Vue d'ensemble du Machine Learning

Plan

- Définition de Machine Learning
- Ce que n'est pas du Machine Learning
- Cas d'application du Machine Learning
- Différents types de Machine Learning
- Différents algorithmes par types de Machine Learning
- Evaluation d'un modèle
- Difficultés à mettre en place le ML
- Processus d'élaboration d'un projet de ML

spidr

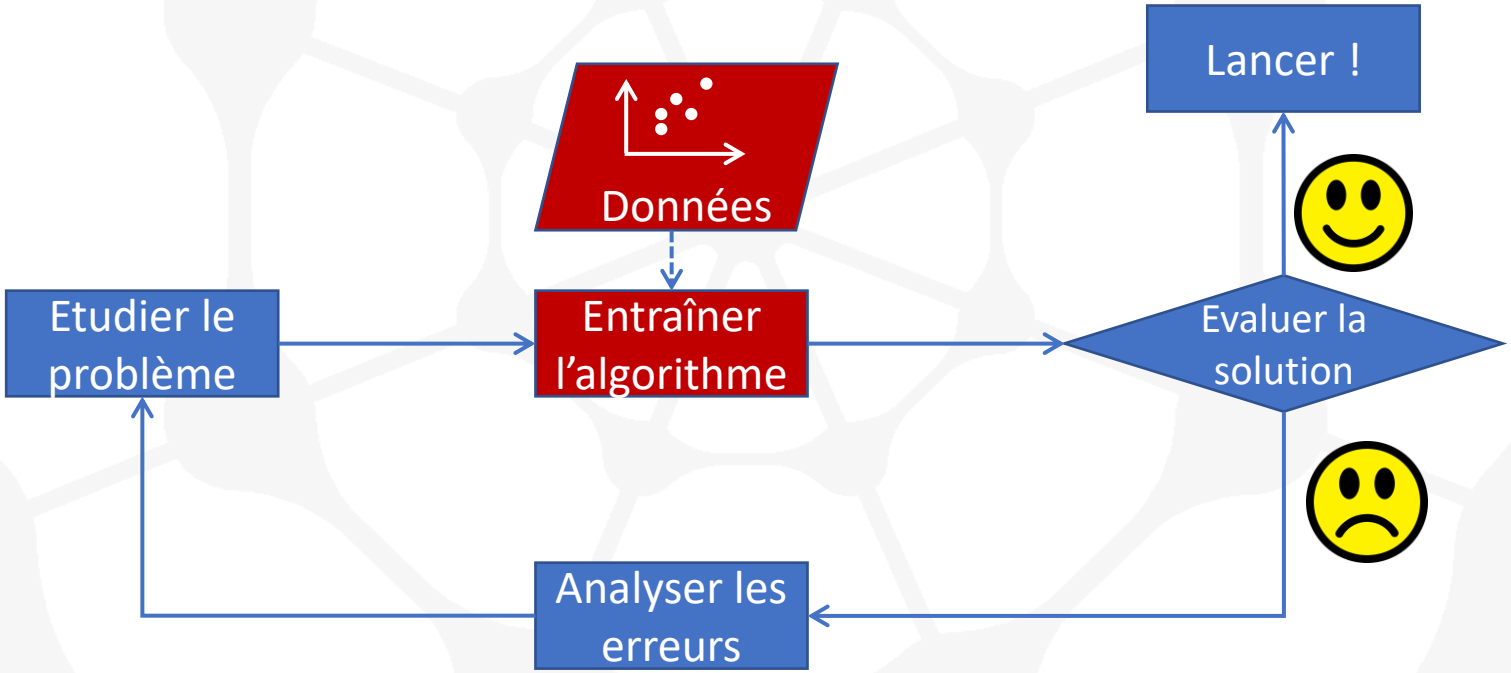
Définition de Machine Learning

- **Définition : « L'apprentissage automatique est la discipline donnant aux ordinateurs la capacité d'apprendre (*via des données et des algorithmes*) sans qu'ils soient explicitement programmés .» Arthur Samuel, 1959**

- **L'apprentissage automatique tourne autour de trois notions :**
  - **Un tâche T : *Problématique métier***
  - **Une performance P**
  - **Une expérience E : *Les données***

