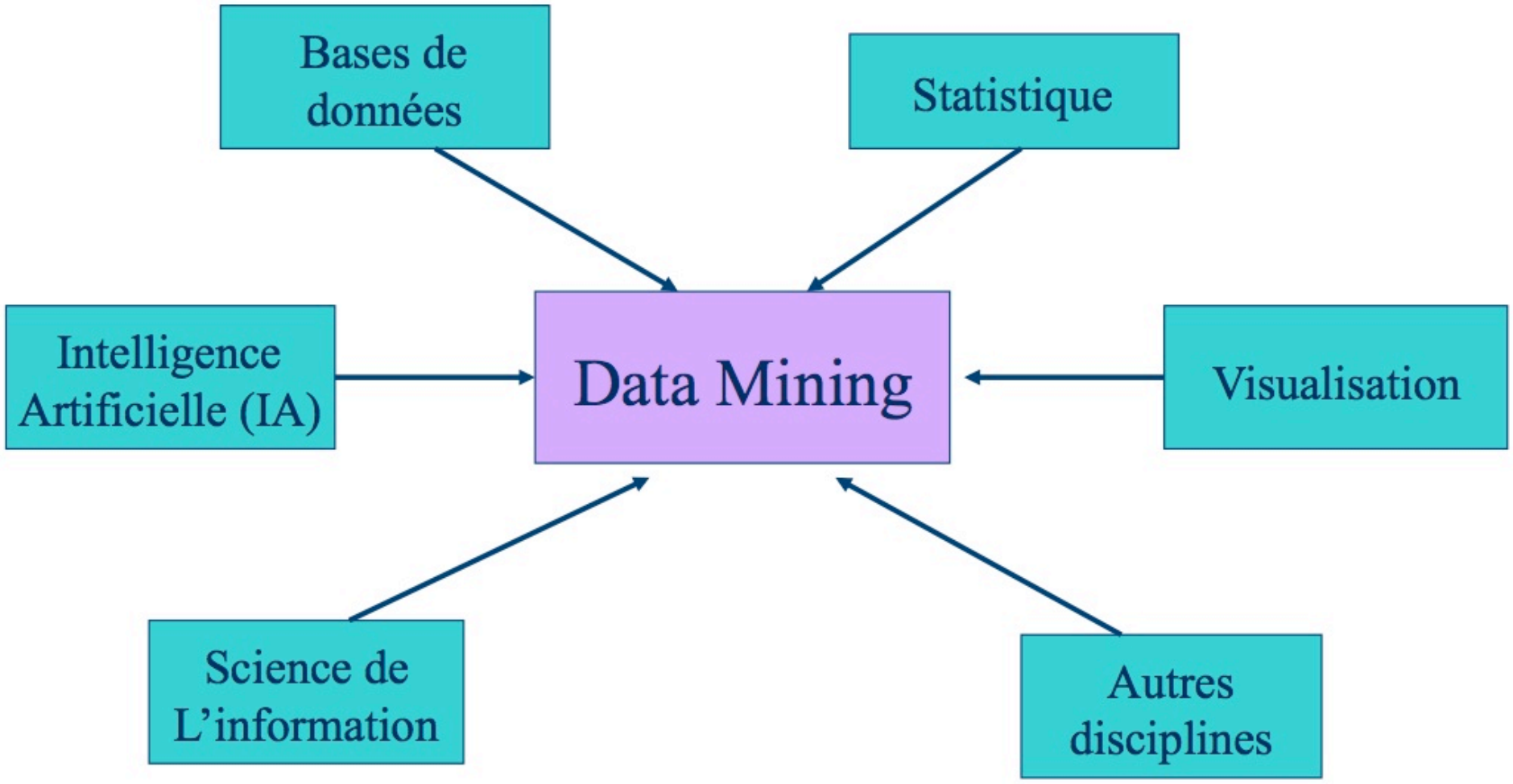
b. DM vs Statistiques et IA

c. Les deux grands types d'approches en DM

d. Variables numériques et catégorielles

UNIVERSITÉ MONTPELLIER 2
SCIENCES ET TECHNIQUES

Liens entre DM et d'autres disciplines



Data Mining vs Statistiques

- En statistique
  - Quelques centaines d'individus
  - Quelques variables
  - Fortes hypothèses sur les lois statistiques
  - Importance accordée au calcul
  - Échantillon aléatoire

- En Data mining
  - Des millions d'individus
  - Des centaines de variables
  - Données recueillies sans étude préalable
  - Nécessité de calculs rapides
  - Corpus d'apprentissage.

Data Mining vs Statistiques et IA

DM	Stat. Tableau individu -variable Calculs numériques	IA Formalisme de la logique Induction/déduction
Recherche de règles de classement	Méthodes de discrimination Réseaux de neuronnes Segmentation	Apprentissage supervisé/ex. -Génèr° de règles -Constr° d'arbre de décision -Raisonnement à base de cas
Régression	Méthodes de régression Réseaux de neuronnes	—
Classification automatique	Classif° automatique hiérarchique Partitionnement Réseaux de neuronnes	Apprentissage non supervisé -Classif° conceptuelle
Description synthétique	Stat. Élémentaire (histogramme, moy, écart-type) Outils d'interprét° de classes Méthodes factorielles (ACP)	Apprentissage non supervisé -Généralisation
Recherche de dépendances	Corrélations Analyse factorielles des corr. (AFC) Réseaux bayésiens	Apprentissage non supervisé -Généralisation -Recherche d'associations
Détection de déviations	Test stat sur les écarts	—

Ce qui est nouveau en Data Mining

- Expression et résolution des règles d'association
  – analyse de la consommation depuis <Num. transaction, articles achetés>
- Extension de SQL par des requêtes inductives (ex. DMQL) – intro. de connaissances tq hiérarchie de concepts et définition des seuils
- Nouveaux algorithmes
  – ex. pour le clustering utilise des techniques d'indexation de bases de données pour l'efficacité sur de GROS volumes de données ϖ
- L'intégration de l'OLAP et du data mining
  – Par exemple, génération de hiérarchies de dimension par la classification automatique hiérarchique.

.

Les deux types de techniques en Data Mining

- Les techniques descriptives (recherche de « patterns ») :
  - Décrire : visent à mettre en évidence des informations présentes mais cachées par le volume des données (c'est le cas des segmentations de clientèle et des recherches d'associations de produits sur les tickets de caisse)
  - réduisent, résument, synthétisent les données
  - il n'y a pas de variable à expliquer
  - On les appelle aussi : technique non supervisées.
  - Elles produisent des modèles de classement : typologie, méta-typologie.

Les deux types d'approches en Data Mining

- Les techniques descriptives (recherche de « patterns ») :
  - Décrire : visent à mettre en évidence des informations présentes mais cachées par le volume des données (c'est le cas des segmentations de clientèle et des recherches d'associations de produits sur les tickets de caisse)
  - réduisent, résument, synthétisent les données
  - il n'y a pas de variable à expliquer
  - On les appelle aussi : technique non supervisées.
  - Elles produisent des modèles de classement : typologie, méta-typologie.
- Les techniques prédictives (modélisation) :
  - Prédire : visent à extrapoler de nouvelles informations à partir des informations présentes (c'est le cas du scoring)
  - expliquent les données
  - il y a une variable à expliquer
  - On les appelle aussi : techniques supervisées
  - demandent plus d'historique que les techniques descriptives.
  - Elles produisent des modèles de prédiction.

Les variables numériques et catégorielles

- Les variables numériques

  permettent de faire des résumés, des synthèses: moyenne, minimum, maximum, écart type, etc.

- Les variables catégorielles

  permettent de faire des regroupement par catégories, c'est-à-dire des classements.

II. Les six grands types de techniques du Data Mining

a. La description
b. La classification
c. L'association
d. L'estimation
e. La segmentation
f. La prévision

UNIVERSITÉ MONTPELLIER 2
SCIENCES ET TECHNIQUES

Les six grands types de techniques du Data Mining

Techniques Descriptives			Techniques prédictives		
1. Description	2. Classification	3. Association	4. Estimation	5. Segmentation	6. Prévision

Les techniques concrètes

Le data mining utilise des techniques concrètes qui peuvent être limitées à un type de technique spécifique ou être partagées par plusieurs types de techniques :

- Exemple de méthodes descriptives : la classification hiérarchique, la classification des K moyennes, les réseaux de Kohonen, les règles d'association.

- Exemples de méthodes prédictives : les méthodes de régression, les arbres de décision, les réseaux de neurones, les K plus proches voisins.

1. La description (technique descriptive)

Principe

La description consiste à mettre au jour

- Pour une variable donnée : la répartition de ses valeurs (tri, histogramme, moyenne, minimum, maximum, etc.).
- Pour deux ou trois variables données : des liens entre les répartitions des valeurs des variables. Ces liens s'appellent des « tendances ».

Intérêt

- Favoriser la connaissance et la compréhension des données.

Méthode

- Méthodes graphiques pour la clarté : analyse exploratoire des données.

Exemples

- Répartition des votes par âge (lien entre les variables « vote » et « âge »).

2. La classification (technique descriptive)

Principe

- Aussi appelée clustering ou segmentation : consiste à créer des classes (sous-ensembles) de données similaires entre elles et différentes des données d'une autre classe

  →l'intersection des classes entre elles doit toujours être vide).

- il s'agit pour n variables de créer des sous-ensembles disjoints de données. On dit aussi «segmenter» l'ensemble entier des données.
- Elle définit les types de regroupement / distinction : on parle de métatypologie (type de type).
- Elle permet une vision générale de l'ensemble (de la clientèle, par exemple).

Intérêt

- Favoriser, grâce à la métatypologie, la compréhension et la prédiction
- Fixer des segments qui serviront d'ensemble de départ pour des analyses approfondies
- Réduire les dimensions, c'est-à-dire le nombre d'attributs, quand il y en a trop au départ

Méthode

- Classification hiérarchique ; Classification des K moyennes ; Réseaux de Kohonen ; Règles d'association.

Exemples : Métatypologie d'une clientèle en fonction de l'âge, les revenus, le caractère urbain ou rural, la taille des villes, etc.

3. L'association (technique descriptive)

Principe

- consiste à trouver quelles valeurs des variables sont corrélées ensemble. Par exemple, telle valeur d'une variable intervient avec telle valeur d'une autre variable.
- Les règles d'association sont de la forme : si antécédent, alors conséquence.
- L'association ne fixe pas de variable cible. Toute les variables peuvent à la fois être prédicteurs et variable cible.
- On appelle aussi ce type d'analyse une « analyse d'affinité ».

Intérêt

- Mieux connaître les comportements.

Méthode

- Algorithme a priori : Algorithme du GRI (induction de règles généralisées)

Exemples

- Analyse du panier de la ménagère (si j'achète des fraises, alors j'achète des cerises).
- Étudier quelle configuration contractuelle d'un abonné d'une compagnie de téléphone portable conduit plus facilement à un changement d'opérateur.

4. L'estimation (technique prédictive)

Principe

- L'estimation consiste à définir le lien entre un ensemble de prédicteurs et une variable cible. Ce lien est défini à partir de données « complètes », c'est-à-dire dont les valeurs sont connues tant pour les prédicteurs que pour la variable cible. Ensuite, on peut déduire une variable cible inconnue de la connaissance des prédicteurs.
- À la différence de la segmentation (technique prédictive suivante) qui travaille sur une variable cible catégorielle, l'estimation travaille sur une variable cible numérique.

Intérêt

- Permettre l'estimation de valeurs inconnues

Méthode

- Analyse statistique classique : régression linéaire simple, corrélation, régression multiple, intervalle de confiance, estimation de points.
- Réseaux de neurones

Exemples

- Estimer la pression sanguine à partir de l'âge, le sexe, le poids et le niveau de sodium dans le sang.
- Estimer les résultats dans les études supérieures en fonction de critères sociaux.

5. La segmentation (technique prédictive)

Principe

- La segmentation est une estimation qui travaille sur une variable cible catégorielle.
- On parle de segmentation car chaque valeur possible pour la variable cible va définir un segment (ou type, ou classe, ou catégorie) de données.
- La segmentation peut être vue comme une classification supervisée.

Intérêt

- Permettre l'estimation de valeurs inconnues

Méthode

- Graphiques et nuages de points ; Méthode des k plus proches voisins ; Arbres de décision ; Réseau de neurones.

Exemples

- Segmentation par tranche de revenus : élevé, moyen et faible (3 segments). On cherche les caractéristiques qui conduisent à ces segments.
- Déterminer si un mode de remboursement présente un bon ou un mauvais niveau de risque crédit (deux segments).

6. La prévision (technique prédictive)

Principe

- La prévision est similaire à l'estimation et à la segmentation mise à part que pour la prévision, les résultats portent sur le futur.

Intérêt

- Permettre l'estimation de valeurs inconnues

Méthode

- Celles de l'estimation ou de la segmentation

Exemples

- Prévoir le prix d'action à trois mois dans le futur
- Prévoir le temps qu'il va faire
- Prévoir le gagnant du championnat de football, par rapport à une comparaison des résultats des équipes

II. Algorithmes de ce cours pour la classification

a. Classification supervisée :
   - Méthode de Bayes naïf
   - k plus proches voisins
   - Arbres de décision
   - Réseaux de neurones
b. Classification non supervisée : k-means
c. Évaluation des méthodes
d. Règles d'association et motifs séquentiels

UNIVERSITÉ MONTPELLIER 2
SCIENCES ET TECHNIQUES

Rappel : Algorithmes supervisés et non supervisés

- Apprentissage supervisé
  - On dispose d'un fichier décrivant des données alliant une description et une classe
  - On cherche une fonction de classification permettant d'induire la classe en fonction d'une description
- Apprentissage non supervisé
  - On dispose d'un fichier de description des données sans classes connues a priori
  - On cherche à diviser ces données en catégories

Problématiques associées

- données pas forcément très propres
  - données bruitées
  - données manquantes
  - Données aberrantes
  - Doublons
- type des données : données numériques, symboliques, etc.

- pré-traitements ??
- post-traitements ??

Apprentissage supervisé : Méthode de Bayes naïf

- Comment classer un nouvel exemple en fonction d'un ensemble d'exemples pour lesquels on connaît la classe ?

- Soit un exemple d = (d1, ..., dn) et c classes k = 1, ..., c

$$\text{Classe}(d) = \underset{k}{\text{argmax}} \prod_{i} P(d_i|k) \cdot P(k)$$

$P(d_i|k)$: proportion d'exemples di (respectant une caractéristique) parmi ceux de la classe k

$P(k)$: proportion d'exemples de la classe k

→ La classe choisie sera celle pour qui la probabilité est maximale

Méthode de Bayes naïf : Exemple : Va-t-on jouer au tennis?

	TEMPS	HUMIDITE	VENT	TENNIS
Ex1	Soleil	Haute	Oui	Oui
Ex2	Soleil	Basse	Non	Non
Ex3	nuageux	Basse	Oui	Oui
Ex4	pluvieux	Haute	Oui	Non
Ex5	pluvieux	Basse	Oui	Non
Ex6	Soleil	Basse	Oui	Oui
Ex7	pluvieux	Basse	Non	Non
	Soleil	***haute***	***Non***	***?***

Va-t-on jouer s'il y a du soleil, beaucoup d'humidité et pas de vent ?

Apprentissage supervisé : Méthode de Bayes naïf

- Apprentissage Bayes naïf

Classe Oui			Classe Non		
Temps	Soleil	2/3	Temps	Soleil	1/4
	Nuageux	1/3		Nuageux	0/4
	pluvieux	0/3		pluvieux	3/4
Humidité	Haute	1/3	Humidité	Haute	1/4
	Basse	2/3		Basse	3/4
Vent	Oui	3/3	Vent	Oui	2/4
	Non	0/3		Non	2/4

- P (Oui) = 3/7
- P (Non) = 4/7

Test : classifier (Temps = soleil, humidité = haute, Vent = non)

Max (3/7 * 2/3 * 1/3 * 0/3, 4/7 * ¼ * ¼ * 2/4) = max (0, 0.017) = 0.017

→ Pas de tennis aujourd'hui !!

Méthode de Bayes naïf : conclusions

- Avantages
  - Méthode très répandue
  - facile à mettre en œuvre

- Inconvénients
  - hypothèse naïve très souvent fausse mais modèle souvent fiable

Méthode de Bayes naïf : Exemple 2 : Va-t-on jouer au tennis?

	TEMPS	HUMIDITE	VENT	TENNIS
Ex1	Soleil	Haute	Oui	Oui
Ex2	Soleil	Basse	Non	Non
Ex3	nuageux	Basse	Oui	Oui
Ex4	pluvieux	Haute	Oui	Non
Ex5	pluvieux	Basse	Oui	Non
Ex6	Soleil	Basse	Oui	Oui
Ex7	pluvieux	Basse	Non	Non
Ex8	Soleil	haute	Non	Non

Va-t-on jouer s'il y a du soleil, beaucoup d'humidité et du vent ?

Pondération des caractéristiques : temps = 2, humidité = 1 et vent = 1.

