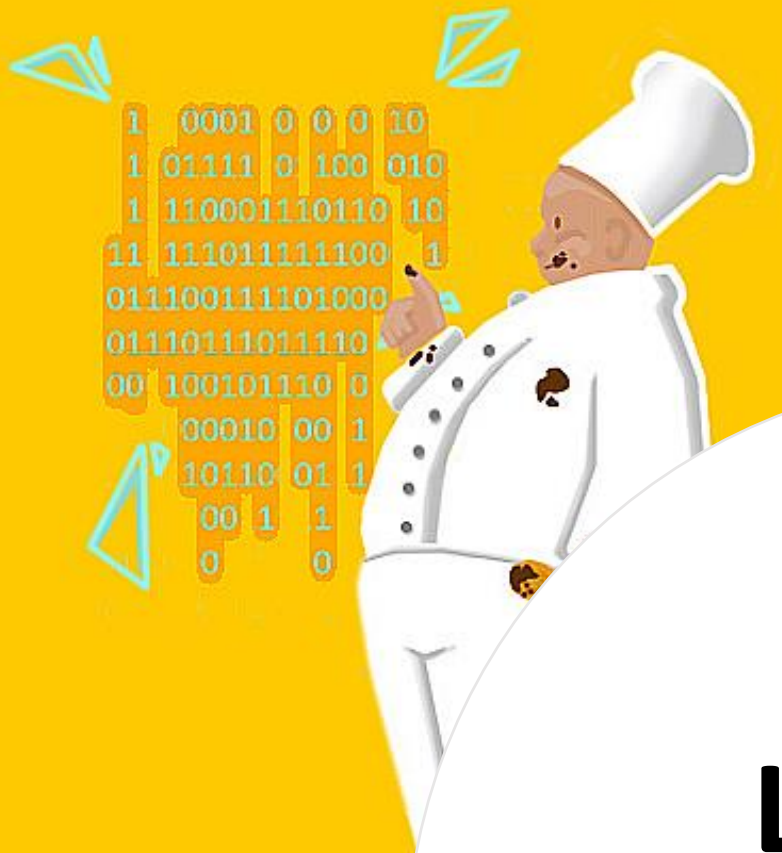
BeeBryte
Energy Intelligence

10 min



Le Machine Learning… tous aux fourneaux !

Lucas BOYER, Maxime LAHARRAGUE, Pierre-Jean LARPIN

16 avril 2018

Le Machine Learning… tous aux fourneaux !

Introduction aux principaux concept-clés

Lorsqu'on évoque les notions de « Machine Learning » et « d'intelligence artificielle », on entend souvent des réponses bien différentes d'un interlocuteur à l'autre : une technologie révolutionnaire ? Une robotisation des tâches quotidiennes ? En réalité, derrière ces buzzwords assez génériques se cachent des maths parfois vieilles de plus de 200 ans. En effet, Gauss et Legendre formulèrent dès 1829 la méthode des moindres carrés, base de certains algorithmes de prédiction utilisés aujourd'hui. Mais donc pourquoi tant d'années se sont-elles écoulées avant l'arrivée du Machine Learning dans nos vies ? Cela peut s'expliquer en partie par les immenses quantités de données et les puissances de calcul que nécessitent la pleine exploitation de ces algorithmes, apportées par les avancées informatiques de la fin du XXe siècle.

Première d'une série de quelques articles sur le sujet, cette publication se veut introduire les principaux concepts de Machine Learning en les illustrant d'exemples parlants.

QU'EST-CE QUE LE MACHINE LEARNING ?

Le Machine Learning est un domaine particulier du vaste éventail de solutions regroupées sous le nom générique d'intelligence artificielle, au même titre que les systèmes experts, les systèmes experts, ou de nombreuses solutions d'optimisation. Il est généralement lui-même subdivisé en trois sous-domaines : l'apprentissage supervisé (*Supervised Learning),* l'apprentissage non-supervisé (*Unsupervised Learning*) et l'apprentissage par renforcement (*Reinforcement Learning*).

Dans cet article, nous nous bornerons à l'étude du *Supervised Learning*, qui consiste à concevoir un modèle mathématique (régresseur, classifieur, etc…) autonome capable d'identifier par apprentissage des relations parfois très complexes (corrélation, causalité, …) entre un grand nombre de données contextuelles (les « features ») et certaines observables (les « cibles »). Une fois cet apprentissage effectué (et qui peut être approfondi régulièrement au gré de l'enrichissement des données) le modèle est capable de **généralisation**, c'est-à-dire : anticiper, reproduire ou extrapoler les valeurs ou catégories observables dans des contextes inédits.

