Introduction au Data-Mining

Alain Rakotomamonjy - Gilles Gasso

INSA Rouen -Département ASI
Laboratoire PSI

Data-Mining : Kèkecé ?

- Traduction : Fouille de données.
- Terme récent (1995) représentant un mélange d'idées et d'outils provenant de la Statistique, l'Intelligence Artificielle et l'Informatique.
- La définition exacte reste peu claire et les terminologies associées au Data-Mining sont encore floues.

Une définition suivant un critère égocentré :

Le data-mining est un processus de **découverte** de règle, relations, corrélations et/ou dépendances à travers une grande quantité de données, grâce à des méthodes statistiques, mathématiques et de reconnaissances de formes.

Autres définitions :

Le data-mining est un processus d'extractions **automatique** d'informations **predictives** à partir de grandes bases de données.

Data-Mining : les raisons du développement

pourquoi ca s'est développé ?

- Interêt économique : du produit aux clients.
- Technologie de l'information : faible coût de stockage de données, saisie automatique de transaction (code bar, click, données de localisation GPS, internet)
- Augmentation de la puissance de calculs des ordinateurs (loi de Moore)

$\Rightarrow$ Extraire de la connaissance à partir de grandes bases de données devient possible

Exemples d'applications

- Entreprise et Relation Clients : système de création de profils clients, ciblage de clients potentiels et nouveaux marchés
- Finances : minimisation de risque financiers
- Bioinformatique : Analyse du génome, mise au point de médicaments, ...
- Internet : spam, e-commerce, détection d'intrusion etc...
- Sécurité

Exemples d'applications : E-commerce

Dell

- **Problème** : 50% des clients de Dell achètent leurs machines à travers le site Web. Mais seulement 0.5% des visiteurs du site deviennent clients.

- **Solution** : Stocker les séquences de clicks des visiteurs, analyser les caractéristiques des acheteurs et lors de la visite d'un client potentiel, adapter le contenu du site pour maximiser la probabilité d'un achat.

Amazon

- **Opportunité** : la liste des achats des clients est stockée en mémoire et par ailleurs, les utilisateurs du site notent les produits ! Comment tirer profit des choix d'un utilisateur pour proposer des produits à un autre client ?

- **Solutions** : technique dit de filtrage collaboratif permettant de regrouper des clients ayant les mêmes "goûts"

Exemples d'applications : Analyse des risques

Détection de fraudes pour les assurances

- Analyse des déclarations des assurés par un expert afin d'identifier les cas de fraudes.
- Extraction de caractéristiques à partir de ces déclarations (type d'accident, de blessures, etc...)
- Applications de méthodes statistiques pour identifier les caractéristiques des déclarations fortement corrélées à la fraude.

Prêt Bancaire

- **Objectif des banques** : réduire le risque des prêts bancaires.
- Créer un modèle à partir de caractérisques des clients pour discriminer les clients à risque des autres.

Exemples d'applications : Commerce

Organisation de rayonnage

- **Objectifs** : Identifier les produits que les gens sont susceptibles d'acheter conjointement afin d'organiser les rayonnages

- **Données** : Code-Barre des produits. **Méthodes :** Extractions de règles

- Exemples :
  - résultats logiques : les boissons alcoolisées et les biscuits apéritifs sont souvent proches.
  - resultats étranges : dans une étude américaine, la vente de bière est plus importante si le rayon des couches n'est pas trop loin, et si sur le chemin il y a des chips, cela permet d'augmenter la vente des 3 produits.

Mise en oeuvre d'un projet d'un projet de DM

1. Comprendre et analyser les objectifs de l'application
2. Créer une base de données pour la mise au point de l'application.
3. Prétraitement et nettoyage des données
4. Analyse statistique des données (réduction de la dimension, projection, etc...)
5. Identifier le type de problèmes (discrimination, clustering, etc...) et choisir un algorithme.
6. Evaluer les performances de l'algorithme.
7. Réitérer les étapes précédentes si necessaire.
8. Déployer l'application.

