

Cours 6

Fouille de données (datamining)

BDA Master GL/RT- 2019- 2020

OUARED Abdelkader

Plan

- Motivation
- Données vs connaissances
- Datamining ou fouilles de données
- Apprentissage
- Principales étapes de fouilles de données(KDD)
- Tâches de datamining
- Principaux outils

- Les données forment le cœur des processus de base dans la plupart des entreprises.
- L'archivage des données crée la mémoire de l'entreprise.
- L'exploitation des données crée l'intelligence de l'entreprise (on parle d'intelligence « économique » ou d'« affaires »).

Motivations

- Volume de données trop grands, explosion des données
- Comment explorer des millions d'enregistrements avec des milliers d'attributs ?
- Faible pourcentage de données analysées
- Besoins de répondre rapidement aux opportunités
- Besoin de traitement en temps réel de ces données
- Requêtes traditionnelles (SQL) impossibles
- ...

Real World Example

Major telecommunications company
30 days of *call data records* (1TB)

Motivations

- Améliorer la productivité
 - Besoin de prendre des décisions stratégiques efficaces
 - Exploiter les données historiques pour prédire le futur et anticiper le marché

Exemples d'applications

Secteur de distribution (1)

- **Nature du problème** : Découvrir les règles d'association d'achat = Cerner les habitudes des clients
- **Objectif** :
 - Optimiser l'organisation des rayons.
 - Améliorer la planification des offres spéciales et des promotions.
 - Anticiper les attentes des consommateurs de la chaîne.



Secteur de distribution (2)

- **Paramètres pris en compte** : Tickets de caisse, nature des offres spéciales du jour, données météo, période de l'année.
- **Source des données** : tickets de caisse, enquêtes consommateurs.
- **Exemple** : Dans les supermarchés américains, il a été possible de mettre en évidence des **corrélations** entre achat d'un boisson et achat de couches bébé avant le week-end ! remarque justifiée par le comportement des jeunes pères américains qui préparent leur week-end en préparant leur provision de bière pour regarder la télévision et qui font les achats pour bébé au même moment.

