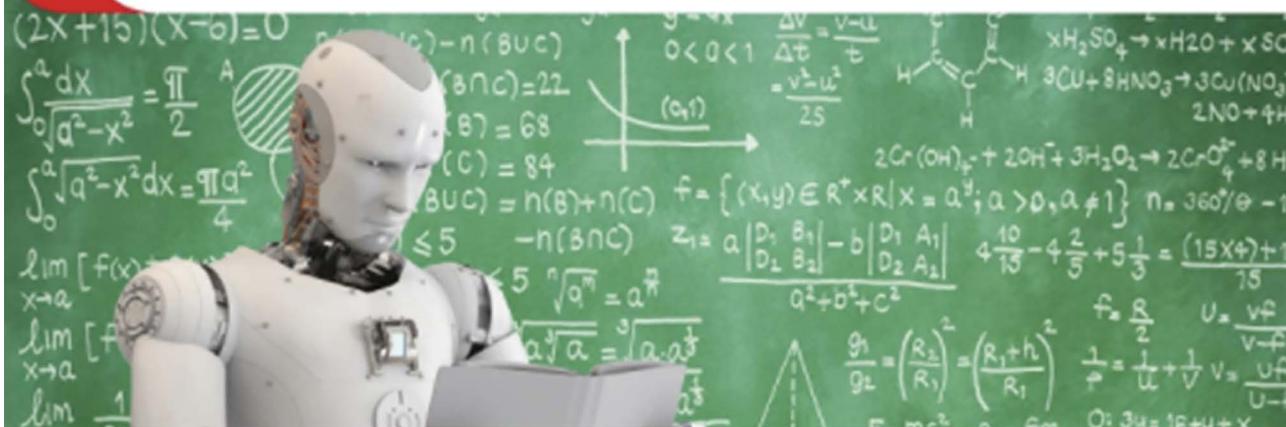


Introduction au Machine Learning



Chloé-Agathe Azencott

DUNOD

Introduction au Machine Learning

Chez le même éditeur

Big Data et Machine Learning

3^e édition

Pirmin Lemberger, Marc Batty, Médéric Morel, Jean-Luc Raffaëlli

272 pages

Dunod, 2019

Machine Learning avec Scikit-Learn

2^e édition

Aurélien Géron

288 pages environ

Dunod, 2019

Deep Learning avec Keras et TensorFlow

2^e édition

Aurélien Géron

392 pages environ

Dunod, 2019

Introduction au Machine Learning

Chloé-Agathe Azencott

Maîtresse de conférences au CBIO (Centre de bio-informatique)
de MINES ParisTech, de l’Institut Curie et de l’INSERM

DUNOD



Illustration de couverture : © PhonlamaiPhoto - iStock

© Dunod, 2018
Nouvelle présentation avec corrections 2019
11 rue Paul Bert, 92240 Malakoff
www.dunod.com
ISBN 978-2-10-080450-4

Le Code de la propriété intellectuelle n'autorisant, aux termes de l'article L. 122-5, 2^e et 3^e a), d'une part, que les « copies ou reproductions strictement réservées à l'usage privé du copiste et non destinées à une utilisation collective » et, d'autre part, que les analyses et les courtes citations dans un but d'exemple et d'illustration, « toute représentation ou reproduction intégrale ou partielle faite sans le consentement de l'auteur ou de ses ayants droit ou ayants cause est illicite » (art. L. 122-4).

Cette représentation ou reproduction, par quelque procédé que ce soit, constituerait donc une contrefaçon sanctionnée par les articles L. 335-2 et suivants du Code de la propriété intellectuelle.

PRÉAMBULE

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 objective possible les performances de plusieurs algorithmes de machine learning pour une application particulière.

Public visé : Ce livre s'adresse à des étudiants en informatique ou maths 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 (voir par exemple <http://tinyurl.com/ma2823-2017>) et sur OpenClassrooms (voir <https://openclassrooms.com/fr/paths/148-ingeneur-machine-learning>) et suppose les prérequis suivants :

Préambule

- 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. Les bases sont posées avec des éléments d'inférence bayésienne (chapitre 4), qui seront 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). Sui-vront 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.

