Fouille de données - Introduction

G. Forestier

Fouille de données

C. Wemmert, S. Lebre

Pourquoi ?

- Disponibilité croissante de quantité énorme de données :
  - La technologie est disponible
  - Collecte de données : code barre, scanners, satellites, logs des serveurs, etc.
  - Pour aider à stocker : base de données, data warehouses, bibliothèques numériques, www
- Pourquoi l'extraction de connaissances ?
  - Nécessité économique
  - E-commerce
  - Haut degré de concurrence
  - Personnalisation, fidélisation de la clientèle, market segmentation
- Données sur les clients
- Numérisation de textes, images, vidéo, voix, etc.
- Internet et catalogues en ligne

Pourquoi ?

- Données en trop grandes quantités pour être traitées manuellement ou par des algorithmes classiques :
  - Nombre d'enregistrements en millions ou milliards
  - Données de grandes dimensions souvent trop clairsemées
  - Sources de données hétérogènes
- Utilisateur est gavé de données mais en manque de connaissances
  - *"The greatest problem of today is how to teach people to ignore the irrelevant, how to refuse to know things, before they are suffocated. For too many facts are as bad as none at all."* (W.H. Auden)
- De quoi a-t-on besoin ?
  - Extraire des connaissances intéressantes et utiles à partir des données : règles, régularités, irrégularités, motifs, contraintes

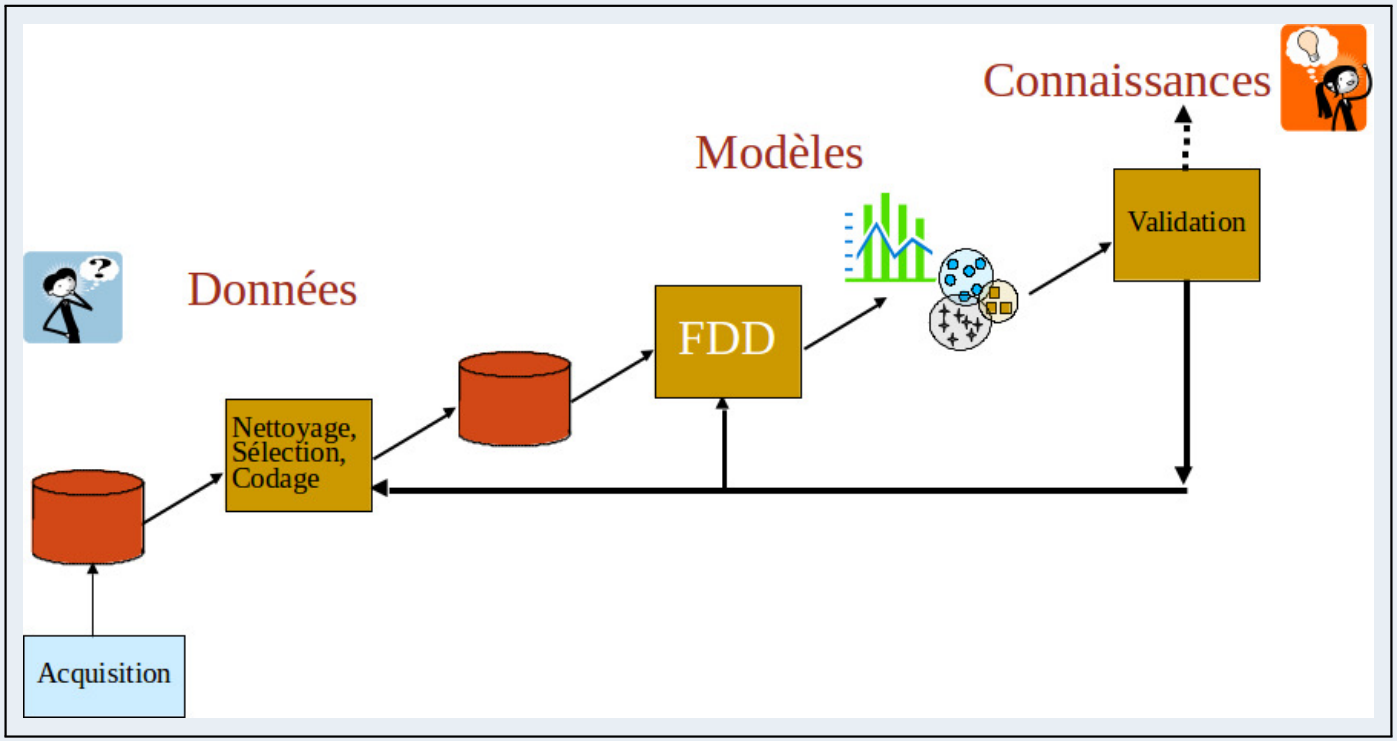
Extraction de Connaissances à partir de Données

- ECD : Extraction d'informations originales (non triviales) implicites, inconnues auparavant et potentiellement utiles à partir de grandes bases de données :
  - Non triviale : sinon la connaissance n'est pas utile
  - Implicite : la connaissance cachée est difficile à observer
  - Inconnue jusqu'alors : évident !
  - Potentiellement utile : utilisable, compréhensible
- ECD : ensemble du processus de découvertes et d'interprétation de régularités dans des données
- Autres appellations :
  - Knowledge Discovery in Databases (KDD)
  - Knowledge extraction
  - Data/pattern analysis
  - Data Analytics
  - Big Data

Le processus de découverte de connaissances dans les données

1. Poser le problème
2. Recherche des données
3. Nettoyage des données
4. Codage des données, actions sur les variables
5. Recherche d'un modèle, de connaissances, etc.
6. Validation et interprétation du résultat, avec retour possible sur les étapes précédentes
7. Intégration des connaissances apprises

Le processus de découverte de connaissances dans les données



ECD - Préparation des données

- Données existantes ou à constituer
  - Fichiers : information contenue dans un ou plusieurs fichiers indépendants
  - BD relationnelles : information contenue dans plusieurs fichiers unis par une 'clé' commune
  - Base de données transactionnelles
- Nettoyage :
  - doublons, erreurs de saisie, valeurs aberrantes, informations manquantes (ignorer l'observation, valeur moyenne, valeur moyenne sur la classe, régression, etc.)

ECD - Préparation des données

- Data Warehouses : entrepôt de données collectées de sources multiples souvent hétérogènes
- Les données sont enregistrées, nettoyées, transformées et intégrées
- Habituellement modélisées par une structure de données multidimensionnelle (cube) :
  - Les données sont structurées suivant plusieurs axes d'analyses (dimensions du cube) comme le temps, la localisation, etc.
  - Une cellule est l'intersection des différentes dimensions.
  - Le calcul de chaque cellule est réalisé au chargement.
  - Le temps de réponse est ainsi stable quelque soit la requête

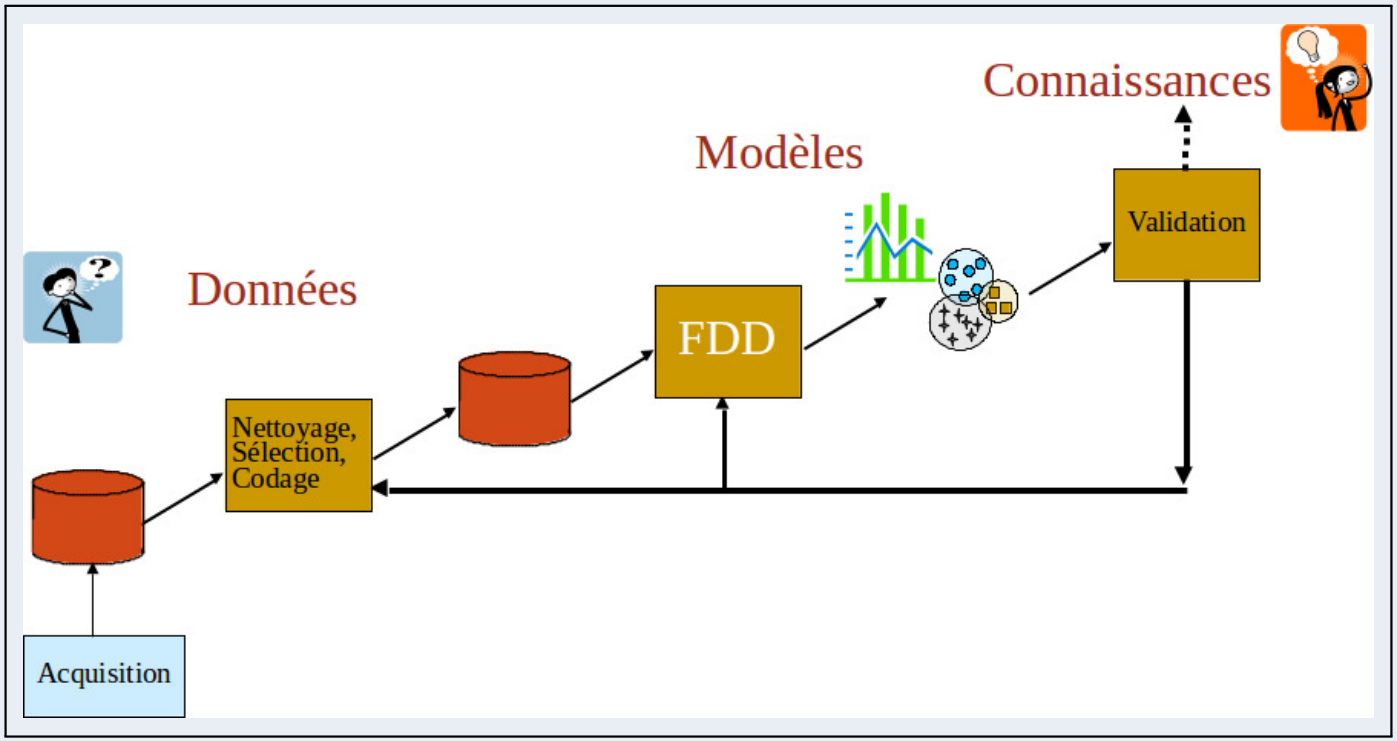
ECD - Préparation des données

- *Data Warehouses* : entrepôt de données collectées de sources multiples souvent hétérogènes
- Les cubes sont bien adaptés aux requêtes rapides et à l'analyse des données : On-Line Analytical Processing (OLAP) :
  - *Quel est le nombre de paires de chaussures vendues par le magasin "OnVendDesChaussuresIci" en mai 2016 ET Comparer les ventes avec le même mois de 2015 et 2014.*
  - *Quelles sont les composantes des machines de production ayant eu le plus grand nombre d'incidents imprévisibles au cours de la période 2015-2016 ?*
- Les réponses aux requêtes OLAP peuvent prendre de quelques secondes à plusieurs minutes.

