Initiation au machine learning

Laurent Signac – CC-BY-SA – 20-12-20 2324 ac1b3f6d3c70f2e5b41b

Le machine learning (apprentissage automatique ou apprentissage machine en français) relève des statistiques, de l'intelligence artificielle, de l'informatique, des mathématiques appliquées (automatique…).

Les applications à succès ont connu une forte croissance depuis quelques années :

- recommandations (musique, film, nourriture…) ;
- diagnostic médical (des systèmes experts à l'analyse d'images médicales) ;
- reconnaissance de formes (visages, véhicules autonomes…) ;
- …

1 Deux types d'IA

Très schématiquement, les méthodes d'intelligence artificielle peuvent être divisées en deux catégories : règles fixes ou apprentissage automatique.

1.1 Règles fixes

Ce sont des règles du type : `si ceci alors, conclure cela`

Certains filtres antispam fonctionnent ainsi : si tel mot est présent, alors c'est du spam

Ces systèmes (≈ systèmes experts qui datent des années 60) fonctionnent bien dans certains cas mais ont deux défauts majeurs :

- le système est à revoir au moindre changement d'environnement (modification, ajout de règles) ;
- il nécessite une grande expertise de la part d'un humain (capable d'inventer et d'édicter les règles).

1.2 Apprentissage automatique

L'idée est de *présenter* ce qu'il faut faire (par exemple en apportant des exemples) et le système doit ensuite se débrouiller avec ça.

Exemple : pour différencier une femme d'un homme avec une photo de visage, on fournit des exemples de ce visage, plutôt que d'essayer de trouver des mesures discriminantes (couleur des yeux, cheveux, forme du visage, taille de la bouche etc…) et les règles qui vont avec.

Ce type de système est connu depuis assez longtemps, et il nécessite beaucoup d'exemples pour fonctionner. Une des nouveautés, outres les avancées théoriques, est la *disponibilité des exemples.*

2 Apprentissage supervisé / non supervisé

En apprentissage **supervisé**, on fournit les entrées **et les sorties** désirées sur des exemples connus, et on attend du système qu'il généralise. Par exemple, on pourrait fournir en entrée des images de chats et de chiens, et le tag chat ou chien en sortie.

Certains succès de l'apprentissage supervisé sont en partie dûs à la grande disponibilité (récente) de données.

- Exemples de tâches :
  - Identifier un code postal manuscrit (entrée : chiffres manuscrits / sorties : le code postal)
  - Dépistage de tumeurs (entrées : images médicales (irm par ex) / sorties : tag bénin/malin)
  - …

En apprentissage **non supervisé**, on ne connaît pas les sorties désirées, on ne fournit que des entrées, et on attend du système qu'il fournisse des sorties pertinentes (classification par exemple).

- Exemple de tâche :

  - création de groupes : on dispose de profils d'achat d'utilisateurs. On souhaite les segmenter en groupes ayant des comportements similaires et ainsi découvrir des types de profils
  - …

Dans les deux cas, la méthode d'apprentissage peut être dite *en ligne* : les données arrivent en permanence et le modèle doit être continuellement adapté (au fur et à mesure que les données arrivent). Dans le cas contraire (apprentissage *hors ligne*), la phase d'apprentissage précède la phase d'exploitation.

3 Les données

Les données doivent naturellement être représentables informatiquement (image en couleurs ou niveaux de gris, caractéristiques des objets, ou valeurs numériques, lieux géographiques…), et elles doivent souvent être très nombreuses (pour que le système fonctionne).

Dans les applications métier, il est souvent très important que la personne qui réalise l'analyse *comprenne* effectivement les données, c'est à dire fasse partie du bon corps de métiers, en plus d'avoir les connaissances nécessaires en apprentissage.

Les données sont généralement représentées en tableau :

- chaque ligne est un échantillon / un exemple
- chaque colonne est une caractéristique (feature) / une variable

Voici un exemple montrant quelques échantillons de la base de données *Iris*. Chaque échantillon est composé de 4 mesures sur la fleur (longueur et largeur des sépales et des pétales) ainsi que d'une colonne indiquant l'espèce :

5.1	3.8	1.6	0.2	Iris-setosa
4.6	3.2	1.4	0.2	Iris-setosa
5.3	3.7	1.5	0.2	Iris-setosa
5.0	3.3	1.4	0.2	Iris-setosa
7.0	3.2	4.7	1.4	Iris-versicolor
6.4	3.2	4.5	1.5	Iris-versicolor
6.9	3.1	4.9	1.5	Iris-versicolor
5.5	2.3	4.0	1.3	Iris-versicolor

L'opération qui consiste à collecter / nettoyer / choisir les données, et souvent difficile et pourtant cruciale.

4 Un domaine en plein effervescence

Le domaine de l'apprentissage automatique est en pleine effervescence, mais il n'est pas nouveau. Il existe une multitude de méthodes, qui marchent sur certains problèmes et pas sur d'autres. En conséquence, obtenir un résultat est *en partie* expérimental.

