Introduction au machine learning

3e édition

Chloé-Agathe AZENCOTT
Professeure au CBIO (Centre de bio-informatique)
de Mines Paris, de l'Institut Curie et de l'INSERM

DUNOD

11, rue Paul Bert, 92240 Malakoff
www.dunod.com

ISBN 978-2-10-088668-5

TABLE DES MATIÈRES

AVANT-PROPOS

Le *machine learning* (apprentissage automatique) est au cœur de la science des données et de l'intelligence artificielle. Que l'on parle de transformation numérique des entreprises, de Big Data ou de stratégie nationale ou européenne, le machine learning est devenu incontournable. Ses applications sont nombreuses et variées, allant des moteurs de recherche et de la reconnaissance de caractères à la recherche en génomique, l'analyse des réseaux sociaux, la publicité ciblée, la vision par ordinateur, la traduction automatique ou encore le trading algorithmique.

À l'intersection des statistiques et de l'informatique, le machine learning se préoccupe de la modélisation des données. Les grands principes de ce domaine ont émergé des statistiques fréquentistes ou bayésiennes, de l'intelligence artificielle ou encore du traitement du signal. Dans ce livre, nous considérons que le machine learning est la science de l'apprentissage automatique d'une fonction prédictive à partir d'un jeu d'observations de données étiquetées ou non.

Ce livre se veut une introduction aux concepts et algorithmes qui fondent le machine learning, et en propose une vision centrée sur la minimisation d'un risque empirique par rapport à une classe donnée de fonctions de prédictions.

Objectifs pédagogiques : Le but de ce livre est de vous accompagner dans votre découverte du machine learning et de vous fournir les outils nécessaires à :

1. identifier les problèmes qui peuvent être résolus par des approches de machine learning ;
2. formaliser ces problèmes en termes de machine learning ;
3. identifier les algorithmes classiques les plus appropriés pour ces problèmes et les mettre en œuvre ;
4. implémenter ces algorithmes par vous-même afin d'en comprendre les tenants et aboutissants ;
5. évaluer et comparer de la manière la plus pertinente possible les performances de plusieurs algorithmes de machine learning pour une application particulière.

Public visé : Ce livre s'adresse à des étudiantes ou étudiants en informatique ou mathématiques appliquées, niveau L3 ou M1 (ou deuxième année d'école d'ingénieur), qui cherchent à comprendre les fondements des principaux algorithmes utilisés en machine learning. Il se base sur mes cours à CentraleSupélec, à Mines Paris - PSL et sur OpenClassrooms et suppose les prérequis suivants :

- algèbre linéaire (inversion de matrice, théorème spectral, décomposition en valeurs propres et vecteurs propres) ;
- notions de probabilités (variable aléatoire, distributions, théorème de Bayes).

Plan du livre : Ce livre commence par une vue d'ensemble du machine learning et des différents types de problèmes qu'il permet de résoudre. Il présente comment ces problèmes peuvent être formulés mathématiquement comme des problèmes d'optimisation (chapitre 1) et pose en annexe les bases d'optimisation convexe nécessaires à la compréhension des algorithmes présentés par la suite. La majeure partie de ce livre concerne les problèmes d'apprentissage supervisé ; le chapitre 2 détaille plus particulièrement leur formulation et introduit les notions d'espace des hypothèses, de risque et perte, et de généralisation. Avant d'étudier les algorithmes d'apprentissage supervisé les plus classiques et fréquemment utilisés, il est essentiel de comprendre comment évaluer un modèle sur un jeu de données, et de savoir sélectionner le meilleur modèle parmi plusieurs possibilités, ce qui est le sujet du chapitre 3.

Il est enfin pertinent à ce stade d'aborder l'entraînement de modèles prédictifs supervisés. Le livre aborde tout d'abord les modèles paramétriques, dans lesquels la fonction modélisant la distribution des données ou permettant de faire des prédictions a une forme analytique explicite. Des éléments d'apprentissage bayésien, présentés dans le chapitre 4, complémenté par une annexe qui pose les bases de l'estimation ponctuelle, sont ensuite appliqués à des modèles d'apprentissage supervisé paramétriques (chapitre 5). Le chapitre 6 présente les variantes régularisées de ces algorithmes. Enfin, le chapitre 7 sur les réseaux de neurones propose de construire des modèles paramétriques beaucoup plus complexes et d'aborder les bases du deep learning.

