Inégalités Oracles pour le hold-out

Joseph Rynkiewicz

11 janvier 2021

1 Introduction

La technique du hold-out (sélection de modèle sur un ensemble de validation) est largement répandue. Cela revient à choisir un meilleur modèle parmi un nombre fini de modèle.

Nous allons voir, que pour la classification, le hold-out vérifie des inégalités oracles qui justifient son importance dans la pratique.

Notons, que bien que nous ne parlerons que de classification, on pourrait généraliser de nombreuses inégalités à d'autres fonctions de perte bornées.

2 Cadre de la classification

Dans toute la suite on considère une suite de variables aléatoires indépendantes et identiquement distribuées (i.i.d.)

$$\left(\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \cdots, \begin{pmatrix} X_n \\ Y_n \end{pmatrix}\right)$$

où $Y_i \in \{1, \cdots K\}$, est une variable qualitative à expliquer (par exemple la classe d'une image) et $X_i \in \mathbb{R}^d$ la variables explicative (par exemple les pixels de l'image).

Soit $\mathcal{F}$, l'ensemble des fonctions de $\mathbb{R}^d \to \{1, \cdots K\}$ et f une fonction de classification, on définit la perte théorique $L(f) = \mathbb{E}\left(\mathbf{1}_{\{Y_i \neq f(X_i)\}}\right)$, la probabilité de mauvaise classification de la fonction f. On notera f^* la meilleure fonction de classification :

$$f^* = \arg\min_{f \in \mathcal{F}} L(f).$$

f^* est l'oracle que fait des prédictions parfaites à partir de $\left(\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \cdots, \begin{pmatrix} X_n \\ Y_n \end{pmatrix}\right)$. Cette fonction n'est pas facilement atteignable, pour l'approximer on se donne un ensemble de fonctions $\mathcal{G}$ (par exemple un ensemble de réseaux profonds) et on cherche à estimer

$$\tilde{f} = \arg\min_{f \in \mathcal{G}} L(f).$$

Si l'ensemble $\mathcal{G}$ est bien choisi, on peut espéré que la perte $L\left(\tilde{f}\right)$ soit proche de la perte de l'oracle $L\left(f^*\right)$.

Evidemment, $L\left(f\right)$ n'est pas calculable directement mais elle peut être approximée par le risque empirique

$$\hat{L}_n\left(f\right) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{Y_i \neq f(X_i)\}}.$$

Si l'ensemble de fonctions $\mathcal{G}$ est trop grand, il se peut que le risque empirique et le risque théorique soient très différents (c'est le phénomène de surapprentissage). Le hold-out résoud ce problème en coupant le problème en deux. On cherche d'abord de bonnes fonctions de classification grâce à l'ensemble d'apprentissage, on obtient ainsi un ensemble fini de fonctions candidates. Ensuite, sur un ensemble indépendant (l'ensemble de validation) on choisit la meilleure fonction de classification parmi cet ensemble fini de fonctions.

Plus formellement, si on dispose d'une collection finie de fonctions $\{f_1, \cdots, f_N\}$ on choisira comme fonction :

$$f_{\hat{k}} = \arg \min_{k \in \{1,\cdots,N\}} \hat{L}_n\left(f_k\right).$$

Si on savait calculer la perte théorique, on aurait plutôt choisi

$$f_{\tilde{k}} = \arg \min_{k \in \{1,\cdots,N\}} L\left(f_k\right).$$

On va donc étudier la différence qu'il peut y avoir entre $L\left(f_{\tilde{k}}\right)$ et $L\left(f_{\hat{k}}\right)$. Mais pour cela on a besoin de quelques résultats simples et utilisés constamment dans toute la suite.

2.1 Quelques astuces de calcul

On va donner des astuces dans le cas particulier des variables bornée par 1, mais des généralisations sont souvent simples. On rappelle que, si Y est une variable aléatoire positive, bornée par 1, on aura, par Fubini :

$$E\left(Y\right) = \int_0^1 P\left(Y > t\right) dt.$$

— Soit $\lambda > 0$ et $C \geq 1$, si on a une variable aléatoire Z , avec $|Z| \leq 1$, qui, pour tout $t > 0$, vérifie l'inégalité

$$P\left(Z > t\right) \leq C \exp\left(-\lambda t\right) = \exp\left(\ln C - \lambda t\right)$$

ainsi

$$P\left(Z - \frac{\ln C}{\lambda} > \varepsilon\right) = P\left(Z > \varepsilon + \frac{\ln C}{\lambda}\right) \leq \exp\left(\ln C - \lambda\varepsilon - \ln\left(C\right)\right) = \exp\left(-\lambda\varepsilon\right).$$

