

**Mastère Professionnel : Business Intelligence**

# **Apprentissage automatique**

## **Généralités**

**Hind Elouedi**

**Semestre 2 - 2018**



- Définir la notion d'apprentissage automatique (Machine Learning).
- Enumérer les différents domaines d'application.
- Connaître les différents types d'apprentissage automatique.
- Comprendre la notion de classification.
- Comprendre la notion d'évaluation

# Points abordés



- Introduction
- Définitions
- Domaines d'application
- Types d'apprentissage
- Classification
- Evaluation



- Sources de données :
  - Données financières : Bourse, Banque, etc.
  - Données scientifiques : données géologiques, biologiques, images satellite, etc.
  - Business transactions : code à barre, e-commerce, etc.
  - Données personnelles / statistiques : recensement, dossier médical, profil client, données démographiques, etc.
  - World Wide Web et répertoires Online : BD Online, emails, news, images, vidéos, Web documents, librairies digitales, user registrations, etc.

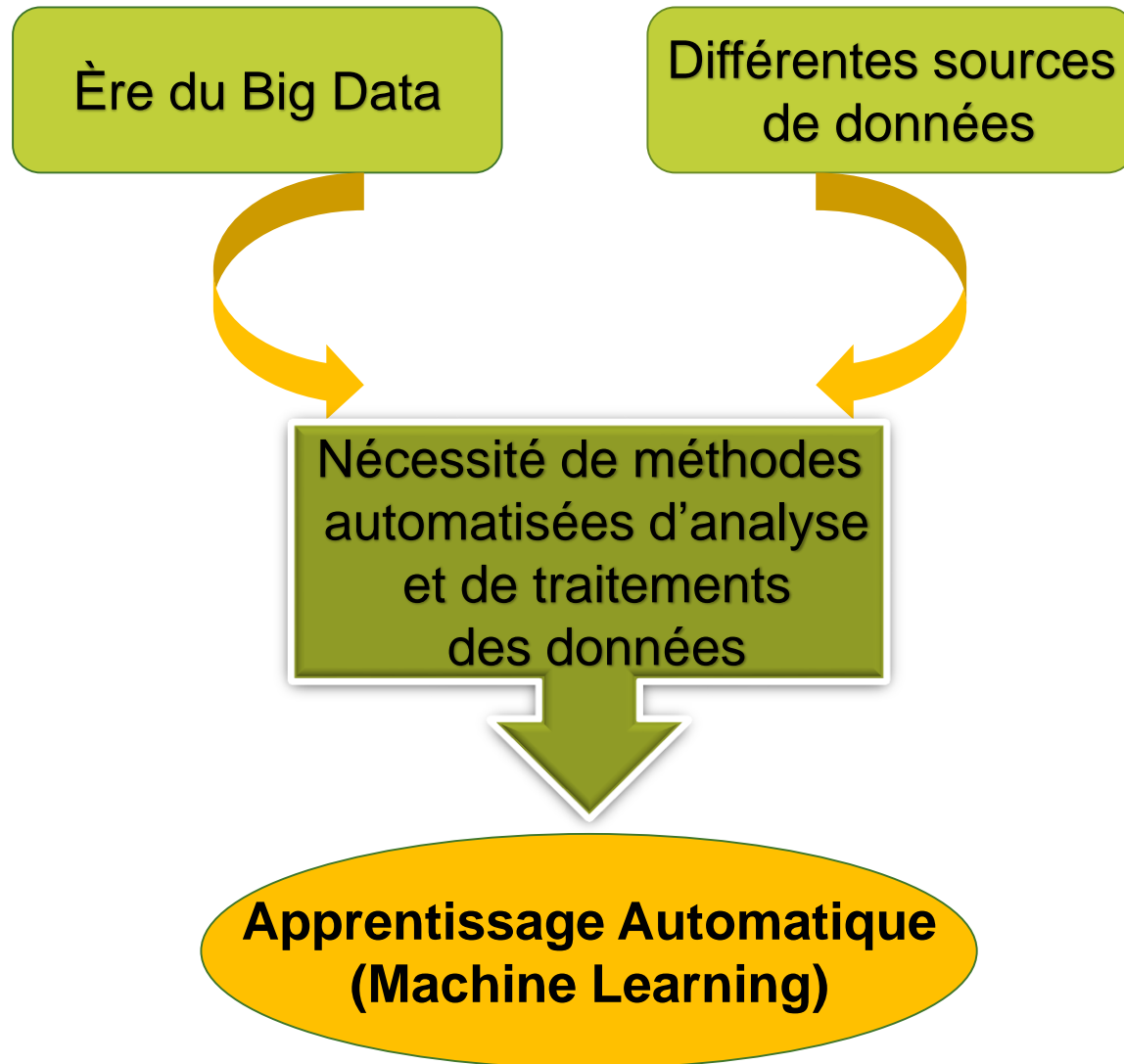
# Introduction (2)



