Apprentissage Statistique
M2 Probabilités et Statistiques, Université Paris-Sud

Cours 2 : Théorie de l'apprentissage statistique: de Vapnik à la localisation (2/2)

SYLVAIN ARLOT ET FRANCIS BACH
NOTES DE COURS INITIALEMENT PRISES PAR NICOLAS DROUGARD (2012)

Table des matières

2.4. Pénalisation 2
2.5. Majoration de la pénalité idéale "globale" 2
2.6. Bornes inférieures 6
3. Localisation et condition de marge 7
3.1. Condition de marge 7
3.2. Bornes supérieures sur le risque : vitesses rapides 8
3.3. Bornes inférieures sous condition de marge 8
3.4. Sélection de modèles sous condition de marge 9
Références 9

Rappels :

- Observations : $D_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$,
  X_i variable explicative et Y_i étiquette associée
- Prédicteur : application mesurable $t : \mathcal{X} \to \mathcal{Y}$
- $\mathbb{S}$ l'ensemble des préditeurs t
- Contraste : $\gamma : \left\{ \begin{array}{ccc} \mathbb{S} \times (\mathcal{X} \times \mathcal{Y}) & \to & \mathbb{R} \\ t, (x, y) & \mapsto & \gamma(t, (x, y)) \end{array} \right.$
- Perte $P\gamma : \left\{ \begin{array}{ccc} \mathbb{S} & \to & \mathbb{R} \\ t & \mapsto & \mathbb{E}_{(X,Y)\sim P} \left[\gamma(t, (X, Y)) \right] \end{array} \right.$
- Prédicteur de Bayes :
$$s^* \in \operatorname*{argmin}_{t \in \mathbb{S}} P\gamma(t)$$
- Perte relative $\ell(s^*, t) = P\gamma(t) - P\gamma(s^*)$
- Risque empirique $P_n\gamma(t) = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} \gamma(t, (x_i, y_i))$

Date: 9 Février 2015.

2.4. **Pénalisation.** Nous rappelons que le minimiseur du risque empirique (ERM pour Empirical Risk Minimiseur) sur le modèle $S_m \subset \mathbb{S}$ est noté $\widehat{s}_m$. La sélection de modèle nous amène à minimiser en $m \in \mathcal{M}$ (indices des modèles), un critère proche du risque de l'ERM . Une première idée serait de choisir le risque empirique, *i.e.* d'utiliser le critère $crit(m) = P_n\gamma(\hat{s}_m)$.
Cependant, si $m_1 < m_2$, c'est-à-dire si le modèle S_{m_2} est plus complexe que S_{m_1}, le minimum du risque empirique sur S_{m_2} est plus petit que sur S_{m_1} : $P_n\gamma(\hat{s}_{m_1}) \geqslant P_n\gamma(\hat{s}_{m_2})$. Cela nous amènerait alors à choisir les modèles les plus complexes, et comme expliqué en section (2.2), mènerait à une explosion de l'erreur d'estimation : le problème du sur-apprentissage (overfitting).

La pénalisation consiste à contourner le problème en pénalisant les modèles les plus complexes : nous considérons les critères de la forme

$$\begin{aligned} &\mathrm{crit}(m) = P_n\gamma\,(\,\widehat{s}_m\,) + \mathrm{pen}(m) \\ \Rightarrow \quad &\widehat{m} \in \arg\min_{m\in\mathcal{M}} \{\, P_n\gamma\,(\,\widehat{s}_m\,) + \mathrm{pen}(m)\,\} \end{aligned}$$