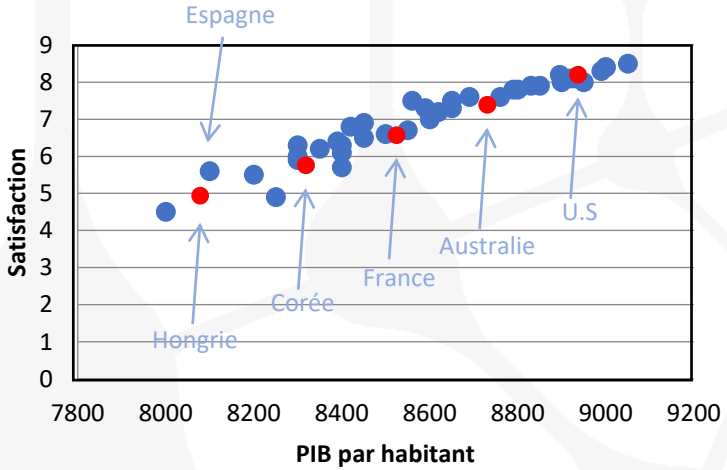
- **L'objectif est de réaliser la tâche T avec la meilleure performance possible P en apprenant de l'expérience E**

Définition de Machine Learning



Définition de Machine Learning

- **Apprentissage à partir d'un modèle**
  - Cette méthode consiste à construire un modèle à partir des données qui permettra de prédire sur les nouvelles données



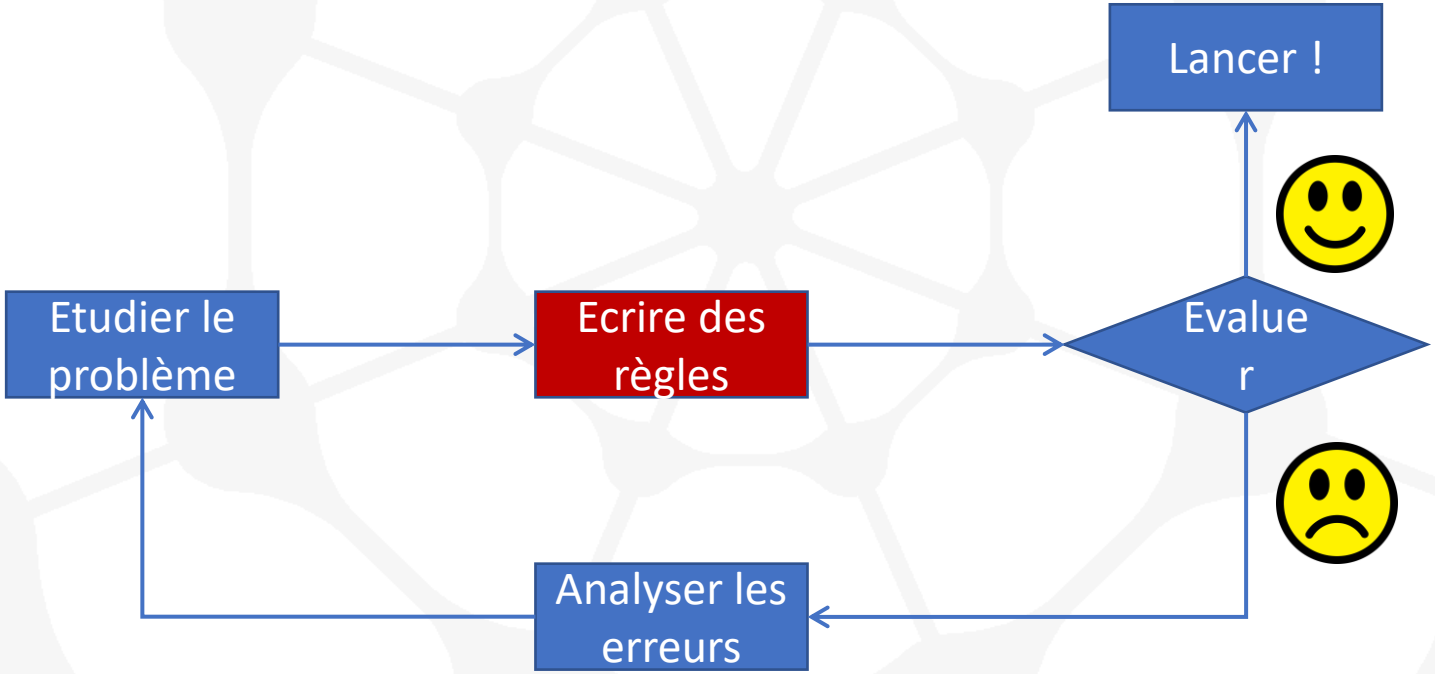
Modèle : *Satisfaction ~ $\vartheta_0 + \vartheta_1$ x PIB par habitant*

