Introduction to Artificial Intelligence for Biology

Jean-Paul Comet (projet Bioinfo Formelle)

EPU dept GB-BIMB - 5ème année
Université Côte d'Azur

23 septembre 2024

inspired by : Frédéric Précioso, Eric Debreuve, Rémi Eyraud, Cécile Capponi, Marie Cottrell, Patrick Gallinari, Philippe Preux, Nicolas P. Rougier, Jérémy Fix, Hervé Frezza-Buet, Matthieu Geist, Frédéric Pennerath, Ricco Rakotomalala, et bien d'autres...

Teaching organization

- 8 sessions of 3 hours
- 50% lectures + 50% TDs
- teachers :
  Jean-Paul Comet `Jean-Paul.Comet@univ-cotedazur.fr`

	sessions	hours	teacher
1	23 septembre 2024	8h30-11h45	JPC
2	7 octobre 2024	8h30-11h45	JPC
3	21 octobre 2024	8h30-11h45	JPC
4	4 novembre 2024	13h30-16h45	JPC
5	18 novembre 2024	8h30-11h45	JPC
6	2 dcembre 2024	8h30-11h45	JPC
7	4 décembre 2023	8h30-11h45	JPC
8	18 décembre 2023	8h30-11h45	JPC
–	22 janvier 2025	9h-12h	Presentations

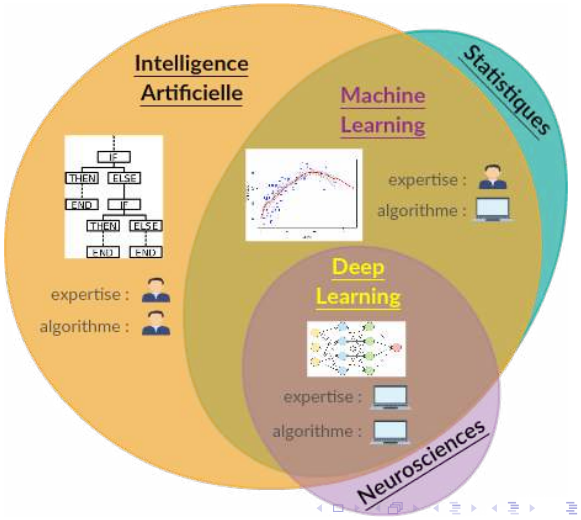
- Evaluation :
  TDs to be handed in (4 TDs each out of 5 points)
  A project : written report + oral presentation in January (out of 20)

$$\text{Final mark} = \frac{1}{3} \times \text{TDs' marks} \quad + \quad \frac{2}{3} \times \text{Project mark}$$

- Installing python packages on your machines : `scikit-learn`
- Supports : `https://www.i3s.unice.fr/~comet/SUPPORTS/`

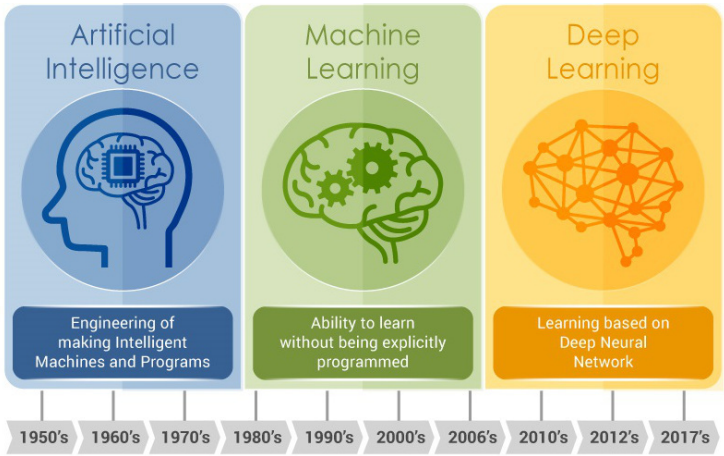
AI is a very broad field

« all the theories and techniques used to create machines capable of simulating human cognitive abilities »
« A set of methods and techniques for solving complex problems that would otherwise be the preserve of human intelligence. »



