Algorithmes

2e édition

Massih-Reza AMINI

Préface de Francis Bach

Machine Learning

Programmes libres (GPLv3) essentiels au développement de solutions big data

Éditions EYROLLES

2e édition

Machine Learning

Machine Learning et intelligence artificielle

Le Machine Learning est l'un des domaines de l'intelligence artificielle qui a pour but de concevoir des programmes qui ne sont pas explicitement codés pour s'acquitter d'une tâche particulière. Les concepts de ce domaine sont fondés sur la logique inférentielle et tentent de dégager des règles générales à partir d'un nombre fini d'observations.

Un ouvrage de référence

Cet ouvrage présente les fondements scientifiques de la théorie de l'apprentissage supervisé, les algorithmes les plus répandus développés suivant ce domaine ainsi que les deux cadres de l'apprentissage semi-supervisé et de l'ordonnancement, à un niveau accessible aux étudiants de master et aux élèves ingénieurs. La première édition, connue sous le nom *Apprentissage machine*, fut traduite en chinois par les éditions iTuring. Dans cette deuxième édition, un nouveau chapitre est dédié au *Deep Learning*, sur les réseaux de neurones artificiels, et nous avons réorganisé les autres chapitres pour un exposé cohérent reliant la théorie aux algorithmes développés dans cette sphère. Vous trouverez également dans cette édition quelques programmes des algorithmes classiques, écrits en langages Python et C (langages à la fois simples et populaires), et à destination des lecteurs qui souhaitent connaître le fonctionnement de ces modèles désignés parfois comme des boîtes noires. Ces programmes libres (GPLv3) essentiels au développement de solutions big data sont déposés progressivement sur ce gitlab (https://gricad-gitlab.univ-grenoble-alpes.fr/aminima/machine-learning-tools).

À qui s'adresse ce livre ?

- Aux élèves ingénieurs, étudiants de master et doctorants en mathématiques appliquées, algorithmique, recherche opérationnelle, gestion de production, aide à la décision.
- Aux ingénieurs, enseignants-chercheurs, informaticiens, industriels, économistes et décideurs ayant à résoudre des problèmes de classification, de partitionnement et d'ordonnancement à large échelle.

Massih-Reza Amini, *professeur d'informatique à l'université Grenoble Alpes (UGA), est titulaire d'une thèse sur l'étude de nouveaux cadres et modèles d'apprentissage statistiques pour les nouvelles applications émergentes issues d'Internet. Il est co-auteur de plus d'une centaine d'articles scientifiques parus parmi les actes de conférences et de revues les plus prestigieux des domaines de l'apprentissage automatique et de la recherche d'information. Il est également co-auteur des ouvrages* Recherche d'information *et* Data science *parus aux éditions Eyrolles.*

Code éditeur : G67947
ISBN : 978-2-212-67947-2

Conception graphique : Tai-Marc Le Thanh

Éditions EYROLLES

www.editions-eyrolles.com
Éditions Eyrolles | Diffusion Geodif

Machine Learning

Programmes libres (GPLv3) essentiels au développement de solutions big data

Sur le même thème

P. Bosc, M. Guyomard, L. Miclet. – **Conception d'algorithmes.**
N°67728, 2e édition, 2019, 852 pages.

M.-R. Amini, R. Blanch, M.Clausel, J.-B. Durand, E. Gaussier, J. Malick, C.Picard, C. Quema, G. Quenot. – **Data science : cours et exercices.**
N°67410, 2018, 272 pages.

A. Cornuejols, L. Miclet. – **Apprentissage artificiel.**
N°67522, 3e édition, 2018, 912 pages.

E. Biernat, M. Lutz. – **Data science : fondamentaux et études de cas.**
N°14243, 2015, 296 pages.

W. Mckinney. – **Analyse de données en Python.**
N°14109, 2015, 488 pages.

M.-R. Amini, E. Gaussier. – **Recherche d'information.**
N°13532, 2013, 234 pages.