Préambule

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 à ceux qui m'ont enseigné le machine learning, et plus particulièrement Pierre Baldi, Padhraic Smyth, et Max Welling ; à ceux 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 ParisTech, et bien d'autres encore qu'il serait difficile de tous nommer ici ; à ceux 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 ceux auxquels je l'ai enseigné, en particulier les étudiants du cours Data Mining in der Bioinformatik de l'université de Tübingen qui ont subi ma toute première tentative d'enseignement des méthodes à noyaux en 2012, et les étudiants centraliens qui ont essayé 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 des éditions Dunod, et à tous 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.

TABLE DES MATIÈRES

CHAPITRE 1 • PRÉSENTATION DU MACHINE LEARNING	1
1.1 Qu'est-ce que le machine learning ?	1
1.2 Types de problèmes de machine learning	4
1.3 Ressources pratiques	10
1.4 Notations	11
Exercices	12
Solutions	13
CHAPITRE 2 • APPRENTISSAGE SUPERVISÉ	14
2.1 Formalisation d'un problème d'apprentissage supervisé	14
2.2 Espace des hypothèses	17
2.3 Minimisation du risque empirique	18
2.4 Fonctions de coût	20
2.5 Généralisation et sur-apprentissage	26
Exercices	31
Solutions	32
CHAPITRE 3 • SÉLECTION DE MODÈLE ET ÉVALUATION	33
3.1 Estimation empirique de l'erreur de généralisation	33
3.2 Critères de performance	37
Exercices	47
Solutions	48
CHAPITRE 4 • INFÉRENCE BAYÉSIENNE	49
4.1 Modèles génératifs pour la classification binaire	49
4.2 Règles de décision	52
4.3 Estimation de densité	58
4.4 Classification naïve bayésienne	62
4.5 Sélection de modèle bayésien	64
Exercices	65
Solutions	67

Table des matières

CHAPITRE 5 • RÉGRESSIONS PARAMÉTRIQUES	70
5.1 Apprentissage supervisé d'un modèle paramétrique	70
5.2 Régression linéaire	72
5.3 Régression logistique	75
5.4 Régression polynomiale	77
Exercices	78
Solutions	79
CHAPITRE 6 • RÉGULARISATION	81
6.1 Qu'est-ce que la régularisation ?	81
6.2 La régression ridge	82
6.3 Le lasso	85
6.4 Elastic net	88
Exercices	90
Solutions	91
CHAPITRE 7 • RÉSEAUX DE NEURONES ARTIFICIELS	92
7.1 Le perceptron	92
7.2 Perceptron multi-couche	98
Exercices	106
Solutions	106
CHAPITRE 8 • MÉTHODE DES PLUS PROCHES VOISINS	109
8.1 Méthode du plus proche voisin	109
8.2 Méthode des plus proches voisins	111
8.3 Distances et similarités	114
8.4 Filtrage collaboratif	119
Exercices	121
Solutions	122
CHAPITRE 9 • ARBRES ET FORÊTS	124
9.1 Arbres de décision	124
9.2 Comment faire pousser un arbre	126
9.3 Méthodes ensemblistes : la sagesse des foules	130
Exercices	137
Solutions	138

Table des matières

CHAPITRE 10 • MACHINES À VECTEURS DE SUPPORT ET MÉTHODES À NOYAUX.....	140
10.1 Le cas linéairement séparable : SVM à marge rigide	140
10.2 Le cas linéairement non séparable : SVM à marge souple ..	145
10.3 Le cas non linéaire : SVM à noyau.....	149
10.4 Régression ridge à noyau	154
Exercices.....	158
Solutions	160
CHAPITRE 11 • RÉDUCTION DE DIMENSION	164
11.1 Motivation	164
11.2 Sélection de variables	166
11.3 Extraction de variables	169
Exercices.....	182
Solutions	183
CHAPITRE 12 • CLUSTERING	185
12.1 Pourquoi partitionner ses données	185
12.2 Évaluer la qualité d'un algorithme de clustering	186
12.3 Clustering hiérarchique.....	190
12.4 Méthode des K-moyennes.....	193
12.5 Clustering par densité	196
Exercices.....	200
Solutions	201
APPENDICE • NOTIONS D'OPTIMISATION CONVEXE	203
A.1 Convexité	203
A.2 Problèmes d'optimisation convexe	205
A.3 Optimisation convexe sans contrainte	207
A.4 Optimisation convexe sous contraintes	218
INDEX	225

PRÉSENTATION DU MACHINE LEARNING

INTRODUCTION

Le machine learning est un domaine captivant. Issu de nombreuses disciplines comme les statistiques, l'optimisation, l'algorithme 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 d'apprendre 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

Chapitre 1 · Présentation du machine learning

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.