AI is a very broad field

« all the theories and techniques used to create machines capable of simulating human cognitive abilities »
« A set of methods and techniques for solving complex problems that would otherwise be the preserve of human intelligence. »



different areas of AI applications

- **Daily life :** camera, smartphone, fingerprint, GPS, etc.
- **Health :** obtain a precise diagnosis, techniques based on Deep learning (+ large quantity of data), monitoring of the treatment of certain pathologies in patients, expert systems for diagnosis, etc.
- **Banking/finance/insurance :** market forecasting, assessment of loan granting risks, adjusted calculation of insurance premiums, fraud detection, etc.
- **Industries :** logistics, supply, production, intelligent robots, monitoring, predicting breakdowns before they occur, ...
- **Security :** facial recognition, cyberattack prediction, speech Recognition, ...
- **Transport :** towards autonomous vehicles (taxis in circulation in the USA), piloting trains, planes, shuttles, detecting signs of fatigue on a driver's face, regulating traffic lights, etc.
- **Commerce, marketing, services :** inventory management, customer targeting, personalization of messages, behavioral analysis, recommendation (Netflix for example), etc.
- **Research :** finding unsuspected links in data, ...

Data Mining / Fouille de données

From Statistics... to Data mining

From Statistics...

- A few hundred individuals,
- Some variables,
- Strong assumptions about the statistical laws followed,
- Importance given to calculation,
- Random sample.

... to Data mining

- Millions of individuals,
- Hundreds of variables,
- Data collected without prior study,
- Need for rapid calculations,
- Not a random sample.

Information systems and data mining

- Data forms the core of basic processes in most companies.
- Data archiving creates corporate memory.
- Data mining creates business intelligence.

Definition of Data Mining : *Data mining, [...], is the exploration and analysis of large quantities of data in order to discover meaningful patterns and rules. [...] Many of the techniques discussed in this book started out in the fields of statistics, artificial intelligence, or machine learning.*
(Berry et Linoff, 1997)

L'exploration des données : une réalité industrielle

- Les techniques d'exploration des données existent depuis des années.
- L'utilisation de ces techniques dans l'industrie est cependant beaucoup plus récente :
  - Les données sont produites électroniquement,
  - Les données sont archivées,
  - La puissance de calcul nécessaire est abordable,
  - Le contexte est ultra-concurrentiel,
  - De nombreux algorithmes pour l'exploration des données ont émergés.

Data mining ou ...

- KDD (Knowledge Discovery in Databases)
- Fouille de données (terme français)
- Extraction automatique de connaissances à partir de données (ECD)
- Recherche d'Information (Information Retrieval)

But : Extraire de la connaissances dans de grands volumes de données :

- Extraction d'informations originales auparavant inconnues, potentiellement utiles
- découverte de nouvelles corrélations, tendances et modèles ▸ Spurious
- processus d'aide à la décision en cherchant des modèles d'interprétation des données

---

Data mining : « La famille »

- Statistiques, analyse des données
- Apprentissage automatique
- Reconnaissance de formes
- Bases de données
- Entrepôt de données (Data Warehouse)
- Visualisation des données

Domaine d'application du data mining (1)

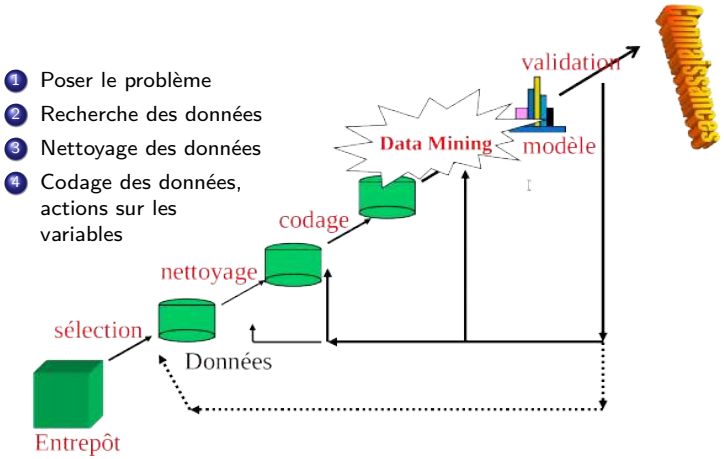
- Activités commerciales : grande distribution, vente par correspondance, banque, assurances
  - segmentation de la clientèle
  - détermination du profil du consommateur
  - analyse du panier de la ménagère
  - mise au point de stratégies de rétention de la clientèle
  - prédiction des ventes
  - détection des fraudes
  - identification de clients à risques
- Activités financières
  - recherche de corrélations entre les indicateurs financiers
  - maximiser le retour sur investissement de portefeuilles d'actions
- Activités de gestion des ressources humaines
  - prévision du plan de carrière
  - aide au recrutement

