Principes de base en apprentissage supervisé

Massih-Reza Amini

Principes de base en apprentissage supervisé

Massih-Reza Amini

Résumé

Ce document constitue le premier chapitre de l'ouvrage [1], présentant la théorie de l'apprentissage machine selon le cadre de [19] et qui a servi de base dans la description des algorithmes d'apprentissage décrits dans les chapitres suivants. Plus particulièrement, nous présentons ici 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. L'étude de la consistance de ce principe, présentée dans la section 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 [8, 7, 16, 3]. 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 [9, 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 [19]. 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 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 [19]. 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. 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.

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 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 **??**, 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 :

$$\mathfrak{L}(f) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}} \mathbf{e}(f(\mathbf{x}), y) = \int_{\mathcal{X}\times\mathcal{Y}} \mathbf{e}(f(\mathbf{x}), y) d\mathcal{D}(\mathbf{x}, y) \tag{1}$$

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 :

$$\hat{\mathfrak{L}}(f, S) = \frac{1}{m} \sum_{i=1}^{m} \mathbf{e}(f(\mathbf{x}_i), y_i) \tag{2}$$

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).

1.2 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 [4]**

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.

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.

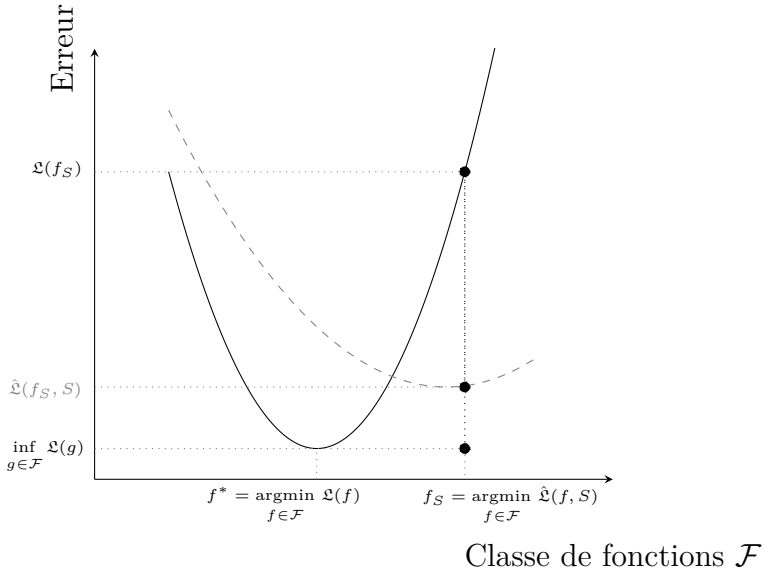

Figure 1 - Description schématique de la notion de consistance. L'axe des abscisses représente la classe des fonctions $\mathcal{F}$ et les courbes d'erreurs empirique (en pointillé) et de généralisation (en trait plein) sont montrées en fonction de $f \in \mathcal{F}$. Le principe MRE consiste à trouver la fonction f_S dans la classe $\mathcal{F}$ qui minimise l'erreur empirique sur une base d'entraînement S. Ce principe est consistant si en probabilité $\hat{\mathfrak{L}}(f_S, S)$ converge vers $\mathfrak{L}(f_S)$ et $\inf_{g \in \mathcal{F}} \mathfrak{L}(g)$.

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).

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

$$|\mathfrak{L}(f_S) - \hat{\mathfrak{L}}(f_S, S)| \leq \sup_{g \in \mathcal{F}} |\mathfrak{L}(g) - \hat{\mathfrak{L}}(g, S)| \tag{3}$$

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 :

$$\sup_{g \in \mathcal{F}} |\mathfrak{L}(g) - \hat{\mathfrak{L}}(g, S)| \xrightarrow{\mathbb{P}} 0 \tag{4}$$

Cette condition suffisante pour généraliser est une considération au pire cas et, d'après (équation 3), elle implique une convergence uniforme bilatérale pour toutes les fonctions de la classe $\mathcal{F}$. De plus, la condition (équation 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).

2.2 Étude de pire cas

Le résultat fondamental de la théorie de l'apprentissage [19, théorème 2.1, p.38] exhibe une autre relation faisant intervenir le supremum sur la classe de fonctions concernant la consistance du principe MRE (et non pas la capacité à généraliser) sous forme d'une convergence uniforme unilatérale au pire cas :
Le principe MRE est consistant si et seulement si :

$$\forall \epsilon > 0, \lim_{m \to \infty} \mathbb{P}\left(\sup_{f \in \mathcal{F}} \left[\mathfrak{L}(f) - \hat{\mathfrak{L}}(f, S)\right] > \epsilon\right) = 0 \tag{5}$$

Cette condition stipule que le principe MRE est consistant si et seulement si la convergence uniforme unilatérale est assurée pour la fonction de la classe de fonctions $\mathcal{F}$ pour laquelle la différence entre son erreur de généralisation et son erreur empirique est la plus grande, ce qui lui a valu le nom de « étude de pire »(ou *worst case study*).

L'implication directe de ce résultat est une borne uniforme sur l'erreur de généralisation de toute fonction de prédiction $f \in \mathcal{F}$ apprise sur une base d'entraînement S de taille m et qui est de la forme suivante :

$$\forall \delta \in]0, 1], \mathbb{P}\left(\forall f \in \mathcal{F}, (\mathfrak{L}(f) - \hat{\mathfrak{L}}(f, S)) \leq \mathfrak{C}(\mathcal{F}, m, \delta)\right) \geq 1 - \delta \tag{6}$$

où $\mathfrak{C}$ est un terme qui dépend de la *taille* de la classe de fonctions utilisée, de la taille de la base d'entraînement et de la précision $\delta \in]0, 1]$ souhaitée. La recherche en apprentissage machine a étudié différentes façons pour mesurer la taille d'une classe de fonctions et la mesure utilisée à cette fin est communément appelée la complexité ou la capacité de la classe de fonctions. Dans ce chapitre et le suivant, nous allons présenter deux de ces mesures (à savoir la dimension VC et la complexité de Rademacher) menant à différents types de bornes de généralisation et aussi à un nouveau principe d'apprentissage appelé : minimisation du risque structurel (MRS).

3 Principe de la Minimisation du Risque Structurel

Avant de présenter une borne de généralisation estimée sur la base d'entraînement qui a servi à trouver la fonction de prédiction, nous allons considérer dans un premier temps l'estimation de l'erreur de généralisation d'une fonction sur une base de test [12]. Le but ici est de montrer qu'il est possible d'estimer une borne supérieure sur l'erreur de généralisation en utilisant un ensemble de test et que l'erreur empirique sur cet ensemble converge, en probabilité, vers l'erreur de généralisation lorsque le nombre d'exemples de test tend vers l'infini et ceci indépendamment de la capacité de la classe de fonctions considérée.

