Supervised Learning Algorithms

*

Les algorithmes de l'apprentissage supervisé

*

- **Introduction**
- **Chapitre 1 :** Tutoriel de régression linéaire
- **Chapitre 2 :** Tutoriel sur le classificateur de régression logistique
- **Chapitre 3 :** Arbres de décision et forêts aléatoires pour les débutants
- **Chapitre 4 :** Tutoriel sur le classificateur de machines à vecteurs de support
- **Chapitre 5 :** Évaluation et optimisation du modèle

Introduction

Bienvenue dans le monde fascinant du machine learning, où les algorithmes, les mathématiques et la créativité s'unissent pour donner vie à des systèmes capables d'apprendre et de s'adapter à partir de données. Dans ce cours, nous allons explorer ensemble les fondements théoriques et les applications pratiques de cette discipline, qui est au cœur de nombreux progrès technologiques actuels, du diagnostic médical à la reconnaissance d'images, en passant par la traduction automatique et les recommandations personnalisées.

Au fur et à mesure de notre progression, vous découvrirez comment les algorithmes de machine learning permettent de transformer des données brutes en informations utiles et exploitables. Vous serez amenés à comprendre et à concevoir des modèles capables de déceler des tendances, de faire des prédictions et même de prendre des décisions. Ce voyage dans le monde du machine learning vous fournira non seulement les compétences techniques nécessaires, mais aussi la capacité à raisonner sur les enjeux éthiques et les implications de l'utilisation de ces technologies dans la société.

Préparez-vous à relever des défis intellectuels et à développer des solutions créatives aux problèmes du monde réel. Vous êtes sur le point de découvrir comment les machines peuvent non seulement traiter les données, mais aussi "apprendre" d'elles. Bienvenue dans cette aventure stimulante, qui vous permettra de devenir acteurs et actrices de la révolution numérique actuelle !

Les techniques qu'on va apprendre dans ce cours vont nous permettre beaucoup de choses comme estimer le prix d'un appartement, prédire le cours de la bourse, détecter un objet sur une photo ou même calculer vos chances de survie dans une catastrophe comme celle du Titanic. Avant tout ça, voyons d'abord qu'est-ce que c'est : *machine learning* (*apprentissage supervisé*).

Le *machine learning*, c'est l'art de donner à une machine la capacité d'apprendre sans la programmer de façon explicite. Ça, c'est la définition historique.

Maintenant, dans les faits, le *machine learning* consiste à développer un modèle mathématique à partir de données expérimentales.

Pour cela on distingue 3 techniques de *Machine Learning* :

- L'apprentissage supervisé,
- L'apprentissage non supervisé,
- L'apprentissage par renforcement.

Dans l'*apprentissage supervisé*, la machine reçoit des données caractérisées par des variables x (couleur : x_1, longueur : x_2, largeur : x_3) et annotées d'une

variable y. Dans le contexte du *machine learning*, on appelle les variables x des "features" (caractéristiques), la variable y, quant à elle, est nommée "label" (étiquette).

Dans l'*apprentissage supervisé*, le but est que la machine apprenne à prédire la valeur y en fonction des "features" x qu'on lui donne. C'est pour cela que y est également appelée "target" (variable cible), ce qui veut dire objectif. Pour réussir à faire cela, on commence par donner plein de données à la machine.

En fournissant beaucoup de données à la machine, on constitue un *dataset*.

Ensuite, on spécifie quel type de modèle la machine doit apprendre, en précisant les hyperparamètres du modèle (dans le *machine learning*, un hyperparamètre est un paramètre dont la valeur est utilisée pour contrôler le processus d'apprentissage).

Est-ce qu'il s'agit d'un modèle linéaire ? d'un modèle polynomial ? d'un arbre de décision ou bien d'un réseau de neurones ? On spécifie les hyperparamètres de notre modèle : par exemple, combien de branches il doit y avoir dans un arbre de décision.

Une fois qu'on a fait cela, la machine va commencer à travailler. Elle va utiliser un algorithme d'optimisation pour trouver quels sont les paramètres du modèle qui nous donnent les meilleures performances pour les données de notre *dataset*.

C'est ce qu'on appelle la phase d'entraînement.

Une fois cette phase terminée, notre modèle de *machine learning* est prêt à être utilisé ; lorsque la machine reçoit de nouvelles données (sans étiquette cette fois-ci), elle va utiliser le modèle pour prédire quelle est la valeur de y.