Domaine d'application du data mining (2)

- Activités industrielles
  - détection et diagnostic de pannes et de défauts
  - analyse des flux dans les réseaux de distribution
- Activités scientifiques
  - diagnostic médical, santé publique : ex, étude du génome
  - analyse chimique, biologique et pharmaceutique
  - exploration de données astronomiques
  - Recherche d'information dans les grands volumes de données multimédia

Le processus d'extraction de connaissances

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, d'information (Data mining)
6. Validation et interprétation du résultat, avec retour possible sur les étapes précédentes
7. Intégration des connaissances apprises

1- Poser le problème

- C'est comprendre
  - le domaine d'application,
  - la connaissance déjà existante et
  - les buts de l'utilisateur final.
- Quel type de problème a-t-on à traiter ?
  - on connaît les classes, on veut identifier les facteurs d'affectation, ou
  - on veut créer les classes facteurs de différentiation.
  - on veut trouver des tendances,
  - on veut exhiber les facteurs de risque ...
- Si on met en évidence de nombreux groupes de clients, dans une étude de marketing, pourra-t-on revoir les processus marketing pour chaque groupe ?

1. Poser le problème : un exemple

- Un éditeur vend 5 sortes de magazines : sport, voiture, maison, musique et BD. Les questions qu'il se pose :
  1. Combien de personnes ont pris un abonnement à un magazine de sport cette année ?
  2. A-t-on vendu plus d'abonnements sport cette année que l'année dernière ?
  3. Est-ce les acheteurs de magazines de BD sont aussi amateurs de sport ?
  4. Quelles sont les caractéristiques principales de mes lecteurs de magazines de voiture ?
  5. Peut-on prévoir les pertes de clients et prévoir des mesures pour les diminuer ?

  1 et 2 sont de simples requêtes.
  Dans 2 on a la notion de temps donc les données doivent être historisées.
  Pour 3, la réponse pourrait être une valeur estimant la proba que la règle soit vraie. 3 peut être généralisée : on peut chercher des associations fréquentes entre acheteurs de magazine.
  4 est plus ouverte, 5 aussi c'est vraiment le domaine de la fouille de données

2. Recherche des données

- Données existantes ou à constituer :
  Entrepôt de données (Data Warehouse), magasin de données, Bases de données relationnelles, Bases de données temporelles, Web,...
- Échantillon ou travail sur toutes les données :
  dépend des données disponibles, de la puissance machine, de la fiabilité souhaitée. Très souvent le travail sur un échantillon est bien adapté au data mining qui est un processus itératif.

Entrepôt de données (Data Warehouse) :
« Collection de données orientées pour un sujet, intégrées, non volatiles et historisées, organisées pour le support du processus d'aide à la décision ».
Base de données dans laquelle sont déposées après nettoyage et homogénéisation les informations en provenance des différents systèmes de production de l'entreprise.
L'entrepôt facilite le data mining mais le data mining peut se faire aussi sur des données extraites pour l'occasion.

Types de bases de données : plusieurs types de structure de BD :

- flat file : Toute l'information du client est contenue dans un même fichier qui peut être de longueur variable
- Relationnelle : L'information du client est contenu dans plusieurs fichiers unis par une « clef » commune, par exemple le numéro du client

3- Nettoyage des données

- Doublons, erreurs de saisie, pannes de capteurs...
- Valeurs aberrantes : rechercher les pics, les valeurs en dehors d'un espace déterminé par la moyenne et un certain nombre d'écarts-type, outils de visualisation : histogrammes, nuages de points, ...
  Ignorer l'observation / Utiliser la valeur moyenne (la pire ! !)/ Utiliser la valeur moyenne pour les exemples d'une même classe / Utiliser la régression (plus précise mais plus complexe)
  Stratégie pour traiter les valeurs aberrantes / Utilisation de ces valeurs :
  - si l'objectif est de prévoir les taux de fréquentation et les revenus de rencontres sportives, il faut certainement éliminer les chiffres de fréquentations anormales dues à des événements particuliers, grève des transports, etc...
  - dans le cas de la détection de fraudes, il peut être pertinent de se concentrer sur certaines valeurs aberrantes car elles sont peut-être la représentation de transactions frauduleuses.
- Informations manquantes : exclure les enregistrements incomplets, remplacer les données manquantes (valeur moyenne, valeur par défaut), garder les manquants si la méthode de fouille sait les gérer