Apprentissage supervisé : Méthode de Bayes naïf

- Apprentissage Bayes naïf

Classe Oui			Classe Non		
Temps	Soleil	2/3	Temps	Soleil	2/5
	Nuageux	1/3		Nuageux	0/5
	pluvieux	0/3		pluvieux	3/5
Humidité	Haute	1/3	Humidité	Haute	2/5
	Basse	2/3		Basse	3/5
Vent	Oui	3/3	Vent	Oui	2/5
	Non	0/3		Non	3/5

- P (Oui) = 3/8 et P (Non) = 5/8

Test : classifier (Temps = soleil, humidité = haute, Vent = oui) ;

pondération : temps = 2 ; humidité = 1 ; Vent = 1

Max(3/8 * 2/3 * 2/3 * 1/3 * 3/3, 5/8 * 2/5 * 2/5 * 2/5 * 2/5) = max(0.037, 0.016) =0.037

→Un peu de sport ne fait pas de mal !!

II. Algorithmes de ce cours pour la classification

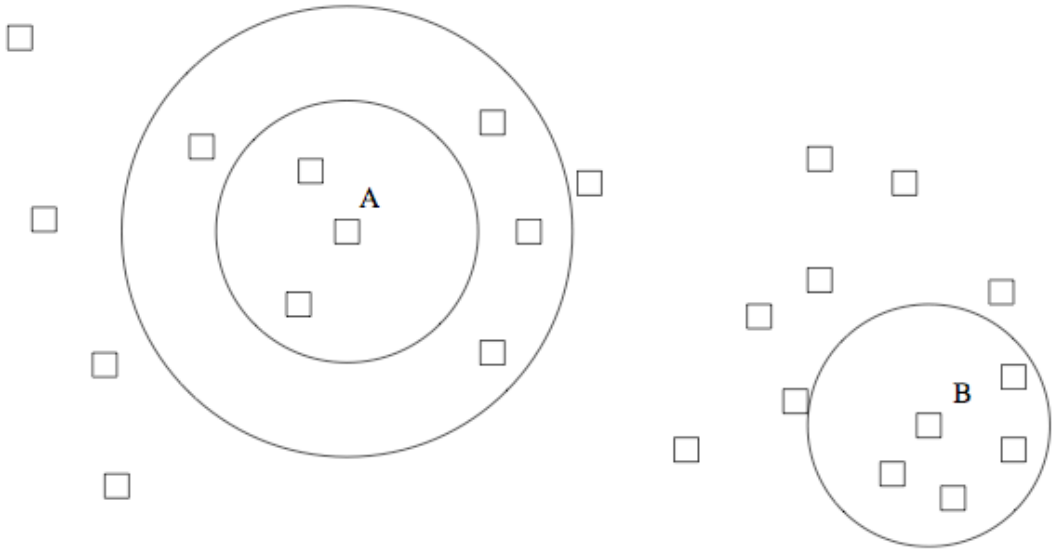
a. Classification supervisée :
   - Méthode de Bayes naïf
   - **k plus proches voisins**
   - Arbres de décision
   - Réseaux de neurones
b. Classification non supervisée : k-means
c. Évaluation des méthodes
d. Règles d'association et motifs séquentiels

UNIVERSITÉ MONTPELLIER 2
SCIENCES ET TECHNIQUES

Apprentissage supervisé : K plus proches voisins

- Objectifs et Principe :
  - Objectif : Pouvoir prédire simplement la classe d'un nouvel exemple
  - Principe : utiliser les exemples déjà connus

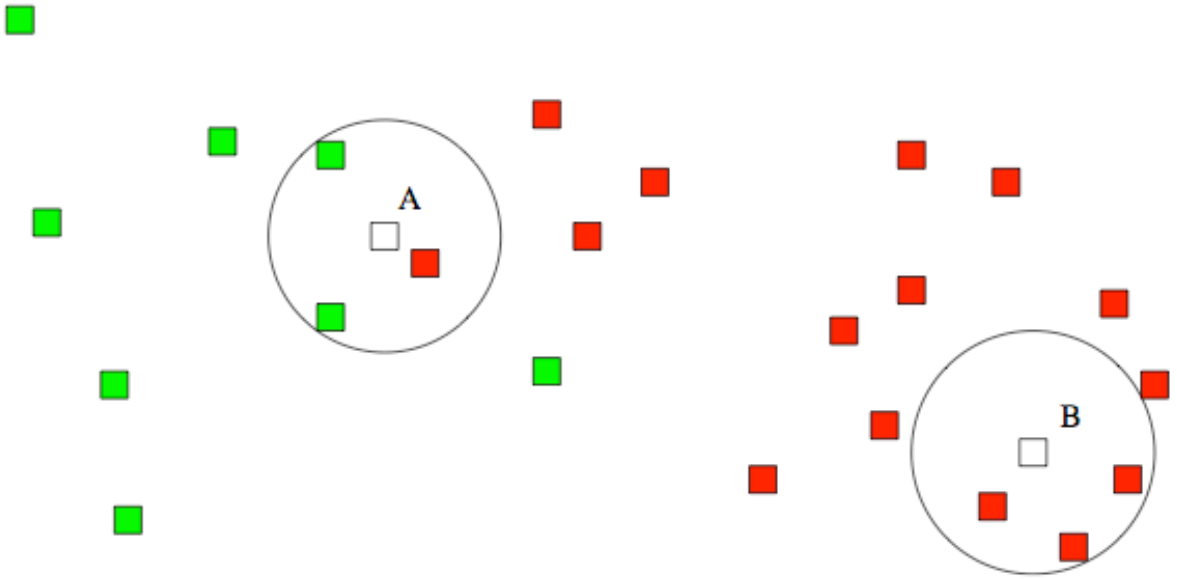
Exemple



Apprentissage supervisé : K plus proches voisins

- Principe
  - Regarder la classe des k exemples les plus proches (k = 1, 3, . . .)
  - Affecter la classe majoritaire au nouvel exemple

Exemple : deux classes, k = 3



Apprentissage supervisé : K plus proches voisins

- Notations
  - Soit L = {(x , c)|x ∈ R , c ∈ C} l'ensemble d'apprentissage
  - Soit x l'exemple dont on souhaite déterminer la classe

Algorithme

début
  pour chaque *(exemple* $(x', c) \in \mathcal{L}$*)* **faire**
    Calculer la distance $D(x, x')$
  fin
  pour chaque $\{x' \in kppv(x)\}$ **faire**
    compter le nombre d'occurences de chaque classe
  fin
  Attribuer à x la classe la plus fréquente;
fin

Apprentissage supervisé : K plus proches voisins

- Raisonnement à partir de cas
- Utilisation des cas similaires pour prendre une décision
- Etapes :
  1. On considère que l'on dispose d'une base d'apprentissage constituée de X éléments dont on connaît la classe,

Apprentissage supervisé : K plus proches voisins

- Raisonnement à partir de cas
- Utilisation des cas similaires pour prendre une décision
- Etapes :

1. On considère que l'on dispose d'une base d'apprentissage constituée de X éléments dont on connaît la classe,
2. Dès que l'on reçoit un nouvel élément à classifier on calcule sa distance avec tous les éléments de la base d'apprentissage

Rappels de distance : $d(A,A) = 0$; $d \Leftrightarrow d(A,B) = d(B,A)$; $d(A,C) < d(A,B) + d(B,C)$

Distance sur chacun des attributs : d(x,y) = |x-y| ou d(x,y) = |x-y| / distance_max

puis combinaison. distance euclidienne :

$$d(x,y) = \sqrt{[d1(x1,y1)\ 2 + \ldots + dn(xn,yn)2]}$$

Apprentissage supervisé : K plus proches voisins

- Raisonnement à partir de cas
- Utilisation des cas similaires pour prendre une décision
- Etapes :

1. On considère que l'on dispose d'une base d'apprentissage constituée de X éléments dont on connaît la classe,
2. Dès que l'on reçoit un nouvel élément à classifier on calcule sa distance avec tous les éléments de la base d'apprentissage

   Rappels de distance : $d(A,A) = 0$; $d \Leftrightarrow d(A,B) = d(B,A)$; $d(A,C) < d(A,B) + d(B,C)$

   Distance sur chacun des attributs : d(x,y) = |x-y| ou d(x,y) = |x-y| / distance_max

   puis combinaison. distance euclidienne :

$$d(x,y) = \sqrt{[d1(x1,y1)\ 2 + \ldots + dn(xn,yn)2]}$$

3. On sélectionne ensuite les k voisins les plus proches
   - Décider du nombre de voisins à utiliser k (souvent k = nbre d'attributs + 1)

Apprentissage supervisé : K plus proches voisins

- Raisonnement à partir de cas
- Utilisation des cas similaires pour prendre une décision
- Etapes :

1. On considère que l'on dispose d'une base d'apprentissage constituée de X éléments dont on connaît la classe,
2. Dès que l'on reçoit un nouvel élément à classifier on calcule sa distance avec tous les éléments de la base d'apprentissage

   Rappels de distance : $d(A,A) = 0$; $d \Leftrightarrow d(A,B) = d(B,A)$; $d(A,C) < d(A,B) + d(B,C)$

   Distance sur chacun des attributs : d(x,y) = |x-y| ou d(x,y) = |x-y| / distance_max

   puis combinaison. distance euclidienne :

$$d(x,y) = \sqrt{[d1(x1,y1)\ 2 + \ldots + dn(xn,yn)2]}$$

3. On sélectionne ensuite les k voisins les plus proches
   - Décider du nombre de voisins à utiliser k (souvent k = nbre d'attributs + 1)
4. On attribue au nouvel élément la classe parmi ces k voisins
   - Classe majoritaire ou classe pondérée

Apprentissage supervisé : K plus proches voisins

- Le résultat change en fonction de tous ces choix (distance, combinaison, calcul de la classe)

- Exemple : va-t-on jouer au tennis avec cette méthode ?
  - on choisit k = 4
  - distance euclidienne
    - d(A,A)=0
    - d(A,B)=1
  - calcul des voisins
  - combinaison des classes des voisins

Méthode des K plus proches voisins : Exercice : Va-t-on jouer au tennis?

	TEMPS	HUMIDITE	VENT	TENNIS
Ex1	Soleil	Haute	Oui	Oui
Ex2	Soleil	Basse	Non	Non
Ex3	nuageux	Basse	Oui	Oui
Ex4	pluvieux	Haute	Oui	Non
Ex5	pluvieux	Basse	Oui	Non
Ex6	Soleil	Basse	Oui	Oui
Ex7	pluvieux	Basse	Non	Non
	Soleil	***haute***	***Non***	***?***

Va-t-on jouer s'il y a du soleil, beaucoup d'humidité et pas de vent ?

Méthode des K plus proches voisins : Exemple : Va-t-on jouer au tennis? Résultats

Exemples	Combinansons				Distances			
	Temps	Humidite	Vent	Tennis	Temps	Humidite	Vent	Résultat
Ex1	Soleil	Haute	Oui	**Oui**	0	0	1	**1**
Ex2	Soleil	Basse	Non	**Non**	0	1	0	**1**
Ex3	nuageux	Basse	Oui	Oui	1	1	1	1,73
Ex4	Pluvieux	Haute	Oui	**Non**	1	0	1	**1,41**
Ex5	Pluvieux	Basse	Oui	Non	1	1	1	1,73
EX6	Soleil	Basse	Oui	**Oui**	0	1	1	**1,41**
EX7	Pluvieux	Basse	Non	Non	1	1	0	1,41

Va-t-on jouer s'il y a du soleil, beaucoup d'humidité et pas de vent ?

→ **La question reste entière… des solutions?**

Apprentissage supervisé : K plus proches voisins

Quelle décision prendre en cas d'égalité ?

- Augmenter la valeur de k de 1 pour trancher. L'ambiguïté peut persister
- Tirer au hasard la classe parmi les classes ambiguës.
- Pondération des exemples par leur distance au point x

II. Algorithmes de ce cours pour la classification

a. Classification supervisée :
   - Méthode de Bayes naïf
   - k plus proches voisins
   - **Arbres de décision**
   - Réseaux de neurones
b. Classification non supervisée : k-means
c. Évaluation des méthodes
d. Règles d'association et motifs séquentiels

UNIVERSITÉ MONTPELLIER 2
SCIENCES ET TECHNIQUES

Apprentissage supervisé : Arbre de décisions

- Représentation graphique d'une procédure de décision
- Représentation compréhensive ⇒ règles

- Définition

  Un arbre de décision est un arbre au sens informatique du terme :

  - les noeuds d'un arbre sont repérés par des positions qui sont des mots sur {1,...,p}*, où p est l'arité maximale des noeuds.

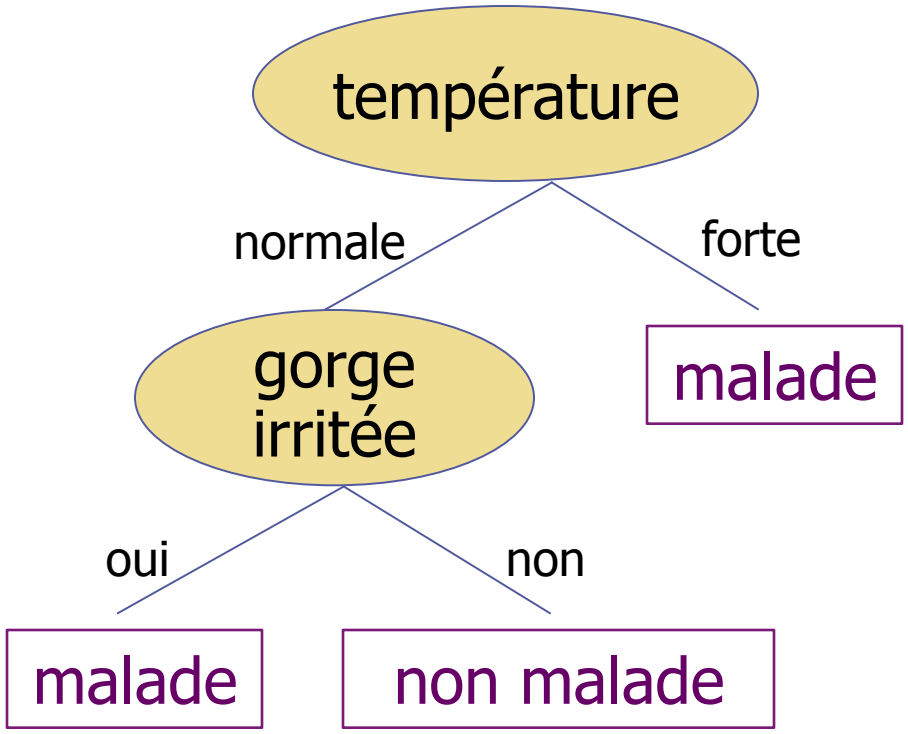
Apprentissage supervisé : Arbre de décisions

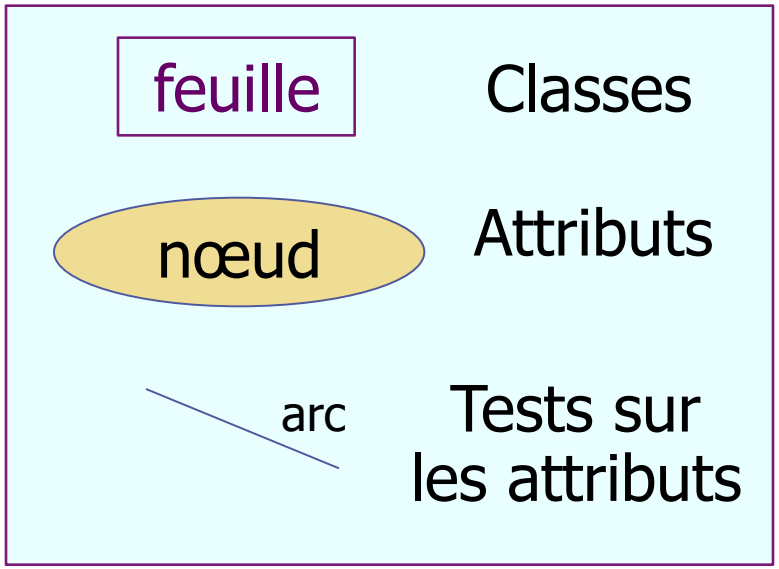
Exemple :

Si on note le mot vide par e, les positions pour l'arbre suivant sont :

- e étiquetée par le test Température ; 1 étiquetée par le test Gorge irritée,
- 2 étiquetée par la feuille malade ; 2.1 étiquetée par la feuille malade,

- 2.2 étiquetée par la feuille bien portant.





Apprentissage supervisé : Arbre de décisions

Définitions :

- Un arbre de décision est la représentation graphique d'une procédure de classification
  - à toute description complète est associée une seule feuille de l'arbre de décision.
  - Cette association est définie en commençant à la racine de l'arbre et en descendant dans l'arbre selon les réponses aux tests qui étiquettent les noeuds internes.
  - La classe associée est alors la classe par défaut associée à la feuille qui correspond à la description.
  - La procédure de classification obtenue a une traduction immédiate en terme de règles de décision.
  - Les systèmes de règles obtenus sont particuliers car l'ordre dans lequel on examine les attributs est fixé et les règles de décision sont mutuellement exclusives.

Apprentissage supervisé : Arbre de décisions

Définitions :

- Les noeuds
- Les noeuds internes sont appelés noeuds de décision : test qui peut être appliqué à toute description d'un individu de la population.
- En général, chaque test examine la valeur d'un unique attribut de l'espace des descriptions.
- Les réponses possibles au test correspondent aux labels des arcs issus de ce noeud.
    - Pour les noeuds de décision binaires :
        - les labels des arcs sont omis et, par convention,
        - l'arc gauche correspond à une réponse positive au test.
        - Les feuilles sont étiquetées par une classe appelée classe par défaut.

