Bornes PAC-Bayésiennes, classification et clustering en grande dimension

Olivier Catoni

CNRS, INRIA (projet CLASSIC)
Département de Mathématiques et Applications,
ENS, 45 rue d'Ulm, 75 230 Paris Cedex 05,
`Olivier.Catoni@ens.fr`

JOURNÉE THÉMATIQUE IMAGE ET APPRENTISSAGE
Séminaire Méthodes Mathématiques du Traitement d'Images
Laboratoire Jacques Louis Lions,
Université Pierre et Marie Curie & CNRS

jeudi 5 juillet 2012

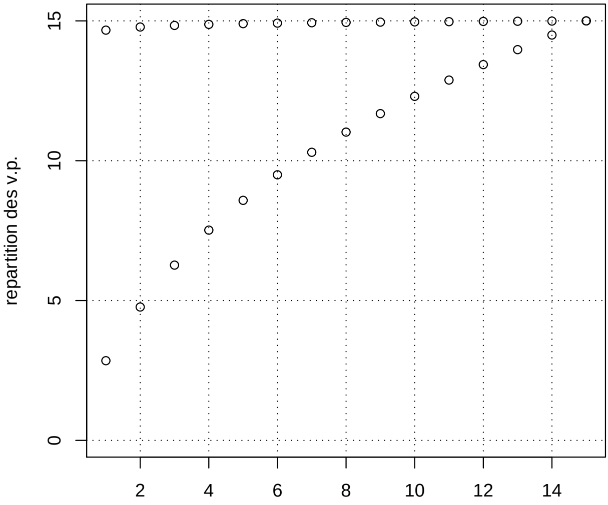
On considère :

- des données (par exemple des images) $(X_1, \dots, X_n) \in \mathcal{F}$, i.i.d. de loi $\mathbb{P}$, à valeurs dans un *espace de formes* $\mathcal{F}$;
- un noyau symétrique positif $k : \mathcal{F} \times \mathcal{F} \to \mathbb{R}_+$, normalisé, $\Big(k(x,y) = k(y.x), \quad \sum_{i,j} \alpha_i k(x_i, x_j) \alpha_j \geq 0, \quad k(x,x) = 1 \Big)$;
- pour tout $x \in \mathcal{F}$ la fonction $K_x = y \mapsto k(x,y) : \mathcal{F} \to \mathbb{R}$;
- le produit scalaire $\langle K_x, K_y \rangle = k(x,y)$;
- l'espace de Hilbert $\mathcal{H}$ complété de $\mathbf{Vect}\{K_x, x \in \mathcal{F}\}$ pour ce produit scalaire ;
- le plongement $\Psi = x \mapsto K_x : \mathcal{F} \to \mathcal{H}$ (construction de Moore-Aronszajn).

Exemple type : $\mathcal{F} \subset \mathbb{R}^d, \quad k(x,y) = \exp(-\lambda \|x - y\|^2)$.

Effet du choix du noyau :

- pour la classification : La règle linéaire dans $\mathcal{H}$ (introduite par Vapnik sous le nom de Support Vector Machine) $\{x \in \mathcal{F} : \langle x, \theta \rangle > 0\}$, où $\theta \in \mathcal{H}$, a une frontière dont la régularité dépend du choix de k ;
- pour le clustering : le choix de k influence la décroissance des valeurs propres de l'opérateur $y \mapsto \int \langle y, x \rangle x \mathrm{d}\mathbb{P}(x) : \mathcal{H} \to \mathcal{H}$, dont les vecteurs propres v_k définissent une partition de $\mathcal{F}$ suivant la valeur de $\arg\max_k |\langle x, v_k \rangle|$.

repartition des v.p.
0
5
10
15
2
4
6
8
10
12
14

Enjeux statistiques

Classification supervisée : on observe un échantillon étiqueté, (X_i, Y_i), où $Y_i \in \{-1, +1\}$. On souhaite relier l'erreur empirique

$$\int \mathbb{1}\Big(\langle x, \theta\rangle y \leq 0\Big) \mathrm{d}\overline{\mathbb{P}}(x, y)$$

où $\overline{\mathbb{P}} = \frac{1}{n} \sum_{i=1}^{n} \delta_{(X_i, Y_i)}$, à son espérance $\int \mathbb{1}\Big(\langle x, \theta\rangle y \leq 0\Big) \mathrm{d}\mathbb{P}(x, y)$.