On aura donc,

$E\left(Z-\frac{\ln C}{\lambda}\right)\leq E\left(\max\left(Z-\frac{\ln C}{\lambda},0\right)\right)=\int_0^1 P\left(\max\left(Z-\frac{\ln C}{\lambda},0\right)>t\right)dt\leq$
$\int_0^1 \exp -\lambda t dt=\left[-\frac{1}{\lambda}\exp\left(-\lambda t\right)\right]_0^1=\frac{1}{\lambda}\left(1-\exp\left(-\lambda\right)\right)\leq\frac{1}{\lambda}.$

On en déduit que

$$E\left(Z\right)\leq\frac{1}{\lambda}\left(\ln C+1\right) \tag{1}$$

— Soit $\lambda>0$ et $C\geq 1$, si on a une variable aléatoire $Z>0$, avec $Z\leq 1$, qui, pour tout $t>0$, vérifie l'inégalité

$$P\left(Z>t\right)\leq C\exp\left(-\lambda t^2\right)=\exp\left(\ln C-\lambda t^2\right)$$

ainsi

$$P\left(Z^2>t\right)\leq\exp\left(\ln C-\lambda t\right)$$

$P\left(Z^2-\frac{\ln C}{\lambda}>\varepsilon\right)=P\left(Z^2>\varepsilon+\frac{\ln C}{\lambda}\right)\leq$
$\exp\left(\ln C-\lambda\varepsilon-\ln\left(C\right)\right)=\exp\left(-\lambda\varepsilon\right).$

On aura donc,

$E\left(Z^2-\frac{\ln C}{\lambda}\right)\leq E\left(\max\left(Z^2-\frac{\ln C}{\lambda},0\right)\right)=\int_0^1 P\left(\max\left(Z^2-\frac{\ln C}{\lambda},0\right)>t\right)dt\leq$
$\int_0^1 \exp -\lambda t dt=\left[-\frac{1}{\lambda}\exp\left(-\lambda t\right)\right]_0^1=\frac{1}{\lambda}\left(1-\exp\left(-\lambda\right)\right)\leq\frac{1}{\lambda}.$

On en déduit que

$$E\left(Z^2\right)\leq\frac{1}{\lambda}\left(\ln C+1\right)$$

et comme, par l'inégalité de Jensen, $\left(E\left(Z\right)\right)^2\leq E\left(Z^2\right)$, on aura :

$$E\left(Z\right)\leq\sqrt{\frac{1}{\lambda}\left(\ln C+1\right)} \tag{2}$$

$$E\left(Z\right)\leq\sqrt{\frac{1}{\lambda}\left(\ln C+1\right)}.$$

— Si on a une collection $Z_1,\cdots,Z_N$ de variables aléatoires i.i.d., on a la borne de l'union (une borne plutôt grossière, mais utilisée constamment) :

$$P\left(\max_{k\in\{1,\cdots N\}}Z_k>\varepsilon\right)\leq P\left(\cup_{k\in\{1,\cdots,N\}}\left\{Z_k>\varepsilon\right\}\right)\leq NP\left(Z_k>\varepsilon\right) \tag{3}$$

3 Un premier Résultat en $O\left(\frac{1}{\sqrt{n}}\right)$

3.1 Inégalités exponentielles

Commençons par un résultat de Hoeffding (1963), on pourra en trouver une démonstration dans [Lugosi, 2002] :

Lemme 1 Soit $S(f) = \sum_{i=1}^{n} \mathbf{1}_{\{Y_i \neq f(X_i)\}}$, avec $\mathbb{E}(S(f)) = nL(f) \geq 0$, on a $|S(f) - nL(f)| \leq n$ et l'inégalité de Hoeffding permet d'écrire :

$$P(S(f) > nL(f) + nt) \leq \exp\left(-2nt^2\right), \tag{4}$$

soit

$$P\left(\frac{1}{n}S(f) - L(f) > t\right) \leq \exp\left(-2nt^2\right), \tag{5}$$

et

$$P\left(L(f) - \frac{1}{n}S(f) > t\right) \leq \exp\left(-2nt^2\right), \tag{6}$$

Ce résultat permettrait d'avoir la borne sur l'espérance de $\frac{1}{n}S(f)$, en utilisant les résultats de la section 2.1, on aura

$$E\left(\sup_{k\in\{1,\cdots,N\}} \frac{1}{n}S(f_k) - L(f_k)\right) \leq \sqrt{\frac{\ln N + 1}{2n}}.$$