R. Bruchez. – **Les bases de données NoSQL et le Big Data.**
N°14155, 2015, 322 pages.

Massih-Reza Amini

Préface de Francis Bach

Machine Learning

Programmes libres (GPLv3) essentiels au développement de solutions big data

2e édition

Éditions EYROLLES

ÉDITIONS EYROLLES
61, bd Saint-Germain
75240 Paris Cedex 05
www.editions-eyrolles.com

 ISBN : 978-2-212-67947-2

Préface

Depuis quelques années, dans les domaines scientifiques, industriels et personnels, la presence de données numériques et leur utilisation ont explosé. Certaines de ces données sont massives, nécessitent des outils et architectures spécifiques, comme en astronomie ou pour les moteurs de recherche, et constituent les problèmes dits de «big data».

D'autres données ne sont pas si massives, comme les photos ou vidéos familiales, mais constituent toujours un défi algorithmique. Le grand changement récent est non seulement la taille, mais aussi le côté omniprésent de ces données, qui sont utilisées quotidiennement.

Depuis une vingtaine d'années, l'apprentissage statistique («machine learning» en anglais) s'est considérablement développé, à l'interface entre l'informatique et les statistiques, et constitue le cœur méthodologique des algorithmes modernes de traitement de données. Même si les recherches en apprentissage sont toujours en plein essor, un socle méthodologique et algorithmique a émergé.

Ce livre constitue une introduction équilibrée aux concepts et outils les plus importants de l'apprentissage supervisé et de ses extensions. Un accent remarquable est mis sur des résultats théoriques élégants, simples mais puissants, des algorithmes efficaces qui ont montré leurs preuves en pratique, et des codes permettant de reproduire les expériences.

Francis Bach

octobre 2014

Table des matières

Annexe B

Table des figures

Liste des algorithmes

Avant-propos

L'apprentissage machine est l'un des domaines phares de l'intelligence artificielle. Il concerne l'étude et le développement de modèles quantitatifs permettant à un ordinateur d'accomplir des tâches sans qu'il soit explicitement programmé à les faire. Apprendre dans ce contexte revient à reconnaître des formes complexes et à prendre des décisions intelligentes. Compte tenu de toutes les entrées existantes, la difficulté d'accomplir cette tâche réside dans le fait que l'ensemble des décisions possibles est généralement très complexe à énumérer. Pour contourner cette difficulté, les algorithmes en apprentissage machine ont été conçus dans le but d'acquérir de la connaissance sur le problème à traiter en se basant sur un ensemble de données limitées issues de ce problème.

Concepts étudiés

Pour illustrer ce principe, considérons le cadre de l'apprentissage supervisé que nous allons en partie traiter dans cet ouvrage. Suivant ce cadre, la décision à prendre sur une entrée donnée est prise d'après la sortie d'une fonction de prédiction qui est inférée en utilisant un ensemble d'exemples étiquetés (ou base d'entraînement), où chacun de ces exemples est une paire constituée du vecteur représentatif d'une observation dans un espace vectoriel donné, et d'une réponse associée à l'exemple (aussi appelée sortie désirée ou sortie réelle). Après la phase d'estimation ou d'apprentissage, la fonction renvoyée par l'algorithme doit permettre de prédire la réponse associée à de nouvelles observations. L'hypothèse sous-jacente dans ce cas est que les exemples sont, d'une façon générale, représentatifs du problème de prédiction sur lequel la fonction sera appliquée. En pratique, une fonction d'erreur mesure l'écart entre la prédiction du modèle sur un exemple et sa sortie désirée. À partir d'un ensemble d'entraînement donné, l'algorithme d'apprentissage choisit alors une fonction, issue d'un ensemble de fonctions défini au préalable, qui réalise l'erreur moyenne la plus faible sur les exemples de la base d'entraînement.

