Introduction à la classification

Sylvain Arlot
sous la direction de Pascal Massart

13 Octobre 2004

La classification binaire s'inscrit dans le cadre de la théorie de l'apprentissage statistique, fondée notamment par Vapnik [Vap82, Vap98]. Elle a pour objectif de déduire d'un nombre fini d'observations indépendantes une «classification» de l'espace en deux domaines. Une bonne méthode de classification doit être capable de s'adapter aux propriétés du domaine qu'elle cherche à déterminer (en particulier sa complexité et le niveau de bruit «près du bord»), sans toutefois les connaître *a priori*. Après avoir défini un cadre théorique rigoureux, nous verrons comment réaliser une telle estimation et quelle précision on peut en attendre. Nous esquisserons enfin les moyens susceptibles de rendre cette estimation adaptative. Cet objectif n'est pas atteint à l'heure actuelle, mais les nombreuses pistes inexplorées nous donnent bon espoir d'y arriver, ce qui aurait des répercussions dans de nombreux domaines, allant de la détection de gènes à la reconnaissance de formes.

1 Problème de la classification

On observe n réalisations indépendantes $(X_i,Y_i)_{1\leq i\leq n}$ d'un même couple de variables aléatoires (X,Y), de loi inconnue P. On a $X \in \mathcal{X}$ un espace mesurable et $Y \in \mathcal{Y} = \{0,1\}$ est un label attaché à X. On souhaite pouvoir déterminer le label $y \in \mathcal{Y}$ associé à un point quelconque x de $\mathcal{X}$.

Définition 1.1 (Classifieur). On appelle *classifieur* toute application mesurable $t : \mathcal{X} \mapsto \mathcal{Y}$, et on note $\mathbb{S}$ l'ensemble des classifieurs.

Pour évaluer l'erreur commise par un classifieur, on doit tout d'abord définir une *fonction de perte* $\gamma : \mathbb{S} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$. Un choix usuel pour γ est

$$\gamma(t,(x,y)) = \mathbb{1}_{t(x)\neq y}.$$

Le *risque* du classifieur est alors

$$P(\gamma(t,\cdot)) = P(Y \neq t(X)).$$

Le problème de la classification (binaire) est celui de la minimisation de ce risque.

Connaissant P, le classifieur optimal est le *classifieur de Bayes* $s^\star$, défini par

$$s^\star(x) = \mathbb{1}_{\eta(x)\geq\frac{1}{2}} \text{ avec } \eta(x) = P(Y = 1|X = x).$$

Démonstration. Soit t un autre classifieur. On a alors

$$\begin{aligned} P(Y \neq t(X)|X) &= (1-\eta(X))\mathbb{1}_{t(X)=1} + \eta(X)\mathbb{1}_{t(X)=0} \\ &\geq (1-\eta(X)) \wedge \eta(X) \\ &= P(Y \neq s^\star(X)|X). \end{aligned}$$

À la seconde ligne, on a noté $a \wedge b$ pour $\min(a,b)$. On notera également $a \vee b$ pour $\max(a,b)$. Le résultat s'en déduit en passant aux espérances. □

Notons que la définition de $s^\star$ sur l'événement $\{\eta(X) = 1/2\}$ ne change rien pour son risque. La *perte relative*

$$l(s^\star,t) = P[Y \neq t(X)] - P[Y \neq s^\star(X)]$$

est positive et mesure l'erreur commise par le classifieur t. On souhaite construire un classifieur $\hat{s}$ estimant $s^\star$ avec pour seules données les $(X_i,Y_i)_{1\leq i\leq n}$. La qualité d'un estimateur est mesurée par $\mathbb{E}_P\left[l(s^\star,\hat{s})\right]$.

2 Minimisation du risque empirique

2.1 Définition

Une méthode naturelle est de substituer dans le problème de minimisation du risque $P\left[\gamma(t,\cdot)\right]$ la loi P par la *loi empirique*

$$P_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{(X_i,Y_i)}.$$

Cela revient à minimiser le risque empirique $\gamma_n(t)$, qui est égal à l'erreur commise par t sur l'échantillon des $(X_i,Y_i)_{1\leq i\leq n}$:

$$\gamma_n(t) = \frac{1}{n}\sum_{i=1}^{n} \gamma(t,(X_i,Y_i)).$$

Définition 2.1 (Modèle). Un modèle S est un ensemble de classifieurs. Dans le cadre de la classification, on écrit aussi

$$S = \{\mathbb{1}_A / A \in \mathcal{A}\}$$

où $\mathcal{A}$ est une famille de parties de $\mathcal{X}$.

Définition 2.2 (ERM). Étant donné un *modèle* S, l'*estimateur par minimisation du risque empirique* (ERM) sur S est

$$\hat{s} \in \arg\min_{t\in S} \gamma_n(t).$$

Notons que lorsque cette borne inf n'est pas atteinte, ou simplement quand on ne peut pas la calculer en pratique, on se contente d'un minimiseur approché du risque empirique.