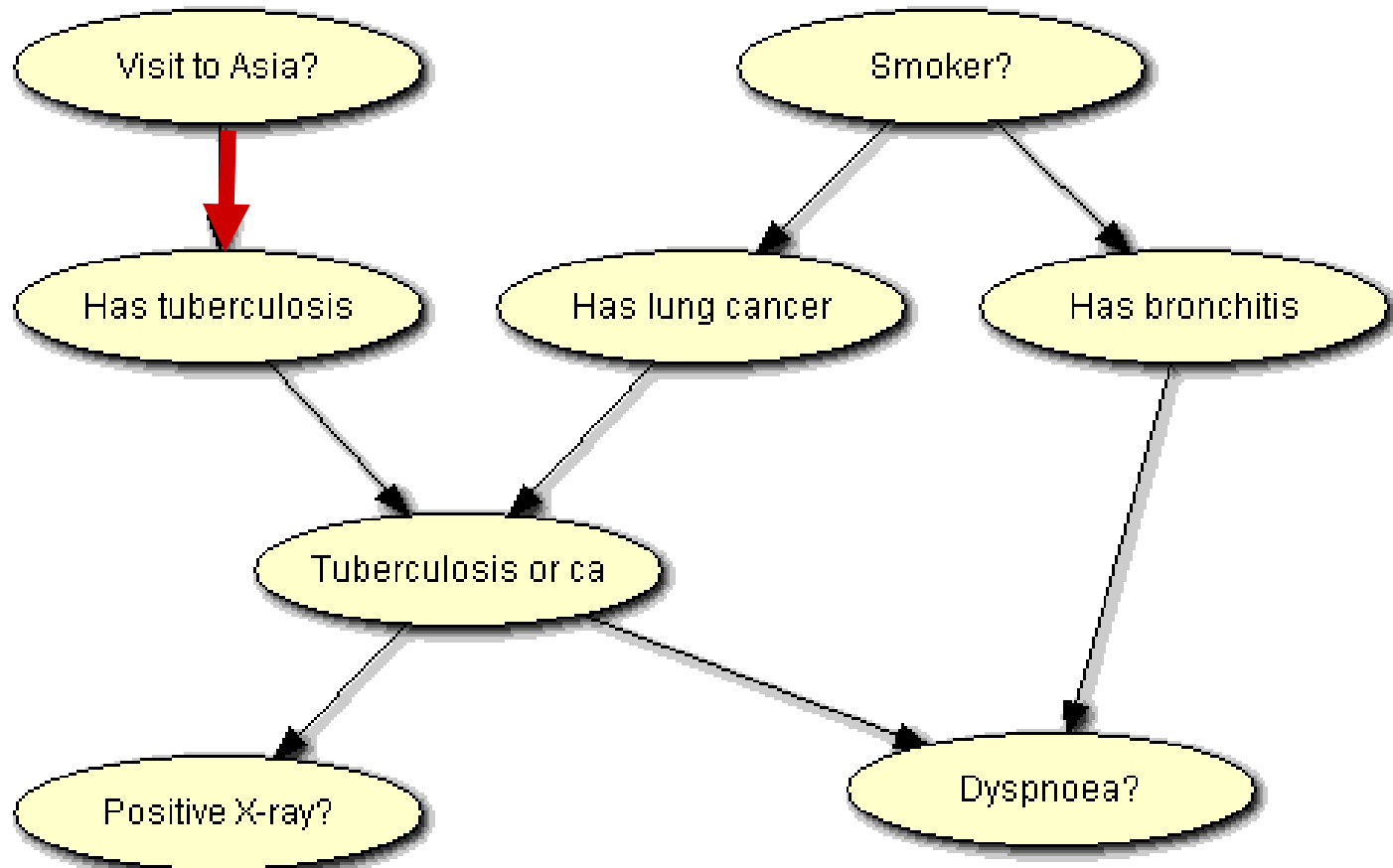


+



Domaine médicale (2)

- Par la technique réseaux bayésiens on obtient le réseau suivant:
- Si on demande aux experts de donner les liens par leur expériences il y a certains liens qu'ils ne détectent pas (exemple lien Asie – Tuberculose)



Secteur bancaire

- **Nature du problème** : Définition des comportements des titulaires de compte dans une banque
- **Objectif** : Détecter les clients fidèles .
- **Paramètres pris en compte** :
 - **Caractéristiques personnelles du client** (âge, sexe, situation familiale, salaire, divers indicateurs liés aux comptes bancaires du client, à l'historique des événements, aux produits déjà souscrits.)
 - **Type d'habitat** : Zone rurale, urbaine, sous-urbaine.
 - **Saisonnalité** : fêtes particulières, congés, événements liés à l'actualité..
- **Source des données** : BD internes historisées (calendrier de toutes les opérations réalisées), coupons-réponses, mailings.

Détection d'intrusions

(objet de votre projet Fouille de données)

- **Nature du problème :** Caractériser les connexions sur un réseau informatique
- **Objectif**
 - Détecter toute violation de la politique de sécurité en vigueur sur un système informatique
 - Distinguer les connexions normales des attaques
- **Paramètres pris en compte**
 - Paramètres définissant la connexion: exemple durée de la connexion, type du protocole (tcp, udp, icmp,..), service réseau (destination) (http, telnet,...), statut de la connexion, etc (41 attributs)
- **Source des données :**
 - Données Off line (exemple Darpa, KDD'99),
 - BD contenant une description des attaques connues, ...

Secteur télécom

Vous êtes gestionnaire marketing
d'un opérateur de
télécommunications mobiles :

- Les clients reçoivent un téléphone gratuit (valeur 150€) avec un contrat d'un an ; vous payer une commission de vente de 250€ par contrat
- Problème : Taux de renouvellement (à la fin du contrat) est de 25% !!!
- Donner un nouveau téléphone à toute personne ayant expiré son contrat coûte cher.
- Faire revenir un client après avoir quitté est difficile et coûteux.



Secteur télécom



Solution : Trois mois
avant l'expiration du
contrat, prédire les
clients qui vont quitter :

**« Si vous voulez les
garder, offrir un
nouveau téléphone ».**

Points communs

Données importantes, hétérogènes, coûteuses en stockage et **inexploitées**



Datamining

Données $\xrightarrow{?}$ connaissances

Définition de datamining (1)

Le data mining est défini comme un ensemble d'outils et de techniques pour l'aide à la décision où les utilisateurs cherchent des modèles d'interprétation dans les données.