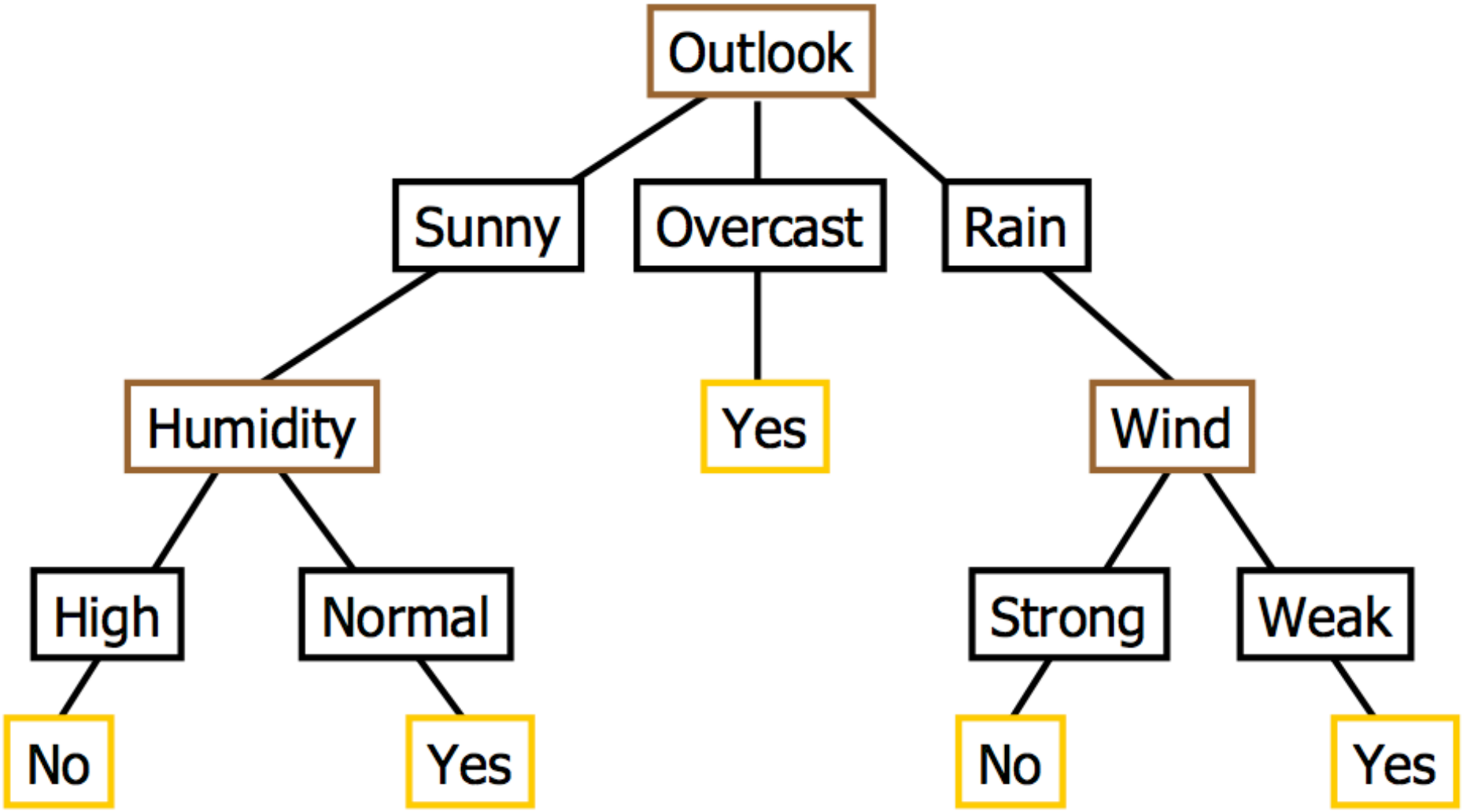
Apprentissage supervisé : Arbre de décisions

Définitions :

- Plusieurs algorithmes d'apprentissage par arbres de décision :
- Notations :
  - Étant donnés un échantillon S, un ensemble de classes {1,...,c} et un arbre de décision t,
  - à chaque position p de t correspond un sous-ensemble de l'échantillon qui est l'ensemble des exemples qui satisfont les tests de la racine jusqu'à cette position.
  - Par conséquent, on peut définir, pour toute position p de t, les quantités suivantes :
    - N(p) est le cardinal de l'ensemble des exemples associé à p,
    - N(k/p) est le cardinal de l'ensemble des exemples associé à p qui sont de classe k,
    - P(k/p) = N(k/p)/N(p) la proportion d'éléments de classe k à la position p.

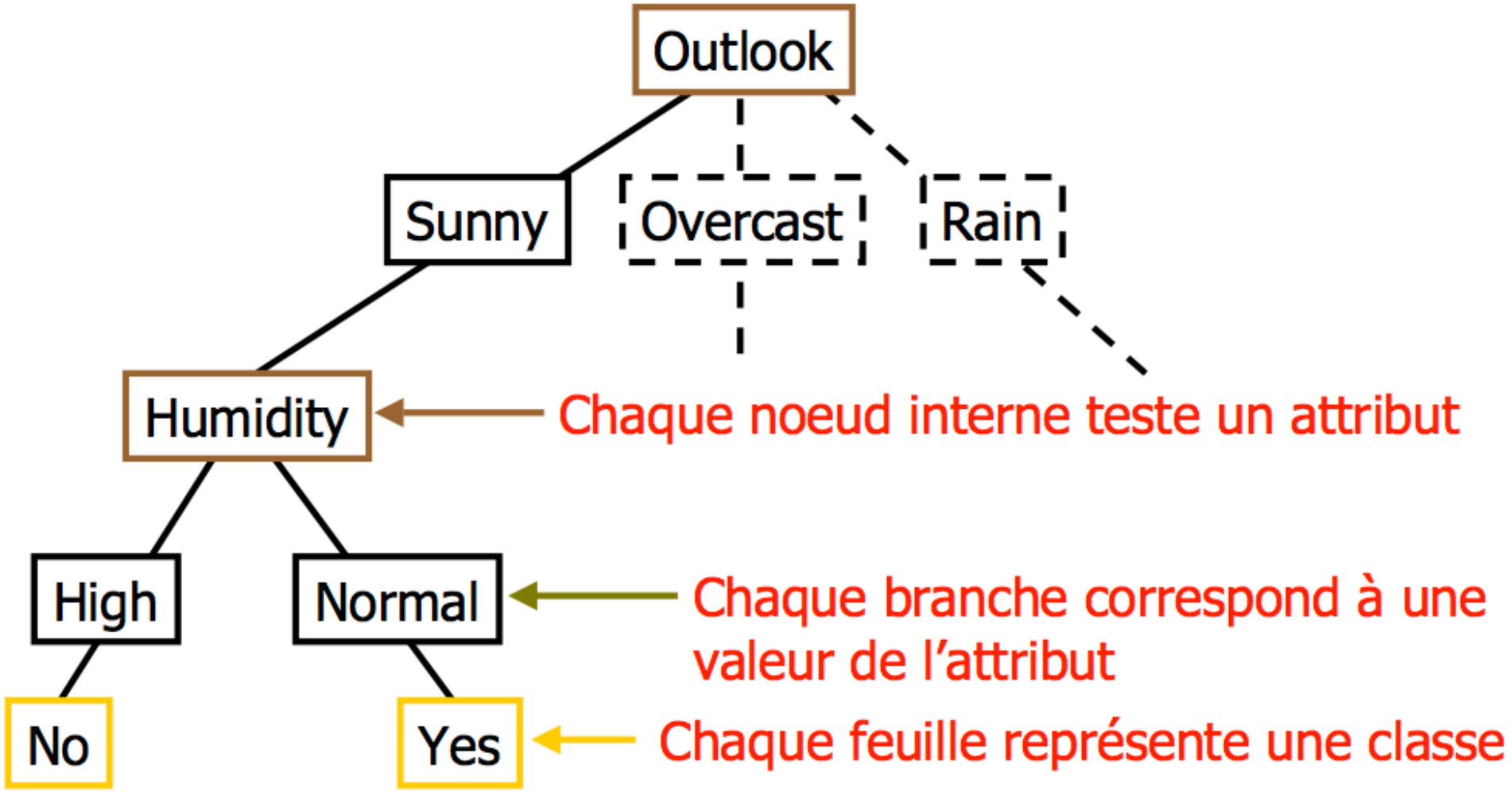
Apprentissage supervisé : Arbre de décisions

Exemples :



Apprentissage supervisé : Arbre de décisions

Exemples



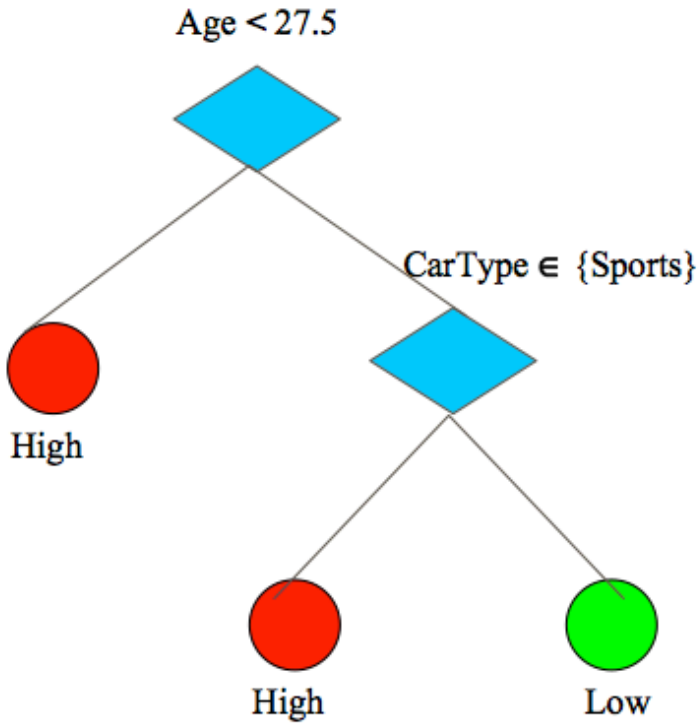
Apprentissage supervisé : Arbre de décisions

Exemples

Risque - Assurances

Tid	Age	Car Type	Class
0	23	Family	High
1	17	Sports	High
2	43	Sports	High
3	68	Family	Low
4	32	Truck	Low
5	20	Family	High

Numérique Enumératif



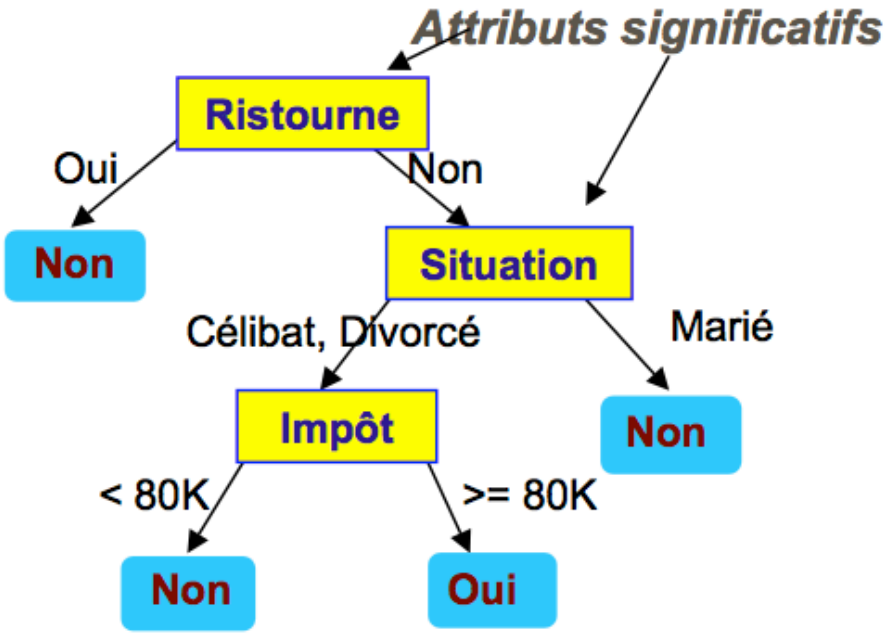
Age=40, CarType=Family ⟹ Class=Low

Apprentissage supervisé : Arbre de décisions

Exemples : détection de la fraude fiscale

énumératif énumératif numérique classe

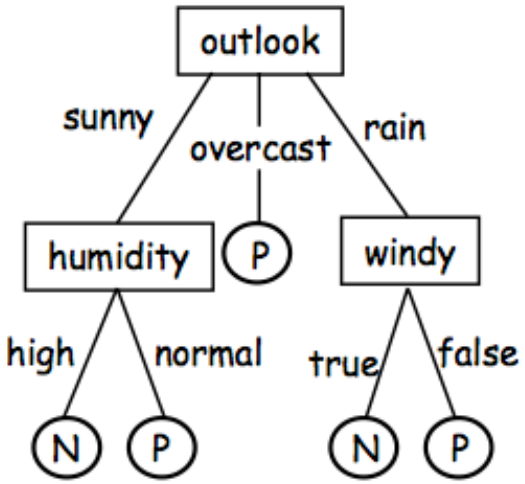
Id	Ristourne	Situation famille	Impôt revenu	Fraude
1	Oui	Célibat.	125K	Non
2	Non	Marié	100K	Non
3	Non	Célibat.	70K	Non
4	Oui	Marié	120K	Non
5	Non	Divorcé	95K	Oui
6	Non	Marié	60K	Non
7	Oui	Divorcé	220K	Non
8	Non	Célibat.	85K	Oui
9	Non	Marié	75K	Non
10	Non	Célibat.	90K	Oui



- **L'attribut significatif à un noeud est déterminé en se basant sur l'indice Gini.**
- **Pour classer une instance : descendre dans l'arbre selon les réponses aux différents tests. Ex = (Ristourne=Non, Situation=Divorcé, Impôt=100K) ➔ Oui**

Apprentissage supervisé : Arbre de décisions

De l'arbre de décision aux règles de classification :

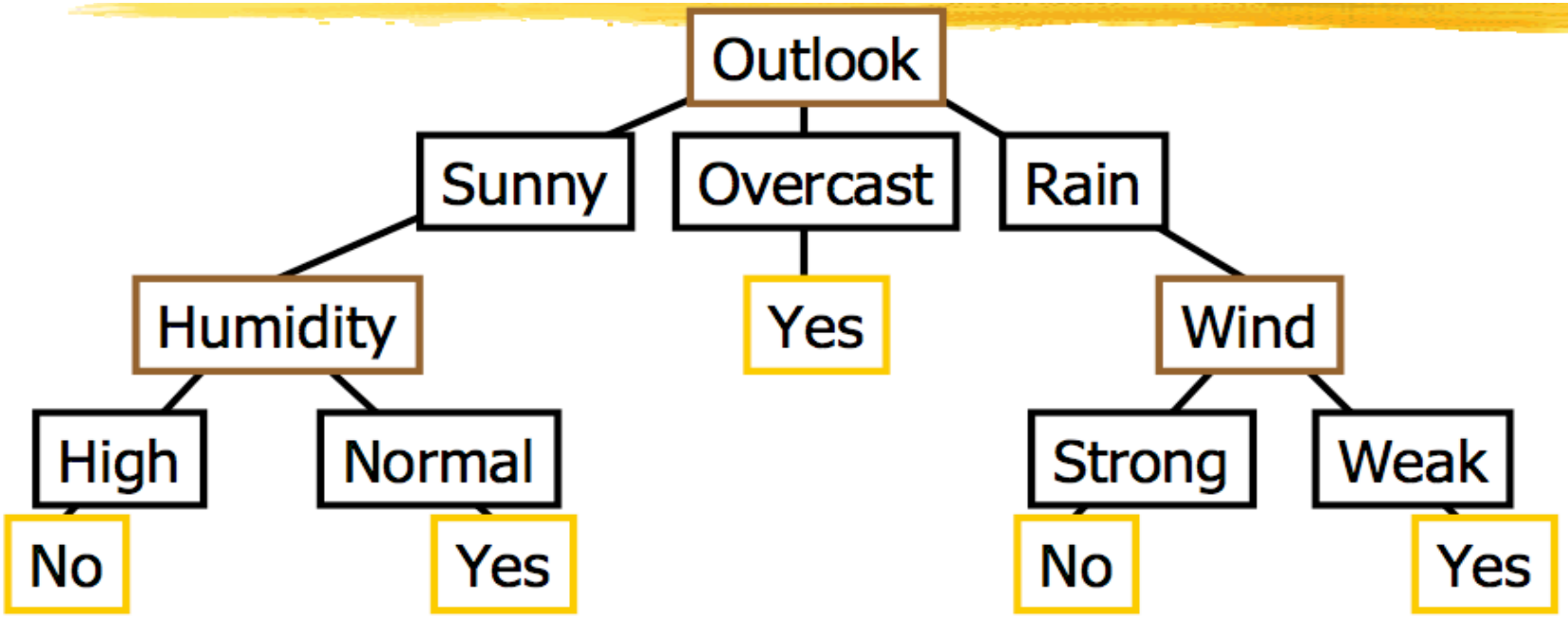


Si outlook=sunny
Et humidity=normal
Alors play tennis

- une **règle** est générée pour chaque **chemin** de l'arbre (de la racine à une feuille)
- Les paires attribut-valeur d'un chemin forment une conjonction
- Le noeud terminal représente la classe prédite
- Les règles sont généralement plus faciles à comprendre que les arbres

Apprentissage supervisé : Arbre de décisions

Des arbres de décision aux règles



R_1: If (Outlook=Sunny) $\wedge$ (Humidity=High) Then PlayTennis=No
R_2: If (Outlook=Sunny) $\wedge$ (Humidity=Normal) Then PlayTennis=Yes
R_3: If (Outlook=Overcast) Then PlayTennis=Yes
R_4: If (Outlook=Rain) $\wedge$ (Wind=Strong) Then PlayTennis=No
R_5: If (Outlook=Rain) $\wedge$ (Wind=Weak) Then PlayTennis=Yes

Apprentissage supervisé : Arbre de décisions

Génération de l'arbre de décision :

- Deux phases dans la génération de l'arbre :
  - Construction de l'arbre
    Arbre peut atteindre une taille élevée
  - Elaguer l'arbre (Pruning)

    Identifier et supprimer les branches qui représentent du "bruit" Améliorer le taux d'erreur

Apprentissage supervisé : Arbre de décisions

Algorithme de classification : Construction de l'arbre

1. Au départ, toutes les instances d'apprentissage sont à la racine de l'arbre

2. Sélectionner un attribut et choisir un test de séparation (split) sur l'attribut, qui sépare le "mieux" les instances.