Grâce à l'apprentissage supervisé, on peut à la fois résoudre des problèmes de *régression* (quand y est une variable continue, c'est-à-dire une variable quantitative) et des problèmes de *classification* (quand y est une variable discrète, c'est-à-dire une variable qualitative).

L'apprentissage supervisé peut être divisé en deux types de problèmes lors de l'exploration de données : la classification et la régression.

La classification utilise un algorithme pour attribuer, avec précision, les données de test à des catégories spécifiques. Il reconnaît des entités spécifiques au sein de l'ensemble de données et tente de tirer des conclusions sur la manière dont ces entités doivent être étiquetées ou définies. Les algorithmes de classification courants sont les classificateurs linéaires, les machines à vecteur de support

(SVM), les arbres de décision et la forêt aléatoire.

La régression est utilisée pour comprendre la relation entre les variables dépendantes et indépendantes. Elle est couramment utilisée pour faire des projections, par exemple pour le chiffre d'affaires d'une entreprise donnée. La régression linéaire, la régression logistique et la régression polynomiale sont les algorithmes les plus populaires.

Chapitre 1 : Tutoriel de régression linéaire

*

La régression linéaire est utilisée pour identifier la relation entre une variable dépendante et une ou plusieurs variables indépendantes et est généralement exploitée pour faire des prédictions sur les résultats futurs. Lorsqu'il n'y a qu'une seule variable indépendante et une variable dépendante, on parle de régression linéaire simple. À mesure que le nombre de variables indépendantes augmente, on parle de régression linéaire multiple.

Pour chaque type de régression linéaire, on cherche à tracer une ligne de meilleur ajustement calculée par la méthode des moindres carrés.

Méthode des moindres carrés

Une situation courante en sciences biologiques est d'avoir à sa disposition deux ensembles de données de taille n : $\{y_1, y_2, \ldots, y_n\}$ et $\{x_1, x_2, \ldots, x_n\}$, obtenus expérimentalement ou mesurés sur une population. Le problème de la régression consiste à rechercher une relation pouvant éventuellement exister entre les x et les y, par exemple de la forme $y = f(x)$.

Lorsque la relation recherchée est affine, c'est-à-dire de la forme $y = ax + b$, on parle de régression linéaire. Mais même si une telle relation est effectivement présente, les données mesurées ne vérifient pas en général cette relation exactement.

La droite des moindres carrés

Pour tenir compte dans le modèle mathématique des erreurs observées, on considère les données $\{y_1, \ldots, y_n\}$ comme autant de réalisations d'une variable aléatoire Y et parfois aussi les données $\{x_1, \ldots, x_n\}$ comme autant de réalisations d'une variable aléatoire X. On dit que la variable Y est la variable dépendante ou variable expliquée, et que la variable X est la variable explicative.

Les données $\{(x_i, y_i) : i = 1, \ldots, n\}$ peuvent être représentées par un nuage de n points dans le plan (x, y), le diagramme de dispersion. Le centre de gravité de ce nuage peut se calculer facilement : il s'agit du point de coordonnées

$$(\overline{x}, \overline{y}) = \left(\frac{1}{n} \sum_{i=1}^{n} x_i, \frac{1}{n} \sum_{i=1}^{n} y_i \right).$$

Rechercher une relation affine entre les variables X et Y revient à rechercher une droite qui s'ajuste le mieux possible à ce nuage de points. Parmi toutes les droites possibles, on retient celle qui rend minimale la somme des carrés des écarts des valeurs observées y_i à la droite $\hat{y}_i = ax_i + b$. Si ε_i représente cet écart, appelé aussi résidu, le principe des moindres carrés consiste à choisir les valeurs

de a et b qui minimisent :

$$E = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - (ax_i + b)\right)^2 .$$

Un calcul montre que les valeurs notées a et b sont égales à :

$$\hat{a} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \text{et} \quad \hat{b} = \bar{y} - \hat{a}\,\bar{x}$$