Clustering : on souhaite relier l'énergie empirique dans la direction $\theta \in \mathcal{H}$

$$e(\theta) = \int \langle \theta, x\rangle^2 \mathrm{d}\overline{\mathbb{P}}(x), \text{ et son espérance } \mathcal{E}(\theta) = \int \langle \theta, x\rangle^2 \mathrm{d}\mathbb{P}(x).$$

Technique de perturbation du paramètre θ : on suppose pour le moment que $\mathcal{H} = \mathbb{R}^d$. On introduit la perturbation Gaussienne

$$\frac{\mathrm{d}\mu_\theta}{\mathrm{d}\theta'}(\theta') = \left(\frac{\beta}{2\pi}\right)^{d/2} \exp\left(-\frac{\beta\|\theta' - \theta\|^2}{2}\right), \qquad \theta \in \mathbb{R}^d.$$

Elle vérifie $\mathcal{K}(\mu_\theta, \mu_0) = \int \log(\mathrm{d}\mu_\theta / \mathrm{d}\mu_0) \mathrm{d}\mu_\theta = \frac{\beta}{2}\|\theta\|^2$.

Fonction d'erreur perturbée :

$$\int \mathbb{1}\Big(\langle x,\theta'\rangle y \leq 0\Big)\mathrm{d}\mu_\theta(\theta') = \varphi\left(\frac{\sqrt{\beta}\langle x,\theta\rangle y}{\|x\|}\right),$$

où $\varphi(a) = \displaystyle\int_a^{+\infty} (2\pi)^{-1/2}\exp(-t^2/2)\mathrm{d}t$

$\left(\leq \dfrac{1}{2}\exp(-a^2/2), \text{ pour tout } a \geq 0\right)$.

On prend un peu de marge :

$$
\begin{aligned}
&\int \mathbb{1}\Big(\|x\|^{-1}\langle x,\theta'\rangle y \leq 1\Big)\mathrm{d}\mu_\theta(\theta') \\
&\qquad = \varphi\Big[\sqrt{\beta}\Big(\|x\|^{-1}\langle x,\theta\rangle y - 1\Big)\Big] \geq \varphi(-\sqrt{\beta})\mathbb{1}\Big(\langle x,\theta\rangle y \leq 0\Big) \\
&\qquad\qquad\qquad\qquad\qquad\qquad \Big[1-\varphi(\sqrt{\beta})\Big]\mathbb{1}\Big(\langle x,\theta\rangle y \leq 0\Big),
\end{aligned}
$$

pour aboutir à

$$
\begin{aligned}
&\int \mathbb{1}\Big(\langle x,\theta\rangle y \leq 0\Big)\mathrm{d}\mathbb{P}(x,y) \leq \Big[1-\varphi(\sqrt{\beta})\Big]^{-1} \\
&\qquad\qquad \times \int \underbrace{\mathbb{1}\Big(\|x\|^{-1}\langle x,\theta'\rangle y \leq 1\Big)}_{\text{erreur avec marge}} \underbrace{\mathrm{d}\mu_\theta(\theta')}_{\text{perturbation}} \mathrm{d}\mathbb{P}(x,y).
\end{aligned}
$$

Inégalité PAC-Bayésienne :

$$\exp\left(\int h(\theta')\mu_\theta(\theta') - \mathcal{K}(\mu_\theta, \mu_0)\right) \leq \int \exp\left[h(\theta')\right] \mathrm{d}\mu_0(\theta').$$