On peut faire un peu mieux grâce à un résultat que l'on trouvera par exemple dans [Massart, 2003], p. 196. :

Lemme 2 Soit $(f_k)_{1\leq k\leq N}$, une collection finie de fonctions. Notons $Z(f_k) = \sum_{i=1}^{n}\left(\mathbf{1}_{\{Y_i\neq f_k(X_i)\}} - E\left(\mathbf{1}_{\{Y_i\neq f_k(X_i)\}}\right)\right)$, on a $E(Z(f_k)) = 0$, et $-nE\left(\mathbf{1}_{\{Y_i\neq f_k(X_i)\}}\right) \leq Z(f_k) \leq n - nE\left(\mathbf{1}_{\{Y_i\neq f_k(X_i)\}}\right)$, alors

$$E\left(\sup_{k\in\{1,\cdots,N\}} \frac{1}{n}Z(f_k)\right) \leq \sqrt{\frac{\ln N}{2n}}.$$

3.2 Premiere inegalité oracle

Soit $(f_1, \cdots, f_N)$ une collection finie de fonctions obtenues sur l'ensemble d'apprentissage (indépendant de l'ensemble de validation). On note

$$\tilde{k} = \arg\min_{k\in\{1,\cdots,N\}} L(f_k) \text{ et } \hat{k} = \arg\min_{k\in\{1,\cdots,N\}} \hat{L}_n(f_k).$$

On a

$$L\left(f_{\hat{k}}\right) - L\left(f_{\tilde{f}}\right) = L\left(f_{\hat{k}}\right) - \hat{L}_n\left(f_{\hat{k}}\right) + \hat{L}_n\left(f_{\hat{k}}\right) - \hat{L}_n\left(f_{\tilde{k}}\right) + \hat{L}_n\left(f_{\tilde{k}}\right) - L\left(f_{\tilde{f}}\right) \tag{7}$$

Puisque, $\hat{L}_n\left(f_{\hat{k}}\right) = \min_{k\in\{1,\cdots,N\}}\left(\hat{L}_n(f_k)\right)$, alors

$$L\left(f_{\hat{k}}\right) - \hat{L}_n\left(f_{\hat{k}}\right) \leq \max_{k\in\{1,\cdots,N\}}\left(L(f_k) - \hat{L}_n(f_k)\right) \tag{8}$$

De la même façon, puisque $\tilde{k} = \arg\min_{k\in\{1,\cdots,N\}} L(f_k)$, alors

$$\hat{L}_n\left(f_{\tilde{k}}\right) - L\left(f_{\tilde{f}}\right) \leq \max_{k\in\{1,\cdots,N\}}\left(\hat{L}_n(f_k) - L(f_k)\right) \tag{9}$$

Pour $\varepsilon > 0$, la borne de l'union donnera

$$P\left(\max_{k\in\{1,\cdots,N\}}\left(L\left(f_k\right)-\hat{L}_n\left(f_k\right)\right)>\varepsilon\right)\leq$$
$$N\times P\left(\left(L\left(f_k\right)-\hat{L}_n\left(f_k\right)\right)>\varepsilon\right)\leq Ne^{-2n\varepsilon^2}$$

et, symétriquement :

$$P\left(\max_{k\in\{1,\cdots,N\}}\left(\hat{L}_n\left(f_k\right)-L\left(f_k\right)\right)>\varepsilon\right)\leq$$
$$N\times P\left(\left(L\left(f_k\right)-\hat{L}_n\left(f_k\right)\right)>\varepsilon\right)\leq Ne^{-2n\varepsilon^2}$$

Finalement, par l'équation (4), puisque $\hat{L}_n\left(f_{\hat{k}}\right)-\hat{L}_n\left(f_{\tilde{k}}\right)<0$, on aura :

$$P\left(L\left(f_{\hat{k}}\right)-L\left(f_{\tilde{f}}\right)>\varepsilon\right)\leq 2Ne^{-2n\varepsilon^2}.$$

Et utilisant le Lemme 2, on conclut que :

$$E\left(L\left(f_{\hat{k}}\right)-L\left(f_{\tilde{f}}\right)\right)\leq E\left(\max_{k\in\{1,\cdots,N\}}\left(L\left(f_k\right)-\hat{L}_n\left(f_k\right)\right)\right)+$$
$$E\left(\max_{k\in\{1,\cdots,N\}}\left(\hat{L}_n\left(f_k\right)-L\left(f_k\right)\right)>\varepsilon\right)\leq 2\sqrt{\tfrac{\ln N}{2n}}$$

.