On exprime souvent $\hat{a}$ au moyen de la variance de x, S_x^2, et de la covariance des variables aléatoires x et y comme :

$$\hat{a} = \frac{\mathrm{cov}_{xy}}{S_x^2} \quad \text{avec} \quad S_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad \text{et}$$

$$\mathrm{cov}_{xy} = \tfrac{1}{n}\textstyle\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

$$E = \sum_{i=1}^{n}(y_i - ax_i - b)^2$$

Le minimum de cette expression est trouvé quand les deux dérivées partielles $\frac{\partial E}{\partial a}$ et $\frac{\partial E}{\partial b}$ sont égales à 0 :

$$\frac{\partial E}{\partial a} = \sum_{i=1}^{n} 2(y_i - ax_i - b)(-x_i) = 0$$

$$\frac{\partial E}{\partial b} = \sum_{i=1}^{n} 2(y_i - ax_i - b)(-1) = 0$$

ce qui donne le système d'équations suivant :

$$a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i$$

$$a\sum_{i=1}^{n} x_i + nb = \sum_{i=1}^{n} y_i$$

En forme matricielle, on a :

$$\begin{pmatrix} \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & n \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} x_i y_i \\ \sum_{i=1}^{n} y_i \end{pmatrix}$$

$$\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & n \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^{n} x_i y_i \\ \sum_{i=1}^{n} y_i \end{pmatrix}$$

$$\begin{pmatrix} \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & 1 \end{pmatrix}^{-1} = \frac{1}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2}$$
$$\begin{pmatrix} n & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix}$$

$$a = \frac{n\sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2}$$
$$b = \frac{\sum_{i=1}^n x_i \left(\sum_{i=1}^n x_i y_i\right) + n\sum_{i=1}^n y_i}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2}$$

Évaluation de la qualité de la régression

Pour mesurer la qualité de l'approximation d'un nuage $(x_i, y_i)_{i=1,\ldots,n}$ par sa droite des moindres carrés (comparatifs : faire passer une droite par n'importe quel nuage), on calcule son coefficient de corrélation linéaire défini par :

$$r_{xy} = \frac{\mathrm{cov}_{xy}}{s_x s_y}$$

C'est un nombre compris entre -1 et $+1$ qui vaut $+1$ (resp. -1) si les points du nuage sont exactement alignés sur une droite de pente positive (resp. négative). Ce coefficient est une mesure de la dispersion du nuage. On considère que l'approximation d'un nuage par sa droite des moindres carrés est de bonne qualité lorsque $|r_{xy}|$ est proche de 1 (donc r_{xy} proche de $+1$ ou de -1) et de médiocre qualité lorsque $|r_{xy}|$ est proche de 0.

En pratique, on estime souvent la régression acceptable lorsque

$$|r_{xy}| > \frac{\sqrt{3}}{2} = \sqrt{0,75} \approx 0,8666.$$

Parfois, on préfère calculer non plus r_{xy} mais son carré noté R^2 avec la relation suivante :

$$R^2 = r_{xy} \cdot r_{xy}$$

car on a la relation suivante :

$$\sum (y_i - \overline{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \overline{y})^2$$

qui exprime que la dispersion totale de y (DT) est égale à la dispersion autour de la régression (DA), plus la dispersion due à la régression (DR). Or, on peut vérifier que l'on a

$$R^2 = \frac{\mathrm{DR}}{\mathrm{DT}}$$

c'est-à-dire que le R^2 représente la part de la dispersion totale de y que l'on peut expliquer par la régression. Ainsi, si l'on obtient une valeur de $R^2 = 0{,}85$ (et donc $r = 0{,}92$), cela signifie que la modélisation par la droite des moindres carrés explique 85% de la variation totale, ce qui est un très bon résultat.

Cependant, même avec un R^2 excellent (proche de 1), notre modèle linéaire peut encore être rejeté. En effet, pour être assuré que les formules données a et b fournissent de bonnes estimations de la pente et de l'ordonnée à l'origine de la droite de régression, il est nécessaire que les résidus ε_i soient indépendants et distribués aléatoirement autour de 0. Ces hypothèses ne sont pas forcément faciles à vérifier. Un tracé des résidus et un examen de leur histogramme permettent de détecter une anomalie grossière, mais il faut faire appel à des techniques statistiques plus élaborées pour tester réellement ces hypothèses (ce que nous ne ferons pas ici).

Prévisions

Si $y = \hat{a}x + \hat{b}$ est la droite des moindres carrés d'un nuage de points $(x_i, y_i)_{i=1,\ldots,n}$, on appelle valeurs prédites de y par le modèle les valeurs $\hat{y_i} = \hat{a}x_i + \hat{b}$.

Notons cependant qu'il peut sembler naturel d'inférer une valeur prédite pour compléter les données initiales dans l'intervalle des valeurs de x. On se gardera de procéder sans de multiples précautions supplémentaires aux valeurs de x en dehors de cet intervalle. En effet, il se peut que la relation entre x et y ne soit pas du tout linéaire, mais qu'elle nous soit apparue comme telle à tort parce que les x_i sont proches les uns des autres.