Le livre aborde ensuite les modèles non paramétriques, à commencer par une des plus intuitives de ces approches, la méthode des plus proches voisins (chapitre 8). Suivront ensuite les approches à base d'arbres de décision, puis les méthodes à ensemble qui permettront d'introduire deux des algorithmes de machine learning supervisé les plus puissants à l'heure actuelle : les forêts aléatoires et le boosting de gradient (chapitre 9). Le chapitre 10 sur les méthodes à noyaux, introduites grâce aux machines à vecteurs de support, permettra de voir comment construire des modèles non linéaires via des modèles linéaires dans un espace de redescription des données.

Enfin, le chapitre 11 présentera la réduction de dimension, supervisée ou non-supervisée, et le chapitre 12 traitera d'un des problèmes les plus importants en apprentissage non supervisé : le clustering.

Chaque chapitre sera conclu par quelques exercices.

Comment lire ce livre : Ce livre a été conçu pour être lu linéairement. Cependant, après les trois premiers chapitres, il vous sera possible de lire les suivants dans l'ordre qui vous conviendra, à l'exception du chapitre 6, qui a été écrit dans la continuité du chapitre 5. De manière générale, des références vers les sections d'autres chapitres apparaîtront si nécessaire.

Remerciements : Cet ouvrage n'aurait pas vu le jour sans Jean-Philippe Vert, qui m'a fait découvrir le machine learning, avec qui j'ai enseigné et pratiqué cette discipline pendant plusieurs années, et qui m'a fait, enfin, l'honneur d'une relecture attentive.

Ce livre doit beaucoup aux personnes qui m'ont enseigné le machine learning, et plus particulièrement Pierre Baldi, Padhraic Smyth, et Max Welling ; à celles avec qui je l'ai pratiqué, notamment les membres du Baldi Lab à UC Irvine, du MLCB et du département d'inférence empirique de l'Institut Max Planck à Tübingen, et du CBIO à Mines Paris, et bien d'autres encore qu'il serait difficile de nommer exhaustivement ici ; à celles avec qui je l'ai enseigné, Karsten Borgwardt, Yannis Chaouche, Frédéric Guyon, Fabien Moutarde, mais aussi Judith Abecassis, Eugene Belilovsky, Joseph Boyd, Peter Naylor, Benoît Playe, Mihir Sahasrabudhe, Jiaqian Yu, et Luc Bertrand ; et enfin à celles auxquelles je l'ai enseigné, en particulier dans le cadre du cours Data Mining in der Bioinformatik de l'université de Tübingen (ma toute première tentative d'enseignement des méthodes à noyaux en 2012 !) et les élèves de Centrale qui ont essuyé les plâtres de la première version de ce cours à l'automne 2015.

Mes cours sont le résultat de nombreuses sources d'inspirations accumulées au cours des années. Je remercie tout particulièrement Ethem Alpaydin, David Barber, Christopher M. Bishop, Stephen Boyd, Hal Daumé III, Jerome Friedman, Trevor Hastie, Tom Mitchell, Bernhard Schölkopf, Alex Smola, Robert Tibshirani, Lieven Vandenberghe, et Alice Zhang pour leurs ouvrages.

Parce que tout serait différent sans scikit-learn, je remercie chaleureusement tous ses core-devs, et en particulier Alexandre Gramfort, Olivier Grisel, Gaël Varoquaux et Nelle Varoquaux.

Je remercie aussi Matthew Blaschko, qui m'a poussée à l'eau, et Nikos Paragios, qui l'y a encouragé.