ECD - Préparation des données

- Sélection des données :
  - Échantillonnage
  - Sélection de sources
- Réduction dimensionnalité :
  - Sélection ou transformation d'attributs
  - Pondération
- Codage :
  - Agrégation (somme, moyenne), discrétisation, codage des attributs discrets, uniformisation d'échelle ou standardisation

Le processus de découverte de connaissances dans les données



Fouille de données

- But : apprendre quelque chose de nouveau !
  - **Concepts** : regroupements basés sur le partage de caractéristiques
  - **Associations** : corrélations entre attributs ou données
- Principes
  - Obtenir le plus haut niveau d'abstraction possible
  - Règles ou vérités qui sont les bases pour d'autres vérités

Fouille de données

Différentes approches :

1. **Estimation** : créer un modèle qui décrit au mieux une variable de prévision liée à des données réelles
2. **Classification** : créer une fonction qui classe un élément parmi plusieurs classes prédéfinies existantes
3. **Regroupement (clustering)** : rechercher à identifier un ensemble fini de catégories ou groupes en vue de décrire les données
4. **Modélisation des dépendances** : trouver un modèle qui décrit des dépendances significatives entre les variables

Fouille de données - Apprentissage

Apprentissage supervisée :

- Modèle inductif où l'apprenant considère un ensemble d'exemples la cible "à apprendre" est connue (classe d'appartenance, propriété, etc.) : les exemples sont étiquetés préalablement

Data Mining prédictif :

- Diviser/regrouper les instances dans des classes spécifiques pour des prédictions futures
- Prédire des valeurs inconnues ou manquantes

Algorithmes :

- Arbres de décision, classifications, algorithmes génétiques, régression linéaire et non linéaire

Fouille de données - Apprentissage

Induction :

- C'est une technique communément utilisée
- Généralisation d'une observation ou d'un raisonnement établie à partir de cas singuliers.
- Elle consiste à tirer des conclusions à partir d'une série de faits

Exemple :

- induction : l'eau, l'huile et le lait se congèlent sous l'influence du froid, nous en inférons que tous les liquides doivent se congeler, pourvu que le froid soit assez intense
- déduction : tous les liquides sont susceptibles de se congeler; or, si le mercure est un liquide, il peut se congeler

Fouille de données - Apprentissage

Apprentissage non supervisé :

- Construction d'un modèle et découverte des relations dans les données sans référence à d'autres données
- On ne dispose d'aucune information a priori sur le données

Data mining explicatif :

- Regrouper les instances dans des classes spécifiques en se basant sur leur ressemblance ou sur le partage de propriétés. Les classes sont inconnues et sont donc créées : elles servent à "expliquer" ou résumer les données
- Mise en relation des données

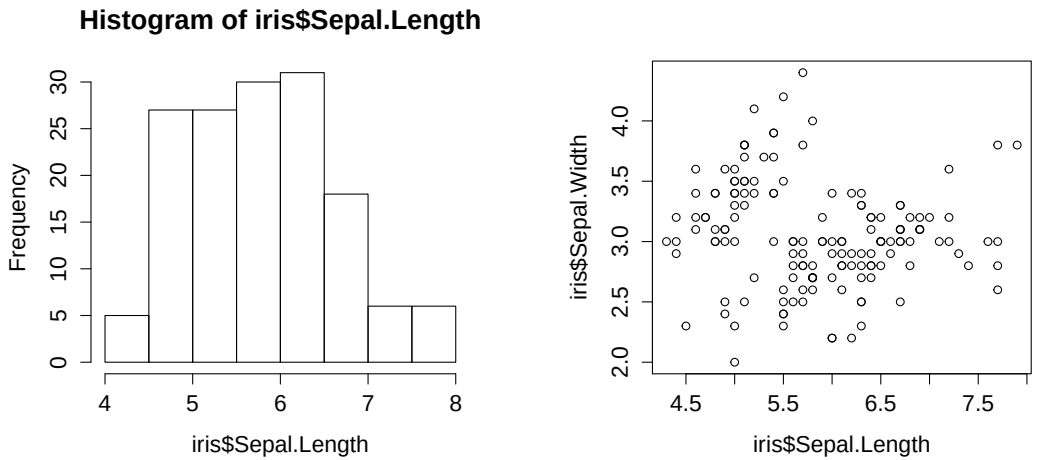
Algorithmes :

- Segmentation, regroupement, découverte d'associations et de règles

Visualisation de données

- Obtenir une représentation visuelle des données
- Pas toujours possible en fonction du type de données
- Pas toujours possible en fonction de la quantité de données

Exemple pratique :



Fouille de données

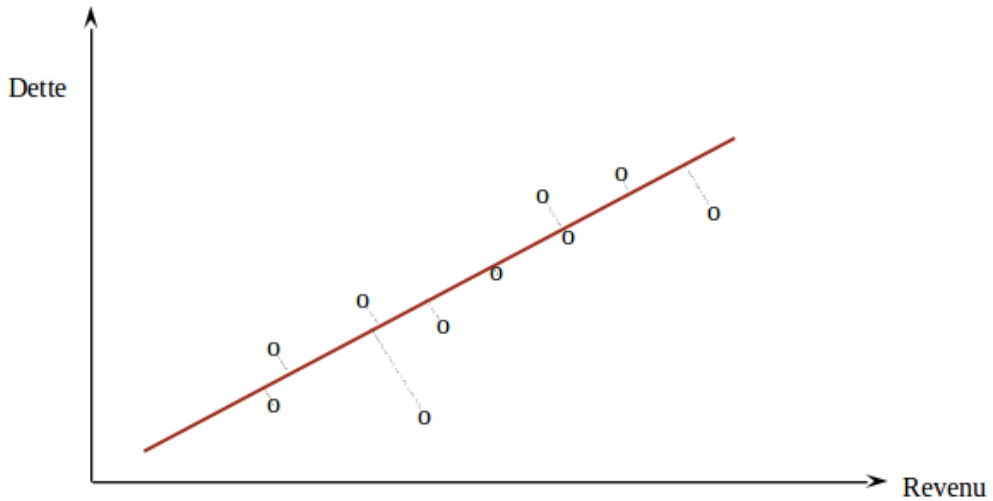
Différentes approches :

1. **Estimation** : créer un modèle qui décrit au mieux une variable de prévision liée à des données réelles
2. **Classification** : créer une fonction qui classe un élément parmi plusieurs classes prédéfinies existantes
3. **Regroupement (clustering)** : rechercher à identifier un ensemble fini de catégories ou groupes en vue de décrire les données
4. **Modélisation des dépendances** : trouver un modèle qui décrit des dépendances significatives entre les variables

Fouille de données - Estimation

Régression :

- Analyser la relation d'une variable vs. une ou plusieurs autres
- Méthode des moindres carrés

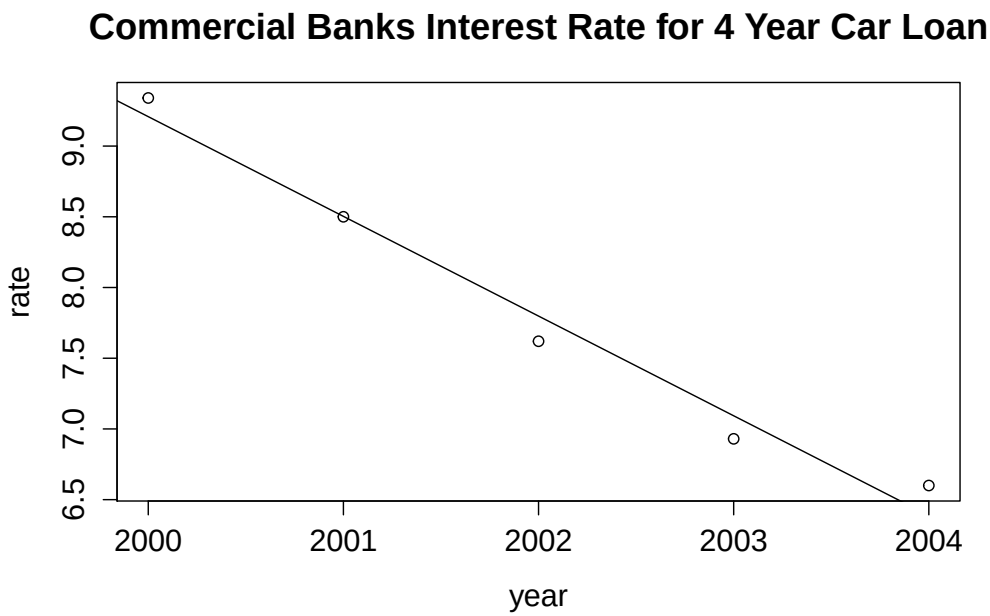


Fouille de données - Estimation

Régression :

- Analyser la relation d'une variable vs. une ou plusieurs autres
- Méthode des moindres carrés

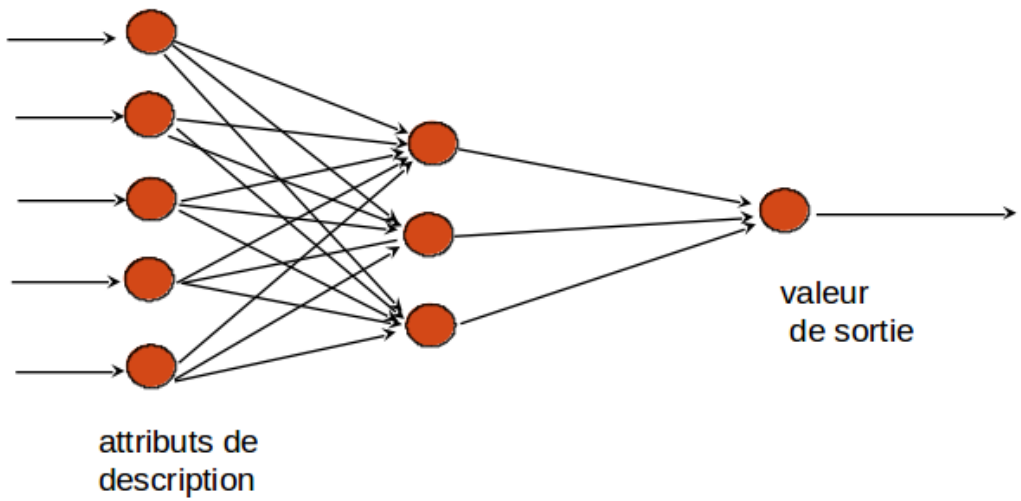
Exemple pratique :



Fouille de données - Estimation

Régression :

- Analyser la relation d'une variable vs. une ou plusieurs autres
- Réseau de neurones



Fouille de données

Différentes approches :

