

Apprentissage PAC-Bayésien : de la classification à la régression

Olivier Catoni

CNRS

INRIA - CLASSIC

Département de Mathématiques et Applications,

École Normale Supérieure

45 rue d'Ulm,

75 230 Paris Cedex 05,

`Olivier.Catoni@ens.fr`

ENS Paris,

mercredi 16 mars 2011

Comment faire le moins d'hypothèses possibles ?

(C'est le thème général de CLASSIC !)

- Suites individuelles :

minimiser $\sum_{i=1}^n \ell[f_i(x_1, \dots, x_{i-1}), x_i]$, voir l'exposé de Gilles Stoltz.

- Échantillon i.i.d. $(W_1, \dots, W_n) \sim \mathbb{P}^{\otimes n}$ de loi inconnue.

L'indépendance peut être assurée par le plan d'expérience, c'est une hypothèse plus réaliste que la connaissance précise des propriétés de \mathbb{P} .

Objectif : estimer $\theta(\mathbb{P}) \in \arg \min_{\theta \in \Theta} \int L(W, \theta) d\mathbb{P}$ par $\hat{\theta}(W_1, \dots, W_n)$

en cherchant à minimiser

$$\ell[\hat{\theta}, \theta(\mathbb{P})] = \int L(W, \hat{\theta}) d\mathbb{P} - \inf_{\theta \in \Theta} \int L(W, \theta) d\mathbb{P}.$$

Exemples :

- 1 Classification : $W = (X, Y) \in \mathcal{X} \times \mathcal{Y}$ où $\mathcal{Y} = \{-1, +1\}$, ou plus généralement $|\mathcal{Y}| < \infty$, $\theta : \mathcal{X} \rightarrow \mathcal{Y}$ et $L(W, \theta) = \mathbb{1}[Y \neq \theta(X)]$.
- 2 Régression quadratique : $W = (X, Y) \in \mathbb{R}^d \times \mathbb{R}$, $\theta \in \Theta \subset \mathbb{R}^d$ et $L(W, \theta) = (Y - \langle \theta, X \rangle)^2$.
- 3 Estimation de densité : $W \in \mathcal{W}$, $\theta \in \mathcal{M}_+^1(\mathcal{W})$, $\theta \ll \pi \in \mathcal{M}_+^1(\mathcal{W})$ et $L(W, \theta) = -\log \left[\frac{d\theta}{d\pi}(W) \right]$. Si $\mathbb{P} \ll \pi$, $\theta(\mathbb{P}) = \mathbb{P}$.

Dans tous les cas, le choix d'un estimateur $\hat{\theta}(W_1, \dots, W_n)$ est lié au comportement du **processus empirique**

$$f \mapsto \int f(W) d\bar{\mathbb{P}} : \mathcal{F} \rightarrow \mathbb{R},$$

où $\bar{\mathbb{P}} = \frac{1}{n} \sum_{i=1}^n \delta_{W_i}$ et la classe de fonctions considérée est

$$\mathcal{F} = \left\{ f_\theta : \mathcal{W} \rightarrow \mathbb{R}; \quad f_\theta(w) = L(w, \theta), \theta \in \Theta \right\}.$$

Comment faire peu d'hypothèses sur \mathbb{P} ?

→ Minimisation structurelle du risque (Vapnik).

$\theta \in \Theta_m$, modèle restreint + sélection du modèle m .

$$\inf \left\{ \int L(W, \theta) d\mathbb{P}; \theta \in \Theta_m \right\} - \inf \left\{ \int L(W, \theta) d\mathbb{P}; \theta \in \bigcup_k \Theta_k \right\}$$

augmente quand Θ_m décroît, alors que

$$\left| \inf \left\{ \int L(W, \theta) d\mathbb{P}; \theta \in \Theta_m \right\} - \inf \left\{ \int L(W, \theta) d\bar{\mathbb{P}}; \theta \in \Theta_m \right\} \right|$$

décroît quand Θ_m décroît :

c'est ce que l'on appelle parfois le **compromis biais variance**, ou compromis entre l'erreur de modélisation et l'erreur d'estimation.

