FONDEMENTS MATHÉMATIQUES DE L'APPRENTISSAGE STATISTIQUE

par

Christophe Giraud

Résumé. L'objectif d'un algorithme de classification est de prédire au mieux la classe d'un objet à partir d'observations de cet objet. Un exemple typique est le filtre à spam des messageries électroniques qui prédisent (plus ou moins bien) si un courriel est un spam ou non. Nous introduisons dans ces notes les principaux concepts fondamentaux de la théorie de la classification statistique supervisée et quelques uns des algorithmes de classification les plus populaires. Nous soulignons chemin faisant l'importance de certains concepts mathématiques, parmi lesquels la symétrisation, la convexification, les inégalités de concentration, le principe de contraction et les espaces de Hilbert à noyau reproduisant.

1. Introduction

En ce début de siècle, nous observons un accroissement phénoménal de l'utilisation des mathématiques, à la fois dans l'industrie et dans les laboratoires scientifiques. Cet essor de l'importance des mathématiques va de pair avec l'explosion des volumes de données collectées et l'accroissement de la puissance de calculs des ordinateurs. Dans l'industrie, la modélisation mathématique apparait à tous les stades de la vie d'un produit. Depuis la conception technique, avec force de simulations numériques, via la production, avec l'optimisation des ressources et des flux, jusqu'au marketing et la distribution avec des prévisions basées sur l'analyse de grandes bases de données. Dans les laboratoires scientifiques, la modélisation mathématique devient de plus en plus cruciale, en particulier en biologie et médecine où les scientifiques doivent extraire des informations pertinentes des

données massives qu'ils produisent grâce aux récents développements biotechnologiques.

La classification automatique est peut-être l'un des usages quotidiens les plus invasifs des mathématiques. L'objectif de la classification automatique est de prédire au mieux la classe y d'un objet x à partir d'observations de ce dernier. Un exemple typique est le filtre à spam de notre messagerie électronique qui prédit (plus ou moins bien) si un courriel est un spam ou non. La classification automatique est omniprésente dans notre quotidien, filtrant nos courriels, lisant automatiquement les codes postaux sur nos lettres ou reconnaissant les visages sur les photos que nous publions sur les réseaux sociaux. Elle est aussi extrêmement importante en sciences, par exemple en médecine pour effectuer des diagnostics précoces à partir de données à hauts débits, ou pour la recherche *in silico* de médicaments efficaces.

Nous introduisons dans ces notes les principaux concepts fondamentaux de l'apprentissage statistique (supervisé). Nous décrivons dans la partie 2 la modélisation mathématique d'un problème générique de classification. Dans la partie 3, nous analysons la précision prédictive d'un algorithme universel de classification et dans la partie 4 nous dérivons de cet algorithme théorique des algorithmes numériquement implémentables et populaires. Les appendices rassemblent des résultats techniques nécessaires pour la définition et l'analyse des algorithmes de classification.

2. Modélisation mathématique

Par soucis de simplicité, nous allons nous restreindre au cas où il y a seulement deux classes (comme pour le filtre à spam). Le problème de la classification automatique peut alors être modélisé de la manière suivante. Soit $\mathscr{X}$ un espace mesuré. On observe conjointement un point $X \in \mathscr{X}$ et une étiquette $Y \in \{-1, +1\}$. Notre objectif est de construire une fonction $h : \mathscr{X} \to \{-1, +1\}$, appelée *classifieur*, telle que $h(X)$ prédit au mieux l'étiquette Y.

Supposons que le couple $(X, Y) \in \mathscr{X} \times \{-1, +1\}$ est issu d'un tirage selon une loi $\mathbb{P}$. Pour un classifieur $h : \mathscr{X} \to \{-1, +1\}$ donné, la probabilité de mauvaise classification est

$$L(h) = \mathbb{P}(Y \neq h(X)).$$

Comme $|Y-h(X)| \in \{0,2\}$, nous avons par le théorème de Pythagore

$$\begin{aligned} L(h) &= \frac{1}{4}\,\mathbb{E}\left[(Y-h(X))^2\right] \\ &= \frac{1}{4}\,\mathbb{E}\left[(Y-\mathbb{E}\,[Y|X])^2\right] + \frac{1}{4}\,\mathbb{E}\left[(\mathbb{E}\,[Y|X]-h(X))^2\right]. \end{aligned}$$

En conséquence, $L(h)$ est minimal pour le classifieur de Bayes

$$h_*(X) = \mathrm{sign}(\mathbb{E}\,[Y|X]) \quad \text{où} \quad \mathrm{sign(x)} = \mathbf{1}_{x>0} - \mathbf{1}_{x\leqslant 0} \quad \text{pour } x \in \mathbb{R}.$$

Lorsque la loi $\mathbb{P}$ est connue, il suffit donc d'utiliser le classifieur de Bayes h_* pour avoir la plus petite probabilité possible de mauvaise classification.

Malheureusement, la loi $\mathbb{P}$ est en général inconnue et on ne peut donc pas calculer le classifieur de Bayes h_*. En pratique, nous avons uniquement accès à des données dites d'apprentissage $(X_i, Y_i)_{i=1,\dots,n}$ i.i.d. de loi $\mathbb{P}$ et notre objectif est de construire à partir de ces données d'apprentissage un classifieur $\widehat{h} : \mathscr{X} \to \{-1,+1\}$ tel que $L(\widehat{h}) - L(h_*)$ est aussi petit que possible.

3. Minimisation du risque empirique

Comme la loi $\mathbb{P}$ est inconnue, nous ne pouvons pas calculer la probabilité de mauvaise classification $L(h)$. On peut cependant calculer à la place la probabilité empirique de mauvaise classification

$$\widehat{L}_n(h) := \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{Y_i \neq h(X_i)} = \widehat{\mathbb{P}}_n(Y \neq h(X)),$$

où $\widehat{\mathbb{P}}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{(X_i,Y_i)}$. À un ensemble $\mathscr{H}$ de classifieurs, appelé *dictionnaire*, on peut associer le classifieur de minimisation du risque empirique

$$\widehat{h}_{\mathscr{H}} \in \operatorname*{argmin}_{h\in\mathscr{H}} \widehat{L}_n(h). \tag{1}$$

La définition de ce classifieur est très naturelle, cependant deux questions fondamentales se posent immédiatement : quel dictionnaire $\mathscr{H}$ choisir ? et comment $\widehat{h}_{\mathscr{H}}$ se comporte-t-il comparativement à h_* ? Ces deux questions sont bien évidemment fortement reliées. En décomposant la différence entre les probabilités de mauvaise classification

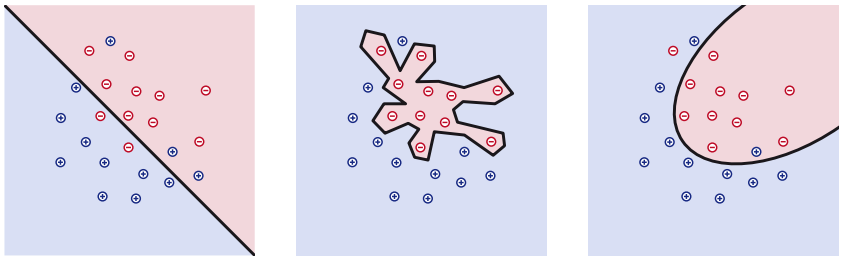

FIGURE 1. Exemples de classifications produites par différents dictionnaires. Gauche : avec des classifieurs linéaires $\mathscr{H}_{\mathsf{lin}}$. Centre : avec des classifieurs polygonaux $\mathscr{H}_{\mathsf{poly}}$. Droite : avec des classifieurs basés sur des formes quadratiques.

$L(\widehat{h}_{\mathscr{H}})$ et $L(h_*)$, on obtient

$$0 \leqslant L(\widehat{h}_{\mathscr{H}}) - L(h_*) = \underbrace{\min_{h\in\mathscr{H}} L(h) - L(h_*)}_{\text{erreur d'approximation}} + \underbrace{L(\widehat{h}_{\mathscr{H}}) - \min_{h\in\mathscr{H}} L(h)}_{\text{erreur stochastique}}.$$

Le premier terme mesure la qualité de l'approximation de h_* par un classifieur $h \in \mathscr{H}$. Cette erreur d'approximation est purement déterministe et élargir le dictionnaire $\mathscr{H}$ ne peut que la réduire. Le second terme mesure l'erreur résultant de la minimisation sur $h \in \mathscr{H}$ de la probabilité empirique de mauvaise classification $\widehat{L}_n(h)$ au lieu de la vraie probabilité de mauvaise classification $L(h)$. Ce terme est stochastique et il tend à augmenter lorsque $\mathscr{H}$ grossit. Ce phénomène est illustré Figure 1. Dans cette illustration dans $\mathscr{X} = \mathbb{R}^2$, les classifieurs du dictionnaire $\mathscr{H}_{\mathsf{lin}} = \{h(x) = \mathrm{sign}(\langle w, x\rangle) : \|w\| = 1\}$ ne sont pas suffisamment flexibles et ils produisent une classification de mauvaise qualité. Dans ce cas, l'erreur d'approximation est grande. À l'opposé, les classifieurs du dictionnaire $\mathscr{H}_{\mathsf{poly}} = \{h(x) = 2\,\mathbf{1}_A(x) - 1 : A \text{ polygone dans } \mathscr{X}\}$ sont très flexibles et ils peuvent toujours reproduire exactement la classification des données $(X_i, Y_i)_{i=1,\dots,n}$. L'erreur empirique $\widehat{L}_n(\widehat{h}_{\mathscr{H}_{\mathsf{poly}}})$ est donc toujours nulle, mais $\widehat{h}_{\mathscr{H}_{\mathsf{poly}}}$ tend a mal classifier une nouvelle donnée (X, Y) et l'erreur stochastique $L(\widehat{h}_{\mathscr{H}}) - \min_{h\in\mathscr{H}} L(h)$ est grande. Le dernier exemple, basé sur un ensemble moins flexible de classifieurs quadratiques produit de meilleurs résultats.

Pour choisir un bon dictionnaire $\mathscr{H}$, il faut donc trouver un bon équilibre entre les propriétés d'approximation de $\mathscr{H}$ et sa taille. La

première étape pour développer un principe de sélection de dictionnaire $\mathscr{H}$ est de quantifier la probabilité de mauvaise classification du minimiseur du risque empirique $\widehat{h}_{\mathscr{H}}$.

3.1. Probabilité de mauvaise classification avec $\widehat{h}_{\mathscr{H}}$

Comme mentionné ci-dessus, augmenter la taille de $\mathscr{H}$ tend à augmenter l'erreur stochastique $L(\widehat{h}_{\mathscr{H}}) - \min_{h\in\mathscr{H}} L(h)$. En fait, ce n'est pas exactement la taille du dictionnaire qui compte, mais plutôt sa flexibilité en terme de classification. Par exemple, nous ne pouvons pas classifier correctement les trois points étiquetés $\big\{\big((0,1),+1\big),\big((1,1),-1\big),\big((1,0),+1\big)\big\}$ avec les classifieurs dans $\mathscr{H}_{\mathsf{lin}}$. A l'inverse, pour tout ensemble $(x_i,y_i)_{i=1,\dots,n}$ de points étiquetés, il existe $h\in\mathscr{H}_{\mathsf{poly}}$ tel que $h(x_i)=y_i$.

Nous quantifions cette flexibilité de classification par le coefficient d'éclatement

$$\mathbb{S}_n(\mathscr{H}) = \max_{(x_1,\dots,x_n)\in\mathscr{X}^n} \operatorname{card}\{(h(x_1),\dots,h(x_n)) : h\in\mathscr{H}\}\ , \tag{2}$$

qui donne le nombre maximum d'étiquetages différents de n points qui peuvent être produits par les classifieurs dans $\mathscr{H}$. Par exemple, comme on peut obtenir à partir de $\mathscr{H}_{\mathsf{poly}}$ tous les étiquetages possibles de n points (distincts), nous avons $\mathbb{S}_n(\mathscr{H}_{\mathsf{poly}}) = 2^n$. À l'inverse, le nombre d'étiquetages possibles de n points avec les classifieurs dans $\mathscr{H}_{\mathsf{lin}}$ est plus limité. En effet, la Proposition 1 Partie 3.3 implique que $\mathbb{S}_n(\mathscr{H}_{\mathsf{lin}}) \leqslant (n+1)^2$. Le prochain théorème nous fournit une borne supérieure de l'erreur stochastique et un intervalle de confiance pour la probabilité de mauvaise classification $L(\widehat{h}_{\mathscr{H}})$ en terme du coefficient d'éclatement.

