Examen du cours d'Apprentissage statistique

Enseignants :
Francis Bach – Olivier Catoni – Rémi Lajugie – Guillaume Obozinski

Session de juin 2013

Les exercices qui suivent sont indépendants et peuvent donc être abordés dans un ordre arbitraire.

Ils ne sont pas classés par ordre de difficulté.

Merci de rendre chaque exercice sur une feuille séparée !

Durée : 3 heures – **Tous documents autorisés**

1 Fonction de perte de Huber et régression quadratique bornée

Soit (X, Y) un couple de variables aléatoires, où $X \in \mathbb{R}^d$ et $Y \in \mathbb{R}$.
On considère n copies indépendantes (X_i, Y_i), $1 \leqslant i \leqslant n$ du couple (X, Y).
On suppose que $\mathbb{E}(Y^2) < +\infty$
et qu'il existe un réel positif R tel que $\mathbb{P}(\|X\| \leqslant R) = 1$.

Introduisons la fonction de Huber $\psi : \mathbb{R} \to \mathbb{R}$ définie par

$$\psi(z) = \begin{cases} z^2/2, & |z| \leqslant 1, \\ |z| - 1/2, & |z| \geqslant 1. \end{cases}$$

Pour tout paramètre d'échelle $S \in \mathbb{R}_+^*$, on définit de même $\psi_S(z) = S^2\psi(z/S)$.
On s'intéresse à la fonction de perte

$$L(x, y, \theta) = \psi_S\big(y - \langle \theta, x\rangle\big), \quad \text{où } x \in \mathbb{R}^d, \theta \in \mathbb{R}^d, y \in \mathbb{R}, \langle \theta, x\rangle = \sum_{j=1}^{d} \theta_j x_j.$$

Soit $\mathbb{B}_d = \{\theta \in \mathbb{R}^d : \|\theta\| = 1\}$ la boule unité de $\mathbb{R}^d$ et

$$\theta_* \in \arg\min_{\theta \in \mathbb{B}_d} \mathbb{E}\big[L(X, Y, \theta)\big]$$

le minimiseur du risque de Huber sur la boule unité.
Considérons l'estimateur $\widehat{\theta} \in \mathbb{B}_d$ de θ_* défini par

$$\widehat{\theta} \in \arg\min_{\theta \in \mathbb{B}_d} \sum_{i=1}^{n} L(X_i, Y_i, \theta).$$

1. Montrer que la proposition 2.5 page 12 du cours s'applique (cette proposition est rappelée à la page suivante).
   Plus précisément, montrer que pour tout $x \in \mathbb{R}^d$ tel que $\|x\| \leqslant R$, tout $y \in \mathbb{R}, \theta \in \mathbb{R}^d$,

   $$\big|L(x, y, \theta') - L(x, y, \theta)\big| \leqslant RS\|\theta' - \theta\|,$$
   $$\big|L(x, y, \theta') - L(x, y, \theta) - \langle \nabla L(x, y, \theta), \theta' - \theta\rangle\big| \leqslant \frac{R^2}{2}\|\theta' - \theta\|^2,$$
   $$\text{où } \nabla L(x, y, \theta) = -S\psi'\big[S^{-1}\big(y - \langle \theta, x\rangle\big)\big]x.$$

   On pourra considérer la fonction $g(t) = L\big[x, y, (1-t)\theta + t\theta'\big]$ et borner successivement $|g(1) - g(0)|$ et $|g(1) - g(0) - g'(0)|$ en utilisant la formule de Taylor avec reste intégral.
2. On suppose de plus que $\mathbb{P}(|Y| \leqslant R) = 1$ et que $S = 2R$.
   (a) Montrer que
   $$\theta_* = \arg\min_{\theta \in \mathbb{B}_d} \mathbb{E}\big[\big(Y - \langle \theta, X\rangle\big)^2\big].$$
   (b) Montrer que pour tout $\theta \in \mathbb{B}_d$,
   $$\mathbb{E}\big[L(X, Y, \theta)\big] - \mathbb{E}\big[L(X, Y, \theta_*)\big] \geqslant \frac{1}{2}\mathbb{E}\big[\langle \theta - \theta_*, X\rangle^2\big].$$
   Pour cela, on pourra considérer la fonction $g(t) = \mathbb{E}\big[L(X, Y, t\theta + (1-t)\theta_*)\big]$ et montrer que $g'(0) \geqslant 0$.