Objectifs du cours : Etude des méthodes pour les étapes 4 à 6

Caractérisation des méthodes de Data-Mining

Types d'apprentissage

- Apprentissage supervisé
- Apprentissage non-supervisé
- Apprentissage semi-supervisé

Caractérisation des méthodes de Data-Mining

Apprentissage supervisé

- Objectifs : à partir d'un ensemble d'observations $\{x_1, \cdots, x_n\} \in \mathcal{X}^d$ et de mesures $\{y_i\} \in \mathcal{Y}$, on cherche à estimer les dépendances entre l'ensemble $\mathcal{X}$ et $\mathcal{Y}$.

  Exemple : on cherche à estimer les liens entre les habitudes alimentaires et le risque d'infarctus. x_i est un patient décrit par d caractéristiques concernant son régime et y_i une catégorie (risque, pas risque).

  On parle d'apprentissage *supervisé* car les y_i permettent de guider le processus d'estimation

- Exemples de méthodes : Méthode du plus proche voisin, réseaux de neurones, Séparateurs à Vastes Marges, CART etc..

- Exemples d'applications : détection de fraude, marketing téléphonique, changement d'opérateurs téléphonique etc...

Caractérisation des méthodes de Data-Mining

Apprentissage non-supervisé

- Objectifs : Comme seules les observations $\{x_1, \cdots, x_n\} \in \mathcal{X}^d$ sont disponibles, l'objectif est de décrire comment les données sont organisées et d'en extraire des sous-ensemble homogènes.

  Exemple : On cherche à étudier le panier de la mènagère dans une certaine zone démographique en fonction de certains critères sociaux. x représente un individu à travers ses caractéristiques sociales et ses habitudes lors des courses

- Exemples de méthodes : Classification hierarchique, Carte de Kohonen, K-means, extractions de règles...

- Exemples d'applications : identification de segments de marchés, identification de document similaires,

Caractérisation des méthodes de Data-Mining

Apprentissage semi-supervisé

- Objectifs : parmi les observations $\{x_1, \cdots, x_n\} \in \mathcal{X}^d$, seulement un petit nombre d'entre elles ont un label $\{y_i\}$. L'objectif est le même que pour l'apprentissage supervisé mais on aimerait tirer profit des observations non labelisées.

- Exemple : pour la discrimination de pages Web, le nombre d'exemples peut être très grand mais leur associer un label est coûteux.

- Exemples de méthodes : méthodes bayesiennes, Séparateur à Vastes Marges, etc...

Ensemble de données

Dans un problème de Data-Mining, les informations caractérisant une étude (un client pour un problème de e-commerce ou un déclaration dans le cas d'un détection de fraudes) sont présentées sous la forme d'attributs et d'exemples.

Attributs

- Un attribut est un descripteur d'une entité. On l'appele également variable, champs, caratéristiques ou observations

Exemple

- Un exemple est une entité caractérisant un objet et est donc constitué d'attributs.
- synonymes : point, vecteur (surtout si les exemples sont dans $\mathbb{R}^n$)

