INTRODUCTION AU MACHINE LEARNING

DAKILAB

WWW.DAKILAB.FR

19/08/2020

DAKILAB

INTRODUCTION AU MACHINE LEARNING

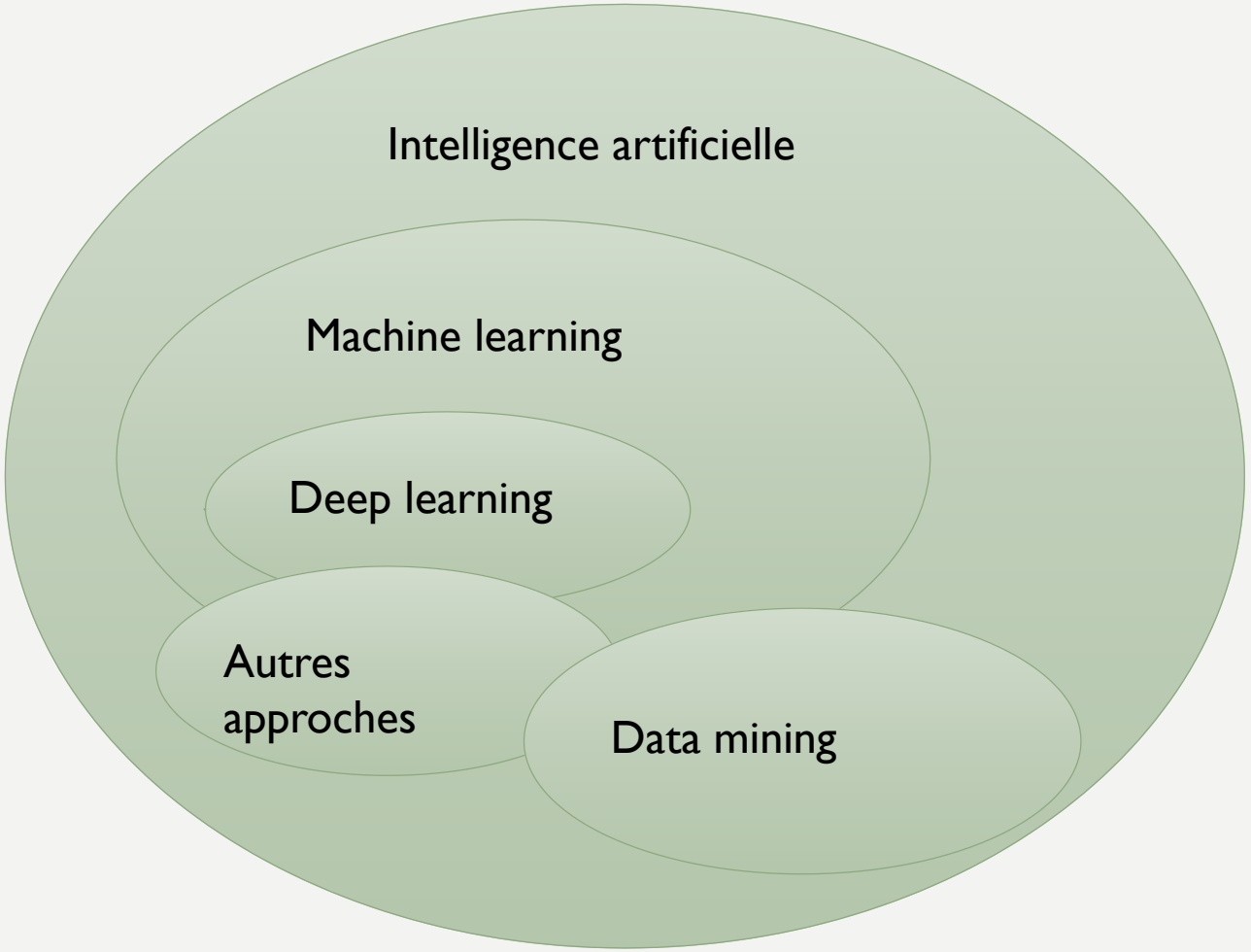
MOURAD NACHAOUI

Plan du chapitre

Introduction à l'apprentissage automatique

- L'apprentissage automatique
- Quelques domaines d'application
- Quelques méthodes simples en guise d'illustration
- Différentes caractéristiques des méthodes d'apprentissage supervisé
- (Quelques) Problèmes théoriques en apprentissage automatique
- Evaluation et comparaison de modèles en apprentissage supervisé