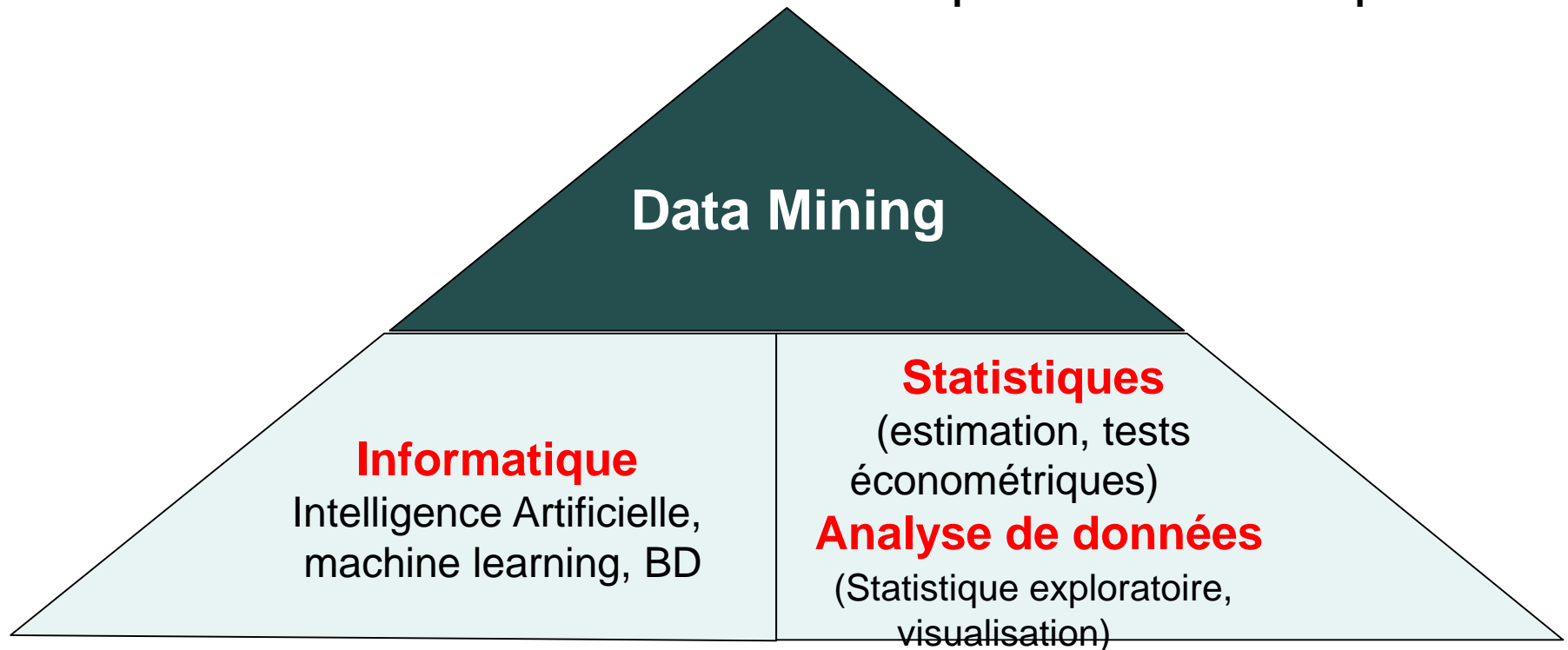
Définition de datamining (2)

Le data mining est l'ensemble des :

- techniques et méthodes
- ... destinés à l'exploration et l'analyse
- ... de (souvent) grandes bases de données informatiques
- ... en vue de détecter dans ces données des règles, des associations, des tendances inconnues (non fixées a priori), des structures particulières restituant de façon concise l'essentiel de l'information utile
- ... pour l'aide à la décision

Principaux outils

- Les principaux outils utilisés en data mining sont dérivés de méthodes informatiques ou statistiques



What is the difference between data science and data mining?

Données vs connaissances

Données - Connaissances

Décision

- Promouvoir le produit P dans la région R durant la période N
- Réaliser un mailing sur le produit P aux familles de profil F

Connaissance

- Une quantité Q du produit P est vendue en région R
- Les familles de profil F utilisent M% de P durant la période N

Information (ou faits personnalisés)

- X habite la région R
- Y a A ans
- Z dépense son argent dans la ville V de la région R

Données (brutes)

- Consommateurs
- Magasins
- Ventes
- Démographie
- Géographie

Données vs Connaissance

- **Comment les transformer en connaissances**
 - **apprentissage?**

Apprentissage

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



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

Comment peut-elle apprendre ?

Comment :

- On apprend de plusieurs façons
- Pratique d'une technique ou d'une habileté
- Observation
- Réflexion sur les expériences
- Lecture et méditation
- Essais et erreurs
- Conditionnement
- Imitation
- Simulation réelle ou imagée
- Feed Feed--bac



Apprentissage chez l'enfant!

- L'enfant apprend à reconnaître l'odeur de sa mère, puis sa voix,... apprend à coordonner **ses perceptions** (**vue, toucher et mvmt**)... par des essais gratifiants ou pénalisants, apprend à marcher ... apprend à segmenter à catégoriser des sons, à leur associer des significations ... apprend la structure de sa langue maternelle et acquiert un répertoire organisés de connaissances sur le monde... puis il apprend à lire, puis à maîtriser des concepts de plus en plus abstraits enfin, ... et autres
- Puis il apprend l'informatique, la sécurité Informatique et la fouille de données, java, etc ...

Apprentissage automatique (Machine Learning)

- Apprentissage artificiel (par opposition à l'apprentissage naturel)
- Toute méthode permettant de construire un modèle de la réalité à partir de données,
 - soit en améliorant un modèle partiel ou moins général,
 - soit en créant complètement le modèle.

Apprentissage supervisé (1)

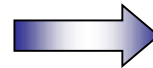
- On dispose d'un ensemble de paires d'entrée(s)/sortie(s) de la forme (x_i, y_i) ,
 - x_i : entrée(s) possible(s)  Descriptions ou situations
 - y_i : sortie(s) associée(s) à x_i  Actions ou prédictions
- Les paires d'entrée(s)/sortie(s) sont appelées **les exemples** qui proviennent d'**une fonction inconnue**.
- Il s'agit de trouver une bonne approximation d'une fonction f dont on connaît le résultat que pour un certain nombre d'exemples.