3.1 Estimation de l'erreur de généralisation sur un ensemble de test

Rappelons que les exemples d'une base de test sont générés i.i.d. suivant la même distribution de probabilité $\mathcal{D}$ qui a servi à générer les exemples d'une base d'entraînement. Considérons f_S une fonction apprise sur une base d'entraînement S. Soit $T = \{(\mathbf{x}_i, y_i); i \in \{1, \ldots, n\}\}$ une base de test de taille n. Comme les exemples de cet ensemble n'interviennent pas dans la phase d'apprentissage, la fonction f_S ne dépend pas des valeurs des erreurs instantanées des exemples $(\mathbf{x}_i, y_i)$ de cet ensemble, et les variables aléatoires $(f_S(\mathbf{x}_i), y_i) \mapsto \mathbf{e}(f_S(\mathbf{x}_i), y_i)$ peuvent être considérées comme des copies indé-

pendantes d'une même variable aléatoire, soit :

$$\begin{aligned}\mathbb{E}_{T\sim\mathcal{D}^n}\hat{\mathfrak{L}}(f_S,T) &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{T\sim\mathcal{D}^n}\mathbf{e}(f_S(\mathbf{x}_i),y_i)\\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\mathbf{e}(f_S(\mathbf{x}),y) = \mathfrak{L}(f_S)\end{aligned}$$

Ainsi, l'erreur empirique de f_S sur la base de test $\hat{\mathfrak{L}}(f_S,T)$ est un estimateur non biaisé de son erreur de généralisation.

De plus, pour chaque exemple $(\mathbf{x}_i,y_i)$, désignons par X_i la variable aléatoire définie par $\frac{1}{n}\mathbf{e}(f_S(\mathbf{x}_i),y_i)$. Comme les variables aléatoires $X_i, i\in\{1,\ldots,n\}$ sont indépendantes et qu'elles ont des valeurs dans $\{0,\frac{1}{n}\}$, en remarquant que $\hat{\mathfrak{L}}(f_S,T)=\sum_{i=1}^{n}X_i$ et que $\mathfrak{L}(f_S)=\mathbb{E}\left(\sum_{i=1}^{n}X_i\right)$, nous avons d'après l'inégalité de [10] :

$$\forall\epsilon>0,\mathbb{P}\left(\left[\mathfrak{L}(f_S)-\hat{\mathfrak{L}}(f_S,T)\right]>\epsilon\right)\leq e^{-2n\epsilon^2} \tag{7}$$

Pour mieux appréhender ce résultat, résolvons l'équation $e^{-2n\epsilon^2}=\delta$ en fonction de ϵ, soit $\epsilon=\sqrt{\frac{\ln 1/\delta}{2n}}$, et considérons l'évènement opposé. Nous avons ainsi :

$$\forall\delta\in]0,1],\mathbb{P}\left(\mathfrak{L}(f_S)\leq\hat{\mathfrak{L}}(f_S,T)+\sqrt{\frac{\ln 1/\delta}{2n}}\right)\geq 1-\delta \tag{8}$$

Pour un δ faible, nous avons d'après (équation 8) et avec une forte probabilité l'inégalité $\mathfrak{L}(f_S)\leq\hat{\mathfrak{L}}(f_S,T)+\sqrt{\frac{\ln 1/\delta}{2n}}$, qui se tient pour pratiquement tous les ensembles de test possibles de taille n. D'après ce résultat, nous avons ainsi une borne sur l'erreur de généralisation de la fonction apprise qui peut se calculer sur un ensemble de test quelconque et qui, dans le cas où n est suffisamment grand, donne une bonne approximation de cette dernière.

Exemple **Estimation de l'erreur de généralisation sur un ensemble de test [12]**
Supposons que l'erreur empirique d'une fonction de prédiction f_S sur une base de test T de taille $n=1000$ est $\hat{\mathfrak{L}}(f_S,T)=0.23$. Pour $\delta=0.01$, i.e. $\sqrt{\frac{\ln(1/\delta)}{2n}}\approx 0.047$, nous avons l'erreur de généralisation de la fonction f_S qui est ainsi majorée par 0.277 avec une probabilité d'au moins 0.99.

3.2 Borne uniforme sur l'erreur de généralisation

Pour une fonction de prédiction donnée, nous savons d'après le résultat précédent comment borner son erreur de généralisation, en utilisant une

base de test sur laquelle les paramètres de la fonction n'ont pas été trouvés. Dans le cadre de l'étude de la consistance du principe MRE, nous aimerions maintenant établir une borne uniforme sur l'erreur de généralisation d'une fonction apprise en fonction de son erreur empirique sur une base d'entraînement. Nous ne pouvons pas répondre à cette question en utilisant le développement précédent. Ceci est principalement dû au fait que, lorsque la fonction apprise f_S a eu connaissance des données de la base d'entraînement $S = \{(\mathbf{x}_i, y_i); i \in \{1, \ldots, m\}\}$, les variables aléatoires $X_i = \frac{1}{m}\mathbf{e}(f_S(\mathbf{x}_i), y_i); i \in \{1, \ldots, m\}$, qui interviennent dans le calcul de l'erreur empirique de la fonction f_S sur S, sont toutes dépendantes les unes des autres. En effet, si on change un exemple de la base d'entraînement, la fonction choisie f_S change aussi, ainsi que les erreurs instantanées pondérées de tous les autres exemples. De ce fait, comme les variables aléatoires X_i ne peuvent plus être considérées indépendamment distribuées, nous ne sommes plus en mesure d'appliquer l'inégalité de [10].

Dans la suite, nous allons exposer une borne uniforme sur l'erreur de généralisation en suivant le cadre de [19]. Dans la section suivante, nous présentons un autre cadre plus récent, développé au début des années 2000, en montrant le lien avec les travaux de [19].

Pour la borne uniforme, notre point de départ est la majoration de la probabilité $\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left[\mathfrak{L}(f) - \hat{\mathfrak{L}}(f, S)\right] > \epsilon\right)$ de (équation 5). À ce stade, deux cas se présentent, le cas des ensembles de fonctions finis et infinis.

3.2.1 Cas des ensembles finis de fonctions

Considérons une classe de fonctions $\mathcal{F} = \{f_1, \ldots, f_p\}$ de taille $p = |\mathcal{F}|$. La borne de généralisation consiste ainsi à estimer, étant donné $\epsilon > 0$, la probabilité que pour une base d'apprentissage de taille m, $\max_{j\in\{1,\ldots,p\}}\left[\mathfrak{L}(f_j) - \hat{\mathfrak{L}}(f_j, S)\right]$ soit supérieure à ϵ.