Donne

$$\int \exp\Bigg\{\sup_{\theta \in \mathbb{R}^d} n\lambda \int \Bigg[\Phi_\lambda\left(\int \mathbb{1}\left[\|x\|^{-1}\langle x, \theta'\rangle y \leq 1\right] \mathrm{d}\mathbb{P}(x,y)\right) \\ - \int \mathbb{1}\left(\|x\|^{-1}\langle x, \theta'\rangle y \leq 1\right) \mathrm{d}\overline{\mathbb{P}}(x,y)\Bigg] \mathrm{d}\mu_\theta(\theta') \\ - \mathcal{K}(\mu_\theta, \mu_0)\Bigg\} \mathrm{d}\mathbb{P}^{\otimes n} \leq 1,$$

où $\Phi_\lambda(p) = -\lambda^{-1} \log\left[1 - p + p\exp(-\lambda)\right]$.

Avec probabilité au moins $1-\epsilon$, pour tout $\theta \in \mathbb{R}^d$,

$$
\begin{aligned}
\int \mathbb{1}\big(\langle x,\theta\rangle y \leq 0\big)\mathrm{d}\mathbb{P}(x,y) \leq \big[1-\varphi(\sqrt{\beta})\big]^{-1} \\
\times \inf_{\lambda\in\Lambda} \Phi_\lambda^{-1}\bigg\{\int \varphi\big[\sqrt{\beta}\big(\|x\|^{-1}\langle x,\theta\rangle y - 1\big)\big]\mathrm{d}\overline{\mathbb{P}}(x,y) \\
+ \frac{\beta\|\theta\|^2/2 + \log\big(|\Lambda|/\epsilon\big)}{n\lambda}\bigg\} \\
\leq \big[1-\varphi(\sqrt{\beta})\big]^{-1} \\
\times \inf_{\lambda\in\Lambda} \Phi_\lambda^{-1}\bigg\{\int \Big(2-\|x\|^{-1}\langle\theta,x\rangle y\Big)_+ \mathrm{d}\overline{\mathbb{P}}(x,y) + \frac{\beta\|\theta\|^2}{2n\lambda} \\
+ \frac{\log\big(|\Lambda|/\epsilon\big)}{n\lambda} + \varphi(\sqrt{\beta})\bigg\},
\end{aligned}
$$

indépendamment de la dimension !

Résultat dépendant de la dimension :
Considérons l'approximation suivante de l'identité

$$\psi(a) = \begin{cases} -\log\left(1 - a + a^2/2\right), & 0 \leq a \leq 1, \\ \log(2), & a \geq 1, \\ -\psi(-a), & a \leq 0. \end{cases}$$

Soit $e(\theta)$ la solution de

$$\int \psi\Big(\lambda\Big[\langle x, \theta\rangle^2 e(\theta)^{-1} - 1\Big]\Big) \mathrm{d}\overline{\mathbb{P}}(x) = 0.$$

En perturbant θ comme précédemment et en utilisant le changement de variable $G^{1/2}\theta, G^{-1/2}x$ (qui laisse le produit scalaire invariant), on obtient avec probabilité au moins $1-\epsilon$

$$\sup_{\theta \in \mathbb{R}^d \setminus \{0\}} |\mathcal{E}(\theta)/e(\theta) - 1| \leq 2(\eta + \gamma),$$

où

$$\eta = \sqrt{\frac{2(\kappa^2-1)}{n}\left[\log(2/\epsilon) + \frac{10 s_4^2}{9\kappa}\right]},$$

$$\gamma = 6\sqrt{\frac{5 s_4^2 \kappa}{n}},$$

$$s_4^2 = \sqrt{\int \| G^{-1/2}x \|^4 \, \mathrm{d}\mathbb{P}(x)} \geq d = \int \| G^{-1/2}x \|^2 \, \mathrm{d}\mathbb{P}(x),$$

$$\kappa = \sup\left\{ \sqrt{\int \langle x, \theta \rangle^4 \mathrm{d}\mathbb{P}(x)}, \theta \in \mathbb{R}^d, \int \langle x, \theta \rangle^2 \mathrm{d}\mathbb{P}(x) \leq 1 \right\}.$$