***Théorème 1* (Contrôle de l'erreur stochastique).** *Pour tout $t>0$, nous avons avec probabilité au moins $1-e^{-t}$*

$$L(\widehat{h}_{\mathscr{H}}) - \min_{h\in\mathscr{H}} L(h) \leqslant 4\sqrt{\frac{2\log(2\,\mathbb{S}_{\mathscr{H}}(n))}{n}} + \sqrt{\frac{2t}{n}} \tag{3}$$

et

$$\big|L(\widehat{h}_{\mathscr{H}}) - \widehat{L}_n(\widehat{h}_{\mathscr{H}})\big| \leqslant 2\sqrt{\frac{2\log(2\,\mathbb{S}_{\mathscr{H}}(n))}{n}} + \sqrt{\frac{t}{2n}} \tag{4}$$

Démonstration. Nous segmentons la preuve en trois lemmes. Le premier lemme montre que les termes de gauche dans (3) et (4) peuvent être bornés à l'aide du maximum sur $\mathscr{H}$ de la différence entre l'erreur empirique de mauvaise classification et la vraie erreur de mauvaise classification

$$\widehat{\Delta}_n(\mathscr{H}) = \sup_{h\in\mathscr{H}} \big|\widehat{L}_n(h) - L(h)\big|. \tag{5}$$

Lemme 1.1. *Pour tout dictionnaire $\mathscr{H}$ et entier n, nous avons les bornes supérieures*

$$L(\widehat{h}_{\mathscr{H}}) - \min_{h\in\mathscr{H}} L(h) \leqslant 2\,\widehat{\Delta}_n(\mathscr{H}) \quad \textit{et} \quad \big|L(\widehat{h}_{\mathscr{H}}) - \widehat{L}_n(\widehat{h}_{\mathscr{H}})\big| \leqslant \widehat{\Delta}_n(\mathscr{H}).$$

Démonstration. Pour tout $h \in \mathscr{H}$, nous avons $\widehat{L}_n(\widehat{h}_{\mathscr{H}}) \leqslant \widehat{L}_n(h)$ et donc

$$\begin{aligned} L(\widehat{h}_{\mathscr{H}}) - L(h) &= L(\widehat{h}_{\mathscr{H}}) - \widehat{L}_n(\widehat{h}_{\mathscr{H}}) + \widehat{L}_n(\widehat{h}_{\mathscr{H}}) - L(h) \\ &\leqslant L(\widehat{h}_{\mathscr{H}}) - \widehat{L}_n(\widehat{h}_{\mathscr{H}}) + \widehat{L}_n(h) - L(h) \\ &\leqslant 2\,\widehat{\Delta}_n(\mathscr{H}). \end{aligned}$$

Comme cette inégalité est vraie pour tout $h \in \mathscr{H}$, nous en déduisons directement la première borne du Lemme 1.1. La seconde borne est évidente. □

Pour démontrer le Théorème 1, il reste à prouver qu'on a

$$\widehat{\Delta}_n(\mathscr{H}) \leqslant 2\sqrt{\frac{2\log(2\,\mathbb{S}_{\mathscr{H}}(n))}{n}} + \sqrt{\frac{t}{2n}}$$

avec probabilité au moins $1-e^{-t}$. Nous segmentons la preuve de cette borne en deux lemmes.

Lemme 1.2. *Avec probabilité $1-e^{-t}$, nous avons*

$$\widehat{\Delta}_n(\mathscr{H}) \leqslant \mathbb{E}\left[\widehat{\Delta}_n(\mathscr{H})\right] + \sqrt{\frac{t}{2n}}\ .$$

Démonstration. Nous avons $\widehat{\Delta}_n(\mathscr{H}) = F((X_1,Y_1),\dots,(X_n,Y_n))$ avec

$$\begin{aligned} F:\quad (\mathscr{X}\times\{-1,+1\})^n \quad &\longrightarrow \quad \mathbb{R} \\ ((x_1,y_1),\dots,(x_n,y_n)) &\longmapsto \frac{1}{n}\sup_{h\in\mathscr{H}}\Big|\sum_{i=1}^n \mathbf{1}_{y_i\neq h(x_i)} - L(h)\Big|. \end{aligned}$$

Pour tout $(x_1, y_1), \ldots, (x_n, y_n), (x'_i, y'_i) \in \mathscr{X} \times \{-1, +1\}$, nous avons

$$\big|F((x_1, y_1), \ldots, (x'_i, y'_i), \ldots, (x_n, y_n)) - F((x_1, y_1), \ldots, (x_i, y_i), \ldots, (x_n, y_n))\big| \leqslant \frac{1}{n}\,.$$

En conséquence, l'inégalité de concentration de McDiarmid (voir Théorème 4 de l'Appendice B), nous assure qu'avec probabilité au moins $1-e^{-2ns^2}$, nous avons $\widehat{\Delta}_n(\mathscr{H}) \leqslant \mathbb{E}\left[\widehat{\Delta}_n(\mathscr{H})\right] + s$. Le Lemme 1.2 en découle (avec le changement de variable $s = \sqrt{t/(2n)}$). □

Il reste à borner l'espérance de $\widehat{\Delta}_n(\mathscr{H})$ à l'aide de $\mathbb{S}_{\mathscr{H}}(n)$.

Lemme 1.3. *Pour tout dictionnaire $\mathscr{H}$, nous avons l'inégalité*

$$\mathbb{E}\left[\widehat{\Delta}_n(\mathscr{H})\right] \leqslant 2\sqrt{\frac{2\log(2\,\mathbb{S}_{\mathscr{H}}(n))}{n}}\,.$$

Démonstration. La preuve du Lemme 1.3 est basée sur un argument classique et élégant de symétrisation.

La première étape de la symétrisation est de représenter la probabilité de mauvaise classification $L(h)$ comme l'espérance de la probabilité empirique de mauvaise classification

$$L(h) = \mathbb{P}(Y \neq h(X)) = \widetilde{\mathbb{E}}\left[\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\widetilde{Y}_i \neq h(\widetilde{X}_i)}\right],$$

où $(\widetilde{X}_i, \widetilde{Y}_i)_{i=1,\ldots,n}$ est indépendant de $(X_i, Y_i)_{i=1,\ldots,n}$ et est identiquement distribué. Dans la suite, nous noterons $\widetilde{\mathbb{E}}$ pour l'espérance par rapport aux variables $(\widetilde{X}_i, \widetilde{Y}_i)_{i=1,\ldots,n}$ et $\mathbb{E}$ pour l'espérance par rapport aux variables $(X_i, Y_i)_{i=1,\ldots,n}$. Les inégalités de Jensen et Fatou nous donnent

$$\begin{aligned}\mathbb{E}\left[\widehat{\Delta}_n(\mathscr{H})\right] &= \mathbb{E}\left[\sup_{h\in\mathscr{H}}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{Y_i\neq h(X_i)} - \widetilde{\mathbb{E}}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{\widetilde{Y}_i\neq h(\widetilde{X}_i)}\right]\right|\right] \\ &\leqslant \mathbb{E}\widetilde{\mathbb{E}}\left[\sup_{h\in\mathscr{H}}\Big|\frac{1}{n}\sum_{i=1}^{n}\Big(\mathbf{1}_{Y_i\neq h(X_i)} - \mathbf{1}_{\widetilde{Y}_i\neq h(\widetilde{X}_i)}\Big)\Big|\right].\end{aligned}$$

La seconde étape consiste à capitaliser sur la symétrie des variables $\mathbf{1}_{Y_i\neq h(X_i)} - \mathbf{1}_{\widetilde{Y}_i\neq h(\widetilde{X}_i)}$. Nous introduisons n variables aléatoires i.i.d. $(\sigma_i)_{i=1,\ldots,n}$ indépendantes de $(X_i, Y_i, \widetilde{X}_i, \widetilde{Y}_i)_{i=1,\ldots,n}$ et

de loi $\mathbb{P}_\sigma(\sigma_i = 1) = \mathbb{P}_\sigma(\sigma_i = -1) = 1/2$. Par symétrie, on remarque que $\left(\sigma_i(\mathbf{1}_{Y_i \neq h(X_i)} - \mathbf{1}_{\widetilde{Y}_i \neq h(\widetilde{X}_i)})\right)_{i=1,\dots,n}$ a la même loi que $\left(\mathbf{1}_{Y_i \neq h(X_i)} - \mathbf{1}_{\widetilde{Y}_i \neq h(\widetilde{X}_i)}\right)_{i=1,\dots,n}$, donc nous obtenons

$$\begin{aligned}
\mathbb{E}\left[\widehat{\Delta}_n(\mathscr{H})\right] &\leqslant \mathbb{E}\widetilde{\mathbb{E}}\mathbb{E}_\sigma\left[\sup_{h\in\mathscr{H}}\Big|\frac{1}{n}\sum_{i=1}^n \sigma_i\left(\mathbf{1}_{Y_i\neq h(X_i)} - \mathbf{1}_{\widetilde{Y}_i\neq h(\widetilde{X}_i)}\right)\Big|\right] \\
&\leqslant 2\mathbb{E}\mathbb{E}_\sigma\left[\sup_{h\in\mathscr{H}}\Big|\frac{1}{n}\sum_{i=1}^n \sigma_i \mathbf{1}_{Y_i\neq h(X_i)}\Big|\right] \\
&\leqslant 2\max_{y\in\{-1,+1\}^n}\max_{x\in\mathscr{X}^n}\mathbb{E}_\sigma\left[\sup_{h\in\mathscr{H}}\Big|\frac{1}{n}\sum_{i=1}^n \sigma_i \mathbf{1}_{y_i\neq h(x_i)}\Big|\right],
\end{aligned}$$

où la seconde inégalité découle de l'inégalité triangulaire. Pour tout $(x, y) \in \mathscr{X}^n \times \{-1, +1\}^n$, nous introduisons l'ensemble

$$\mathscr{V}_{\mathscr{H}}(x, y) = \left\{(\mathbf{1}_{y_1\neq h(x_1)}, \dots, \mathbf{1}_{y_n\neq h(x_n)}) : \ h \in \mathscr{H}\right\}.$$

La dernière majoration de $\mathbb{E}\left[\widehat{\Delta}_n(\mathscr{H})\right]$ peut s'écrire comme

$$\mathbb{E}\left[\widehat{\Delta}_n(\mathscr{H})\right] \leqslant \frac{2}{n} \times \max_{y\in\{-1,+1\}^n}\max_{x\in\mathscr{X}^n}\mathbb{E}_\sigma\left[\sup_{v\in\mathscr{V}_{\mathscr{H}}(x,y)}|\langle\sigma, v\rangle|\right],$$

où $\langle x, y\rangle$ est le produit scalaire canonique sur $\mathbb{R}^n$. Remarquons au passage qu'entre le membre de gauche et le membre de droite nous avons remplacé l'espérance sous la probabilité inconnue $\mathbb{P}$ par une espérance sous la probabilité connue $\mathbb{P}_\sigma$. Pour tout $y \in \{-1, +1\}^n$ remarquons aussi qu'il y a une bijection entre $\mathscr{V}_{\mathscr{H}}(x, y)$ et l'ensemble $\{(h(x_1), \dots, h(x_n)) : h \in \mathscr{H}\}$. En conséquence, nous avons la majoration

$$\max_{y\in\{-1,+1\}^n}\max_{x\in\mathscr{X}^n}\operatorname{card}(\mathscr{V}_{\mathscr{H}}(x, y)) \leqslant \mathbb{S}_n(\mathscr{H}).$$

Au regard des deux dernières inégalités, il suffit de démontrer que

$$\mathbb{E}_\sigma\left[\sup_{v\in\mathscr{V}}|\langle\sigma, v\rangle|\right] \leqslant \sqrt{2n\log\left(2\operatorname{card}(\mathscr{V})\right)}, \tag{6}$$

pour tout ensemble $\mathscr{V} \subset \{-1, 0, +1\}^n$

pour conclure la preuve du Lemme 1.3. Démontrons (6). Avec la notation $\mathscr{V}^{\#} = \mathscr{V} \cup -\mathscr{V}$, l'inégalité de Jensen nous donne pour tout $s > 0$

$$\mathbb{E}_\sigma \left[\sup_{v \in \mathscr{V}} |\langle \sigma, v \rangle| \right] = \mathbb{E}_\sigma \left[\sup_{v \in \mathscr{V}^{\#}} \langle \sigma, v \rangle \right] \leqslant \frac{1}{s} \log \mathbb{E}_\sigma \left[\sup_{v \in \mathscr{V}^{\#}} e^{s\langle \sigma, v \rangle} \right]$$

$$\leqslant \frac{1}{s} \log \left(\sum_{v \in \mathscr{V}^{\#}} \mathbb{E}_\sigma \left[e^{s\langle \sigma, v \rangle} \right] \right). \tag{7}$$