L'ERM présente l'avantage d'être un cas particulier d'estimation par *minimum de contraste*. On pourra lui appliquer les résultats déjà obtenus dans ce cadre.

2.2 Dimension de Vapnik

On peut estimer le risque de l'ERM $\mathbb{E}_P[l(s^\star,\hat{s})]$ en utilisant la mesure suivante de la complexité du modèle S.

Définition 2.3 (Classe de Vapnik-Chervonenkis). Si $\mathcal{A}$ est une famille de parties mesurables de $\mathcal{X}$, on note $m_{\mathcal{A}}(N) = \sup_{C\in\mathfrak{P}(X),\,\mathrm{Card}\,C\leq N} \mathrm{Card}\,\{A\cap C / A\in\mathcal{A}\}$. On dit que $\mathcal{A}$ est une *classe de Vapnik-Chervonenkis* lorsque

$$V = \sup\{N / m_{\mathcal{A}}(N) = 2^N\} < \infty$$

et V est appelée la *dimension de Vapnik* de $\mathcal{A}$.

Exemple 2.1. L'ensemble $\mathcal{A}_c$ des parties convexes de $\mathbb{R}^2$ n'est pas une classe de Vapnik car $\forall k,\ m_{\mathcal{A}_c}(k) = 2^k$.

Exemple 2.2. L'ensemble des parties à moins de V éléments de $\mathcal{X}$ est une classe de Vapnik de dimension V.

Exemple 2.3. Si $\mathcal{A} = \{]-\infty;a]/a \in \mathbb{R}\}$ et $\mathcal{X} = \mathbb{R}$, alors $m_{\mathcal{A}}(2) = 3 < 2^2$ donc c'est une classe de Vapnik de dimension 1.

Exemple 2.4. L'ensemble des demi-espaces affines de $\mathbb{R}^d$ est une classe de Vapnik de dimension $d+1$.

La dimension de Vapnik de $\mathcal{A}$ décrit la capacité de S à distinguer une partie d'un ensemble fini de points de $\mathcal{X}$. En particulier, si $V \geq n$, le risque empirique de l'ERM est toujours nul lorsque $\{X_1,\ldots,X_n\}$ réalise le sup dans $m_{\mathcal{A}}(n)$, quelle que soit la loi de Y sachant X. Cette situation est à éviter pour une bonne estimation de $s^\star$, ne serait-ce que parce qu'il peut alors y avoir de nombreux minimiseurs possibles, et nous n'avons pas d'informations pour choisir entre eux.

2.3 Risque, risque minimax

Borne du risque Soit $\mathcal{P}(S)$ la famille des lois P telles que $s^\star \in S$. Lugosi [Lug02] a montré qu'il existe une constante $k_1 > 0$ telle que

$$\sup_{P\in\mathcal{P}(S)} \mathbb{E}_P\left[l(s^\star,\hat{s})\right] \leq k_1\sqrt{\frac{V}{n}}. \tag{2.1}$$

Notons que si $s^\star \notin S$, on a la même majoration avec le terme supplémentaire $l(s^\star,S) = \inf_{t\in S} l(s^\star,t)$. Il suffit en effet de remplacer $s^\star$ par $\pi(s^\star) \in S$ qui réalise l'inf et de poser $l(\pi(s^\star),t) = l(s^\star,t) - l(s^\star,\pi(s^\star))$.

Risque minimax Le *risque minimax* donne un moyen de définir l'optimalité d'un estimateur. Si $\mathcal{P}$ est une famille de lois, on pose

$$\mathcal{R}_{\min\max}(\mathcal{P}) = \inf_{\widetilde{s}} \max_{P\in\mathcal{P}} \mathbb{E}_P\left[l(s^\star,\widetilde{s})\right]$$

où l'inf est pris sur tous les estimateurs $\widetilde{s}$.

Des arguments combinatoires permettent une estimation du risque minimax dans le cas où S est une classe de Vapnik et $P = \mathcal{P}(S)$. Devroye et Lugosi [DL95] ont ainsi montré qu'il existe un constante $k_2 > 0$ telle que

$$\mathcal{R}_{\min\max}(\mathcal{P}(S)) \geq k_2\sqrt{\frac{V}{n}} \tag{2.2}$$

À une constante près, l'ERM est donc optimal au sens minimax pour $\mathcal{P} = \mathcal{P}(S)$.

On pourrait penser qu'ici s'achève l'étude du risque de l'ERM, mais il n'en est rien. Considérer la famille $\mathcal{P}(S)$ est en effet très pessimiste. Ainsi, dans la situation (très optimiste) où $Y = \eta(X)$, on montre que le risque est en $\frac{V}{n}$. Nous verrons plus loin que des hypothèses plus raisonnables sur P permettent d'obtenir des bornes intermédiaires sur le risque de l'ERM.