Mais à l'image de la grande cuisine, le Machine Learning ne se limite pas à une simple recette scientifique. Il suppose aussi une bonne dose « d'art » ; derrière son clavier en guise de piano (de cuisson), le data-scientist aguerri compose, teste, ajuste. Les données sont ses ingrédients, les algorithmes ses ustensiles. Il choisit les meilleures combinaisons, élabore sa recette, utilise ses connaissances et son expérience pour tirer le meilleur de ce dont il dispose.

Tous les cuisiniers vous le diront, un plat réussi nécessite en premier lieu de bons ingrédients, de bons ustensiles, mais aussi et surtout le « coup de patte » forgé sur l'expérience pour enchaîner les préparations, les cuissons subtiles et les mélanges équilibrés. De la même manière, un programme de Machine Learning pertinent et efficace requiert des données en quantité et qualité suffisantes, des techniques et des moyens informatiques adaptés, et toute l'expertise d'un data-scientist pour assurer la sélection et l'enchainement des traitements et des méthodes les plus appropriés pour extraire toute la substantifique moelle (l'information) contenue dans les jeux d'apprentissage dont il dispose, et optimiser la performance de la fonctionnalité recherchée (prédiction, détection, classification, etc.).

Lucas BOYER, Maxime LAHARRAGUE, Pierre-Jean LARPIN 16 avril 2018

Le Machine Learning… tous aux fourneaux !

1ere ETAPE : Le choix et la préparation des ingrédients

Il s'agit dans un premier temps de « faire son marché » et sélectionner avec le plus grand soin ses données. En particulier, il convient de procéder par étape et de rejeter les données ne véhiculant a priori aucune information, celles caractérisées par une absence totale de lien de causalité avec la cible, ou encore celles véhiculant de l'information redondante.

Si on cherche par exemple à prédire la consommation d'un bâtiment, il est a priori peu pertinent de prendre en compte le revenu moyen des employés du bâtiment, comme il est parfaitement inutile de prendre en compte à la fois des relevés de températures en degré Celsius, et les mêmes relevés exprimés en degré Fahrenheit.

Il s'agit ensuite de s'assurer de la bonne mise en forme, propre et homogène des données. Cette étape est cruciale car dans « la vraie vie », les données provenant de mesures terrains présentent bien souvent des qualités disparates (trous, valeurs aberrantes, fréquences d'acquisition différentes, etc.). Cela passe par le traitement de chaque paramètre individuellement – cohérence des données, choix d'un format et d'une unité adaptés, normalisation éventuelles (mise à l'échelle des paramètres entre eux). Il faut également distinguer les paramètres quantitatifs (ex : la température ; 20°C est supérieur à 15°C) des paramètres qualitatifs (ex : le jour de la semaine ; Un Vendredi n'est pas supérieur à un Mardi) qui doivent être traités différemment.

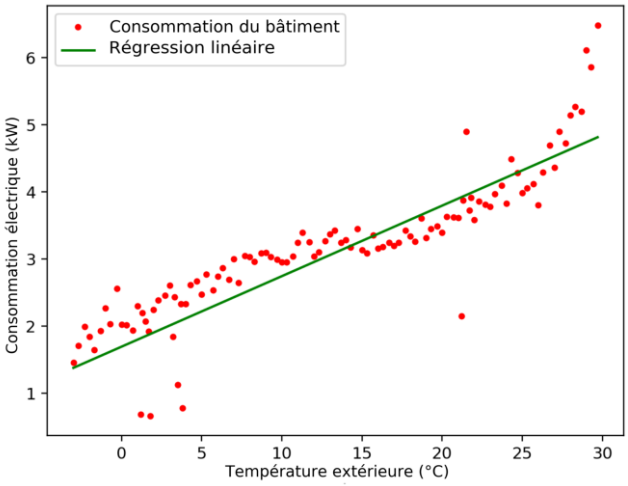
2ème ETAPE : Le choix des ustensiles

La seconde étape consiste à choisir le modèle de régression (ou estimateur) le plus adapté, et ses paramètres de réglage a priori (Eh oui, comme les épices, un plat de data-cuisine n'exprimera son plein potentiel que correctement assaisonné...). Ces choix doivent se faire en fonction du type de cible : quantitative ou qualitative, et parmi les nombreux modèles existants, nous proposons de survoler les trois plus classiques : la régression linéaire, les arbres de décision et les réseaux de neurones.

LA REGRESSION LINEAIRE

Dans sa version la plus simple, ce modèle linéaire se résume à trouver une combinaison linéaire des features permettant d'approcher « au mieux » l'estimation de la cible. Malgré sa simplicité, cette technique est remarquablement efficace et même généralisable, en utilisant diverses « bidouilles » mathématiques (typiquement le « kernel trick »), que nous ne développerons pas ici, et qui permettent de ramener au cas linéaire des systèmes très fortement non-linéaires.

