

Chapitre 4

Principaux algorithmes du Machine Learning

1. Ce que nous allons découvrir et les prérequis

Dans le chapitre précédent, nous avons découvert ou redécouvert les fondamentaux de l'analyse statistique descriptive qui, nous le verrons par la pratique, nous permettront de comprendre et de préparer nos données avant l'apprentissage. Nous allons à présent faire connaissance avec les principaux algorithmes du Machine Learning qui vont nous permettre de réaliser cet apprentissage.

Attention, notre objectif en écrivant cet ouvrage est de vulgariser les concepts de l'intelligence artificielle. Par conséquent, **nous n'y aborderons pas les explications théoriques et mathématiques de chaque algorithme d'apprentissage.**

Nous nous contenterons d'une explication la plus explicite possible illustrée par un ou plusieurs exemples le cas échéant. Si nous devons faire un parallèle avec le monde du bricolage, nous allons vous présenter les différents outils à utiliser en fonction du travail à réaliser, mais nous ne vous expliquerons pas comment ils ont été fabriqués.

Nous vous conseillons de considérer ce chapitre comme un aide-mémoire dans lequel vous pourrez venir vous référer au fur et à mesure de votre lecture afin de comprendre pourquoi nous utilisons tel ou tel algorithme et en comprendre son fonctionnement dans les grandes lignes.

■ Remarque

Prérequis nécessaires pour bien aborder ce chapitre : avoir lu le chapitre Des statistiques pour comprendre les données

2. Supervisé ou non supervisé ? Régression ou classification ?

Réaliser un apprentissage supervisé consiste à fournir à la machine des données étiquetées (labellisées) et propices à l'apprentissage. C'est-à-dire que nous allons analyser et préparer les données et leur donner une signification. C'est à partir de cette signification que la machine va réaliser son apprentissage. L'objectif étant d'indiquer à la machine que pour une série de données et pour une observation précise, la valeur à prédire est un chat, un chien ou bien une autre valeur.

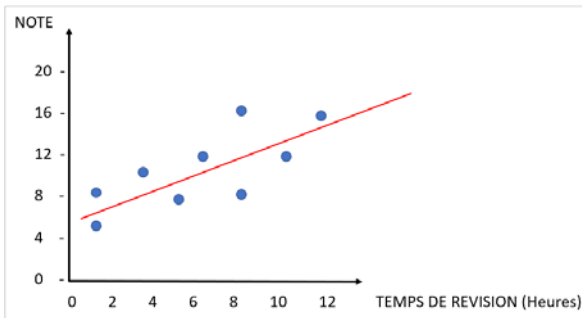
Lorsqu'il s'agit de prédire une valeur, nous parlerons alors de **régression**, dans le cas contraire, nous parlerons de **classification**. Prédire le pourcentage de réussite d'une équipe de football lors d'un match est une régression, prédire que la photo affichée est un chat ou un chien est une classification.

3. Les algorithmes d'apprentissage supervisés pour la régression (prédiction de valeurs)

3.1 La régression linéaire univariée (linear regression)

Cet algorithme cherche à établir, sous forme d'une droite, une relation entre une variable expliquée et une variable explicative. Par exemple, prédire une note à un examen (variable expliquée) en fonction du nombre d'heures de révisions (variable explicative).

En d'autres termes, les données d'une série d'observations sont représentées sous forme d'un nuage de points et l'on cherche à trouver une droite passant au plus près de ces points.



Régression linéaire univariée

Ainsi, connaissant le nombre d'heures de révisions, il nous est possible de prédire approximativement la note que l'on obtiendra au prochain examen.

3.2 La régression linéaire multiple (Multiple Linear Regression-MLR)

Là où nous utilisons une seule variable explicative pour expliquer une autre variable (une note en fonction d'un temps de révision), dans la régression linéaire multivariée nous allons utiliser plusieurs variables explicatives.

Par exemple, nous allons chercher à prédire le temps que va mettre un cycliste pour remporter une étape du tour de France, en fonction de son âge, du temps qu'il a réalisé à la précédente étape, de son classement dans le peloton...

Une étape importante lors de l'utilisation de multiples variables explicatives est leur **normalisation** (mise à l'échelle). Dans notre exemple, le temps réalisé en minutes lors de la précédente étape peut éventuellement varier entre 160 à 200, la position dans le peloton entre 1 et 80 en fonction du nombre de participants au tour de France. Nous ne sommes donc pas sur la même échelle pour chacune des variables explicatives (160 à 200 vs 1 à 80).

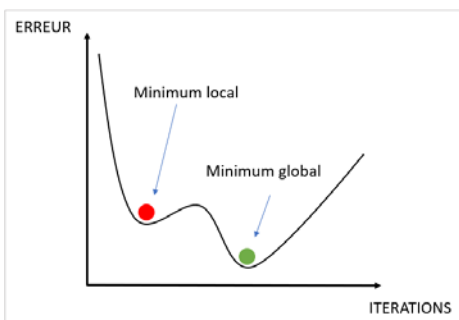
La mise à l'échelle (*scaling*) va donc consister à faire en sorte que la moyenne de chaque série d'observations soit égale à 0, que la variance et l'écart-type soient égaux à 1. Cette méthode est également appelée centrage de réduction.