3 Rôle de la concentration

C'est la théorie de la concentration de la mesure qui nous fournit les outils nécessaires à l'élaboration de telles bornes. Avant d'énoncer un théorème général de Massart et Nédélec [MN03], introduisons les inégalités de concentration qui en sont à la base. On trouvera une bonne introduction à la concentration dans un cadre statistique dans [Mas03].

3.1 Inégalité de concentration

L'objet de la théorie de la concentration de la mesure est de quantifier comment une fonction Z de n variables aléatoires indépendantes (de même loi ou non) se concentre autour de son espérance. Il s'agit d'estimer l'ordre de grandeur de $|Z - \mathbb{E}(Z)|$, avec $Z = \zeta(X_1, \ldots, X_n)$. Le théorème central limite en est un exemple rudimentaire, Z étant la moyenne des X_i, indépendantes et de même loi. En classification, on aimerait pouvoir considérer (par exemple) des variables de la forme

$$Z = \sup_{f \in \mathcal{F}} \sum_{i=1}^{n} f(X_i).$$

Idéalement, on voudrait atteindre des résultats du type

$$P\left[Z - \mathbb{E}(Z) \geq \sqrt{x \operatorname{var} Z}\right] \leq e^{-\eta x}$$

pour tout $x > 0$, avec une constante $\eta > 0$. En-dehors du cas gaussien, on se contente souvent de

$$P\left[Z - \mathbb{E}(Z) \geq \sqrt{xv}\right] \leq e^{-\eta x}$$

pour tout $x_0 \geq x > 0$, avec $v \geq \operatorname{var} Z$ et une constante $\eta > 0$.

L'inégalité de Talagrand traite le cas où Z est le supremum de processus empiriques. Bousquet [Bou02] en a montré une version où les constantes sont optimales.

Proposition 3.1 (Inégalité de Talagrand, Bousquet). *Soient $\xi_1, \ldots, \xi_n$ des variables aléatoires i.i.d., $(f_t)_{t \in T}$ une famille dénombrable de fonctions mesurables et telles que*

$$\forall t \in T, \|f_t - \mathbb{E}\left[f_t(\xi_1)\right]\|_\infty \leq b.$$

On pose $v = \sup_{t \in T} \operatorname{var}\left(f_t(\xi_1)\right)$ et

$$Z = \sup_{t \in T} \frac{1}{n} \sum_{i=1}^{n} \left(f_t(\xi_i) - \mathbb{E}\left[f_t(\xi_i)\right]\right).$$

Alors, pour tout $x > 0$,

$$\mathbb{P}\left[Z - \mathbb{E}(Z) \geq \sqrt{2x\left(\frac{v}{n} + 2b\mathbb{E}(Z)\right)} + \frac{bx}{3}\right] \leq e^{-x}.$$

3.2 Inégalité maximale

Les inégalités maximales visent à estimer l'espérance du supremum d'un processus. Combinées avec des inégalités de concentration, elles permettent donc de contrôler ce type de variable aléatoire sur un événement de grande probabilité. Nous utiliserons de cette manière l'inégalité maximale suivante.

Proposition 3.2 (Inégalité maximale pour un processus pondéré). *Soit S dénombrable, $u \in S$, $\omega : S \to \mathbb{R}^+$ tel que $\omega(u) = \inf_{t\in S} \omega(t)$. Soit Z un processus indexé par S tel que $\forall \epsilon > 0$, $\mathbb{E}\left[\sup_{t\in B_u(\epsilon)}(Z(u) - Z(t))\right] < \infty$ avec $B_u(\epsilon) = \{t \in S/\omega(t) \leq \epsilon\}$. Soit $\psi : \mathbb{R}^+ \to \mathbb{R}^+$ telle que $x \mapsto \frac{\psi(x)}{x}$ est décroissante sur $\mathbb{R}^+$ et $\exists \epsilon_\star > 0$, $\forall \epsilon \geq \epsilon_\star$, $\mathbb{E}\left[\sup_{t\in B_u(\epsilon)}(Z(u) - Z(t))\right] \leq \psi(\epsilon)$.*

Alors, $\forall x \geq \epsilon_\star$,

$$\mathbb{E}\left[\sup_{t\in S} \frac{Z(u) - Z(t)}{\omega(t)^2 + x^2}\right] \leq \frac{4\psi(x)}{x^2}.$$

3.3 Théorème général

Afin d'établir des bornes plus fines sur le risque de l'ERM dans de nombreux cas (dont la classification), Massart et Nédélec [MN03] ont établi un théorème général. Nous en présentons ici une forme un peu simplifiée, adaptée au cadre de la classification. Ce théorème reste toutefois très abstrait, aussi la section suivante consacrée à ses applications permet-elle de mieux en saisir la portée.

Les notations sont celles qui ont été introduites précédemment. On note

$$\overline{\gamma}_n(t) = \gamma_n(t) - P[\gamma(t,\cdot)]$$

le risque empirique centré.