Structuration du Machine learning



Machine Learning

Motivation

On veut répondre automatiquement à des questions comme :

- le patient aura-t-il un accident cardio-vasculaire ?
- la molécule que je désire commercialiser est-elle cancérigène ?
- qui est l'auteur de cette page HTML ?
- cette phrase est-elle grammaticalement correcte ?
- quelle sera la taille de cet enfant à l'âge adulte ?

Ne pas écrire des programmes qui répondent à ces questions... mais les découvrir automatiquement, par apprentissage (observation d'exemples et de contre-exemples) en vue de prédire (classer de nouveaux exemples).

Machine Learning

Aujourd'hui : la motivation de la découverte

- INDANA [Colombet, 2002] :
  - prédiction du risque cardio-vasculaire après un examen minimal ;
  - des économies réalisées...
- Skicat [Fayyad, 1995] :
  - En astronomie, quel secteur du ciel regarder ? plusieurs téraoctets de données ;
  - 40 fois plus d'objets découverts par nuit d'observation, dépasse l'humain sur les objets faiblement lumineux ;
- molécules cancérigènes [Srinivasan et al., 1994] :
  - décider si un produit peut être diffusé, expérimentations de plusieurs années sur des animaux ;
  - bonnes performances, au croisement de plusieurs disciplines.
- Plus de 5 590 000 de publications indexées sur Google scholar

Les points à définir

1. Ce que l'on veut prédire ;
2. Modalités d'obtention des exemples ;
3. Nature des exemples, nature des classiffieurs ;
4. Mode d'évaluation des prédictions ;
5. Méthode d'apprentissage.

Machine Learning

Qu'est-ce que l'apprentissage automatique?

- De manière générale, un programme informatique tente de résoudre un problème pour lequel nous avons la solution. Par exemple : calculer la moyenne générale des étudiants, classer les étudiants selon leur moyenne…
- Pour certains problèmes, nous ne connaissons pas de solution exacte et donc nous ne pouvons pas écrire de programme informatique. Par exemple : reconnaître automatiquement des chiffres écrits à la main à partir d'une image scannée, déterminer automatiquement une typologie des clients d'une banque, jouer automatiquement aux échecs contre un humain ou un autre programme…
- En revanche, pour ces problèmes il est facile d'avoir une base de données regroupant de nombreuses instances du problème considéré.

L'apprentissage automatique consiste alors à programmer des algorithmes permettant d'apprendre automatiquement de données et d'expériences passées, un algorithme cherchant à résoudre au mieux un problème considéré.

Machine Learning

Qu'est-ce que l'apprentissage automatique?

Selon Wikipedia (septembre 2020) :
« L'apprentissage automatique (en anglais machine learning*, littéralement « apprentissage machine ») ou apprentissage statistique est un champ d'étude de l'intelligence artificielle qui se fonde sur des approches mathématiques et statistiques pour donner aux ordinateurs la capacité d'« apprendre » à partir de données, c'est-à-dire d'améliorer leurs performances à résoudre des tâches sans être explicitement programmés pour chacune. Plus largement, il concerne la conception, l'analyse, l'optimisation, le développement et l'implémentation de telles méthodes. »*

L'objectif du cours est de donner un sens à cette définition. Que signifie « apprendre » à partir de données, ou « ne pas être explicitement programmé » pour résoudre une tâche?