On demande au système de généraliser

Exemples

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

0	→	0
1	→	1
4	→	64
5	→	125



$$h(x) = x^3$$

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

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

- Dans un jeu de cartes:

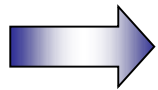
- les cartes gagnantes sont: 9♥, Roi ♥ et 7♦.
- les cartes perdantes sont: 3♠, 4♣ et 6♣.



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

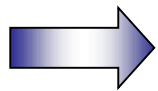
Quelle méthode pour un apprentissage supervisé? (2)

- Apprentissage supervisé avec **variable réponse continue.**



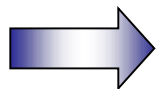
Régression, Estimation de densité

- Apprentissage supervisé avec **variable réponse discrète.**




Classification ou Analyse discriminante

- Apprentissage supervisé avec **variable réponse booléenne.**



Apprentissage de concept

Apprentissage non supervisé

- On ne dispose pas d'un ensemble de paires d'entrée(s)/sortie(s).
 On dispose uniquement d'un ensemble d'entrées.
- Regrouper les entrées en un nombre fixe de groupes (**clusters**):
 - 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** dans les données (ex: Réseaux Bayésiens).

Les tâches de l'apprentissage:

Tâche de l'apprentissage

Classification

Prédiction

Caractéristique

Discrimination

Association

Clustering

..

Principales étapes de fouilles de données

Le processus KDD

Knowledge Discovery in Databases (KDD) est un processus non trivial d'identification de modèles valides, originaux, potentiellement utiles, et finalement compréhensibles à partir des données.

[FAYYAD, PLATETSKY-SHAPIRO, SMYTH, 96]

Les étapes du processus de KDD

- Étape 1 : Poser le problème
- Étape 2 : Rechercher les données
- Étape 3 : Sélectionner les données pertinentes
- Étape 4 : Nettoyer les données
- Étape 5 : Transformer les données
- Étape 6 : Rechercher les modèles
- Étape 7 : Évaluer et valider les résultats
- Étape 8 : Extraire la connaissance

Étape 1: Poser le problème

Un exemple : Détection d'intrusions réseaux

Connaissant :

- Les caractéristiques des connexions

Est-il possible de classer les connexions en cours (sujet du projet)

Étape 2 : Rechercher les données

- Identifier les informations
- Identifier les sources (différents fichiers .log)
- Vérifier leur qualité
- Vérifier leur facilité d'accès
 - Documents papier
 - Supports électroniques
 - Fichiers internes ou externes
 - Sources multiples, Data Warehouse ou Data Mart

Étape 3 : Sélectionner les données pertinentes

- **Feature engineering** is the process of using domain knowledge of the data to create features that make machine learning algorithms work

Étape 3 : Sélectionner les données pertinentes

- Sélectionner les attributs (**features**) les plus pertinents.
- Pas tous les contenus des « paquets » sont pertinents
- Réduire les dimensions
 - Expertise humaine
 - Analyses graphiques
 - Analyses de corrélation
 - Analyse en composantes principales

Étape 4 : Nettoyer les données

- Vérifier l'origine des données
- Traiter les valeurs aberrantes
- Traiter les valeurs manquantes
- Traiter les valeurs nulles

Étape 4 : Nettoyer les données

- Traiter les valeurs aberrantes/ manquantes
 - Min, Max, random
 - Ajouter nouvelle colonne
 - Données aberrantes/non aberrantes
 - Algo. Pour les données abérantes
 - Box Plots

Étape 5 : Transformer les données

- Mettre le résultat d'un « tcp-dump » sous une forme exploitable.
- Coder les informations qualitatives
 - Coder en ratios (pourcentages)
 - Normaliser les données
 - Transformer les dates en durées

Étape 5 : Transformer les données

- Transcoder les données,
- exemple : code postal en coordonnées géographiques
- Exprimer des fréquences
 - Exprimer des tendances
 - Réaliser des combinaisons de variables
 - ...

Étape 5 : Transformer les données


Types de variables

- Numériques (Poids, Taille, ...)
- Binaires
- Catégoriques (Couleur, Situation familiale, ...)
- Ordinales (Résultat d'un concours, Qualité d'un produit, ...)