Parce que je n'aurais pas pu écrire ce livre seule, merci à Jean-Luc Blanc et Matthieu Daniel des éditions Dunod, et à celles et ceux qui ont relu tout ou partie de cet ouvrage, en particulier Judith Abecassis, Luc Bertrand, Caroline Petitjean, Denis Rousselle, Erwan Scornet.

La relecture attentive de Jean-Marie Monier, ainsi que les commentaires d'Antoine Brault, ont permis d'éliminer de nombreuses coquilles et approximations de la première version de ce texte.

Merci à Alix Deleporte, enfin, pour ses relectures et son soutien.

La deuxième et la troisième édition de cet ouvrage se sont particulièrement enrichies de mes enseignements à Mines Paris - PSL, de mes interactions avec les élèves et des cours donnés avec, en plus de certaines des personnes suscitées, Katia Antonenko, Thomas Bonte, Alice Blondel, Jesus Bujalance Martin, Julie Cartier, Lucia Clarotto, Nicolas Desassis, Paul Etheimer, Thibault Faney, Adeline Fermanian, Bruno Figliuzzi, Joseph Gesnouin, Vivien Goepp, Gwenn Guichaoua, Arthur Imbert, Victor Laigle, Tristan Lazard, Matthieu Najm, Asma Nouira, Thibaud Martinez, Romain Ménégaux, Thomas Walter, et Daniel Zyss. Merci de vos questions, de vos

commentaires et de ce que vous m'avez appris. Merci aussi à Stéphane Canu, Laure Reboul et Joseph Salmon pour les documents mis en ligne, sur lesquels je me suis appuyée pour compléter cette nouvelle édition. Merci enfin à Christopher Shirley pour les coquilles et imprécisions qu'il m'a signalées.

1 Présentation du machine learning

INTRODUCTION

Le machine learning est un domaine captivant. Issu de nombreuses disciplines comme la statistique, l'optimisation, l'algorithmique ou le traitement du signal, c'est un champ d'études en mutation constante qui s'est maintenant imposé dans notre société. Déjà utilisé depuis des décennies dans la reconnaissance automatique de caractères ou les filtres anti-spam, il sert maintenant à protéger contre la fraude bancaire, recommander des livres, films, ou autres produits adaptés à nos goûts, identifier les visages dans le viseur de notre appareil photo, ou traduire automatiquement des textes d'une langue vers une autre.

Dans les années à venir, le machine learning nous permettra vraisemblablement d'améliorer la sécurité routière (y compris grâce aux véhicules autonomes), la réponse d'urgence aux catastrophes naturelles, le développement de nouveaux médicaments, ou l'efficacité énergétique de nos bâtiments et industries.

Le but de ce chapitre est d'établir plus clairement ce qui relève ou non du machine learning, ainsi que des branches de ce domaine dont cet ouvrage traitera.

OBJECTIFS

- Définir le machine learning.
- Identifier si un problème relève ou non du machine learning.
- Donner des exemples de cas concrets relevant de grandes classes de problèmes de machine learning.

1.1 Qu'est-ce que le machine learning ?

Qu'est-ce qu'apprendre, comment apprend-on, et que cela signifie-t-il pour une machine ? La question de l'*apprentissage* fascine les spécialistes de l'informatique et des mathématiques tout autant que neurologues, pédagogues, philosophes ou artistes.

Une définition qui s'applique à un programme informatique comme à un robot, un animal de compagnie ou un être humain est celle proposée par Fabien Benureau (2015) : « *L'apprentissage est une modification d'un comportement sur la base d'une expérience* ».

Dans le cas d'un programme informatique, qui est celui qui nous intéresse dans cet ouvrage, on parle d'*apprentissage automatique*, ou *machine learning*, quand ce programme a la capacité de se modifier lui-même sans que cette modification ne soit explicitement programmée. Cette définition est celle donnée par Arthur Samuel (1959). On peut ainsi opposer un programme *classique*, qui utilise une procédure

et les données qu'il reçoit en entrée pour produire en sortie des réponses, à un programme *d'apprentissage automatique*, qui utilise les données et les réponses afin de produire la procédure qui permet d'obtenir les secondes à partir des premières.