   La sélection des attributs est basée sur une heuristique ou une mesure statistique.

3. Partitionner les instances entre les noeuds fils suivant la satisfaction des tests logiques Traiter chaque noeud fils de façon récursive

4. Répéter jusqu'à ce que tous les noeuds soient des terminaux. Un noeud courant est terminal si :

   - Il n'y a plus d'attributs disponibles

   - Le noeud est "pur", i.e. toutes les instances appartiennent à une seule classe,

   - Le noeud est "presque pur", i.e. la majorité des instances appartiennent à une seule classe (Ex : 95%)

   - Nombre minimun d'instances par branche (Ex : algorithme C5 évite la croissance de l'arbre, k=2 par défaut)

5. Etiqueter le noeud terminal par la classe majoritaire

Apprentissage supervisé : Arbre de décisions

Elaguer l'arbre obtenu (pruning)

- Supprimer les sous-arbres qui n'améliorent pas l'erreur de la classification (accuracy) arbre ayant un meilleur pouvoir de généralisation, même si on augmente l'erreur sur l'ensemble d'apprentissage
- Eviter le problème de sur-spécialisation (over- fitting), i.e., on a appris "par coeur" l'ensemble d'apprentissage, mais on n'est pas capable de généraliser
    - L'arbre généré peut sur- spécialiser l'ensemble d'apprentissage
        - Plusieurs branches
        - Taux d'erreur important pour les instances inconnues
    - Raisons de la sur-spécialisation
        - bruits et exceptions
        - Peu de donnée d'apprentissage
        - Maxima locaux dans la recherche gloutonne

Apprentissage supervisé : Arbre de décisions

Comment éviter l'overfitting ?

- Deux approches :
  - Pré-élagage : Arrêter de façon prématurée la construction de l'arbre
  - Post-élagage :
    - Supprimer des branches de l'arbre complet ("fully grown")
    - Convertir l'arbre en règles ; élaguer les règles de façon indépendante (C4.5)

Apprentissage supervisé : Arbre de décisions

Problématiques associées :

- Choix des attributs tests (divisions successives de la base d'apprentissage)
- Critère d'arrêt
- But : construire un arbre le plus petit possible
- Heuristique. Algorithme glouton.
- Plusieurs algorithmes (ID3, C4.5)

Apprentissage supervisé : Arbre de décisions

Algorithme de construction :

- Nœud Courant ← racine
- Répéter
  - Si le nœud courant est terminal
    - Alors l'étiqueter Nœud Courant ← Classe
  - Sinon
    - Sélectionner un attribut test
    - Créer le sous-arbre
  - Passer au nœud suivant non exploré
- Jusqu'à obtention d'un arbre

Apprentissage supervisé : Arbre de décisions

Critère d'arrêt :

- Plusieurs tests possibles pour décider si le nœud courant est terminal :
  - il n'y a plus assez d'exemples
  - les exemples ne sont pas trop mélangés (une classe se dégage). seuil d'impureté.

- On étiquette avec la classe majoritaire

Apprentissage supervisé : Arbre de décisions

Sélection de l'attribut test :

- Quel est l'attribut dont la connaissance nous aide le plus sur la classe ?
- Plusieurs critères possibles : test de Gini, gain d'information, entropie, ...
- ID3 : entropie de Shannon

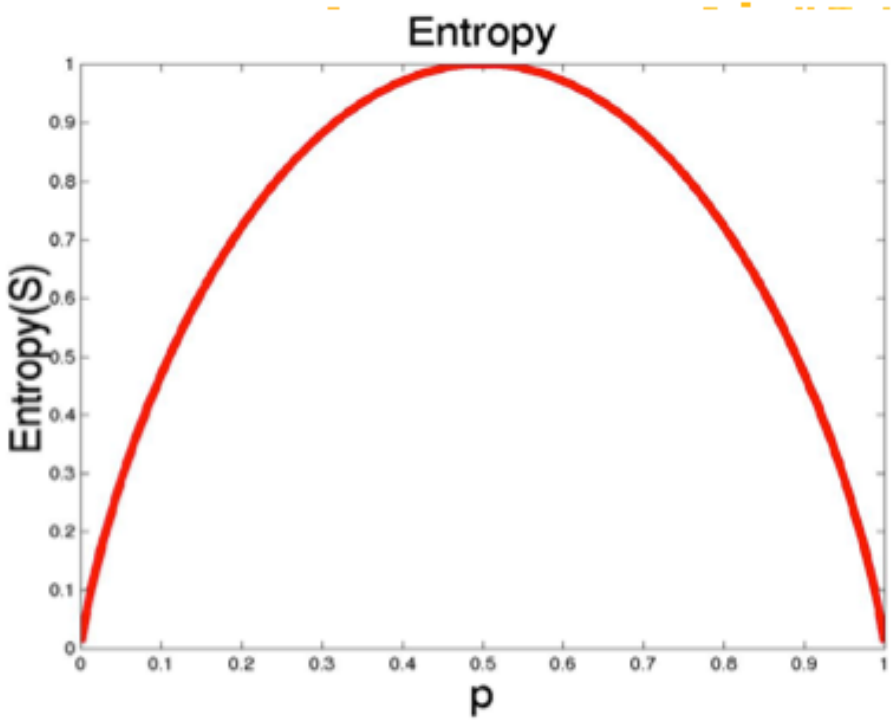
Apprentissage supervisé : Arbre de décisions

Gain d'information

- Sélectionner l'attribut avec le plus grand gain d'information
- Soient P et N deux classes et S un ensemble d'instances avec p éléments de P et n éléments de N
- L'information nécessaire pour déterminer si une instance prise au hasard fait partie de P ou N est (entropie) :

$$I(p,n) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n}$$

Apprentissage supervisé : Arbre de décisions



- S est l'ensemble d'apprentissage
- p+ est la proportion d'exemples positifs (P)
- p- est la proportion d'exemples négatifs (N)
- Entropie mesure l'impureté de S
- Entropie(S) = -p+ log2 p+ - p- log2 p-

Apprentissage supervisé : Arbre de décisions

Gain d'information

- Soient les ensembles {S , S , ..., S } formant une partition de l'ensemble S, en utilisant l'attribut A
- Toute partition Si contient pi instances de P et ni instances de N
- L'entropie, ou l'information nécessaire pour classifier les instances dans les sous-arbres S est :

$$E(A) = \sum_{i=1}^{\nu} \frac{p_i + n_i}{p+n} I(p_i, n_i)$$

- Le gain d'information par rapport au branchement sur A est

$$\text{Gain}(A) = I(p, n) - E(A)$$

- Choisir l'attribut qui maximise le gain → besoin d'information minimal

Apprentissage supervisé : Arbre de décisions

Exemple

- Sélectionner l'attribut avec le plus grand gain d'information
- Soient P et N deux classes et S un ensemble d'instances avec p éléments de P et n éléments de N
- L'information nécessaire pour déterminer si une instance prise au hasard fait partie de P ou N est (entropie) :

$$I(p,n) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n}$$

Apprentissage supervisé : Arbre de décisions

Exemple: Hypothèses :

- Classe P : jouer_tennis = “oui” ; Classe N : jouer_tennis = “non”
- Information nécessaire pour classer un exemple donné est :
  - I (p, n) = I (9,5) = 0.940

- Calculer l’entropie pour l’attribut outlook :

outlook	p_i	n_i	$I(p_i, n_i)$
sunny	2	3	0,971
overcast	4	0	0
rain	3	2	0,971

Apprentissage supervisé : Arbre de décisions

Exemple: Hypothèses :

- Classe P : jouer_tennis = “oui” ; Classe N : jouer_tennis = “non”
- Information nécessaire pour classer un exemple donné est :
  - I (p, n) = I (9,5) = 0.940

outlook	p_i	n_i	$I(p_i, n_i)$
sunny	2	3	0,971
overcast	4	0	0
rain	3	2	0,971

- Calculer l’entropie pour l’attribut outlook :
- On a :

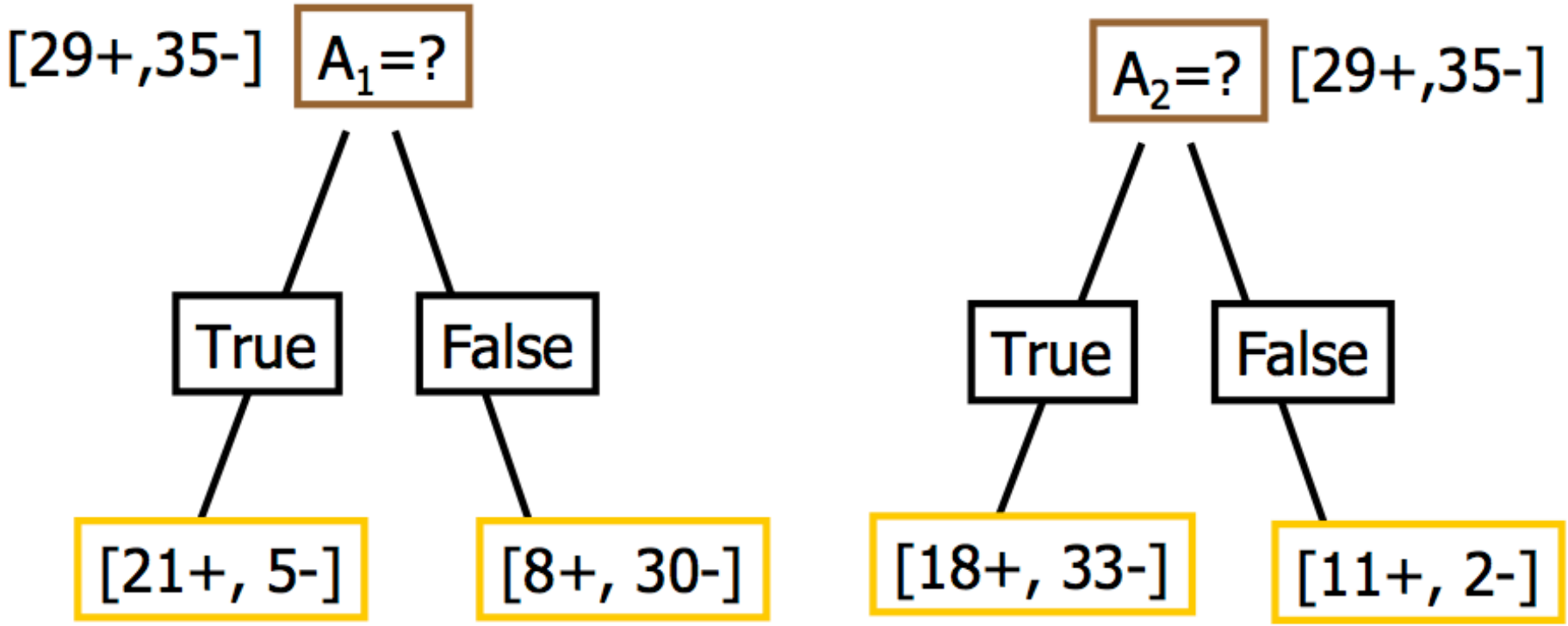
$$E(outlook) = \frac{5}{14} I(2,3) + \frac{4}{14} I(4,0) + \frac{5}{14} I(3,2) = 0.694$$

- Alors

  Gain (outlook) = I (9,5) − E (outlook) = 0.246
- De manière similaire
  - Gain(temperature) = 0.029 ; Gain(humidity) = 0.151 ; Gain(windy) = 0.048

Apprentissage supervisé : Arbre de décisions

Quel Attribut est ”meilleur” ?

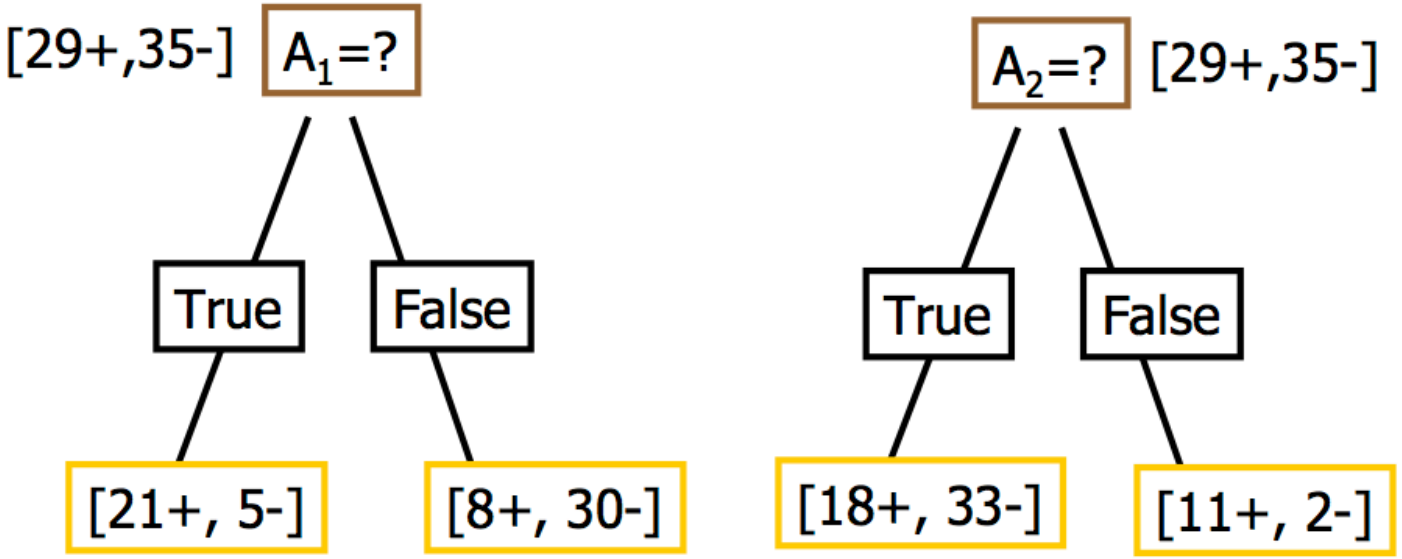


Apprentissage supervisé : Arbre de décisions

Gain(S,A) : réduction attendue de l'entropie dûe au branchement de S sur l'attribut A

$$Gain(S,A)=Entropie(S) - \sum_{v \in values(A)} |S_v|/|S| \; Entropie(S_v)$$

$$Entropie([29+,35-]) = -29/64 \log_2 29/64 - 35/64 \log_2 35/64 = 0.99$$



Apprentissage supervisé : Arbre de décisions

Entropie([21+,5-]) = 0.71

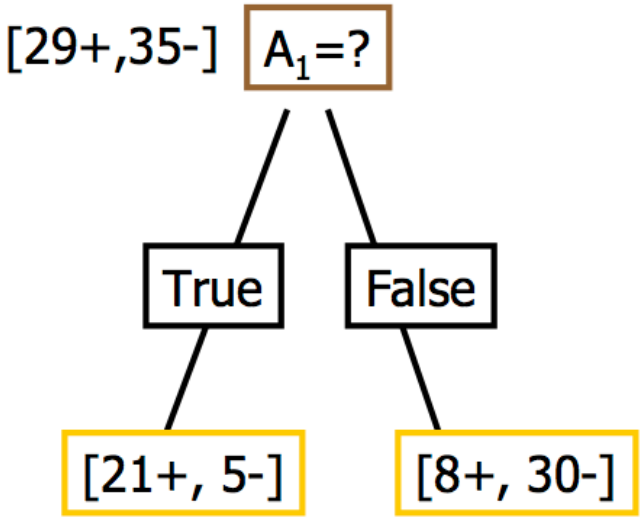
Entropie([8+,30-]) = 0.74

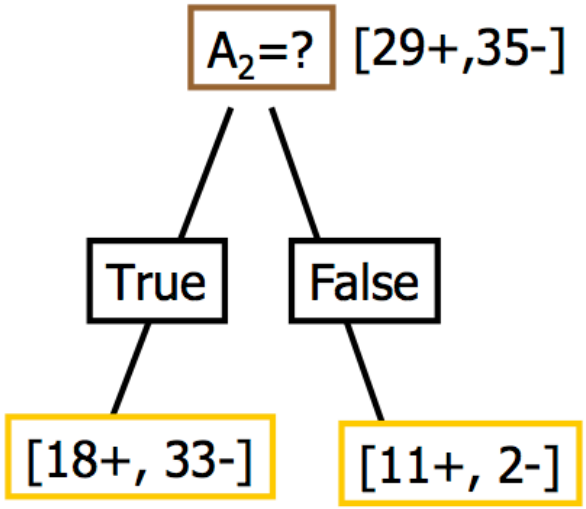
Gain(S,A1)=Entropie(S)
-26/64*Entropie([21+,5-])
-38/64*Entropie([8+,30-])
=0.27

Entropie([18+,33-]) = 0.94

Entropie([8+,30-]) = 0.62

Gain(S,A2)=Entropie(S)
-51/64*Entropie([18+,33-])
-13/64*Entropie([11+,2-])
=0.12





Apprentissage supervisé : Arbre de décisions

Exemple : va-t-on jouer au tennis avec cette méthode ?:

	TEMPS	HUMIDITE	VENT	TENNIS
Ex1	Soleil	Haute	Oui	Oui
Ex2	Soleil	Basse	Non	Non
Ex3	nuageux	Basse	Oui	Oui
Ex4	pluvieux	Haute	Oui	Non
Ex5	pluvieux	Basse	Oui	Non
Ex6	Soleil	Basse	Oui	Oui
Ex7	pluvieux	Basse	Non	Non
	Soleil	***haute***	***Non***	***?***

Va-t-on jouer s'il y a du soleil, beaucoup d'humidité et pas de vent ?