Si $p = 1$, le seul choix de sélectionner la fonction de prédiction dans $\mathcal{F} = \{f_1\}$ se restreint à prendre f_1 et ceci avant même de considérer les observations de n'importe quel échantillon S de taille m. Nous pouvons dans ce cas appliquer directement la borne (7) précédente issue de l'inégalité de [10], soit :

$$\forall\epsilon > 0, \mathbb{P}\left(\max_{j=1}\left[\mathfrak{L}(f_j) - \hat{\mathfrak{L}}(f_j, S)\right] > \epsilon\right) = \mathbb{P}\left(\left[\mathfrak{L}(f_1) - \hat{\mathfrak{L}}(f_1, S)\right] > \epsilon\right) \leq e^{-2m\epsilon^2}$$

L'interprétation de ce résultat est que, pour une fonction f fixée et un $\epsilon > 0$ donné, la proportion des ensembles de m observations possibles pour lesquelles $\mathfrak{L}(f) - \hat{\mathfrak{L}}(f, S) > \epsilon$ est inférieure à $e^{-2m\epsilon^2}$.

Si $p > 1$, nous pouvons d'abord remarquer que :

$$\max_{j\in\{1,\ldots,p\}} \left[\mathfrak{L}(f_j) - \hat{\mathfrak{L}}(f_j, S)\right] > \epsilon \Leftrightarrow \exists f \in \mathcal{F}, \mathfrak{L}(f) - \hat{\mathfrak{L}}(f, S) > \epsilon \tag{9}$$

pour un $\epsilon > 0$ fixé et chaque fonction $f_j \in \mathcal{F}$; considérons l'ensemble des échantillons de taille m pour lesquels l'erreur de généralisation de f_j est plus grande, de plus de ϵ, par rapport à son erreur empirique :

$$\mathfrak{S}_j^\epsilon = \{S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\} : \mathfrak{L}(f_j) - \hat{\mathfrak{L}}(f_j, S) > \epsilon\}$$

Étant donné $j \in \{1, \ldots, p\}$ fixé, nous avons d'après l'interprétation précédente que la probabilité sur les échantillons S pour que $\mathfrak{L}(f_j) - \hat{\mathfrak{L}}(f_j, S) > \epsilon$ est inférieure à $e^{-2m\epsilon^2}$, soit :

$$\forall j \in \{1, \ldots, p\}; \mathbb{P}(\mathfrak{S}_j^\epsilon) \leq e^{-2m\epsilon^2} \tag{10}$$

D'après l'équivalence (équation 9), nous avons par ailleurs :

$$\begin{aligned}
\forall \epsilon > 0, \mathbb{P}\left(\max_{j\in\{1,\ldots,p\}} \left[\mathfrak{L}(f_j) - \hat{\mathfrak{L}}(f_j, S)\right] > \epsilon\right) &= \mathbb{P}\left(\exists f \in \mathcal{F}, \mathfrak{L}(f) - \hat{\mathfrak{L}}(f, S) > \epsilon\right) \\
&= \mathbb{P}\left(\mathfrak{S}_1^\epsilon \cup \ldots \cup \mathfrak{S}_p^\epsilon\right)
\end{aligned}$$

Nous allons maintenant borner l'égalité précédente en utilisant le résultat de (équation 10) et la borne de l'union qui est l'outil de base pour dériver des bornes de généralisation :

$$\begin{aligned}
\forall \epsilon > 0, \mathbb{P}\left(\max_{j\in\{1,\ldots,p\}} \left[\mathfrak{L}(f_j) - \hat{\mathfrak{L}}(f_j, S)\right] > \epsilon\right) &= \mathbb{P}\left(\mathfrak{S}_1^\epsilon \cup \ldots \cup \mathfrak{S}_p^\epsilon\right) \\
&\leq \sum_{j=1}^{p} \mathbb{P}(\mathfrak{S}_j^\epsilon) \leq pe^{-2m\epsilon^2}
\end{aligned}$$

En résolvant cette borne pour $\delta = pe^{-2m\epsilon^2}$, soit $\epsilon = \sqrt{\frac{\ln(p/\delta)}{2m}} = \sqrt{\frac{\ln(|\mathcal{F}|/\delta)}{2m}}$, et en considérant l'évènement opposé, il vient :

$$\forall \delta \in]0, 1], \mathbb{P}\left(\forall f \in \mathcal{F}, \mathfrak{L}(f) \leq \hat{\mathfrak{L}}(f, S) + \sqrt{\frac{\ln(|\mathcal{F}|/\delta)}{2m}}\right) \geq 1 - \delta \tag{11}$$

En comparaison avec la borne de généralisation obtenue sur un ensemble de test (équation 8), nous voyons bien que l'erreur empirique sur une base test est un meilleur estimateur de l'erreur de généralisation que l'erreur empirique sur une base d'entraînement. En outre, plus l'ensemble de fonctions contient

des fonctions différentes, plus il y a de chances que l'erreur empirique sur la base d'entraînement soit une sous-estimation importante de l'erreur de généralisation.

En effet, l'interprétation de la borne (équation 11) est que pour un $\delta \in]0, 1]$ fixé et pour une fraction plus grande que $1 - \delta$ des ensembles d'entraînement possibles, toutes les fonctions de l'ensemble fini $\mathcal{F}$ (y compris la fonction qui minimise l'erreur empirique) ont une erreur de généralisation inférieure à leur erreur empirique plus le terme résiduel $\sqrt{\frac{\ln(|\mathcal{F}|/\delta)}{2m}}$. De plus, la différence au pire cas entre l'erreur de généralisation et l'erreur empirique tend vers 0 lorsque le nombre d'exemples tend vers l'infini, et ceci sans aucune hypothèse particulière sur la distribution $\mathcal{D}$ générant les données. Ainsi, pour tout ensemble fini de fonctions, le principe MRE est consistant pour n'importe quelle distribution de probabilité $\mathcal{D}$.

En Récapitulatif

Les deux étapes qui ont mené à la borne de généralisation présentées dans le développement précédent sont ainsi :

1. Pour toute fonction $f_j \in \{f_1, \ldots, f_p\}$ fixée et un $\epsilon > 0$ donné, borner la probabilité sur les échantillons S pour que $\mathfrak{L}(f_j) - \hat{\mathfrak{L}}(f_j, S) > \epsilon$.
2. Utiliser la borne de l'union pour passer de cette probabilité pour une seule fonction à la probabilité pour toutes les fonctions de la classe $\mathcal{F}$ en même temps.