Étape 5 : Transformer les données

Le Codage doit préserver les relations

- Codage Binaire: 0,1
 - **Sexe : M-F**
- Codage simple : 1,2,3 → L'ordre est important ??
 - **Taille** internationale, XXS, XS, **S**, **S/M**, **M**, **M/L**
 - **risk levels: High, Medium and Low.**
- **One hot encoding** is a binary representation of a categorical data



color		color_red	color_blue	color_green
red		1	0	0
green		0	0	1
blue		0	1	0
red		1	0	0

Étape 5 : Transformer les données (Feature Scaling)

Exemple: Normalisation des variables continues

$$x'_i = \frac{x_i - \min x_j}{\max x_j - \min x_j}$$

- x'_i sont entre 0 et 1
- $x'_i = 0$ si $x_i = \min x_j$
- $x'_i = 1$ si $x_i = \max x_j$

Normalisation des variables continues

$$x'_i = \frac{x_i - \min x_i}{\max x_i - \min x_i}$$

	Age	Nombre Enfants	Salaire
P1	40	2	1000
P2	75	4	600
P3	50	3	1100
P4	35	1	550



	Age	Nombre Enfants	Salaire
P1	0.125	0.333	0.818
P2	1	1	0.09
P3	0.375	0.666	1
P4	0	0	0

Étape 5 : Transformer les données

Quelques techniques

- La fonction **countplot()**: Tester l'équilibre des données (**balanced data set**)
(Oui/Non)
- La fonction **box plot** ou **boxplot** est une méthode pour représenter graphiquement des groupes de données numériques
 - **Ex.** Valeur > seuil ☹ (*log, racine, exp,...*)
- La fonction **distplot()** nous permet de réaliser les graphiques de **distribution**
 - Python Probability **Distributions** – **Normal, Binomial, Poisson, Bernoulli**
- Analyse contradictoire entre deux colonnes:
 - **Ex.** Age: 12, Nombre d'enfants: 2
- Nombre de valeurs distinctes sous forme histogramme, etc. Pour interpréter et conclure.
- Afficher les valeurs les plus fréquentes
 - Ex. Job: **111** catégories → **10** plus fréquents et autre
- Data augmentation → Techniques

Exercice

Variables binaires:

- Comment transformer les variables binaires !!!

$$O_i = (1, 0, 1, 1, 1)$$

$$O_j = (1, 0, 1, 0, 0)$$

Étape 6 : Recherche de modèles

- Choix d'une méthode ou d'une technique :
 - Réseaux de neurones
 - Règles d'association
 - Arbres de décision
 - Logique Floue
 - Algorithmes génétiques
 - Le raisonnement à base de cas
 - Les réseaux Bayésiens
 - ...

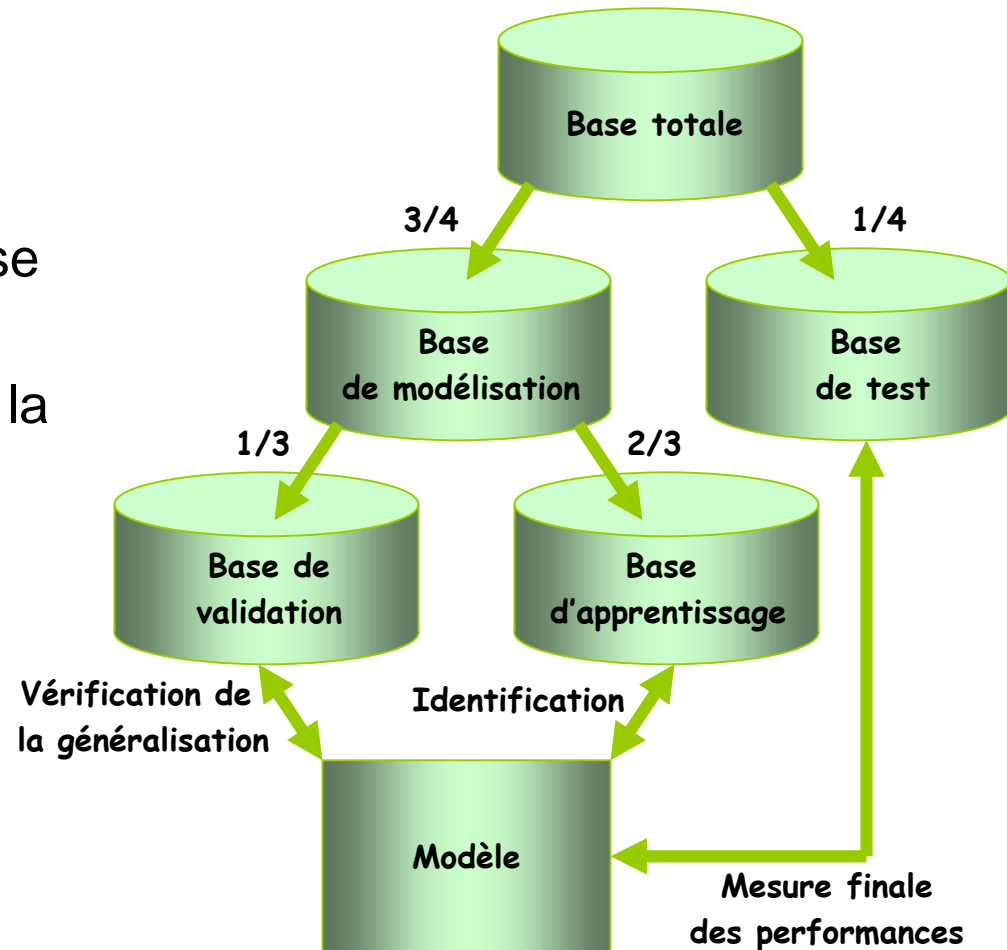
Étape 7 : Évaluer et valider les résultats