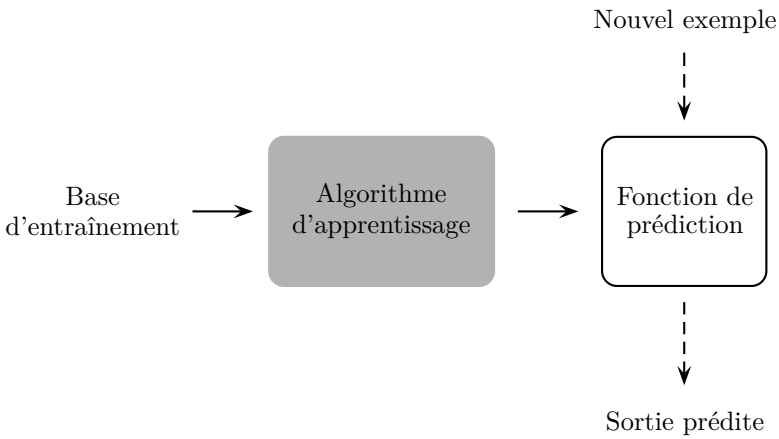

Figure I - Illustration des deux phases d'un problème d'apprentissage supervisé. Dans la phase d'apprentissage (schématisée par les traits pleins), une fonction minimisant l'erreur empirique sur une base d'entraînement est trouvée parmi une classe de fonctions prédéfinies. Dans la phase de test (schématisée par les traits pointillés), les sorties de nouveaux exemples sont prédites par la fonction de prédiction.

Cette erreur n'est généralement pas représentative de la performance de l'algorithme sur de nouveaux exemples. Il est alors nécessaire de disposer d'un second ensemble d'exemples étiquetés, ou une base de test, auxquels l'algorithme n'avait pas accès et d'estimer l'erreur moyenne de la fonction produite lors de la phase d'estimation qui sera cette fois représentative de son erreur de généralisation. On attend de l'algorithme d'apprentissage qu'il trouve une fonction ayant de bonnes performances en généralisation et non celle qui sera capable de reproduire parfaitement les réponses associées aux exemples d'entraînement (voir figure I). Les garanties d'apprenabilité du procédé de la minimisation du risque empirique ont été étudiées dans la théorie de l'apprentissage machine largement initiée par Vapnik (1999). Elles dépendent de la taille de la base d'entraînement et de la complexité de la classe de fonctions où on cherche la fonction de prédiction.

Historiquement, les deux tâches principales du cadre de l'apprentissage supervisé étaient la classification et la régression. Ces tâches sont similaires à la différence de l'espace des sorties désirées des exemples. Dans le cas de la classification, l'espace de sortie est discret alors qu'en régression cet espace est réel.

À la fin des années 1990 et sous l'impulsion de nouvelles technologies, notamment celles liées au développement d'Internet, de nouveaux cadres d'apprentissage ont vu le jour. Un de ces cadres est l'apprentissage avec des données partiellement

étiquetées, ou l'apprentissage semi-supervisé, dont le développement est motivé par l'effort qu'il faut consentir à constituer des bases d'apprentissage étiquetées et le constat que les données étiquetées sont chères à obtenir alors que les données non étiquetées sont foison et qu'elles contiennent de l'information sur le problème que l'on cherche à résoudre. De ce constat sont nés plusieurs travaux qui avaient pour objectif d'employer une petite quantité de données étiquetées, simultanément avec une grande quantité de données non étiquetées, pour apprendre une fonction de prédiction.

L'autre cadre qui a suscité de nombreux travaux dans la communauté d'apprentissage depuis les années 2000 concerne le développement de modèles d'ordonnancement. Ce cadre a formalisé dans un premier temps les problèmes de la Recherche d'Information et il a été par la suite étendu à d'autres problèmes plus généraux.

Depuis de nombreuses années, les algorithmes d'apprentissage développés suivant ces cadres ont été appliqués avec succès à une grande variété de problèmes, incluant la reconnaissance de la parole et de l'écriture manuscrite, la vision par ordinateur, la prédiction de la structure des protéines, les systèmes de recommendations, la classification documentaire, les moteurs de recherche, etc.

Organisation du livre

Cet ouvrage présente les fondements scientifiques de la théorie de l'apprentissage supervisé, les algorithmes les plus répandus développés suivant ce cadre ainsi que les deux cadres de l'apprentissage évoqués plus haut, à un niveau accessible aux étudiants de master et aux élèves ingénieurs. Notre souci a été de proposer un exposé cohérent reliant la théorie aux algorithmes développés dans ce domaine. En outre, cette étude ne se limite pas à l'exposé de ces fondements, mais présente aussi quelques programmes des algorithmes classiques proposés dans ce manuscrit, écrits dans un langage informatique simple et populaire qui est le langage C [1], et à destination des lecteurs qui cherchent à connaître le fonctionnement de ces modèles désignés parfois comme des boîtes noires. Ce livre est organisé en six chapitres principaux et deux annexes. L'enchaînement des idées présentées dans chacun d'eux est le suivant :