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 Qu'est-ce que le machine learning ?

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 vendeur en ligne 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.

Chapitre 1 · Présentation du machine learning

- Des chercheurs 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 les statistiques, 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.

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.

1.2 TYPES DE PROBLÈMES DE MACHINE LEARNING

Le machine learning est un champ assez vaste, et nous dressons dans cette section une liste des plus grandes classes de problèmes auxquels il s'intéresse.

1.2.1 Apprentissage supervisé

L'apprentissage supervisé est peut-être le type de problèmes de machine learning le plus facile à appréhender : son but est d'apprendre à faire des *predictions*, à partir d'une liste d'exemples *étiquetés*, c'est-à-dire accompagnés de la valeur à prédire (voir figure 1.1). Les étiquettes servent de « professeur » et supervisent l'apprentissage de l'algorithme.

1.2 Types de problèmes de machine learning

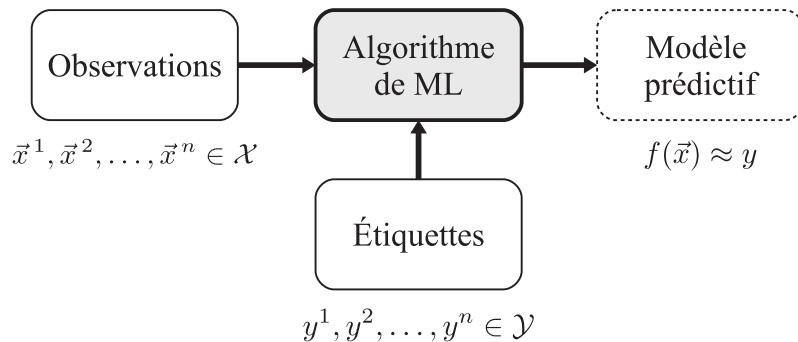


Figure 1.1 – Apprentissage supervisé.

Définition 1.1 (Apprentissage supervisé)

On appelle *apprentissage supervisé* la branche du machine learning qui s'intéresse aux problèmes pouvant être formalisés de la façon suivante : étant données n observations $\{\vec{x}^i\}_{i=1,\dots,n}$ décrites dans un espace \mathcal{X} , et leurs étiquettes $\{y^i\}_{i=1,\dots,n}$ décrites dans un espace \mathcal{Y} , on suppose que les étiquettes peuvent être obtenues à partir des observations grâce à une fonction $\phi : \mathcal{X} \rightarrow \mathcal{Y}$ fixe et inconnue : $y^i = \phi(\vec{x}^i) + \epsilon_i$, où ϵ_i est un bruit aléatoire. Il s'agit alors d'utiliser les données pour déterminer une fonction $f : \mathcal{X} \rightarrow \mathcal{Y}$ telle que, pour tout couple $(\vec{x}, \phi(\vec{x})) \in \mathcal{X} \times \mathcal{Y}$, $f(\vec{x}) \approx \phi(\vec{x})$.

L'espace sur lequel sont définies les données est le plus souvent $\mathcal{X} = \mathbb{R}^p$. Nous verrons cependant aussi comment traiter d'autres types de représentations, comme des variables binaires, discrètes, catégoriques, voire des chaînes de caractères ou des graphes.

Classification binaire

Dans le cas où les étiquettes sont *binaires*, elles indiquent l'appartenance à une *classe*. On parle alors de *classification binaire*.

Définition 1.2 (Classification binaire)

Un problème d'apprentissage supervisé dans lequel l'espace des étiquettes est binaire, autrement dit $\mathcal{Y} = \{0, 1\}$ est appelé un problème de *classification binaire*.

Exemple

Voici quelques exemples de problèmes de classification binaire :

- Identifier si un email est un spam ou non.
- Identifier si un tableau a été peint par Picasso ou non.