- Évaluation quantitative :
 - PCC : Pourcentage de classification correcte .
 - Erreur à base de distance
 - Coefficient de corrélation

Étape 7 : Évaluer et valider les résultats

- Évaluation par le test

- Base de test 25% de la base totale
- Base de validation 25% de la base totale
- Base d'apprentissage 50% de la base totale



Exemples

Rechercher les règles associatives

Algorithmes d'extraction des items fréquents

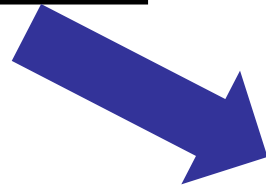
- **APRIORI**
- **Close**
- **OCD**
- **Partition**
- **DIC**

Rechercher les règles associatives

Algorithme APRIORI

- Génération d'ensembles d'items
- Calcul des fréquences des ensembles d'items
- On garde les ensembles d'items avec un support minimum: les ensembles d'items

fréquents

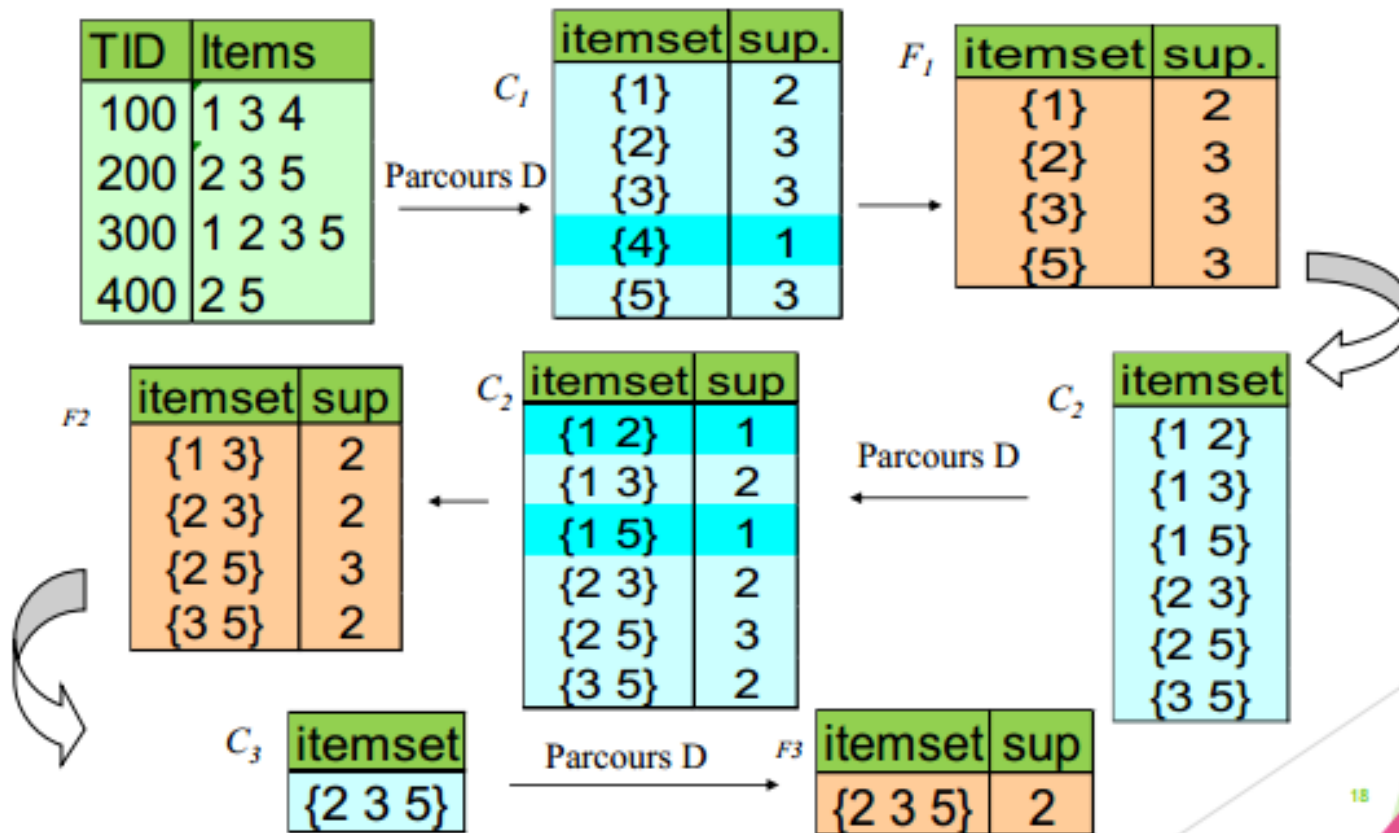


La force d'une règle d'association

Rechercher les règles associatives

Algorithme APRIORI

Exemple avec le support minimum = 2



Classification supervisée

a méthode des k plus proches voisins est une méthode de **d'apprentissage supervisé**.