**Ne pas oublier que l'on cherche un modèle qui minimise l'erreur de prédiction*

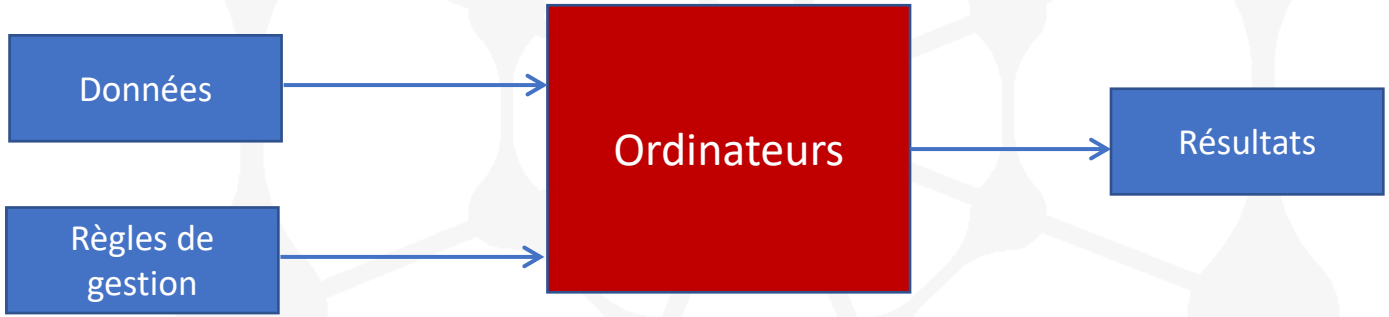
A l'aide de mon niveau de PIB je peux prédire le niveau de satisfaction pour un autre pays

Ce que n'est pas du ML

Ce que n'est pas du ML



Ce que n'est pas du ML



Ce que n'est pas du ML

- **Apprentissage à partir d'observations**
  - Il s'agit de la forme la plus banale d'apprentissage
  - On apprend sur un ensemble de données (par exemple les spams)
  - Le système apprend les exemples par cœur , puis il généralise à de nouveaux cas en utilisant une mesure de similarité

Quel niveau de similarité choisir ?

Pourquoi utiliser l'apprentissage automatique?

En résumé, l'apprentissage automatique est excellent pour :