Théorème 3.3 (Massart, Nédélec 2003). *Soit S un modèle dénombrable, Soit d la pseudo-distance $L^2(P_X)$ sur $\mathbb{S}$.*

Soient w et ϕ appartenant à la classe $\mathcal{C}_1$ des fonctions ψ croissantes continues de $\mathbb{R}^+$ dans $\mathbb{R}^+$ telles que $x \to \psi(x)/x$ est décroissante et $\psi(1) \geq 1$.

On suppose que w contrôle le module de continuité uniforme de d par rapport à l :

$$\forall \epsilon > 0, \sup_{t\in\mathbb{S},\, l(s^\star,t)\leq\epsilon^2} d(s^\star,t) \leq w(\epsilon)$$

et ϕ le module stochastique de continuité uniforme de $\overline{\gamma}_n$ par rapport à d :

$$\forall u \in S, \forall \sigma > 0 \text{ tel que } \phi(\sigma) \leq \sqrt{n}\sigma^2,$$
$$\sqrt{n}\mathbb{E}\left[\sup_{t\in S, d(u,t)\leq\sigma} \overline{\gamma}_n(u) - \overline{\gamma}_n(t)\right] \leq \phi(\sigma).$$

Soit $\epsilon_\star$ l'unique solution de l'équation $\sqrt{n}\epsilon_\star^2 = \phi(w(\epsilon_\star))$ et $\hat{s} \in S$ tel que $\gamma_n(\hat{s}) \leq \rho + \inf_{t\in S} \gamma_n(t)$ avec $\rho > 0$.

Alors, il existe une constante absolue $\kappa > 0$ telle que

$$\forall y \geq 1, \mathbb{P}\left[l(s^\star,\hat{s}) > 2\rho + 2l(s^\star,S) + \kappa y\epsilon_\star^2\right] \leq e^{-y}.$$

En particulier,

$$\mathbb{E}\left[l(s^\star,\hat{s})\right] \leq 2\left(\rho + l(s^\star,S) + \kappa\epsilon_\star^2\right).$$

Les termes ρ et $l(s^\star,S)$ sont inévitables. Finalement, seul $\epsilon_\star^2$ quantifie la complexité du modèle S. Les deux fonctions ψ et w jouent ainsi un rôle fondamental, via la détermination de $\epsilon_\star$.

Remarquons que l'hypothèse S dénombrable peut être remplacée par une hypothèse de séparabilité que nous ne détaillons pas ici par souci de simplicité.

Éléments de preuve. On peut dégager trois idées principales dans la démonstration de ce théorème.

1. Introduction d'une pondération. Soit $\pi(s^\star) \in S$ tel que $l(s^\star,\pi(s^\star)) \leq l(s^\star,S) + \epsilon_\star^2$. On a alors

$$\begin{aligned} l(s^\star,\hat{s}) &= l(s^\star,\pi(s^\star)) + \gamma_n(\hat{s}) - \gamma_n(\pi(s^\star)) + \overline{\gamma}_n(\pi(s^\star)) - \overline{\gamma}_n(\hat{s}) \\ &\leq [\rho + l(s^\star,\pi(s^\star))] + [\overline{\gamma}_n(\pi(s^\star)) - \overline{\gamma}_n(\hat{s})] \end{aligned}$$

   Pour estimer le deuxième terme, on introduit une pondération, avant de remplacer $\hat{s}$ par $t \in S$ quelconque et de passer au sup. Posons $x = \sqrt{\kappa' y}\epsilon_\star$, $\kappa' \geq 1$ étant à choisir plus tard, et

$$V_x = \sup_{t \in S} \frac{\overline{\gamma}_n(\pi(s^\star)) - \overline{\gamma}_n(t)}{l(s^\star,t) + x^2 + \epsilon_\star^2}.$$

   On a alors

$$\mathbb{P}\left[l(s^\star,\hat{s}) \geq 2\left(\rho + l(s^\star,S)\right) + x^2 + 3\epsilon_\star^2\right] \leq \mathbb{P}\left[V_x \geq \frac{1}{2}\right]$$

   d'où le résultat annoncé avec $\kappa = \kappa' + 3$ si cette dernière probabilité est majorée par e^{-y}.