- Dans le chapitre 1, nous décrivons les concepts fondamentaux de la théorie de l'apprentissage statistique de Vapnik (1999). Nous exposons la notion de consistance du principe de la minimisation du risque empirique selon lequel la plupart des algorithmes en apprentissage supervisé ont été développés. L'étude de cette consistance nous mènera à l'exposé du second principe fondamental en apprentissage qui est la minimisation du risque structurel, ouvrant le champ au développement de nouveaux modèles en apprentissage machine. En particulier,

1. http://www.tiobe.com/index.php/content/paperinfo/tpci/index.html

nous présentons dans ce chapitre la notion de borne sur l'erreur de généralisation en décrivant les hypothèses et les outils nécessaires pour l'obtenir.

- Dans le chapitre 2, nous allons présenter des bornes sur l'erreur de généralisation qui peuvent être estimées sur une base d'entraînement qui sert à apprendre un modèle. Ces bornes sont basées sur une notion de complexité de classes dépendante des données, appelée complexité de Rademacher. Avec cette notion il est aussi possible de dériver facilement des bornes sur l'erreur de généralisation pour la classification multi-classe. Nous exposons les approches multi-classe basées sur la classification binaire, appelées approches combinées, et dérivons une borne sur l'erreur de généralisation des classifieurs linéaires dans ce cas.
- C'est dans le chapitre 3 que nous nous intéressons aux algorithmes de base issus du domaine de l'optimisation pour la minimisation d'une fonction de risque convexe majorant le risque empirique appelés *algorithmes à direction de descente*. En particulier, nous présentons les conditions nécessaires à vérifier un algorithme à direction de descente pour converger vers le minimiseur d'une fonction objectif convexe et nous décrivons quelques variantes simples et efficaces de cet algorithme.
- Le chapitre 4 présente les principaux modèles qui sont issus des travaux sur la modélisation numérique des neurones du système nerveux, appelés plus communément réseaux de neurones artificiels ou *Deep Learning*. Ces modèles étaient précurseurs au développement des méthodes quantitatives en Intelligence Artificielle et un soin particulier a été porté à la présentation de ces modèles dans leur contexte historique.
- Dans le chapitre 5, nous présentons les séparateurs à vaste marge (SVM) qui sont issus du principe de la minimisation du risque structurel. Ces modèles sont devenus très populaires grâce à leurs justifications théoriques. En particulier, nous verrons comment utiliser l'astuce de noyau pour plonger l'espace d'entrée dans un espace de plus grande dimension dans lequel le problème d'apprentissage devient plus simple à résoudre et nous présentons l'extension de ces au cas multi-classe.
- Le chapitre 6 présente l'algorithme Adaboost issus des travaux de Schapire (1999). Cet algorithme combine plusieurs classifieurs de base, appelés apprenants faibles, pour construire un classifieur final, appelé apprenant fort, qui est plus performant que chacun de ces classifieurs de base. En particulier, nous ferons le lien entre cet algorithme et le principe de la minimisation du risque empirique énoncé par Vapnik (1999). Nous exposons en outre l'extension de cet algorithme au cas multi-classe.
- Nous exposons ensuite le cadre de l'apprentissage semi-supervisé dans le chapitre 7. Nous commençons ce chapitre par présenter les algorithmes EM et CEM développés dans le cadre de l'apprentissage non supervisé, en détaillant quelques cas particuliers menant à des modèles non supervisés bien connus

comme l'algorithme des k-moyennes. Nous présentons ensuite les hypothèses de base en apprentissage semi-supervisé en détaillant les trois approches génératives, discriminantes et graphiques développées suivant ce cadre.

- C'est dans le chapitre 8 que nous décrivons formellement le cadre de l'apprentissage de fonctions d'ordonnancement (ou learning to rank en anglais) en focalisant sur deux formes particulières d'ordonnancement appelées ordonnancement d'alternatives et d'ordonnancement d'instances. Nous exposons ensuite quelques algorithmes développés suivant les approches classiques de l'apprentissage de fonctions d'ordonnancement. Nous terminons ce chapitre par montrer la réduction de quelques problèmes d'ordonnancement à la classification binaire de paires d'observations. Cette réduction ouvre la voie à l'apprentissage de classifieurs avec des exemples interdépendants que nous analysons avec le résultat de Janson (2004).
- En annexe A, nous donnons quelques rappels des outils de base en probabilité que nous employons dans cet ouvrage.
- En annexe B, nous donnons les codes programmes de quinze algorithmes présentés dans les différents chapitres, en détaillant les structures de données utilisées et en liant les différentes parties des programmes aux points correspondants abordés dans cet ouvrage.