> ***Exemple***
>
> Supposons qu'une entreprise veuille connaître le montant total dépensé par un client ou une cliente à partir de ses factures. Il suffit d'appliquer un algorithme classique, à savoir une simple addition : un algorithme d'apprentissage n'est pas nécessaire.
>
> Supposons maintenant que l'on veuille utiliser ces factures pour déterminer quels produits le client est le plus susceptible d'acheter dans un mois. Bien que cela soit vraisemblablement lié, nous n'avons manifestement pas toutes les informations nécessaires pour ce faire. Cependant, si nous disposons de l'historique d'achat d'un grand nombre d'individus, il devient possible d'utiliser un algorithme de machine learning pour qu'il en tire un modèle prédictif nous permettant d'apporter une réponse à notre question.

Ce point de vue informatique sur l'apprentissage automatique justifie que l'on considère qu'il s'agit d'un domaine différent de celui de la statistique. Cependant, nous aurons l'occasion de voir que la frontière entre inférence statistique et apprentissage est souvent mince. Il s'agit ici, fondamentalement, de modéliser un phénomène à partir de données considérées comme autant d'observations de celui-ci.

1.1.1 Pourquoi utiliser le machine learning ?

Le machine learning peut servir à résoudre des problèmes

- que l'on ne sait pas résoudre (comme dans l'exemple de la prédiction d'achats ci-dessus) ;
- que l'on sait résoudre, mais dont on ne sait formaliser en termes algorithmiques comment nous les résolvons (c'est le cas par exemple de la reconnaissance d'images ou de la compréhension du langage naturel) ;
- que l'on sait résoudre, mais avec des procédures beaucoup trop gourmandes en ressources informatiques (c'est le cas par exemple de la prédiction d'interactions entre molécules de grande taille, pour lesquelles les simulations sont très lourdes).

Le machine learning est donc utilisé quand les *données* sont abondantes (relativement), mais les *connaissances* peu accessibles ou peu développées.

Ainsi, le machine learning peut aussi aider les humains à apprendre : les modèles créés par des algorithmes d'apprentissage peuvent révéler l'importance relative de certaines informations ou la façon dont elles interagissent entre elles pour résoudre un problème particulier. Dans l'exemple de la prédiction d'achats, comprendre le modèle peut nous permettre d'analyser quelles caractéristiques des achats passés permettent de prédire ceux à venir. Cet aspect du machine learning est très utilisé dans

la recherche scientifique : quels gènes sont impliqués dans le développement d'un certain type de tumeur, et comment ? Quelles régions d'une image cérébrale permettent de prédire un comportement ? Quelles caractéristiques d'une molécule en font un bon médicament pour une indication particulière ? Quels aspects d'une image de télescope permettent d'y identifier un objet astronomique particulier ?

1.1.2 Ingrédients du machine learning

Le machine learning repose sur deux piliers fondamentaux :

- d'une part, les *données*, qui sont les exemples à partir duquel l'algorithme va apprendre ;
- d'autre part, l'*algorithme d'apprentissage*, qui est la procédure que l'on fait tourner sur ces données pour produire un modèle. On appelle *entraînement* le fait de faire tourner un algorithme d'apprentissage sur un jeu de données.

Ces deux piliers sont aussi importants l'un que l'autre. D'une part, aucun algorithme d'apprentissage ne pourra créer un bon modèle à partir de données qui ne sont pas pertinentes – c'est le concept *garbage in, garbage out* qui stipule qu'un algorithme d'apprentissage auquel on fournit des données de mauvaise qualité ne pourra rien en faire d'autre que des prédictions de mauvaise qualité. D'autre part, un modèle appris avec un algorithme inadapté sur des données pertinentes ne pourra pas être de bonne qualité.

Cet ouvrage est consacré au deuxième de ces piliers – les algorithmes d'apprentissage. Néanmoins, il ne faut pas négliger qu'une part importante du travail de *machine learner* ou de *data scientist* est un travail d'ingénierie consistant à préparer les données afin d'éliminer les données aberrantes, gérer les données manquantes, choisir une représentation pertinente, etc.