- Inégalités oracles (ici non asymptotiques et en déviations) : du type

$$\mathbb{P}^{\otimes n} \left\{ \int L(W, \theta) d\mathbb{P} \leq \int L(W, \theta) d\bar{\mathbb{P}} + p_\epsilon(\theta, n); \theta \in \Theta \right\} \geq 1 - \epsilon,$$

où la pénalité $p_\epsilon(\theta, n)$ est souvent de la forme $p_\epsilon(\theta, n) = p_\epsilon(m, n)$, $\theta \in \Theta_m$ (ce qui signifie que l'on borne uniformément les déviations du processus empirique sur chaque Θ_m).

- Le point de vue PAC-Bayésien :

- On va borner les déviations de $\int L(W, \theta) d\bar{\mathbb{P}}$ par rapport à $\int L(W, \theta) d\mathbb{P}$ en s'aidant d'une mesure de probabilité $\pi \in \mathcal{M}_+^1(\Theta)$ sur l'espace des paramètres (à interpréter comme une mesure de codage).
- On considère tout d'abord des estimateurs perturbés : au lieu d'être une fonction déterministe de (W_1, \dots, W_n) , $\hat{\theta}$ est une variable aléatoire dépendant de (W_1, \dots, W_n) , dont la loi conditionnelle $\mathcal{L}(\hat{\theta} | W_1, \dots, W_n) = \rho(W_1, \dots, W_n) \in \mathcal{M}_+^1(\Theta)$ est décrite par la **loi a posteriori** ρ .

- Les débuts historiques de la théorie (due à McAllester) : le cas de la classification où $L(W, \theta) \in \{0, 1\}$. On s'appuie sur l'inégalité

$$\sum_{\lambda \in \Lambda} \int \exp \left\{ \sup_{\rho} n \lambda \int \left[\Phi_{\lambda} \left[\int L(W, \theta) d\mathbb{P} \right] - \int L(W, \theta) d\bar{\mathbb{P}} \right] d\rho - \mathcal{K}(\rho, \pi) - \log |\Lambda| \right\} d\mathbb{P}^{\otimes n} \leq 1,$$

où $\Phi_{\lambda}(p) = -\frac{1}{\lambda} \log [1 - p + p \exp(-\lambda)]$ et

$$\mathcal{K}(\rho, \pi) = \begin{cases} \int \log(d\rho/d\pi) d\rho, & \rho \ll \pi, \\ +\infty, & \text{sinon.} \end{cases}$$

On en déduit qu'avec probabilité au moins $1 - \epsilon$, pour tout $\rho : \mathcal{W}^n \rightarrow \mathcal{M}_+^1(\Theta)$,

$$\iint L(W, \theta) d\mathbb{P} d\rho \leq \inf_{\lambda \in \Lambda} \Phi_\lambda^{-1} \left\{ \iint L(W, \theta) d\bar{\mathbb{P}} d\rho + \frac{\mathcal{K}(\rho, \pi) + \log(|\Lambda|/\epsilon)}{n\lambda} \right\}.$$

REMARQUES : • Posons

$$K(q, p) = q \log(q/p) + (1 - q) \log[(1 - q)/(1 - p)], \quad p, q \in [0, 1],$$

$$\begin{aligned} \inf_{\lambda \in \mathbb{R}_+} \Phi_\lambda^{-1} \left(q + \frac{\delta}{\lambda} \right) &= \sup \left\{ p \in [0, 1]; K(q, p) \leq \delta \right\} \\ &\leq \begin{cases} q + \sqrt{2\delta q(1 - q)} + 2\delta(1 - 2q), & q \in [0, 1/2], \\ q + \sqrt{q/2}, & q \in [1/2, 1]. \end{cases} \end{aligned}$$