Maintenant, si on considère l'oracle f^*, on aura :

$$E\left(L\left(f_{\hat{k}}\right)-L\left(f^*\right)\right)\leq L\left(f_{\tilde{f}}\right)-L\left(f^*\right)+2\sqrt{\frac{\ln N}{2n}}.$$

Le terme $L\left(f_{\tilde{f}}\right)-L\left(f^*\right)$ est un terme de biais qui dépend du choix de l'ensemble $\{f_1,\cdots,f_N\}$. Le terme $2\sqrt{\frac{\ln N}{2n}}$ est le terme de variance provenant du choix de $f_{\hat{k}}$ avec le risque empirique sur l'ensemble de validation.

Remarque sur le nombre N de fonctions. On peut voir que le terme de variance de l'inégalité est en $O\left(\sqrt{\frac{\log N}{2n}}\right)$, on aurait donc pu prendre un nombre de fonctions N qui soit une fonction du nombre d'observations n, par exemple $N = V\log(n)$, où V est une constante positive. On aurait alors une borne de la forme $O\left(\sqrt{\frac{V\log n}{2n}}\right)$, qui converge vers 0 quand $n\to 0$. C'est exactement ce qui arrive si l'ensemble de fonctions n'est pas fini, mais est une classe de Vapnik. La constante V est alors liée à la dimension de Vapnik de l'ensemble de fonctions (cf [Lugosi, 2002]).

4 Convergence rapide

On peut obtenir des vitesses de convergence plus rapides grâce à l'inégalité de Bernstein :

Lemme 3 Soit $S(f) = \sum_{i=1}^{n} \mathbf{1}_{\{Y_i \neq f(X_i)\}}$, avec $\mathbb{E}(S(f)) = nL(f) \geq 0$, on a $|S(f) - nL(f)| \leq n$ et l'inégalité de Bernstein (cf [Lugosi, 2002]) :

Pour $t > 0$, on aura

$$P(S_n - nL(f) > t) \leq \exp\left(-\frac{t^2}{2n\sigma^2 + \frac{2}{3}t}\right) \tag{10}$$

et

$$P(nL(f) - S_n > t) \leq \exp\left(-\frac{t^2}{2n\sigma^2 + \frac{2}{3}t}\right) \tag{11}$$

4.1 Application pour un risque $L(f)$ petit.

Si, on augmente par un petit facteur $\alpha L(f)$ l'écart entre le risque empirique et le risque théorique, on peut obtenir une borne plus rapide, en $O\left(\frac{1}{n}\right)$.

Avec les notations précédentes, $E\left(n\hat{L}_n(f)\right) = nL(f) \geq 0, Var\left(n\hat{L}_n(f)\right) = \sigma^2 = nL(f)(1 - L(f)) \leq nL(f)$ et $\left|n\hat{L}_n(f) - nL(f)\right| \leq n$. Les inégalités de Bernstein permettent d'écrire :

$$\begin{array}{l} P\left(L(f) - \hat{L}_n(f) \geq t\right) \leq \exp\left(-\frac{n^2t^2}{2\left(\sigma^2 + \frac{n}{3}t\right)}\right) \leq \exp\left(-\frac{n^2t^2}{2\left(nL(f) + \frac{n}{3}t\right)}\right) \\ P\left(\hat{L}_n(f) - L(f) \geq t\right) \leq \exp\left(-\frac{n^2t^2}{2\left(\sigma^2 + \frac{n}{3}t\right)}\right) \leq \exp\left(-\frac{n^2t^2}{2\left(nL(f) + \frac{n}{3}t\right)}\right) \end{array}$$

Pour $0 < \alpha < 1$ and $1 \geq \varepsilon > 0$, posons $t = \alpha L(f) + \varepsilon$, alors $2\left(nL(f) + \frac{n}{3}t\right) \leq 2\left(\frac{1}{\alpha} + \frac{1}{3}\right)nt$ et

$$\frac{n^2t^2}{2\left(\frac{1}{\alpha} + \frac{1}{3}\right)nt} = \frac{n(\alpha L(f) + \varepsilon)}{2\left(\frac{1}{\alpha} + \frac{1}{3}\right)} \leq \frac{n(\alpha L(f) + \varepsilon)}{2\left(\frac{3+\alpha}{3\alpha}\right)} \geq \frac{3\alpha(\alpha L(f) + 1)}{2(3+\alpha)}n\varepsilon \geq \frac{3\alpha}{8}n\varepsilon.$$