Bien que l'usage soit souvent d'appeler les deux du même nom, il faut distinguer l'*algorithme d'apprentissage* automatique du *modèle appris* : le premier utilise les données pour produire le second, qui peut ensuite être appliqué comme un programme classique.

Un algorithme d'apprentissage permet donc de *modéliser* un phénomène à partir d'*exemples*. Nous considérons ici qu'il faut pour ce faire définir et *optimiser* un objectif. Il peut par exemple s'agir de minimiser le nombre d'erreurs faites par le modèle sur les exemples d'apprentissage. Cet ouvrage présente en effet les algorithmes les plus classiques et les plus populaires sous cette forme.

Exemple

Voici quelques exemples de reformulation de problèmes de machine learning sous la forme d'un problème d'optimisation. La suite de cet ouvrage devrait vous éclairer sur la formalisation mathématique de ces problèmes, formulés ici très librement.

- Un site marchand peut chercher à *modéliser* des types représentatifs de clientèle, à partir des transactions passées, en *maximisant* la proximité entre clients et clientes affectés à un même type.

- Une compagnie automobile peut chercher à *modéliser* la trajectoire d'un véhicule dans son environnement, à partir d'enregistrements vidéo de voitures, en *minimisant* le nombre d'accidents.
- Des chercheuses en génétique peuvent vouloir *modéliser* l'impact d'une mutation sur une maladie, à partir de données patient, en *maximisant* la cohérence de leur modèle avec les connaissances de l'état de l'art.
- Une banque peut vouloir *modéliser* les comportements à risque, à partir de son historique, en *maximisant* le taux de détection de non-solvabilité.

Ainsi, le machine learning repose d'une part sur les mathématiques, et en particulier la statistique, pour ce qui est de la construction de modèles et de leur inférence à partir de données, et d'autre part sur l'informatique, pour ce qui est de la représentation des données et de l'implémentation efficace d'algorithmes d'optimisation. De plus en plus, les quantités de données disponibles imposent de faire appel à des architectures de calcul et de base de données distribuées. C'est un point important mais que nous n'abordons pas dans cet ouvrage.

1.1.3 Et l'intelligence artificielle, dans tout ça ?

Le machine learning peut être vu comme une branche de l'intelligence artificielle. En effet, un système incapable d'apprendre peut difficilement être considéré comme intelligent. La capacité à apprendre et à tirer parti de ses expériences est en effet essentielle à un système conçu pour s'adapter à un environnement changeant.

L'intelligence artificielle, définie comme l'ensemble des techniques mises en œuvre afin de construire des machines capables de faire preuve d'un comportement que l'on peut qualifier d'intelligent, fait aussi appel aux sciences cognitives, à la neurobiologie, à la logique, à l'électronique, à l'ingénierie et bien plus encore.

Probablement parce que le terme « intelligence artificielle » stimule plus l'imagination, il est cependant de plus en plus souvent employé en lieu et place de celui d'apprentissage automatique.

1.1.4 Questions de société

L'essor récent de l'intelligence artificielle, en particulier à travers les progrès du machine learning, suscite de vifs débats philosophiques, éthiques et moraux. Les biais algorithmiques, et en particulier les biais de l'intelligence artificielle, sont le sujet d'inquiétudes profondes, certaines hélas bien fondées, d'autres plutôt fantasmées.

Les biais algorithmiques font régulièrement les gros titres de la presse et sont un des sujets du règlement européen sur l'intelligence artificielle « AI Act » du 12 juillet 2024. L'importance des ressources computationnelles nécessaires à la mise en œuvre de certains outils, notamment les robots conversationnels basés sur des grands modèles de langage, soulève des questions écologiques. La provenance des données est, elle aussi, source de questionnements : les personnes qu'elles concernent ou les ayant produites, parfois à leur insu, consentent-elles à l'utilisation qui en est faite ? A-t-on recouru à des travailleurs et travailleuses du clic pour les générer ou les annoter ?