- Lien avec l'information mutuelle entre l'estimateur perturbé $\hat{\theta}$ et l'échantillon :

$$\begin{aligned}
 \inf_{\pi} \int \mathcal{K}(\rho, \pi) d\mathbb{P}^{\otimes n} &= \int \mathcal{K}(\rho, \int \rho d\mathbb{P}^{\otimes n}) d\mathbb{P}^{\otimes n} \\
 &= I[\hat{\theta}, (W_1, \dots, W_n)] \stackrel{\text{def}}{=} \mathcal{K}[\mathcal{L}(\hat{\theta}, W_1^n), \mathcal{L}(\hat{\theta}) \otimes \mathcal{L}(W_1^n)] \\
 &= \text{information mutuelle entre } \hat{\theta} \text{ et } W_1^n \\
 &= \text{nbre de bits du codage de } \hat{\theta} \\
 &\quad \text{qui dépendent de l'échantillon } (W_1, \dots, W_n).
 \end{aligned}$$

La valeur de π optimale (théorique car non observable) est $\int \rho d\mathbb{P}^{\otimes n}$. Si l'estimateur est bon, elle est concentrée autour de $\theta(\mathbb{P})$, quand on prend π uniforme sur Θ faute de mieux, on perd de la précision dans la borne (typiquement un facteur $\log(n)$).

Deux directions possibles :

- Minimiser la borne en ρ , pour π fixé. La valeur optimale de ρ est

$$\rho = \pi_{\exp(-n\lambda \int L(W, \theta) d\bar{\mathbb{P}})},$$

où $\frac{d\pi_{\exp(h)}}{d\pi} \stackrel{\text{def}}{=} \frac{\exp(h)}{\int \exp(h) d\pi}$. C'est une conséquence de l'identité

$$\log \left[\int \exp(h) d\pi \right] = \int h d\rho - \mathcal{K}(\rho, \pi) + \mathcal{K}(\rho, \pi_{\exp(h)}).$$

On obtient des estimateurs perturbés appelés estimateurs de Gibbs. La borne devient

$$\Phi_{\lambda}^{-1} \left\{ -\frac{1}{n\lambda} \log \left[\int \exp \left(-n\lambda \int L(W, \theta) d\bar{\mathbb{P}} \right) d\pi \right] - \frac{\log(|\Lambda|/\epsilon)}{n\lambda} \right\}.$$

On peut l'améliorer en remplaçant π par $\pi_{\exp(-n\beta \int L(W, \theta) d\mathbb{P})}$, et en utilisant des bornes relatives, portant sur $L(W, \theta) - L[W, \theta(\mathbb{P})]$.

- Autre direction : choisir des lois π et ρ plus explicites.

Exemple : inégalités de marge pour les SVM.

Classification linéaire : $Y \in \{-1, +1\}$, $X \in \mathbb{R}^d$,

$$L(W, \theta) = \mathbb{1}(\langle \theta, X \rangle Y \leq 0).$$

Support Vector Machines (kernel method)

$$L(W, \theta) = \mathbb{1}(\langle \theta, \Psi(X) \rangle Y \leq 0).$$

où $\Psi : \mathcal{X} \rightarrow \mathcal{H}$, espace de Hilbert. On a besoin uniquement de calculer le noyau positif

$$k(X_1, X_2) = \langle \Psi(X_1), \Psi(X_2) \rangle,$$

l'espace \mathcal{H} peut rester implicite.

Exemples de noyaux (où $X \in \mathbb{R}^d$) :

$$k(X_1, X_2) = (1 + \langle X_1, X_2 \rangle)^s, \quad \dim \mathcal{H} < \infty,$$

$$k(X_1, X_2) = \exp\left(-\|X_1 - X_2\|^2\right), \quad \dim \mathcal{H} = \infty.$$

On va dans la suite se restreindre au cas où $\dim \mathcal{H} < \infty$, mais le passage en dimension ∞ serait possible. On peut alors supposer que $\Psi = \text{identité}$ sans perte de généralité.