En combinant le fait que les σ_i sont indépendants, que $(e^x + e^{-x}) \leqslant 2e^{x^2/2}$ pour tout $x \in \mathbb{R}$ et que $v_i^2 \in \{0, 1\}$ pour tout $v \in \mathscr{V}^{\#}$, nous obtenons

$$\mathbb{E}_\sigma \left[e^{s\langle \sigma, v \rangle} \right] = \prod_{i=1}^{n} \mathbb{E}_\sigma \left[e^{s v_i \sigma_i} \right] = \prod_{i=1}^{n} \frac{1}{2} (e^{s v_i} + e^{-s v_i})$$

$$\leqslant \prod_{i=1}^{n} e^{s^2 v_i^2 / 2} \leqslant e^{n s^2 / 2}.$$

En injectant cette majoration dans (7), nous obtenons

$$\mathbb{E}_\sigma \left[\sup_{v \in \mathscr{V}} |\langle \sigma, v \rangle| \right] \leqslant \frac{\log(\operatorname{card}(\mathscr{V}^{\#})}{s} + \frac{ns}{2} \qquad \text{pour tout } s > 0.$$

Le terme de droite est minimal pour $s = \sqrt{2 \log(\operatorname{card}(\mathscr{V}^{\#}))/n}$, ce qui nous donne la majoration

$$\mathbb{E}_\sigma \left[\sup_{v \in \mathscr{V}} |\langle \sigma, v \rangle| \right] \leqslant \sqrt{2n \log(\operatorname{card}(\mathscr{V}^{\#}))}\ .$$

Nous obtenons finalement la borne (6) en remarquant que $\operatorname{card}(\mathscr{V}^{\#}) \leqslant 2 \operatorname{card}(\mathscr{V})$. La preuve du Lemme 1.3 est complète. □

Les bornes (3) et (4) s'obtiennent directement en combinant les trois lemmes. □

3.2. Sélection de dictionnaire

Soit $\{\mathscr{H}_1, \ldots, \mathscr{H}_M\}$ une collection de dictionnaires de classifieurs. Nous aimerions sélectionner parmi cette collection, le dictionnaire $\mathscr{H}_*$ avec la plus petite probabilité de mauvaise classification $L(\widehat{h}_{\mathscr{H}_*})$. Ce dictionnaire $\mathscr{H}_*$, dit *dictionnaire oracle*, dépend de la probabilité inconnue $\mathbb{P}$, il n'est donc malheureusement pas connu du statisticien. Dans cette partie, nous allons développer à partir du Théorème 1 une

procédure reposant sur les données pour sélectionner parmi la collection $\{\mathscr{H}_1, \dots, \mathscr{H}_M\}$ un dictionnaire $\mathscr{H}_{\widehat{m}}$ ayant des performances similaires à celles de $\mathscr{H}_*$.

Le dictionnaire oracle $\mathscr{H}_*$ est obtenu en minimisant la probabilité de mauvaise classification $L(\widehat{h}_{\mathscr{H}})$ sur $\mathscr{H} \in \{\mathscr{H}_1, \dots, \mathscr{H}_M\}$. Une première idée est de sélectionner $\mathscr{H}_{\widehat{m}}$ en minimisant sur la collection $\{\mathscr{H}_1, \dots, \mathscr{H}_M\}$ la probabilité empirique de mauvaise classification $\widehat{L}_n(\widehat{h}_{\mathscr{H}})$. Cette procédure de sélection ne donne pas de bons résultats car pour tout $\mathscr{H} \subset \mathscr{H}'$ on a toujours $\widehat{L}_n(\widehat{h}_{\mathscr{H}'}) \leqslant \widehat{L}_n(\widehat{h}_{\mathscr{H}})$, donc la procédure tend à sélectionner le dictionnaire le plus gros possible. Pour bâtir une bonne procédure de sélection, nous devons prendre en compte les fluctuations de $\widehat{L}_n(\widehat{h}_{\mathscr{H}})$ autour de $L(\widehat{h}_{\mathscr{H}})$. La borne (4) du Théorème 1 nous offre un contrôle de ces fluctuations. En élaborant à partir de cette borne, nous obtenons le résultat suivant.

***Théorème 2* (Sélection de dictionnaire).** *Considérons la procédure de sélection de dictionnaire*

$$\widehat{m} = \operatorname*{argmin}_{m=1,\dots,M} \left\{ \widehat{L}_n(\widehat{h}_{\mathscr{H}_m}) + \operatorname{pen}(\mathscr{H}_m) \right\}, \qquad \text{avec} \quad \operatorname{pen}(\mathscr{H}) = 2\sqrt{\frac{2\log(2\,\mathbb{S}_n(\mathscr{H}))}{n}}\,.$$

Alors, pour tout $t > 0$, avec probabilité au moins $1 - e^{-t}$ nous avons

$$L(\widehat{h}_{\mathscr{H}_{\widehat{m}}}) \leqslant \min_{m=1,\dots,M} \left\{ \inf_{h \in \mathscr{H}_m} L(h) + 2\operatorname{pen}(\mathscr{H}_m) \right\} + \sqrt{\frac{2\log(M) + 2t}{n}}\,. \tag{8}$$

Avant de prouver le Théorème 2, commentons la borne (8). Comme $\min_{h\in\mathscr{H}} L(h) \leqslant L(\widehat{h}_{\mathscr{H}})$, nous obtenons avec probabilité $1 - e^{-t}$ que

$$L(\widehat{h}_{\mathscr{H}_{\widehat{m}}}) \leqslant L(\widehat{h}_{\mathscr{H}_*}) + 2\operatorname{pen}(\mathscr{H}_*) + \sqrt{\frac{2\log(M) + 2t}{n}}\,.$$

En particulier, cela nous permet de comparer la probabilité de mauvaise classification du classifieur sélectionné à celle du meilleur classifieur parmi la collection $\{\widehat{h}_{\mathscr{H}_1}, \dots, \widehat{h}_{\mathscr{H}_M}\}$.

Remarquons aussi que le second terme de la borne (8) augmente en $\sqrt{2\log(M)/n}$ avec le nombre M de dictionnaires candidats. Enfin, nous noterons que les résultats restent valides si on prend $\operatorname{pen}(\mathscr{H})$ plus grand que $2\sqrt{2\log(2\,\mathbb{S}_n(\mathscr{H}))/n}$.

Démonstration du Théorème 2

Rappelons la notation $\widehat{\Delta}_n(\mathscr{H}) = \sup_{h\in\mathscr{H}} \left|\widehat{L}_n(h) - L(h)\right|$. Les Lemmes 1.2 et 1.3 nous assurent qu'avec probabilité au moins $1 - e^{-t}$ nous avons

$$\widehat{\Delta}_n(\mathscr{H}_m) \leqslant \mathrm{pen}(\mathscr{H}_m) + \sqrt{\frac{\log(M)+t}{2n}}, \quad \textit{simultanément} \text{ pour tout } m = 1,\dots,M. \tag{9}$$

En conséquence, le Lemme 1.1 et la définition du critère de sélection nous donnent qu'avec probabilité au moins $1 - e^{-t}$

$$\begin{aligned} L(\widehat{h}_{\mathscr{H}_{\widehat{m}}}) &\leqslant \widehat{L}_n(\widehat{h}_{\mathscr{H}_{\widehat{m}}}) + \mathrm{pen}(\mathscr{H}_{\widehat{m}}) + \sqrt{\frac{\log(M)+t}{2n}} \\ &\leqslant \min_{m=1,\dots,M} \left\{ \widehat{L}_n(\widehat{h}_{\mathscr{H}_m}) + \mathrm{pen}(\mathscr{H}_m) \right\} + \sqrt{\frac{\log(M)+t}{2n}}\,. \end{aligned} \tag{10}$$

Pour conclure, il nous suffit de contrôler $\widehat{L}_n(\widehat{h}_{\mathscr{H}_m})$ en fonction de $\inf_{h\in\mathscr{H}_m} L(h)$. Cela peut être fait directement en combinant les inégalités (3) et (4), mais la majoration résultante peut être améliorée par le raisonnement ci-dessous.

Pour comparer $\widehat{L}_n(\widehat{h}_{\mathscr{H}_m})$ à $\inf_{h\in\mathscr{H}_m} L(h)$, commençons par remarquer que pour tout $h \in \mathscr{H}_m$ nous avons

$$\widehat{L}_n(\widehat{h}_{\mathscr{H}_m}) \leqslant \widehat{L}_n(h) \leqslant L(h) + \widehat{\Delta}_n(\mathscr{H}_m),$$

donc en prenant l'infimum sur $h \in \mathscr{H}_m$ nous obtenons pour tout $m = 1,\dots,M$

$$\widehat{L}_n(\widehat{h}_{\mathscr{H}_m}) \leqslant \inf_{h\in\mathscr{H}_m} L(h) + \widehat{\Delta}_n(\mathscr{H}_m).$$

En combinant cette majoration avec (9) et (10), on trouve

$$L(\widehat{h}_{\mathscr{H}_{\widehat{m}}}) \leqslant \min_{m=1,\dots,M} \left\{ \inf_{h\in\mathscr{H}_m} L(h) + 2\,\mathrm{pen}(\mathscr{H}_m) \right\} + 2\sqrt{\frac{\log(M)+t}{2n}}\,.$$

La preuve du Théorème 2 est complète. □

Remarque. En combinant (9) avec le Lemme 1.1, on obtient un intervalle de confiance pour la probabilité de mauvaise classification

$$\mathbb{P}\left(L(\widehat{h}_{\mathscr{H}_{\widehat{m}}}) \in \left[\widehat{L}_n(\widehat{h}_{\mathscr{H}_{\widehat{m}}}) - \delta(\widehat{m},t), \widehat{L}_n(\widehat{h}_{\mathscr{H}_{\widehat{m}}}) + \delta(\widehat{m},t) \right] \right) \geqslant 1 - e^{-t}$$

$$\text{avec} \quad \delta(\widehat{m},t) = \mathrm{pen}(\mathscr{H}_{\widehat{m}}) + \sqrt{\frac{\log(M)+t}{2n}}\,.$$

3.3. Dimension combinatoire de Vapnik-Chervonenkis

Le calcul des coefficients d'éclatement $\mathbb{S}_n(\mathscr{H})$ peut être complexe en pratique. Cependant, une propriété combinatoire remarquable de ces coefficients d'éclatement permet d'obtenir une majoration simple de $\mathbb{S}_n(\mathscr{H})$, ne dépendant de $\mathscr{H}$ qu'à travers une unique quantité, la dimension de Vapnik-Chervonenkis de $\mathscr{H}$ définie ci-dessous.

Par convention on posera $\mathbb{S}_0(\mathscr{H}) = 1$. On appelle VC-dimension de $\mathscr{H}$ l'entier $d_{\mathscr{H}}$ défini par

$$d_{\mathscr{H}} = \sup\left\{d \in \mathbb{N} : \mathbb{S}_d(\mathscr{H}) = 2^d\right\} \ \in \ \mathbb{N} \cup \{+\infty\}\,.$$

Elle correspond au nombre maximum de points de $\mathscr{X}$ qui peuvent être classifiés de toutes les façons possibles par les classifieurs de $\mathscr{H}$. La proposition suivante donne une majoration du coefficient d'éclatement $\mathbb{S}_n(\mathscr{H})$ en fonction de la VC-dimension $d_{\mathscr{H}}$.

***Proposition 1* (Lemme de Sauer).** *Soit $\mathscr{H}$ un ensemble de classifieurs de VC-dimension $d_{\mathscr{H}}$ finie. Pour tout $n \in \mathbb{N}$, on a*

$$\mathbb{S}_n(\mathscr{H}) \leqslant \sum_{i=0}^{d_{\mathscr{H}}} C_n^i \leqslant (n+1)^{d_{\mathscr{H}}} \quad \textit{avec} \quad C_n^i = \begin{cases} \dfrac{n!}{i!\,(n-i)!} & \textit{pour } n \geqslant i \\ 0 & \textit{pour } n < i. \end{cases}$$

Démonstration. Nous commençons par prouver par récurrence sur k l'inégalité

$$\mathbb{S}_k(\mathscr{H}) \leqslant \sum_{i=0}^{d_{\mathscr{H}}} C_k^i \tag{11}$$

pour tout $\mathscr{H}$ de VC-dimension finie $d_{\mathscr{H}}$.

Considérons dans un premier temps le cas $k = 1$. Si $d_{\mathscr{H}} = 0$, aucun point ne peut être étiqueté de deux manières différentes, donc tous les points ne peuvent être étiquetés que d'une seule manière. Il en découle que $\mathbb{S}_1(\mathscr{H}) = 1$, qui est égal à C_1^0. Si $d_{\mathscr{H}} \geqslant 1$, nous avons $\mathbb{S}_1(\mathscr{H}) = 2$ qui est égal à $C_1^0 + C_1^1$.