Apprentissage supervisé : Arbre de décisions

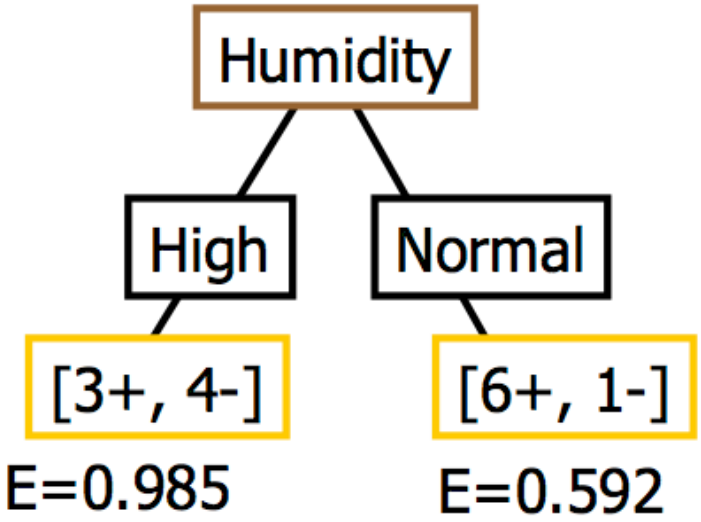
Exemple d'apprentissage

Day	Outlook	Temp.	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Weak	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cold	Normal	Weak	Yes
D10	Rain	Mild	Normal	Strong	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Apprentissage supervisé : Arbre de décisions

Sélection de l'attribut suivant

S=[9+,5-] et E=0.940



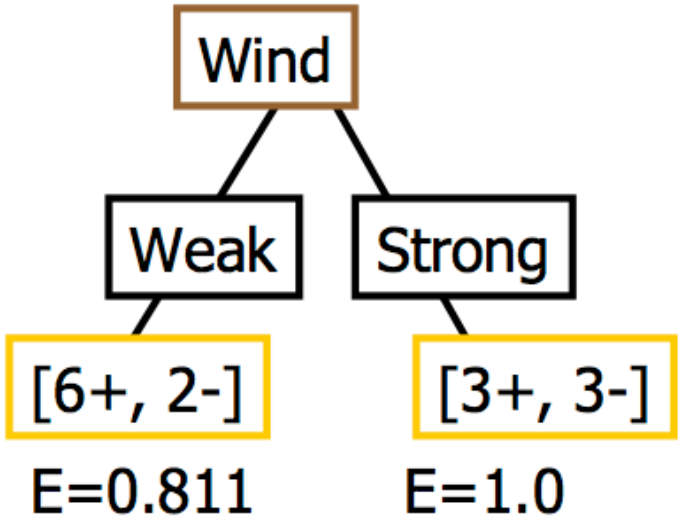
Gain(S,Humidity)

= 0.940-(7/14)*0.985

– (7/14)*0.592

= 0.151

S=[9+,5-] et E=0.940



Gain(S,Wind)

= 0.940-(8/14)*0.811

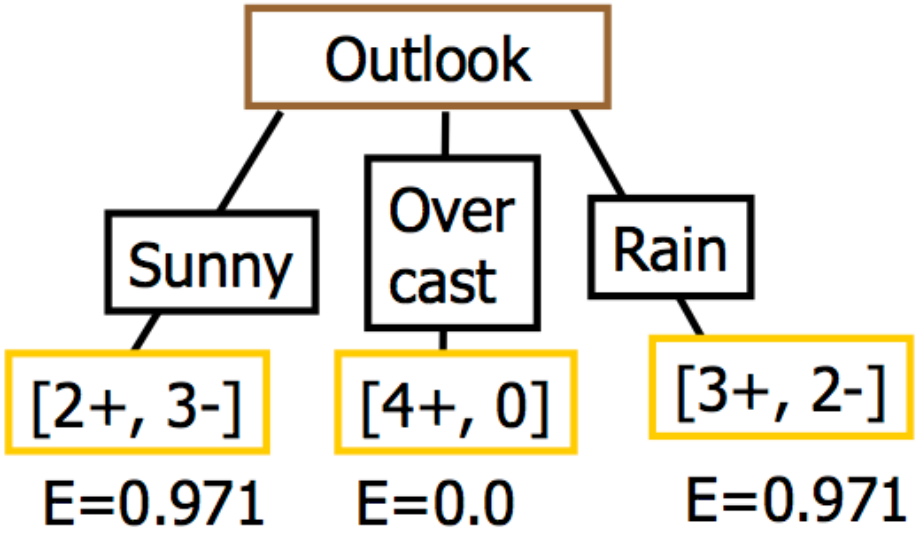
– (6/14)*1.0

= 0.048

Apprentissage supervisé : Arbre de décisions

Sélection de l'attribut suivant

S=[9+,5-] E=0.940



Gain(S,Outlook)

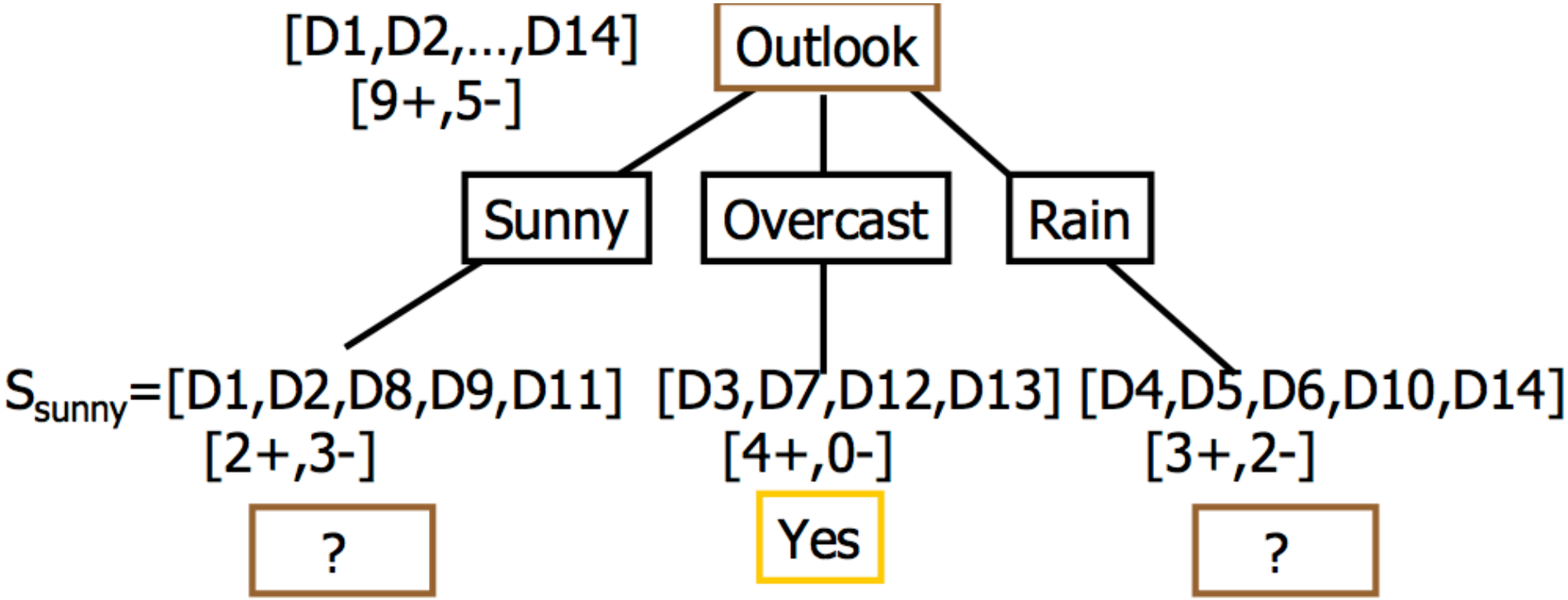
=0.940-(5/14)*0.971

-(4/14)*0.0 – (5/14)*0.0971

=0.247

Apprentissage supervisé : Arbre de décisions

Algorithme ID3



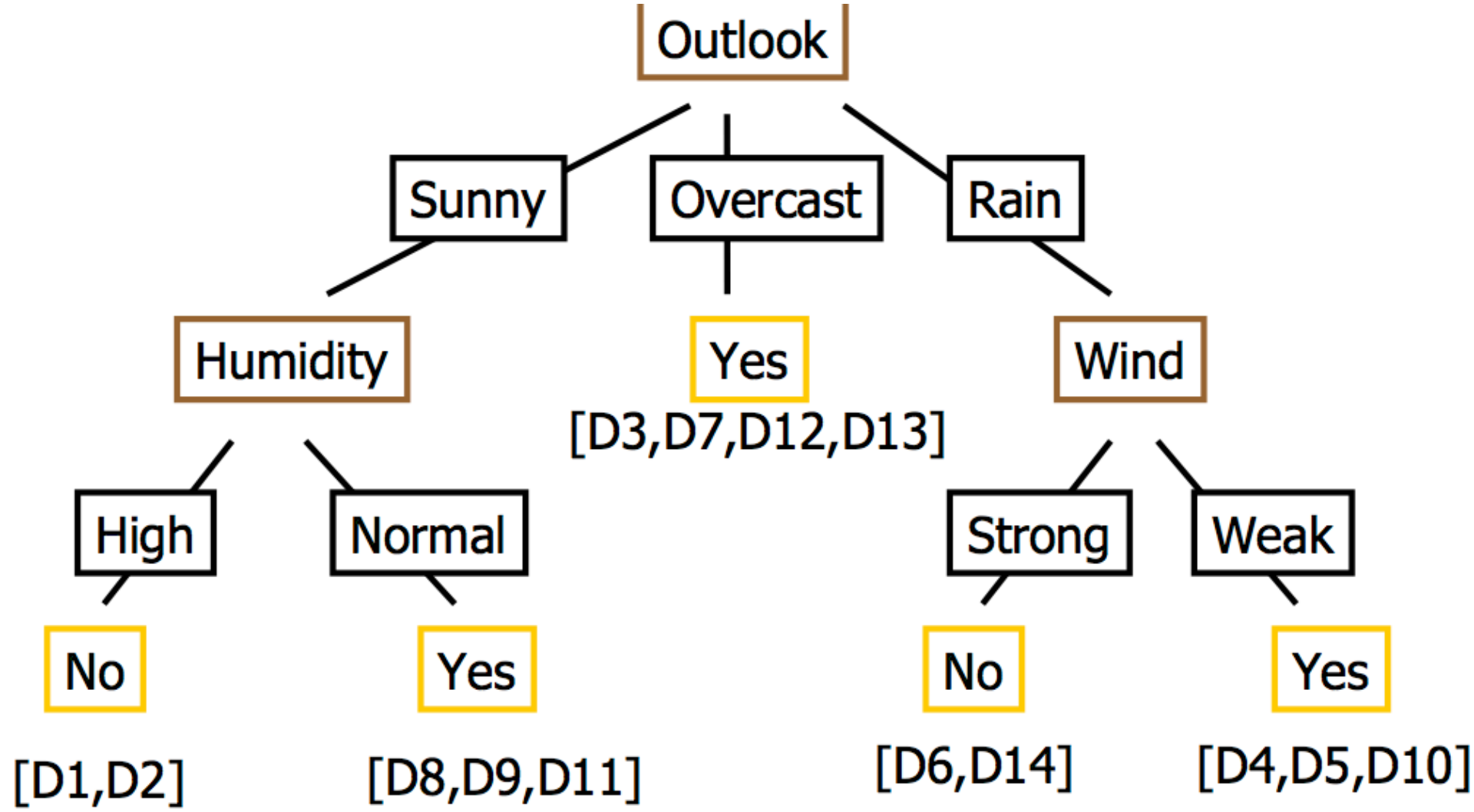
$Gain(S_{sunny}, Humidity) = 0.970 - (3/5)0.0 - 2/5(0.0) = 0.970$

$Gain(S_{sunny}, Temp.) = 0.970 - (2/5)0.0 - 2/5(1.0) - (1/5)0.0 = 0.570$

$Gain(S_{sunny}, Wind) = 0.970 = -(2/5)1.0 - 3/5(0.918) = 0.019$

Apprentissage supervisé : Arbre de décisions

Algorithme ID3



Apprentissage supervisé : Arbre de décisions

Avantages et inconvénients :

- attention au sur-apprentissage ⇒ élagage
- performances moins bonnes si beaucoup de classes
- algorithme non incrémental

- on peut expliquer une décision
- permet la sélection des attributs pertinents (feature selection)
- classification rapide d'un nouvel exemple (parcours d'un chemin d'arbre)

II. Algorithmes de ce cours pour la classification

a. Classification supervisée :
   - Méthode de Bayes naïf
   - k plus proches voisins
   - Arbres de décision
   - **Réseaux de neurones**

b. Classification non supervisée : k-means

c. Évaluation des méthodes

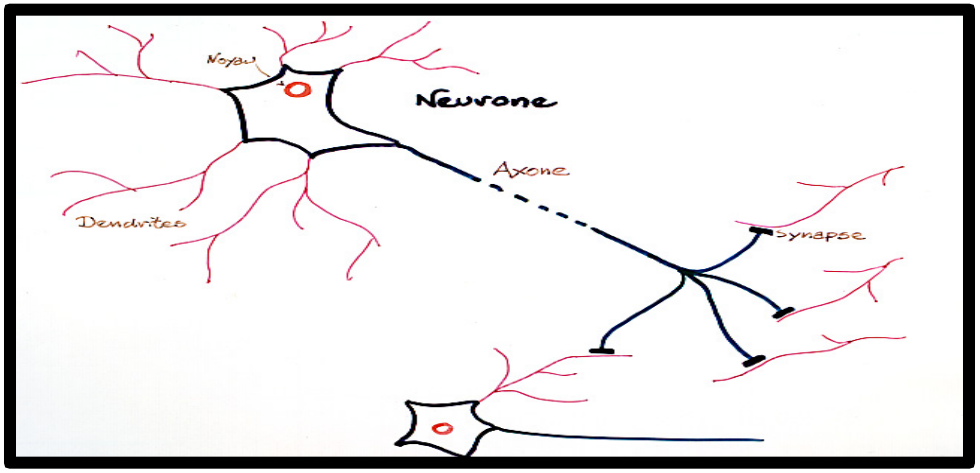
d. Règles d'association et motifs séquentiels

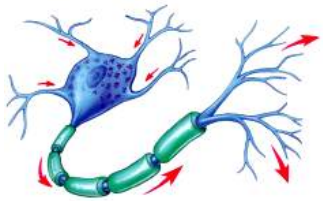
UNIVERSITÉ MONTPELLIER 2
SCIENCES ET TECHNIQUES

Apprentissage supervisé : Réseaux de neurones

Une méthode issue des modèles biologiques :

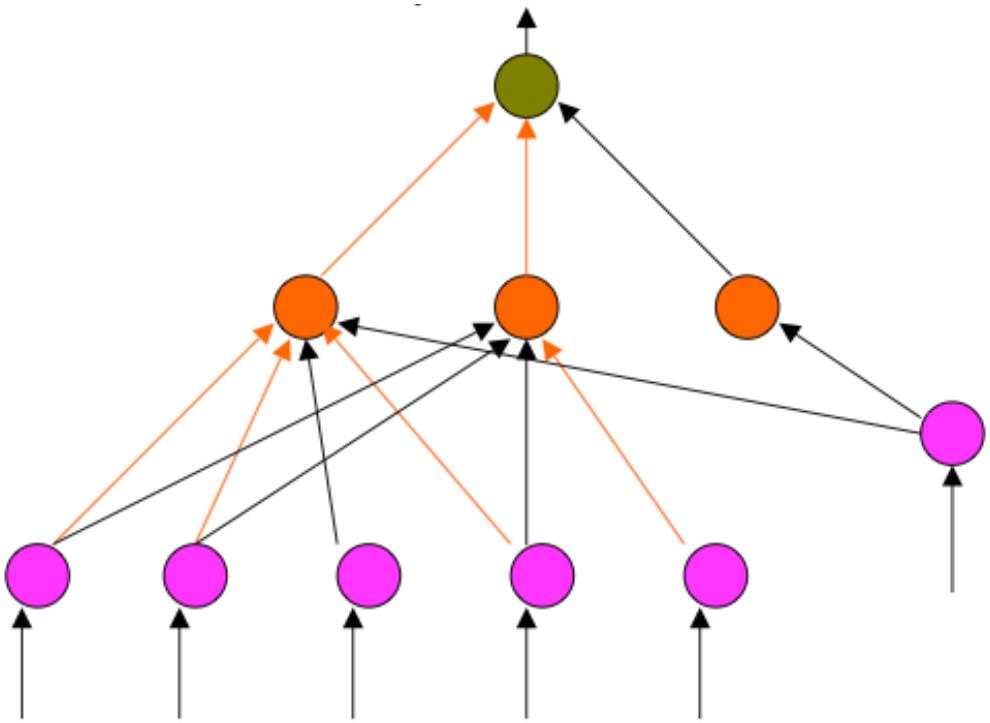
- Chaque neurone disposer en moyenne de 10.000 points de contacts (synapses) avec les neurones qui l'entourent, et jusqu'à 50.000 !
- Nous disposons de quelques dizaines de milliards de ces neurones à l'intérieur de notre cerveau
- De synapse en synapse, l'information transite dans la totalité de notre corps, au travers d'environ 500 000 milliards de synapses





Apprentissage supervisé : Réseaux de neurones

- Réseau neuronal : simule le système nerveux biologique
- Un réseau de neurones est composé de plusieurs neurones interconnectés. Un poids est associé à chaque arc. A chaque neurone on associe une valeur.