Ainsi

$$\begin{array}{l} P\left(\frac{1}{1+\alpha}\hat{L}_n(f) - L(f) > \varepsilon\right) = \\ P\left(\hat{L}_n(f) - L(f) > \alpha L(f) + (1+\alpha)\varepsilon\right) \leq \exp\left(-\frac{3\alpha(1+\alpha)}{8}n\varepsilon\right) \leq \exp\left(-\frac{3\alpha(1-\alpha)}{8}n\varepsilon\right) \end{array}$$

De la même façon :

$$\begin{array}{l} P\left(L(f) - \frac{1}{1-\alpha}\hat{L}_n(f) > \varepsilon\right) = \\ P\left(L(f) - \hat{L}_n(f) > \alpha L(f) + (1-\alpha)\varepsilon\right) \leq \exp\left(-\frac{3\alpha(1-\alpha)}{8}n\varepsilon\right). \end{array}$$

On aura de plus :

$$\begin{array}{l} L\left(f_{\hat{k}}\right) - L\left(f_{\tilde{f}}\right) = L\left(f_{\hat{k}}\right) - \frac{1}{(1-\alpha)}\hat{L}_m\left(f_{\hat{k}}\right) + \frac{1}{(1-\alpha)}\hat{L}_n\left(f_{\hat{k}}\right) - \frac{1}{(1-\alpha)}\hat{L}_n\left(f_{\tilde{k}}\right) + \\ \frac{1}{1+\alpha}\hat{L}_n\left(f_{\tilde{k}}\right) - L\left(f_{\tilde{f}}\right) + 2\frac{\alpha}{1-\alpha^2}\hat{L}_n\left(f_{\tilde{k}}\right) \end{array}$$

Ainsi

$$E\left(L\left(f_{\hat{k}}\right)-L\left(f_{\tilde{f}}\right)\right)\leq E\left(\max_{k\in\{1,\cdots,N\}}\left(L\left(f_k\right)-\frac{1}{1-\alpha}\hat{L}_n\left(f_k\right)\right)\right)+$$
$$E\left(\max_{k\in\{1,\cdots,N\}}\left(\frac{1}{1+\alpha}\hat{L}_n\left(f_k\right)-L\left(f_k\right)\right)\right)+2\frac{\alpha}{1-\alpha^2}L\left(f_{\tilde{f}}\right),$$

et pour $\alpha\in]0;1[$,

$$E\left(L\left(f_{\hat{k}}\right)-L\left(f_{\tilde{f}}\right)\right)\leq 2\left(\frac{8\left(\log N+1\right)}{3\alpha\left(1-\alpha\right)n}+\frac{\alpha}{1-\alpha^2}L\left(f_{\tilde{f}}\right)\right).$$

Maintenant, si on considère l'oracle f^*, on aura :

$$E\left(L\left(f_{\hat{k}}\right)-L\left(f^*\right)\right)\leq\left(1+\frac{2\alpha}{1-\alpha^2}\right)\left(L\left(f_{\tilde{f}}\right)-L\left(f^*\right)\right)+$$
$$2\left(\frac{8(\log N+1)}{3\alpha(1-\alpha)n}+\frac{\alpha}{1-\alpha^2}L\left(f^*\right)\right).$$

4.2 Convergence rapide sous condition de bruit

Si on suppose que le bruit (l'erreur de classification pour l'oracle) a de bonne propriété, on peut obtenir des vitesses de convergence encore plus rapide.

Une application immédiate de l'inégalité de Bernstein est que si Z une variable aléatoire de moyenne μ, de variance σ^2 et bornée : $|Z-\mu|\leq b$, alors :

$$P\left(\mu-\frac{1}{n}\sum_{t=1}^{n}Z_i>t\right)=P\left(n\mu-\sum_{t=1}^{n}Z_i>nt\right)\leq\exp\left(-\frac{n^2t^2}{2\left(n\sigma^2+b\frac{n}{3}t\right)}\right).$$