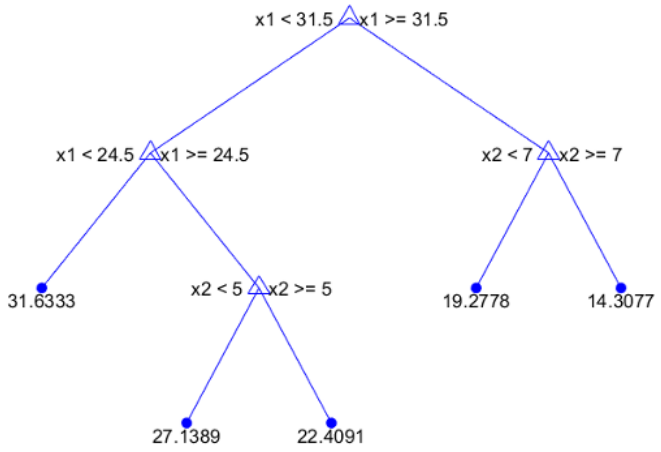
Un exemple de régression linéaire simple est présenté ci-dessous. En identifiant la droite qui passe « au mieux » des points, on dispose d'une bonne approximation de la relation sous-jacente entre la « feature » reportée en abscisse et la cible reportée en ordonnée.



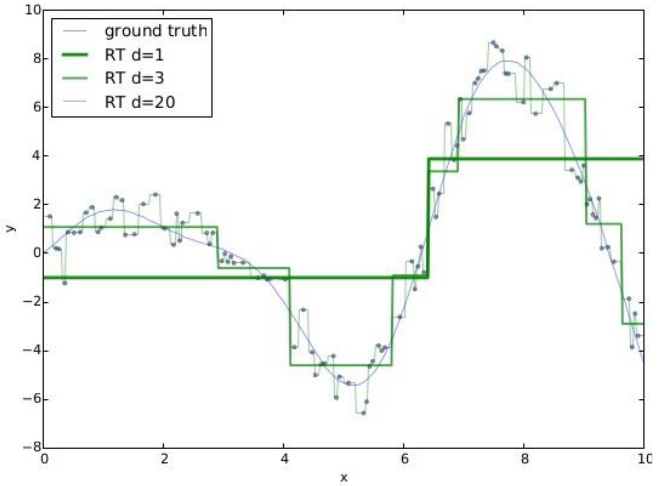
L'ARBRE DE REGRESSION & DE CLASSIFICATION

Un arbre de régression (ou de classification) a pour objectif de partitionner successivement l'échantillon de données selon une série de choix binaires par rapport à un seuil (ex : $x < 3$ ou $x \geq 3$). Le résultat de ce découpage s'organise selon un graph qui n'est pas sans rappeler la forme d'un arbre renversé. Un exemple de paramètre à choisir judicieusement est la profondeur de l'arbre, qui détermine le nombre de « feuilles » finales et donc la précision de la régression ou de la classification.

Lucas BOYER, Maxime LAHARRAGUE, Pierre-Jean LARPIN 16 avril 2018



L'apprentissage consiste à identifier à chaque embranchement le seuil le plus discriminant en termes d'information contenue dans le sous-échantillon résultant du choix binaire précédent. En exploitation, la valeur de « feature » est soumise à l'arbre, qui est parcouru conformément à « l'aiguillage » binaire de chaque embranchement. La « feuille » atteinte en fin de parcours est retournée en tant qu'estimation de la valeur cible. Le graph ci-après présente le résultat de régression par arbre, avec l'évolution de la précision selon la profondeur de l'arbre.



Opérationnellement, les arbres de régression ne sont aujourd'hui jamais utilisés seuls, mais démultipliés sous forme de « forêts », en introduisant de l'aléatoire dans leurs élaborations. L'estimation de la valeur cible se fait par exploitation statistique des résultats au sein de la forêt et se trouve ainsi lissée par effet de foisonnement, et meilleure en généralisation.

LE PERCEPTRON MULTICOUCHES, PLUS CONNU SOUS LE NOM DE « RESEAU DE NEURONES »

Un réseau de neurones est un modèle s'inspirant du fonctionnement des neurones biologiques. Sa force réside dans sa grande versatilité : théoriquement, un réseau de neurones bien dimensionné est capable de reproduire n'importe quelle fonction continue (théorème d'approximation universelle).

Nous ne détaillerons pas plus ici cet ustensile, qui fera prochainement l'objet d'un article dédié.

3ème ETAPE : Le savoir-faire du cuisinier

Le cuisinier de la donnée doit maintenant choisir les paramètres de son modèle afin d'optimiser la précision et la pertinence de ses prévisions, c'est-à-dire minimiser leur éloignement des cibles attendues, sans excès de complexité.