1. **Estimation** : créer un modèle qui décrit au mieux une variable de prévision liée à des données réelles
2. **Classification** : créer une fonction qui classe un élément parmi plusieurs classes prédéfinies existantes
3. **Regroupement (clustering)** : rechercher à identifier un ensemble fini de catégories ou groupes en vue de décrire les données
4. **Modélisation des dépendances** : trouver un modèle qui décrit des dépendances significatives entre les variables

Classification supervisée :

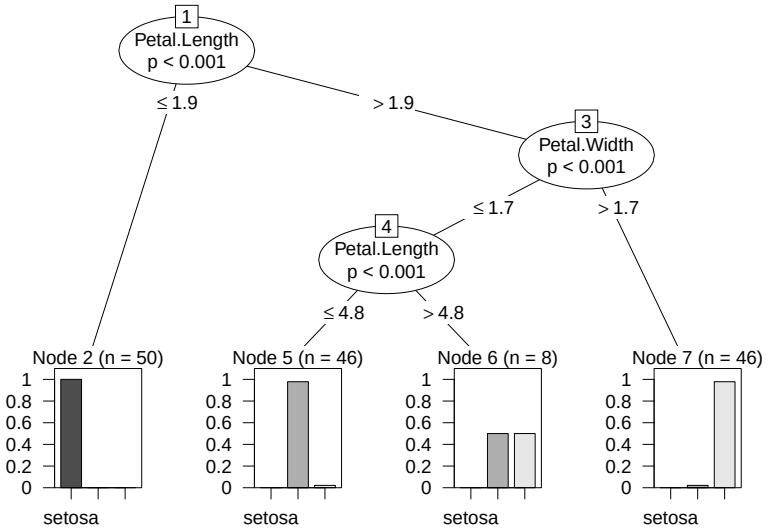
- Division de l'ensemble de données en classes disjointes
- But : recherche d'un ensemble de prédicats caractérisant une classe d'objets et qui peut être appliqué à des objets inconnus pour prévoir leur classe d'appartenance.
- Principales techniques :
  1. Arbres de décision
  2. Classifieur bayésien
  3. K plus proches voisins
  4. Réseaux de neurones
  5. SVM
  6. Algorithmes génétiques

Classification - Arbre de décisions

Arbres de décision :

- Classer les objets en sous-classes par divisions hiérarchiques
- Construction automatique à partir d'un échantillon
- Il existe plusieurs techniques pour construire l'arbre

Exemple pratique :



Fouille de données - Arbre de décisions

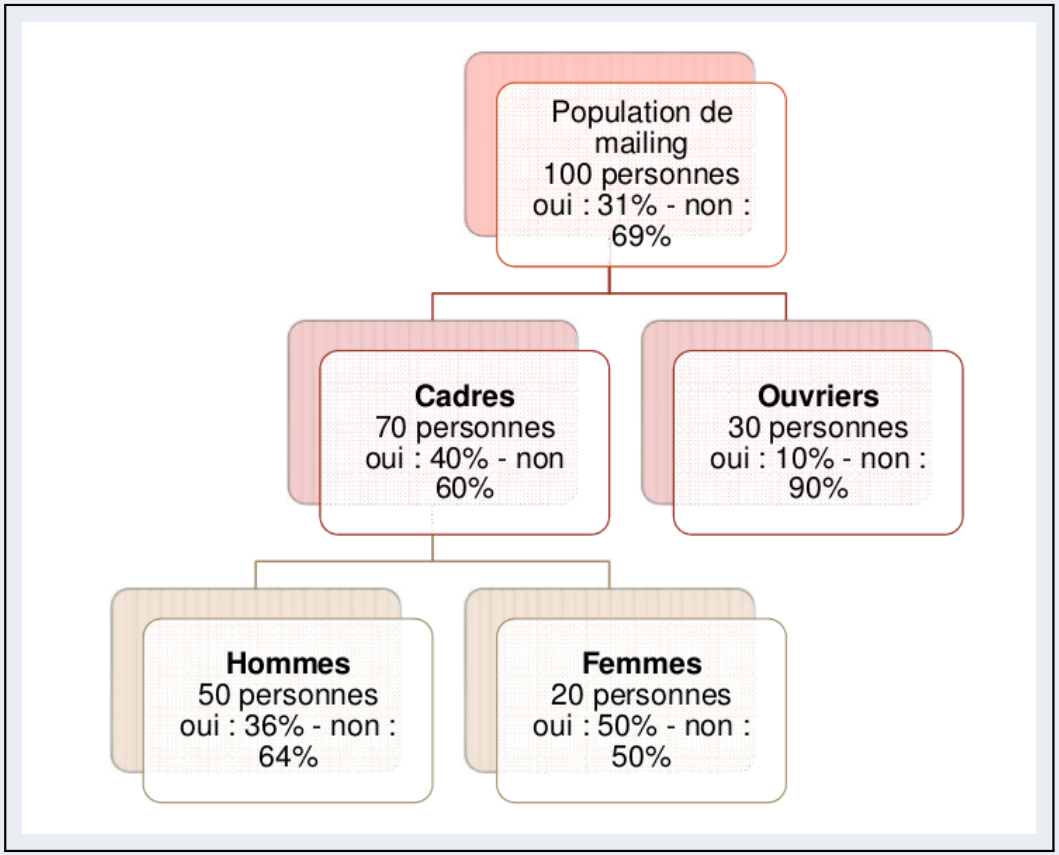
Exemple :

- Un cadeau est envoyé par mailing.
- Un envoi sans réponse coûte 50e et une réponse assure 100e.
- L "oubli" d'un envoi de mailing à un client qui aurait répondu : perte de 100e.
- Tableau des réponses sur un échantillon (taille 100) de la population:

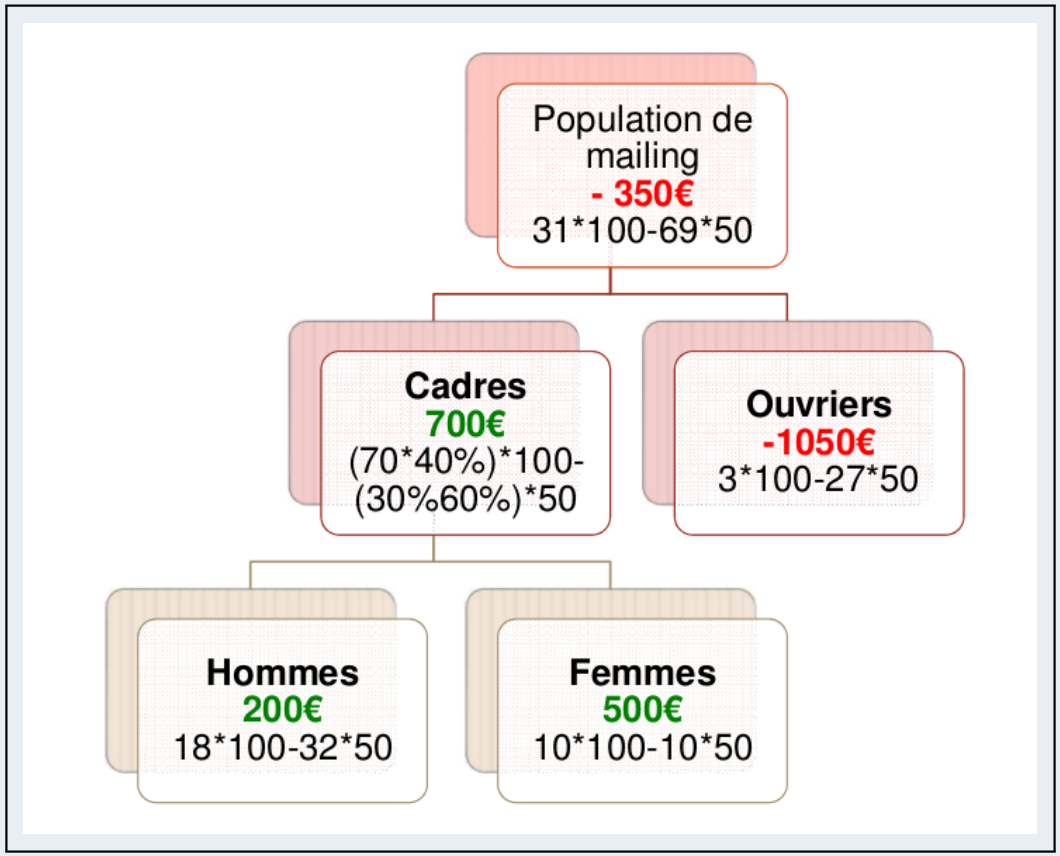
Nom	Prénom	Sexe	Profession	Réponse
Martin	Jeanne	F	Cadre	ok
Berluchette	Huguette	F	Ouvrière	ok
Sarkau	Sy	M	Ouvrier	non
Vil	Dominique	M	Cadre	non
Maitre	Kanter	M	Cadre	ok

- Question: A quelle catégorie de la population faut-il envoyer le mail ?

Fouille de données - Arbre de décisions



Fouille de données - Arbre de décisions



mailing aux cadres ou uniquement aux femmes cadres

Classification - Classifieur bayésien

Classifieur bayésien :

- Cherche à optimiser la probabilité $P(c_k|x)$ c-à-d de $P(x|c_k).P(c_k)/P(x)$ c-à-d $P(x|c_k).P(c_k)$ car $P(x)$ ne dépend pas de $P(c_k)$
- Les attributs sont supposés indépendants : $P(c_k) = n_k/n$ et $P(x|c_k) = \prod P(x_i|c_k)$

Exercice :