Supposons maintenant que (11) est vraie pour tout $k \leqslant n-1$. Considérons $\mathscr{H}$ de VC-dimension $d_{\mathscr{H}}$ finie. Comme on vient de le voir, lorsque $d_{\mathscr{H}} = 0$ tous les points peuvent être étiqueté d'une seule manière donc $\mathbb{S}_k(\mathscr{H}) = 1$ et (11) est vraie pour tout k. Considérons maintenant le cas $d_{\mathscr{H}} \geqslant 1$. Soient $x_1, \ldots, x_n \in \mathscr{X}$ et définissons

$$\mathscr{H}(x_1, \ldots, x_n) = \{(h(x_1), \ldots, h(x_n)) : h \in \mathscr{H}\}\,.$$

L'ensemble $\mathscr{H}(x_1,\dots,x_n)$ ne dépend que des valeurs de h sur $\{x_1,\dots,x_n\}$, donc on peut remplacer $\mathscr{H}$ par $\mathscr{F}=\{h|_{\{x_1,\dots,x_n\}} : h\in\mathscr{H}\}$ dans la définition de $\mathscr{H}(x_1,\dots,x_n)$. Comme $d_{\mathscr{F}}$ est inférieur à $d_{\mathscr{H}}$, on peut supposer sans perte de généralité que $\mathscr{X}=\{x_1,\dots,x_n\}$ et $\mathscr{H}=\mathscr{F}$, ce que nous faisons par la suite. Considérons l'ensemble

$$\mathscr{H}'=\left\{h\in\mathscr{H} : h(x_n)=1 \text{ et } h'=h-2\times\mathbf{1}_{\{x_n\}}\in\mathscr{H}\right\}.$$

Comme $\mathscr{H}(x_1,\dots,x_n)=\mathscr{H}'(x_1,\dots,x_n)\cup(\mathscr{H}\smallsetminus\mathscr{H}')(x_1,\dots,x_n)$ nous avons

$$\begin{aligned}(12)\quad \operatorname{card}\left(\mathscr{H}(x_1,\dots,x_n)\right)\leqslant{}&\operatorname{card}\left(\mathscr{H}'(x_1,\dots,x_n)\right)\\&+\operatorname{card}\left((\mathscr{H}\smallsetminus\mathscr{H}')(x_1,\dots,x_n)\right).\end{aligned}$$

Nous allons majorer séparément le cardinal de $\mathscr{H}'(x_1,\dots,x_n)$ et celui de $(\mathscr{H}\smallsetminus\mathscr{H}')(x_1,\dots,x_n)$.

(1) Pour commencer, remarquons que $\operatorname{card}\left(\mathscr{H}'(x_1,\dots,x_n)\right)=\operatorname{card}\left(\mathscr{H}'(x_1,\dots,x_{n-1})\right)$ car on a $h(x_n)=1$ pour tout $h\in\mathscr{H}'$. Ensuite, on observe que la VC-dimension $d_{\mathscr{H}'}$ de $\mathscr{H}'$ est au plus $d_{\mathscr{H}}-1$. En effet, si d points $x_{i_1},\dots,x_{i_d}$ de $\mathscr{X}=\{x_1,\dots,x_n\}$ peuvent être étiquetés de toutes les manières possibles par $\mathscr{H}'$, alors $x_n\notin\{x_{i_1},\dots,x_{i_d}\}$ vu que $h(x_n)=1$ pour tout $h\in\mathscr{H}'$. De plus, l'ensemble $\{x_{i_1},\dots,x_{i_d},x_n\}$ peut être étiqueté de toutes les manières possibles par $\mathscr{H}$ par définition de $\mathscr{H}'$, donc $d+1\leqslant d_{\mathscr{H}}$, ce qui implique $d_{\mathscr{H}'}\leqslant d_{\mathscr{H}}-1$. En appliquant (11) avec $k=n-1$ on obtient

$$\begin{aligned}(13)\quad \operatorname{card}\left(\mathscr{H}'(x_1,\dots,x_n)\right)&=\operatorname{card}\left(\mathscr{H}'(x_1,\dots,x_{n-1})\right)\\&\leqslant\sum_{i=0}^{d_{\mathscr{H}}-1}C_{n-1}^{i}.\end{aligned}$$

(2) Si $h,h'\in\mathscr{H}\smallsetminus\mathscr{H}'$ vérifient $h(x_i)=h'(x_i)$ pour $i=1,\dots,n-1$, alors ils vérifient aussi $h(x_n)=h'(x_n)$ par définition de $\mathscr{H}'$. En conséquence, on a comme précédemment l'égalité

$$\operatorname{card}\left((\mathscr{H}\smallsetminus\mathscr{H}')(x_1,\dots,x_n)\right)=\operatorname{card}\left((\mathscr{H}\smallsetminus\mathscr{H}')(x_1,\dots,x_{n-1})\right).$$

De plus $d_{\mathscr{H}\setminus\mathscr{H}'}$ est inférieur à $d_{\mathscr{H}}$ car $\mathscr{H}\setminus\mathscr{H}'\subset\mathscr{H}$ donc l'équation (11) avec $k=n-1$ donne

$$\begin{aligned}(14)\quad &\operatorname{card}\left((\mathscr{H}\setminus\mathscr{H}')(x_1,\dots,x_n)\right)\\ &\qquad=\operatorname{card}\left((\mathscr{H}\setminus\mathscr{H}')(x_1,\dots,x_{n-1})\right)\leqslant\sum_{i=0}^{d_{\mathscr{H}}}C_{n-1}^i\,.\end{aligned}$$

En combinant (12), (13) et (14), on obtient

$$\operatorname{card}\left(\mathscr{H}(x_1,\dots,x_n)\right)\leqslant\sum_{i=1}^{d_{\mathscr{H}}}C_{n-1}^{i-1}+\sum_{i=0}^{d_{\mathscr{H}}}C_{n-1}^i=\sum_{i=0}^{d_{\mathscr{H}}}C_n^i\,,$$

car $C_{n-1}^i+C_{n-1}^{i-1}=C_n^i$ pour $i\geqslant 1$. Il en découle que (11) est vrai pour $k=n$ et la récurrence est établie.

La seconde borne de la proposition est obtenue par

$$\sum_{i=0}^{d}C_n^i\leqslant\sum_{i=0}^{d}\frac{n^i}{i!}\leqslant\sum_{i=0}^{d}C_d^i\,n^i=(1+n)^d.$$

La preuve de la Proposition 1 est complète. □

Donnons quelques exemples de VC-dimension pour quelques dictionnaires simples de $\mathscr{X}=\mathbb{R}^d$. Les preuves sont laissées en exercice.

Exemple 1 : classifieurs linéaires. La VC-dimension de l'ensemble $\mathscr{H}=\{h(x)=\operatorname{sign}(\langle w,x\rangle):\|w\|=1\}$ de classifieurs linéaires est d.

Exemple 2 : classifieurs affines. La VC-dimension de l'ensemble

$$\mathscr{H}=\{h(x)=\operatorname{sign}(\langle w,x\rangle+b):\|w\|=1,\ b\in\mathbb{R}\}$$

de classifieurs affines est $d+1$.

Exemple 3 : classifieurs hyper-rectangulaires. La VC-dimension de l'ensemble

$$\mathscr{H}=\left\{h(x)=2\,\mathbf{1}_A(x)-1:A\text{ hyper-rectangle de }\mathbb{R}^d\right\}$$

de classifieurs hyper-rectangulaires est $2d$.

Exemple 4 : classifieurs à polygones convexes. La VC-dimension de l'ensemble

$$\mathscr{H}=\left\{h(x)=2\,\mathbf{1}_A(x)-1:A\text{ polygone convexe de }\mathbb{R}^d\right\}$$

de classifieurs par polygones convexes est $+\infty$ (prendre n points sur la sphère unité : pour tout sous-ensemble de ces points, on peut prendre leur enveloppe convexe comme polygone convexe).

4. De la théorie vers la pratique

4.1. Convexification du risque empirique

Les classifieurs par minimisation du risque empirique analysés précédemment possèdent de bonnes qualités statistiques, mais ils ne peuvent pas être mis en œuvre en pratique à cause de leur coût en temps de calcul. En effet, il n'existe pas de façon efficace de minimiser (1) car ni $\mathscr{H}$ ni $\widehat{L}_n$ ne sont convexes. Un certain nombre des algorithmes de classification les plus populaires sont obtenus par une simple relaxation convexe du problème de minimisation (1). La probabilité empirique de mauvaise classification $\widehat{L}_n$ est remplacée par un substitut convexe et l'ensemble des classifieurs $\mathscr{H}$ est remplacé par un ensemble fonctionnel convexe $\mathscr{F} \subset \mathbb{R}^{\mathscr{X}}$.

Considérons un ensemble convexe $\mathscr{F}$ de fonctions de $\mathscr{X}$ dans $\mathbb{R}$. Une fonction $f \in \mathscr{F}$ n'est pas un classifieur, mais on peut l'utiliser pour classifier en étiquetant les points en fonction du signe de f. En d'autres mots, on peut associer à f le classifieur $\operatorname{sign}(f)$. La probabilité empirique de mauvaise classification de ce classifieur peut être réécrite sous la forme

$$\widehat{L}_n(\operatorname{sign}(f)) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{Y_i \operatorname{sign}(f)(X_i)<0\}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{Y_i f(X_i)<0\}}.$$

Remplaçons maintenant la probabilité empirique de mauvaise classification $\widehat{L}_n$ par un substitut convexe se prêtant bien au calcul numérique. Un moyen simple et efficace d'obtenir un critère convexe est de remplacer la fonction de perte $z \to \mathbf{1}_{z<0}$ par une fonction convexe $z \to \ell(z)$. Nous allons développer cette idée et considérer dans la suite les classifieurs obtenus par la procédure

$$\widehat{h}_{\mathscr{F}} = \operatorname{sign}(\widehat{f}_{\mathscr{F}}) \tag{15}$$
$$\text{où} \quad \widehat{f}_{\mathscr{F}} = \operatorname*{argmin}_{f\in\mathscr{F}} \widehat{L}_n^{\ell}(f) \quad \text{avec} \quad \widehat{L}_n^{\ell}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i f(X_i)).$$

Ce classifieur peut être calculé numériquement de façon efficace car à la fois $\mathscr{F}$ et $\widehat{L}_n^{\ell}$ sont convexes. De nombreux classifieurs populaires

sont obtenus en résolvant (15) pour des choix spécifiques de $\mathscr{F}$ et ℓ, voir par exemple les Parties 4.3 et 4.4 pour quelques exemples.

Quelques fonctions de perte convexes ℓ classiques. Il est naturel de considérer une fonction de perte ℓ qui est décroissante et positive. Habituellement, on demande aussi que $\ell(z) \geqslant \mathbf{1}_{z<0}$ pour tout $z \in \mathbb{R}$ car cela nous permet de donner une majoration de la probabilité de mauvaise classification, voir le Théorème 3. Quelques fonctions de pertes classiques sont

- la perte exponentielle $\ell(z) = e^{-z}$
- la perte logit $\ell(z) = \log_2(1 + e^{-z})$
- la perte hinge $\ell(z) = (1 - z)_+ \quad (\text{avec } (x)_+ = \max(0, x))$

voir la Figure 2 pour un tracé de ces trois fonctions.

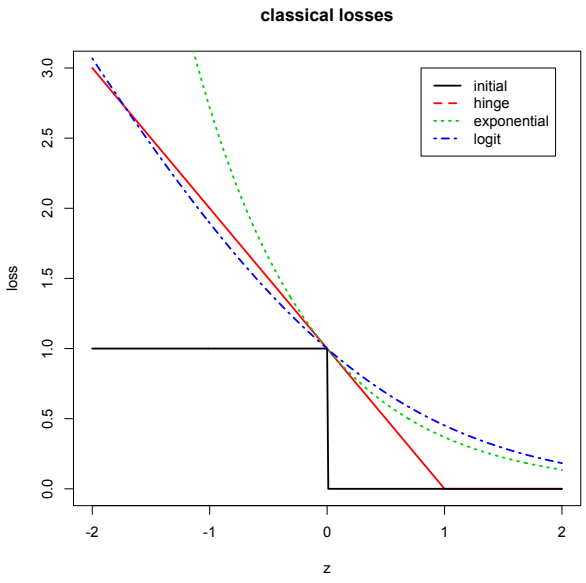

FIGURE 2. Tracé des pertes exponentielle, hinge et logit

Quelques exemples classiques d'ensembles fonctionnels $\mathscr{F}$. Les principaux exemples classiques d'ensembles fonctionnels $\mathscr{F}$ peuvent être regroupés en deux classes.