$\mathrm{pen}(m)$ corrige alors le biais de $P_n\gamma\,(\,\widehat{s}_m\,)$ comme estimateur du risque $\mathbb{E}\,[\,P\gamma\,(\,\widehat{s}_m\,)\,]$.
La pénalité la plus adaptée serait la pénalité idéale

$$\mathrm{pen}_{\mathrm{id}}(m) = (P - P_n)\gamma\,(\,\widehat{s}_m\,)$$

car nous minimiserions directement la perte de l'ERM.
Tout d'abord, nous pouvons remarquer l'inégalité suivante définissant la pénalité idéale globale :

$$\mathrm{pen}_{\mathrm{id}}(m) \leqslant \sup_{t\in S_m} \{\,(P - P_n)\gamma\,(\,t\,)\,\} = \mathrm{pen}_{\mathrm{id,g}}(m)$$

Comme précisé dans la section précédente, nous cherchons un modèle satisfaisant une inégalité oracle. Si $\mathrm{pen}(m) \geqslant \mathrm{pen}_{\mathrm{id}}(m)$ pour tout $m \in \mathcal{M}$, en utilisant le lemme "recette" avec $A(m) = \mathrm{pen}(m) - \mathrm{pen}_{\mathrm{id}}(m)$ et $B(m) = 0$,

$$\begin{aligned} \mathrm{pen}(m) - \mathrm{pen}_{\mathrm{id}}(m) \geqslant 0 \quad &\Leftrightarrow \quad \mathrm{crit}(m) - \mathrm{crit}_{\mathrm{id}}(m) \geqslant 0 \\ (\mathit{Lemme}) \quad &\Rightarrow \quad \ell\,(\,s^\star, \widehat{s}_{\widehat{m}}\,) \leqslant \inf_{m\in\mathcal{M}} \{\,\ell\,(\,s^\star, \widehat{s}_m\,) + \mathrm{pen}(m) - \mathrm{pen}_{\mathrm{id}}(m)\,\} \end{aligned}$$

où l'on rappelle que $s^\star$ est l'estimateur de Bayes.
Exemple : L'inégalité est vraie pour la pénalité idéale $\mathrm{pen}(m) = \mathrm{pen}_{\mathrm{id,g}}(m)$ (la pénalité $\mathrm{pen}_{\mathrm{id,g}}(m)$ mesure la « capacité » de S_m).

2.5. **Majoration de la pénalité idéale "globale".** Référence : Section 8.2 de [6]. Nous souhaitons, comme précédemment, trouver une majoration de la pénalité idéale globale

$$\mathrm{pen}_{\mathrm{id,g}}(m) = \sup_{t\in S_m} \{\,(P - P_n)\gamma\,(\,t\,)\,\}$$

afin de majorer l'erreur d'estimation (en espérance) et de construire une pénalité (voir propriété section 2.2).

Dans le cas où la fonction de contraste γ prend ses valeurs dans [0,1] et S_m est fini, l'inégalité obtenue à l'exercice 3 du premier cours permet d'obtenir

une borne supérieure pour $\mathrm{pen}_{\mathrm{id,g}}(m)$: pour tout $x \geqslant 0$,

$$\mathrm{pen}_{\mathrm{id,g}}(m) < \sqrt{\frac{\ln(\mathrm{Card}(S_m))}{2n}} + \sqrt{\frac{x}{2n}}$$

avec probabilité supérieure à $1 - \mathrm{Card}(S_m)\exp(-x)$, si bien que

$$\mathbb{E}\left[\mathrm{pen}_{\mathrm{id,g}}(m)\right] \leqslant \sqrt{\frac{\ln\left(\mathrm{Card}(S_m)\right)}{2n}}$$

L'objectif de la démonstration ci-dessous est de généraliser ce résultat au cas où S_m est éventuellement infini, mais « se comporte essentiellement comme s'il était fini ».

(1) Première étape : concentration. Le résultat suivant, qui peut être vu comme une généralisation de l'inégalité de Hoeffding, nous permet d'obtenir l'inégalité de concentration désirée :

Théorème 1. *Inégalité de Mc Diarmid* [6, Section 5.2]
Soient $\xi_1, \ldots, \xi_n$ des variables aléatoires indépendantes, $F : \mathbb{R}^n \to \mathbb{R}$ mesurable. Notons $Z = F(\xi_1, \ldots, \xi_n)$.

