Apprentissage statistique
Chapitre 7 : Vitesses rapides

Lucie Le Briquer

13 mars 2018

Table des matières

1 Hypothèse de bruit de Massart **2**

2 Hypothèse de bruit de Tsybakov **4**

1 Hypothèse de bruit de Massart

Définition 1 (bruit de Massart)

Le bruit dans la classification binaire satisfait l'hypothèse de Massart au niveau $\gamma \in]0, \frac{1}{2}]$ si :

$$\forall x \in \mathcal{X}, \ \left|\eta(x) - \frac{1}{2}\right| \geqslant \gamma$$

Rappels 1 (inégalité de Bernstein)

Si les $X_i \leqslant a$ et $\mathrm{Var}(X_i) = \sigma^2$, alors :

$$\mathbb{P}(\overline{X_n} - \mathbb{E}[X] \geqslant \varepsilon) \leqslant \exp\left(-\frac{n\varepsilon^2}{2\sigma^2 + \frac{2}{3}\varepsilon a}\right)$$

Théorème 1

Supposons que $\mathcal{F}$ est finie et que $f^* \in \mathcal{F}$ (où f^* est le classifieur optimal de Bayes) alors, sous l'hypothèse de Massart au niveau γ,

$$\forall \delta > 0, \quad \mathcal{R}(\hat{f}_{\mathrm{ERM}}) - \mathcal{R}^* \leqslant \frac{\log\left(\frac{|\mathcal{F}|}{\delta}\right)}{n\gamma}$$

avec probabilité au moins $1 - \delta$.

Remarque. L'hypothèse $f^* \in \mathcal{F}$ est nécessaire. Contre-exemple : $\mathcal{X} = \{x_1, x_2\}$ et $\eta(x_1) = 1$, $\eta(x_2) = 0$ mais $\mathcal{F} = \{(0,0), (1,1)\}$. Notons $p = \mathbb{P}(X = x_1)$. Alors :

$$\mathcal{R}(0,0) = p \qquad \mathcal{R}(1,1) = 1 - p$$

Le classifieur optimal minimise $\{p, 1-p\}$. Donc $(0,0)$ est optimal si $p < 1 - p \Leftrightarrow p < \frac{1}{2}$ et $(1,1)$ l'est si $p \geqslant \frac{1}{2}$.
Donc pour distinguer le classifieur optimal, il faut au moins $\frac{1}{|1-2p|^2}$ échantillons. D'où si $|1-2p| \leqslant \frac{1}{\sqrt{n}}$, on va sélectionner le mauvais classifieur avec une probabilité constante. L'erreur est de l'ordre de $\frac{1}{\sqrt{n}}$. On a donc une vitesse lente.

Preuve.

$$\begin{aligned}\mathcal{R}(\hat{f}) - \mathcal{R}(f^*) &= \mathcal{R}(\hat{f}) - \hat{\mathcal{R}}_n(\hat{f}) + \underbrace{\hat{\mathcal{R}}_n(\hat{f}) - \hat{\mathcal{R}}_n(f^*)}_{\leqslant 0} + \hat{\mathcal{R}}_n(f^*) - \mathcal{R}(f^*) \\ &\leqslant \left[\hat{\mathcal{R}}_n(f^*) - \hat{\mathcal{R}}_n(\hat{f})\right] - \left[\mathcal{R}(f^*) - \mathcal{R}(\hat{f})\right]\end{aligned}$$

Or,

$$\hat{\mathcal{R}}_n(f^*) - \hat{\mathcal{R}}_n(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n} \underbrace{\left(\mathbb{1}_{f^*(x_i) \neq y_i} - \mathbb{1}_{\hat{f}(x_i) \neq y_i}\right)}_{Z_i(\hat{f})}$$

Notons $Z_i(f) = \mathbb{1}_{f^*(x_i)\neq y_i} - \mathbb{1}_{f(x_i)\neq y_i}$. Il suffit alors de montrer que $\forall f \in \mathcal{F}$,

$$\frac{1}{n}\sum_{i=1}^{n} Z_i(f) - \mathbb{E}[Z_i(f)] \leqslant \frac{\log\left(\frac{M}{\delta}\right)}{\gamma n}$$

avec probabilité $\geqslant 1-\delta$ (en notant $M = |\mathcal{F}|$).
On va utiliser l'inégalité de Bernstein. Pour ce faire, on calcule :

$$\mathrm{Var}(Z_i(f)) \leqslant \mathbb{E}[Z_i^2(f)] = \mathbb{P}\left(f(x_i) \neq f^*(x_i)\right) = \sigma_f^2$$

En utilisant Bernstein, en notant :

$$\varepsilon(f) := \max\left(\sqrt{\frac{2\sigma_f^2 \log\left(\frac{M}{\delta}\right)}{n}}\ ;\ \frac{2}{3}\frac{\log\left(\frac{M}{\delta}\right)}{n}\right)$$

On obtient avec une probabilité $\geqslant 1-\delta$, $\forall f \in \mathcal{F}$:

$$\frac{1}{n}\sum_{i=1}^{n} Z_i(f) - \mathbb{E}[Z_i(f)] \leqslant \varepsilon(f)$$