Une première classe d'ensembles $\mathscr{F}$ est obtenue en prenant des combinaisons linéaires d'une famille finie $\mathscr{H} = \{h_1, \dots, h_p\}$ de classifieurs

$$\mathscr{F} = \left\{ f : f(x) = \sum_{j=1}^{p} \beta_j h_j(x) \text{ avec } \beta_j \in \mathscr{C} \right\}, \tag{16}$$

où $\mathscr{C}$ est un sous-ensemble convexe de $\mathbb{R}^p$. Des choix typiques pour $\mathscr{C}$ sont le simplexe $\left\{\beta \in \mathbb{R}^p : \beta_j \geqslant 0,\ \sum_{j=1}^p \beta_j \leqslant 1\right\}$, la boule $\{\beta \in \mathbb{R}^p : |\beta|_1 \leqslant R\}$ ou tout l'ensemble $\mathbb{R}^p$. Ces choix apparaissent notamment dans les méthodes de "boosting", voir Partie 4.4. Les classifieurs de base $\{h_1, \dots, h_p\}$ sont souvent appelés *weak learners*. Un choix populaire pour les weak learners est $h_j(x) = \text{sign}(x_j - t_j)$ avec $t_j \in \mathbb{R}$.

Un seconde classe classique d'ensembles $\mathscr{F}$ est obtenue en prenant une boule d'une espace de Hilbert à noyaux reproduisant (qu'on appelle RKHS, pour *Reproducing Kernel Hilbert Space* en anglais). On renvoie à l'Appendice A pour une brève introduction aux RKHS. Plus précisément, considérons un RKHS $\mathscr{F}_k$ de noyau k et notons par $\|f\|_{\mathscr{F}}$ la norme hilbertienne de $f \in \mathscr{F}_k$. Pour alléger les notations, dans la suite on écrira simplement $\mathscr{F}$ pour $\mathscr{F}_k$. Minimiser $\widehat{L}_n^\ell$ sur la boule $\{f \in \mathscr{F} : \|f\|_{\mathscr{F}} \leqslant R\}$ est équivalent à minimiser sur $\mathscr{F}$ le problème lagrangien dual

$$\widehat{f}_{\mathscr{F}} = \operatorname*{argmin}_{f \in \mathscr{F}} \widetilde{L}_n^\ell(f) \qquad \text{avec} \quad \widetilde{L}_n^\ell(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i\, f(X_i)) + \lambda \|f\|_{\mathscr{F}}^2, \tag{17}$$

pour un certain $\lambda > 0$. Ce type de classifieur apparaît notamment dans les algorithmes de type *Support Vector Machine*, présentés Partie 4.3. Il s'avère que le minimiseur $\widehat{f}_{\mathscr{F}}$ de (17) est de la forme

$$\widehat{f}_{\mathscr{F}} = \sum_{i=1}^{n} \widehat{\beta}_i\, k(X_i, .). \tag{18}$$

En effet, notons V l'espace linéaire engendré par $k(X_1, .), \dots, k(X_n, .)$ et décomposons $f = f_V + f_{V^\perp}$ sur $V \oplus V^\perp$. Par la propriété de reproduction on a $f(X_i) = \langle f, k(X_i, .)\rangle_{\mathscr{F}} = \langle f_V, k(X_i, .)\rangle_{\mathscr{F}} = f_V(X_i)$,

donc la formule de Pythagore nous donne

$$\widetilde{L}_n^\ell(f_V + f_{V^\perp}) = \frac{1}{n}\sum_{i=1}^n \ell(Y_i f_V(X_i)) + \lambda\|f_V\|_{\mathscr{F}}^2 + \lambda\|f_{V^\perp}\|_{\mathscr{F}}^2 .$$

Comme λ est strictement positif, tout minimiseur de $\widehat{f}$ of $\widetilde{L}_n^\ell$ doit satisfaire $\widehat{f}_{V^\perp} = 0$, donc il est de la forme (18). De plus, la propriété de reproduction nous assure de nouveau que $\langle k(X_i,.), k(X_j,.)\rangle_{\mathscr{F}} = k(X_i, X_j)$ donc

$$\left\| \sum_{j=1}^n \beta_j k(X_j,.) \right\|_{\mathscr{F}}^2 = \sum_{i,j=1}^n \beta_i\beta_j k(X_i, X_j).$$

Le problème de minimisation (17) est donc équivalent à

$$\widehat{f}_{\mathscr{F}} = \sum_{j=1}^n \widehat{\beta}_j\, k(X_j,.) \quad \text{avec} \tag{19}$$

$$\widehat{\beta} = \operatorname*{argmin}_{\beta\in\mathbb{R}^n} \left\{ \frac{1}{n}\sum_{i=1}^n \ell\Big(\sum_{j=1}^n \beta_j Y_i k(X_j, X_i)\Big) + \lambda \sum_{i,j=1}^n \beta_i\beta_j k(X_i, X_j) \right\}.$$

Cette formulation est cruciale en pratique, car elle réduit le problème de minimisation infini-dimensionnel (17) à un problème de minimisation n-dimensionnel qui peut être résolu efficacement. Dans la Partie 4.3 sur les Support Vector Machines, nous donnerons une description plus précise des solutions de ce problème lorsque ℓ est la perte Hinge.

4.2. Propriétés statistiques

Le classifieur $\widehat{h}_{\mathscr{F}}$ donné par (15) avec $\mathscr{F}$ et ℓ convexes a l'avantage de se calculer facilement, mais a-t-il de bonnes propriétés statistiques ?

Lien avec le classifieur de Bayes. Le minimiseur du risque empirique $\widehat{h}_{\mathscr{H}}$ de la Partie 3 était un minimiseur de la version empirique de la probabilité de mauvaise classification $\mathbb{P}(Y \neq h(X))$ sur un ensemble $\mathscr{H}$ de classifieurs. La fonction $\widehat{f}_{\mathscr{F}}$ minimise à la place la version empirique de $\mathbb{E}\left[\ell(Y f(X))\right]$ sur un ensemble fonctionnel $\mathscr{F}$. Le classifieur $\widehat{h}_{\mathscr{H}}$ peut donc être vu comme une version empirique du classifieur de Bayes h_* qui minimise $\mathbb{P}(Y \neq h(X))$ sur l'ensemble des fonctions mesurables $h : \mathscr{X} \to \{-1, +1\}$, alors que la fonction $\widehat{f}_{\mathscr{F}}$ est

une version empirique de la fonction f_*^ℓ qui minimise $\mathbb{E}\,[\ell(Yf(X))]$ sur l'ensemble des fonctions mesurables $f : \mathscr{X} \to \mathbb{R}$. Une première étape est de comprendre le lien entre le classifieur de Bayes h_* et le signe de la fonction f_*^ℓ. Il s'avère que sous des hypothèses très faibles sur ℓ, le signe de f_*^ℓ coïncide exactement avec le classifieur de Bayes h_*, donc $\mathrm{sign}(f_*^\ell)$ minimise la probabilité de mauvaise classification $\mathbb{P}(Y \neq h(X))$. Démontrons maintenant ce résultat.

En conditionnant par X on obtient

$$\begin{aligned}\mathbb{E}\,[\ell(Yf(X)] &= \mathbb{E}\,[\mathbb{E}\,[\ell(Yf(X)|X]] \\ &= \mathbb{E}\Big[\ell(f(X))\mathbb{P}(Y=1|X) + \ell(-f(X))(1-\mathbb{P}(Y=1|X))\Big].\end{aligned}$$

Supposons que ℓ est décroissante, dérivable et strictement convexe (par exemple perte exponentielle ou logit). En minimisant l'expression ci-dessus on obtient que $f_*^\ell(X)$ est solution de

$$\frac{\ell'(-f(X))}{\ell'(f(X))} = \frac{\mathbb{P}(Y=1|X)}{1-\mathbb{P}(Y=1|X)}\,.$$

Comme ℓ est strictement convexe, on a $f(X) > 0$ si et seulement si $\ell'(-f(X))/\ell'(f(X)) > 1$, donc

$$\begin{aligned}f_*^\ell(X) > 0 &\iff \mathbb{P}(Y=1|X) > 1/2 \\ &\iff \mathbb{E}[Y|X] = 2\mathbb{P}(Y=1|X) - 1 > 0\,.\end{aligned}$$

Comme $h_*(X) = \mathrm{sign}(\mathbb{E}\,[Y|X])$ (voir Partie 2), on obtient $\mathrm{sign}(f_*^\ell) = h_*$. Cette égalité reste vraie pour la perte Hinge (exercice !).

Pour résumer la discussion ci-dessus, la fonction cible f_*^ℓ approchée par $\widehat{f}_{\mathscr{F}}$ est intéressante pour le problème de classification car son signe coïncide avec le meilleur classifieur possible h_*.

Majoration de la probabilité de mauvaise classification. Nous nous concentrons maintenant sur la probabilité de mauvaise classification $L(\widehat{h}_{\mathscr{F}})$ du classifieur $\widehat{h}_{\mathscr{F}} = \mathrm{sign}(\widehat{f}_{\mathscr{F}})$ défini par (15). En pratique, il est intéressant d'avoir une majoration de la probabilité de mauvaise classification $L(\widehat{h}_{\mathscr{F}})$ qui puisse être évaluée à partir des données. Le prochain théorème fournit une telle majoration pour des exemples typiques d'ensembles $\mathscr{F}$.

***Théorème 3* (Borne de confiance pour $L(\widehat{h}_{\mathscr{F}})$).** *Pour tout $R > 0$, on définit $\Delta\ell(R) = |\ell(R) - \ell(-R)|$. On suppose que la fonction de perte ℓ est convexe, décroissante, positive, α-Lipschitz sur $[-R, R]$ et vérifie*

$\ell(z) \geqslant \mathbf{1}_{z<0}$ pour tout z dans $\mathbb{R}$. On s'intéresse au classifieur $\widehat{h}_{\mathscr{F}}$ défini par (15).

(a) *Lorsque $\mathscr{F}$ est de la forme (16) avec $\mathscr{C} = \{\beta \in \mathbb{R}^p : |\beta|_1 \leqslant R\}$, nous avons avec probabilité au moins $1 - e^{-t}$*

$$L(\widehat{h}_{\mathscr{F}}) \;\leqslant\; \widehat{L}_n^{\ell}(\widehat{f}_{\mathscr{F}}) + 2\alpha R \sqrt{\frac{2\log(2p)}{n}} + \Delta\ell(R)\sqrt{\frac{t}{2n}}\,. \tag{20}$$

(b) *Soit $\mathscr{F}$ une boule de rayon R d'un RKHS de noyau k vérifiant $k(x,x) \leqslant 1$ pour tout $x \in \mathscr{X}$. Alors, avec probabilité au moins $1-e^{-t}$, on a*

$$L(\widehat{h}_{\mathscr{F}}) \;\leqslant\; \widehat{L}_n^{\ell}(\widehat{f}_{\mathscr{F}}) + \frac{2\alpha R}{\sqrt{n}} + \Delta\ell(R)\sqrt{\frac{t}{2n}}\,. \tag{21}$$

Démonstration. Commençons par prouver une majoration générale de $L(\widehat{h}_{\mathscr{F}})$ très similaire au Théorème 1.

Lemme 3.1. *Supposons que $\sup_{f\in\mathscr{F}} |f(x)| \leqslant R < +\infty$. Alors pour toute fonction de perte ℓ vérifiant les hypothèses du Théorème 3, on a avec probabilité au moins $1 - e^{-t}$*

$$L(\widehat{h}_{\mathscr{F}}) \leqslant \widehat{L}_n^{\ell}(\widehat{f}_{\mathscr{F}}) + \frac{2\alpha}{n} \max_{x\in\mathscr{X}^n} \mathbb{E}_\sigma\left[\sup_{f\in\mathscr{F}} \left|\sum_{i=1}^n \sigma_i f(x_i)\right|\right] + \Delta\ell(R)\sqrt{\frac{t}{2n}} \tag{22}$$

où $\sigma_1, \dots, \sigma_n$ sont des variables aléatoires i.i.d. de loi $\mathbb{P}_\sigma(\sigma_i = 1) = \mathbb{P}_\sigma(\sigma_i = -1) = 1/2$.