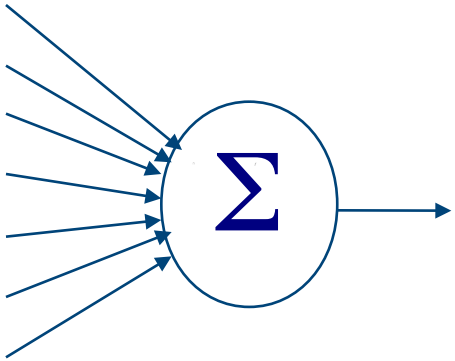


- Temps de ”switch” d’un neurone > 10-3 secs
- Nombre de neurones (humain) ~1010
- Connexions (synapses) par neurone : ~10^4 – 10^5

Apprentissage supervisé : Réseaux de neurones

Caractéristiques

- un neurone (biologique) est un noeud qui a plusieurs entrées et une sortie
- Les entrées proviennent d'autres neurones ou organes sensoriels
- Les entrées sont pondérées
- Les poids peuvent être positifs ou négatifs
- Les entrées sont sommées au niveau du noeud pour produire une valeur d'activation
- Si l'activation est plus grande qu'un certain seuil, le neurone s'active



Apprentissage supervisé : Réseaux de neurones

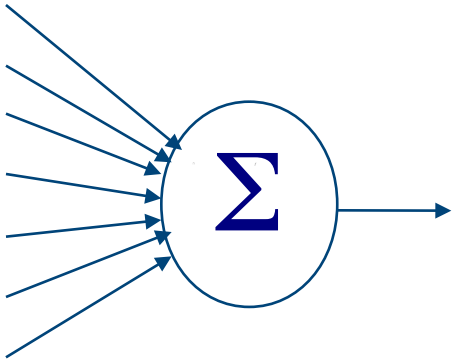
- Méthode de type boîte noire
- Nombreuses applications (notamment reconnaissance d'écriture manuscrite)
- Méthode coûteuse en temps de calcul
- Topologie à connaître

Apprentissage supervisé : Réseaux de neurones

Caractéristiques

- un neurone (biologique) est un noeud qui a plusieurs entrées et une sortie
- Les entrées proviennent d'autres neurones ou organes sensoriels
- Les entrées sont pondérées
- Les poids peuvent être positifs ou négatifs
- Les entrées sont sommées au niveau du noeud pour produire une valeur d'activation
- Si l'activation est plus grande qu'un certain seuil, le neurone s'active

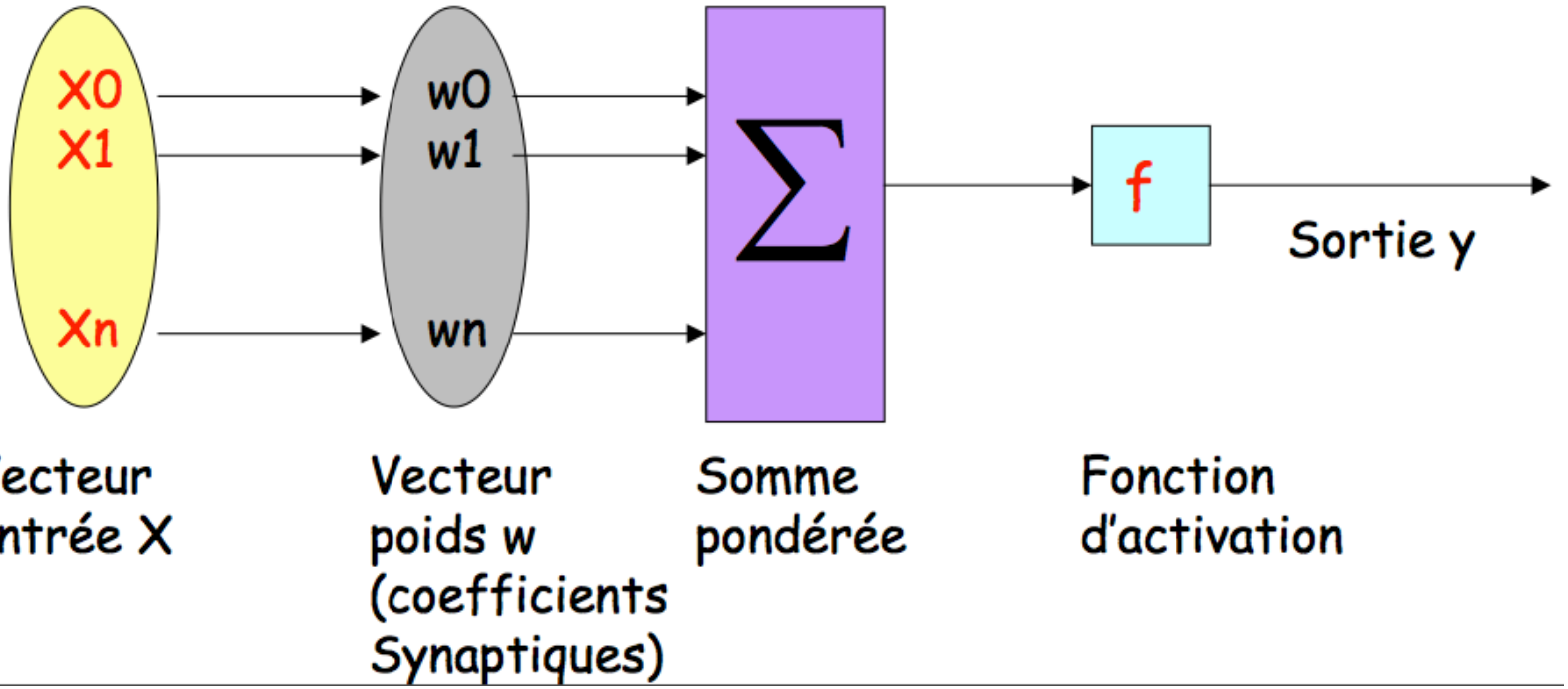
Va-t-on jouer au tennis avec cette méthode ?



Apprentissage supervisé : Réseaux de neurones

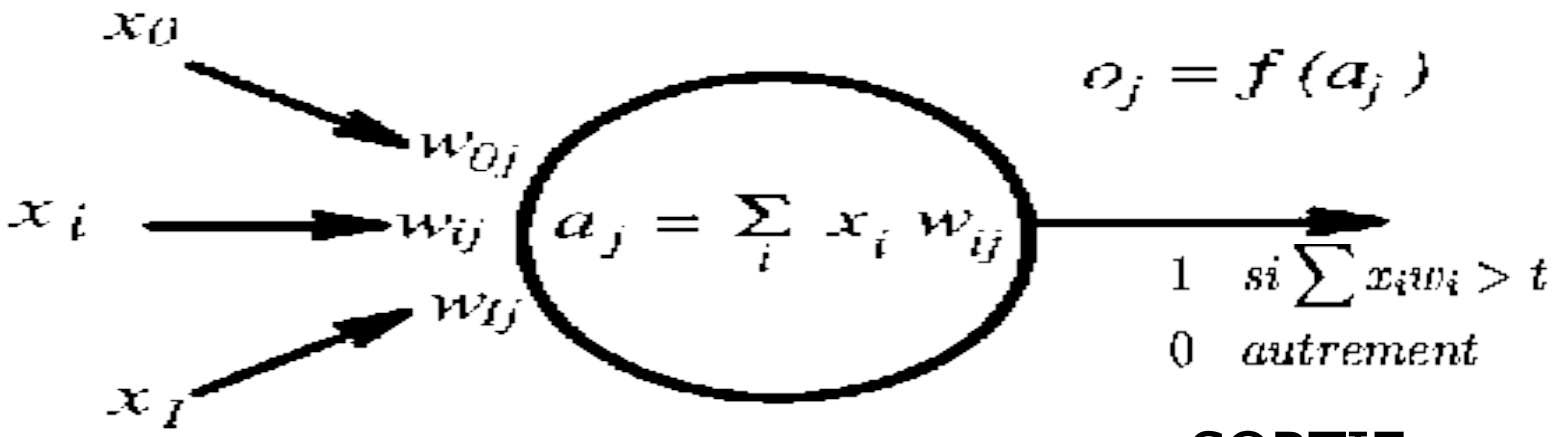
Neurone ou perceptron :

- Neurone = Unité de calcul élémentaire
- Le vecteur d'entrée X est transformé en une variable de sortie y, par un produit scalaire et une fonction de transformation non linéaire

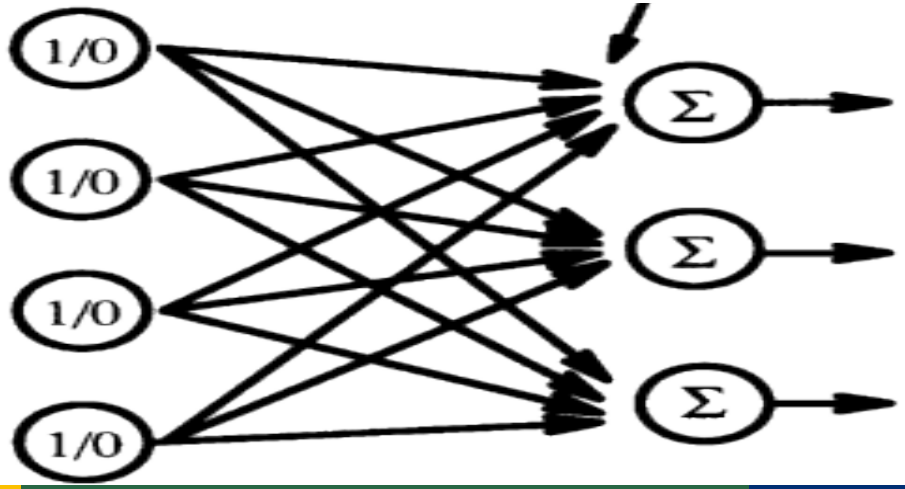


Apprentissage supervisé : Réseaux de neurones

Approche informatique : perception

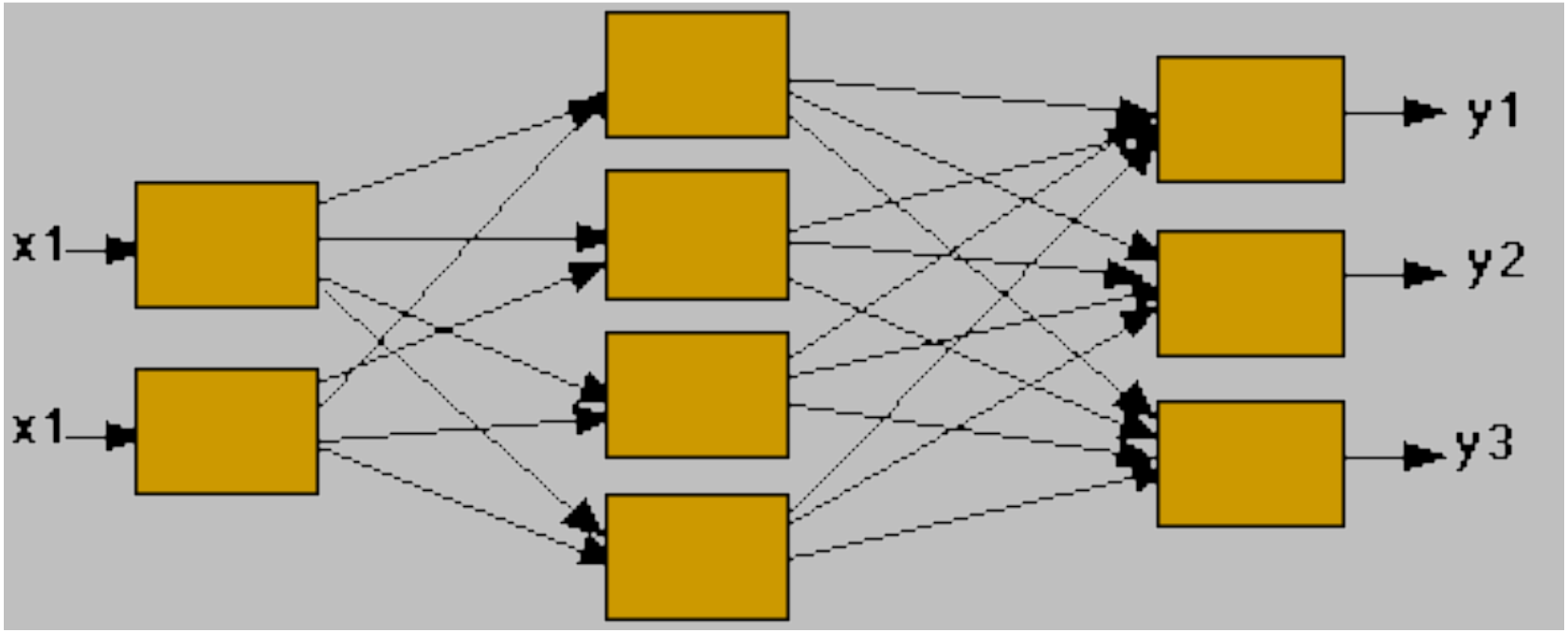


PLUSIEURS CLASSES



Apprentissage supervisé : Réseaux de neurones

Perception multi-couches:



COUCHE D'ENTREE

COUCHE DE SORTIE

COUCHE CACHEE

Apprentissage supervisé : Réseaux de neurones

Algorithme d'apprentissage des poids :

- Initialiser les poids de manière aléatoire
- Répéter

  Pour chaque exemple i

  Si la sortie s n'est pas égale à la sortie attendue a

  Alors poids wi ⇐ wi + (a - s)xi
- Jusqu'à ce que tous les exemples soient bien classés

Apprentissage supervisé : Réseaux de neurones

Algorithme d'apprentissage des poids :

- Capaciité d 'apprentissage : apprendre et changer son comportement en fonction de toute nouvelle expérience.
- Permettent de découvrir automatiquement des modèles complexes.
- Plusieurs modèles de réseaux de neurones : PMC (Perceptron Multi-Couches), RBF (Radial Basis Function), Kohonen,…

Apprentissage supervisé : Réseaux de neurones

Avantages :

- Taux d'erreur généralement bon
- Outil disponible dans les environnements de data mining
- Robustesse (bruit) – reconnaissance de formes (son, images sur une rétine, ...)
- Classification rapide (réseau étant construit)
- Combinaison avec d'autres méthodes (ex : arbre de décision pour sélection d'attributs)

Inconvénients :

- Apprentissage très long
- Plusieurs paramètres (architecture, coefficients synaptiques, ...)
- Pouvoir explicatif faible (boite noire)
- Pas facile d'incorporer les connaissances du domaine.
- Traitent facilement les attributs numériques et binaires
- Evolutivité dans le temps (phase d'apprentissage)