3.2.2 Cas des ensembles infinis de fonctions

Pour le cas d'une classe de fonctions infinie, l'approche précédente n'est pas directement applicable. En effet, étant donné un ensemble de m observations $S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}$, si l'on considère l'ensemble suivant :

$$\mathfrak{F}(\mathcal{F}, S) = \left\{ \Big((\mathbf{x}_1, f(\mathbf{x}_1)), \ldots, (\mathbf{x}_m, f(\mathbf{x}_m)) \Big) \mid f \in \mathcal{F} \right\} \tag{12}$$

la taille de cet ensemble correspond au nombre de manières possibles dont les fonctions de $\mathcal{F}$ peuvent étiqueter les exemples $(\mathbf{x}_1, \ldots, \mathbf{x}_m)$. Comme ces fonctions n'ont que deux sorties possibles (-1 ou $+1$), la taille de $\mathfrak{F}(\mathcal{F}, S)$ est finie, bornée par 2^m, et ceci quelle que soit la classe de fonctions $\mathcal{F}$ considérée. Ainsi, un algorithme d'apprentissage minimisant le risque empirique sur un ensemble d'apprentissage S choisit la fonction parmi $|\mathfrak{F}(\mathcal{F}, S)|$ fonctions de $\mathcal{F}$ qui réalise l'étiquetage des exemples de S aboutissant à la plus petite erreur. Ainsi, il n'y a qu'un nombre fini de fonctions qui vont intervenir dans

le calcul de l'erreur empirique apparaissant dans l'expression de :

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left[\mathfrak{L}(f)-\hat{\mathfrak{L}}(f,S)\right]>\epsilon\right) \tag{13}$$

Cependant, pour un second ensemble S' différent du premier, l'ensemble $\mathfrak{F}(\mathcal{F},S')$ sera différent de $\mathfrak{F}(\mathcal{F},S)$ et il est impossible d'appliquer la borne obtenue pour les ensembles finis en considérant $|\mathfrak{F}(\mathcal{F},S)|$.

La solution proposée par Vapnik et Chervonenkis est une manière élégante de résoudre ce problème. Elle consiste à remplacer la vraie erreur $\mathfrak{L}(f)$ dans l'expression (équation 13) par l'erreur empirique de f sur un autre échantillon de même taille que S, appelé *échantillon virtuel* ou *fantôme* (*ghost sample* en anglais), et elle s'énonce formellement comme suit :

Lemme 1 (Symétrisation de Vapnik et Chervonenkis [19])**.** *Soit $\mathcal{F}$ une classe de fonctions (pouvant être infinie) et S et S' deux échantillons d'apprentissage de même taille m. Pour tout réel $\epsilon>0$, tel que $m\epsilon^2\geq 2$ nous avons alors :*

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left[\mathfrak{L}(f)-\hat{\mathfrak{L}}(f,S)\right]>\epsilon\right)\leq 2\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left[\hat{\mathfrak{L}}(f,S')-\hat{\mathfrak{L}}(f,S)\right]>\epsilon/2\right) \tag{14}$$

Preuve **du lemme de symétrisation**

Soit $\epsilon>0$ et $f_S^*\in\mathfrak{F}(\mathcal{F},S)$ la fonction qui réalise le supremum $\sup_{f\in\mathcal{F}}\left[\mathfrak{L}(f)-\hat{\mathfrak{L}}(f,S)\right]$. D'après la remarque ci-dessus, f_S^* dépend de l'échantillon S. Nous avons :

$$\begin{aligned}\mathbb{1}_{[\mathfrak{L}(f_S^*)-\hat{\mathfrak{L}}(f_S^*,S)]>\epsilon}\mathbb{1}_{[\mathfrak{L}(f_S^*)-\hat{\mathfrak{L}}(f_S^*,S')]<\epsilon/2} &= \mathbb{1}_{[\mathfrak{L}(f_S^*)-\hat{\mathfrak{L}}(f_S^*,S)]>\epsilon\wedge[\hat{\mathfrak{L}}(f_S^*,S')-\mathfrak{L}(f_S^*)]\geq-\epsilon/2}\\ &\leq \mathbb{1}_{\hat{\mathfrak{L}}(f_S^*,S')-\hat{\mathfrak{L}}(f_S^*,S)>\epsilon/2}\end{aligned}$$

En prenant l'espérance sur l'échantillon S' dans l'inégalité précédente il vient :

$$\mathbb{1}_{[\mathfrak{L}(f_S^*)-\hat{\mathfrak{L}}(f_S^*,S)]>\epsilon}\mathbb{E}_{S'\sim\mathcal{D}^m}[\mathbb{1}_{\mathfrak{L}(f_S^*)-\hat{\mathfrak{L}}(f_S^*,S')<\epsilon/2}]\leq\mathbb{E}_{S'\sim\mathcal{D}^m}[\mathbb{1}_{\hat{\mathfrak{L}}(f_S^*,S')-\hat{\mathfrak{L}}(f_S^*,S)>\epsilon/2}]$$

Soit :

$$\mathbb{1}_{[\mathfrak{L}(f_S^*)-\hat{\mathfrak{L}}(f_S^*,S)]>\epsilon}\mathbb{P}(\mathfrak{L}(f_S^*)-\hat{\mathfrak{L}}(f_S^*,S')<\epsilon/2)\leq\mathbb{E}_{S'\sim\mathcal{D}^m}[\mathbb{1}_{\hat{\mathfrak{L}}(f_S^*,S')-\hat{\mathfrak{L}}(f_S^*,S)>\epsilon/2}]$$

Pour chaque exemple $(\mathbf{x}'_i,y'_i)\in S'$ désignons par X_i la variable aléatoire $\frac{1}{m}\mathbf{e}(f_S^*(\mathbf{x}'_i),y'_i)$. Comme f_S^* est indépendante de l'échantillon S', les variables aléatoires $X_i, i\in\{1,\ldots,m\}$ sont indépendantes. La variance de la variable aléatoire $\hat{\mathfrak{L}}(f_S^*,S')$, $\mathbb{V}(\hat{\mathfrak{L}}(f_S^*,S'))$, est ainsi égale à :

$$\mathbb{V}(\hat{\mathfrak{L}}(f_S^*,S'))=\frac{1}{m}\mathbb{V}(\mathbf{e}(f_S^*(\mathbf{x}'),y'))$$

et, d'après l'inégalité de Tchebychev, nous avons :

$$\mathbb{P}(\mathfrak{L}(f_S^*) - \hat{\mathfrak{L}}(f_S^*, S') \geq \epsilon/2) \leq \frac{4\mathbb{V}(\mathbf{e}(f_S^*(\mathbf{x}'), y'))}{m\epsilon^2} \leq \frac{1}{m\epsilon^2}$$

La dernière inégalité est due au fait que $\mathbf{e}(f_S^*(\mathbf{x}'), y'))$ soit une variable aléatoire prenant ses valeurs dans $[0, 1]$ et que sa variance soit inférieure à $1/4$:

$$\left(1 - \frac{1}{m\epsilon^2}\right) \mathbb{1}_{[\mathfrak{L}(f_S^*) - \hat{\mathfrak{L}}(f_S^*, S)] > \epsilon} \leq \mathbb{E}_{S' \sim \mathcal{D}^m}[\mathbb{1}_{\hat{\mathfrak{L}}(f_S^*, S') - \hat{\mathfrak{L}}(f_S^*, S) > \epsilon/2}]$$

Le résultat s'ensuit en prenant l'espérance sur l'échantillon S et en notant que $m\epsilon^2 \geq 2$, i.e. $\frac{1}{2} \leq \left(1 - \frac{1}{m\epsilon^2}\right)$.