- Soit $D = \{(x', c), c \in C\}$ l'ensemble d'apprentissage
- Soit x l'exemple dont on souhaite déterminer la classe

Algorithme

Début

pour chaque $(x', c) \in D$ **faire**

Calculer la distance $dist(x, x')$

fin

pour chaque $\{x' \in kppv(x)\}$ **faire**

compter le nombre d'occurrence de chaque classe

fin

Attribuer à x la classe la plus fréquente;

fin

■ Mesures de distance

souvent utilisées:
la distance Euclidienne

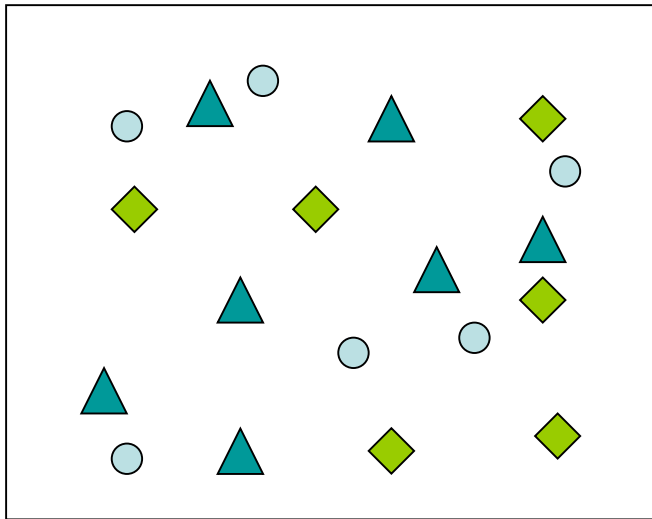
$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

Clustering

Classification non supervisée: Classes non prédéfinies a priori

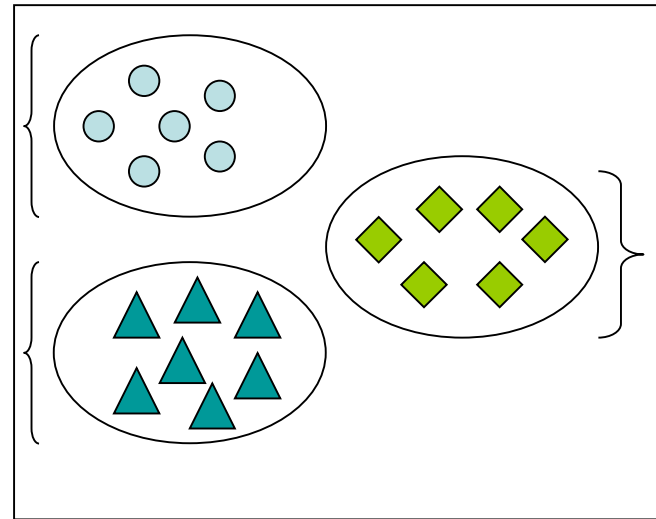


Regroupement des objets en des clusters formant les classes.