II. Algorithmes de ce cours pour la classification

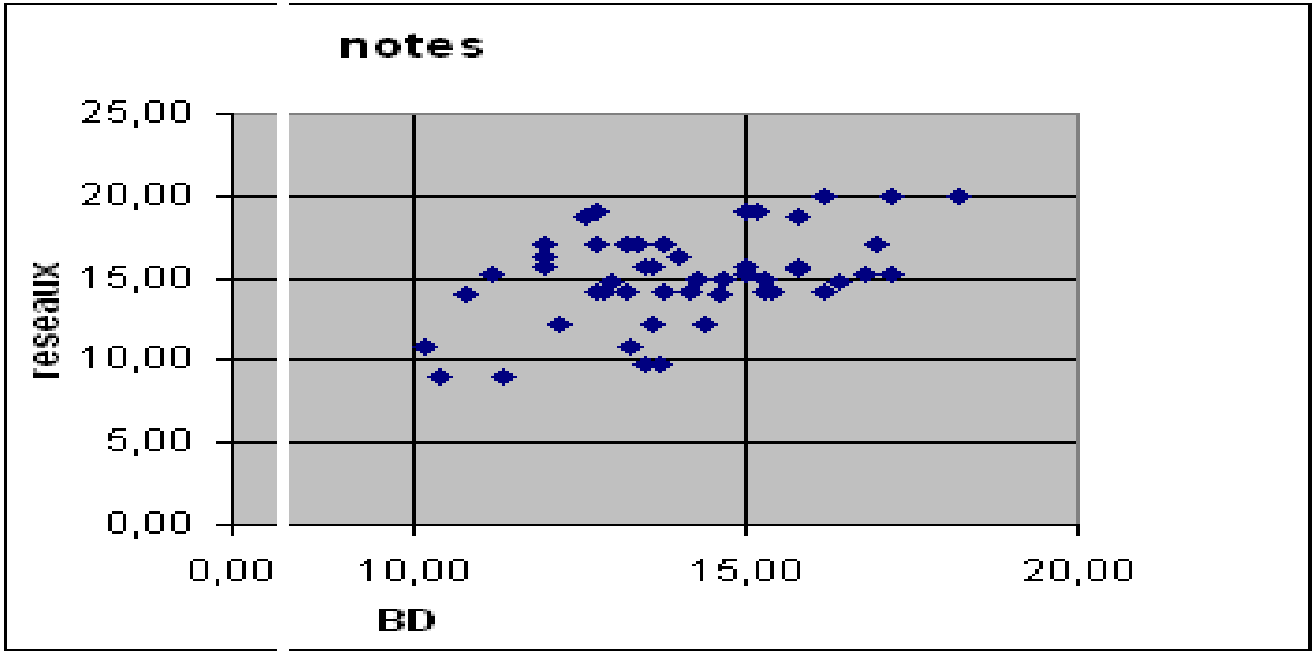
a. Classification supervisée :
   - Méthode de Bayes naïf
   - k plus proches voisins
   - Arbres de décision
   - Réseaux de neurones
b. **Classification non supervisée : k-means**
c. Évaluation des méthodes
d. Règles d'association et motifs séquentiels

UNIVERSITÉ MONTPELLIER 2
SCIENCES ET TECHNIQUES

Apprentissage non supervisé : Segmentation

Objectifs :

- diviser la population en groupes
- Minimiser la similarité intra-groupe
- Maximiser la similarité inter-groupes
- Exemple : notes des IG2 2002-2003



Apprentissage non supervisé : Les k moyennes

Algorithme :

1. Choisir le nombre de groupes à créer k
2. Choisir k centres initiaux c1, .., ck
3. Pour chaque exemple
   - l'affecter au groupe i dont le centre est le plus proche
4. Si aucun exemple ne change de groupe
   - Alors STOP
5. Sinon

   Calculer les nouveaux centres
   - Pour i = 1 à k

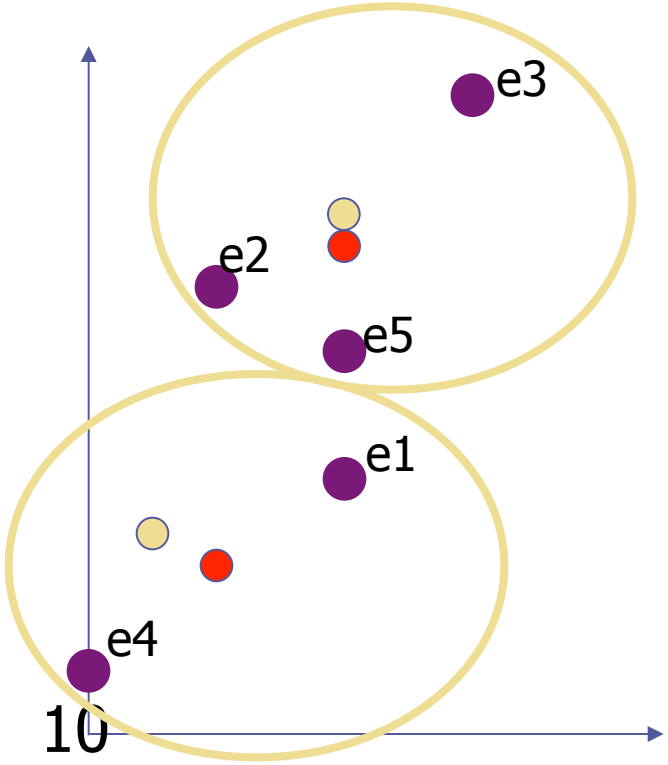
     ci est la moyenne des éléments du groupe
   - Aller en 3.

Fin

Apprentissage non supervisé : Les k moyennes

Exemple : faire 2 groupes d'étudiants

e1	14	14
e2	12	17
e3	16	20
e4	10	11
e5	14	16



- Centres initiaux : c1=(11,13) c2=(14,18)
  - d(e1,c1)= $[(14-11)^2 + (14-13)^2]^{1/2}$= **3.16**
  - d(e1,c2) = $[(14-14)^2 + (14-18)^2]^{1/2} \approx 4$
  - d(e2,c1)= 4.12 d(e2,c2) ≈ **2.24**
  - d(e3,c1) > d(e3,c2)
  - d(e4,c1) < d(e4,c2)
  - d(e5,c1) > d(e5,c2)
- Nouveaux centres :
  - c' 1 = ((14+10)/2,(14+11)/2) = (12,12.5)
  - c' 2 = ((12+16+14)/3),(17+20+16)/3)=(14,17.6)
- calcul de d(e1,c' 1) d(e1,c' 2) …
- résultat inchangé ⇒ FIN

Apprentissage non supervisé : Les k moyennes

Exercice : Nous avons

- 8 points A, …, H de l 'espace euclidéen 2D
- k=2 (2 groupes)

Points	Centre choisis : D (2, 4) et B (2, 2)	Centres ?
A (1, 3)	?	…
B (2, 2)	?	…
C (2, 3)	?	…
D (2, 4)	?	…
E (4, 2)	?	…
F (5, 2)	?	…
G (6, 2)	?	…
H (7, 3)	?	…

Apprentissage non supervisé : Les k moyennes

Solution

points	Centre D(2,4), B(2,2)	Centre D(2,4), I(27/7,17/7)	Centre J(5/3,10/3), K(24/5,11/5)
A(1,3)	B	D	J
B(2,2)	B	I	J
C(2,3)	B	D	J
D(2,4)	D	D	J
E(4,2)	B	I	K
F(5,2)	B	I	K
G(6,2)	B	I	K
H(7,3)	B	I	K

Apprentissage non supervisé : Les k moyennes

Avantages :

- Relativement extensible dans le traitement d'ensembles de taille importante
- Relativement efficace : O(t.k.n),

  où n représente # objets, k # clusters, et t # iterations.

  Normalement, k, t << n.
- Produit généralement un optimum local ; un optimum global peut être obtenu en utilisant d'autres techniques telles que : algorithmes génétiques, …

Inconvénients :

- Applicable seulement dans le cas où la moyenne des objets est définie
- Besoin de spécifier k, le nombre de clusters, a priori
- Incapable de traiter les données bruitées (noisy).
- Non adapté pour découvrir des clusters avec structures non-convexes, et des clusters de tailles différentes
- Les points isolés sont mal gérés (doivent-ils appartenir obligatoirement à un cluster ?) - probabiliste

II. Algorithmes de ce cours pour la classification

a. Classification supervisée :
   - Méthode de Bayes naïf
   - k plus proches voisins
   - Arbres de décision
   - Réseaux de neurones
b. Classification non supervisée : k-means
c. **Évaluation des méthodes**
d. Règles d'association et motifs séquentiels

UNIVERSITÉ MONTPELLIER 2
SCIENCES ET TECHNIQUES

Evaluation des méthodes

Apprentissage supervisé

- évaluation sur une base d'exemples test

Approches non supervisées

- Méthodes de séparation entre les bases d'apprentissage et de test.
- on dispose de deux bases séparées
- on coupe la base en deux
- validation croisée. Leave One Out.

Evaluation des méthodes

Validation croisée

- Découpage de la base d'exemples en n sous-base b1, ... , bn
- n apprentissages :
  - On apprend sur n-1 sous-bases
  - On teste sur la sous-base restante
  - Moyenne des n résultats
- n = 10 fonctionne bien
- Leave one out

Evaluation des méthodes

Validation croisée

- Découpage de la base d'exemples en n sous-base b1, ... , bn
- n apprentissages :
  - On apprend sur n-1 sous-bases
  - On teste sur la sous-base restante
  - Moyenne des n résultats
- n = 10 fonctionne bien
- Leave one out

Evaluation des méthodes

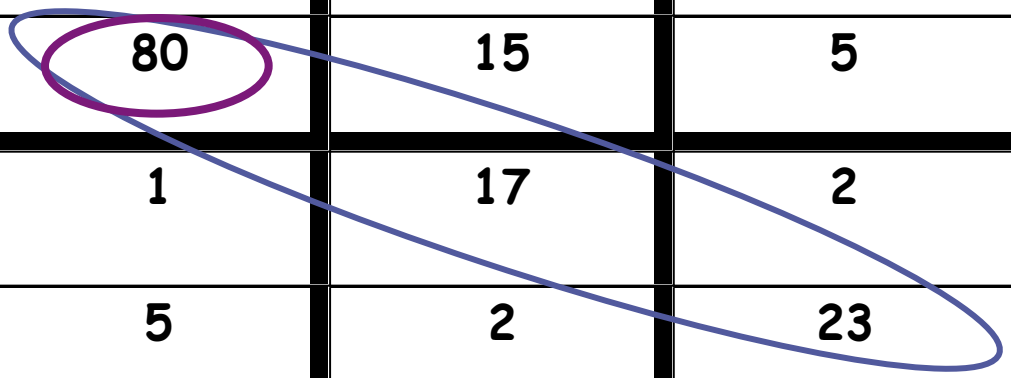
Critères d'évaluation

- Taux de bon apprentissage

  Parmi tous les exemples, quelle proportion est bien classée
- Précision de la classe k
  - Parmi les exemples classés dans la classe k, quelle proportion est effectivement de la classe k ?
- Rappel de la classe k
  - Parmi les exemples de la classe k, quelle proportion se retrouvent classés dans la classe k ?
- Précision contre Rappel
- Matrice de confusion : table de contingence

Evaluation des méthodes

Prédit ↓	OBSERVE			
	Payé	Retardé	Impayé	TOTAL
Payé	80	15	5	100
Retardé	1	17	2	20
Impayé	5	2	23	30
TOTAL	86	34	30	150



- **Validité du modèle (taux d'apprentissage) :** nombre de cas exacts (=somme de la diagonale) divisé par le nombre total : 120/150 = 0.8

- **Rappel** de la classe Payé : nombre de cas prédits et observés « payé » divisé par le nombre total de cas observés « payés » : 80/86 = 0.93
- **Précision** de la classe Payé : nombre de cas observés et prédits « payé » divisé par le nombre total de cas prédits « payés » : 80/100 = 0.8

Evaluation des méthodes

Traitement des données manquantes

- Attention à la sémantique : La donnée peut-elle exister ?
- Plusieurs méthodes :
  1. les oublier
  2. les remplacer :
     - valeurs majoritaire
     - valeur moyenne
  3. ...

II. Algorithmes de ce cours pour la classification

a. Classification supervisée :
   - Méthode de Bayes naïf
   - k plus proches voisins
   - Arbres de décision
   - Réseaux de neurones
b. Classification non supervisée : k-means
c. Évaluation des méthodes
d. **Règles d'association et motifs séquentiels**

UNIVERSITÉ MONTPELLIER 2
SCIENCES ET TECHNIQUES

Règles d'association et motifs séquentiels

Règles d'association

- Panier de la ménagère
- Recherche d'associations
  - recherche de corrélations entre attributs (items)
  - caractéristiques : « panier de la ménagère »
  - de très grandes données
  - limitations : données binaires
    - le client a acheté de la bière ou non

Règles d'association et motifs séquentiels

Motifs séquentiels

- Panier de la ménagère toujours ...
- Recherche de motifs séquentiels
  - recherche de corrélations entre attributs (items) mais en prenant en compte le temps entre items => Comportement
  - beaucoup plus complexe que les règles d'associations
  - énormément d'applications : prise en compte du temps

II. Conclusions

a. Quelques produits

b. Zoom sur Weka

c. Le Data Mining demain

UNIVERSITÉ MONTPELLIER 2
SCIENCES ET TECHNIQUES

Conclusions

- il existe de nombreuses (autres) méthodes
- il n'y a pas de meilleure méthode
- méthode à choisir selon :
  - les données (continues ? manquantes ? volumineuses ? denses ? ...)
  - la tâche
  - le temps de calcul dont on dispose
- règle du rasoir d'Ockham :
  - « pluralitas non est ponenda sine neccessitate »
  - « Les choses essentielles ne doivent pas être multipliées sans nécessité »

Quelques produits

- SAS Entreprise Miner de SAS
  – Statistiques, groupage, arbres de décision, réseaux de neurones, associations, ...
- SPSS Modeller (ex. Clementine)
  – statistiques, classification, réseaux de neurones
- Intelligent Miner d'IBM
  – modélisation prédictive (stat.), groupage, segmentation, analyse d'associations, détection de déviation, analyse de texte libre
- KXEN : Utilise SVM pour le SRM (Structural Risk Minimization)
- Oracle 10g ODM & SQL Server DM

- **Logiciels libres**
- Weka ;
- RapidMiner (Univ. Dortmund) ;
- Orange ;
- SIPINA/Tanagra (Univ. Lyon 2)

Quelques produits

Tanagra

•logiciel gratuit développé à l'Université de Lumiere Lyon 2, laboratoire ERIC, par Ricco Rakotomalala

•destiné à l'enseignement et à la recherche, et téléchargeable à l'adresse : http://chirouble.univ-lyon2.fr/~ricco/cours/index.html

•implémente diverses méthodes de fouilles de données issues du domaine de la statistique exploratoire, de l'apprentissage automatique et des bases de données, …

Orange

•développé par Blaz Zupan, à la Faculty of Computer and Information Science, de l'Université de Ljubljana en Slovenie

•destiné à l'enseignement et à la recherche, et téléchargeable à l'adresse : http://www.ailab.si/orange

•implémente aussi diverses méthodes de fouilles de données issues du domaine de la statistique exploratoire, de l'apprentissage automatique et des bases de données, ...

Logiciels de fouille de données : Weka

Weka (Waikato Environment for Knowledge Analysis) est un ensemble de classes et d'algorithmes en Java développé à l'Université de Waikato en Nouvelle Zélande

•implémente les principaux algorithmes de la fouille, notamment :

- les arbres de décision
- les réseaux de neurones

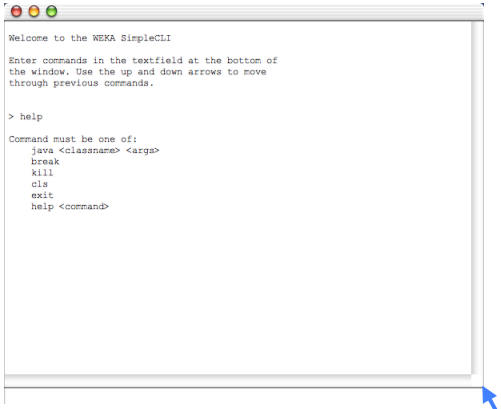
•téléchargeable (Unix et Windows) : http://www.cs.waikato.ac.nz/ml/weka

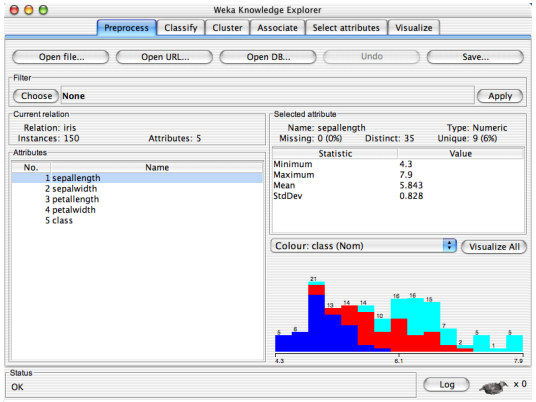
•développé en complément du livre : Data Mining par I. Witten et E. Frank (éditions Morgan Kaufmann).