2. On applique l'inégalité de Bousquet à V_x avec

$$\begin{aligned} v &= n w_1(x)^2, \text{ en notant } w_1(x) = 1 \wedge (2w(x)) \\ b &= 1 \\ f_t(\xi) &= \frac{\gamma(t,\xi) - \gamma(s^\star,\xi) - \mathbb{E}[\gamma(t,\xi) - \gamma(s^\star,\xi)]}{l(s^\star,t) + x^2 + \epsilon_\star^2}, \end{aligned}$$

   d'où $Z = x^2 n V_x$. On obtient alors

$$\mathbb{P}\left[V_x - \mathbb{E}(V_x) \geq \sqrt{\frac{2(w_1(x)^2 x^{-2} + 2\mathbb{E}(V_x))y}{nx^2}} + \frac{y}{3nx^2}\right] \leq e^{-y}.$$

3. On contrôle $\mathbb{E}[V_x]$ à l'aide de l'inégalité maximale (proposition 3.2), avec $u = \pi(s^\star)$, $\omega(t)^2 = l(s^\star,\pi(s^\star)) + (P(\gamma(t,\cdot) - \gamma(\pi(s^\star),\cdot)))_+$ et $\psi(\epsilon) = \phi(2\sqrt{2}w(\epsilon))$, en notant a_+ pour $\max(a,0)$. On obtient alors

$$\mathbb{E}(V_x) \leq \frac{8\sqrt{2}}{\kappa'},$$

   d'où le résultat en prenant κ' assez grand.

□

4 Meilleures bornes pour l'ERM : avec une condition de marge

Munis de ces outils, nous pouvons désormais obtenir une estimation plus fine du risque de l'ERM en faisant des hypothèses supplémentaires sur $\mathcal{P}$ et S. Entre le pessimisme du $\mathcal{O}(n^{-1/2})$ et la vitesse $\mathcal{O}(n^{-1})$ du cas où Y est une fonction déterministe de X, on trouve ainsi toute une variété de risques en $\mathcal{O}(n^{-\alpha})$, $1/2 < \alpha < 1$.

4.1 Condition de marge

Les pires lois P sont celles pour lesquelles η peut être arbitrairement proche de $1/2$. Mammen et Tsybakov [MT99] ont ainsi introduit la *condition de marge* suivante sur P :

$$\forall t \in S, l(s^\star,t) \geq c_0 \|s^\star - t\|^\kappa_{L^1(P_X)}. \tag{4.1}$$

On comprend le lien entre (4.1) et le comportement de η autour de $1/2$ en considérant les deux versions simplifiées ci-dessous.

S'il existe des constantes $C > 0$ et $0 < \alpha \leq \infty$ telles que

$$P\left(0 < |\eta(X) - 1/2| \leq t\right) \leq Ct^\alpha, \tag{4.2}$$

alors (4.1) est vérifiée avec $\kappa = \frac{1+\alpha}{\alpha}$.

Plus simplement encore, s'il existe $h > 0$ tel que

$$P\left(0 < |2\eta(X) - 1| < h\right) = 0 \tag{4.3}$$

alors (4.1) est vérifiée avec $\kappa = 1$ et $c_0 = h$. En effet,

$$\begin{aligned} l(s^\star,t) &= \mathbb{E}_P\left[|2\eta(X) - 1| \cdot |s^\star(X) - t(X)|\right] \\ &\geq h\|s^\star - t\|_{L^1(P_X)} \end{aligned}$$

On notera $\mathcal{P}(S,h)$ l'ensemble des lois P telles que $s^\star \in S$ et la condition (4.3) est vérifiée.

La condition de marge fournit ainsi une mesure du «bruit» induit par P, et permet de considérer des cas moins pessimistes que $P \in \mathcal{P}(S) = \mathcal{P}(S,0)$ qui autorisent η à se concentrer très fortement autour de $1/2$. Le cas le plus optimiste est $\mathcal{P}(S,1)$ et correspond à l'absence d'erreurs.

4.2 Cas des classes de Vapnik

Risque Avec la condition (4.3) et la dimension de Vapnik de $\mathcal{A}$, Massart et Nédélec [MN03] ont obtenu une borne meilleure que (2.1) :

$$\sup_{P \in \mathcal{P}(S,h)} \mathbb{E}_P\left[l(s^\star,\hat{s})\right] \leq \begin{cases} \kappa\sqrt{\frac{V}{n}} & \text{si } h \leq \sqrt{\frac{V}{n}} \\ \kappa\frac{V}{nh}\left(1 + \log\frac{nh^2}{V}\right) & \text{si } h > \sqrt{\frac{V}{n}} \end{cases} \tag{4.4}$$

Éléments de preuve. La condition de marge (4.3) s'écrit $l(s^\star,t) \geq hd^2(s^\star,t)$ et permet de prendre

$$w(\epsilon) = h^{-1/2}\epsilon. \tag{4.5}$$

Notons que le choix $w \equiv 1$ est toujours possible.