Si $\forall x, x' \in \mathbb{R}^n$, $\forall 1 \leqslant i \leqslant n$,

$$\left|F(\xi_1, \ldots, \xi_{i-1}, \xi_i', \xi_{i+1}, \ldots, \xi_n) - F(\xi)\right| \leqslant c_i$$

Alors $\forall x \leqslant 0$,

$$\mathbb{P}(Z - \mathbb{E}[Z] \geqslant x) \leqslant \exp\left(-\frac{2x^2}{\sum_{i=1}^n c_i^2}\right)$$

Utilisons cette inégalité avec $Z = \sup_{t\in\mathbb{S}_m}(P-P_n)\gamma(t) = F((X_1,Y_1),\ldots,(X_n,Y_n))$. Supposons que $\widehat{t}_i$ réalise $\sup_{t\in\mathbb{S}_m}(P-P_n^{'(i)}) = F((X_1,Y_1),\ldots,(X_i',Y_i'),\ldots,(X_n,Y_n))$:

$$\begin{aligned}
\sup_{t\in\mathbb{S}_m}(P-P_n^{'(i)})\gamma(t) - \sup_{t\in\mathbb{S}_m}(P-P_n)\gamma(t) &\leqslant (P-P_n^{'(i)})\gamma(\widehat{t}_i) - (P-P_n)\gamma(\widehat{t}_i) \\
&= (P_n - P_n^{'(i)})\gamma(\widehat{t}_i) \\
&= \frac{1}{n}\gamma(\widehat{t}_i,(X_i,Y_i)) - \frac{1}{n}\gamma(\widehat{t}_i,(X_i,Y_i)) \\
&\leqslant \frac{1}{n}
\end{aligned}$$

si l'on supposons encore que $0 \leqslant \gamma \leqslant 1$. Nous pouvons alors conclure avec l'inégalité de Mc Diarmid : si $\gamma(t;(x,y)) \in [0,1]$,

$$\mathbb{P}\left(\mathrm{pen}_{\mathrm{id,g}}(m) \geqslant \mathbb{E}\left[\mathrm{pen}_{\mathrm{id,g}}(m)\right] + \sqrt{\frac{x}{2n}}\right) \leqslant e^{-x}$$

(2) Deuxième étape : symétrisation. Nous calculons ici une majoration de l'espérance (sur l'échantillon) de la pénalité idéale globale Nous posons $D_n = ((X_1,Y_1),\ldots,(X_n,Y_n))$ iid $\perp$ $D_n' = ((X_1',Y_1'),\ldots,(X_n',Y_n'))$, et $\varepsilon_1,\ldots,\varepsilon_n$ iid $\perp$ D_n, D_n', des variables de Rademacher, *i.e.* $\xi_1 \in$

$\{-1,1\}$ et $\mathbb{P}(\varepsilon=1)=\frac{1}{2}$.

$$\begin{aligned}
\mathbb{E}[\mathrm{pen}_{\mathrm{id,g}}(m)] &= \mathbb{E}[\sup_{t\in\mathbb{S}_m}(P-P_n)\gamma(t)] \\
&= \mathbb{E}[\sup_{t\in\mathbb{S}_m}(\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{D_n'}\left[\gamma(t,(X_i',Y_i'))-\gamma(t,(X_i,Y_i))\,\middle|\, D_n\right])] \\
&\leqslant \mathbb{E}\left[\sup_{t\in\mathbb{S}_m}(\frac{1}{n}\sum_{i=1}^n \gamma(t,(X_i',Y_i'))-\gamma(t,(X_i,Y_i)))\right] \quad (1)\\
&= \mathbb{E}\left[\sup_{t\in\mathbb{S}_m}\left[\frac{1}{n}\sum_{i=1}^n \varepsilon_i(\gamma(t,(X_i',Y_i'))-\gamma(t,(X_i,Y_i)))\right]\right] \quad (2)\\
&\leqslant \mathbb{E}\left[\sup_{t\in\mathbb{S}_m}\frac{1}{n}\sum_{i=1}^n \varepsilon_i\gamma(t,(X_i',Y_i'))\right]+\mathbb{E}\left[\sup_{t\in\mathbb{S}_m}\frac{1}{n}\sum_{i=1}^n -\varepsilon_i\gamma(t,(X_i,Y_i))\right] \\
&= 2\mathbb{E}\left[\sup_{t\in\mathbb{S}_m}\frac{1}{n}\sum_{i=1}^n \varepsilon_i\gamma(t,(X_i',Y_i'))\right]
\end{aligned}$$