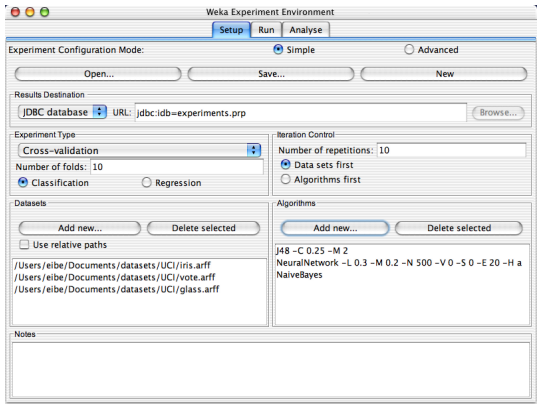
•peut être utilisé de plusieurs façons :

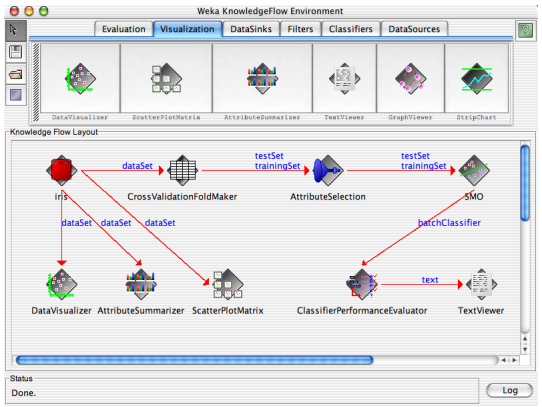
- par l'intermédiaire d'une interface utilisateur (comme utilisée en TP)
- en ligne de commande.
- par l'utilisation des classes fournies pour intégration) des programmes Java (classes documentées)

Weka : interface graphique

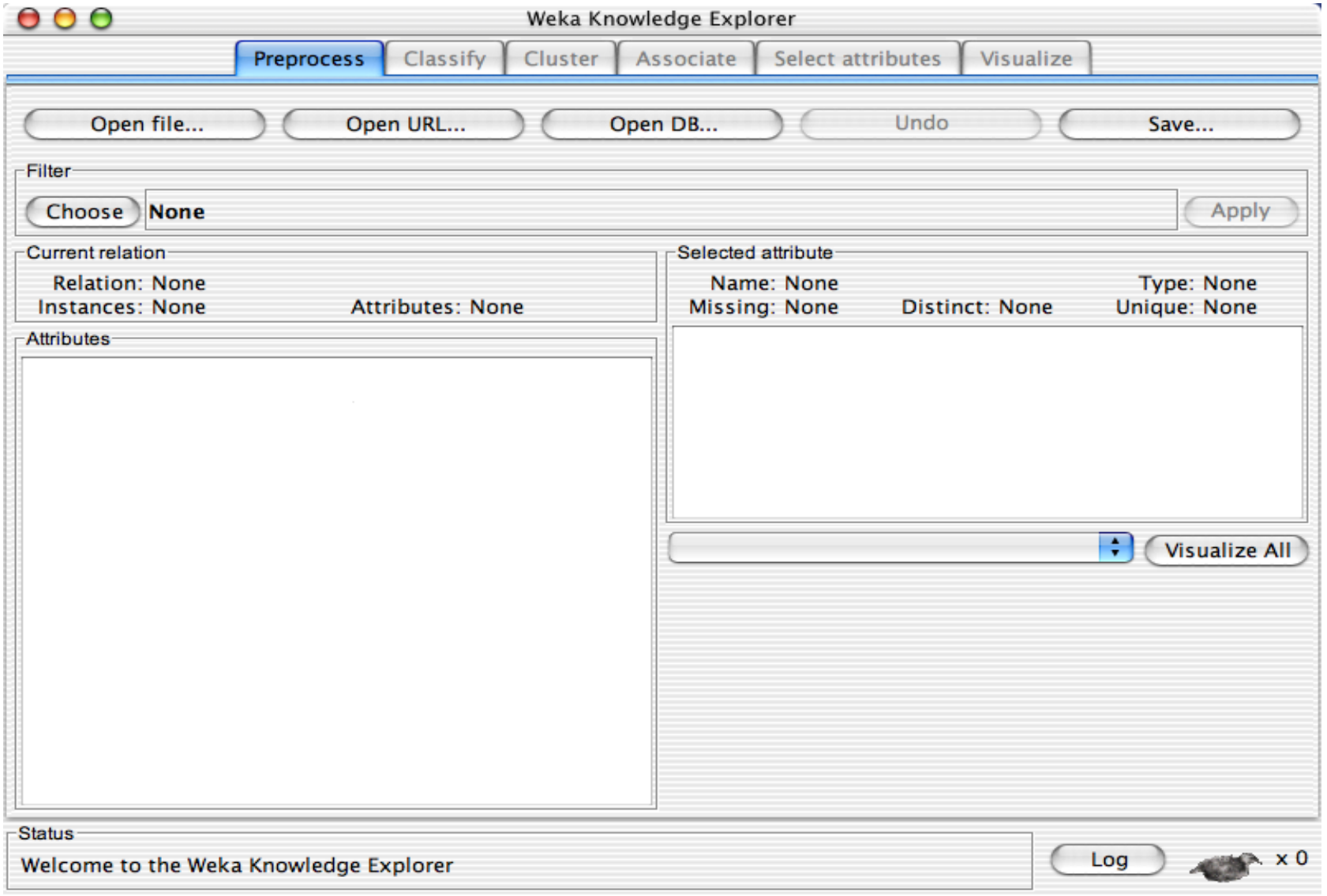




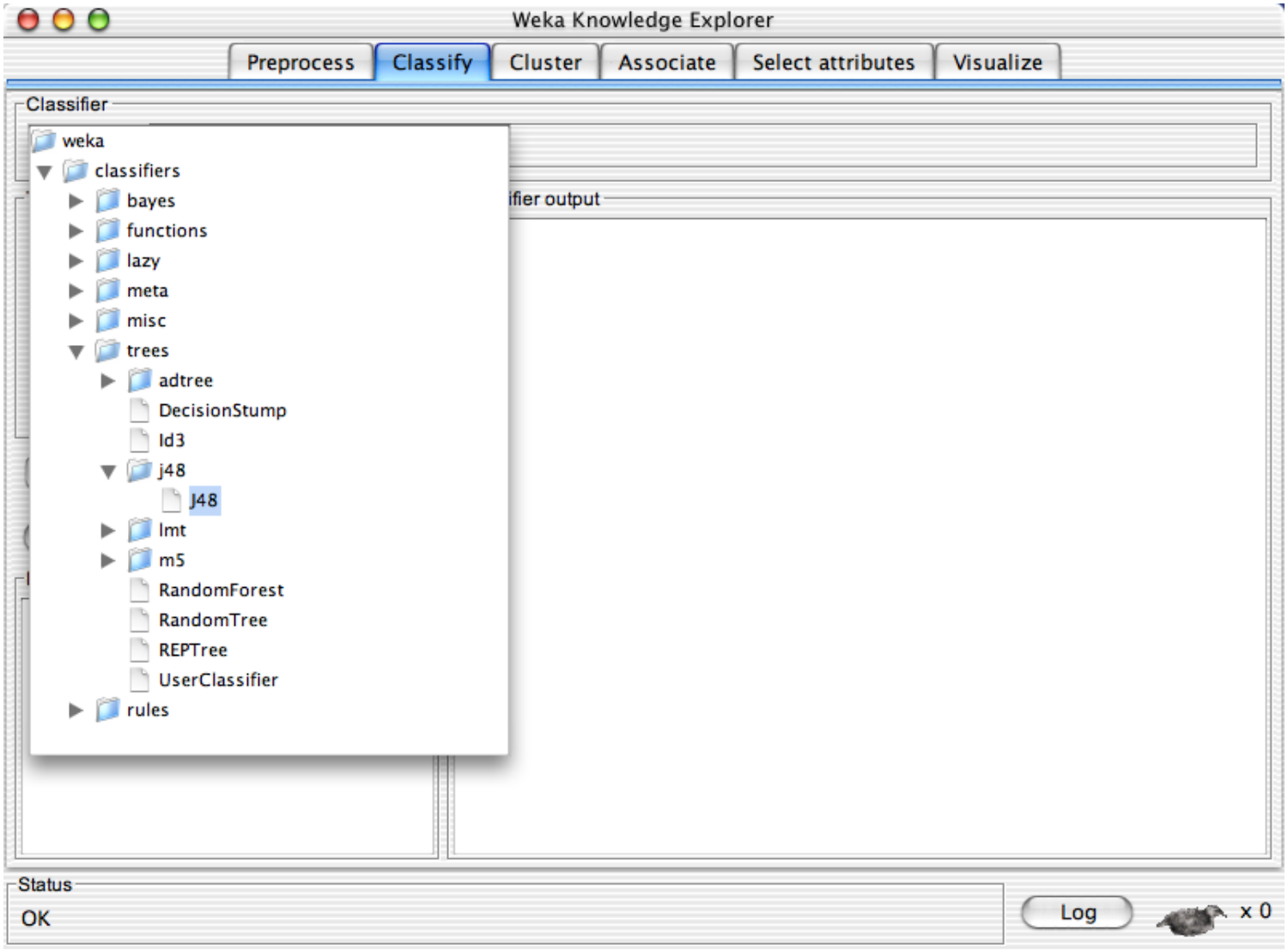




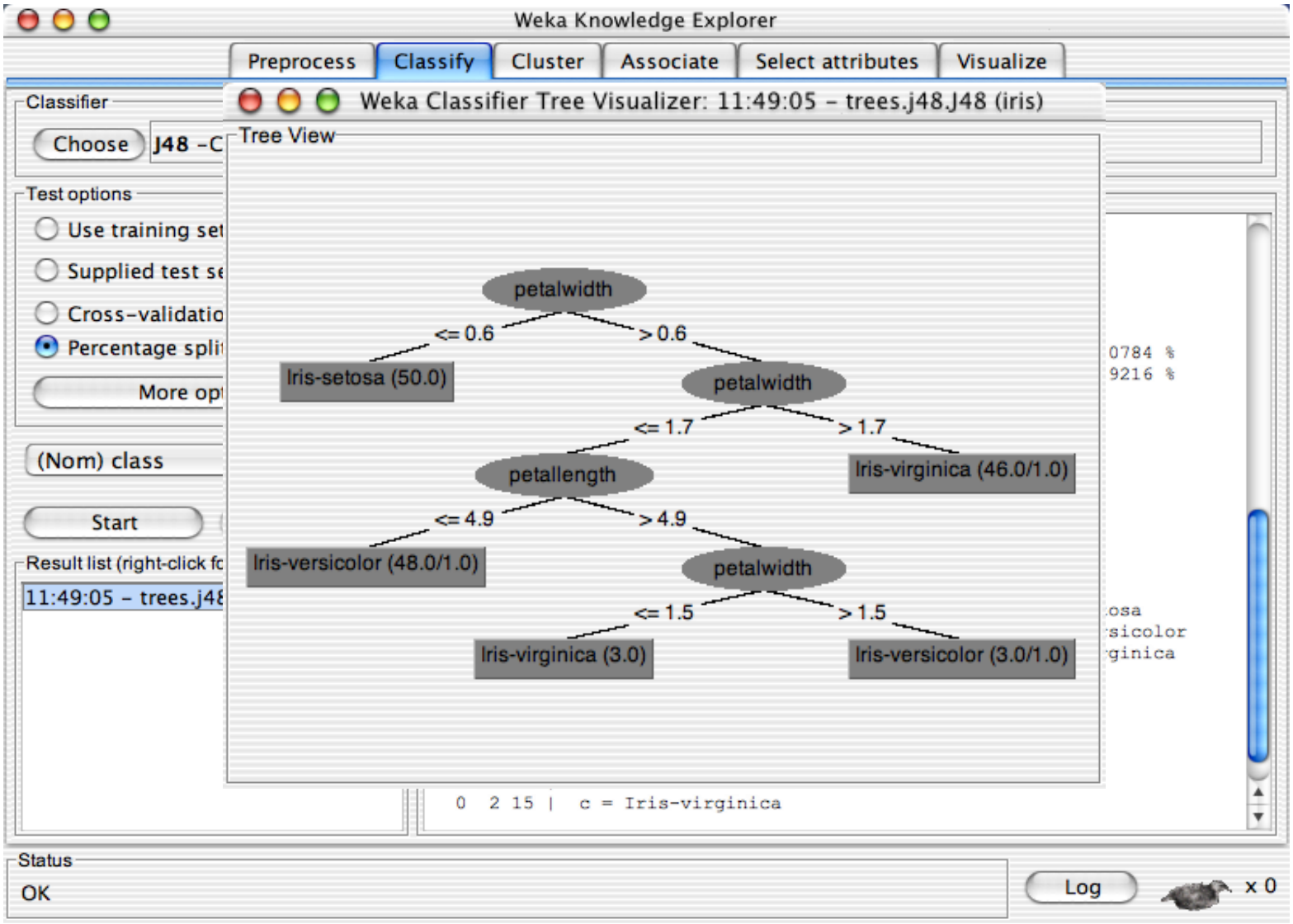
Weka : Explorer



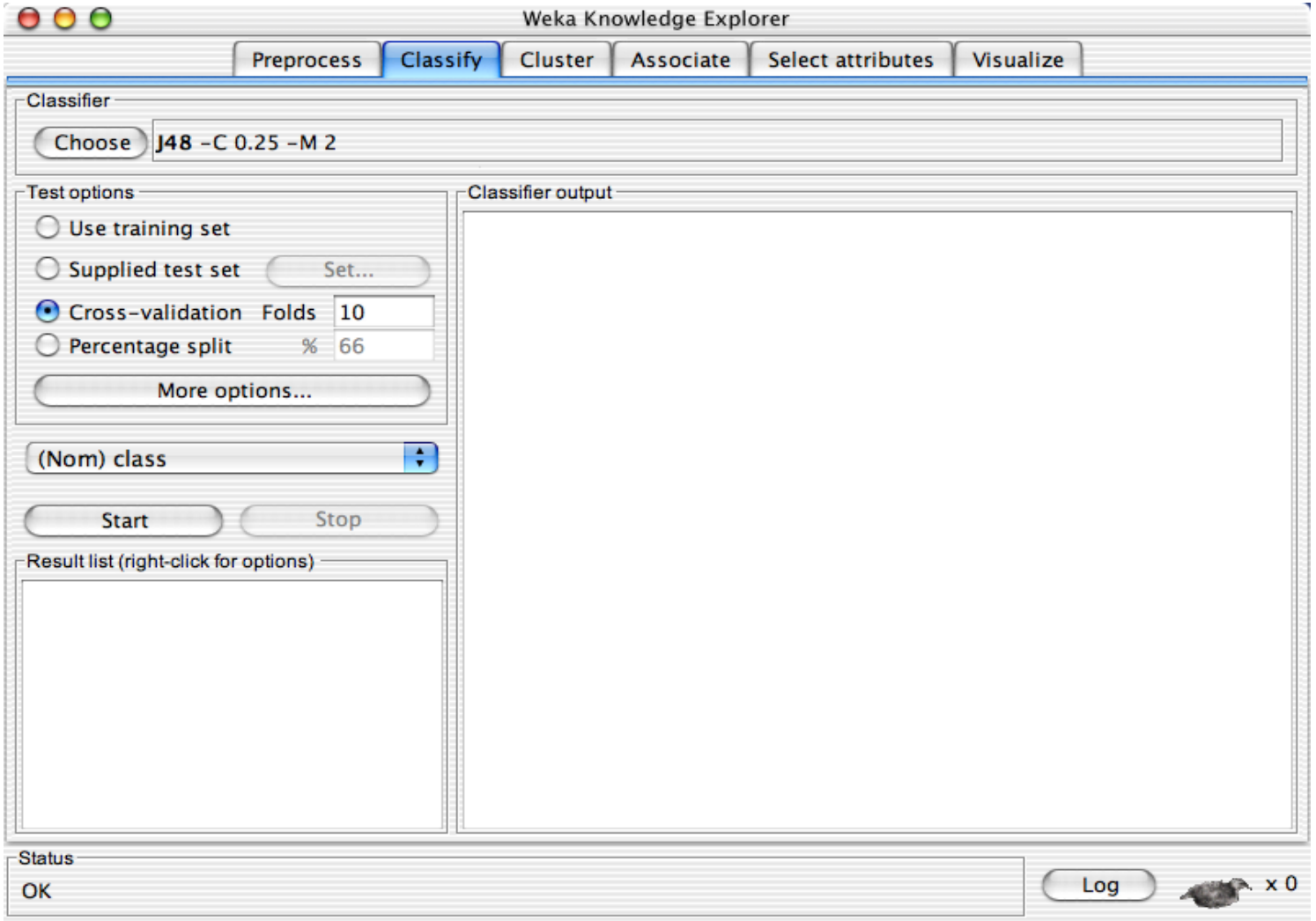
Weka : Classification



Weka : arbres de décision C4.5



Weka : Evaluation



Weka : matrice de confusion

```
Class attribute: play
Classes to Clusters:

 0 1  <-- assigned to cluster
 5 4 | yes
 3 2 | no

Cluster 0 <-- yes
Cluster 1 <-- no

Incorrectly clustered instances : 7.0  50%
```

Le Data Mining demain

- Agrégation de modèles
  - rééchantillonnage bootstrap, bagging, boosting...
- Web mining
  - optimisation des sites
  - meilleure connaissance des internautes
  - croisement avec les bases de données de l'entreprise
- Text mining
  - statistique lexicale pour l'analyse des courriers, courriels, dépêches, comptes-rendus, brevets (langue naturelle)
- Image mining
  - reconnaissance automatique d'une forme ou d'un visage
  - détection d'une échographie anormale, d'une tumeur

Conclusion

« Il est toujours dangereux de faire des prévisions, surtout quand c'est dans l'avenir. »

(Pierre DAC)