Type de données

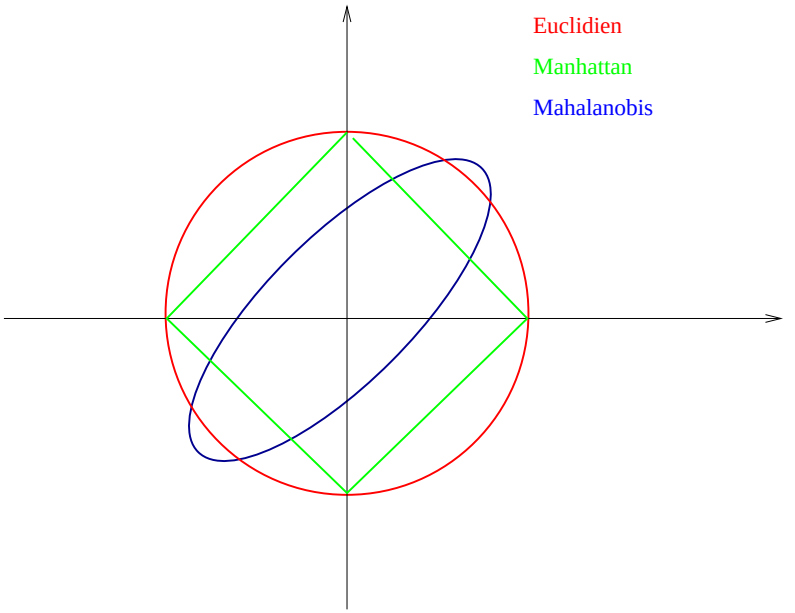
Types

- numérique continue : la valeur de la variable peut prendre une valeur dans $\mathbb{R}$ (par exemple : le montant du compte en banques de B. Spears).
- numérique discrète : la valeur de la variable appartient à $\mathbb{Z}$ ou $\mathbb{N}$ (par exemple : l'age du capitaine)
- catégorie : avec ou sans relation d'ordre (par exemple : { rouge, vert, bleu }).
- binaire
- Chaînes de caractères (par exemple : un texte)
- Arbre : (par exemple Page XML)
- Données structurées : graphe, enregistrement

Données et Métriques

Que ce soit dans le cadre d'un problème d'apprentissage supervisé ou non supervisé, la plupart des algorithmes nécessite une notion de similarité dans l'espace $\mathcal{X}$ des données. La similarité est traduite par la notion de distance.

- distance euclidienne : $x, x' \in \mathbb{R}^d$, on a $d(x, x^{'}) = \|x - x^{'}\|_{\ell_2} = \sqrt{\sum_{i=1}^{d}(x_i - x_i^{'})^2} = \sqrt{(x - x^{'})^t(x - x^{'})}$

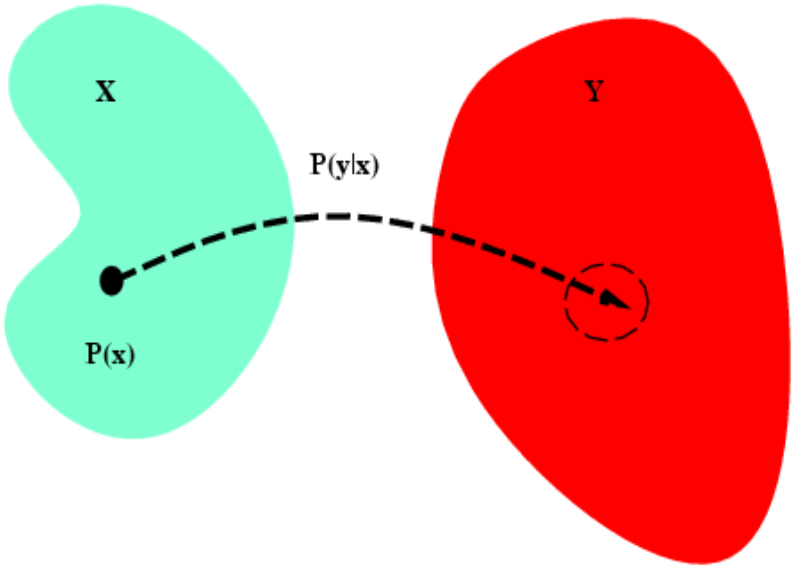


- distance de manhattan $d(x, x^{'}) = \|x - x^{'}\|_{\ell_1} = \sum_{i=1}^{d}\left|(x_i - x_i^{'})\right|$
- distance de mahalanobis $d(x, x^{'}) = \sqrt{(x - x^{'})^t\Sigma^{-1}(x - x^{'})}$

La notion de similarité peut être étendue à d'autres types de données.

Apprentissage supervisé : les concepts

- Supposons que l'on a deux ensembles $\mathcal{X}$ et $\mathcal{Y}$ munis d'une loi de probabilité jointe $p(X, Y)$.