Le choix de ϕ est plus délicat et repose sur des inégalités maximales. Il y a au moins deux manières de mesurer la «taille» de $\mathcal{A}$, et donc deux choix possibles pour ϕ : avec l'entropie combinatoire aléatoire et avec l'entropie métrique universelle. Dans ce second cas, on peut exprimer ϕ à l'aide de la dimension de Vapnik V :

$$\phi(\sigma) = K\sigma\sqrt{V(1 + \log(\sigma^{-1} \vee 1))}. \tag{4.6}$$

En combinant les deux choix pour w et les deux choix pour ϕ, nous avons quatre manières d'appliquer le théorème 3.3. Avec ϕ définie par (4.5), $w \equiv 1$ donne $\epsilon_\star^2 = K\sqrt{\frac{V}{n}}$, tandis que la définition (4.5) nous donne

$$\epsilon_\star^2 \leq K^2\left(\frac{V(1 + \log((nh^2/V) \vee 1))}{nh}\right)$$

On en déduit la borne (4.4). □

Risque minimax Celle-ci est minimax — au facteur logarithmique près — puisque

$$\mathcal{R}_{\min\max}(\mathcal{P}(S,h)) \geq \kappa' \left(\sqrt{\frac{V}{n}} \wedge \frac{V}{nh} \right) \text{ si } 2 \leq V \leq n. \tag{4.7}$$

De plus, lorsque $\mathcal{A}$ est l'ensemble des demi-espaces de $\mathbb{R}^d$ (et donc $V = d+1$), le facteur logarithmique est nécessaire, au facteur $(1-h)$ près :

$$\mathcal{R}_{\min\max}(\mathcal{P}(S,h)) \geq \kappa''(1-h) \left[\frac{d}{nh} \left(1 + \log \frac{nh^2}{d} \right) \right] \text{ si } 2 \leq V \leq n. \tag{4.8}$$

4.3 Cas de l'entropie à crochets

On doit parfois utiliser des modèles qui ne sont pas des classes de Vapnik, aussi est-il utile de disposer d'une autre mesure de la complexité d'un ensemble. C'est pourquoi nous introduisons ici l'*entropie à crochets*.

Définition 4.1 (Entropie à crochets). On note $H_{[.]}(\epsilon,S,\mu)$ le log du nombre minimal de crochets $[f,g] = \{h \in S / f \leq h \leq g\}$ qu'il faut pour recouvrir S, avec $\|f-g\|_{L^1(P_X)} \leq \epsilon$.

On dit alors qu'un modèle S a une complexité $\rho > 0$ s'il existe une constante $A > 0$ telle que

$$\forall 0 < \epsilon \leq 1, H_{[.]}(\epsilon,S,P_X) \leq A\epsilon^{-\rho}. \tag{4.9}$$

On note $\mathcal{P}(S,\kappa,c_0,\rho,A)$ l'ensemble des lois P telles que $s^\star \in S$ et les deux conditions (4.1) et (4.9) soient vérifiées avec les constantes correspondantes. Tsybakov [Tsy04] a obtenu sous ces conditions l'estimation suivante du risque de l'ERM.

Théorème 4.1 (Tsybakov). *Soient $\kappa \geq 1$, $c_0 > 0$, $0 < \rho < 1$, $A > 0$, $a > 0$ et $S \subset \mathbb{S}$ un modèle. Pour toute loi $P \in \mathcal{P} \subset \mathcal{P}(S,\kappa,c_0,\rho,A)$, on considère $\mathcal{N}$ un réseau de maille $a \cdot n^{-\frac{1}{1+\rho}}$ sur S pour la pseudo-distance $L^1(P_X)$ et telle que $\mathcal{N}$ a également une borne de complexité ρ. On suppose de plus que*

$$\lim_{t \to 0} \sup_{P \in \mathcal{P}} P\left(0 < \left| \eta(X) - \frac{1}{2} \right| < t \right) = 0.$$

Alors, l'ERM sur $\mathcal{N}$, noté $\hat{s}_n$, vérifie

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P \left[l(s^\star,\hat{s}_n) \right] = \mathcal{O}\left(n^{-\frac{\kappa}{2\kappa+\rho-1}} \right) \tag{4.10}$$

quand n tend vers $+\infty$.

Dans le cas où $\mathcal{A}$ est la famille des régions bornées de $\mathbb{R}^d$ dont le bord a une régularité de Hölder γ, on montre que la complexité de S est $\rho = \frac{d-1}{\gamma}$, et le risque de l'ERM est alors asymptotiquement optimal au sens minimax.

L'ERM présente ainsi une certaine adaptativité au paramètre de marge κ et à la complexité ρ ou V des modèles puisque son mode de calcul (S étant donné) ne dépend pas de ces paramètres. En revanche, le choix d'un modèle S requiert *a priori* certaines informations sur P si l'on veut optimiser l'estimation de $s^\star$. Pour rendre réellement la procédure adaptative, il faut n'utiliser que les $(X_i,Y_i)_{1 \leq i \leq n}$ dans ce choix. C'est l'objectif de la sélection de modèles.

5 Sélection de modèles

5.1 Principe

Considérons une famille $(S_m)_{m\in\mathcal{M}}$ de modèles et un estimateur $\hat{s}_m$ associé à chacun de ces modèles (par exemple l'ERM sur S_m, mais pas nécessairement). L'objet de la sélection de modèles est de choisir $\hat{m} \in \mathcal{M}$ à partir des données pour construire un estimateur $\widetilde{s} = \hat{s}_{\hat{m}}$.