## Données coûteuses en stockage et inexplorées



# Introduction (3)



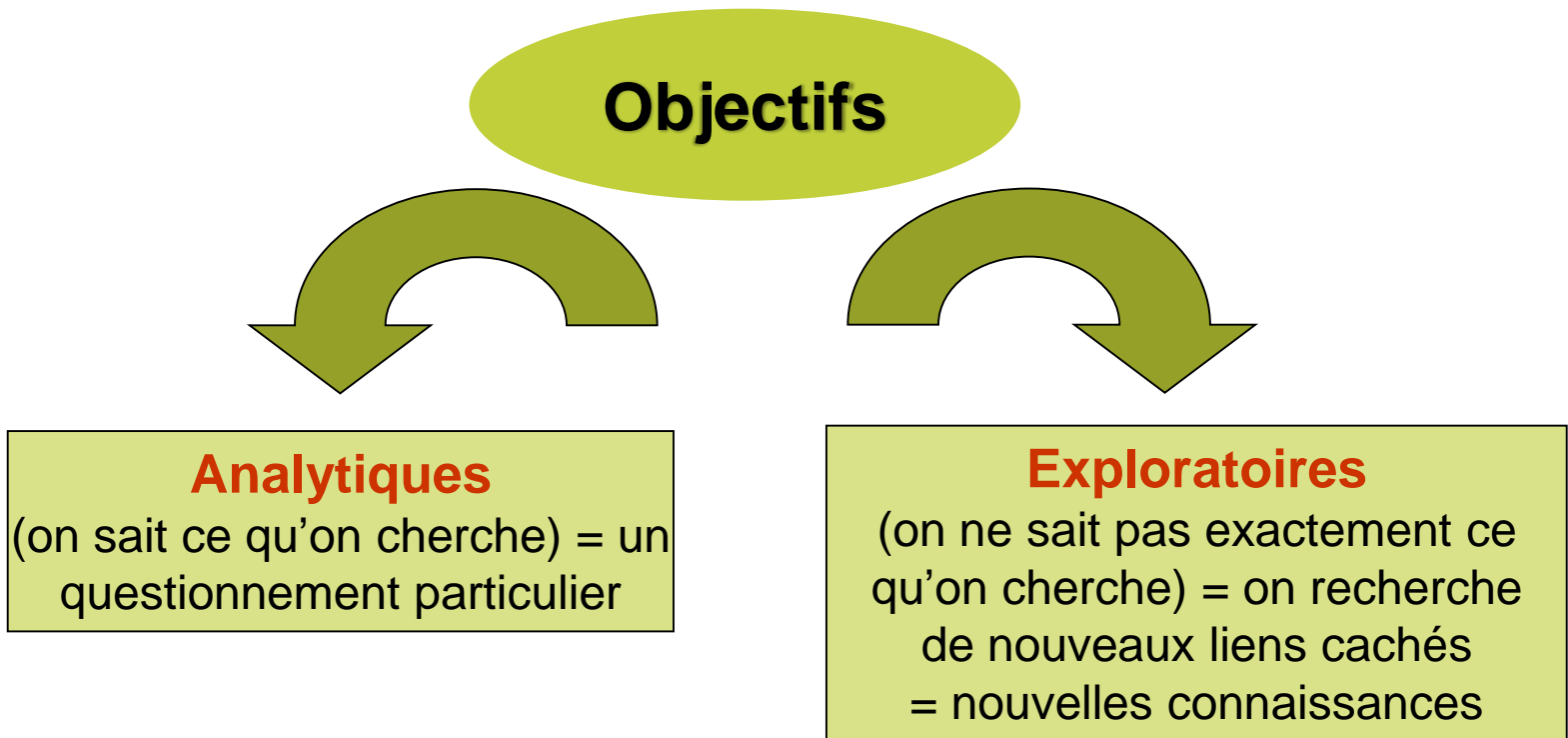


- Informatique décisionnelle
- Apprentissage
- Apprentissage automatique

# Définitions (2)



- Objectifs : Utilisation efficace des données (souvent hétérogènes) en un temps raisonnable pour des prises de décision compétitives.