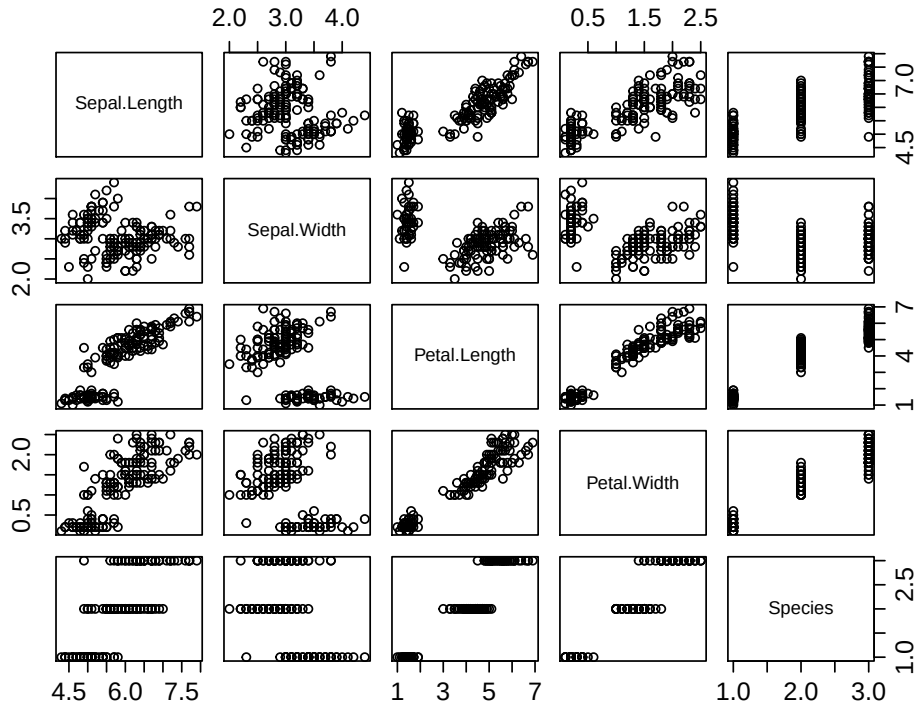
- Deux classes :
  - c1 = 01100, 11001, 10110, 10101, 10010
  - c2 = 01010, 11111, 11010, 11101, 10101
- Classer x = 00111

Classification - Classifieur bayésien

Classifieur bayésien :

- Très usilisé en classification de texte
- Marche avec peu de données, possibilité de mise à jour

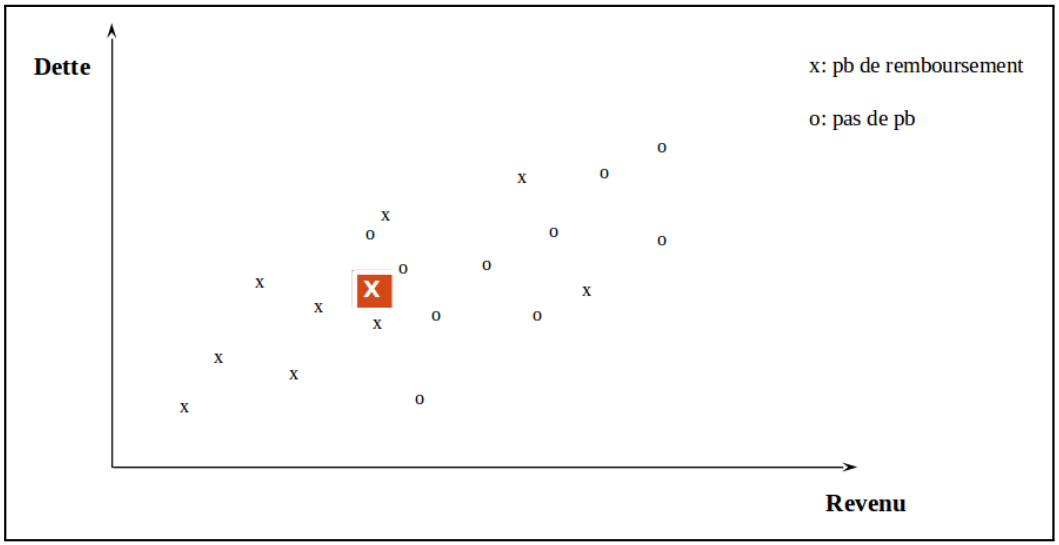
Exemple pratique :



Classification - K plus proches voisins

K plus proches voisins :

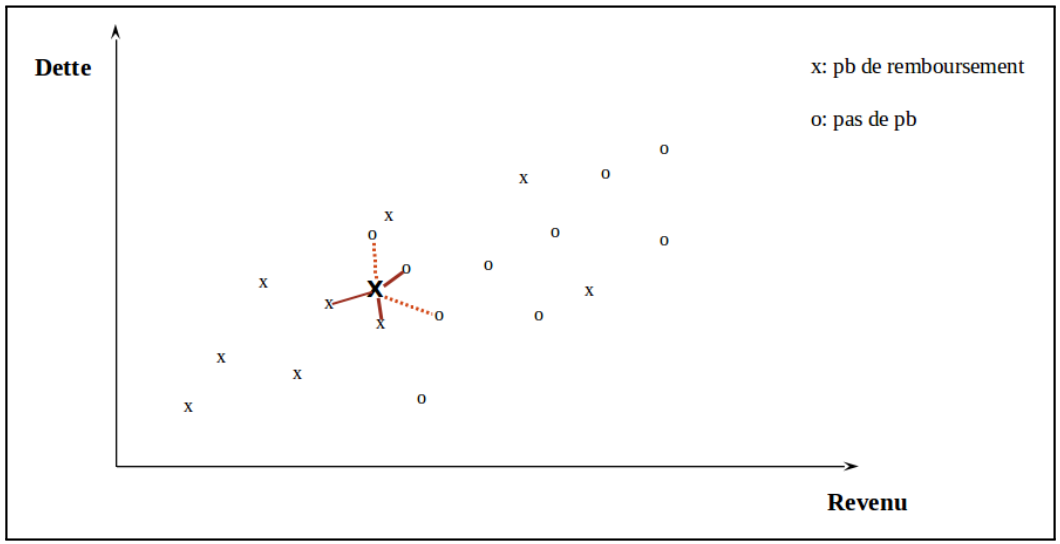
- On calcule toutes les distances entre le point X à classer et tous les points donc on connaît la classe : on conserve les K plus proches. La classe majoritaire dans cet ensemble est attribuée à X.



Classification - K plus proches voisins

K plus proches voisins :

- On calcule toutes les distances entre le point X à classer et tous les points donc on connaît la classe : on conserve les K plus proches. La classe majoritaire dans cet ensemble est attribuée à X.



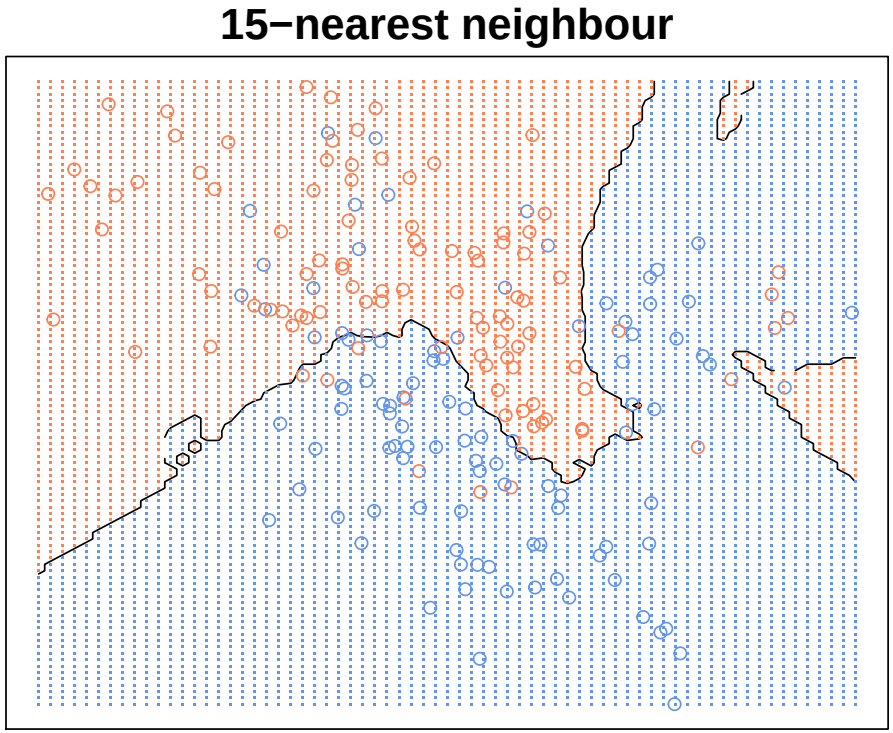
$K = 3 \rightarrow x$, $K = 5 \rightarrow o$

Classification - K plus proches voisins

K plus proches voisins :

- Pas d'hypothèse sur la "forme" des classes
- Complexité croissante avec la taille de la base d'entrainement

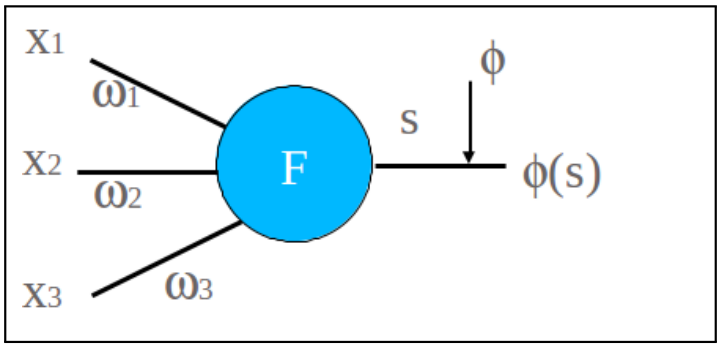
Exemple pratique :



Classification - Réseaux de neurone

Réseaux de neurones :

- Inspirés de la structure du système nerveux
- Un grand nombre de neurones connectés qui traitent l'information
- La réponse du neurone dépend de son état et des poids des connexions
- Les poids (ou forces) sont développées par expérience



Classification – Réseaux de neurones

Principe :

- Construction d'un réseau d'unités calculatoires simples (neurone) liées par des connexions
- Apprentissage des paramètres du réseau (poids des connexions) grâce à un ensemble d'exemples

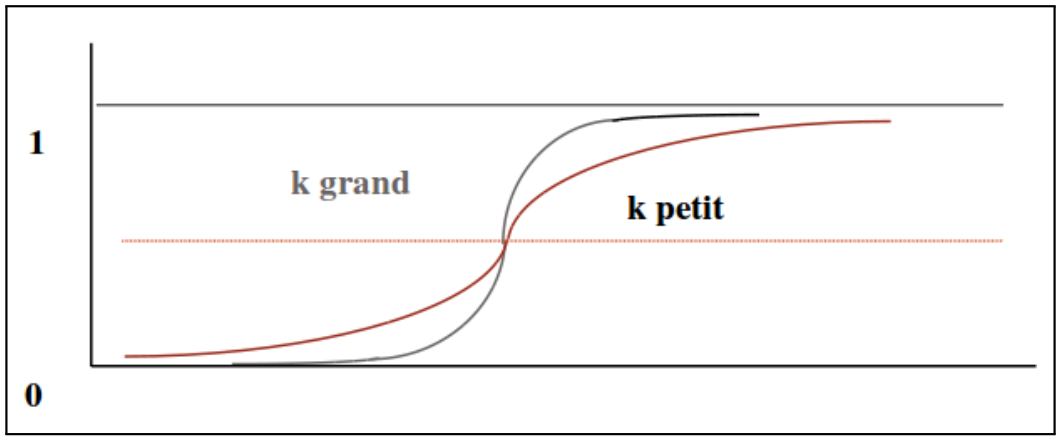
Un neurone est formé :