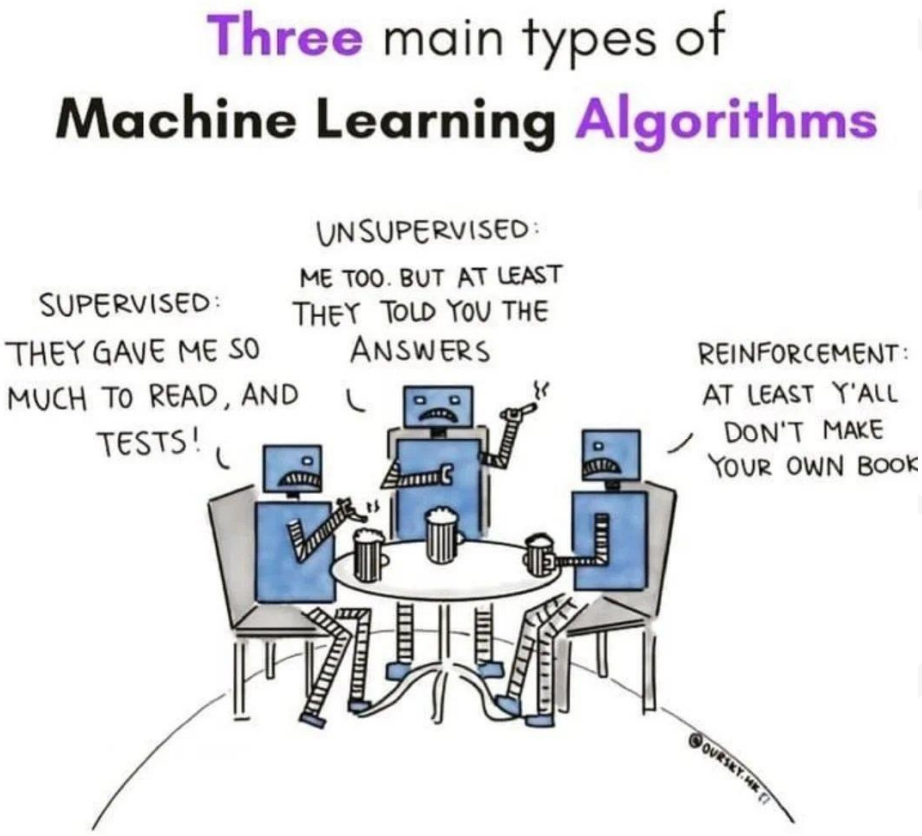
- Les problèmes pour lesquels les solutions existantes requièrent beaucoup d'ajustements manuels ou de longues listes de règles : un algorithme d'apprentissage automatique peut souvent simplifier le code et donner de meilleurs résultats

- Les problèmes complexes pour lesquels il n'existe aucune bonne solution si l'on adopte une approche traditionnelle : les meilleures techniques d'apprentissage automatique peuvent trouver une solution

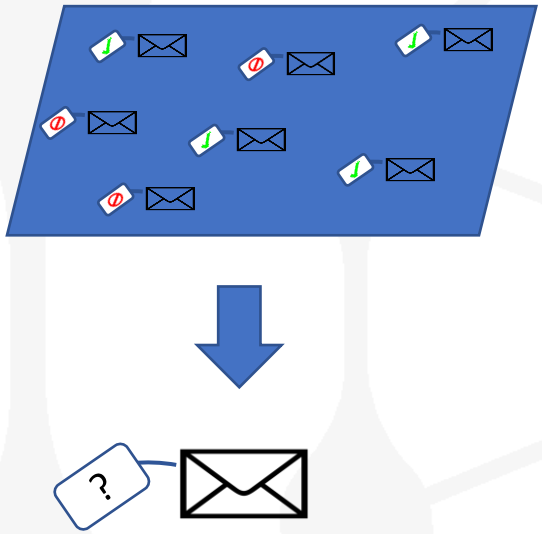
- Les environnements fluctuants : un système d'apprentissage automatique peut s'adapter à de nouvelles données

- L'exploration des problèmes complexes et des gros volumes de données

Types de systèmes d'apprentissage automatique



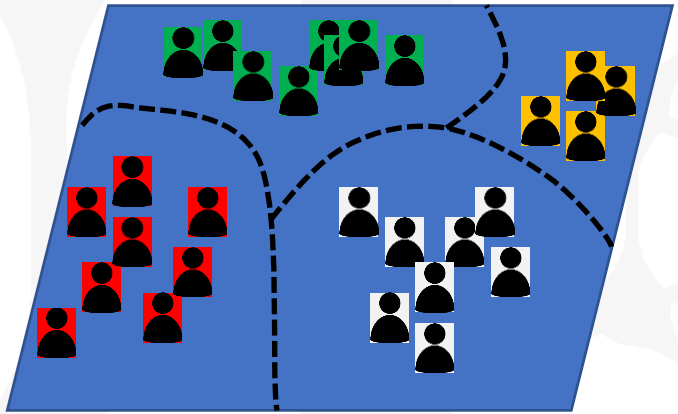
Types de ML : Apprentissage supervisé



Quelques Algo d'apprentissage supervisé:

- K plus proches voisins (KNN)
- Régression linéaire
- Régression logistique
- Machines à vecteurs de support
- Arbres de décision et forêts aléatoires
- Réseaux neuronaux