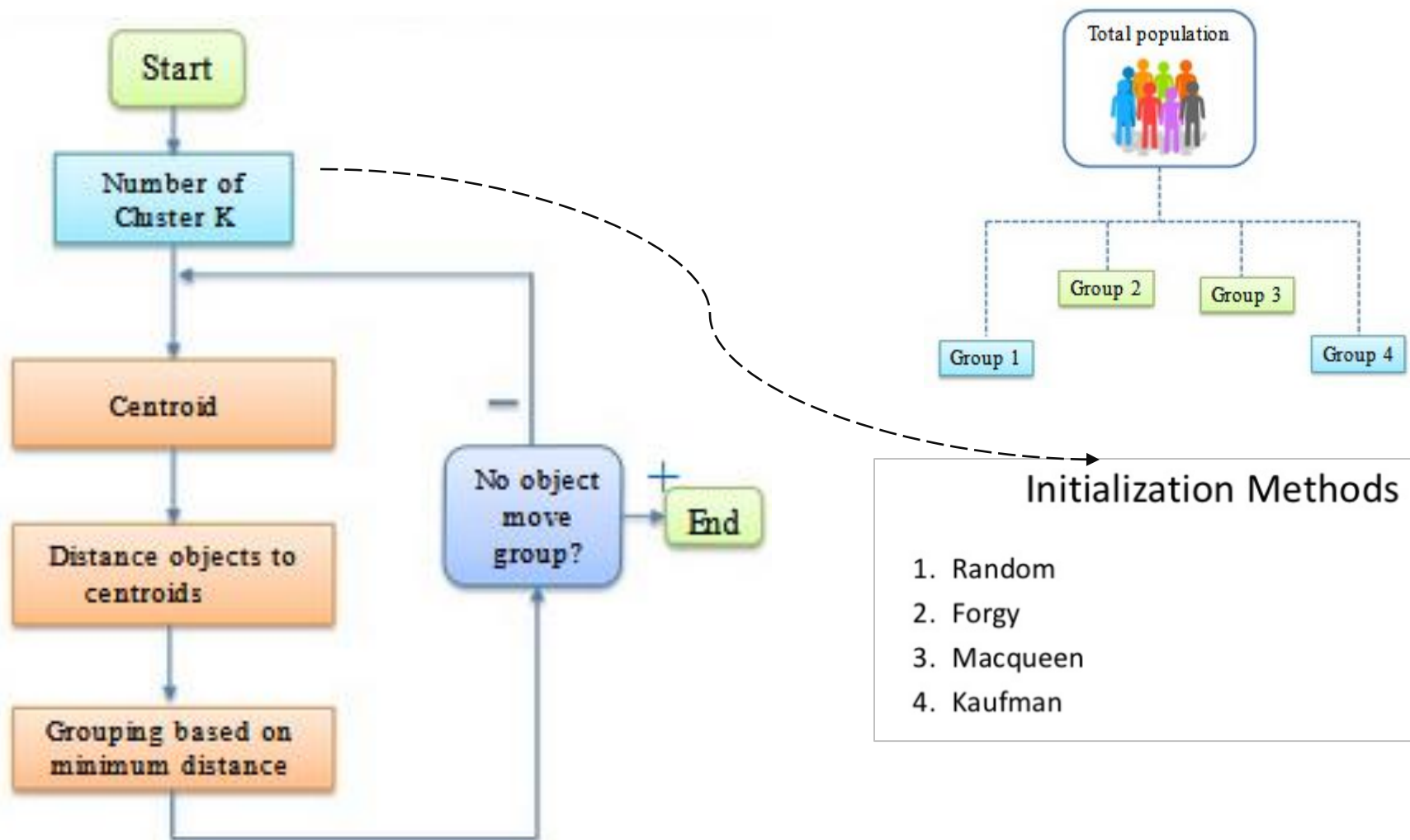
Cluster 1

Cluster 2



Cluster 3

Alg. K-means (Clustering)



Alg. K-means (Clustering)

Exercice : Un enseignant veut segmenter ses étudiants en fonction des notes de contrôles continues pour de créer quatre groupes homogènes (pas forcément du même nombre) afin de faciliter l'affectation des mini projets.

Considérons les résultats suivants : **5, 10, 18, 4, 15, 0, 8, 7**. En appliquant la méthode des k moyennes avec **k=3**, et la fonction de distance $(A,B)=|A-B|$, donnez le résultat de la segmentation en donnant les différentes étapes

1. On choisit les **K** centres de chaque groupe :

$$m_1=5 ; m_2=10 ; m_3=18$$

2. On calcule la distance pour affecter chaque élément au centre le plus proche

<div>éléments</div> <div>centres</div>	5	10	4	15	0	8	18	7
5	0	5	1	10	5	3	13	2
10	5	0	6	5	10	2	10	3
18	13	8	14	3	18	10	0	11

$$K_1=\{5,4,0,7\} \quad K_2=\{10,8\} \quad K_3=\{15,18\}$$

Alg. K-means (Clustering)

Suite Exo :

1. On recalcule les centres de chaque groupe :

$$m_1 = (5+4+0+7)/4 = 4$$

$$m_2 = (10+8)/2 = 9$$

$$m_3 = (15+18)/2 = 16,5$$

1. On affecte les éléments aux nouveaux centres :

éléments \ centres	5	10	4	15	0	8	18	7
4	1	6	0	14	4	4	14	3
9	4	1	5	6	9	1	9	2
16,5	11,5	6,5	12,5	1,5	16,5	8,5	2,5	9,5

$$K_1 = \{5, 4, 0\}$$

$$K_2 = \{10, 8, 7\}$$

$$K_3 = \{15, 18\}$$



Après 6^{ème} itération: On remarque que aucun élément n'a changé son groupe → arrêt de l'algorithme

$$K_1 = \{5, 4, 0\}$$

$$K_2 = \{10, 8, 7\}$$

$$K_3 = \{15, 18\}$$

Clustering: Exercice

	Age	Nombre Enfants	Salaire
P1	40	2	1000
P2	75	4	600
P3	50	3	1100
P4	35	1	550

- segmenter ses employés en fonction de ces trois attributs: Age, nombre d'enfants et salaire ?

Ensemble d'apprentissage

Valeurs des attributs

Attributs			Classes
Revenu	Propriété	Crédit non remboursé	
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

C_1 : Attribuer tout le crédit.

C_2 : Attribuer une partie crédit.

C_3 : Ne pas attribuer le crédit.