- d'entrées (connexions entrantes ou variables d'entrée)
- de poids sur les connexions entrantes
- d'une fonction F qui calcule une sortie en fonction des entrées et des poids sur les entrées
- d'une fonction d'activation Φ qui modifie l'amplitude de la sortie du nœud

Classification – Réseaux de neurones

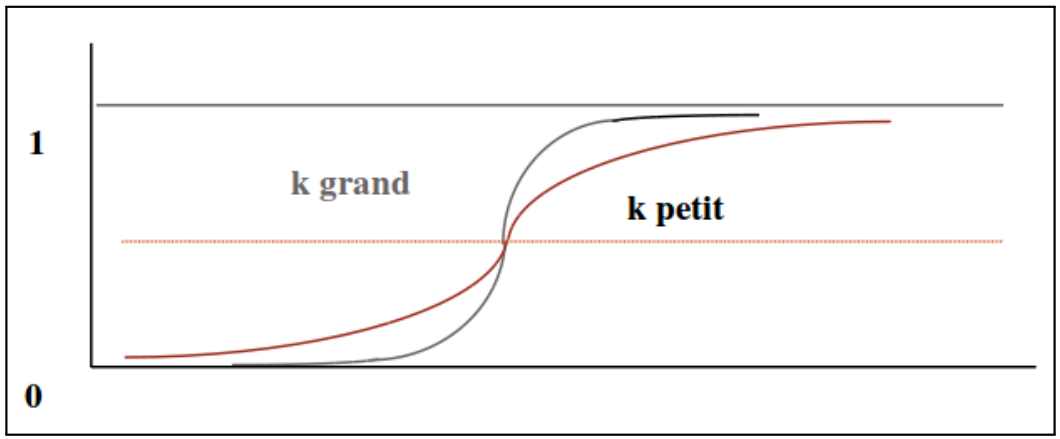
Fonction d'activation :

- $\Phi(s) =$ linéaire
- $\Phi(s) =$ seuil
  - $\Phi(s) = 0 \quad \text{si} \quad s \leq a$
  - $\Phi(s) = 1 \quad \text{si} \quad s \geq a$
- $\Phi(s) = 1/(1 + e^{-ks})$



Classification – Réseaux de neurones

- $\Phi(s) = 1/(1 + e^{-ks})$
- Si le coefficient k est grand, alors la sortie est presque toujours proche de 0 ou de 1 : réseau neuronal relativement symbolique
- Si le coefficient k de $1/(1 + e^{-ks})$ est petit, alors la force de chaque cellule est bien distribuée entre 0 et 1
- Un autre paramètre, implicite, est le centre de la fonction sigmoïde.



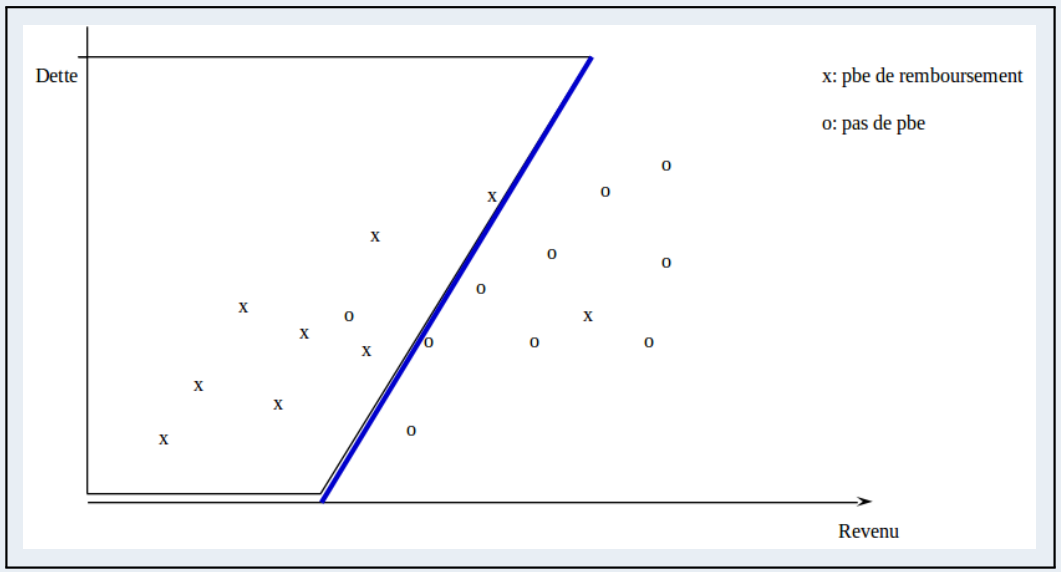
Classification – Réseaux de neurones

Cas le plus simple :

- Un seul neurone
- F = somme pondérée des entrées
- Φ = seuillage
- $\Phi(s) = 1$ si $s > a$ sinon 0
- $\rightarrow$ s = 1 si $w_1 x_1 + w_2 x_2 + ... > a$
- $\rightarrow$ s = 1 si $w_1 x_1 + w_2 x_2 + ... - a > 0$
- $\rightarrow$ équation d'un hyperplan

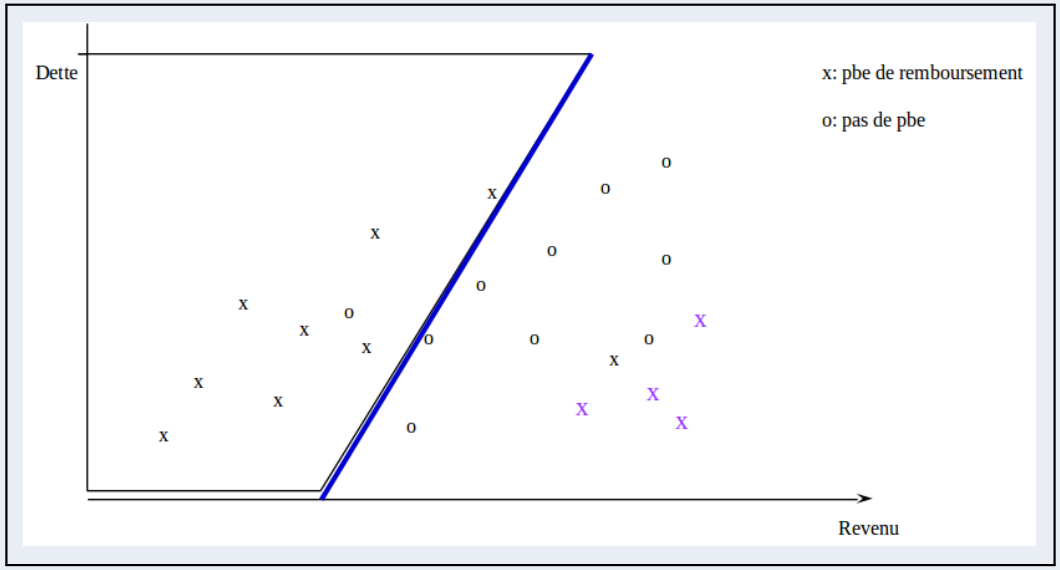
Classification – Réseaux de neurones

Séparation linéaire :



Classification – Réseaux de neurones

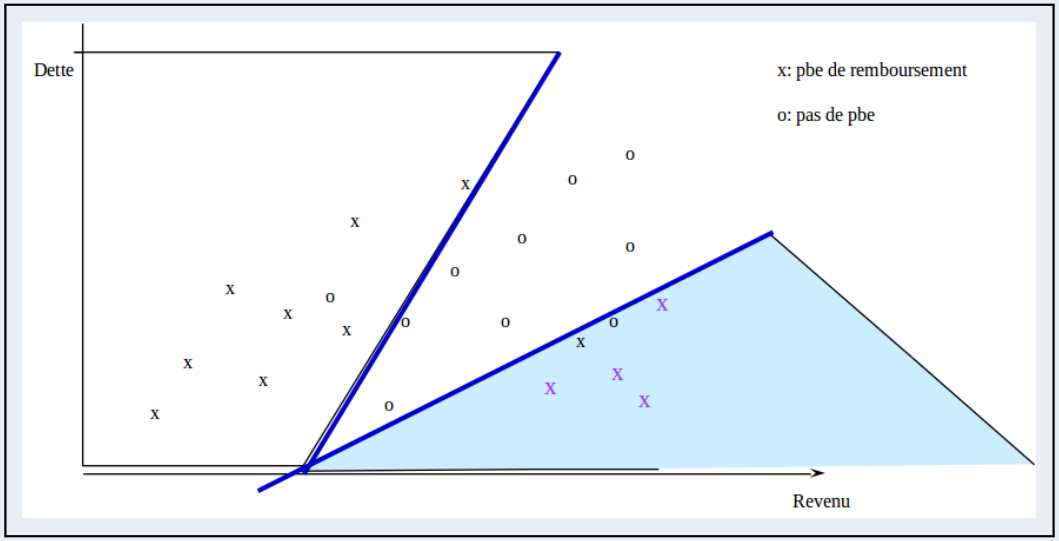
On rajoute des exemples :



Trouver un réseau de neurones discriminant les deux classes ?