(c) Soit $\lambda_{\min} = \inf\{\mathbb{E}(\langle\theta, X\rangle^2), \theta \in \mathbb{R}^d, \|\theta\| = 1\}$ la plus petite valeur propre de la forme quadratique $\theta \mapsto \mathbb{E}(\langle\theta, X\rangle^2)$. On suppose que $\lambda_{\min} > 0$ et on introduit comme dans la proposition 2.5 la fonction

$$\chi(h) = \sup_{\theta\in\mathbb{B}_d}\left(\frac{h}{2}\|\theta-\theta_*\|^2 - \mathbb{E}\big[L(X,Y,\theta) + \mathbb{E}\big[L(X,Y,\theta_*)\big]\right), \qquad h \in \mathbb{R}_+^*.$$

Montrer que $\chi(\lambda_{\min}) = 0$.

(d) Pour tout $\varepsilon > 0$, en déduire des majorants de

$$\mathbb{E}\big[\big(Y - \langle\widehat{\theta}, X\rangle\big)^2\big] - \mathbb{E}\big[\big(Y - \langle\theta_*, X\rangle\big)^2\big]$$
$$\text{et de } \|\widehat{\theta} - \theta_*\|^2$$

valables avec probabilité $1 - \varepsilon$.

Rappel de la proposition 2.5 du cours

Proposition 2.5 *Considérons n variables aléatoires indépendantes X_i, $1 \leqslant i \leqslant n$ à valeurs dans un espace mesurable $\mathcal{X}$, l'espace des paramètres $\Theta = \mathbb{R}^d$ et une fonction mesurable $f : \mathcal{X} \times \Theta \to \mathbb{R}$ telle que $\mathbb{E}\big[f(X_i,\theta)^2\big] < +\infty$, $\theta \in \Theta$, $1 \leqslant i \leqslant n$. Posons*

$$M(\theta) = \frac{1}{n}\sum_{i=1}^n f(X_i,\theta),$$
$$m(\theta) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\big[f(X_i,\theta)\big].$$

Supposons qu'il existe une fonction mesurable $(x,\theta) \mapsto \nabla f(x,\theta) \in \mathbb{R}^d$ et des constantes positives g et H telles que

$$|f(x,\theta) - f(x,\theta')| \leqslant g\|\theta-\theta'\|,$$
$$|f(x,\theta') - f(x,\theta) - \langle\nabla f(x,\theta), \theta'-\theta\rangle| \leqslant \frac{H}{2}\|\theta'-\theta\|^2, \qquad x \in \mathcal{X}, \quad \theta,\theta' \in \mathbb{R}^d.$$

Soit $\theta_* \in \arg\min_{\theta\in\mathbb{B}_d} m(\theta)$. Introduisons pour tout $h > 0$ la fonction

$$\chi(h) = \sup_{\theta\in\mathbb{B}_d}\frac{h}{2}\|\theta-\theta_*\|^2 - m(\theta) + m(\theta_*),$$

Dans ces conditions, le minimiseur empirique $\widehat{\theta} \in \arg\min_{\theta\in\mathbb{B}_d} M(\theta)$ de m sur la boule unité vérifie avec probabilité au moins $1-\varepsilon$

$$\|\widehat{\theta} - \theta_*\|^2 \leqslant \frac{8g^2}{nh^2}\left[\left(\frac{8H}{h}+1\right)d + 2\log\big(\varepsilon^{-1}\big)\right] + \frac{4\chi(h)}{h}$$
$$\text{et } m(\widehat{\theta}) - m(\theta_*) \leqslant \frac{4g^2}{nh}\left[\left(\frac{8H}{h}+1\right)d + 2\log\big(\varepsilon^{-1}\big)\right] + \chi(h).$$

2 Apprentissage du noyau

Dans ce problème, une approche pour apprendre le noyau directement à partir des données sera étudiée.