Machine Learning

Exemples introductifs

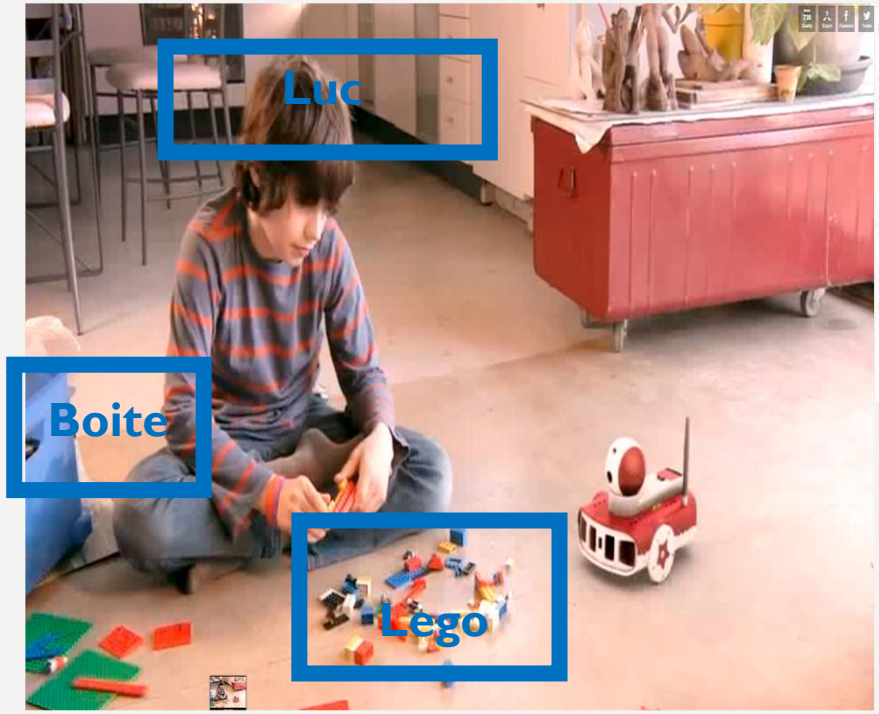
Exemple 1: Supposons que l'on dispose d'une collection d'articles de journaux. Comment identifier des groupes d'articles portant sur un même sujet ?

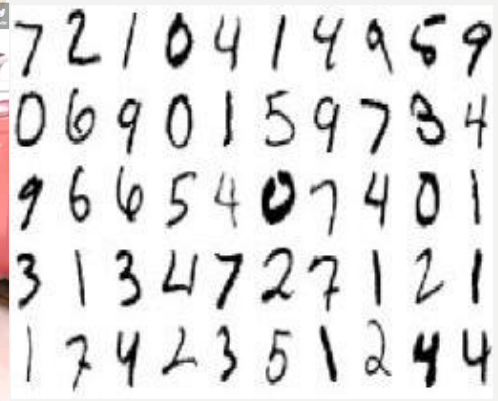
Exemple 2 : Supposons que l'on dispose d'un certain nombre d'images représentant des chiens, et d'autres représentant des chats. Comment classer automatiquement une nouvelle image dans une des catégories « chien » ou « chat » ?

Exemple 3 : Supposons que l'on dispose d'une base de données regroupant les caractéristiques de logements dans une ville : superficie, quartier, étage, prix, année de construction, nombre d'occupants, montant des frais de chauffage. Comment prédire la facture de chauffage à partir des autres caractéristiques pour un logement qui n'appartiendrait pas à cette base ?

Machine Learning

Exemples introductifs





Reconnaissance d'objets à partir d'images :

- Chiffres, lettres, logos
- Biométrie : visages, empreintes, iris,…
- Classes d'objets (voitures, arbres, bateaux, …)
- Objets spécifiques (une voiture, un mug, …)

Machine Learning

Un domaine pluridisciplinaire

- L'apprentissage automatique (AA) (« **Machine Learning** ») est à la croisée de plusieurs disciplines :
  - **Les statistiques** : pour l'inférence de modèles à partir de données.
  - **Les probabilités** : pour modéliser l'aspect aléatoire inhérent aux données et au problème d'apprentissage.
  - **L'intelligence artificielle** : pour étudier les tâches simples de reconnaissance de formes que font les humains (comme la reconnaissance de chiffres par exemple), et parce qu'elle fonde une branche de l'AA dite symbolique qui repose sur la logique et la représentation des connaissances.
  - **L'optimisation** : pour optimiser un critère de performance, soit estimer les paramètres d'un modèle, soit déterminer la meilleure décision à prendre étant donné une instance d'un problème.
  - **L'informatique** : puisqu'il s'agit de programmer des algorithmes et qu'en AA ceux-ci peuvent être de grande complexité et gourmands en termes de ressources de calcul et de mémoire.