Chapitre 1

Principes de base en apprentissage supervisé

Dans ce chapitre, nous exposons la théorie de l'apprentissage machine selon le cadre de Vapnik (1999) qui nous servira de base dans notre description des algorithmes d'apprentissage décrits dans les chapitres suivants. Plus particulièrement, nous présentons la notion de consistance qui garantit l'apprenabilité d'une fonction de prédiction. Les définitions et les hypothèses de base de cette théorie, ainsi que le principe de la minimisation du risque empirique, sont décrits dans la section 1.1. L'étude de la consistance de ce principe, présentée dans la section 1.2, nous mène au second principe de la minimisation du risque structurel, qui stipule que l'apprentissage est un compromis entre une erreur empirique faible et une capacité de la classe de fonctions forte.

Un modèle d'apprentissage construit une fonction de prédiction à partir d'un ensemble fini d'exemples, appelé base d'entraînement ou base d'apprentissage (Fukunaga 1972 ; Duda *et al.* 2001 ; Schölkopf et Smola 2002 ; Boucheron *et al.* 2005). Suivant le cadre supervisé, chaque exemple est un couple constitué généralement du vecteur représentatif d'une observation et de sa réponse associée (aussi appelée sortie désirée). Le but de l'apprentissage est d'induire une fonction qui prédise les réponses associées à de nouvelles observations en commettant une erreur de prédiction la plus faible possible. Cette réponse est généralement une valeur réelle ou une étiquette de classe, comme nous allons le voir dans la suite. L'hypothèse sous-jacente ici est que les données sont stationnaires, c'est-à-dire que les exemples de la base d'entraînement, sur laquelle la fonction de prédiction est apprise, sont en quelque sorte représentatifs du problème général que l'on souhaite résoudre. Nous allons revenir sur cette hypothèse dans la section suivante.

En pratique, parmi une classe de fonctions existante, le modèle d'apprentissage choisit la fonction qui réalise la plus faible erreur moyenne de prédiction (ou erreur empirique) sur une base d'entraînement. La fonction d'erreur quantifie le désaccord entre la prédiction de sortie donnée par la fonction que l'on souhaite apprendre pour une observation de la base d'entraînement et sa réponse associée. Le but de cette recherche n'est pas que le modèle d'apprentissage induise une fonction donnant exactement les sorties désirées des observations de la base d'entraînement (ou faire du surapprentissage), mais de trouver, comme nous venons de l'évoquer, la fonction qui aura de bonnes performances de généralisation.

En logique, ce raisonnement ou procédé de recherche d'une règle générale à partir d'un ensemble d'observations fini est appelé induction (Genesereth et Nilsson 1987, chapitre 7, pp.161-176) [1]. En apprentissage machine, le cadre inductif a été mis en place suivant le principe de la minimisation du risque empirique (MRE) (ou *Empirical Risk Minimisation* en anglais) et ses propriétés statistiques ont été étudiées dans la théorie développée par Vapnik (1999). Le résultat marquant de cette théorie est une borne supérieure de l'erreur de généralisation de la fonction apprise qui s'exprime en fonction de l'erreur empirique de cette dernière sur une base d'entraînement et de la complexité de la classe de fonctions utilisée. Cette complexité traduit la capacité de la classe de fonctions à résoudre le problème de prédiction et elle est d'autant plus grande qu'il y a de possibilités d'assigner des sorties désirées à des observations de la base d'entraînement. En d'autres termes, plus la capacité est grande, plus le risque empirique serait faible et moins on est garanti d'atteindre l'objectif principal de l'apprentissage, qui est d'avoir une faible erreur de généralisation. Cette borne exhibe ainsi le compromis qui existe entre l'erreur empirique et la capacité de la classe de fonctions, et montre une façon de