Types de ML : Apprentissage non supervisé

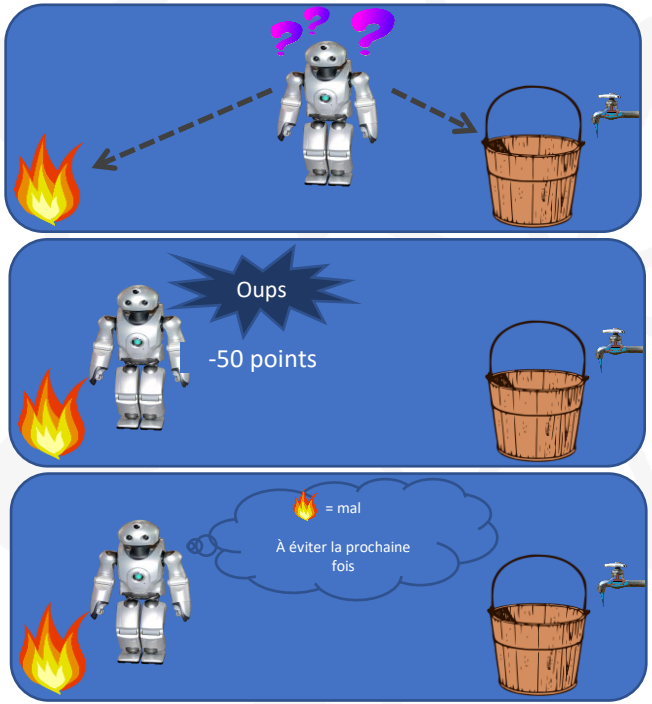


Quelques algorithmes d'apprentissage non supervisé :

- K-means, CAH
- ACP, TNSE
- DBScan
- Plongement localement linéaire
- One-class SVM
- Isolation Forest
- Eclat