Nous remarquons que l'espérance de gauche dans l'inégalité (14) est suivant la distribution d'un échantillon i.i.d. de taille m, alors que celle de droite est suivant la distribution d'un échantillon i.i.d. de taille $2m$.

L'extension de la borne de généralisation pour une classe de fonctions infinie, $\mathcal{F}$, se fait en étudiant le plus grand écart entre les risques empiriques des fonctions de $\mathcal{F}$ sur deux ensembles d'apprentissage S et S' quelconques et de même taille. En effet, la quantité importante qui intervient dans le résultat précédent est le nombre maximal d'étiquetages possibles pour deux ensembles de même taille, m, notée $\mathfrak{G}(\mathcal{F}, 2m)$, où :

$$\mathfrak{G}(\mathcal{F}, m) = \max_{S \in \mathcal{X}^m} |\mathfrak{F}(\mathcal{F}, S)| \tag{15}$$

$\mathfrak{G}(\mathcal{F}, m)$ est appelée la fonction de croissance et elle mesure le nombre maximum d'étiquetages possibles d'une séquence de m points de $\mathcal{X}$ par la classe de fonctions, $\mathcal{F}$. $\mathfrak{G}(\mathcal{F}, m)$ peut ainsi être vue comme une mesure de la taille de la classe de fonctions $\mathcal{F}$ comme le montre le résultat suivant :

Théorème 1 ([20] ; [19], chapitre 3)**.** *Soit $\delta \in]0, 1]$ et S une base d'entraînement de taille m générée i.i.d. suivant une distribution de probabilité $\mathcal{D}$; nous avons avec une probabilité au moins égale à $1 - \delta$:*

$$\forall f \in \mathcal{F}, \mathfrak{L}(f) \leq \hat{\mathfrak{L}}(f, S) + \sqrt{\frac{8 \ln(\mathfrak{G}(\mathcal{F}, 2m)) + 8 \ln(\frac{4}{\delta})}{m}} \tag{16}$$

Preuve **du théorème 1**

Soit ϵ un réel positif. D'après le lemme de symétrisation (équation 1), nous avons :

$$\begin{aligned}\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left[\mathfrak{L}(f)-\hat{\mathfrak{L}}(f,S)\right]>\epsilon\right) &\leq 2\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left[\hat{\mathfrak{L}}(f,S')-\hat{\mathfrak{L}}(f,S)\right]>\epsilon/2\right)\\ &= 2\mathbb{P}\left(\sup_{f\in\mathfrak{F}(\mathcal{F},S\cup S')}\left[\hat{\mathfrak{L}}(f,S')-\hat{\mathfrak{L}}(f,S)\right]>\epsilon/2\right)\end{aligned}$$

D'après la borne de l'union, cela donne :

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left[\mathfrak{L}(f)-\hat{\mathfrak{L}}(f,S)\right]>\epsilon\right)\leq 2\mathfrak{G}(\mathcal{F},2m)\mathbb{P}\left(\left[\hat{\mathfrak{L}}(f,S')-\hat{\mathfrak{L}}(f,S)\right]>\epsilon/2\right)$$

D'après l'inégalité de [10], nous avons :

$$\forall\epsilon>0,\forall f\in\mathfrak{F}(\mathcal{F},S\cup S'),\mathbb{P}\left(\left[\hat{\mathfrak{L}}(f,S')-\hat{\mathfrak{L}}(f,S)\right]>\epsilon/2\right)\leq 2e^{-m(\epsilon/2)^2/2}$$

soit :

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left[\mathfrak{L}(f)-\hat{\mathfrak{L}}(f,S)\right]>\epsilon\right)\leq 4\mathfrak{G}(\mathcal{F},2m)e^{-m\epsilon^2/8}$$

Le résultat s'ensuit en résolvant $4\mathfrak{G}(\mathcal{F},2m)e^{-m\epsilon^2/8}=\delta$ pour ϵ.

Un résultat important de ce théorème est que le principe MRE est consistant dans le cas où $\sqrt{\frac{\ln(\mathfrak{G}(\mathcal{F},2m))}{m}}$ tend vers 0 lorsque m tend vers l'infini. De plus, comme la distribution $\mathcal{D}$ des observations n'intervient pas dans la définition de la fonction de croissance, l'analyse précédente est valide quelle que soit $\mathcal{D}$.

Ainsi, une condition suffisante pour que le principe MRE soit consistant, pour toutes les distributions de probabilité $\mathcal{D}$ et une classe de fonctions infinie, est :

$$\lim_{m\to\infty}\sqrt{\frac{\ln(\mathfrak{G}(\mathcal{F},2m))}{m}}=0$$

En revanche, $\mathfrak{G}(\mathcal{F},m)$ est une quantité non mesurable et la seule certitude dont on dispose est qu'elle soit bornée par 2^m. De plus, dans le cas où la fonction de croissance atteindrait cette borne, $\mathfrak{G}(\mathcal{F},m)=2^m$, cela signifierait qu'il existe un échantillon de taille m tel que la classe de fonctions $\mathcal{F}$ peut générer tous les étiquetages possibles sur cet échantillon, on dit alors que l'échantillon est pulvérisé par $\mathcal{F}$. À partir de ce constat, Vapnik et Chervonenkis ont proposé une quantité auxiliaire, appelée dimension VC, pour étudier la fonction de croissance et qui est définie de la façon suivante.

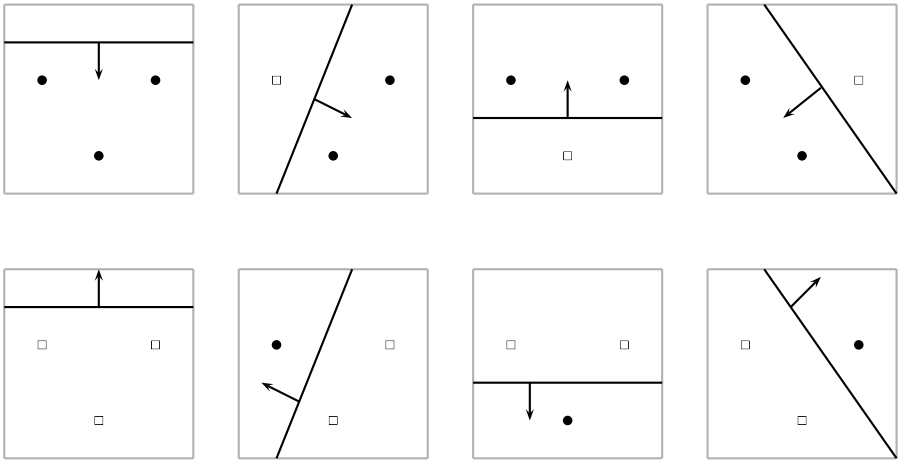

Figure 2 - Pulvérisation des points dans le plan de dimension $d = 2$ par une classe de fonctions linéaires. Chaque classifieur linéaire sépare le plan en deux sous-espaces, avec un vecteur normal qui pointe vers le sous-espace contenant les exemples appartenant à la classe $+1$ (représentés par des cercles pleins). Le nombre maximal de points dans le plan pouvant être pulvérisés par la classe de fonctions linéaires, ou la dimension VC de cette classe de fonctions, est dans ce cas égal à 3.