Machine Learning

AA et matières connexes

- Quelques références et domaines d'application faisant intervenir l'AA :
  - **Les statistiques** (« Statistique Machine Learning ») : modèles d'AA traités sous l'angle des statistiques [Hastie et al., 2011, Dreyfus, 2008].
  - **L'intelligence artificielle** (« Artificial Intelligence ») : modèles d'AA mettant l'accent sur le raisonnement, l'inférence et la représentation des connaissances [Cornuejols and Miclet, 2003, Mitchell, 1997, Alpaydin, 2010].
  - **La fouille de données** (« Data Mining ») : lorsque les objets étudiés sont stockés dans des bases de données volumineuses [Han and Kamber, 2006].
  - **La reconnaissance de formes** (« Pattern Recognition ») : lorsque les objets concernés sont de type « signal » comme les images, les vidéos ou le son [Bishop, 2006].
  - **Le traitement automatique du langage** - TAL (« Natural Langage Processing » - NLP) : lorsque les problèmes concernent l'analyse linguistique de textes [Manning and Schütze, 1999, Clark et al., 2010].

Machine Learning

AA et matières connexes (suite)

- Plus récemment :
  - **La science des données** (« Data science ») : approche(s) pluridisciplinaire(s) pour l'extraction de connaissances à partir de données hétérogènes [Cleveland, 2001, Abiteboul et al., 2014].
  - **Les données massives** (« Big data ») : mettant l'accent sur les problématiques « 4V » (volume, variété, vélocité, véracité) et des éléments de solutions issus du stockage/calcul distribué [Leskovec et al., 2014].
  - Pour plus de ressources, consultez le site

  http://www.kdnuggets.com.

Machine Learning

Plusieurs types de problèmes en AA

- **apprentissage supervisé** : les exemples fournis sont sous la forme de couples entrée-sortie
- (x_i, y_i) avec x_i l'entrée et y_i la sortie. L'objectif est d'inférer la sortie *y* pour une nouvelle entrée x. Si $y_i \in \{-1,1\}$ (voire plus généralement $y_i \in \mathrm{N}$) on parlera de classification, si $y_i \in \mathrm{R}$ on parlera de régression. On dit que qu'il faut apprendre un modèle capable de prédire la bonne valeur cible d'un objet nouveau.
- **apprentissage non-supervisé** : les exemples fournis ne sont que des entrées x_i. L'objectif

est alors de résumer l'espace des x_i possibles (ce qui regroupe notamment l'estimation de densité, la quantification vectorielle, ou encore comment diviser un groupe hétérogène de données en sous-groupes homogènes) ;

Plusieurs types de problèmes en AA

- **apprentissage par renforcement** : les exemples fournis sont sous la forme de transitions (s_i, a_i, r_i, s_{i+1}) où s dénote l'état d'un système dynamique, a une action que l'on peut lui appliquer et r une récompense. L'objectif est celui du contrôle optimal, plus particulièrement dans ce paradigme d'inférer quelle action appliquer pour une configuration donnée du système dynamique à contrôler, ce de façon à maximiser un gain futur dont la récompense est une information locale.
- L'**apprentissage semi-supervisé** est une classe de techniques d'apprentissage automatique qui utilise un ensemble de données étiquetées et non étiquetées. Il se situe ainsi entre l'apprentissage supervisé qui n'utilise que des données étiquetées et l'apprentissage non supervisé qui n'utilise que des données non étiquetées. Il a été démontré que l'utilisation de données non étiquetées, en combinaison avec des données étiquetées, permet d'améliorer significativement la qualité de l'apprentissage.