Donc en particulier,

$$\mathcal{R}(\hat{f}) - \mathcal{R}(f^*) \leqslant \varepsilon(\hat{f})$$

On va utiliser l'hypothèse de bruit pour borner σ_f^2. On va montrer que $\mathcal{R}(f) - \mathcal{R}(f^*) \geqslant 2\gamma\sigma_f^2$.
On sait que :

$$\mathcal{R}(f) - \mathcal{R}(f^*) = \mathbb{E}\left[2\left|\frac{1}{2} - \eta(x)\right| \mathbb{1}_{f(x)\neq f^*(x)}\right] \geqslant 2\gamma\mathbb{P}(f(x) \neq f^*(x)) = 2\gamma\sigma_f^2 \qquad (*)$$

Donc,

$$\mathcal{R}(\hat{f}) - \mathcal{R}(f^*) \leqslant \max\left(\sqrt{\frac{2\sigma_f^2 \log\left(\frac{M}{\delta}\right)}{n}}\ ;\ \frac{2}{3}\frac{\log\left(\frac{M}{\delta}\right)}{n}\right)$$

et

$$\sigma_{\hat{f}}^2 \leqslant \frac{\mathcal{R}(\hat{f}) - \mathcal{R}(f^*)}{2\gamma}$$

Donc finalement,

$$\mathcal{R}(\hat{f}) - \mathcal{R}(f^*) \leqslant \frac{\log\left(\frac{M}{\delta}\right)}{\gamma n}$$

$\square$

2 Hypothèse de bruit de Tsybakov

Définition 2 ($\alpha-$marge de Tsybakov)
L'hypothèse de $\alpha-$marge de Tsybakov est lorsqu'il existe $c_0 > 0$ et $\varepsilon_0 > 0$ tels que :

$$\mathbb{P}\left(\left|\eta(x) - \frac{1}{2}\right| \leqslant \varepsilon\right) \leqslant c_0 \varepsilon^{\frac{\alpha}{1-\alpha}}$$

pour tout $\varepsilon < \varepsilon_0$.

Remarque. Si $\alpha = 0$ la condition est vide. Si $\alpha = 1$, la condition est celle de Massart.

Théorème 2
Sous la condition de $\alpha-$marge et $f^* \in \mathcal{F}$, on a avec probabilité $\geqslant 1 - \delta$:

$$\mathcal{R}(\hat{f}) - \mathcal{R}(f) \leqslant c \left(\frac{\log\left(\frac{M}{\delta}\right)}{n}\right)^{\frac{1}{2-\alpha}}$$

où c est une constante qui ne dépend que de c_0, ε_0 et α.

Preuve.
La preuve est identique à celle de Massart jusqu'à $(*)$.

$$\begin{aligned}
\mathcal{R}(f) - \mathcal{R}(f^*) &= \mathbb{E}\left[2\left|\frac{1}{2} - \eta(x)\right| \mathbb{1}_{f(x)\neq f^*(x)}\right] \\
&\geqslant \mathbb{E}\left[2\left|\frac{1}{2} - \eta(x)\right| \mathbb{1}_{f(x)\neq f^*(x)} \mathbb{1}_{\left|\frac{1}{2}-\eta(x)\right|\geqslant t}\right] \\
&\geqslant 2t\mathbb{P}\left(f(x) \neq f^*(x) \text{ et } \left|\frac{1}{2} - \eta(x)\right| \geqslant t\right) \\
&\geqslant 2t\mathbb{P}(f(x) \neq f^*(x)) - 2t\mathbb{P}\left(\left|\frac{1}{2} - \eta(x)\right| \leqslant t\right) \quad \text{car } \mathbb{P}(A \cap B) \geqslant \mathbb{P}(A) - \mathbb{P}(B^C) \\
&\geqslant 2t\sigma_f^2 - 2tc_0 t^{\frac{\alpha}{1-\alpha}} \quad \text{si } t \leqslant \varepsilon_0
\end{aligned}$$

En prenant $t = c_1 \sigma_f^{2\frac{1-\alpha}{\alpha}}$, avec c_1 tel que $t \leqslant \varepsilon_0$, on obtient :

$$\mathcal{R}(f) - \mathcal{R}(f^*) \geqslant c_2 \sigma_f^{2^{1+\frac{1-\alpha}{\alpha}}} = c_2 \sigma_f^{\frac{2}{\alpha}}$$

Ainsi,

$$\mathcal{R}(\hat{f}) - \mathcal{R}(f^*) \leqslant \max\left(\sqrt{\frac{2\left(\mathcal{R}(\hat{f}) - \mathcal{R}(f^*)\right)^{\alpha} \log\left(\frac{M}{\delta}\right)}{c_2^{\alpha} n}} \;;\; \frac{2}{3}\frac{\log\left(\frac{M}{\delta}\right)}{n}\right)$$

D'où :

$$\mathcal{R}(\hat{f}) - \mathcal{R}(f^*) \leqslant c_3 \left(\frac{\log\left(\frac{M}{\delta}\right)}{n}\right)^{\frac{1}{2-\alpha}}$$

□