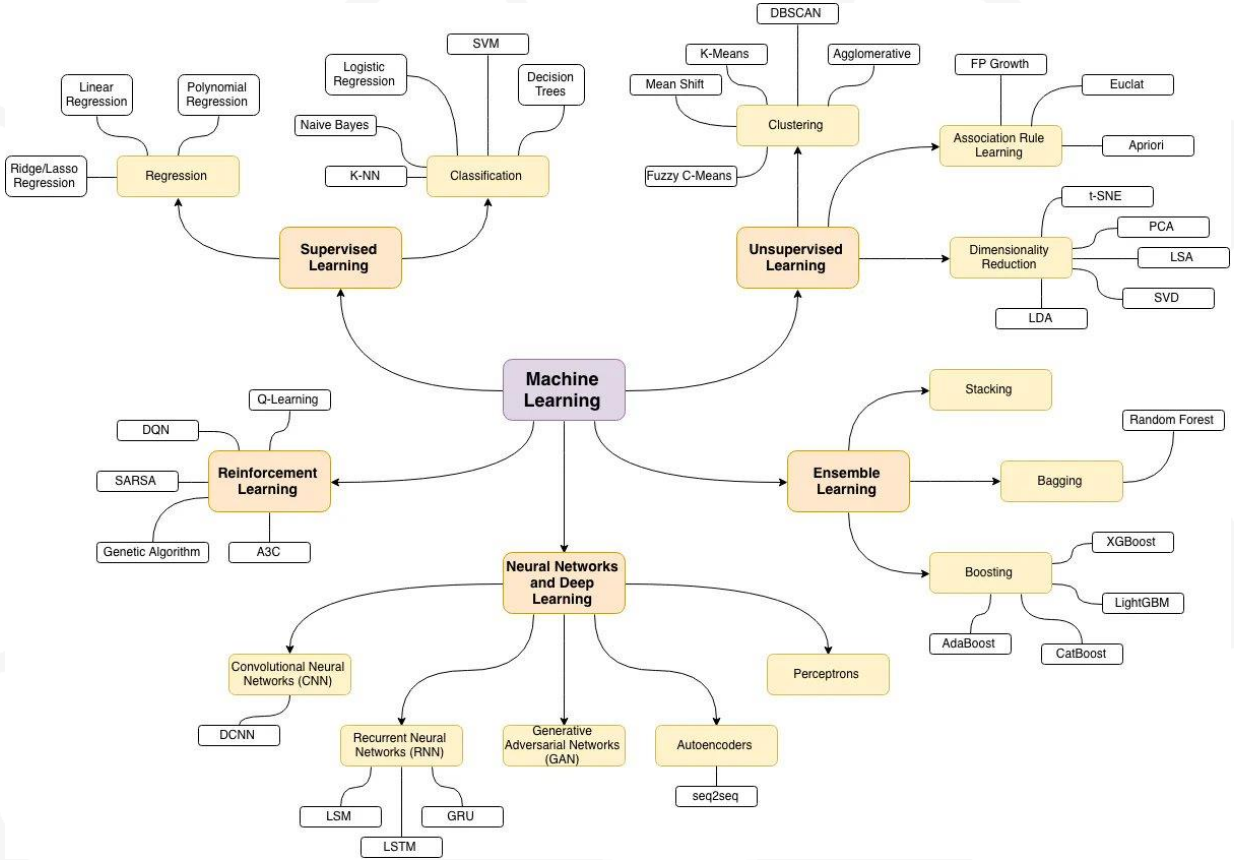
spidr

Types de ML : Apprentissage par renforcement

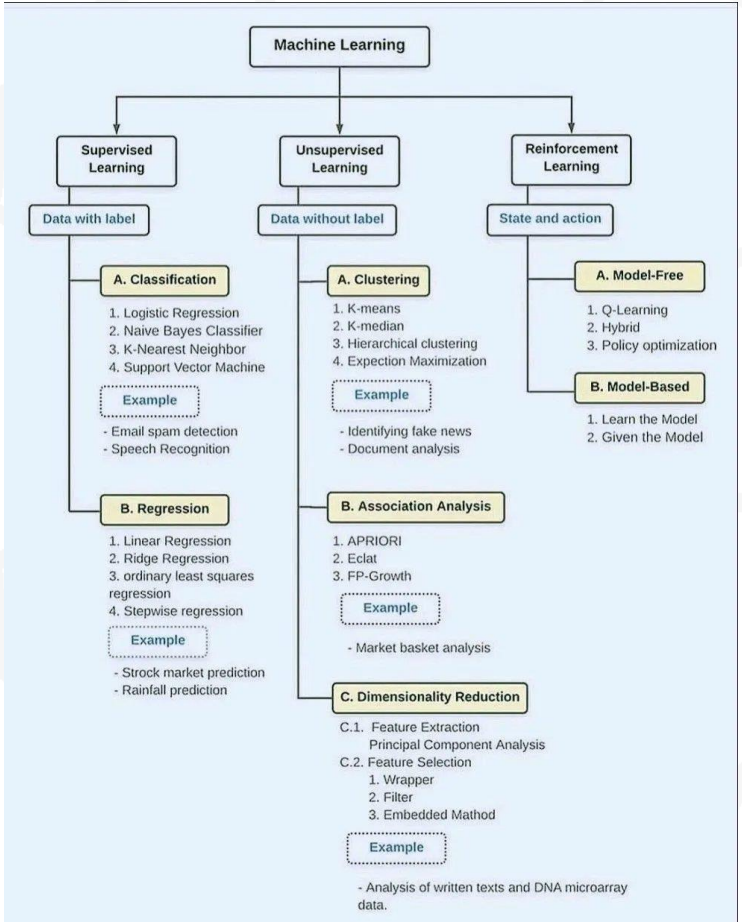


1. Observer
2. Choisir une action d'après la politique
3. Action
4. Récompense ou pénalité
5. Modifier la politique (apprentissage)
6. Itérer jusqu'à obtenir une politique optimale

Différents algorithmes par types de Machine Learning



Différents algorithmes par types de Machine Learning



Cas d'applications du Machine Learning

- Exemples de problèmes de régression :
  - Prédiction du montant des ventes d'une entreprise compte tenu du contexte économique.
  - Prédiction du prix de vente d'une maison en fonction de plusieurs critères.
  - Prédiction de la consommation électrique dans une ville étant donne
  - des conditions météorologiques. . .
  - Etc.