Les données

Comme le suggère la définition proposée par Wikipedia, les algorithmes de l'apprentissage automatique sont basés sur des données. On parle aussi d'échantillons (*samples*), d'observations, ou d'exemples. Concrètement, cela signifie que le jeu de données (*dataset*) est formé d'un certain nombre d'articles de journaux (exemple 1) ou d'images de chiens et chats (exemple 2). Nous noterons chaque observation x_n et la base faite de N observations $(x_n)_{1 \leq n \leq N}$.

- Deux grandes familles de jeux de données peuvent être utilisées :
- les données étiquetées : chaque observation x_n est fournie avec une étiquette (*label*) y_n ;
- les données non-étiquetées : comme le nom l'indique, aucune étiquette n'est fournie.

Machine Learning

Les données

Dans l'exemple 1, les données ne sont pas étiquetées (chaque x_n représente un article de journal), alors qu'elles le sont dans l'exemple 2 (x_n représente une image, et y_n ={chien} ou y_n = {chat}) ou dans l'exemple 3 (x_n représente les informations superficie, quartier, étage, prix, année de construction, nombre d'habitants, et y_n est le montant des frais de chauffage). Il est généralement plus facile de constituer un jeu de données non étiquetées qu'un jeu de données étiquetées.

Dans le premier cas, il « suffit » de collecter des données après prétraitement automatique minimal, alors que dans le second cas une intervention humaine potentiellement coûteuse est souvent nécessaire pour définir les étiquettes.

Machine Learning

Les données

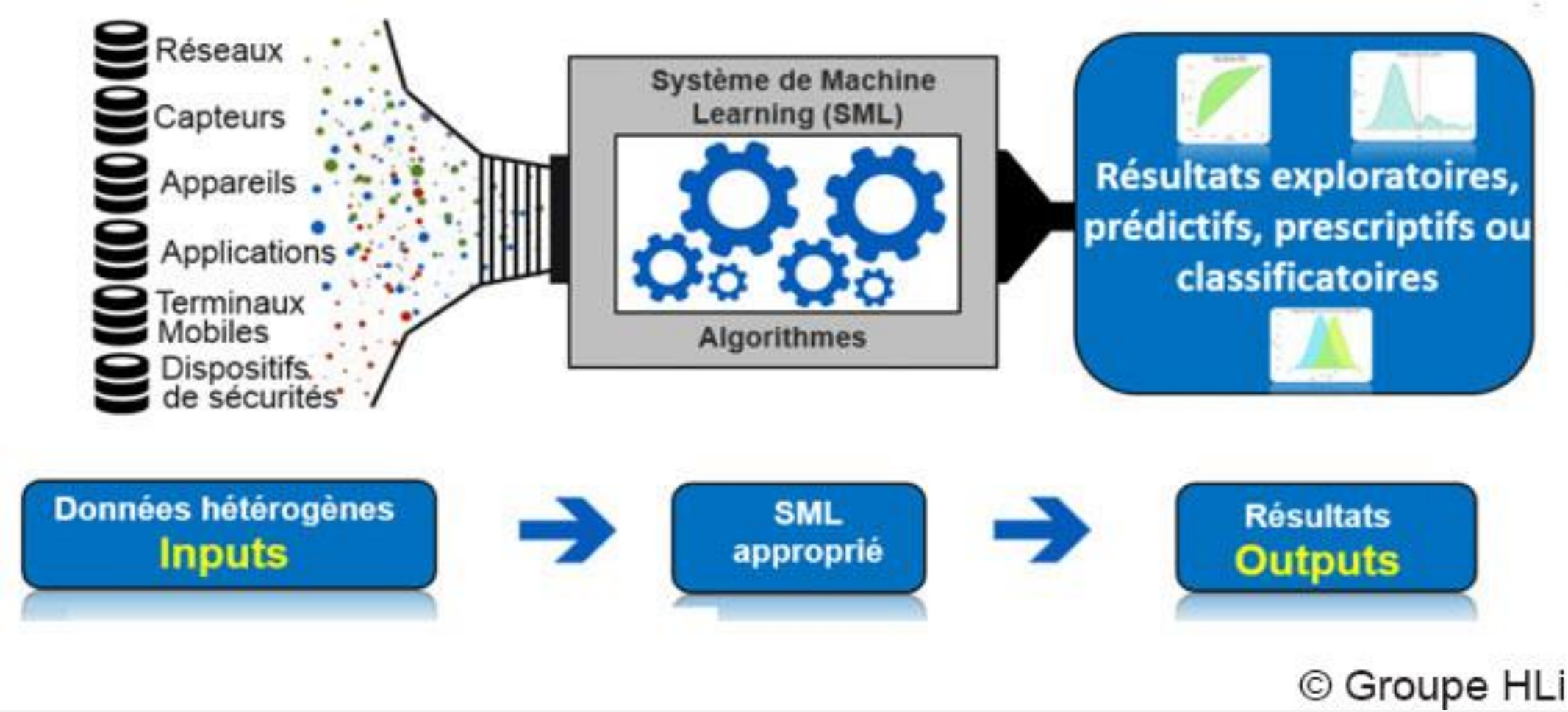
Dans l'exemple 2, les caractéristiques sont le niveau de gris en chaque pixel (pour simplifier, on suppose qu'il s'agit d'images noir et blanc), les observations sont les images, et la dimension *d* peut valoir un million s'il s'agit d'images de taille réaliste, disons 1000x1000 pixels.