4- Codage des données

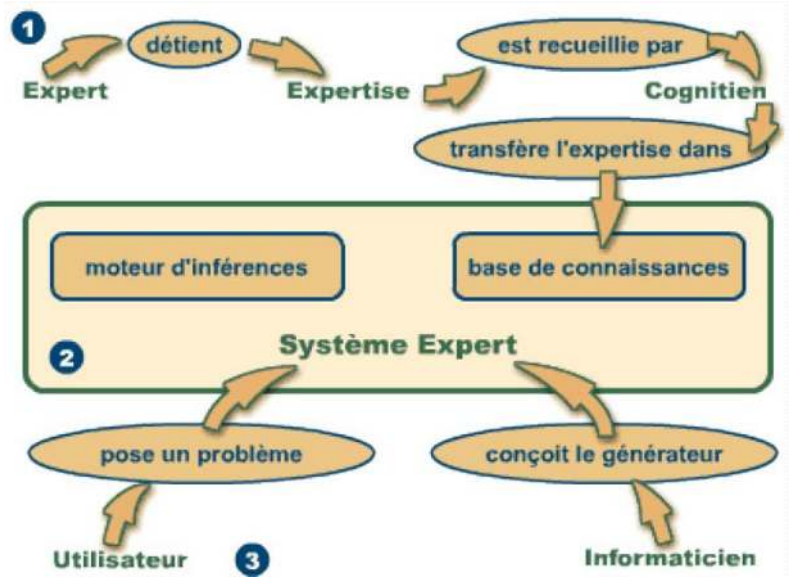
- Agrégation (somme, moyenne)
- Discrétisation (réduire le nombre de valeurs d'une variable continue en divisant le domaine de valeurs en intervalles)
- Codage des attributs discrets
- Uniformisation d'échelle ou standardisation
- Construction de nouvelles variables

⇒ Etape déterminante pour le processus de fouille.

Exemples d'actions sur les variables

- Transformation d'une variable :
  - Transformation des données géographiques : code postal → (longitude,latitude) permettant de prendre en compte la proximité des lieux (géocodage utilisé en géomarketing)
  - Transformation des dates en durées : ancienneté d'un client, durée entre l'envoi d'un catalogue et la 1ère comm.
- Transformation multi-variables :
  - combiner plusieurs variables en une nouvelle variable agrégée, combinaison linéaire ou non-linéaire de plusieurs variables : revenu et nombre d'enfants combinés par Revenu/nbre d'enfants

Systèmes experts



Data mining vs Systèmes Experts

- Déduction : base des systèmes experts
  - schéma logique permettant de déduire un théorème à partir d'axiomes
  - le résultat est sûr, mais la méthode nécessite la connaissance de règles
- Induction : base du data mining
  - méthode permettant de tirer des conclusions à partir d'une série de faits
  - généralisation « un peu abusive »
  - indicateurs de confiance permettant la pondération

Data mining vs Systèmes Experts

Une grande banque internationale voulait contrôler les coûts générés par ses clients lorsqu'ils utilisaient les distributeurs automatiques d'autres banques :

1. Qu'est-ce qui constitue une utilisation excessive par le client des distributeurs automatiques de la concurrence ?
2. Quels sont les clients qui génèrent des coûts excessifs par l'utilisation des distributeurs automatiques de la concurrence ?
3. Quelle est la valeur qu'ils représentent pour notre banque ?
4. A quoi devons-nous prêter attention lorsque nous utilisons ces résultats ?