Classification – Réseaux de neurones

On rajoute des exemples :



Classification – Réseaux de neurones

En pratique :

- On choisit une fonction de calcul et une fonction d'activation
- On choisit une architecture:
  - Nombre d'entrées
  - Nombre de sorties
  - Nombre de couches internes
  - Nombre de neurones de chacune des couches internes
- On choisit une fonction d'erreur
- On définit un critère d'arrêt

Classification – Réseaux de neurones

Avantages des réseaux de neurones :

- Méthode robuste au bruit
- Classement ou estimation rapide une fois le réseau construit
- Disponible dans tous les logiciels de fouille de données

Inconvénients :

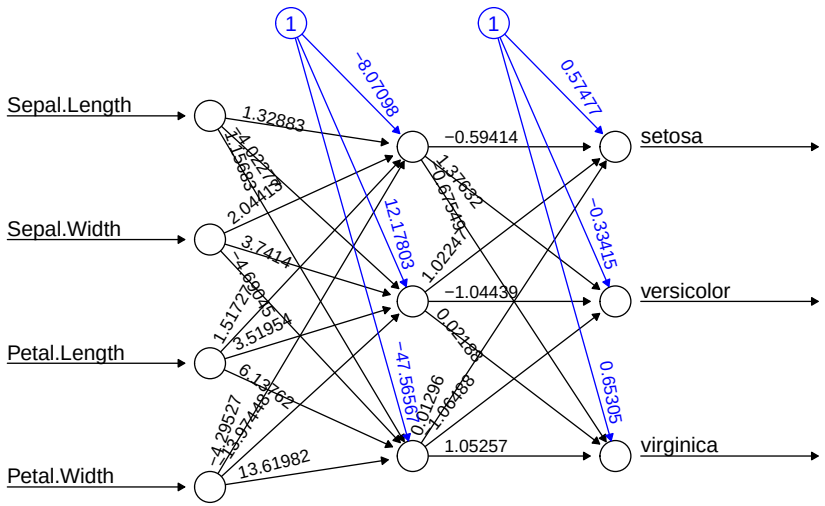
- Boîte noire: difficile d'interpréter le modèle obtenu
- Temps d'apprentissage important
- Difficulté de choix des paramètres

Classification – Réseaux de neurones

Réseaux de neurones :

- Regain de popularité avec l'arrivée du Deep Learning
- Très utilisé en analyse d'images (Convolution Neural Network)

Exemple pratique :

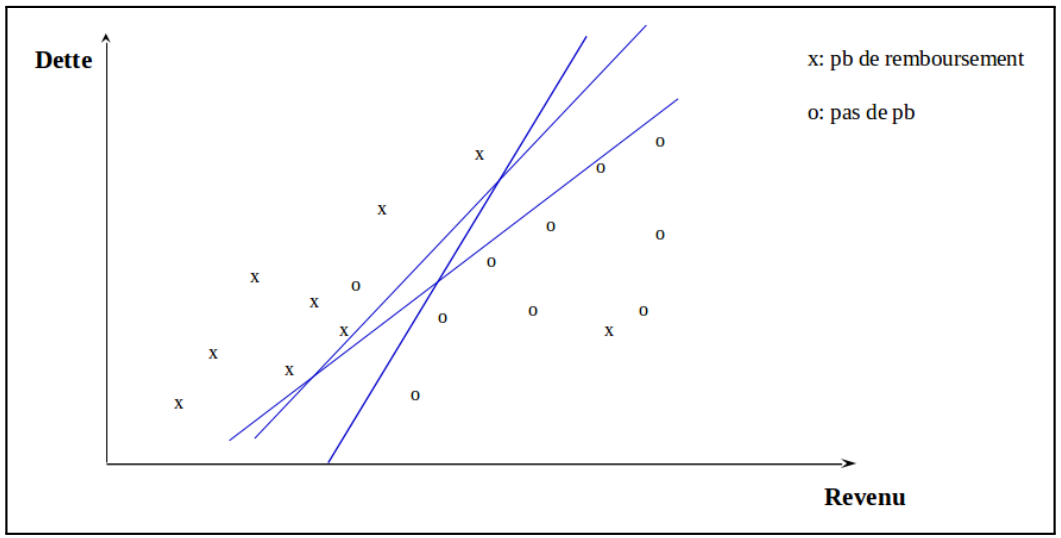


Error: 0.495661 Steps: 7450

Classification - SVM

SVM (Support Vector Machines) :

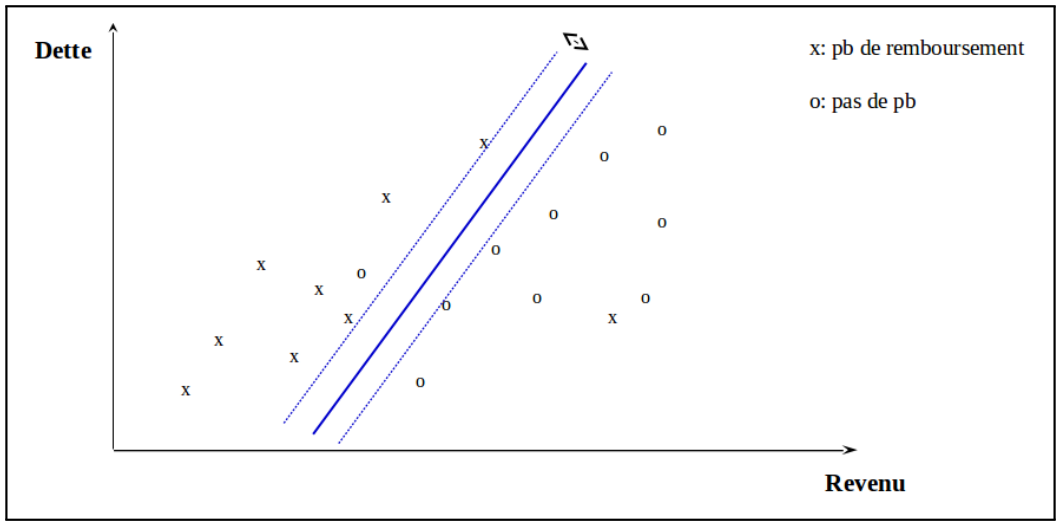
- Principe : séparer les données en deux classes par un hyperplan tout en maximisant l'écart entre cet hyperplan et les données
- Trouver la droite qui maximise l'écart parmi plusieurs possibles



Classification - SVM

SVM (Support Vector Machines) :

- Principe : séparer les données en deux classes par un hyperplan tout en maximisant l'écart entre cet hyperplan et les données
- Trouver la droite qui maximise l'écart parmi plusieurs possibles

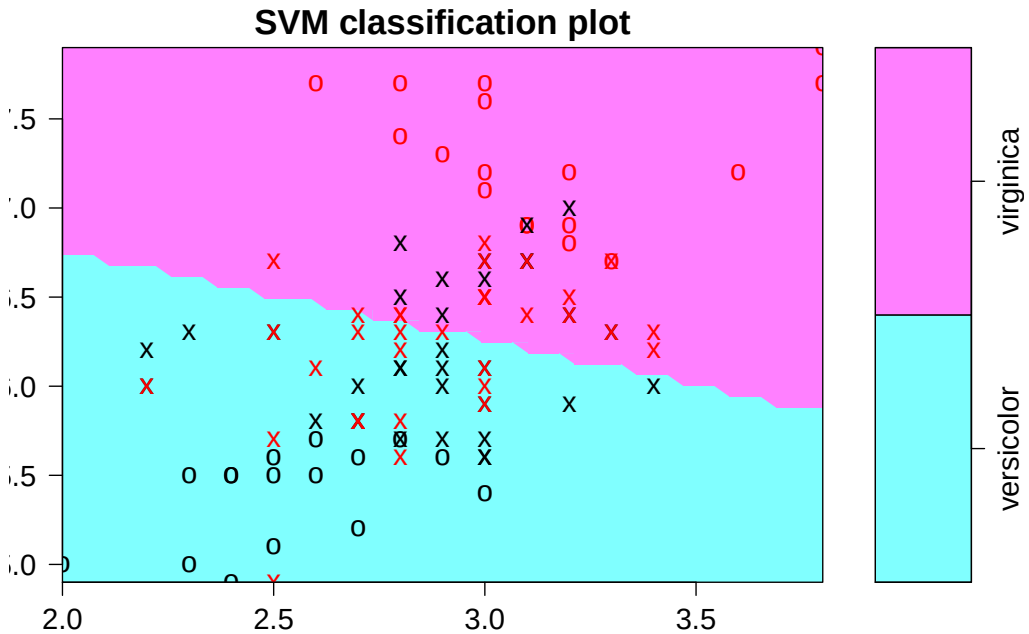


Classification – SVM

SVM (Support Vector Machines) :

- Nécessité de trouver un noyaux adapté pour transformer les données
- Exemple du noyau gaussien $K(x, y) = exp(-\frac{\|x-y\|}{2\sigma^2})$

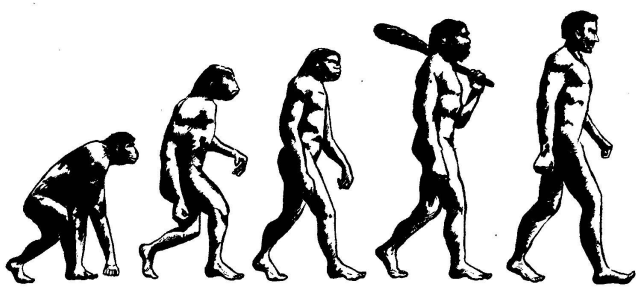
Exemple pratique :



Classification - Algorithmes génétiques

Algorithmes génétiques :

- Inspirés des théories de l'évolution de Darwin, Lamarck ou Baldwin
- Méthode générale d'optimisation
- Peut être utilisé en classification ou estimation



Fouille de données – Approches génétiques

Schéma général :

- On définit les “paramètres” à optimiser : intervalle de valeurs, seuils, etc. On définit le génotype correspondant (chromosomes).
- On définit la fonction de calcul du phénotype et la fonction d'évaluation d'un individu
- On définit les mécanismes et taux de croisement et de mutation
- On définit la fonction de sélection des survivants

Fouille de données – Approches génétiques

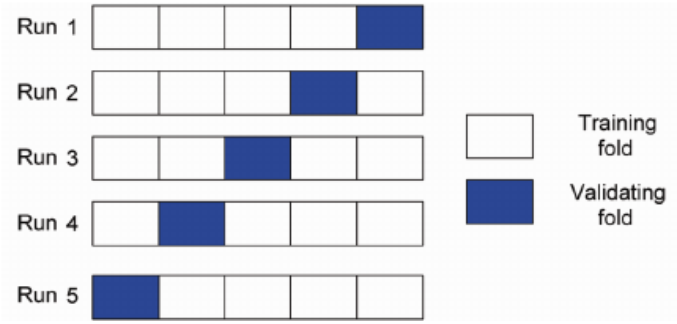
Schéma général :

1. initialiser la population
2. calculer le degré d'adaptation $f(x)$ de chaque individu
3. Tant que non fini ou non convergence
   - 3.1 reproduction des parents :
     - sélectionner 2 individus à la fois
     - appliquer les opérateurs génétiques
   - 3.2 calculer le degré d'adaptation $f(x)$ de chaque enfant
   - 3.3 sélectionner les survivants parmi les parents et les enfants

Classification - Validation

Validation par le test :

- Données = ensemble d'apprentissage + ensemble de test
- Construction d'un modèle sur l'ensemble d'apprentissage et test du modèle sur le jeu de test pour lequel les résultats sont connus
- Cross-validation (validation croisée)