Dans l'exemple 3, chaque observation est composée de 6 caractéristiques, donc *d* =6. Néanmoins, dans l'exemple 1, les observations ne sont pas naturellement éléments d'un espace vectoriel, et on peut se demander comment définir une distance entre documents telle que des documents portant sur le même sujet soient proches au sens de cette distance.

On peut noter que le nombre d'observations disponibles peut grandement varier : on pourra sans doute extraire un grand nombre d'images de chats et chiens de base de photographies comme Flickr (à la date de rédaction du polycopié, 1 101 601 photographies de Flickr portent le label « chien » et 1 991 665 « chat »), mais il est peu probable de disposer d'un aussi grand nombre de données relatives aux appartements. Comme on l'a vu précédemment, la dimension *d* peut également fortement varier.

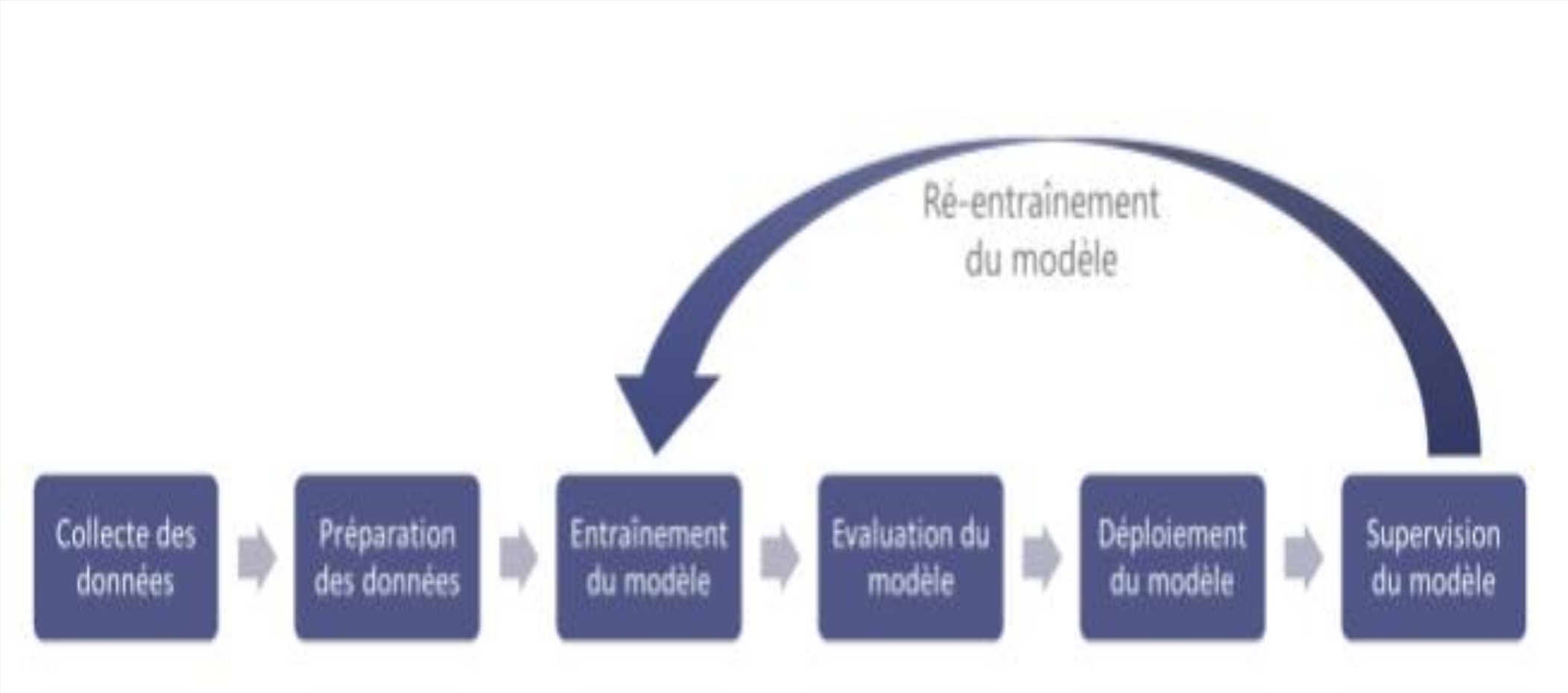
Machine Learning

Schéma général d'apprentissage automatique



Machine Learning

Schéma de fonctionnement d'apprentissage automatique



Machine Learning

Quelle est l'approche machine learning ?

1. Analyser les données,
2. Choisir un modèle,
3. Les modèles sont entraînés avec des données (data mining),
4. Estimer l'erreur du modèle,
5. Mettre à jour le modèle.

→ Les données doivent être de très bonne qualité

→ Le volume des données est important pour capturer l'information

Facteurs de pertinence et d'efficacité