Data mining vs Systèmes Experts

1. La première découverte faite par le laboratoire en Data Mining de Teradata (http ://www.teradata.com/) fut que 10% des clients de la banque généraient 90% des coûts des distributeurs automatiques concurrents. Cette constatation aurait pu être faite à l'aide de moyens traditionnels. Cependant, il a été mis en évidence que sur les 10% des clients qui généraient les coûts, 80% étaient des clients de faible valeur.
2. Plusieurs techniques de fouille de données permettant de prendre en compte de multiples variables ont permis de comprendre la valeur potentielle de chacun des clients de faible valeur et quantifier le concept d'« utilisation excessive ».
3. La fouille de données a permis de répondre à la question de la banque : Fallait-il revoir le service offert à ces 80% de clients de faible valeur ? L'analyse des tendances comportementales a permis de découvrir qu'environ 30% de ces clients de faible valeur étaient des clients à fort potentiel, à savoir : des étudiants. Pas surprenant mais l'étude aurait été difficilement réalisable sans les techniques d'analyse de multiples variables et les données détaillées.
4. La découverte la plus intéressante : un concurrent ciblait les campus universitaires (nouveaux distributeurs). Aucune autre banque ne développait d'actions sur les campus et celle-ci jouissait d'une présence quasiment exclusive.

⇒ Grâce à cette expérience en fouille de données, la banque put répondre à ses questions initiales. Mais ce qui est encore plus important, c'est qu'elle fut en mesure de découvrir la stratégie d un concurrent.

5- Data Mining

Des algorithmes d'inspirations ...

- Mathématiques : statistiques et Analyse de Données
- Calculatoires
  - « Clustering » / Arbres de décision / Règles d'association
  - Programmation dynamique
  - Machines à Vecteurs de Support (SVM)
- Biologiques
  - Réseaux de neurones
  - Algorithmes génétiques

Des tâches différentes :

- Non Supervisées - Apprentissage *a priori* en mode Découverte
  - « Clustering »
  - Algorithmes génétiques
  - Règles d'association
- Supervisées - Apprentissage *a posteriori* en mode Reconnaissance/Prédiction
  - Réseaux de neurones / Machines à Vecteurs de Support
  - Arbres de décision
  - Programmation dynamique

5- Data Mining : classification

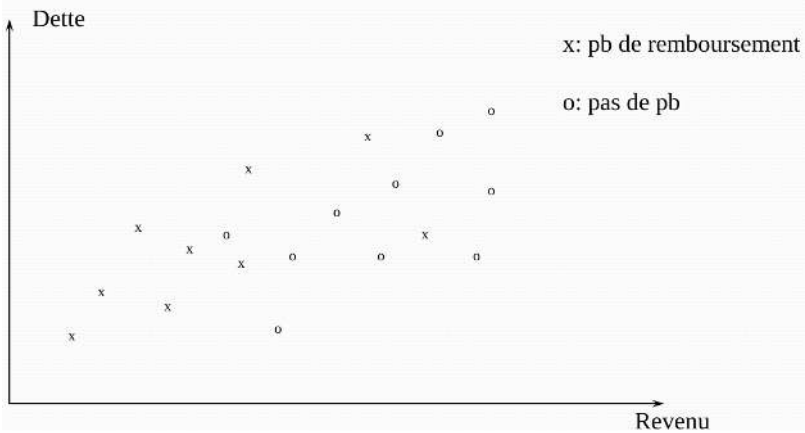
- Affecter un objet à une classe en fonction de ses caractéristiques $A_1, ...A_n$
- Exemple
  - Déterminer si un message est un mail de SPAM ou non (2 classes)
  - Affecter une page web dans une des catégories thématiques de l'annuaire Yahoo (multi-classes)
  - Diagnostic : risque d'accident cérébral ou non (2 classes)
- Si pas de connaissance *a priori* pour définir la classe en fonction de $A_1, ..., A_n$ alors on étudie un ensemble d'exemples pour lesquels on connaît $A_1, ..., A_n$ et la classe associée et on construit un modèle