Définition 1 (Dimension VC, [19]). *Soit $\mathcal{F} = \{f : \mathcal{X} \to \{-1, +1\}\}$ une classe de fonctions à valeurs discrètes. La dimension VC de $\mathcal{F}$ est le plus grand entier $\mathcal{V}$ vérifiant $\mathfrak{G}(\mathcal{F}, \mathcal{V}) = 2^{\mathcal{V}}$. Autrement dit, $\mathcal{V}$ est le plus grand nombre de points que la classe de fonctions arrive à pulvériser. Si un tel entier n'existe pas, la dimension VC de $\mathcal{F}$ est considérée alors comme infinie.*

La figure 2 illustre le calcul de la dimension VC d'une classe de fonctions linéaires dans le plan. D'après la définition précédente, nous voyons bien que plus la dimension VC, $\mathcal{V}$, d'une classe de fonctions est grande, plus la fonction de croissance $\mathfrak{G}(\mathcal{F}, m)$ de cette classe est élevée, et ceci pour n'importe quel $m \geq \mathcal{V}$. Une propriété importante démontrée par [15, 17], est que la dimension VC d'une classe de fonctions $\mathcal{F}$ est une mesure de la capacité de $\mathcal{F}$ elle est exhibée dans le lemme suivant.

Lemme 2 ([20, 15, 17] [2]). *Soit $\mathcal{F}$ une classe de fonctions à valeurs dans $\{-1, +1\}$ et avec une dimension VC finie, $\mathcal{V}$.*

2. Ce lemme est plus connu sous le nom de lemme de Sauer mais il a été énoncé pour la première, et sous une forme légèrement différente, dans [20].

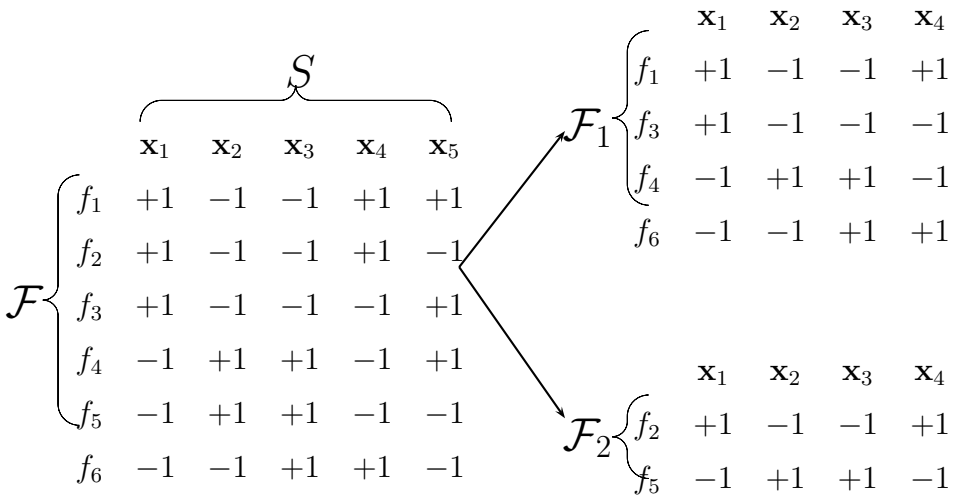

Figure 3 - Construction des ensembles $\mathcal{F}_1$ et $\mathcal{F}_2$ à partir de la classe de fonctions $\mathcal{F}$ pour la preuve du lemme de [15] sur un exemple jouet.

Pour tout entier naturel m, la fonction de croissance $\mathfrak{G}(\mathcal{F}, m)$ est bornée par :

$$\mathfrak{G}(\mathcal{F}, m) \leq \sum_{i=0}^{\mathcal{V}} \binom{m}{i} \tag{17}$$

Et pour tout $m \geq \mathcal{V}$:

$$\mathfrak{G}(\mathcal{F}, m) \leq \left(\frac{m}{\mathcal{V}}\right)^{\mathcal{V}} e^{\mathcal{V}} \tag{18}$$

Il existe différentes preuves de ce lemme [15, 17, 5, 6, 13], dont celle basée sur un raisonnement par récurrence par rapport à $m + \mathcal{V}$ que nous allons présenter dans la suite. Notons d'abord que l'inégalité (17) est vraie pour $\mathcal{V} = 0$ et $m = 0$. En effet :

— Si $\mathcal{V} = 0$, cela signifie que la classe de fonctions n'arrive à pulvériser aucun point, en produisant toujours le même étiquetage, i.e. $\mathfrak{G}(\mathcal{F}, m) = 1 = \binom{m}{0}$.

— Si $m = 0$, cela signifie que nous sommes en face de l'étiquetage trivial de l'ensemble vide, i.e. $\mathfrak{G}(\mathcal{F}, 0) = 1 = \sum_{i=0}^{\mathcal{V}} \binom{0}{i}$.

Supposons maintenant que l'inégalité (17) soit vraie pour tout $m' + \mathcal{V}' < m + \mathcal{V}$. Étant donné un ensemble $S = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ et une classe de fonctions $\mathcal{F}$ avec une dimension de VC, $\mathcal{V}$, montrons que $|\mathfrak{F}(\mathcal{F}, S)| \leq \sum_{i=0}^{\mathcal{V}} \binom{m}{i}$.