La qualité du travail dépendra de facteurs initiaux contraignants, liés à la base de données :

- **Nombre d'*exemples*** (moins il y en a, plus l'analyse est difficile, mais plus il y en a, plus le besoin de mémoire informatique est élevé et plus longue est l'analyse) ;
- **Nombre et qualité des attributs** décrivant ces exemples. La distance entre deux « exemples » numériques (prix, taille, poids, intensité lumineuse, intensité de bruit, etc.) est facile à établir, celle entre deux attributs catégoriels (couleur, beauté, utilité…) est plus délicate ;
- **Pourcentage de données renseignées** et manquantes ;
- **« Bruit »** : le nombre et la « localisation » des valeurs douteuses (erreurs potentielles, valeurs aberrantes…) ou naturellement non-conformes au *pattern* de distribution générale des « exemples » sur leur espace de distribution impacteront sur la qualité de l'analyse.

Concept de prédiction

- De manière générale, nous aimerions prédire une valeur t à partir d'une observation x

$$y = f(x, w)$$

- Si y est continu : régression
- Si y est discret : classification
- Apprentissage des paramètres w à partir de données.
- Performance visée : minimiser erreur de prédiction
- Moyen mis en œuvre : utiliser des données expérimentales pour trouver un modèle prédiction=$f(x, w)$ le plus correct possible

Typologie de l'apprentissage

Capacité d'un système à améliorer ses performances via des interactions avec son environnement