## Informatique décisionnelle

**= Business intelligence**

Exploitation des données dans le but de faciliter la prise de décision par les décideurs, c'est-à-dire la compréhension du fonctionnement actuel et l'anticipation des actions pour un pilotage éclairé.



## Apprentissage

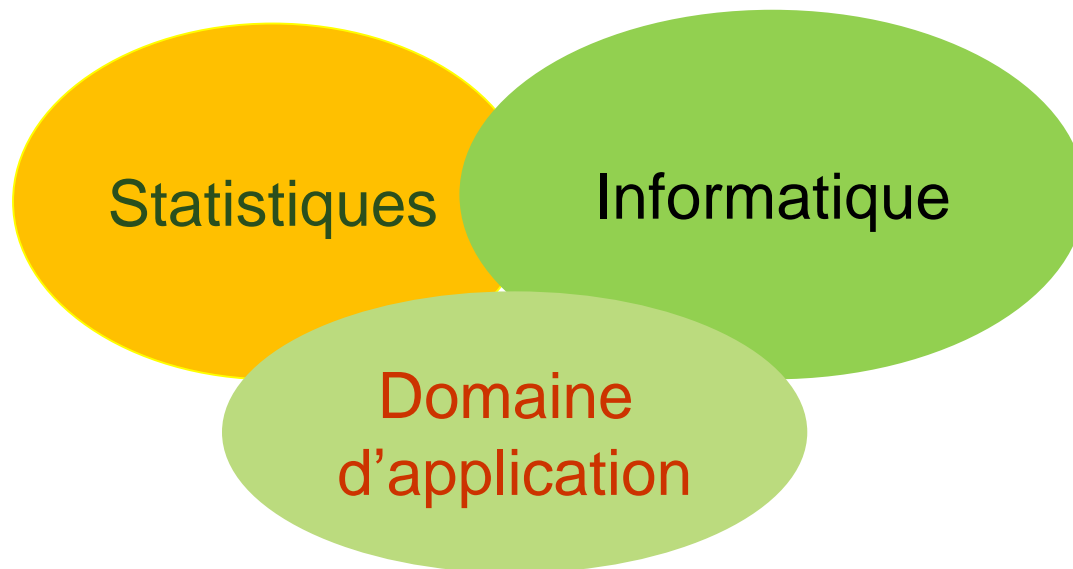
- Acquérir de nouvelles connaissances.
- Contracter de nouvelles habitudes.
- Avoir une connaissance extraite à partir d'un ensemble d'exemples ou d'expériences antérieures.

C'est la capacité d'améliorer l'accomplissement d'une tâche en interagissant avec un environnement.



## Apprentissage automatique

- Simuler la cognition humaine.
- Doter la machine d'un mécanisme d'apprentissage.
- **Machine learning** = Intersection de l'informatique, statistiques et domaines particuliers.



# Domaines d'application



- Data mining : Fouille de données
  - Exploitation des données historiques pour améliorer les décisions.
  - Ensemble de techniques d'exploration de données afin d'en tirer des connaissances (la signification profonde).
- Domaine des banques: Attribution de crédit
  - Utiliser un historique de crédits accordés et non accordés avec la situation personnelle du client.

# Domaines d'application (2)



- Domaine de la médecine: Aide au diagnostic
  - Caractériser les symptômes des anciens patients et de leurs maladies.
- Marketing: Élaboration d'un profil client
  - Faire une segmentation automatique des clients.
- Analyse financière: Prévission d'évolution des marchés.
- Assurance: Analyse des risques.
- Télécoms: Détection des fraudes.
- Sécurité: Détection des intrusions.