Considérons deux sous-ensembles de classes $\mathcal{F}_1$ et $\mathcal{F}_2$, de $\mathcal{F}$, définis sur l'ensemble $S' = S \setminus \{\mathbf{x}_m\}$ de taille $m - 1$. Construisons la classe $\mathcal{F}_1$, en y

ajoutant l'ensemble des fonctions de $\mathcal{F}$ telles que les vecteurs de prédiction de ces fonctions sur S' soient tous différents, et posons la classe $\mathcal{F}_2 = \mathcal{F} \setminus \mathcal{F}_1$. Ainsi, si deux fonctions de $\mathcal{F}$ ont les mêmes vecteurs de prédiction sur S' et dont la seule différence réside dans leurs prédictions sur le seul exemple $\mathbf{x}_m$, l'une de ces fonctions sera mise dans $\mathcal{F}_1$ et l'autre dans $\mathcal{F}_2$. La figure 3 illustre cette construction pour un problème jouet. Les paires de fonctions (f_1, f_2) et (f_4, f_5) ont les mêmes vecteurs de prédiction sur $S' = S \setminus \{\mathbf{x}_5\}$, et les ensembles $\mathcal{F}_1$ et $\mathcal{F}_2$ vont contenir chacun une des fonctions de ces paires. Nous remarquons maintenant que si un ensemble est pulvérisé par la classe $\mathcal{F}_1$, il le sera aussi par la classe $\mathcal{F}$ puisque $\mathcal{F}_1$ contient toutes les fonctions non redondantes sur S' de $\mathcal{F}$, ainsi :

$$\text{dimension de VC}(\mathcal{F}_1) \leq \text{dimension de VC}(\mathcal{F}) = \mathcal{V}$$

De plus, si un ensemble S' est pulvérisé par $\mathcal{F}_2$, l'ensemble $S' \cup \{\mathbf{x}_m\}$ sera aussi pulvérisé par $\mathcal{F}$ puisque, pour toute fonction dans $\mathcal{F}_2$, $\mathcal{F}$ contient aussi l'autre fonction dont la sortie sur $\mathbf{x}_m$ diffère de la première.
Ainsi, dimension de $\text{VC}(\mathcal{F}) \geq$ dimension de $\text{VC}(\mathcal{F}_2) + 1$, soit :

$$\text{dimension de VC}(\mathcal{F}_2) \leq \mathcal{V} - 1$$

D'après l'hypothèse de la récurrence, nous avons :

$$|\mathcal{F}_1| = |\mathfrak{F}(\mathcal{F}_1, S')| \leq \mathfrak{G}(\mathcal{F}_1, m-1) \leq \sum_{i=0}^{\mathcal{V}} \binom{m-1}{i}$$
$$|\mathcal{F}_2| = |\mathfrak{F}(\mathcal{F}_2, S')| \leq \mathfrak{G}(\mathcal{F}_2, m-1) \leq \sum_{i=0}^{\mathcal{V}-1} \binom{m-1}{i}$$

Le raisonnement se termine après un changement de variable et l'utilisation de la formule de Pascal :

$$\begin{aligned} |\mathfrak{F}(\mathcal{F}, S)| &= |\mathcal{F}_1| + |\mathcal{F}_2| \\ &\leq \sum_{i=0}^{\mathcal{V}} \binom{m-1}{i} + \sum_{i=0}^{\mathcal{V}-1} \binom{m-1}{i} \\ &= \sum_{i=0}^{\mathcal{V}} \binom{m-1}{i} + \sum_{i=0}^{\mathcal{V}} \binom{m-1}{i-1} \\ &= \sum_{i=0}^{\mathcal{V}} \binom{m}{i} \end{aligned}$$

Ainsi, comme l'inégalité précédente est vraie pour tout ensemble S de taille m, nous avons :

$$\mathfrak{G}(\mathcal{F}, m) \leq \sum_{i=0}^{\mathcal{V}} \binom{m}{i}.$$

Pour démontrer l'inégalité (18), nous allons utiliser la formule du binôme de Newton. Ainsi, d'après l'inégalité (17) et dans le cas où $\frac{\mathcal{V}}{m} \leq 1$, nous avons :

$$\begin{aligned}\left(\frac{\mathcal{V}}{m}\right)^{\mathcal{V}} \mathfrak{G}(\mathcal{F}, m) &\leq \left(\frac{\mathcal{V}}{m}\right)^{\mathcal{V}} \sum_{i=0}^{\mathcal{V}} \binom{m}{i} \\ &\leq \sum_{i=0}^{\mathcal{V}} \left(\frac{\mathcal{V}}{m}\right)^{i} \binom{m}{i}\end{aligned}$$

En multipliant le terme de droite par $1^{m-i} = 1$ et en utilisant la formule du binôme, il vient :

$$\begin{aligned}\left(\frac{\mathcal{V}}{m}\right)^{\mathcal{V}} \mathfrak{G}(\mathcal{F}, m) &\leq \sum_{i=0}^{\mathcal{V}} \binom{m}{i} \left(\frac{\mathcal{V}}{m}\right)^{i} 1^{m-i} \\ &= \left(1 + \frac{\mathcal{V}}{m}\right)^{m}\end{aligned}$$

Finalement, en utilisant l'inégalité $\forall z \in \mathbb{R}, (1 - z) \leq e^{-z}$, nous avons :

$$\mathfrak{G}(\mathcal{F}, m) \leq \left(\frac{m}{\mathcal{V}}\right)^{\mathcal{V}} \left(1 + \frac{\mathcal{V}}{m}\right)^{m} \leq \left(\frac{m}{\mathcal{V}}\right)^{\mathcal{V}} e^{\mathcal{V}}$$

D'après le résultat précédent, nous voyons bien que les valeurs prises par la fonction de croissance associée à la classe de fonctions $\mathcal{F}$ vont dépendre de l'existence ou non de la dimension VC de $\mathcal{F}$:

$$\forall m, \mathfrak{G}(\mathcal{F}, m) = \begin{cases} O(m^{\mathcal{V}}) & \text{si } \mathcal{V} \text{ est finie,} \\ 2^m & \text{si } \mathcal{V} \text{ est infinie.} \end{cases}$$

De plus, dans le cas où la dimension VC, $\mathcal{V}$, d'une classe de fonctions $\mathcal{F}$ est finie et où on dispose d'assez d'exemples d'entraînement tels que $m \geq \mathcal{V}$, l'évolution de la fonction de croissance devient polynomiale en fonction de m, i.e. $\ln \mathfrak{G}(\mathcal{F}, 2m) \leq \mathcal{V} \ln \frac{2em}{\mathcal{V}}$ (inégalité 18). Ce résultat permet d'exhiber une nouvelle expression pour la borne de généralisation de (équation 16) estimable pour une valeur de $\mathcal{V}$ connue et n'importe quel ensemble d'apprentissage fixé.