- On cherche une fonction $f : \mathcal{X} \rightarrow \mathcal{Y}$ qui a X associe $f(X)$ qui permet d'estimer la valeur y associée à x. f appartient à un espace $\mathcal{H}$ appelé **espace d'hypothèses**.

Apprentissage supervisé : les concepts

- On introduit une notion de **coût** $L(Y, f(X))$ qui permet d'évaluer la pertinence de la prédiction de f, et de pénaliser les erreurs.

- L'objectif est donc de choisir f telle que f minimise

$$R(f) = E_{X,Y}[L(Y, f(X))]$$

où R est appelé le **risque moyen** ou **erreur de généralisation**. Il est également noté $EPE(f)$ pour **expected prediction error**

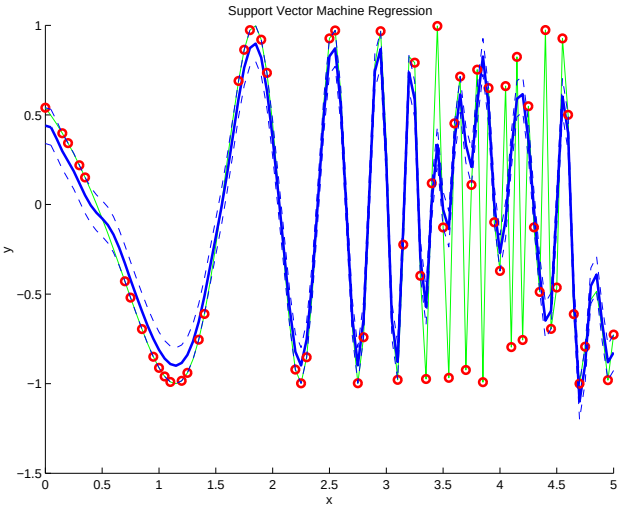
- Exemple de fonction coût et de risque moyen associé.

$$L(Y, f(X)) = (Y - f(X))^2 \quad R(f) = E[(Y - f(X))^2] = \int (y - f(x))^2 p(x, y) dxdy$$
$$L(Y, f(X)) = |Y - f(X)| \quad R(f) = E[|Y - f(X))|] = \int |y - f(x)| p(x, y) dxdy$$

Apprentissage supervisé : les concepts (2)

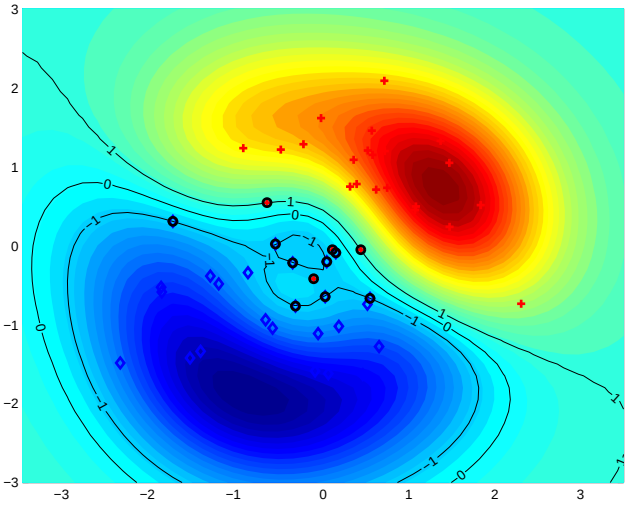
Regression

- On parle de **régression** quand $\mathcal{Y}$ est un sous-espace de $\mathbb{R}^d$.
- Les fonctions de coût typique sont $(y - f(x))^2$ et $|y - f(x)|$



Discrimination

- si $\mathcal{Y}$ est un ensemble discret non-ordonné, (par exemple $\{-1, 1\}$), on parle de **discrimination**.
- La fonction de coût la plus usitée est : $\Theta(-yf(x))$ où Θ est la fonction échelon.