l'inégalité (1) est dûe à l'inégalité de Jensen version conditionnelle, car $\sup_t$ est convexe en γ, et l'égalité (2), la "symétrisation", se comprend facilement en remarquant que les variables (X_i,Y_i) jouent le même rôle à cette étape.

Ainsi, si $\varepsilon_1,\ldots,\varepsilon_n$ sont des variables aléatoires de Rademacher i.i.d.,

$$\mathbb{E}\left[\mathrm{pen}_{\mathrm{id,g}}(m)\right] \leqslant 2\mathbb{E}\left[\frac{1}{n}\sup_{t\in S_m}\left\{\sum_{i=1}^n \varepsilon_i\gamma(t;(X_i,Y_i))\right\}\right]$$

(3) Troisième étape : trois options.
Supposons maintenant que γ est le contraste 0-1 : un Lemme et quelques définitions nous amènent à une majoration de l'espérance de l'étape précédente, conditionnellement à l'échantillon D_n. Notons pour tout $t\in\mathbb{S}$,

$$\begin{aligned}
A_t &= \{\, x\in\mathcal{X} \text{ t.q. } t(x)=1\,\}\subset\mathcal{X} \\
\mathcal{A}_m(D_n) &= \left\{\, (\gamma(t,(X_i,Y_i)))_{1\leqslant i\leqslant n} \text{ t.q. } t\in S_m\,\right\} \\
&= \left\{\, \left(\mathbb{1}_{t(X_i)\neq Y_i}\right)_{1\leqslant i\leqslant n} \text{ t.q. } t\in S_m\right\}\subset\{0,1\}^n
\end{aligned}$$

Alors,

$$\mathrm{Card}(\mathcal{A}_m(D_n))=\mathrm{Card}\,\{\,A_t\cap\{\,X_1,\ldots,X_n\,\} \text{ t.q. } t\in S_m\,\}$$

et

$$\mathbb{E}\left[\sup_{t\in\mathbb{S}_m}\frac{1}{n}\sum_{i=1}^n \varepsilon_i\gamma(t,(X_i,Y_i))\,\middle|\, D_n\right]=\frac{1}{n}\mathbb{E}\left[\sup_{\alpha\in\mathcal{A}_m(D_n)}\sum_{i=1}^n \alpha_i\varepsilon_i\,\middle|\, D_n\right]$$

Nous pouvons donc utiliser le résultat suivant relatif à la concentration de sommes de variables de Rademacher :

Lemme 2 (Section 6.1.1 de [6]). *Si $\mathcal{A} \subset \mathbb{R}^n$ est fini et $(\varepsilon_i)_{i=1,\dots,n}$ sont des variables aléatoires de Rademacher iid, alors*

$$\mathbb{E}\left[\sup_{\alpha\in\mathcal{A}} \sum_{i=1}^{n} \alpha_i\varepsilon_i\right] \leqslant \sqrt{2\log(\mathrm{Card}(\mathcal{A}))\sup_{\alpha\in\mathcal{A}}\sum_{i=1}^{n}\alpha_i^2}$$