1. Le raisonnement contraire appelé déduction se base, quant à lui, sur des axiomes et produit des règles spécifiques (qui sont toujours vraies) comme des conséquences de la loi.

minimiser la borne sur l'erreur de généralisation (et d'avoir ainsi une meilleure estimation de cette erreur) en minimisant l'erreur empirique tout en contrôlant la capacité de l'ensemble de fonctions. Ce principe s'appelle la minimisation du risque structurel ; le principe ERM et lui sont à l'origine d'un grand nombre d'algorithmes d'apprentissage. De plus, ils peuvent expliquer le fonctionnement des algorithmes conçus avant l'établissement de la théorie de Vapnik (1999). La suite de ce chapitre est consacrée à la présentation plus formelle de ces différents concepts suivant le cadre de la classification bi-classe, qui a constitué le cadre initial du développement de cette théorie.

1.1 Principe de la minimisation du risque empirique

Dans cette section, nous allons présenter le principe de minimisation de risque empirique en fixant tout d'abord les notations qui seront utilisées par la suite.

1.1.1 Hypothèse et définitions

Nous supposons que les observations possèdent une représentation numérique dans un espace vectoriel de dimension fixe d, $\mathcal{X} \subseteq \mathbb{R}^d$. Les sorties désirées des observations sont supposées faire partie d'un ensemble de sortie $\mathcal{Y} \subset \mathbb{R}$. Jusqu'au début des années 2000, il y avait deux déclinaisons majeures des problèmes d'apprentissage supervisé ; la classification et la régression. En classification, l'ensemble de sortie $\mathcal{Y}$ est discret et la fonction de prédiction $f : \mathcal{X} \to \mathcal{Y}$ est appelée un classifieur. Lorsque $\mathcal{Y}$ est continu, f est une fonction de régression. Dans le chapitre 8, nous présenterons le cadre d'apprentissage de fonctions d'ordonnancement qui s'est développé récemment dans les communautés de l'apprentissage machine et de la recherche d'information. Un couple $(\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}$ désigne ainsi un exemple étiqueté et $S = (\mathbf{x}_i, y_i)_{i=1}^m \in (\mathcal{X} \times \mathcal{Y})^m$ dénote un ensemble d'exemples d'entraînement. Dans le cas particulier de la classification binaire que l'on considère dans ce chapitre, nous notons l'espace de sortie par $\mathcal{Y} = \{-1, +1\}$ et un exemple $(\mathbf{x}, +1)$ (respectivement $(\mathbf{x}, -1)$) est appelé un exemple positif (respectivement négatif). Par exemple considérons le problème de classification de courriels, consistant à les étiqueter suivant deux classes : sollicité et non sollicité. On représentera les courriels par des vecteurs dans un espace vectoriel donné et on désignera une des classes (par exemple la classe des courriels sollicités) par l'étiquette de classe $+1$ et l'autre classe par l'étiquette de classe -1.

L'hypothèse fondamentale de la théorie de l'apprentissage machine est que tous les exemples sont générés indépendamment et identiquement selon une distribution de probabilité fixe, mais inconnue, notée $\mathcal{D}$. L'hypothèse identiquement distribuée assure que les observations sont stationnaires, alors que l'hypothèse indépendamment distribuée stipule que chaque exemple individuel apporte un

maximum d'information pour résoudre le problème de prédiction. D'après cette hypothèse, les exemples $(\mathbf{x}_i, y_i)$ de tout ensemble d'entraînement S et de test sont supposés être identiquement et indépendamment distribués (i.i.d.) selon $\mathcal{D}$. Autrement dit, chaque ensemble est un échantillon d'exemples i.i.d. selon $\mathcal{D}$.

Cette hypothèse caractérise ainsi la notion de représentativité d'un ensemble d'apprentissage et de test par rapport au problème de prédiction, c'est-à-dire que les exemples d'entraînement ainsi que les observations futures et leur sortie désirée sont supposés être issus d'une même source d'information.