Le meilleur choix possible de m serait l'*oracle* $m^\star$ qui réalise

$$\inf_{m\in\mathcal{M}} \mathbb{E}_P\left[l(s^\star,\hat{s}_m)\right],$$

mais ce choix nécessite la connaissance de P. Une procédure de sélection de modèles idéale atteint une précision d'oracle à constante près :

$$\mathbb{E}_P\left[l(s^\star,\widetilde{s})\right] \leq C \inf_{m\in\mathcal{M}} \mathbb{E}_P\left[l(s^\star,\hat{s}_m)\right].$$

Une telle inégalité est appelée *inégalité d'oracle*.

Souvent, on doit se contenter d'une inégalité plus faible, de la forme

$$\mathbb{E}_P\left[l(s^\star,\widetilde{s})\right] \leq C \inf_{m\in\mathcal{M}} \left(\mathbb{E}_P\left[l(s^\star,\hat{s}_m) + \chi(m)\right)\right]$$

où $\chi(m)$ est un terme de variance (e.g. $\hat{\delta}_n(m)$ ou $pen(m)$, comme défini ci-après).

Par construction, le risque de l'ERM diminue lorsque S grandit, alors que ce n'est pas le cas pour le risque réel de l'ERM au-delà d'une certaine complexité. Il s'agit donc de trouver un compromis entre la minimisation du biais (qui pousse à prendre de gros modèles) et de la variance (qui augmente avec la complexité de ceux-ci).

5.2 Choix par comparaison

Il y a plusieurs méthodes pour choisir $\hat{m}$. Dans le cas de modèles emboîtés $(S_m)_{m\geq 1}$, on peut construire des complexités $\hat{\delta}_n(m)$ dépendant des données et choisir

$$\hat{m} = \inf\left\{m \geq 1/\forall j > m, l(s^\star,\hat{s}_m) - l(s^\star,\hat{s}_j) \leq \hat{\delta}_n(j)\right\}.$$

Un premier pas a été fait par Tsybakov [Tsy04], qui a construit une telle méthode par *comparaison* dans le cas d'une famille finie de modèles emboîtés, avec un risque en

$$\mathcal{O}((\log n)^{\frac{4\kappa}{2\kappa+\rho_m-1}} n^{\frac{-\kappa}{2\kappa+\rho_m-1}})$$

lorsque $P \in \mathcal{P}(S_m,\kappa,c_0,\rho_m,A)$, pour chaque valeur de m. Koltchinskii [Kol03] a montré qu'on pouvait même se passer du facteur logarithmique (et donc être minimax simultanément sur tous les modèles) et considérer un nombre variable N_n de modèles pourvu que N_n ne grandisse pas trop vite. Cependant, on ne peut réellement parler d'adaptativité puisque Tsybakov utilise encore des paramètres dépendant de P dans sa construction, et la méthode de Koltchinskii n'est pas directement applicable.

5.3 Choix par pénalisation

Une autre méthode est la *pénalisation*, c'est-à-dire choisir

$$\hat{m} \in \arg\min_{m \in \mathcal{M}} \{P_n(\gamma(t,\cdot)) + pen(m)\}.$$

La *pénalité* $pen(m)$, qui peut être déterministe ou dépendre également des données, rend compte de la «complexité» du modèle S_m. Koltchinskii [Kol03] a obtenu des inégalités d'oracle en utilisant deux méthodes de pénalisation (en prolongement de Massart [Mas00], d'une part, et de Lugosi et Wegkamp [LW03], d'autre part), mais sans atteindre l'objectif de l'adaptativité. Celle-ci n'a pour l'instant jamais été obtenue, sauf récemment par van de Geer et Tsybakov [TvdG03] dans un cadre très restreint.

Le calcul suivant précise comment l'on doit choisir la pénalité : pour tout $m \in \mathcal{M}$ et $s_m \in S_m$,

$$\gamma_n(\hat{s}_{\hat{m}}) + \mathrm{pen}(\hat{m}) \leq \gamma(\hat{s}_m) + \mathrm{pen}(m) \leq \gamma_n(s_m) + \mathrm{pen}(m).$$

Remarquons que si $t \in \mathbb{S}$,

$$l(s^\star,t) = \gamma_n(t) - \gamma_n(s^\star) - (\overline{\gamma}_n(t) - \overline{\gamma}_n(s^\star)).$$

En appliquant ceci à $t = \hat{s}_{\hat{m}}$ et $t = s_m$, on obtient donc

$$l(s^\star,\hat{s}_{\hat{m}}) \leq [l(s^\star,s_m) + \mathrm{pen}(m)] + [-\mathrm{pen}(\hat{m}) + (\overline{\gamma}_n(s_m) - \overline{\gamma}_n(\hat{s}_{\hat{m}}))].$$

La pénalité doit être assez grande pour quasiment éliminer le second terme, tout en restant de taille raisonnable pour que le premier terme soit de l'ordre de la précision d'oracle. La clé d'une bonne calibration de la pénalité réside ainsi dans la compréhension du processus $\overline{\gamma}_n(s_m) - \overline{\gamma}_n(\hat{s}_{\hat{m}})$. Une idée possible (utilisée par exemple par Birgé et Massart [BM01] dans le cadre de la sélection de modèles gaussienne ; on pourra aussi consulter à ce sujet mon mémoire de maîtrise) est d'introduire une pondération, dans l'esprit de la preuve du théorème 3.3.