Considérons (suivant Langford, Shawe-Taylor, puis McAllester)

$\pi = \mathcal{N}(0, \beta \text{Id})$, $\rho_{\theta_0} = \mathcal{N}(\theta_0, \beta \text{Id})$. Dans ce cas $\mathcal{K}(\rho_{\theta_0}, \pi) = \frac{\beta}{2} \|\theta_0\|^2$.

Posons

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \int_x^{+\infty} \exp\left(-\frac{y^2}{2}\right) dy \leq \exp\left(-\frac{x^2}{2}\right), \quad x \geq 0.$$

Le contraste $L(W, \theta_0)$ ne dépendant pas de la norme de X (ni de celle de θ_0), on peut supposer que $\mathbb{P}(\|X\| = 1) = 1$. Avec probabilité au moins $1 - \epsilon$, pour tout $\theta_0 \in \mathbb{R}^d$,

$$\begin{aligned}
 \int L(W, \theta_0) d\mathbb{P} &\leq [1 - \varphi(\sqrt{\beta})]^{-1} \iint \mathbb{1}(\langle \theta, X \rangle Y \leq 1) d\mathbb{P} d\rho_{\theta_0} \\
 &\leq [1 - \varphi(\sqrt{\beta})]^{-1} \inf_{\lambda \in \Lambda} \Phi_{\lambda}^{-1} \left\{ \iint \mathbb{1}(\langle \theta, X \rangle Y \leq 1) d\bar{\mathbb{P}} d\rho_{\theta_0} \right. \\
 &\quad \left. + \frac{\beta \|\theta_0\|^2 + 2 \log(|\Lambda|/\epsilon)}{2n\lambda} \right\} \\
 &\leq [1 - \varphi(\sqrt{\beta})]^{-1} \inf_{\lambda \in \Lambda} \Phi_{\lambda}^{-1} \left\{ \int \mathbb{1}(\langle \theta_0, X \rangle Y \leq 2) d\bar{\mathbb{P}} + \varphi(\sqrt{\beta}) \right. \\
 &\quad \left. + \frac{\beta \|\theta_0\|^2 + 2 \log(|\Lambda|/\epsilon)}{2n\lambda} \right\} \stackrel{\text{def}}{=} \inf_{\lambda \in \Lambda} B(\lambda, \theta_0).
 \end{aligned}$$

Choix de l'estimateur : $\hat{\theta} \in \arg \min_{\theta \in \mathbb{R}^d} \inf_{\lambda \in \Lambda} B(\lambda, \theta)$.

Pour tous $\lambda_* \in \Lambda$ et θ_* déterministes $\int L(W, \hat{\theta}) d\mathbb{P} \leq B(\lambda_*, \theta_*)$, d'où en majorant cette borne par une borne déterministe,

Proposition (O.C., travail en cours)

Avec probabilité au moins $1 - \epsilon$,

$$\begin{aligned} \int L(W, \hat{\theta}) d\mathbb{P} &\leq \inf_{\lambda \in \Lambda, \theta \in \mathbb{R}^d} [1 - \varphi(\sqrt{\beta})]^{-1} \\ &\times \Phi_{\lambda}^{-1} \left\{ \sup \left\{ p \in [0, 1] : K[p, \int \mathbf{1}(\langle \theta, X \rangle Y \leq 2) d\mathbb{P}] \right. \right. \\ &\leq \log[(|\Lambda| + 1)/\epsilon] / n \left. \right\} \\ &+ \varphi(\sqrt{\beta}) + \frac{\beta \|\theta\|^2 + 2 \log[(|\Lambda| + 1)/\epsilon]}{2n\lambda} \left. \right\}. \end{aligned}$$

De la classification à la régression

On veut maintenant traiter le cas d'une fonction de contraste générale $L(W, \theta) \in \mathbb{R}$. Idée pour remplacer la log-Laplace d'une Bernoulli Φ : introduire une fonction d'influence ψ .