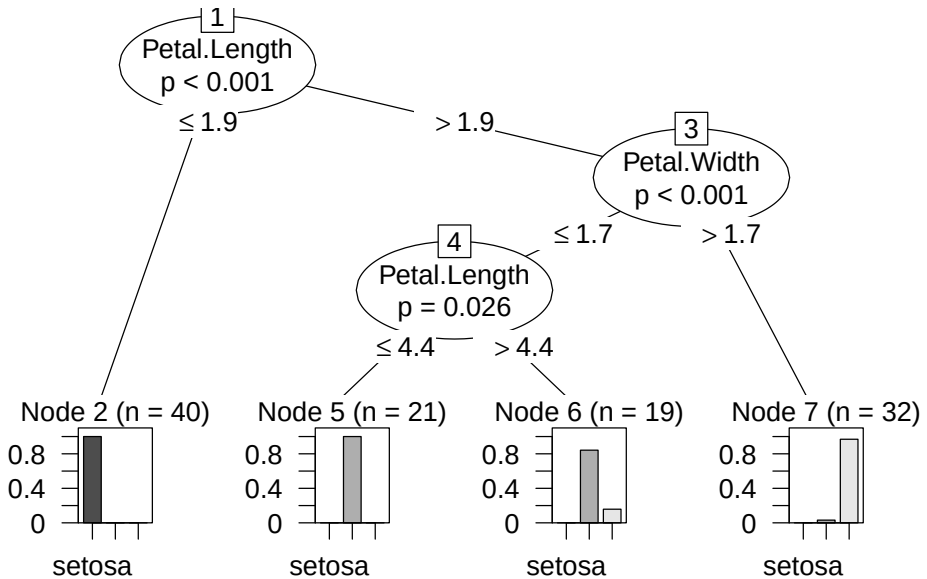


Classification – SVM

Validation :

- Split test / train, en général plus de données pour l'apprentissage
- Le nombre de fold de la cross-validation dépend du volume des données

Exemple pratique :



Fouille de données

Différentes approches :

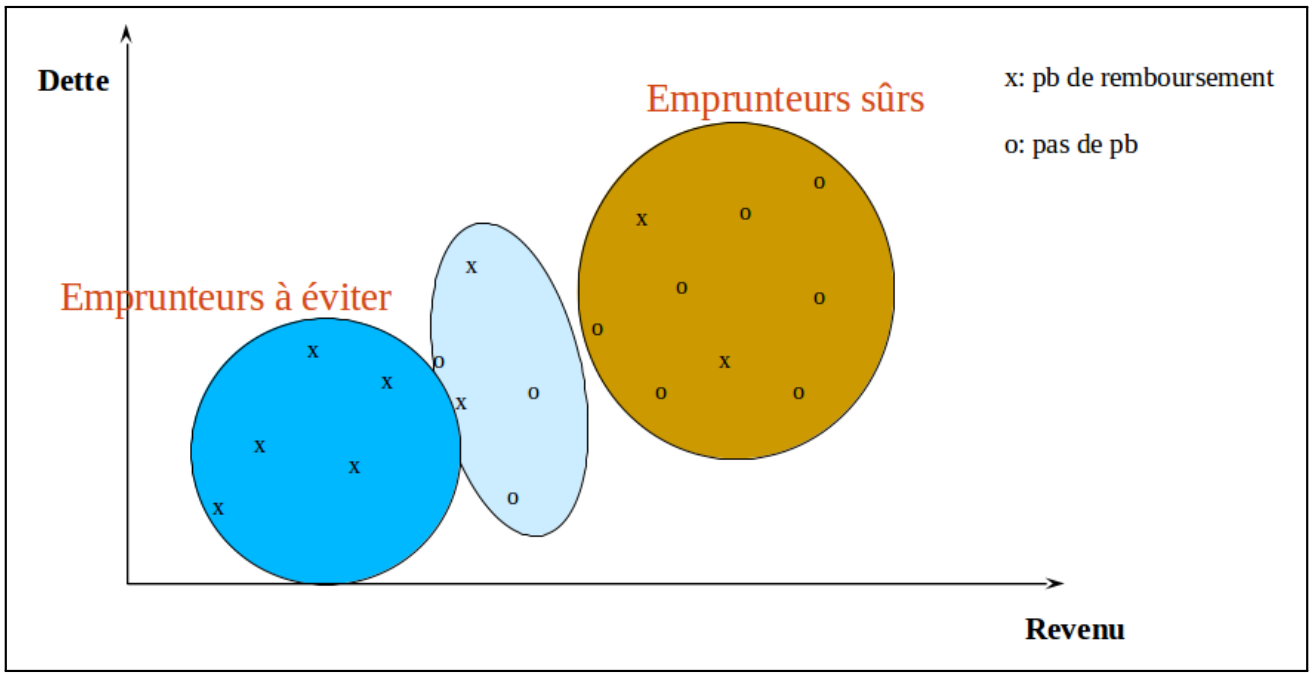
1. **Estimation** : créer un modèle qui décrit au mieux une variable de prévision liée à des données réelles
2. **Classification** : créer une fonction qui classe un élément parmi plusieurs classes prédéfinies existantes
3. **Regroupement (clustering)** : rechercher à identifier un ensemble fini de catégories ou groupes en vue de décrire les données
4. **Modélisation des dépendances** : trouver un modèle qui décrit des dépendances significatives entre les variables

Fouille de données – Clustering

- But du clustering : obtenir une représentation simplifiée (structuration) des données initiales
- Organisation d'un ensemble d'objets en un ensemble de regroupements homogènes et/ou naturelles

Fouille de données – Clustering

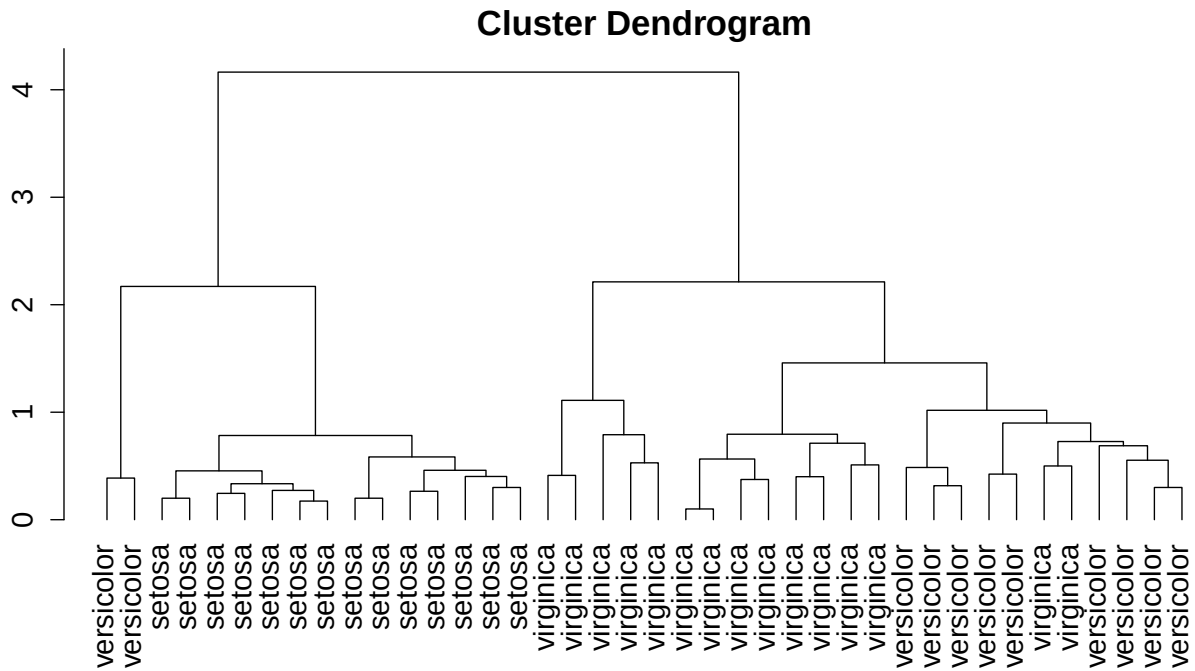
- Partitionnement automatique à partir des données
- Aucune interprétation sémantique a priori



Fouille de données – Clustering

- Clusterirng hiérarchique

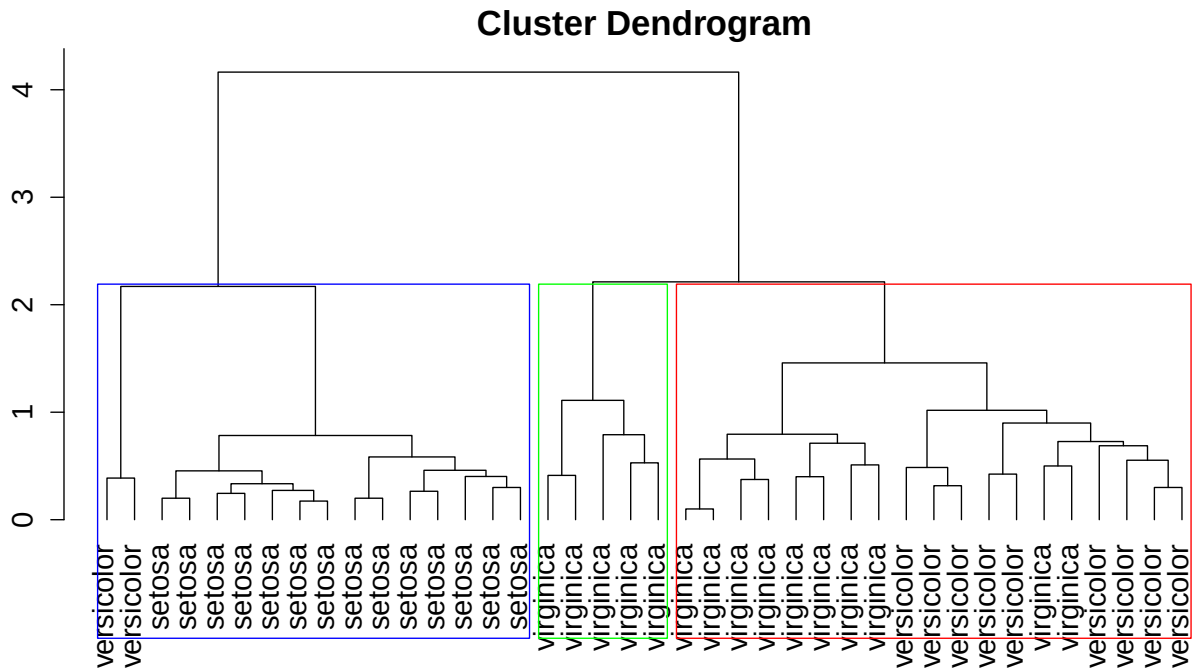
Exemple pratique :



Fouille de données – Clustering

- Clusterirng hiérarchique

Exemple pratique :



Fouille de données

Différentes approches :

1. **Estimation** : créer un modèle qui décrit au mieux une variable de prévision liée à des données réelles
2. **Classification** : créer une fonction qui classe un élément parmi plusieurs classes prédéfinies existantes
3. **Regroupement (clustering)** : rechercher à identifier un ensemble fini de catégories ou groupes en vue de décrire les données
4. **Modélisation des dépendances** : trouver un modèle qui décrit des dépendances significatives entre les variables

Fouille de données – Associations

- Règles d'associations : analyse du panier de courses
- *"Le jeudi, les clients achètent souvent en même temps des packs de bière et des couches"*
- Y-a-t-il des liens de causalité entre l'achat d'un produit P et d'un autre produit P' ?