Cas d'applications du Machine Learning

- Exemples de problèmes de catégorisation :
  - Prédiction de l'état sain/malade d'un patient par rapport a une
  - maladie et compte tenu de divergents facteurs.
  - Prédiction de l'accord ou du refus d'un crédit a un client d'une banque
  - en fonction de ses caractéristiques.
  - Prédiction du chiffre correct a partir d'une image scannée d'un chiffre écrit a la main. . .
  - etc

Cas d'applications du Machine Learning

- Exemples de problèmes de catégorisation :
  - Analyse et classification des termes présents dans un texte
  - Analyse et prédiction de la trajectoire des étudiants après leur formation..
  - etc

Principales difficultés de l'apprentissage automatique

Données d'entrainement non représentatives

Si l'on apprend un modèle sur des données non représentatives, la prédiction pour de nouvelle données sera forcément entachée d'erreurs…

En 1936, pour l'élection présidentielle aux états unis, un institut de sondage prédisait Landon vainqueur sur Roosevelt avec 57% de votes. Dans les faits, Roosevelt l'a emporté avec 62% des votes.

Pourquoi?

L'institut de sondage a commis deux erreurs :

- Ils ont obtenu les adresses dans des annuaires téléphoniques, des listes d'abonnés, des listes de membres de clubs, etc. Une population plutôt aisée et votant majoritairement républicain (Landon). Il s'agit d'un biais d'échantillonnage
- Ils ont envoyé le sondage à 10 millions de personnes et moins de 25% avaient répondu. Il s'agit d'un biais de non réponse

Il est important de récolter de la données pour apprendre un modèle… Mais il faut le faire intelligemment

Principales difficultés de l'apprentissage automatique

▪ Données de mauvaise qualité

Il arrive que les données soient remplies d'erreurs, de valeurs aberrantes et de bruit (mauvaise qualité de mesure). Cela aura une incidence directe sur la qualité d'apprentissage et de prédiction d'un modèle.

Citons deux des cas les plus présents :

- **Les valeurs aberrantes/outliers** : Il s'agit des valeurs extrêmes, par exemple, une voiture ayant roulée 1 000 000 de kms…
- **Les valeurs manquantes** : il s'agit de « trous » pour certaines variables de notre jeu de données. Par exemple, dans le cas d'un sondage, environ 5% des répondants n'ont pas renseigné leur âge…

Principales difficultés de l'apprentissage automatique

Variables non pertinentes

Un des autres grands aspects du travail du data scientist va être de conserver les variables pertinentes pour avoir le modèle donnant les meilleurs résultats. C'est ce qu'on appelle le **FEATURE ENGINEERING**.

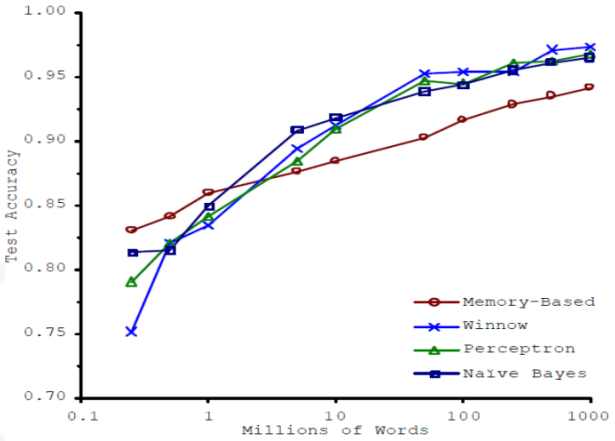
Cela concerne trois grands axes :

- La sélection des variables
- L'extraction des variables
- La collecte de nouvelles variables

spidr

Principales difficultés de l'apprentissage automatique

- **Données d'apprentissage en nombre insuffisant**
  - Un algorithme n'apprend pas comme un bébé, il lui faut beaucoup de données