Ce résultat est appliqué à $(\alpha_i)_{1\leqslant i\leqslant n} = (\gamma(t,(X_i,Y_i)))_{1\leqslant i\leqslant n} \in \mathcal{A}_m(D_n)$ pour D_n fixé. Nous remarquons que la somme des carrés des α_i est toujours inférieure à n (car nous considérons ici le contraste $0-1$), et donc

$$\begin{aligned}\mathbb{E}\left[\sup_{t\in\mathbb{S}_m}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\gamma(t,(X_i,Y_i))\,\middle|\, D_n\right] &= \frac{1}{n}\mathbb{E}\left[\sup_{\alpha\in\mathcal{A}_m(D_n)}\sum_{i=1}^{n}\alpha_i\varepsilon_i\,\middle|\, D_n\right]\\ &\leqslant \sqrt{\frac{2\log(\mathrm{Card}(\mathcal{A}_m))}{n}}\end{aligned}$$

En définissant l'entropie combinatoire empirique H_m comme suit

$$H_m(X_1,\dots,X_n) = \ln\left(\mathrm{Card}\left(\mathcal{A}_m(D_n)\right)\right)$$

nous pouvons écrire la majoration combinatoire

$$\frac{2}{n}\mathbb{E}\left[\sup_{t\in S_m}\left\{\sum_{i=1}^{n}\varepsilon_i\gamma(t;(X_i,Y_i))\right\}\,\middle|\, D_n\right] \leqslant \frac{2\sqrt{2}}{\sqrt{n}}\sqrt{H_m(X_1,\dots,X_n)} \qquad (3)$$

Ce sont maintenant que trois options s'offrent à nous :

(a) première option : la majoration (3) est facilement exploitable pour le cas des classes de Vapnik-Chervonenkis.

Définition 1. *Une **classe de Vapnik-Chervonenkis (VC)** est un modèle S_m tel que la quantité*

$$V_m := \sup\left\{ k\in\mathbb{N} \;\; t.q. \underbrace{\sup_{x_1,\dots,x_k\in\mathcal{X}} \exp\left(H_m(x_1,\dots,x_k)\right)}_{=\sup_{D_n}\mathrm{Card}(\mathcal{A}_m(D_n))} = 2^k \right\},$$

appelée dimension, est finie.

Par exemple, le modèle S_m associé à l'ensemble des demi-espaces de $\mathbb{R}^d$ est une classe de Vapnik-Chervonenkis de dimension $d+1$. Le Lemme de Sauer, qui s'applique à ce type de classe, permet de majorer l'entropie combinatoire pour un échantillon assez grand :

Lemme 3. *Si S_m est une classe VC, alors $\forall n \geqslant V_m$,*

$$H_m(x_1,\dots,x_n) \leqslant V_m\ln(\frac{en}{V_m})$$

Ainsi, en reprenant le résultat de l'étape 1 ainsi que (3), nous obtenons

$$\mathrm{pen}_{\mathrm{id,g}}(m) \leqslant \frac{2\sqrt{2}}{\sqrt{n}}\sqrt{V_m\left(1+\ln\left(\frac{n}{V_m}\right)\right)} + \sqrt{\frac{2x_m}{n}} =: \mathrm{pen}(m)$$

avec probabilité $1-\exp(-x_m)$. Cette stratégie peut mener à une inégalité oracle avec forte probabilité (voir discussion section 2.4). C'est aussi un bon résultat, à m fixé, pour borner l'erreur d'estimation du modèle. (Pour en savoir plus, voir par exemple [3, Chapitre 13].)

(b) deuxième option : la troisième étape peut nous inciter à utiliser l'entropie combinatoire H dans la pénalité, c'est-à-dire à prendre

$$\mathrm{pen}(m) = C\sqrt{\frac{H_m(X_1,\dots,X_n)}{n}} \ ,$$

en utilisant le fait que $H_m(X_1,\dots,X_n)$ se concentre autour de son espérance [6, Section 5.3.2].