Nombre de ces méthodes sont implémentées dans des boîtes à outils et divers langages (R, Scala, ou Python par exemple). Réaliser un système qui fonctionne nécessite parfois seulement d'utiliser ces boîtes à outils. Parfois, il faut aussi comprendre comment elles fonctionnent, éventuellement les modifier, ou en créer de nouvelles.

5 Problèmes et méthodes

Il existe deux grands types de problème d'apprentissage, selon le type de sortie à laquelle on s'intéresse :

- Classification (sortie discrète)
  - binaire (seulement 2 classes)
  - multi classe
- Régression (sortie continue)

Les méthodes en apprentissage supervisé ou non sont très nombreuses. Voici quelques exemples :

- Apprentissage supervisé :
  - régressions (linéaires ou non)
  - **k plus proches voisins**
  - arbres de décision
  - **réseaux de neurones (et le sous-type : deep learning)**
  - …
- Apprentissage non supervisé :
  - réduction de dimension (analyse en composantes principales…)
  - clustering (cartes de Kohonen…)
  - k-means
  - …

Malgré la multiplicité des méthodes, celles-ci partagent certains points communs, en particulier sur la constitution des jeux de données, et sur les problèmes de sur/sous apprentissage.

5.1 Constitution des jeux de données

On doit généralement disposer de deux jeux de données :

- base d'exemples, utilisée pour l'apprentissage
- base de test, pour éprouver la généralisation de l'algorithme

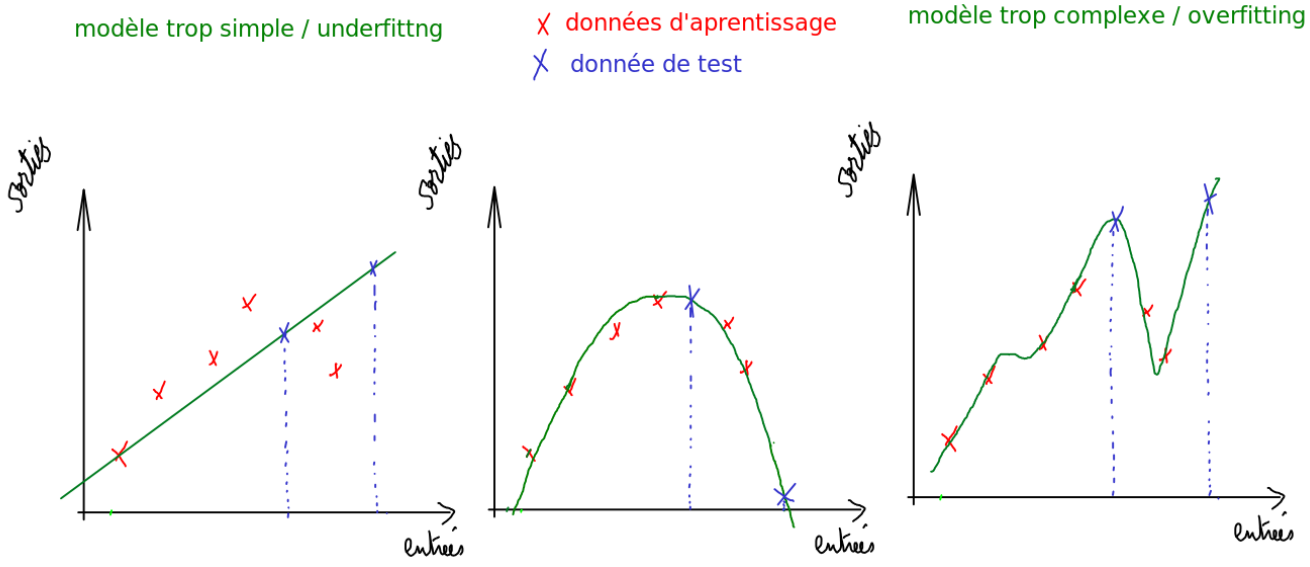
5.2 Sur/sous apprentissage (overfitting / underfitting)

Ces deux problèmes sont parfois désignés avec les noms *underfitting*/*overfitting* ou sous ajustement / sur ajustement :

- un modèle complexe est capable d'être particulièrement exact sur le jeu d'apprentissage (fig droite), mais il généralise mal (points bleus sur la figure de droite)
- un modèle trop simple (fig gauche) n'est pas forcément capable de comprendre des données complexes (underfiting)

En règle générale, plus les données sont nombreuses, plus on peut se permettre d'avoir un modèle complexe

Le choix du modèle pour ne pas avoir des problèmes de sur ou sous ajustement est loin d'être évident (généralement, on ne peux pas visualiser les données, car il y a trop de dimensions).



6 Algorithme des K-plus proches voisins

C'est un algorithme de classification multi-classe, en apprentissage supervisé.

L'idée générale est très simple. L'apprentissage est simplement la donnée des exemples. Le modèle consiste, pour taguer un nouveau point, à regarder les k-plus proches voisins dans la base : le vote majoritaire donne le tag à proposer

Sur la figure suivante, les croix sont les points appris (en haut à gauche). Le cercle est jugé en fonction des tags des voisins les plus proches (3 autres figures)