Démonstration du Lemme 3.1. La preuve de ce lemme repose sur exactement les mêmes arguments que la preuve du Théorème 1. La première étape est de remarquer que l'inégalité $\ell(z) \geqslant \mathbf{1}_{z<0}$ pour tout réel z nous donne

$$\begin{aligned} L(\widehat{h}_{\mathscr{F}}) = \mathbb{P}(Y\widehat{f}_{\mathscr{F}}(X) < 0) \leqslant L^{\ell}(f) \quad &\text{avec } L^{\ell}(f) = \mathbb{E}\left[\ell(Yf(X))\right] \\ &\leqslant \widehat{L}_n^{\ell}(f) + \widehat{\Delta}_n^{\ell}(\mathscr{F}) \\ &\text{où } \widehat{\Delta}_n^{\ell}(\mathscr{F}) = \sup_{f\in\mathscr{F}} |\widehat{L}_n^{\ell}(f) - L^{\ell}(f)|. \end{aligned}$$

Comme dans le Lemme 1.2, l'inégalité de concentration de McDiarmid (Théorème 4 de l'Appendice B) assure qu'avec probabilité au moins

$1 - e^{-t}$ on a

$$\widehat{\Delta}_n^\ell(\mathscr{F}) \leqslant \mathbb{E}\left[\widehat{\Delta}_n^\ell(\mathscr{F})\right] + \Delta\ell(R)\sqrt{\frac{t}{2n}}\,.$$

Pour conclure la preuve du lemme, il suffit de montrer que

$$\mathbb{E}\left[\widehat{\Delta}_n^\ell(\mathscr{F})\right] \leqslant \frac{2\alpha}{n} \max_{x\in\mathscr{X}^n} \mathbb{E}_\sigma\left[\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^n \sigma_i f(x_i)\right|\right]. \tag{23}$$

En suivant exactement les mêmes arguments que dans la preuve du Lemme 1.3 (en remplaçant $\mathbf{1}_{Y_i\neq h(X_i)}$ par $\ell(Y_i f(X_i))$) on obtient

$$\mathbb{E}\left[\widehat{\Delta}_n^\ell(\mathscr{F})\right] \leqslant \frac{2}{n} \max_{y\in\{-1,+1\}^n} \max_{x\in\mathscr{X}^n} \mathbb{E}_\sigma\left[\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^n \sigma_i \ell(y_i f(x_i))\right|\right].$$

Finalement, on utilise que ℓ est α-Lipschitz pour conclure : le principe de Contraction (Proposition 5 de l'Appendice B) nous donne

$$\begin{aligned}\mathbb{E}_\sigma\left[\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^n \sigma_i \ell(y_i f(x_i))\right|\right] &\leqslant \alpha\,\mathbb{E}_\sigma\left[\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^n \sigma_i y_i f(x_i)\right|\right] \\ &= \alpha\,\mathbb{E}_\sigma\left[\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^n \sigma_i f(x_i)\right|\right].\end{aligned}$$

En combinant les deux dernières majorations on obtient (23) et le Lemme 3.1 est démontré. □

(a) Prouvons maintenant la borne (20). L'application $\beta \to \sum_{i=1}^n \sigma_i \sum_{j=1}^p \beta_j h_j(x_i)$ est linéaire, donc elle atteint à l'un des sommets de $\mathscr{C}$ son maximum et son minimum sur la boule $\mathscr{C}$. En conséquence, on a

$$\mathbb{E}_\sigma\left[\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^n \sigma_i f(x_i)\right|\right] = R\,\mathbb{E}_\sigma\left[\max_{j=1,\dots,p}\left|\sum_{i=1}^n \sigma_i h_j(x_i)\right|\right].$$

En appliquant l'inégalité (6) avec

$$\mathscr{V} = \{(h_j(x_1),\dots,h_j(x_n)) : j = 1,\dots,p\}\,,$$

dont le cardinal est inférieur à p, on obtient

$$\mathbb{E}_\sigma\left[\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^n \sigma_i f(x_i)\right|\right] \leqslant R\,\sqrt{2n\log(2p)}.$$

La borne (20) découle du Lemme 3.1.

(b) Démontrons maintenant la borne (21). On note $\|.\|_{\mathscr{F}}$ la norme dans le RKHS. La formule de reproduction et l'inégalité de Cauchy-Schwartz nous donne

$$\left|\sum_{i=1}^n \sigma_i f(x_i)\right| = \left|\left\langle f, \sum_{i=1}^n \sigma_i k(x_i,.)\right\rangle_{\mathscr{F}}\right| \leqslant \|f\|_{\mathscr{F}} \left\|\sum_{i=1}^n \sigma_i k(x_i,.)\right\|_{\mathscr{F}}.$$

En appliquant de nouveau l'inégalité de Cauchy-Schwartz, on obtient

$$\begin{aligned}\mathbb{E}_\sigma\left[\sup_{\|f\|_{\mathscr{F}}\leqslant R}\left|\sum_{i=1}^n \sigma_i f(x_i)\right|\right] &\leqslant R\ \mathbb{E}_\sigma\left[\left\|\sum_{i=1}^n \sigma_i k(x_i,.)\right\|_{\mathscr{F}}\right] \\ &\leqslant R\ \sqrt{\mathbb{E}_\sigma\left[\left\|\sum_{i=1}^n \sigma_i k(x_i,.)\right\|_{\mathscr{F}}^2\right]} \\ &\leqslant R\ \sqrt{\sum_{i=1}^n k(x_i,x_i)\mathbb{E}_\sigma[\sigma_i^2]}\ \leqslant\ R\sqrt{n},\end{aligned}$$

où on a utilisé $k(x,x) \leqslant 1$ dans la dernière inégalité et $\mathbb{E}[\sigma_i\sigma_j] = 0$ pour $i \neq j$ dans l'avant-dernière. En combinant de nouveau la propriété de reproduction avec l'inégalité de Cauchy-Schwartz, on obtient

$$|f(x)| = |\langle f, k(x,.)\rangle_{\mathscr{F}}| \leqslant R\sqrt{k(x,x)} \leqslant R.$$

Le Lemme 3.1 donne alors

$$L(\widehat{h}_{\mathscr{F}}) \leqslant \widehat{L}_n^\ell(\widehat{f}_{\mathscr{F}}) + \frac{2\alpha R}{\sqrt{n}} + \Delta\ell(R)\sqrt{\frac{t}{2n}}\ ,$$

ce qui achève la preuve du Théorème 3. □

Il est possible de démontrer des bornes de risque similaire à (3) pour $L(\widehat{h}_{\mathscr{H}})$, nous renvoyons à Bousquet, Boucheron et Lugosi [**1**] pour un panorama de tels résultats. Dans la dernière partie de ces notes, nous allons décrire deux algorithmes de classification très populaires : les Support Vector Machines et AdaBoost.

4.3. Support Vector Machine

Les Support Vector Machines (SVM) correspondent à l'estimateur (17) avec la perte Hinge $\ell(z) = (1-z)_+$. La classification finale s'effectue avec $\widehat{h}_{\mathscr{F}}(x) = \mathrm{sign}(\widehat{f}_{\mathscr{F}}(x))$. Nous allons donner une interprétation géométrique de la solution $\widehat{f}_{\mathscr{F}}$, interprétation qui est à l'origine du nom "Support Vector Machine".

***Proposition 2* (Vecteurs supports).** *La solution de (17) avec* $\ell(z) = (1-z)_+$ *est de la forme* $\widehat{f}_{\mathscr{F}}(x) = \sum_{i=1}^n \widehat{\beta}_i\, k(X_i, x)$ *avec*

$$\begin{cases} \widehat{\beta}_i = 0 & \text{si } Y_i \widehat{f}_{\mathscr{F}}(X_i) > 1 \\ \widehat{\beta}_i = Y_i/(2\lambda n) & \text{si } Y_i \widehat{f}_{\mathscr{F}}(X_i) < 1 \\ 0 \leqslant Y_i \widehat{\beta}_i \leqslant 1/(2\lambda n) & \text{si } Y_i \widehat{f}_{\mathscr{F}}(X_i) = 1\,. \end{cases}$$

Les vecteurs X_i *d'indice* i *tel que* $\widehat{\beta}_i \neq 0$ *sont appelés vecteurs supports.*

Démonstration. Notons K la matrice $[k(X_i, X_j)]_{i,j=1,\dots,n}$. Nous savons par (19) que la solution de (17) est de la forme $\widehat{f}_{\mathscr{F}} = \sum_{j=1}^n \widehat{\beta}_j\, k(X_j, .)$ avec

$$\widehat{\beta} = \operatorname*{argmin}_{\beta \in \mathbb{R}^n} \left\{ \frac{1}{n} \sum_{i=1}^n \left(1 - Y_i [K\beta]_i\right)_+ + \lambda \beta^T K \beta \right\}.$$

Le problème de minimisation ci-dessus n'est pas régulier, donc on introduit des variables de relâchement $\widehat{\xi}_i = (1 - Y_i[K\widehat{\beta}]_i)_+$ et on reformule le problème de minimisation comme suit

$$(\widehat{\beta}, \widehat{\xi}) = \operatorname*{argmin}_{\substack{\beta, \xi \in \mathbb{R}^n \text{ tels que} \\ \xi_i \geqslant 1 - Y_i [K\beta]_i \\ \xi_i \geqslant 0}} \left\{ \frac{1}{n} \sum_{i=1}^n \xi_i + \lambda \beta^T K \beta \right\}. \tag{24}$$

Ce problème est maintenant régulier et convexe et les conditions de Karush-Kuhn-Tucker pour le problème lagrangien dual

$$(\widehat{\beta}, \widehat{\xi}) = \operatorname*{argmin}_{\beta, \xi \in \mathbb{R}^n} \left\{ \frac{1}{n} \sum_{i=1}^n \xi_i + \lambda \beta^T K \beta - \sum_{i=1}^n \left(\alpha_i (\xi_i - 1 + Y_i [K\beta]_i) + \gamma_i \xi_i \right) \right\}.$$

donnent les formules

conditions de 1er ordre :

$$2\lambda [K\widehat{\beta}]_j = \sum_{i=1}^n K_{ij} \alpha_i Y_i \quad \text{et} \quad \alpha_j + \gamma_j = \frac{1}{n},$$

conditions de relâchement :

$$\min(\alpha_i, \widehat{\xi}_i - 1 + Y_i [K\widehat{\beta}]_i) = 0 \quad \text{et} \quad \min(\gamma_i, \widehat{\xi}_i) = 0.$$

On déduit de la première condition de premier ordre que $\widehat{\beta}_i = \alpha_i Y_i/(2\lambda)$. Comme $\widehat{f}_{\mathscr{F}}(X_i) = [K\widehat{\beta}]_i$, la première condition de relâchement donne $\widehat{\beta}_i = 0$ si $Y_i \widehat{f}_{\mathscr{F}}(X_i) > 1$. La seconde condition de relâchement couplée à la seconde condition de premier ordre

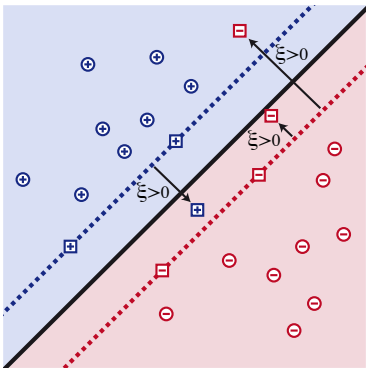

FIGURE 3. Classification par un SVM linéaire : l'hyperplan frontière $\left\{x \in \mathbb{R}^d : \langle \widehat{w}, x\rangle = 0\right\}$ est représenté en noir, les hyperplans de marge $\left\{x \in \mathbb{R}^d : \langle \widehat{w}, x\rangle = +1\right\}$ et $\left\{x \in \mathbb{R}^d : \langle \widehat{w}, x\rangle = -1\right\}$ sont représentés en pointillés bleus et rouges respectivement. Les vecteurs supports sont représentés par des carrés.

implique que $\widehat{\beta}_i = Y_i/(2\lambda n)$ si $\widehat{\xi}_i > 0$ et $0 \leqslant Y_i\widehat{\beta}_i \leqslant 1/(2\lambda n)$ sinon. Pour conclure la preuve de la proposition, on remarque que lorsque que $\widehat{\xi}_i > 0$ on a $\widehat{\beta}_i$ et α_i non nuls, et donc $Y_i\widehat{f}_{\mathscr{F}}(X_i) = 1 - \widehat{\xi}_i < 1$ vu la première condition de relâchement. □

Interprétons géométriquement la Proposition 2.