# Types d'apprentissage





- Apprentissage supervisé
- Apprentissage non supervisé
- Apprentissage semi-supervisé
- Apprentissage par renforcement

# Types d'apprentissage (2)



## Apprentissage supervisé

- C'est une technique d'apprentissage automatique où l'on cherche à produire automatiquement des règles à partir d'une base de données d'apprentissage contenant des « exemples ».
- On dispose d'un ensemble de paires d'E/S de la forme:  $(x_i, y_i)$ 
  - $x_i$  : entrée(s) possible(s)  Descriptions ou situations
  - $y_i$  : sortie(s) associée(s) à  $x_i$   Actions ou prédictions
- Les paires d'E/S sont appelées les **exemples** qui proviennent d'une **fonction inconnue**.
- Il s'agit de trouver une bonne approximation d'une fonction **f** dont on ne connaît le résultat que pour un certain nombre d'exemples.

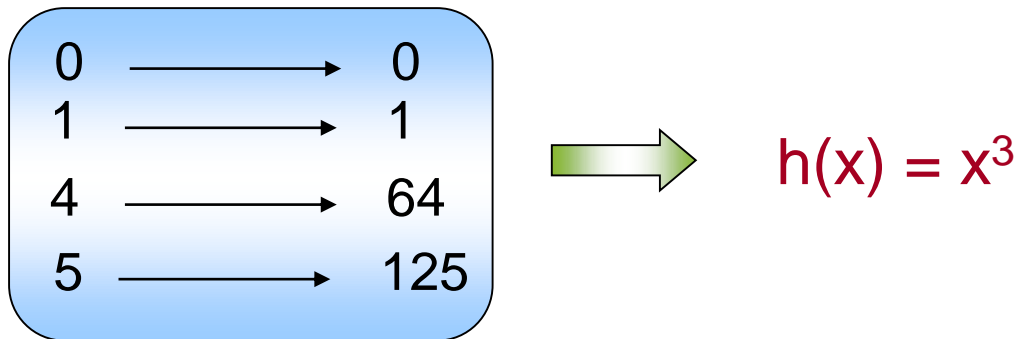
On demande au système de généraliser

# Apprentissage supervisé (2)



## Exemples

- Une fonction **h** aussi proche que possible de **f** où  $f(x_i) = y_i$



- Une distribution de probabilité  $P(x_i, y_i)$

*Quelle est la probabilité qu'un client achète un tel produit?*

- Dans un jeu de carte:
  - Les cartes gagnantes sont: 9♥, Roi ♥ et 7♦.
  - Les cartes perdantes sont: 3♠, 4♣ et 6♣.



Les cartes rouges sont gagnantes et les cartes noires numériques sont perdantes.



# Apprentissage supervisé (3)



- Apprentissage supervisé avec **variable réponse continue**  
➡ Régression, Estimation de densité
- Apprentissage avec **variable réponse discrète**  
➡ Classification ou analyse discriminante
- Apprentissage avec **variable réponse booléenne**  
➡ Apprentissage de concept

# Apprentissage non supervisé



- On dispose uniquement d'un ensemble d'entrées.
- Regrouper les entrées en un ensemble fixe de groupes: **Clustering**.
  - Les entrées de chaque groupe sont proches les uns des autres.
  - On utilise une certaine métrique dans l'espace des entrées.
- Découvrir de nouvelles relations au niveau des données: Ex. **Réseaux bayésiens**.

# Apprentissage non supervisé (2)



## Exemples

- Segmentation du marché:

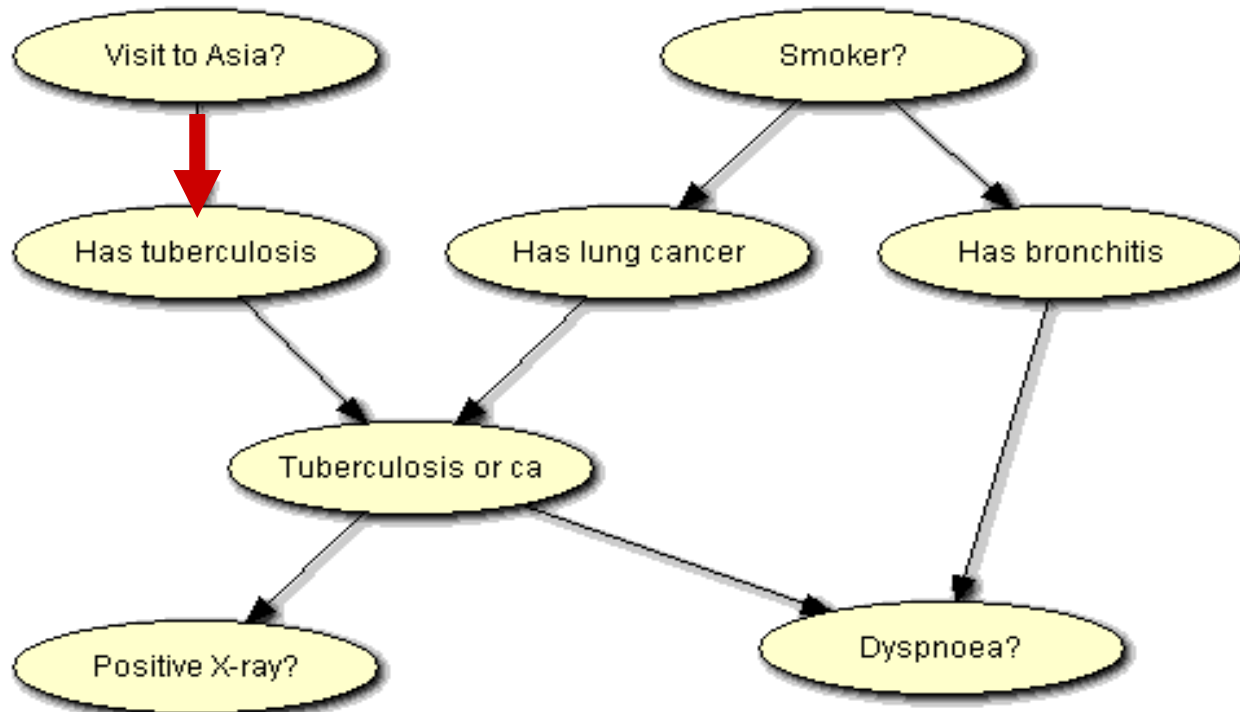
Quelles sont les catégories principales des clients typiques dans le domaine vestimentaire?

- Enfants, adolescent, adultes, etc.
- Habillé, sport, classique, etc.

# Apprentissage non supervisé (3)



- Domaine médical: Découverte de nouveaux liens
  - Il y a un lien entre visiter l'Asie et attraper la tuberculose



# Apprentissage semi-supervisé



- L'apprentissage semi-supervisé utilise un ensemble de données étiquetées et non-étiquetées.



Apprentissage semi-supervisé peut améliorer les performances en combinant les données avec labels et sans labels

# Apprentissage par renforcement



- L'algorithme d'apprentissage doit trouver une stratégie d'actions pour obtenir éventuellement une récompense (ou pénalité).



La récompense ou la pénalité arrive (généralement) suite à un ensemble d'actions.



Maximiser le gain (ou inversement) à long terme  
(apprentissage de réflexes, apprentissage de planification,...)

# Apprentissage par renforcement (2)



## Exemple

- Jeu d'échec :
  - On joue contre un adversaire.
  - Il y a une stratégie d'actions (en fonction du jeu).
  - C'est en fin de la partie qu'on va avoir le résultat de nos actions :
    - Victoire.
    - Nul.
    - Défaite.