- Quel « système » ?
  - types de modèle (Ad hoc ? Issu d'une famille particulière de fonctions mathématiques [tq splines, arbre de décision, réseau de neurones, arbre d'expression, machine à noyau…] ?)
- Quelles « interactions avec l'environnement » ?
  - apprentissage « hors-ligne » v.s. « en-ligne »
  - apprentissage « supervisé » ou non, « par renforcement »
- Quelles « performances » ?
  - fonction de coût, objectif, critère implicite, …
- *Comment* améliorer ?
  - type d'algorithme (gradient, résolution exacte problème quadratique, heuristique, …)

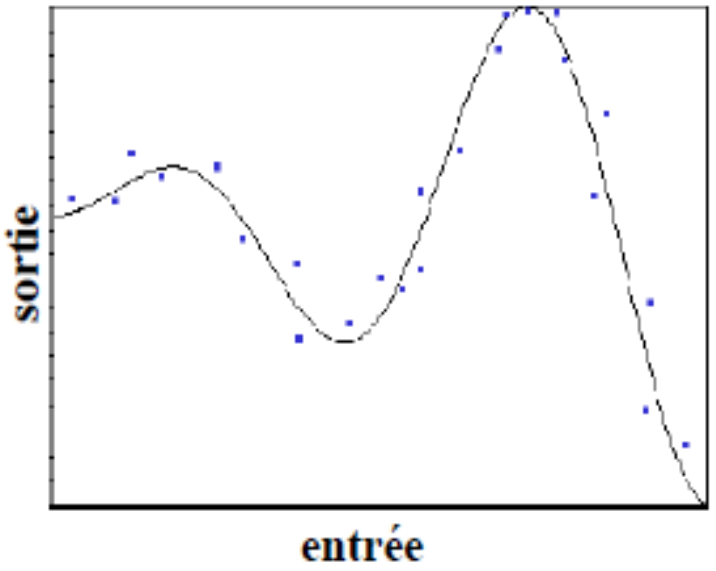
Principaux types d'algorithmes

- Résolution système linéaire (régression, Kalman, …)
- Algos classiques d'optimisation
  - Descente de gradient, gradient conjugué, …
  - Optimisation sous contrainte
  - …
- Heuristiques diverses :
  - Algorithme d'auto-organisation non supervisée de Kohonen
  - Algorithmes évolutionnistes (GA, GP, …)
  - « colonies de fourmis » (Ant Colony Optimization)
  - Optimisation par Essaim Particulaire (OEP)
  - Renforcement (Q-learning, …)

APPRENTISSAGE SUPERVISÉ :

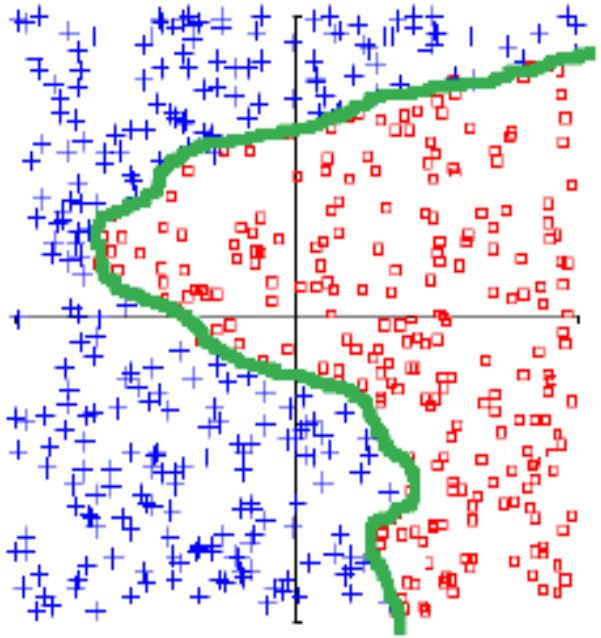
régression et classification

Régression
(approximation)



points = exemples ➔ courbe = régression

Classification
(y_i = « étiquettes »)



entrée = position point

sortie désirée = classe (□ =-1,+=+1)

⬇

Fonction étiquette=f(x) (et frontière de séparation)

MERCI POUR VOTRE ATTENTION

nachaoui@gmail.com