Soit $L'(W, \theta, \theta') = L(W, \theta) - L(W, \theta')$ et ψ croissante au sens large, telle que

$$-\log(1 - x + x^2/2) \leq \psi(x) \leq \log(1 + x + x^2/2).$$

On peut s'appuyer sur l'inégalité

$$\begin{aligned} & \int \exp \left\{ n \int \psi \left[\lambda L'(W, \theta, \theta') \right] d\bar{\mathbb{P}} \right\} d\mathbb{P}^{\otimes n} \\ & \leq \left[1 + \lambda \int L'(W, \theta, \theta') d\mathbb{P} + \frac{\lambda^2}{2} \int L'(W, \theta, \theta')^2 d\mathbb{P} \right]^n \\ & \leq \exp \left[n\lambda \int L'(W, \theta, \theta') d\mathbb{P} + \frac{n\lambda^2}{2} \int L'(W, \theta, \theta')^2 d\mathbb{P} \right]. \end{aligned}$$

En tirant θ suivant une loi a posteriori et en prenant $\theta' = \theta(\mathbb{P}) = \theta_*$, on obtient comme dans le cas de la classification, avec probabilité au moins $1 - \epsilon$, pour toute loi a posteriori ρ ,

$$\begin{aligned} \iint L(W, \theta) d\mathbb{P} d\rho &\leq \int L(W, \theta_*) d\mathbb{P} \\ &+ \iint L(W, \theta) d\bar{\mathbb{P}} d\rho - \int L(W, \theta_*) d\bar{\mathbb{P}} \\ &+ \frac{\lambda}{2} \iint L'(W, \theta, \theta_*)^2 (d\mathbb{P} + d\bar{\mathbb{P}}) d\rho \\ &+ \frac{\mathcal{K}(\rho, \pi) + \log(\epsilon^{-1})}{n\lambda}. \end{aligned}$$

On peut alors choisir des lois ρ et π Gaussiennes et dans le cas quadratique faire des calculs explicites en choisissant comme norme sur \mathbb{R}^d $\|\theta - \theta_*\|^2 = \int [L(W, \theta) - L(W, \theta_*)] d\mathbb{P}$, pour obtenir

Proposition (J.-Y. Audibert, O.C.)

Considérons le cas de la régression quadratique, où

$L(W, \theta) = (\langle \theta, X \rangle - Y)^2$, avec $X \in \mathbb{R}^d$ et $Y \in \mathbb{R}$. Soit Θ un convexe fermé de \mathbb{R}^d ($\Theta = \mathbb{R}^d$ est autorisé). Supposons que $\int \|X\|^4 d\mathbb{P} < +\infty$ et $\int \|X\|^2 [\langle \theta_*, X \rangle - Y]^2 d\mathbb{P} < +\infty$.

Soit $\hat{\theta} \in \arg \min_{\theta \in \Theta} \int L(W, \theta) d\bar{\mathbb{P}}$.

Pour tout ϵ , il existe n_ϵ tel que pour tout $n \geq n_\epsilon$, avec probabilité au moins $1 - \epsilon$,

$$\int L(W, \hat{\theta}) d\mathbb{P} \leq \inf_{\theta \in \Theta} \int L(W, \theta) d\mathbb{P} + \text{ess sup} \int (Y - \langle \theta_*, X \rangle)^2 \mathbb{P}(dY | X) \frac{30d + 1000 \log(3/\epsilon)}{n}.$$



Jean-Yves Audibert and Olivier Catoni.

Robust linear least squares regression.

2010.



Olivier Catoni.

Pac-Bayesian Supervised Classification : The Thermodynamics of Statistical Learning, volume 56 of *IMS Lecture Notes Monograph Series*.

Institute of Mathematical Statistics, 2007.

pages i-xii, 1-163.