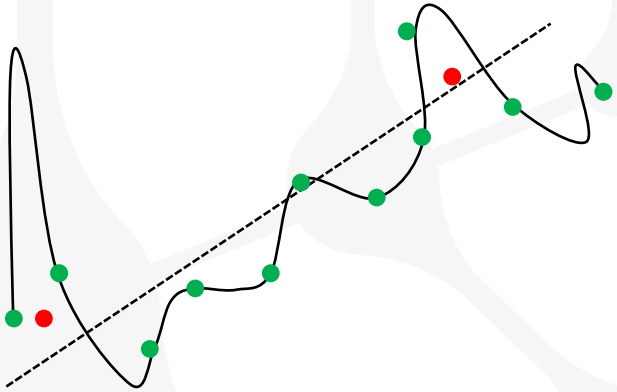


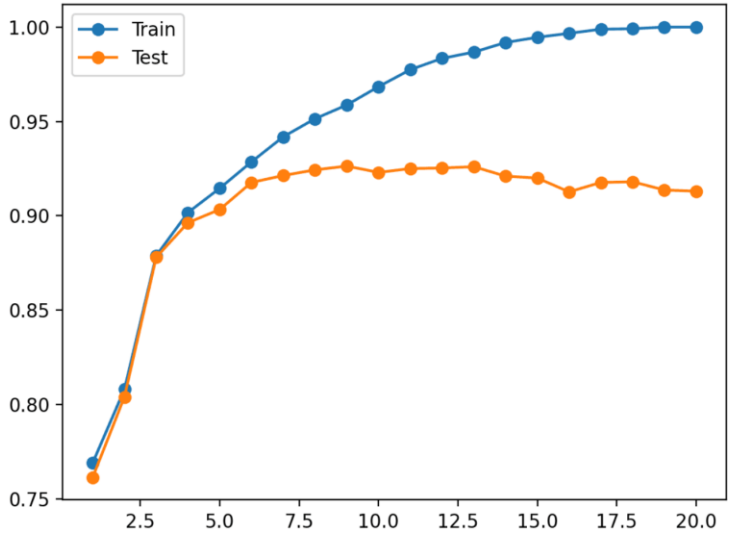
Banko et Brill (en 2001) ont montré que l'algorithme choisi avait moins d'impact sur le résultat que le nombre de données en entrée du modèle

Principales difficultés de l'apprentissage automatique

▬ Sur-apprentissage des données d'entraînement

Le Sur-apprentissage se produit lorsque le modèle est trop complexe par rapport à la quantité de données d'apprentissage et au bruit qu'elles contiennent.





spidr

Principales difficultés de l'apprentissage automatique

▪ Sur-apprentissage des données d'entraînement

Il existe plusieurs solutions possibles :

- Simplifier le modèle en sélectionnant moins de paramètres, en réduisant le nombre d'attributs des données d'entraînement ou en imposant des contraintes au modèle
- Rassembler davantage de données d'apprentissage
- Réduire le bruit dans ces données

Principales difficultés de l'apprentissage automatique

▪ Sous-apprentissage des données d'entraînement

Contrairement au surajustement, le sous-ajustement se produit lorsque le modèle est trop simple par rapport aux données d'apprentissage.

Il existe plusieurs solutions possibles :

- choisir un modèle plus puissant, avec plus de paramètres
- fournir de meilleures variables à l'algorithme d'apprentissage – en les transformant au besoin
- réduire les contraintes sur le modèle

Principales difficultés de l'apprentissage automatique

Un peu de recul

Résumons ce que nous avons vu :

- L'apprentissage automatique consiste à rendre une machine capable de mieux accomplir une tâche grâce à un entraînement sur des données, plutôt que d'avoir à explicitement coder les règles
- Il existe de nombreux types différents de systèmes d'apprentissage automatique supervisé ou non, en différé ou en ligne, à partir d'observations ou à partir de modèles, et ainsi de suite

Principales difficultés de l'apprentissage automatique

- Dans un tel projet, vous rassemblez des données dans un jeu d'entraînement :
  - si l'apprentissage s'effectue à partir d'un modèle, l'algorithme ajuste ce modèle aux données et on espère qu'il sera capable d'effectuer de bonnes prédictions
- Le système ne donnera pas de bons résultats
  - si le jeu de données est trop petit ,
  - si les données ne sont pas représentatives, entachées de bruit ou polluées.
- Le système ne doit ni être trop simple (sous-apprentissage) ou trop compliqué (sur-apprentissage).

Une question se pose… comment juger la qualité de notre modèle?

Lewis Hounkpevi

0695335936
lewis.dumesnil@gmail.com

spidr