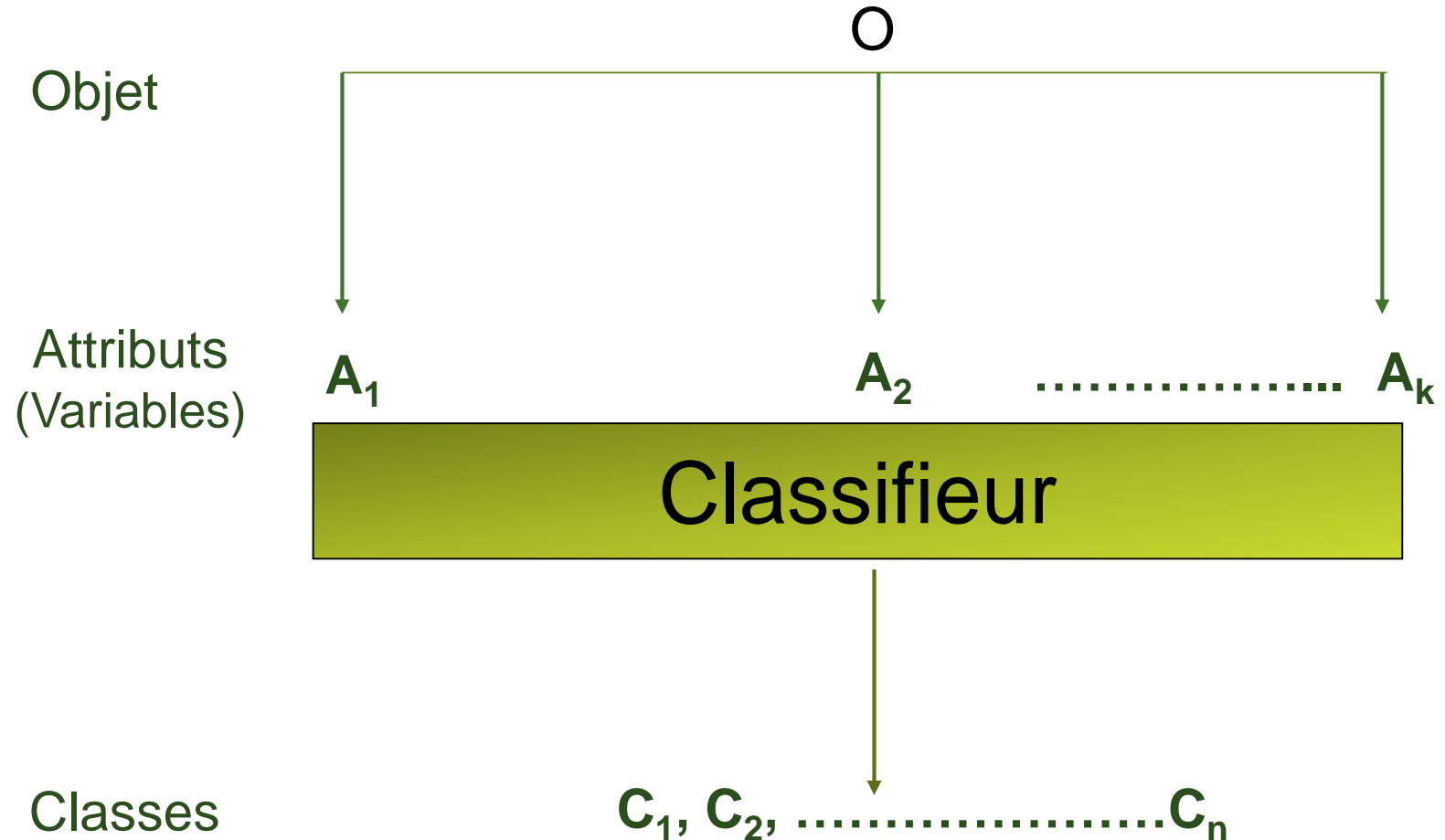
- Notion de classification
- Apprentissage par l'exemple
- Approche paramétrique
- Approche non paramétrique
- Types de classification



# Notion de classification



- L'une des tâches de l'apprentissage est la **Classification**



# Apprentissage par l'exemple



- On dispose d'un grand ensemble d'exemples (objets).
- On cherche à trouver une structure relative à ces exemples pour obtenir un modèle.
- Ce modèle permet de:
  - Extraire une procédure de **classification** à partir d'**exemples**.
  - **Classer** un nouvel exemple.
  - 
  - 
  - 
  - Prévoir une valeur numérique.
  - Comprendre la structure des exemples.

# Apprentissage par l'exemple (2)



- Ensemble d'apprentissage

Attributs

Revenu	Propriété	Crédit non remboursé	Classe
Elevé	Supérieur	Non	$C_1$
Elevé	Supérieur	Oui	$C_2$
Elevé	Supérieur	Non	$C_1$
Elevé	Inférieur	Oui	$C_2$
Moyen	Supérieur	Non	$C_1$
Moyen	Supérieur	Oui	$C_2$
Moyen	Inférieur	Non	$C_2$
Moyen	Inférieur	Oui	$C_2$
Faible	Inférieur	Non	$C_3$
Faible	Inférieur	Oui	$C_3$

Valeurs des attributs

$C_1$ : Attribuer tout le crédit -  $C_2$ : Attribuer une partie crédit -  $C_3$ : Ne pas attribuer le crédit.