$$Classe = f(A_1, ..., A_n)$$

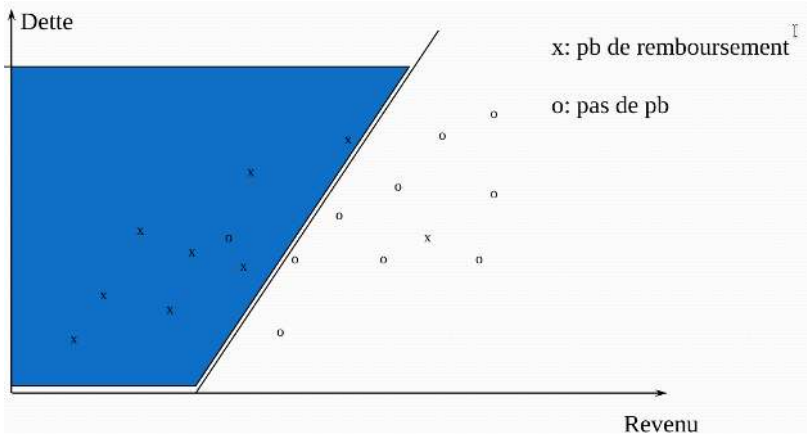
  - Analyse discriminante
  - Arbres de classification
  - Machines à noyaux

Un exemple schématique avec 2 classes

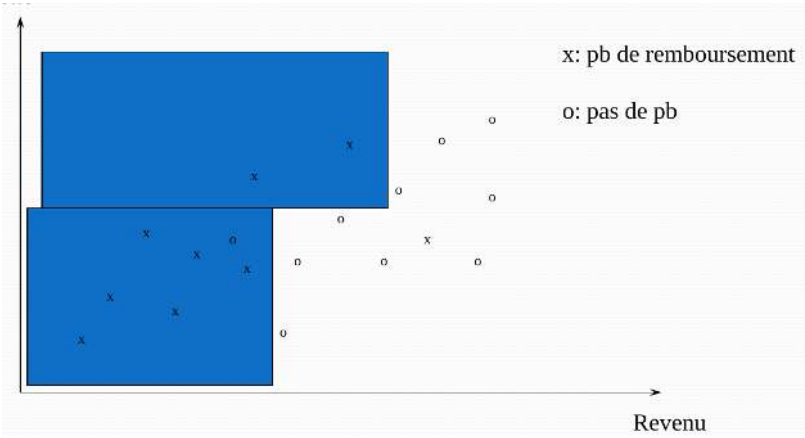
La classification : apprendre une fonction qui permet d'affecter un nouvel individu dans une classe ou une autre.



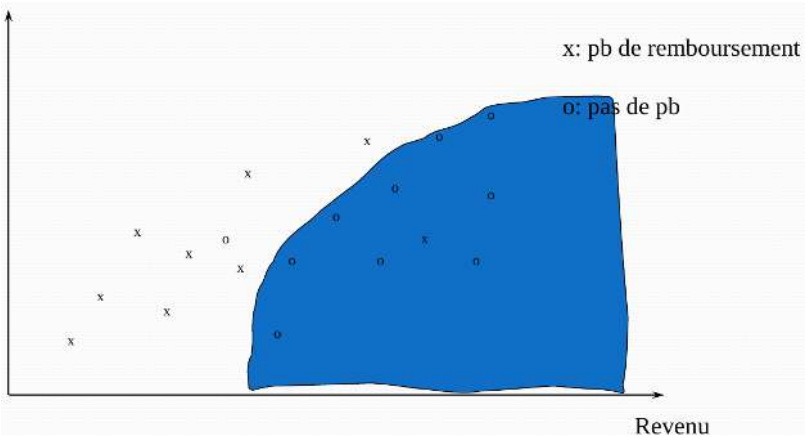
Classification par analyse discriminante