Interprétation géométrique : noyau linéaire. Considérons le noyau le plus simple $k(x, y) = \langle x, y\rangle$ pour tout $x, y \in \mathbb{R}^d$. Le RKHS associé est l'espace des formes linéaires $\mathscr{F} = \left\{\langle w, .\rangle : w \in \mathbb{R}^d\right\}$. Dans ce cas

$$\widehat{f}_{\mathscr{F}}(x) = \sum_{i=1}^{n} \widehat{\beta}_i \langle X_i, x\rangle = \langle \widehat{w}, x\rangle \quad \text{avec} \quad \widehat{w} = \sum_{i=1}^{n} \widehat{\beta}_i X_i\,,$$

donc le classifieur $\widehat{h}_{\mathscr{F}}(x) = \text{sign}(\langle \widehat{w}, x\rangle)$ classe les points en fonction de leur position par rapport à l'hyperplan $\left\{x \in \mathbb{R}^d : \langle \widehat{w}, x\rangle = 0\right\}$. La normale à l'hyperplan $\widehat{w}$ est une combinaison linéaire des vecteurs supports, qui sont les points X_i tels que $Y_i\langle \widehat{w}, X_i\rangle \leqslant 1$. Ils sont représentés par des carrés sur la Figure 3. Les hyperplans $\left\{x \in \mathbb{R}^d : \langle \widehat{w}, x\rangle = +1\right\}$ et $\left\{x \in \mathbb{R}^d : \langle \widehat{w}, x\rangle = -1\right\}$ sont habituellement appelés hyperplans de marge. Pour finir, notons la propriété suivante des SVM. Si on ajoute à l'ensemble d'apprentissage $(X_i, Y_i)_{i=1,\dots,n}$ un point X_{n+1} vérifiant $Y_{n+1}\langle \widehat{w}, X_{n+1}\rangle > 1$, alors le

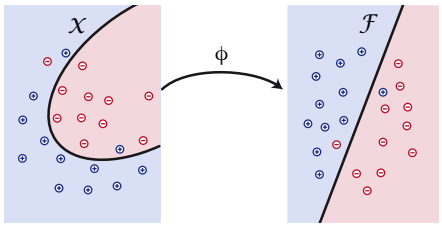

FIGURE 4. Classification avec un noyau non linéaire : la classification linéaire dans $\mathscr{F}$ produit une classification non linéaire dans $\mathscr{X}$ via l'image réciproque de ϕ.

vecteur $\widehat{w}$ et le classifieur $\widehat{h}_{\mathscr{F}}$ ne changent pas. En d'autres mots, seules les données mal classifiées ou classifiées avec une marge insuffisante (i.e. $Y_i\langle \widehat{w}, X_i\rangle \leqslant 1$) influencent l'hyperplan frontière $\left\{x \in \mathbb{R}^d : \langle \widehat{w}, x\rangle = 0\right\}$.

Interprétation géométrique : noyau défini positif quelconque. Notons par $\phi : \mathscr{X} \to \mathscr{F}$ l'application $\phi(x) = k(x, .)$. La propriété de reproduction combinée à la Proposition 2 nous donne

$$\widehat{f}_{\mathscr{F}}(x) = \langle \widehat{f}_{\mathscr{F}}, \phi(x)\rangle_{\mathscr{F}} = \left\langle \sum_{i=1}^{n} \widehat{\beta}_i \phi(X_i), \phi(x) \right\rangle_{\mathscr{F}}.$$

Un point $x \in \mathscr{X}$ est classifié selon le signe du produit scalaire ci-dessus. En conséquence, les points $\phi(x) \in \mathscr{F}$ sont classifiés selon le classifieur linéaire sur $\mathscr{F}$

$$f \mapsto \operatorname{sign}\left(\langle \widehat{w}_\phi, f\rangle_{\mathscr{F}}\right) \quad \text{où } \widehat{w}_\phi = \sum_{i=1}^{n} \widehat{\beta}_i \phi(X_i).$$

La frontière $\{x \in \mathscr{X} : \widehat{f}_{\mathscr{F}}(x) = 0\}$ du classifieur $\widehat{h}_{\mathscr{F}}$ est donc l'image réciproque par ϕ de l'hyperplan $\{f \in \mathscr{F} : \langle \widehat{w}_\phi, f\rangle_{\mathscr{F}} = 0\}$, comme représenté Figure 4. Nous observons que le noyau k délinéarise le SVM, dans le sens où il produit un classifieur non linéaire $\widehat{h}_{\mathscr{F}}$ pour le même coût numérique qu'un classifieur linéaire dans $\mathbb{R}^n$.

Vous pouvez observer les SVM à l'œuvre avec la petite démo :
`http://svm.dcs.rhbnc.ac.uk/pagesnew/GPat.shtml`

Pourquoi les RKHS sont-ils utiles ? Il y a principalement deux raisons pour utiliser un RKHS. La première raison est de permettre de délinéariser un algorithme en envoyant $\mathscr{X}$ dans $\mathscr{F}$ avec $\phi : x \to k(x, .)$,

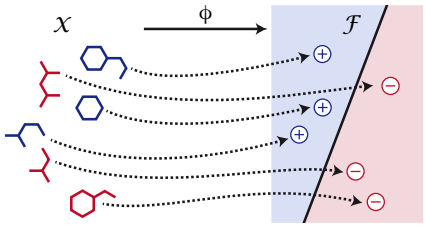

FIGURE 5. Classification de molécules à l'aide d'un SVM.

comme représenté dans la Figure 4. Cela permet d'obtenir un algorithme non linéaire pour un coût similaire à celui d'un algorithme linéaire.

La seconde raison pour utiliser un RKHS est de pouvoir employer sur n'importe quel ensemble $\mathscr{X}$ des algorithmes définis pour des vecteurs. Imaginons par exemple qu'on veuille classifier des protéines ou des molécules en fonction de leur propriétés thérapeutiques. Notons par $\mathscr{X}$ notre ensemble de molécules. Pour tout $x, y \in \mathscr{X}$, représentons par $k(x, y)$ leur similarité (selon des critères à définir). Si le noyau résultant $k : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ est défini positif, on peut alors directement appliquer un SVM pour les classifier, comme illustré Figure 5. Bien sûr, le point clef dans ce cas est de construire le noyau k. En général, le noyau $k(x, y)$ est défini en fonction de certaines propriétés de x, y qui sont connues pour avoir de l'importance pour le problème de classification. Par exemple, le nombre de courtes séquences communes est un indice utile pour quantifier la similarité entre deux protéines. La complexité du calcul de $k(x, y)$ est aussi un critère crucial pour les applications sur des données complexes. Nous renvoyons aux excellents slides de Jean-Philippe Vert pour la description d'applications prometteuses en biologie et médecine :
`http://cbio.ensmp.fr/~jvert/talks/120302ensae/ensae`

4.4. AdaBoost

AdaBoost est un algorithme qui tend à calculer une solution approximative de l'estimateur (15) avec la perte exponentielle $\ell(z) = e^{-z}$ et l'espace fonctionnel $\mathscr{F} = \operatorname{span}\{h_1, \dots, h_p\}$ où $h_1, \dots, h_p$ sont p classifieurs donnés.

Le principe de l'algorithme AdaBoost est de réaliser une minimisation dite agressive de (15). Plus précisément, AdaBoost produit une suite de fonctions $\widehat{f}_m$ pour $m = 0, \dots, M$ en partant de $\widehat{f}_0 = 0$ puis en résolvant récursivement pour $m = 1, \dots, M$

$$\widehat{f}_m = \widehat{f}_{m-1} + \beta_m h_{j_m}$$
$$\text{où} \quad (\beta_m, j_m) = \operatorname*{argmin}_{\substack{j=1,\dots,p \\ \beta\in\mathbb{R}}} \frac{1}{n}\sum_{i=1}^{n} \exp\left(-Y_i\big(\widehat{f}_{m-1}(X_i) + \beta h_j(X_i)\big)\right).$$

La classification finale est effectuée à l'aide de $\widehat{h}_M(x) = \operatorname{sign}(\widehat{f}_M(x))$ qui est une approximation de $\widehat{h}_{\mathscr{H}}$ défini par (15).

La perte exponentielle permet de calculer (β_m, j_m) très efficacement. En effet, en posant $w_i^{(m)} = n^{-1}\exp(-Y_i\widehat{f}_{m-1}(X_i))$, on peut écrire

$$\frac{1}{n}\sum_{i=1}^{n} \exp\left(-Y_i\big(\widehat{f}_{m-1}(X_i) + \beta h_j(X_i)\big)\right)$$
$$= (e^{\beta} - e^{-\beta})\sum_{i=1}^{n} w_i^{(m)} \mathbf{1}_{h_j(X_i)\neq Y_i} + e^{-\beta}\sum_{i=1}^{n} w_i^{(m)}.$$

Lorsque la condition suivante

$$\operatorname{err}_m(j) = \frac{\sum_{i=1}^n w_i^{(m)} \mathbf{1}_{h_j(X_i)\neq Y_i}}{\sum_{i=1}^n w_i^{(m)}} \leqslant \frac{1}{2} \quad \text{pour tout } j = 1, \dots, p,$$

est satisfaite, les minimiseurs (β_m, j_m) sont donnés par

$$j_m = \operatorname*{argmin}_{j=1,\dots,p} \operatorname{err}_m(j) \quad \text{et} \quad \beta_m = \frac{1}{2}\log\left(\frac{1 - \operatorname{err}_m(j_m)}{\operatorname{err}_m(j_m)}\right).$$

En remarquant que $-Y_i h(X_i) = 2\mathbf{1}_{Y_i\neq h(X_i)} - 1$ on obtient la formulation standard de l'algorithme AdaBoost.

AdaBoost

Init : $w_i^{(1)} = 1/n$, for $i = 1, \dots, n$

Iterate : For $m = 1, \dots, M$ do

$$j_m = \operatorname*{argmin}_{j=1,\dots,p} \mathrm{err}_m(j)$$

$$2\beta_m = \log(1 - \mathrm{err}_m(j_m)) - \log(\mathrm{err}_m(j_m))$$

$$w_i^{(m+1)} = w_i^{(m)} \exp(2\beta_m \mathbf{1}_{h_{j_m}(X_i) \neq Y_i}), \quad \text{for } i = 1, \dots, n$$

STOP if $\min_{j=1,\dots,p} \mathrm{err}_{m+1}(j) > 1/2$

Output : $\widehat{f}_M(x) = \sum_{m=1}^{M} \beta_m h_{j_m}(x)$.

On remarque que l'algorithme AdaBoost donne de plus en plus de poids dans $\mathrm{err}_m(j)$ aux points X_i qui sont mal classifiés à l'étape m.

Vous pouvez observer AdaBoost à l'œuvre (pour des classifieurs h_j par demi-espaces) avec l'application récréative :
`http://cseweb.ucsd.edu/~yfreund/adaboost/`

5. Pour aller au-delà de cette introduction

Nous renvoyons le lecteur intéressé à aller au-delà des concepts présentés dans ces notes à l'article de synthèse de Boucheron, Bousquet et Lugosi [**1**]. Cet article comprend notamment une bibliographie assez exhaustive. D'un point de vue plus pratique et appliqué, l'incontournable livre de Hastie, Tibshirani et Friedman [**5**] décrit et discute de très nombreux algorithmes opérationnels. Enfin, nous soulignons que les concepts introduits ici apparaissent aussi pour le problème de *ranking* (ordonner au mieux des données, l'exemple le plus célèbre étant les moteurs de recherche sur internet), voir Clémençon, Lugosi et Vayatis [**3**].

Appendice A. Espace de Hilbert à noyaux reproduisants (RKHS)

Les espaces de Hilbert à noyaux reproduisants, dit *Reproducing Kernel Hilbert Spaces* (RKHS) en anglais, sont des espaces de Hilbert fonctionnels dont la norme décrit la régularité des fonctions. Les RKHS vérifient aussi une propriété de reproduction particulière, qui

s'avère cruciale en pratique car elle permet une mise œuvre numérique efficace.

Une fonction $k : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ est dite définie positive si elle est symétrique ($k(x,y) = k(y,x)$ pour tout $x, y \in \mathscr{X}$) et si pour tout $N \in \mathbb{N}$, $x_1, \ldots, x_N \in \mathscr{X}$ et $a_1, \ldots, a_N \in \mathbb{R}$ on a

$$\sum_{i,j=1}^{N} a_i a_j k(x_i, x_j) \geqslant 0. \tag{25}$$

Exemples de noyaux définis positifs :

- noyau linéaire : $k(x,y) = \langle x, y \rangle$
- noyau gaussien : $k(x,y) = e^{-\|x-y\|^2/2\sigma^2}$
- noyau histogramme ($d = 1$) : $k(x,y) = \min(x,y)$
- noyau exponentiel : $k(x,y) = e^{-\|x-y\|/\sigma}$

On peut associer à un noyau défini positif k un sous-espace de Hilbert $\mathscr{F}$ de $\mathbb{R}^{\mathscr{X}}$ appelé espace de Hilbert de noyau reproduisant k.