(c) troisième option : Inspirés par la deuxième étape, nous pouvons aussi utiliser la complexité de Rademacher (globale) de S_m :

$$R_n(S_m; D_n) = \frac{1}{n}\mathbb{E}\left[\sup_{t\in S_m}\sum_{i=1}^{n}\varepsilon_i\gamma(t;(X_i,Y_i))\,\middle|\, D_n\right]$$

qui se concentre autour de son espérance, comme H_m [6, Section 5.3.2]. La pénalité correspondante sera alors deux fois cette complexité.

Ces stratégies ont un défaut commun : l'utilisation de $\mathrm{pen}_{\mathrm{id,g}}(m)$, qui est souvent d'un ordre de grandeur plus gros que $\mathrm{pen}_{\mathrm{id}}(m)$ (parce qu'il s'agit d'un supremum sur tout S_m, alors qu'$\widehat{s}_m$ n'est pas « n'importe où » dans S_m), et qui a des déviations en $1/\sqrt{n}$.
L'idée de la localisation est de remettre en cause la première majoration $\mathrm{pen}_{\mathrm{id}}(m) \leqslant \mathrm{pen}_{\mathrm{id,g}}(m)$.

2.6. **Bornes inférieures.** Référence : [4] ou [3, Chapitre 14]. Soit $\mathcal{P}$ une famille de lois sur $\mathcal{X}\times\mathcal{Y}$. Nous définissons dans cette section le risque minimax. Ce risque quantifie l'espérance de la perte relative "en pire cas sur $P\in\mathcal{P}$", loi de l'échantillon D_n, minimisée en l'estimateur construit à partir de l'échantillon $\widetilde{s}:(\mathcal{X},\mathcal{Y})^n\to\mathbb{S}$. Formellement,

Définition 2 (Risque minimax)**.**

$$\mathcal{R}_{\min\max}(\mathcal{P},n) := \inf_{\widetilde{s}\ \textit{estimateur}}\sup_{P\in\mathcal{P}}\left\{\mathbb{E}_{D_n\sim P^{\otimes n}}\left[\ell\left(s^\star,\widetilde{s}(D_n)\right)\right]\right\}.$$

À titre d'exemple, le « No Free Lunch Theorem » (NFLT) exposé dans le premier cours, nous assure que lorsque $\mathcal{X}$ est infini, $\mathcal{Y}=\{0,1\}$ et $\gamma=\gamma_{0-1}$, ce risque minimax sur l'ensemble $\mathcal{P}$ des lois de probabilité sur $\mathcal{X}\times\mathcal{Y}$ est supérieur ou égal à $\frac{1}{2}$.

Le théorème qui suit nous donne une minoration de ce risque lorsque l'on sait que $s^\star$ appartient à une classe VC connue (à comparer avec la majoration du risque de l'ERM sur S_m que l'on peut déduire de la Section 2.5).

Théorème 4 (Devroye & Lugosi, [4]). *Si S_m est une classe de Vapnik-Chervonenkis de dimension $V_m \geqslant 2$ et*

$$\mathcal{P}(S_m) = \{\, P \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y}) \text{ t.q. } s^\star \in S_m \,\} \quad ,$$

alors

$$\mathcal{R}_{\min\max}(\mathcal{P}(S_m), n) \geqslant \kappa \sqrt{\frac{V_m - 1}{n}} \quad .$$

Aide pour la preuve : choisir $x_1, \dots, x_n$ qui maximisent l'entropie $H_m(x_1, \dots, x_n)$, et prendre la loi uniforme sur ces valeurs pour X, et $\eta(x_i) = \frac{1}{2} + \varepsilon_i h$ pour Y avec $h \propto \frac{1}{\sqrt{n}}$, et $\varepsilon_1, \dots, \varepsilon_n$ v.a. de Rademacher iid.