Corollaire 1 (Borne de généralisation avec la dimension VC)**.** *Soit $\mathcal{X} \in \mathbb{R}^d$ un espace vectoriel, $\mathcal{Y} = \{-1, +1\}$ un espace de sortie et $\mathcal{F}$ une classe de fonctions à valeurs dans $\mathcal{Y}$ et de dimension VC, $\mathcal{V}$. Supposons que les paires d'exemples $(\boldsymbol{x}, y) \in \mathcal{X} \times \mathcal{Y}$ sont générées i.i.d. suivant une distribution de probabilité $\mathcal{D}$. Pour toute $\delta \in]0, 1]$, nous avons pour toute fonction $f \in \mathcal{F}$ et*

tout ensemble $S \in (\mathcal{X} \times \mathcal{Y})^m$ *de taille* $m \geq \mathcal{V}$ *générés i.i.d. suivant la même distribution de probabilité, l'inégalité suivante qui se tient avec une probabilité au moins égale à* $1 - \delta$ *:*

$$\mathfrak{L}(f) \leq \hat{\mathfrak{L}}(f,S) + \sqrt{\frac{8\mathcal{V} \ln \frac{2em}{\mathcal{V}} + 8 \ln \frac{4}{\delta}}{m}} \tag{19}$$

Ainsi, comme $\lim_{m \to \infty} \frac{8\mathcal{V} \ln \frac{2em}{\mathcal{V}} + 8 \ln \frac{4}{\delta}}{m} = 0$, nous pouvons déduire du résultat précédent une condition suffisante sur la consistance du principe MRE qui s'énonce :

Pour une classe de fonctions binaires $\mathcal{F}$ *donnée, si la dimension VC de* $\mathcal{F}$ *est finie, alors, le principe MRE est consistant pour toutes les distributions* $\mathcal{D}$ *générant les exemples.*
[19] démontre de plus que pour que le principe MRE soit consistant pour toutes les distributions $\mathcal{D}$, il est aussi nécessaire que la dimension VC de la classe de fonctions considérée soit finie. Ainsi, nous avons le résultat principal suivant :

Notion centrale

Quelle que soit la distribution de probabilité générant les exemples, le principe MRE est consistant si et seulement si la dimension VC de la classe de fonctions considérée est finie.

3.3 Énoncé du principe

D'après l'étude précédente, nous voyons bien que plus (respectivement moins) la capacité d'une classe de fonctions est grande, plus (respectivement moins) on a de possibilité d'étiqueter un ensemble d'apprentissage et moins (respectivement plus) l'erreur empirique d'une fonction de cette classe sur une base d'entraînement sera élevée, et ceci sans qu'on puisse garantir une plus faible erreur de généralisation. La difficulté de l'apprentissage est ainsi de réaliser un compromis entre une faible erreur empirique et une faible capacité de l'ensemble de fonctions pour arriver à minimiser l'erreur de généralisation. Ce compromis s'appelle la minimisation du risque structurel [20] et s'énonce comme suit (algorithme 1).

Algorithm 1 : Principe de la minimisation du risque structurel

Entrée :

— problème de prédiction, issu d'un domaine d'application.

Procédé :

— avec une connaissance a priori sur le domaine d'application, choisir une classe de fonctions (par exemple des fonctions polynomiales) ;

— diviser la classe de fonctions en une hiérarchie de sous-ensembles de fonctions imbriqués (par exemple des polynômes de degré croissant) ;

— sur une base d'entraînement, apprendre une fonction de prédiction par sous-ensemble de fonctions considéré en appliquant le principe MRE ;

Sortie : parmi l'ensemble des fonctions apprises, choisir la fonction pour laquelle nous avons la meilleure estimation de sa borne de généralisation (fonction réalisant le meilleur compromis).

En Récapitulatif

On vient de voir que :

1. Pour généraliser, il faudrait maîtriser la capacité de la classe de fonctions.
2. Le principe MRE est consistant pour toutes les distributions $\mathcal{D}$ générant les données, si et seulement si la dimension VC de la classe de fonctions considérée est finie.
3. L'étude de la consistance du principe MRE a mené au deuxième principe fondamental en apprentissage qui est le principe de la minimisation du risque structurel (MRS).
4. L'apprentissage est un compromis entre une erreur empirique faible et une capacité de la classe de fonctions forte.

Les autres chapitre de l'ouvrage [1] exposent des outils statistiques pour la dérivation de bornes de généralisations, l'optimisation convexe non-contrainte ; ainsi que la présentation des modèles classiques comme les réseaux de neurones profonds, les séparateurs à vaste marge, Boosting et les cadres d'apprentissage non-supervisé [14], semi-supervisé [2, 11] et ordonnancement [18].

Références

[1] M.R. Amini. *Machine Learning*. Eyrolles (2eme édition), 2020.

[2] M.R. Amini, N. Usunier, and F. Laviolette. A transductive bound for the voted classifier with an application to semi-supervised learning. In *Advances in Neural Information Processing Systems 21*, pages 65–72, 2009.

[3] S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification : a survey of some recent advances. *ESAIM : Probability and Statistics*, pages 323–375, 2005.

[4] O. Bousquet, S. Boucheron, and G. Lugosi. Introduction to statistical learning theory. In *Advanced Lectures on Machine Learning*, pages 169–207, 2003.

[5] H. Brönnimann and M.T. Goodrich. Almost optimal set covers in finite vc-dimension. *Discrete and Computational Geometry*, 14(4) :463–479, 1995.

[6] N. Cesa-Bianchi and D. Haussler. A graph-theoretic generalization of the sauer-shelah lemma. *Discrete Applied Mathematics*, 86 :27–35, 1998.

[7] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification*. Wiley, 2001.

[8] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, New York, USA, 1972.

[9] M.R. Genesereth and N.J. Nilsson. *Logical Foundations of Artificial Intelligence*. Morgan Kaufmann Publishers Inc., 1987.

[10] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58 :13–30, 1963.

[11] A. Krithara, M.R. Amini, J.M. Renders, and C. Goutte. Semi-supervised document classification with a mislabeling error model. In *Advances in Information Retrieval , 30th European Conference on IR Research, ECIR 2008*, pages 370–381, 2008.

[12] J. Langford. Tutorial on practical prediction theory for classification. *Journal of Machine Learning Research*, 6 :273–306, December 2005.

[13] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of Machine Learning*. MIT Press, 2012.

[14] J.-F. Pessiot, Y.-M. Kim, M.R. Amini, and P. Gallinari. Improving document clustering in a learned concept space. *Information Processing & Management*, 46(2) :180–192, 2010.

[15] N. Sauer. On the density of families of sets. *Journal of Combinatorial Theory*, 13(1) :145–147, 1972.

[16] B. Schölkopf and A.J. Smola. *Learning with kernels : support vector machines, regularization, optimization, and beyond*. MIT Press, 2002.

[17] S. Shelah. A combinatorial problem : Stability and order for models and theories in infinity languages. *Pacific Journal of Mathematics*, 41 :247–261, 1972.

[18] N. Usunier, M.R. Amini, and P. Gallinari. A data-dependent generalisation error bound for the AUC. In *ICML'05 workshop on ROC Analysis in Machine Learning*, 2005.

[19] V. N. Vapnik. *The nature of statistical learning theory (second edition)*. Springer-Verlag, 1999.

[20] V. N. Vapnik and A. J. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16 :264–280, 1971.