Classification par arbre de décision



Classification par machine à noyaux

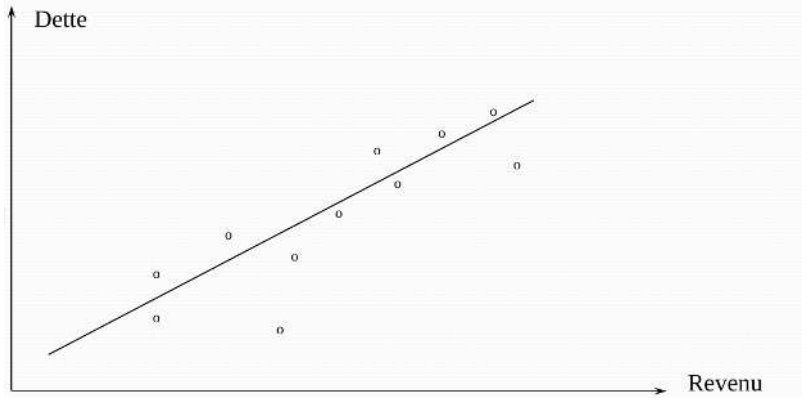


5- Data Mining : estimation

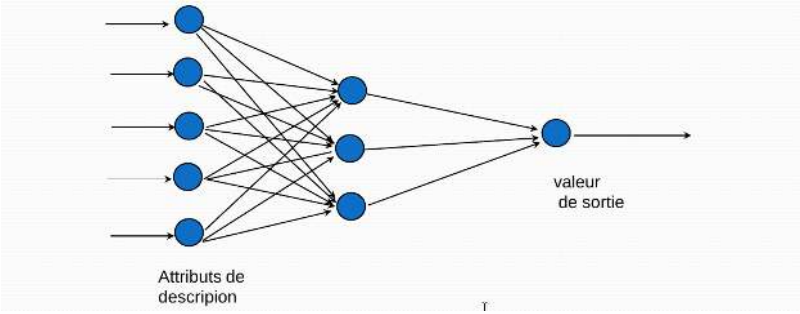
- Estimer (prédire) la valeur d'une variable à valeurs continues à partir des valeurs d'autres attributs
  - Régression
  - Machines à noyaux

Régression linéaire simple

La régression explique les variations d'une variable par une fonction des autres variables : ici la dette est représentée comme une fonction du revenu, le résultat est médiocre car il y a peu de corrélation.



Machines à noyaux : Réseau de neurones & SVM



- La couche d'entrées correspond aux entrées, la couche de sortie(s) au résultat
- Système non-linéaire
- L'apprentissage va ajuster les poids des connexions mais l'architecture et le nombre de neurones dans la couche cachée est un choix arbitraire.

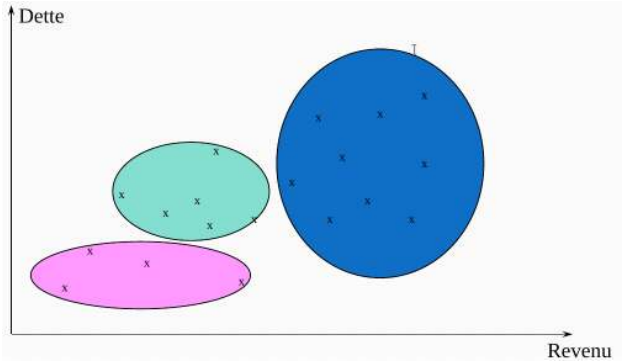
5- Data Mining : Recherche de règles

- Règles d'associations : analyse du panier de la ménagère
  - « 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' ?

5- Data Mining : Segmentation / partitionnement

- Apprentissage non supervisé : les données ne sont pas classées, on isole des sous-groupes d'enregistrements similaires (nuées dynamiques ou agrégation)
- Un fois les clusters détectés, on pourra appliquer des techniques de modélisation sur chaque cluster

Pas d'affectation à une classe connue au départ : regroupement des individus par leur proximité (les groupes peuvent se recouper).



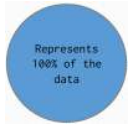
Supervisé vs NON supervisé

- La classification, la régression logistique sont des tâches supervisées
  - Data mining prédictif (on dispose d'une variable dépendante à prédire ou à estimer notée généralement par Y).
- Le clustering, la recherche de règles d'associations sont des tâches non supervisées
  - Data mining explicatif (on cherche plus à expliquer les relations entre les variables sans disposer d'une variable dépendante).