Un autre concept de base en apprentissage est la notion de coût, aussi appelé risque ou erreur. Pour une fonction de prédiction f donnée, le désaccord entre la sortie désirée y d'un exemple $\mathbf{x}$ et la prédiction $f(\mathbf{x})$ est mesurée grâce à une fonction de coût instantané définie par :

$$\mathbf{e} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$$

D'une manière générale, cette fonction est une distance sur l'ensemble de sortie $\mathcal{Y}$ et elle mesure l'écart entre la réponse réelle et la réponse prédite par la fonction de prédiction pour une observation donnée. En régression, les fonctions de coût instantané usuelles sont les normes ℓ_1 et ℓ_2 de la différence entre les réponses réelle et prédite d'une observation donnée. En classification bi-classe, l'erreur instantanée communément envisagée est le coût 0/1, qui pour une observation $(\mathbf{x}, y)$ et une fonction de prédiction f est définie par :

$$\mathbf{e}(f(\mathbf{x}), y) = \mathbb{1}_{f(\mathbf{x}) \neq y}$$

où $\mathbb{1}_\pi$ vaut 1 si le prédicat π est vrai et 0 sinon. En pratique, et dans le cas de la classification bi-classe, la fonction apprise $h : \mathcal{X} \to \mathbb{R}$ est une fonction à valeurs réelles et le classifieur associé $f : \mathcal{X} \to \{-1, +1\}$ est défini en prenant la fonction signe sur la sortie de h. Dans ce cas, l'erreur instantanée équivalente au coût 0/1, définie pour la fonction h est :

$$\begin{array}{rcl} \mathbf{e}_0 : \mathbb{R} \times \mathcal{Y} & \to & \mathbb{R}^+ \\ (h(\mathbf{x}), y) & \mapsto & \mathbb{1}_{y \times h(\mathbf{x}) \leq 0} \end{array}$$

À partir d'un coût instantané et de la génération i.i.d. des exemples selon la distribution $\mathcal{D}$, on peut définir l'erreur de généralisation d'une fonction apprise $f \in \mathcal{F}$ comme :

$$\mathfrak{L}(f) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}} \mathbf{e}(f(\mathbf{x}), y) = \int_{\mathcal{X}\times\mathcal{Y}} \mathbf{e}(f(\mathbf{x}), y) d\mathcal{D}(\mathbf{x}, y) \tag{1.1}$$

où $\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}} X(\mathbf{x}, y)$ est l'espérance de la variable aléatoire X lorsque $(\mathbf{x}, y)$ suit la distribution de probabilité $\mathcal{D}$. Comme $\mathcal{D}$ est inconnue, cette erreur de généralisation ne peut pas être estimée exactement, et pour mesurer la performance

d'une fonction f, on utilise souvent un ensemble d'exemples S de taille m sur lequel on calcule l'erreur empirique de f définie par :

$$\hat{\mathfrak{L}}(f, S) = \frac{1}{m} \sum_{i=1}^{m} \mathbf{e}(f(\mathbf{x}_i), y_i) \tag{1.2}$$

Ainsi, pour résoudre un problème de classification pour lequel nous disposons d'un ensemble d'entraînement S, il est naturel de choisir une classe de fonctions $\mathcal{F}$ et de chercher le classifieur f_S qui minimise l'erreur empirique sur S (puisque cette erreur est un estimateur non biaisé de l'erreur de généralisation de f_S que l'on ne peut pas mesurer).

1.1.2 Énoncé du principe

Cette méthode d'apprentissage, appelée le principe de minimisation du risque empirique (MRE), est à l'origine des tout premiers modèles d'apprentissage machine.

La question fondamentale qui se pose alors est : suivant le cadre MRE, *peut-on générer dans tous les cas une fonction de prédiction qui généralise bien à partir d'un ensemble d'observations fini ?* La réponse à cette question est bien évidemment non. Pour s'en convaincre, considérons le problème jouet de classification binaire suivant.

Exemple **Surapprentissage (Bousquet *et al.* 2003)**

Supposons que la dimension d'entrée est $d = 1$. Prenons l'espace des observations $\mathcal{X}$; l'intervalle $[a, b] \subset \mathbb{R}$ où a et b sont des réels tels que $a < b$ et l'espace des sorties est $\{-1, +1\}$. De plus, supposons que la distribution $\mathcal{D}$ générant les couples d'exemples $(\mathbf{x}, y)$ est une distribution uniforme sur $[a, b] \times \{-1\}$. Autrement dit, les exemples sont choisis de façon aléatoire sur l'intervalle $[a, b]$ et, pour chaque observation, la sortie désirée est -1.