***Proposition 3* (Espace de Hilbert à noyau reproduisant (RKHS))**
À tout noyau défini positif k sur $\mathscr{X}$, on peut associer un (unique) espace de Hilbert $\mathscr{F} \subset \mathbb{R}^{\mathscr{X}}$ vérifiant

- *$k(x,.) \in \mathscr{F}$ pour tout $x \in \mathscr{X}$*
- propriété de reproduction : *$f(x) = \langle f, k(x,.) \rangle_{\mathscr{F}}$ pour tout $x \in \mathscr{X}$ et $f \in \mathscr{F}$.*

L'espace $\mathscr{F}$ est appelé espace de Hilbert de noyau reproduisant k.

Démonstration. Considérons l'espace linéaire $\mathscr{F}_0$ engendré par la famille $\{k(x,.) : x \in \mathscr{X}\}$

$$\mathscr{F}_0 = \Big\{ f : \mathscr{X} \longrightarrow \mathbb{R} : f(x) = \sum_{i=1}^{N} a_i k(x_i, x),\ N \in \mathbb{N}, \\ x_1, \ldots, x_N \in \mathscr{X},\ a_1, \ldots, a_N \in \mathbb{R} \Big\}.$$

À tout $f = \sum_{i=1}^{N} a_i k(x_i,.)$ et $g = \sum_{j=1}^{M} b_j k(y_j,.)$ on associe

$$\langle f, g \rangle_{\mathscr{F}_0} := \sum_{i=1}^{N} \sum_{j=1}^{M} a_i b_j k(x_i, y_j) = \sum_{i=1}^{N} a_i g(x_i) = \sum_{j=1}^{M} b_j f(y_j).$$

Les deux dernières égalités indiquent que $\langle f, g\rangle_{\mathscr{F}_0}$ ne dépend pas du choix de la décomposition de f et g, donc cette quantité est bien définie. De plus, l'application $(f, g) \to \langle f, g\rangle_{\mathscr{F}_0}$ est bilinéaire, symétrique, positive (vu (25)) et on a la propriété de reproduction

$$f(x) = \langle f, k(x, .)\rangle_{\mathscr{F}_0} \quad \text{pour tout } x \in \mathscr{X} \quad \text{et } f \in \mathscr{F}_0.$$

L'inégalité de Cauchy-Schwartz $\langle f, g\rangle_{\mathscr{F}_0} \leqslant \|f\|_{\mathscr{F}_0}\|g\|_{\mathscr{F}_0}$ combinée à la formule de reproduction donne

$$|f(x)| \leqslant \sqrt{k(x, x)}\ \|f\|_{\mathscr{F}_0}. \tag{26}$$

En conséquence $\|f\|_{\mathscr{F}_0} = 0$ implique $f = 0$ donc $\langle f, g\rangle_{\mathscr{F}_0}$ est un produit scalaire. L'espace $\mathscr{F}_0$ est donc un espace pré-hilbertien et pour obtenir $\mathscr{F}$ il nous suffit de le compléter.

Considérons deux suites $(x_i) \in \mathscr{X}^{\mathbb{N}}$ et $(a_i) \in \mathbb{R}^{\mathbb{N}}$ vérifiant $\sum_{i,j\geqslant 1} a_i a_j k(x_i, x_j) < +\infty$. L'inégalité (26) nous assure que pour tout $M < N$ et $x \in \mathscr{X}$ on a

$$\Big|\sum_{i=M+1}^{N} a_i k(x_i, x)\Big| \leqslant \sqrt{k(x, x)}\ \sum_{i,j=M+1}^{N} a_i a_j k(x_i, x_j).$$

Lorsque $\sum_{i,j\geqslant 1} a_i a_j k(x_i, x_j)$ est fini, le membre de droite tend vers 0 lorsque M, N tendent vers l'infini, donc les séries partielles $\sum_{i=1}^{N} a_i k(x_i, x)$ sont de Cauchy et convergent lorsque $N \to \infty$. On peut donc définir l'espace

$$\mathscr{F} = \Big\{ f : \mathscr{X} \longrightarrow \mathbb{R} : f(x) = \sum_{i=1}^{\infty} a_i k(x_i, x),\ (x_i) \in \mathscr{X}^{\mathbb{N}},\ (a_i) \in \mathbb{R}^{\mathbb{N}},\ \sum_{i,j\geqslant 1} a_i a_j k(x_i, x_j) < +\infty \Big\}$$

et la forme bilinéaire

$$\langle f, g\rangle_{\mathscr{F}} := \sum_{i,j=1}^{\infty} a_i b_j k(x_i, y_j) = \sum_{i=1}^{\infty} a_i g(x_i) = \sum_{j=1}^{\infty} b_j f(y_j)$$

pour $f = \sum_{i=1}^{\infty} a_i k(x_i, .)$ et $g = \sum_{j=1}^{\infty} b_j k(y_j, .)$. Exactement comme précédemment, l'application $(f, g) \to \langle f, g\rangle_{\mathscr{F}}$ est un produit scalaire vérifiant la propriété de reproduction

$$f(x) = \langle f, k(x, .)\rangle_{\mathscr{F}} \quad \text{pour tout } x \in \mathscr{X} \text{ et } f \in \mathscr{F}.$$

Finalement, l'espace $\mathscr{F}$ muni du produit scalaire $\langle .,. \rangle_{\mathscr{F}}$ est la complétion dans $\mathbb{R}^{\mathscr{X}}$ de $\mathscr{F}_0$ muni de $\langle .,. \rangle_{\mathscr{F}_0}$, donc c'est un espace de Hilbert à noyau reproduisant. □

La norme d'une fonction f dans un RKHS $\mathscr{F}$ est fortement reliée à la régularité de f. Cette propriété apparait clairement dans l'inégalité

$$|f(x) - f(x')| = |\langle f, k(x,.) - k(x',.)\rangle_{\mathscr{F}}| \leqslant \|f\|_{\mathscr{F}} \, \|k(x,.) - k(x',.)\|_{\mathscr{F}}.$$

Nous illustrons ce point en décrivant le RKHS associé aux noyaux histogrammes et gaussiens.

Exemple 1 : RKHS associé au noyau histogramme

L'espace de Sobolev

$$\mathscr{F} = \Big\{ f \in C([0,1],\mathbb{R}) : f \text{ est p.p. différentiable} \\ \text{avec } f' \in L^2([0,1]) \text{ et } f(0) = 0 \Big\}$$

muni du produit scalaire $\langle f, g\rangle_{\mathscr{F}} = \int_0^1 f'g'$ est un RKHS de noyau reproduisant $k(x,y) = \min(x,y)$ sur $[0,1]$. En effet, $k(x,.) \in \mathscr{F}$ pour tout $x \in [0,1]$ et

$$f(x) = \int_0^1 f'(y)\mathbf{1}_{y \leqslant x}\, dy = \langle f, k(x,.)\rangle_{\mathscr{F}}\,, \text{ pour tout } f \in \mathscr{F} \text{ et } x \in [0,1].$$

Dans ce cas, la norme $\|f\|_{\mathscr{F}}$ correspond simplement à la norme L^2 de la dérivée de f. Plus cette norme est petite, plus f est régulière.

Exemple 2 : RKHS associé au noyau gaussien. Notons par $\mathbf{F}[f]$ la transformée de Fourier dans $\mathbb{R}^d$ de normalisation

$$\mathbf{F}[f](\omega) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} f(t) e^{-\mathbf{i}\langle \omega, t\rangle}, \\ \text{pour } f \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d) \text{ et } \omega \in \mathbb{R}^d.$$

Pour tout $\sigma > 0$, l'espace fonctionnel

$$\mathscr{F}_\sigma = \Big\{ f \in C_0(\mathbb{R}^d) \cap L^1(\mathbb{R}^d) \text{ tel que } \int_{\mathbb{R}^d} \big|\mathbf{F}[f](\omega)\big|^2 e^{\sigma|\omega|^2/2}\, d\omega < +\infty \Big\},$$

muni du produit scalaire

$$\langle f, g\rangle_{\mathscr{F}_\sigma} = (2\pi\sigma^2)^{-d/2} \int_{\mathbb{R}^d} \overline{\mathbf{F}[f](\omega)}\mathbf{F}[g](\omega) e^{\sigma|\omega|^2/2}\, d\omega$$

est un RKHS associé au noyau gaussien

$$k(x,y) = \exp(-\|y-x\|^2/2\sigma^2).$$

En effet, pour tout $x \in \mathbb{R}^d$ la fonction $k(x,.)$ appartient à $\mathscr{F}_\sigma$ et un calcul direct donne

$$\langle k(x,.), f\rangle_{\mathscr{F}_\sigma} = \mathbf{F}^{-1}[\mathbf{F}[f]](x) = f(x) \quad \text{pour tout } f \in \mathscr{F} \text{ et } x \in \mathbb{R}^d.$$

L'espace $\mathscr{F}_\sigma$ est composé de fonctions très régulières et la norme $\|f\|_{\mathscr{F}_\sigma}$ contrôle directement la régularité de f. Remarquons que lorsque σ augmente l'espace $\mathscr{F}_\sigma$ rétrécit pour ne contenir que des fonctions de plus en plus régulières.

Appendice B. Inégalités probabilistes

Les analyses non asymptotiques en apprentissage statistique reposent très souvent sur l'inégalité de concentration de McDiarmid [**6**].

Théorème 4 **(McDiarmid (1989)).** *Soit $\mathscr{X}$ un ensemble mesurable et considérons $F : \mathscr{X}^n \to \mathbb{R}$ telle qu'il existe $\delta_1, \dots, \delta_n$ vérifiant*

$$\big|F(x_1, \dots, x_i', \dots, x_n) - F(x_1, \dots, x_i, \dots, x_n)\big| \leqslant \delta_i,$$

pour tout $x_1, \dots, x_n, x_i' \in \mathscr{X}$, pour $i = 1, \dots, n$. Alors, pour tout $t > 0$ et pour toutes variables indépendantes $X_1, \dots, X_n$, on a

$$\mathbb{P}\big(F(X_1, \dots, X_n) > \mathbb{E}\left[F(X_1, \dots, X_n)\right] + t\big) \leqslant \exp\left(-\frac{2t^2}{\delta_1^2 + \dots + \delta_n^2}\right).$$

Nous renvoyons le lecteur au Chapitre 9 du livre de Devroye, Györfi et Lugosi [**4**] pour la preuve classique de ce résultat, preuve qui combine simplement l'inégalité de Markov avec un argument martingale (par conditionnements télescopiques). Une preuve plus conceptuelle basée sur la méthode d'entropie est donnée dans le livre de Boucheron, Lugosi et Massart [**2**], Chapitre 6.

Un second concept important en apprentissage statistique est le principe de contraction.

Théorème 5 **(Principe de contraction).** *Soit $\mathscr{Z}$ un sous-ensemble borné de $\mathbb{R}^n$ et $\varphi : \mathbb{R} \to \mathbb{R}$ une fonction α-Lipschitz vérifiant $\varphi(0) = 0$. Pour des variables aléatoires $\sigma_1, \dots, \sigma_n$ i.i.d. de loi*

$$\mathbb{P}_\sigma(\sigma_i = 1) = \mathbb{P}_\sigma(\sigma_i = -1) = 1/2,$$

nous avons

$$\mathbb{E}_\sigma\left[\sup_{z\in\mathscr{Z}}\left|\sum_{i=1}^n \sigma_i\varphi(z_i)\right|\right] \leqslant \alpha\,\mathbb{E}_\sigma\left[\sup_{z\in\mathscr{Z}}\left|\sum_{i=1}^n \sigma_i z_i\right|\right].$$

Un preuve concise est donnée dans le Chapitre 11 du livre de Boucheron, Lugosi et Massart [**2**].

Références

[1] Boucheron, S., Bousquet, O. and Lugosi, G. *Theory of classification : some recent advances.* ESAIM Probability & Statistics, **9** (2005) : 323–375.

[2] Boucheron, S., Lugosi, G. and Massart, P. *Concentration Inequalities.* Oxford University Press, 2013.

[3] Clémençon, S., Lugosi, G. and Vayatis, N. *Ranking and empirical risk minimization of U-statistics.* The Annals of Statistics, **36** (2008) : 844–874.

[4] Devroye, L., Györfi, L. and Lugosi, G. *A Probabilistic Theory of Pattern Recognition.* Springer, 1996.

[5] Hastie, T., Tibshirani, R. and Friedman, J. *The element of statistical learning.* Springer, 2009.

[6] McDiarmid, C. *On the Method of Bounded Differences.* Surveys in Combinatorics **141** (1989) : 148–188.

Christophe Giraud, Département de Mathématiques, Bât. 425, Faculté des Sciences d'Orsay, Université Paris-Sud, F-91405 Orsay Cedex
CMAP, UMR CNRS 7641, Ecole Polytechnique, Route de Saclay, 91128 Palaiseau Cedex • *E-mail :* christophe.giraud@math.u-psud.fr
Url : http://www.cmap.polytechnique.fr/~giraud/