Toute la subtilité de l'apprentissage réside alors dans l'ajustement du modèle pour atteindre le juste équilibre entre (1) l'exploitation au maximum de l'information contenue dans l'échantillon d'apprentissage et qui est propre au modèle sous-jacent qu'on cherche à approcher (2) le rejet de l'information « parasite » (bruits, autres facteurs influents, etc…) qui n'est pas directement liée au modèle mais contingent à l'échantillon lui-même.

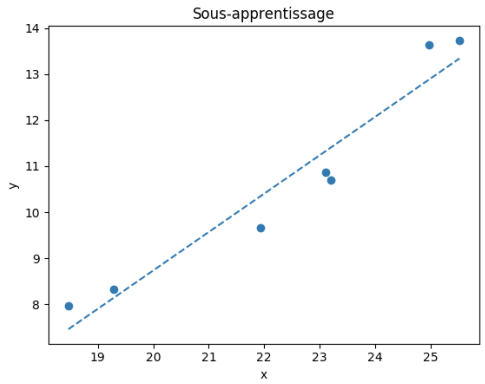
En bref : séparer le général du spécifique, pour justement obtenir à la fois la meilleure précision et la meilleure généralisation possible du modèle approché résultant. Nous allons illustrer ce point sur un exemple très simplifié :

UNDERFITTING

Lorsqu'on n'exploite pas suffisamment l'information contenue dans l'échantillon d'apprentissage, typiquement en ayant recours à un modèle trop simpliste, l'estimateur résultant a de bonne propriété de généralisation mais une précision insatisfaisante. On parle alors de sous-apprentissage ou

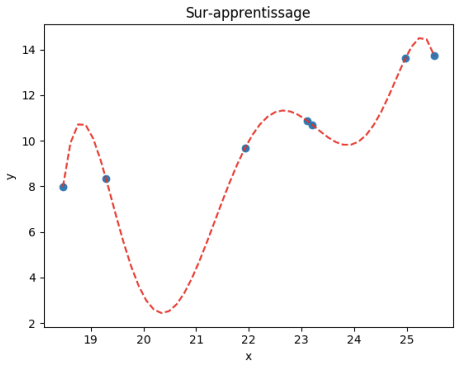
Le Machine Learning… tous aux fourneaux !

« Underfitting ». Le modèle possède un biais élevé et une variance faible.



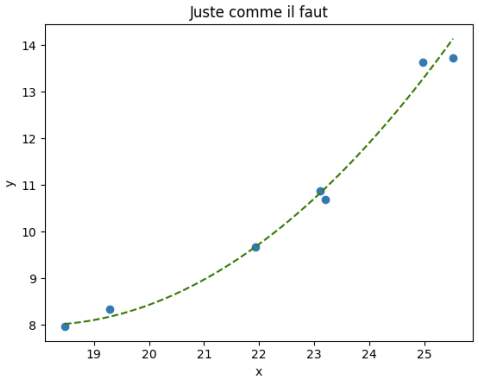
OVERFITTING

Lorsqu'on n'écarte pas suffisamment l'information spécifique à l'échantillon d'apprentissage, le modèle colle bien, même trop bien, aux données mais perd en qualité de prédiction, et renvoie des incohérences lors de l'application à des features éloignées du domaine de l'échantillon d'entraînement. On parle alors de sur-apprentissage ou « Overfitting ». Le modèle possède un biais faible mais une variance élevée puisqu'il cherche à être précis pour un nombre important de données.



CAS OPTIMAL

In fine, en data-science comme en cuisine, tout est affaire d'équilibre et un bon apprentissage doit conduire à un modèle bien balancé entre précision et généralisation, entre biais et variance.



A travers cet article et ses quelques digressions culinaires, nous espérons avoir éclairci quelques concepts de base du Machine Learning et avoir éveillé votre curiosité pour ce vaste domaine ! Dans le prochain article, nous nous pencherons sur ces fascinants outils que sont les réseaux de neurones, qui sont à la base du deep learning et à l'origine de la formidable accélération des performances du machine learning ces dernières années.

Bee hungry, Bee foolish !

BeeBryte développe des solutions autour de l'Intelligence Artificielle pour que les bâtiments industriels et commerciaux, les stations de recharge de véhicules électriques et les éco-quartiers consomment l'énergie de manière plus intelligente, moins chère et plus efficacement tout en réduisant leur empreinte carbone! BeeBryte a une équipe de 20 personnes basée en France et à Singapour et est soutenue par BPI-i Lab & l'ADEME. Depuis sa création en 2015, ses solutions ont reçu de nombreux prix, tels que EDF Pulse, Start-up Energy Transition Award, et le label GreenTech Verte.

contact@beebryte.com
www.twitter.com/BeeBryteGroup
www.beebryte.com