Ainsi, pour $1>\delta>0$,

$$P\left(n\mu-\sum_{t=1}^{n}Z_i>\sqrt{n}\sqrt{2\sigma^2\log\left(\frac{1}{\delta}\right)}+\frac{2b\log\left(\frac{1}{\delta}\right)}{3}\right)\leq$$

$$\exp\left(-\frac{\left(\sqrt{n}\sqrt{2\sigma^2\log\left(\frac{1}{\delta}\right)}+\frac{2b\log\left(\frac{1}{\delta}\right)}{3}\right)^2}{2\left(n\sigma^2+\frac{b}{3}\left(\sqrt{n}\sqrt{2\sigma^2\log\left(\frac{1}{\delta}\right)}+\frac{2b\log\left(\frac{1}{\delta}\right)}{3}\right)\right)}\right)=$$

$$\exp\left(-\frac{2n\sigma^2\log\left(\frac{1}{\delta}\right)+2\sqrt{n}\sqrt{2\sigma^2\log\left(\frac{1}{\delta}\right)}\frac{2b\log\left(\frac{1}{\delta}\right)}{3}+\left(\frac{2b\log\left(\frac{1}{\delta}\right)}{3}\right)^2}{2n\sigma^2+\frac{2b}{3}\sqrt{n}\sqrt{2\sigma^2\log\left(\frac{1}{\delta}\right)}+\frac{4b^2\log\left(\frac{1}{\delta}\right)}{9}}\right)=$$

$$\exp\left(-\log\left(\frac{1}{\delta}\right)\frac{2n\sigma^2+\frac{4b}{3}\sqrt{n}\sqrt{2\sigma^2\log\left(\frac{1}{\delta}\right)}+\frac{4b^2\log\left(\frac{1}{\delta}\right)}{9}}{2n\sigma^2+\frac{2b}{3}\sqrt{n}\sqrt{2\sigma^2\log\left(\frac{1}{\delta}\right)}+\frac{4b^2\log\left(\frac{1}{\delta}\right)}{9}}\right)\leq\delta$$

et finalement :

$$P\left(\mu - \frac{1}{n}\sum_{t=1}^{n} Z_i \leq \sqrt{\frac{2\sigma^2 \log\left(\frac{1}{\delta}\right)}{n}} + \frac{2b\log\left(\frac{1}{\delta}\right)}{3n}\right) \geq 1-\delta \qquad (12)$$

On aura alors le théorème suivant :

Théorème Soit $(f_k)_{1\leq k\leq N}$ un ensemble de fonctions obtenues grâce à un ensemble d'apprentissage. On définit :

$$\tilde{k} = \arg \min_{k\in(1,\cdots,N)} L\left(f_k\right).$$

et

$$\hat{k} = \arg \min_{k\in(1,\cdots,N)} \hat{L}_n\left(f_k\right).$$

On rappelle que f^* est la meilleure fonction de classification (l'oracle). Soit la fonction $w\left(.\right)$ telle que pour toute fonction de classification f,

$$\sqrt{Var\left(\mathbf{1}_{\{f(X)\neq f^*(X)\}}\right)} \leq w\left(L\left(f\right) - L\left(f^*\right)\right)$$

et telle que $\frac{w(x)}{\sqrt{x}}$ ne soit pas croissante. Soit τ^* la plus petite solution positive de $w\left(\varepsilon\right) = \sqrt{n}\varepsilon$, si $\theta \in]0,1[$, alors

$$E\left(L\left(f_{\hat{k}}\right) - L\left(f^*\right)\right) \leq (1+\theta)\left(\left(L\left(f_{\tilde{k}}\right) - L\left(f^*\right)\right) + \left(\frac{8}{3n} + \frac{4\tau^*}{\theta}\right)\left(\log\left(N\right)+1\right)\right).$$

Preuve Puisque $\left|\hat{L}_n\left(f_k\right) - \hat{L}_n\left(f^*\right) - \left(L\left(f_k\right) - L\left(f^*\right)\right)\right| \leq 2$, par l'équation (12) et la borne de l'union, avec probabilité au moins $1-\delta$, pour toutes fonctions f_k,

$$L\left(f_k\right) - L\left(f^*\right) \leq \hat{L}_n\left(f_k\right) - \hat{L}_n\left(f^*\right) + \sqrt{\frac{2\log\frac{N}{\delta}}{n}} w\left(L\left(f_k\right) - L\left(f^*\right)\right) + \frac{4\log\left(\frac{N}{\delta}\right)}{3n},$$

et

$$L\left(f^*\right) - L\left(f_{\tilde{k}}\right) \leq \hat{L}_n\left(f^*\right) - \hat{L}_n\left(f_{\tilde{k}}\right) + \sqrt{\frac{2\log\frac{N}{\delta}}{n}} w\left(L\left(f_{\tilde{k}}\right) - L\left(f^*\right)\right) + \frac{4\log\left(\frac{N}{\delta}\right)}{3n}.$$