3. Localisation et condition de marge

Référence : [2]

3.1. Condition de marge.

Définition 3 (condition de marge ; Mammen et Tsybakov, 1999). *Pour $\beta \geqslant 0$,*

$$\exists B > 0,\ \forall \varepsilon \geqslant 0, \quad \mathbb{P}\left(0 < |2\eta(X) - 1| \leqslant \varepsilon\right) \leqslant B\varepsilon^\beta \qquad (M_\beta)$$

$$\exists h > 0, \quad \mathbb{P}\left(0 < |2\eta(X) - 1| < h\right) = 0 \qquad (M_\infty)$$

Ces conditions se comprennent bien en pensant à la perte relative en classification binaire avec un contraste 0-1. En effet cette perte vaut $\ell(s^\star, t) = \mathbb{E}\left[\mathbb{1}_{t(X) \neq s^*(X)} \, |2\eta(X) - 1|\right]$, ainsi la condition rend toute erreur de classification plus grave, et ainsi facilite ce problème lorsque $\beta > 0$.

Définition 4. *pour $\alpha \in [0; 1]$,*

$$\exists C > 0,\ \forall t \in \mathbb{S}, \quad \mathbb{E}_{(X,Y)\sim P}\left[\left(\gamma(t; (X, Y)) - \gamma(s^\star; (X, Y))\right)^2\right] \leqslant C\left(\ell(s^\star, t)\right)^\alpha \qquad (N_\alpha)$$

En notant $v = \mathrm{var}(\gamma(t, .) - \gamma(s^\star, .))$ et en utilisant l'inégalité de Bernstein, nous obtenons

$$(P - P_n)(\gamma(t) - \gamma(s^\star)) \leqslant \sqrt{\frac{2x}{n} v} + \frac{x}{3n}$$

avec probabilité $1 - e^{-x}$. Sous la condition (N_α) nous obtenons la majoration

$$\sqrt{\frac{2x}{n} C \ell(s^\star, t)^\alpha} + \frac{x}{3n} \quad .$$

Exercice 1.

(1) Pour tout $\beta \geqslant 0$, $(M_\beta) \Rightarrow (N_{\beta/(1+\beta)})$.
(2) Pour tout $\alpha \in [0, 1[$, $(N_\alpha) \Rightarrow (M_{\alpha/(1-\alpha)})$.
(3) $(M_\infty) \Leftrightarrow (N_1)$.

3.2. **Bornes supérieures sur le risque : vitesses rapides.** Référence : [7] ou [6, Section 8.3]. Pour un argument informel, voir [2, Section 5.3].

Dans cette section, l'inégalité de concentration clef s'appelle l'inégalité de Talagrand. [6, Section 5.3.4].

Théorème 5 (Talagrand, puis Bousquet)**.**
Soit $b > 0$ et $\mathcal{F}$ un ensemble de fonctions $\Xi \mapsto \mathbb{R}$ telles que $\forall f \in \mathcal{F}$,
$P(f) - f \leqslant b$, où $P(f) = \mathbb{E}_{\xi\sim P}[f(\xi)]$.
Soit $Z = \sup_{f\in\mathcal{F}} \{(P - P_n)(f)\}$.

Alors, pour tout $x \geqslant 0$, avec probabilité au moins $1 - e^{-x}$,

$$Z \leqslant \mathbb{E}[Z] + \sqrt{\frac{2\left(\sup_{f\in\mathcal{F}}\{\mathrm{var}_P(f)\} + 2b\mathbb{E}[Z]\right)x}{n}} + \frac{bx}{3n} \quad .$$