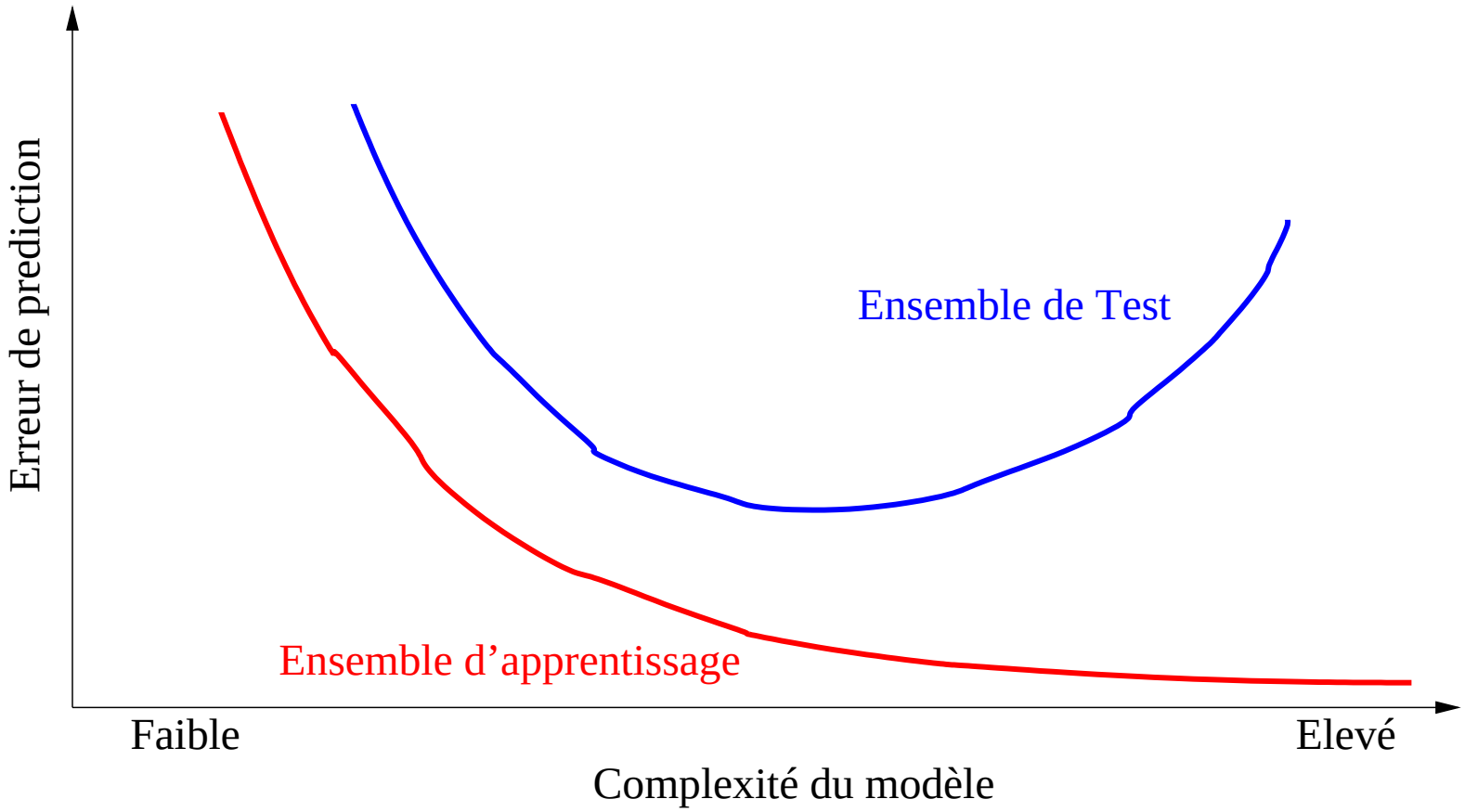
Apprentissage supervisé : les concepts (3)