En additionant les deux inégalités, puisque, pour tout $k \in \{1,\cdots N\}$,

$Var\left(\mathbf{1}_{\left\{f_{\tilde{k}}(X)\neq f^*(X)\right\}}\right) \leq Var\left(\mathbf{1}_{\{f_k(X)\neq f^*(X)\}}\right)$, on obtient, avec probabilité au moins $1-\delta$,

$$L\left(f_k\right) - L\left(f_{\tilde{k}}\right) \leq \hat{L}_n\left(f_k\right) - \hat{L}_n\left(f_{\tilde{k}}\right) + 2\sqrt{\frac{2\log\frac{N}{\delta}}{n}} w\left(L\left(f_k\right) - L\left(f^*\right)\right) + \frac{8\log\left(\frac{N}{\delta}\right)}{3n}.$$

Comme, par définition, $\hat{L}_n\left(f_{\hat{k}}\right) - \hat{L}_n\left(f_{\tilde{k}}\right) \leq 0$, avec probabilité plus grande que $1-\delta$,

$$L\left(f_{\hat{k}}\right) - L\left(f_{\tilde{k}}\right) \leq 2\sqrt{\frac{2\log\frac{N}{\delta}}{n}} w\left(L\left(f_{\hat{k}}\right) - L\left(f^*\right)\right) + \frac{8\log\left(\frac{N}{\delta}\right)}{3m}.$$

Soit τ^*, comme défini dans le théorème. Si $L\left(f_{\hat{k}}\right) - L\left(f^*\right) \geq \tau^*$, alors

$$\frac{w\left(L\left(f_{\hat{k}}\right) - L\left(f^*\right)\right)}{\sqrt{n}} \leq \sqrt{L\left(f_{\hat{k}}\right) - L\left(f^*\right)\tau^*},$$

et on aura, avec probabilité au moins $1-\delta$,

$$L\left(f_{\hat{k}}\right) - L\left(f_{\tilde{k}}\right) \leq 2\sqrt{2\log\frac{N}{\delta}} \times \sqrt{\tau^*} \times \sqrt{L\left(f_{\hat{k}}\right) - L\left(f^*\right)} + \frac{8\log\left(\frac{N}{\delta}\right)}{3n}.$$

On a pour tout $\theta \in \mathbb{R}$,

$$\frac{\theta^2}{2}\left(L\left(f_{\hat{k}}\right) - L\left(f^*\right)\right) - 2\sqrt{2\log\frac{N}{\delta}} \times \sqrt{\tau^*} \times \sqrt{L\left(f_{\hat{k}}\right) - L\left(f^*\right)}\theta + 4\tau^* \log\frac{N}{\delta} \geq 0,$$

car le discriminant réduit de l'équation du second degrée est nul.

On aura ainsi

$$2\sqrt{2\log\frac{N}{\delta}} \times \sqrt{\tau^*} \times \sqrt{L\left(f_{\hat{k}}\right) - L\left(f^*\right)} \leq \frac{\theta}{2}\left(L\left(f_{\hat{k}}\right) - L\left(f^*\right)\right) + \frac{4\tau^*}{\theta}\log\frac{N}{\delta},$$

ce qui implique :

$$L\left(f_{\hat{k}}\right) - L\left(f_{\tilde{k}}\right) \leq 2\sqrt{2\log\tfrac{N}{\delta}} \times \sqrt{\tau^*} \times \sqrt{L\left(f_{\hat{k}}\right) - L\left(f^*\right)} + \tfrac{8\log\left(\frac{N}{\delta}\right)}{3n} \leq \tfrac{\theta}{2}\left(L\left(f_{\hat{k}}\right) - L\left(f^*\right)\right) + \tfrac{4\tau^*}{\theta}\log\tfrac{N}{\delta} + \tfrac{8\log\left(\frac{N}{\delta}\right)}{3n}.$$

Ainsi, avec probabilité au moins $1-\delta$,

$$\left(1 - \frac{\theta}{2}\right)\left(L\left(f_{\hat{k}}\right) - L\left(f^*\right)\right) \leq L\left(f_{\tilde{k}}\right) - L\left(f^*\right) + \frac{4\tau^*}{\theta}\log\frac{N}{\delta} + \frac{8\log\left(\frac{N}{\delta}\right)}{3n},$$

et

$$\left(L\left(f_{\hat{k}}\right) - L\left(f^*\right)\right) \leq \tfrac{1}{\left(1-\frac{\theta}{2}\right)}\left(L\left(f_{\tilde{k}}\right) - L\left(f^*\right) + \tfrac{4\tau^*}{\theta}\left(\log N + \log\left(\tfrac{1}{\delta}\right)\right) + \tfrac{8\left(\log N + \log\left(\frac{1}{\delta}\right)\right)}{3n}\right),$$

donc

$$\left(L\left(f_{\hat{k}}\right) - L\left(f^*\right)\right) \leq \tfrac{1}{\left(1-\frac{\theta}{2}\right)}\left(L\left(f_{\tilde{k}}\right) - L\left(f^*\right) + \tfrac{4\tau^*}{\theta}\left(\log N\right) + \tfrac{8(\log N)}{3n} + \left(\tfrac{4\tau^*}{\theta} + \tfrac{8}{3n}\right)\log\left(\tfrac{1}{\delta}\right)\right).$$