6 Perspectives et enjeux

On souhaite obtenir une méthode de sélection de modèles, dans le cadre de la classification ou dans un cadre un peu plus général, qui soit réellement adaptative à la condition de marge (et à la complexité des modèles). Il s'agit d'estimer aussi finement que possible le risque de l'ERM en utilisant au mieux les données, puis d'utiliser ces bornes pour calibrer des pénalités aléatoires adaptatant la sélection de modèles au paramètre de marge.

On attachera une grande importance à la possibilité de mettre en pratique une telle méthode, en particulier en n'utilisant pas implicitement des paramètres que l'on ne peut calculer que si l'on connaît la loi P. Les modèles considérés doivent également être adaptés à la résolution effective d'un problème de minimisation pour calculer chacun des estimateurs parmi lesquels on opère la sélection. Une part importante de ce travail devra être consacrée à des tests numériques de la méthode construite, notamment pour optimiser d'éventuelles constantes laissées libres et pour s'assurer que la vitesse asymptotique est réellement atteinte pour des valeurs raisonnables de n.

Résoudre un tel problème constituerait une avancée significative, non seulement dans le domaine théorique de l'apprentissage statistique mais aussi par ses très nombreuses applications pratiques. L'emploi d'algorithmes d'apprentissage rapide tels que les support vector

machines [SS01] permet ainsi de résoudre des problèmes biologiques (e.g. l'analyse d'une séquence d'ADN ou de données de puces à ADN), informatiques (e.g. la reconnaissance de formes), etc. Être adaptatif s'avèrerait être un atout fondamental, dans la mesure où l'on ne sait absolument pas quelles hypothèses peuvent être légitimement faites.

Références

[BBM02] Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Localized Rademacher complexities. In *Computational learning theory (Sydney, 2002)*, volume 2375 of *Lecture Notes in Comput. Sci.*, pages 44–58. Springer, Berlin, 2002.

[BBM04] Stéphane Boucheron, Olivier Bousquet, and Pascal Massart. Data-driven penalties : heuristics and results. Preprint, February 2004.

[BM01] Lucien Birgé and Pascal Massart. Gaussian model selection. *J. Eur. Math. Soc. (JEMS)*, 3(3):203–268, 2001.

[Bou02] Olivier Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. *C. R. Math. Acad. Sci. Paris*, 334(6):495–500, 2002.

[DL95] L. Devroye and G. Lugosi. Lower bounds in pattern recognition. *Pattern recognition*, 28:1011–1018, 1995.

[Kol03] Vladimir Koltchinskii. Local rademacher complexities and oracle inequalities in risk minimization. Preprint, September 2003.

[Lug02] G. Lugosi. Pattern classification and learning theory. In *Principles of nonparametric learning (Udine, 2001)*, volume 434 of *CISM Courses and Lectures*, pages 1–56. Springer, Vienna, 2002.

[LW03] G. Lugosi and M. Wegkamp. Complexity regularization via localized random penalties. Ann. Stat., to appear, 2003.

[Mas00] Pascal Massart. Some applications of concentration inequalities to statistics. *Ann. Fac. Sci. Toulouse Math. (6)*, 9(2):245–303, 2000. Probability theory.

[Mas03] Pascal Massart. St-flour lecture notes. Empirical processes and Adaptive estimation, 2003.

[MN03] Pascal Massart and Élodie Nédélec. Risk bounds for statistical learning. Preprint, December 2003.

[MT99] Enno Mammen and Alexandre B. Tsybakov. Smooth discrimination analysis. *Ann. Statist.*, 27(6):1808–1829, 1999.

[SS01] Bernhard Scholkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* MIT Press, 2001.

[Tsy04] Alexandre B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Ann. Statist.*, 32(1):135–166, 2004.

[TvdG03] Alexandre B. Tsybakov and S. van de Geer. Square root penalty : adaptation to the margin in classification and in the edge estimation. Preprint, 2003.

[Vap82] Vladimir Vapnik. *Estimation of dependences based on empirical data.* Springer Series in Statistics. Springer-Verlag, New York, 1982. Translated from the Russian by Samuel Kotz.

[Vap98] Vladimir N. Vapnik. *Statistical learning theory.* Adaptive and Learning Systems for Signal Processing, Communications, and Control. John Wiley & Sons Inc., New York, 1998. A Wiley-Interscience Publication.