Remarque : nous retrouvons l'inégalité de Bernstein lorsque $\mathrm{Card}(\mathcal{F}) = 1$. L'idée est d'appliquer l'inégalité de Talagrand à $\Xi = \mathcal{X} \times \mathcal{Y}$ et $\mathcal{F}$ un sous-ensemble de

$$\mathcal{F}_m = \{\gamma(t;\cdot) - \gamma(s^\star;\cdot) \text{ t.q. } t \in S_m\} \quad ,$$

par exemple,

$$\mathcal{F} = \mathcal{F}_m(r) = \{f \in \mathcal{F}_m,\ \mathrm{var}_P(f) \leqslant r\} \quad .$$

Principe de la localisation : si (M_∞) est satisfaite et si $\ell(s^\star, \widehat{s}_m) \leqslant r_0$, alors l'inégalité de Talagrand appliquée à $\mathcal{F}_m(r_0/h)$ donne une nouvelle borne $r_1 < r_0$ sur $\ell(s^\star, \widehat{s}_m)$ avec grande probabilité. En itérant le processus, on peut obtenir des bornes meilleures que $n^{-1/2}$ pour $\ell(s^\star, \widehat{s}_m(D_n))$.

Théorème 6 (Massart et Nédélec, [7])**.** *Si S_m est une classe de Vapnik de dimension $V_m \geqslant 1$, $s^\star \in S_m$ et si (M_∞) est vérifiée avec $h \geqslant \sqrt{V/n}$, alors tout minimiseur $\widehat{s}_m$ du risque empirique sur S_m vérifie l'inégalité*

$$\mathbb{E}[\ell(s^\star, \widehat{s}_m)] \leqslant C\frac{V}{nh}\left[1 + \ln\left(\frac{nh^2}{V}\right)\right]$$

où C est une constante numérique.

3.3. **Bornes inférieures sous condition de marge.** Référence : [7].

Théorème 7 (Massart et Nédélec, [7])**.** *Si S_m est une classe de Vapnik de dimension $V_m \geqslant 2$ et*

$$\mathcal{P}(S_m, h) = \{P \in \mathcal{M}_1(\mathcal{X}\times\mathcal{Y}) \text{ t.q. } s^\star \in S_m \text{ et } (M_\infty) \text{ est vérifiée}\} \quad ,$$

alors

$$\mathcal{R}_{\min\max}(\mathcal{P}(S_m), n) \geqslant \kappa \min\left\{\frac{V_m}{nh}, \sqrt{\frac{V_m}{n}}\right\} \quad .$$

⇒ Adaptation du minimiseur du risque empirique $\widehat{s}_m$ au paramètre de marge h (au facteur $\ln(n)$ près, parfois nécessaire).

3.4. **Sélection de modèles sous condition de marge.** Références pour en savoir plus : [5] et [1].

- Pour avoir une inégalité-oracle tenant compte de vitesses rapides, on doit prendre $\mathrm{pen}(m)$ plus proche de $\mathrm{pen}_{\mathrm{id}}(m)$ que $\mathrm{pen}_{\mathrm{id,g}}(m)$.
- Il existe des pénalités « locales » (par exemple, les complexités de Rademacher locales), reposant sur une majoration de $(P-P_n)(\gamma(\widehat{s}_m)-\gamma(s^\star))$ utilisant le principe de localisation (en particulier l'inégalité de Talagrand).
- Défauts de telles pénalités.

Références

[1] Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local Rademacher complexities. *Ann. Statist.*, 33(4) :1497–1537, 2005.

[2] Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification : a survey of some recent advances. *ESAIM Probab. Stat.*, 9 :323–375 (electronic), 2005.

[3] Luc P. Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, 1996.

[4] Luc P. Devroye and Gábor Lugosi. Lower bounds in pattern recognition. *Pattern recognition*, 28 :1011–1018, 1995.

[5] Vladimir Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist.*, 34(6) :2593–2656, 2006.

[6] Pascal Massart. *Concentration Inequalities and Model Selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

[7] Pascal Massart and Élodie Nédélec. Risk bounds for statistical learning. *Ann. Statist.*, 34(5) :2326–2366, 2006.

URL: `http://www.di.ens.fr/~arlot/`