# Apprentissage par l'exemple (3)



- Ensemble test

Revenu	Propriété	Crédit non remboursé	Classes
Elevé	Supérieur	Oui	?
Moyen	Inférieur	Non	?
Elevé	Supérieur	Oui	?
Moyen	Supérieur	Oui	?
Faible	Inférieur	Oui	?
Nul	Inférieur	Oui	?
Elevé	Supérieur	Non	?
Moyen	Inférieur	Oui	?



On cache les vraies classes

# Approche paramétrique



- Proposition d'un modèle dont on estime ses paramètres à partir des exemples (phase d'apprentissage).
- Les hypothèses que l'on fait sur les lois de probabilité font partie d'une famille de distributions.
  - Si on sait que **P** est une distribution normale, il suffit d'estimer ses deux paramètres:
    - Sa moyenne.
    - Son écart type.

➡ Avoir une bonne **approximation** de la distribution **P**.

➡ Déterminer une procédure de **classification**.

# Approche non paramétrique



- Pas d'hypothèses sur le modèle que suivent les données.
- Les problèmes à résoudre sont plus complexes que ceux traités par les méthodes paramétriques.
  - Méthodes statistiques.
  - Méthodes issues de l'intelligence artificielle.

# Types de classification



## ● Classification supervisée

- Les classes sont définies **a priori** (à l'avance).
- Découverte de règles ou formules pour ranger les données dans des classes prédéfinies.
  - Construction d'un modèle sur les données dont la classe est connue (ensemble d'apprentissage).
  - Utilisation des nouveaux objets pour classification.

## ● Exemples

- Arbres de décision.
- Méthode K plus proches voisins.
- Réseaux de neurones.
- Machines à vecteurs supports (SVM).

# Types de classification (2)



## ● Classification non supervisée

- Les instances d'apprentissage ne sont pas fournies avec des classes.

➔ L'ensemble d'apprentissage n'est pas étiqueté (on ne connaît pas les classes a priori).

- Intuitivement les objets de même classe sont “proches” les uns des autres.

➔ Mesure de similarité ou de distance

Regrouper les exemples similaires:  
Segmentation et cluster

## ● Exemples

- Centres mobiles.
- Réseaux de Kohonen.





- Utilisation d'un ensemble test.
- Pourcentage de Classification Correcte (PCC).
- Taux d'erreur de la classification (déduit du PCC).
- Utilisation de la validation croisée.

# Evaluation (2)



$$\text{PCC} = \frac{\text{Nombre d'objets correctement classés}}{\text{Nombre total des objets tests}}$$

## Ensemble test

Revenu	Propriété	Crédit non remboursé	Classes prédites	Vraies classes
Elevé	Supérieur	Oui	$C_1$	$C_1$
Moyen	Inférieur	Non	$C_2$	$C_2$
Elevé	Supérieur	Oui	$C_1$	$C_1$
Moyen	Supérieur	Oui	$C_3$	$C_2$
Faible	Inférieur	Oui	$C_1$	$C_3$
Nul	Inférieur	Oui	$C_3$	$C_3$
Elevé	Supérieur	Non	$C_1$	$C_1$
Moyen	Inférieur	Oui	$C_2$	$C_2$

$$\text{PCC} = \frac{6}{8} = 75\%$$

Taux d'erreur = 25%

# Evaluation (3)



- Matrice de confusion

Classifieur			
Prédites \ Vraies	$C_1 (4)$	$C_2 (2)$	$C_3 (2)$
$C_1 (3)$	3	0	0
$C_2 (3)$	0	2	1
$C_3 (2)$	1	0	1

- Bon classifieur sur la diagonale.
- Identifier les classes mal comprises (appprises).



- **Validation croisée**
  - Partition de l'ensemble d'apprentissage  $T$  en  $n$  ensembles disjoints  $(T_1, T_2, \dots, T_n)$  de même taille  $|T_i|$
  - Pour chaque  $i = 1, 2, \dots, n$ 
    - On fait l'apprentissage sur  $T - \{T_i\}$
    - On teste sur  $T_i$
    - On calcule le PCC sur  $T_i$
  - On fait la moyenne des PCC.

À suivre...



# Arbres de décision