Une fois cette étape réalisée, nous pouvons passer à la prédiction grâce à **la méthode de descente de gradient** ou bien encore **la méthode des moindres carrés**. Ces deux méthodes prenant en compte les différentes variables explicatives mises à l'échelle dans le but de prédire la variable expliquée.

3.3 La méthode de descente de gradient

Cette notion est essentielle, car elle est appliquée dans divers algorithmes d'apprentissage du Machine Learning et du Deep Learning que nous verrons un peu plus loin dans cet ouvrage.

Lorsqu'un système est en phase d'apprentissage, il commet des erreurs. Le taux d'erreur diminue au fur et à mesure de l'apprentissage, mais il se peut qu'à un moment donné l'erreur augmente pour à nouveau rediminuer et atteindre un niveau d'erreur plus bas que le précédent qui est le niveau optimal d'apprentissage.



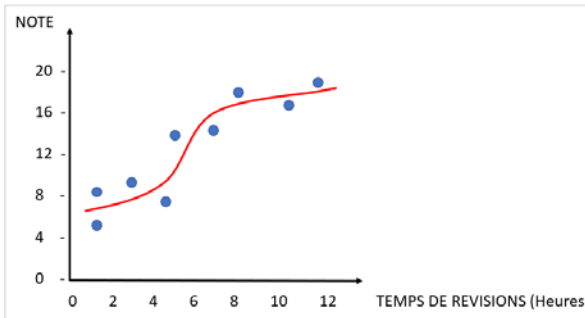
La descente de gradient

Sur la figure précédente, on constate qu'en début d'apprentissage, l'erreur diminue progressivement pour ensuite remonter. Nous aurions donc tendance à dire que le niveau optimal d'apprentissage a été atteint puisque de nouvelles erreurs apparaissent. Cependant, on peut s'apercevoir qu'après de nouvelles itérations d'apprentissage, l'erreur continue de diminuer pour atteindre un niveau plus bas que le précédent appelé minimum global ! Le niveau optimal d'apprentissage n'était donc pas atteint.

L'algorithme du gradient consiste donc à trouver par itérations successives le minimum global de la fonction de coût (erreur). Par analogie souvent reprise dans la littérature, imaginez-vous en haut d'une montagne avec pour objectif d'atteindre la plaine en contre bas. À chaque pas, vous analysez votre situation et décidez d'avancer de quelques pas, quitte à remonter pour prendre le chemin qui mène au but. Le gradient correspondant à la pente du sol que vous êtes en train de parcourir. Le "pas" porte également le nom de taux d'apprentissage dont nous verrons la mise en pratique dans le chapitre Un neurone pour prédire.

3.4 Régression polynomiale (polynomial regression)

Il est parfois difficile de trouver une droite pouvant passer parmi les points de la série d'observations de façon optimale. Cependant, il est parfois possible de trouver un lien entre les variables à l'aide d'une courbe. C'est ce que permet la régression polynomiale en ajoutant des plis à la courbe à l'aide d'éléments appelés polynômes.



Régression polynomiale

3.4.1 Monôme et polynôme

Un monôme est une expression mathématique s'exprimant sous cette forme :

$$\alpha x^n$$

Où

- α (alpha) est un nombre réel ou complexe appelé coefficient du monôme.
- n est un entier naturel représentant le degré du monôme.

Ainsi, $5x^2$ est un monôme de coefficient 5 et de degré 2.

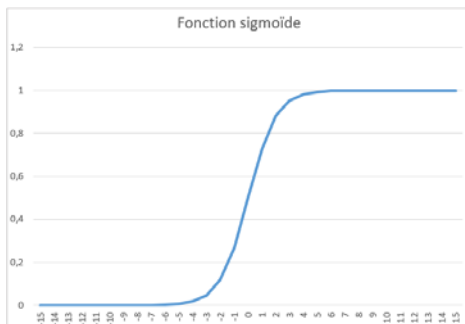
Un polynôme est une somme de monômes. On peut donc dire que $5x^2 + 2x$ est un polynôme.

3.5 Régression logistique

Comme nous venons de le voir, lorsque les données ne sont pas linéairement séparables, il est possible d'utiliser des polynômes pour donner à notre droite la possibilité de réaliser des virages afin de séparer nos observations.

La régression logistique utilise, quant à elle, une fonction logistique encore appelée sigmoïde ou courbe en S. Ce type d'algorithme est à appliquer dans des problèmes de classification.

À noter que nous croiserons plus en détail la fonction sigmoïde lorsque nous traiterons en pratique les réseaux de neurones.



Fonction sigmoïde ou courbe en S