- En pratique, on a un ensemble de données appelé **ensemble d'apprentissage** obtenu par échantillonage indépendant de $p(X, Y)$ que l'on ne connaît pas.
- On cherche une fonction f, appartenant à $\mathcal{H}$ qui minimise le **risque empirique** :

$$R_{emp}(f) = \frac{1}{\ell} \sum_{i=1}^{\ell} L(y_i, f(x_i))$$

- Le risque empirique ne permet pas d'évaluer la pertinence d'un modèle car il est possible de choisir f de sorte que le risque empirique soit nul mais que l'erreur en généralisation soit élevée. On parle alors de **sur-apprentissage**

Illustration du sur-apprentissage



Sélection de modèles

Problèmatique :

- On cherche une fonction f qui minimise un risque empirique donné. On suppose que f appartient à une classe de fonctions paramétrées par α. Comment choisir α pour que f minimise le risque empirique et généralise bien ?
- Exemple : On cherche un polynôme de degré α qui minimise un risque $R_{emp}(f_\alpha) = \sum_{i=1}^{\ell}(y_i - f_\alpha(x_i))^2$.
- Objectifs :
  1. proposer une méthode d'estimation d'un modèle afin de choisir (approximativement) le meilleur modèle appartenant à l'espace hypothèses.
  2. une fois le modèle choisi, calculer son erreur de généralisation.

Sélection de modèles

Cas idéal :

- On est dans un cas où les données abondent.
- Dans ce cas, on sépare les données en 3 ensembles : données d'apprentissage, données de validation et données de test. Le premier sert à construire un modèle, le deuxième à estimer l'erreur de ce modèle. Le troisième ne sert qu'**une fois** : à estimer l'erreur en généralisation du modèle final.

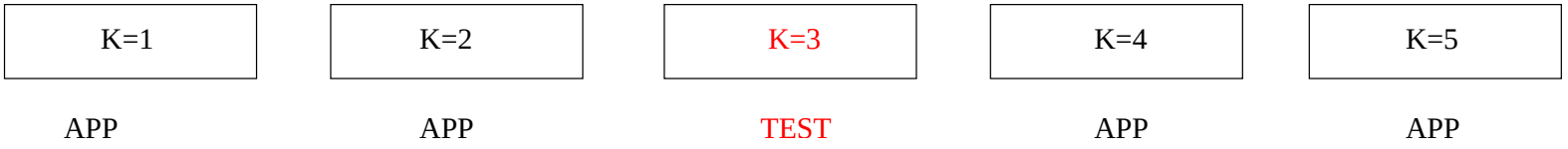
Cas usuel :

- On est pauvre en données.
- Utilisation de méthodes analytiques (AIC, BIC, etc ...) ou de rééchantillonage pour remplacer l'étape de validation.

Sélection de modèles : Validation Croisée

- Méthode d'estimation de l'erreur en généralisation d'une fonction f par rééchantillonage.

- Principe
  1. Séparer les données en K ensembles de part égales.
  2. Pour chaque K, apprendre un modèle en utilisant les $K-1$ autres ensemble de données et évaluer le modèle sur la K-ième partie.
  3. Moyenner les K estimations de l'erreur obtenues pour avoir l'erreur de validation croisée.



Sélection de modèles : Validation Croisée (2)

- Détails :

$$CV = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{N_k} \sum_{i=1}^{N_k} L(y_i^k, f^{-k}(x_i^k))$$

où f^{-k} est le modèle f appris sur l'ensemble des données sauf la k-ième partie.

- Propriétés : Si $K = \ell$, CV est approximativement un estimateur sans biais de l'erreur en généralisation. L'inconvénient est qu'il faut apprendre $\ell - 1$ modèle.

- typiquement, on choisit $K = 5$ ou $K = 10$ pour un bon compromis entre le biais et la variance de l'estimateur.

Conclusions

Pour bien mener un projet de DM

- Identifier et énoncer clairement les besoins.
- Créer ou obtenir des données représentatives du problème
- Identifier le contexte de l'apprentissage
- Analyser et réduire la dimension des données
- Choisir un algorithme et/ou un espace d'hypothèses.
- Choisir un modèle en appliquant l'algorithme aux données prétraitées.
- Valider les performances de la méthode.