Chapitre 1 · Présentation du machine learning

- Identifier si une image contient ou non une girafe.
- Identifier si une molécule peut ou non traiter la dépression.
- Identifier si une transaction financière est frauduleuse ou non.

Classification multi-classe

Dans le cas où les étiquettes sont *discrètes*, et correspondent donc à plusieurs¹ *classes*, on parle de *classification multi-classe*.

Définition 1.3 (Classification multi-classe)

Un problème d'apprentissage supervisé dans lequel l'espace des étiquettes est discret et fini, autrement dit $\mathcal{Y} = \{1, 2, \dots, C\}$ est appelé un problème de *classification multi-classe*. C est le nombre de classes.

Exemple

Voici quelques exemples de problèmes de classification multi-classe :

- Identifier en quelle langue un texte est écrit.
- Identifier lequel des 10 chiffres arabes est un chiffre manuscrit.
- Identifier l'expression d'un visage parmi une liste prédéfinie de possibilités (colère, tristesse, joie, etc.).
- Identifier à quelle espèce appartient une plante.
- Identifier les objets présents sur une photographie.

Régression

Dans le cas où les étiquettes sont à valeurs *réelles*, on parle de *régression*.

Définition 1.4 (Régression)

Un problème d'apprentissage supervisé dans lequel l'espace des étiquettes est $\mathcal{Y} = \mathbb{R}$ est appelé un problème de *régression*.

Exemple

Voici quelques exemples de problèmes de régression :

- Prédire le nombre de clics sur un lien.
- Prédire le nombre d'utilisateurs et utilisatrices d'un service en ligne à un moment donné.
- Prédire le prix d'une action en bourse.
- Prédire l'affinité de liaison entre deux molécules.
- Prédire le rendement d'un plant de maïs.

1. Nous utilisons ici la définition bourbakiste de « plusieurs », c'est-à-dire strictement supérieur à deux.

1.2 Types de problèmes de machine learning

Régression structurée

Dans le cas où l'espace des étiquettes est un espace structuré plus complexe que ceux évoqués précédemment, on parle de *régression structurée* – en anglais, *structured regression*, ou *structured output prediction*. Il peut par exemple s'agir de prédire des vecteurs, des images, des graphes, ou des séquences. La régression structurée permet de formaliser de nombreux problèmes, comme ceux de la traduction automatique ou de la reconnaissance vocale (text-to-speech et speech-to-text, par exemple). Ce cas dépasse cependant le cadre du présent ouvrage, et nous nous concentrerons sur les problèmes de classification binaire et multi-classe, ainsi que de régression classique.

L'apprentissage supervisé est le sujet principal de cet ouvrage, et sera traité du chapitre 2 au chapitre 9.

1.2.2 Apprentissage non supervisé

Dans le cadre de l'apprentissage *non supervisé*, les données ne sont pas étiquetées. Il s'agit alors de modéliser les observations pour mieux les comprendre (voir figure 1.2).



Figure 1.2 – Apprentissage non supervisé.

Définition 1.5 (Apprentissage non supervisé)

On appelle *apprentissage non supervisé* la branche du machine learning qui s'intéresse aux problèmes pouvant être formalisés de la façon suivante : étant données n observations $\{\vec{x}^i\}_{i=1,\dots,n}$ décrites dans un espace \mathcal{X} , il s'agit d'apprendre une fonction sur \mathcal{X} qui vérifie certaines propriétés.

Cette définition est très vague, et sera certainement plus claire sur des exemples.

Clustering

Tout d'abord, le *clustering*, ou *partitionnement*, consiste à identifier des *groupes* dans les données (voir figure 1.3). Cela permet de comprendre leurs caractéristiques générales, et éventuellement d'inférer les propriétés d'une observation en fonction du groupe auquel elle appartient.

Définition 1.6 (Partitionnement)

On appelle *partitionnement* ou *clustering* un problème d'apprentissage non supervisé pouvant être formalisé comme la recherche d'une partition $\bigcup_{k=1}^K \mathcal{C}_k$ des n observations $\{\vec{x}^i\}_{i=1,\dots,n}$. Cette partition doit être pertinente au vu d'un ou plusieurs critères à préciser.