En prenant l'espérance (car $\hat{k}$ est aléatoire),

$$E\left(L\left(f_{\hat{k}}\right)-L\left(f^{*}\right)\right)\leq\frac{1}{\left(1-\frac{\theta}{2}\right)}\left(L\left(f_{\tilde{k}}\right)-L\left(f^{*}\right)+\left(\frac{4\tau^{*}}{\theta}+\frac{8}{3n}\right)\left(\log N+1\right)\right)$$

Finalement, puisque si $\theta\in]0,1[$, $\frac{1}{\left(1-\frac{\theta}{2}\right)}\leq1+\theta$

$$E\left(L\left(f_{\hat{k}}\right)-L\left(f^{*}\right)\right)\leq(1+\theta)\left(L\left(f_{\tilde{k}}\right)-L\left(f^{*}\right)+\left(\frac{4\tau^{*}}{\theta}+\frac{8}{3n}\right)\left(\log N+1\right)\right)$$

■

Remarque Dans le cadre d'une classification binaire : $Y_k\in\{0,1\}$, on peut expliciter la fonction w et la solution τ^*.

D'abord, définissons la condition de bruit de Mammem Tsybakov (cf par exemple [Boucheron et al., 2005]). Soit $\alpha\in[0,1]$, et soit f^* la meilleure fonction de classification. La condition de Mammem Tsybakov est définie par : $\exists\beta>0,\forall g\in\{0,1\}^{\mathcal{X}}$, $E\left(\mathbf{1}_{g(X)\neq g^{*}(X)}\right)\leq\beta\left(L\left(g\right)-L\left(g*\right)\right)^{\alpha}$. De plus, dans le cas de la classification binaire, si $\eta\left(X\right)=E\left(Y=1|X\right)$, et il existe $s>0$ tel que $\left|2\eta\left(X\right)-1\right|>s$ presque sûrement, alors le cas $\alpha=1$ est réalisé.

Si une telle condition est vérifiée par le bruit, le théorème précédent implique le corollaire suivant :

Corollaire Supposons que condition de Mammem-Tsybakov soit vérifiéz avec un nombre α. Alors, on peut choisir $w\left(r\right)=\left(\frac{r}{h}\right)^{\frac{\alpha}{2}}$ pour une constante positive h et $\tau^{*}=\left(\frac{1}{nh^{\alpha}}\right)^{-\frac{1}{2-\alpha}}$. On déduit alors du précédent théorème :

$$\begin{array}{l}E\left(L\left(f_{\hat{k}}\right)-L\left(f^{*}\right)\right)\leq\\(1+\theta)\left(\left(L\left(f_{\tilde{k}}\right)-L\left(f^{*}\right)\right)+\left(\frac{8}{3n}+\frac{4}{\theta(nh^{\alpha})^{\frac{1}{2-\alpha}}}\right)\left(\log\left(N\right)+1\right)\right)\end{array}$$

et la convergence rapide en $\frac{1}{n}$ est atteinte si $\alpha=1$. Ce résultat permet de dire que le hold-out s'adapte au conditions de bruit (il est adaptatif), cf [Blanchard and Massart, 2006].

Références

[Blanchard and Massart, 2006] Blanchard, G. and Massart, P. (2006). Discussion : Local rademacher complexities and oracle inequalities in risk minimization. *Annals of statistics*, 34 :2664–2671.

[Boucheron et al., 2005] Boucheron, S., Bousquet, O., and Lugosi, G. (2005). Theory of classification : A survey of some recent advances. *ESAIM : PS*, 9 :323–375.

[Lugosi, 2002] Lugosi, G. (2002). Pattern classification and learning theory. In Gyorfi, L., editor, *Principles of Nonparametric Learning*. Springer-Verlag.

[Massart, 2003] Massart, P. (2003). *Concentration inequalities and model selection*. Ecole d'Eté de Probabilités de Saint-Flour XXXIII. Springer-Verlag.