Fouille de données – Associations

Identifiant	Transaction
1	beurre fruits lait pain
2	fruits lait pain
3	beurre fromage pain pâtes viande vin
4	fromage fruits lait légumes pain pâtes poisson
5	beurre fruits lait légumes pain pâtes poisson viande
6	beurre fromage légumes pain pâtes viande vin
7	beurre fromage lait légumes pain pâtes viande vin
8	fruits légumes poisson
9	beurre fromage lait pain pâtes viande vin
10	beurre fromage fruits lait légumes pain poisson viande

- Règle d'association : prémisse $\rightarrow$ conclusion
- Questions :
  - beurre $\rightarrow$ pain ?
  - poisson viande $\rightarrow$ lait ?
  - fromage pâtes $\rightarrow$ vin ?

Fouille de données – Associations

Formellement :

- Etant donné un ensemble de transactions D, trouver toutes les règles d'association $X \rightarrow Y$ ayant un **support** et une **confiance** supérieurs aux seuils minimaux prédéfinis par l'utilisateur
- Un transaction est un ensemble d'attributs (beurre, fruit, lait, pain)
- Support : % de transactions dans D qui contiennnent X et Y
- Confiance : % de transactions qui contiennnent X parmi celle contenant Y

Fouille de données – Associations

Interprétation :

- $R : X \rightarrow Y$ (A%, B%)
- A% de toutes les transactions montrent que X et Y ont été achetés en même temps (support de la règle) et B% des clients qui ont acheté X ont aussi acheté Y (confiance dans la règle).

Fouille de données – Associations

Deux sous-problèmes :

- Trouver tous les ensembles fréquents (*itemsets*) ayant un support supérieur ou égal à une valeur minimale minsup donnée : FIS
- A partir des FIS, engendrer l'ensemble des règles d'association ayant une confiance supérieure ou égale à minconf

Fouille de données – Associations

Tickets	Produits achetés
1	beurre fruits lait pain
2	fruits lait pain
3	beurre fromage pain pâtes viande vin
4	fromage fruits lait légumes pain pâtes poisson
5	beurre fruits lait légumes pain pâtes poisson viande
6	beurre fromage légumes pain pâtes viande vin
7	beurre fromage lait légumes pain pâtes viande vin
8	fruits légumes poisson
9	beurre fromage lait pain pâtes viande vin
10	beurre fromage fruits lait légumes pain poisson viande

$$\text{Support} = \frac{\square \text{ et } \square}{\text{Tous tickets}}$$

$$\text{Confiance} = \frac{\square \text{ et } \square}{\square + \bigcirc}$$

		Support	Confiance
beurre	→ pain	70%	100%
poisson viande	→ lait	20%	100%
fromage pâtes	→ vin	40%	80%

Fouille de données – Associations

Tickets	Produits achetés
1	beurre fruits lait pain
2	fruits lait pain
3	beurre fromage pain pâtes viande vin
4	fromage fruits lait légumes pain pâtes poisson ~~vin~~
5	beurre fruits lait légumes pain pâtes poisson viande
6	beurre fromage légumes pain pâtes viande vin
7	beurre fromage lait légumes pain pâtes viande vin
8	fruits légumes poisson
9	beurre fromage lait pain pâtes viande vin
10	beurre fromage fruits lait légumes pain poisson viande

$$\text{Support} = \frac{\square \text{ et } \square}{\text{Tous tickets}}$$

$$\text{Confiance} = \frac{\square \text{ et } \square}{\square + \bigcirc}$$

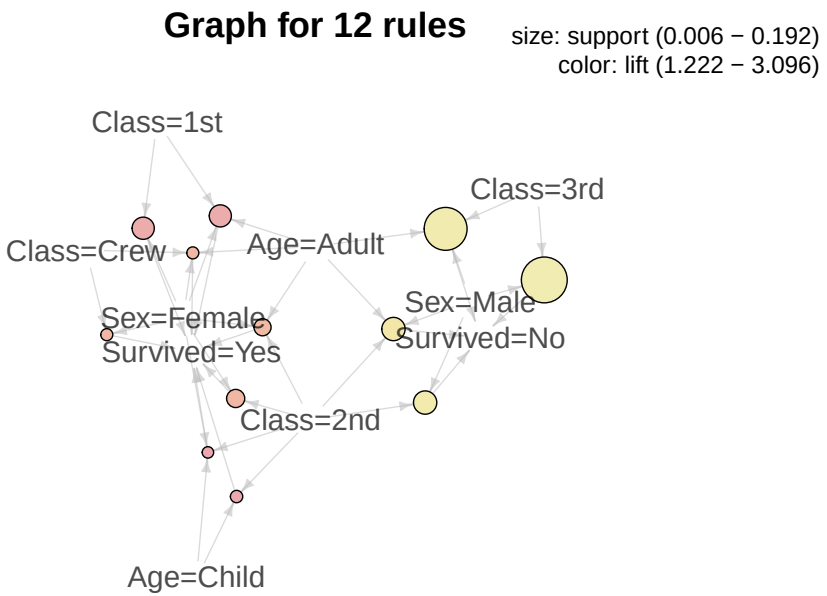
		Support	Confiance
beurre	→ pain	70%	100%
poisson viande	→ lait	20%	100%
fromage pâtes	→ vin	40%	80%

Fouille de données – Associations

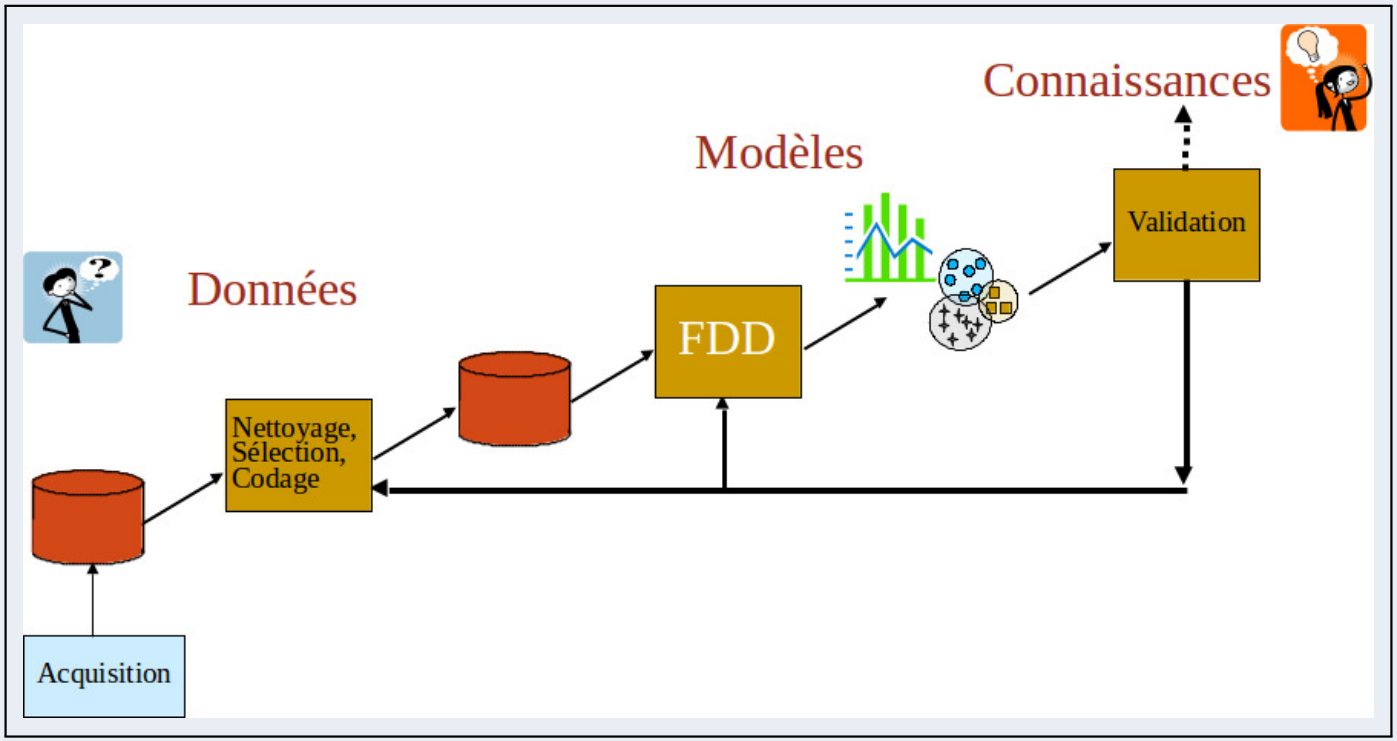
Règles d'associations :

- Nombreux critères pour évaluer l'intérêt d'une règle
- Difficile de passer à l'échelle sur de gros volumes de données

Exemple pratique :



Le processus de découverte de connaissances dans les données



ECD - Validation

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

ECD - Validation

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

Conclusion – Quelques fausses idées

- Méthodes plus inductives que basées sur des hypothèses car il n'y a aucun a priori sur les données

Faux: condition d'application des méthodes, choix des données, codage des données, choix des variables explicatives, des variables à expliquer, ordre d'entrée des variables dans l'algorithme,...

Conclusion – Quelques fausses idées

- Il faut utiliser systématiquement toutes les données informatiquement disponibles ainsi tout va apparaître

Faux: codage des données, ordre d'entrée des variables dans l'algorithme, effectifs irréguliers, outliers, influence des redondances, des corrélations, du modèle de données informatique, saturation, instabilité...

Conclusion – Quelques fausses idées

- Avec toutes ces techniques, on va toujours faire des découvertes incroyables

Faux: il faut trouver des solutions conformes au bon sens (spécialistes, experts métier). En fait, trouver la meilleure solution (parmi n) pour une problématique donnée

Conclusion – Quelques fausses idées

- Le data mining est révolutionnaire

Faux: analyses de données traditionnelles + méthodes plus spécifiques (réseaux de neurones). Optimisation des techniques car grand nombre de données.

Conclusion

Question :

- Pourquoi tant d'algorithmes ?

Réponse :

- Parce qu'aucun n'est optimal dans tous les cas

Comme ils s'avèrent en pratique complémentaires les uns des autres, en les combinant intelligemment (en construisant des méta modèles) il est possible d'obtenir des gains de performance très significatifs