Considérons maintenant un algorithme d'apprentissage minimisant le risque empirique, en choisissant une fonction dans la classe des fonctions $\mathcal{F} = \{f : [a, b] \to \{-1, +1\}\}$ de la façon suivante ; après avoir pris connaissance d'un ensemble d'apprentissage $S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}$ l'algorithme produit la fonction de prédiction f_S telle que :

$$f_S(\mathbf{x}) = \begin{cases} -1, & \text{si } \mathbf{x} \in \{\mathbf{x}_1, \ldots, \mathbf{x}_m\} \\ +1, & \text{sinon.} \end{cases}$$

Dans ce cas, le classifieur produit par l'algorithme d'apprentissage a un risque empirique égal à 0, et ceci pour n'importe quel ensemble d'apprentissage donné. Cependant, comme le classifieur fait une erreur sur tout l'ensemble infini $[a, b]$ sauf pour les exemples d'une base d'entraînement finie, de mesure nulle, son erreur de généralisation est toujours égale à 1.

1.2 Consistance du principe MRE

La question sous-jacente à la question précédente est : *dans quel cas le principe MRE est-il susceptible de générer une règle générale d'apprentissage ?* La réponse à cette question réside dans une notion statistique appelée consistance.

1.2.1 Définition

Ce concept indique les deux conditions qu'un algorithme d'apprentissage doit remplir, à savoir *(a)* l'algorithme doit renvoyer une fonction dont l'erreur empirique reflète son erreur de généralisation lorsque la taille de la base d'entraînement tend vers l'infini et, *(b)* dans le cas asymptotique, l'algorithme doit permettre de trouver une fonction qui minimise l'erreur de généralisation dans la classe de fonctions considérée. De façon formelle :

(a) $\forall \epsilon > 0, \lim_{m \to \infty} \mathbb{P}(|\hat{\mathfrak{L}}(f_S, S) - \mathfrak{L}(f_S)| > \epsilon) = 0$, noté, $\hat{\mathfrak{L}}(f_S, S) \xrightarrow{\mathbb{P}} \mathfrak{L}(f_S)$

(b) $\hat{\mathfrak{L}}(f_S, S) \xrightarrow{\mathbb{P}} \inf_{g \in \mathcal{F}} \mathfrak{L}(g)$

Ces deux conditions impliquent ainsi la convergence en probabilité de l'erreur empirique $\hat{\mathfrak{L}}(f_S, S)$ de la fonction de prédiction trouvée par l'algorithme d'apprentissage sur la base d'entraînement S, f_S, vers son erreur de généralisation $\mathfrak{L}(f_S)$ et $\inf_{g \in \mathcal{F}} \mathfrak{L}(g)$ (figure 1.1).

Une façon naturelle d'analyser la condition *(a)* de la consistance, exprimant le concept de la généralisation, est d'utiliser l'inégalité suivante :

$$|\mathfrak{L}(f_S) - \hat{\mathfrak{L}}(f_S, S)| \leq \sup_{g \in \mathcal{F}} |\mathfrak{L}(g) - \hat{\mathfrak{L}}(g, S)| \tag{1.3}$$

Nous voyons bien d'après cette inégalité qu'une condition suffisante pour généraliser est qu'asymptotiquement, l'erreur empirique de la fonction de prédiction, dont l'écart en valeur absolue entre cette erreur et son erreur de généralisation parmi toutes les autres fonctions d'une classe de fonctions $\mathcal{F}$ donnée est la plus grande, tend vers l'erreur de généralisation de la fonction, soit :

$$\sup_{g \in \mathcal{F}} |\mathfrak{L}(g) - \hat{\mathfrak{L}}(g, S)| \xrightarrow{\mathbb{P}} 0 \tag{1.4}$$

Cette condition suffisante pour généraliser est une considération au pire cas et, d'après (équation 1.3), elle implique une convergence uniforme bilatérale pour toutes les fonctions de la classe $\mathcal{F}$. De plus, la condition (équation 1.4) ne dépend pas de l'algorithme considéré mais uniquement de la classe de fonctions $\mathcal{F}$. Ainsi, une condition nécessaire pour que le principe MRE soit consistant est que la classe de fonctions considérée soit restreinte (voir l'exemple sur le surapprentissage de la section précédente).