6. Validation dans le cas supervisé

- Validation par le test
  Données → (Ensemble d'apprentissage, Ensemble de test)
  - Construction d'un modèle sur l'ensemble d'apprentissage
  - test du modèle sur le jeu de test pour lequel les résultats sont connus
- Evaluation quantitative (ne pas oublier les intervalles de confiance)

Supervised Classification : Performance Evaluation

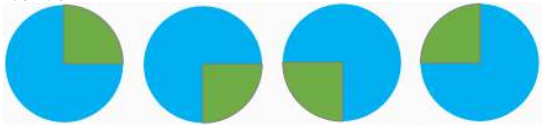


- Available data
  - Potential learning set
  - However, how do we check if the classifier performs well ?
    On the learning set ? Cannot be trusted (see later)

- Performance estimation using cross-validation
  - Choose number of partitions : say 4
  - Round 1
    - Split data into : learning set$_1$ and test set$_1$
    - Learn classifier$_1$ on learning set$_1$
    - Predict the classes of the test set 1 samples and compare with true classes
    - Compute pred_accuracy$_1$
  - Round 2
    - Split data into : learning set$_2$ and test set$_2$
    - Learn classifier$_2$ on learning set$_2$
    - Predict the classes of the test set 2 samples and compare with true classes
    - Compute pred_accuracy$_2$

Supervised Classification : Performance Evaluation

- The 4 rounds :



- Compute average prediction accuracy
  - avg_pred_accuracy = (pred_accuracy$_1$ + ... + pred_accuracy$_4$) / 4
  - Gives a good idea on how the classifier will perform on unseen data
  - If satisfied with this performance
    - Means that the learning procedure is appropriate
    - A final classifier can be learned on the whole available data
      To be used on newly acquired data
- If not enough samples
  - Leave-one-out cross-validation
    - Partition the n samples into :
      - $n-1$ samples for learning and
      - 1 sample for test
    - Repeat n times

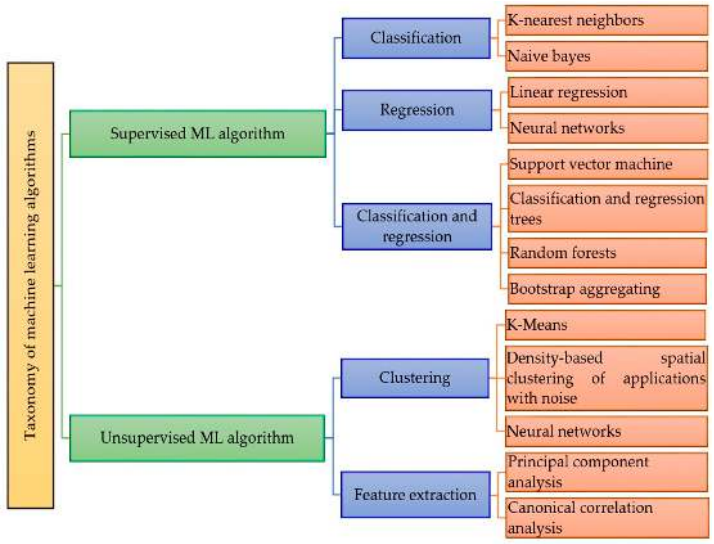
7. Intégration de la connaissance

- Prise de décision grâce aux connaissances extraites
- **Les experts métiers sont essentiels pour donner du sens aux informations extraites !**

Cinq Mythes sur le Data Mining

1. Le data mining pourrait instantanément prévenir l'avenir, à la manière d'une boule de cristal.
2. Le data mining ne serait pas encore viable pour des applications professionnelles.
3. Le data mining exigerait une base de données distincte et dédiée.
4. Il faudrait être polytechnicien pour faire du data mining.
5. Le data mining serait réservé aux grandes entreprises disposant d'un large volume de données client.

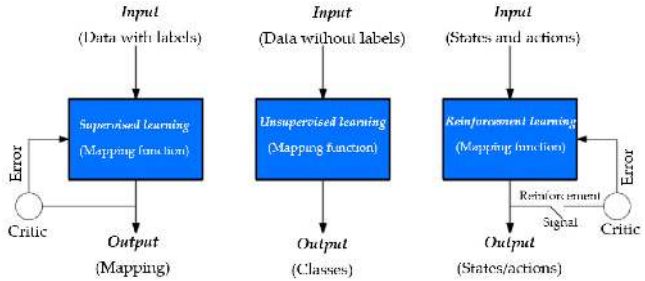
Les principales méthodes du ML



CÔTE D'AZUR

Les principales méthodes du ML

AI for bio
J-P Comet
Introduction
Poser le problème
Recherche des données
Nettoyage des données
Codage des données
Data Mining
Trees
Clustering
kNN
Bayes
SVM
NN
Evaluation
Data Mining Project



40/227