Etant donné n observations $(x_i, y_i) \in \mathcal{X} \times \mathbb{R}$ et une application $\Phi : \mathcal{X} \to \mathbb{R}^d$ ("feature map"), on considère le problème de la régression "ridge" :

$$\min_{w \in \mathbb{R}^d} \frac{1}{2n} \sum_{i=1}^{n} (y_i - w^\top \Phi(x_i))^2 + \frac{\lambda}{2} w^\top w. \tag{1}$$

(a) En introduisant des variables auxiliaires $u_i = w^\top \Phi(x_i)$, déterminer un problème dual à (1) et montrer qu'il y a dualité forte. On montrera notamment que le problème fait intervenir la matrice de noyau K, qui est de taille $n \times n$ et telle que $K_{ij} = \Phi(x_i)^\top \Phi(x_j)$.

(b) On appelle $F(K)$ la valeur optimale commune pour le problème (1) et son dual. Sans calculer $F(K)$, montrer que F est une fonction convexe. Résoudre le problème dual pour obtenir $F(K)$.

(c) On suppose maintenant que m matrices de noyau $K^1, \dots, K^m$ sont disponibles. Montrer que pour tout $\eta \in \mathcal{S} = \left\{\eta \in \mathbb{R}^m,\ \eta \geqslant 0,\quad 1^\top \eta = 1\right\}$, $K(\eta) = \sum_{j=1}^m \eta_j K^j$ est une matrice semi-définie positive. On note $G(\eta) = F(K(\eta))$ et on va maintenant chercher à minimiser $G(\eta)$ par rapport à η.

(d) Calculer le gradient de G. On pourra montrer (en utilisant une méthode de démonstration vue en cours) et utiliser que la différentielle de la fonction $M \mapsto M^{-1}$ en une certaine matrice symmétrique définie positive M est égale à $N \mapsto -M^{-1}NM^{-1}$.

(e) Décrire un algorithme pour la projection orthogonale sur le simplexe $\mathcal{S}$, i.e., pour determiner le minimum global de $\min_{\eta \in \mathcal{S}} \frac{1}{2}\|\eta - \zeta\|^2$. On introduira un Lagrangien et une variable duale uniquement pour la contrainte $1^\top \eta = 1$, et on utilisera (en le montrant) un résultat intermédiare $\min_{\eta_i \geqslant 0} \frac{1}{2}(\eta_i - \zeta_i)^2 + \lambda \eta_i = \frac{1}{2}\zeta_i^2 - \frac{1}{2}(\zeta_i - \lambda)_+^2$, où $c_+ = \max\{c, 0\}$.

(f) Décrire un algorithme pour la minimisation de G sur $\mathcal{S}$.

3 Apprentissage d'une fonction Lipschitz en design fixe

Soit $f^* : [0,1] \to \mathbb{R}$ une fonction Lipschitz de constante L, c'est-à-dire telle que

$$\forall x, y \in [0,1], \qquad |f^*(x) - f^*(y)| \leqslant L|x-y|.$$

On souhaite apprendre f^* à partir d'observations bruitées en des points x_i qui prennent la forme

$$Y_i = f^*(x_i) + \varepsilon_i, \tag{2}$$

où les variables aléatoires ε_i sont indépendantes et vérifient $\mathbb{E}[\varepsilon_i] = 0$ et $\mathbb{E}[\varepsilon_i^2] = \sigma^2 < \infty$. Plus précisément on souhaite minimiser le risque quadratique correspondant à une loi uniforme des données d'entrées sur l'intervalle $[0,1]$, ou de façon équivalente, l'excès de risque que l'on peut écrire

$$\mathcal{E}(f) := \mathcal{R}(f) - \mathcal{R}(f^*) = \int_0^1 (f(x) - f^*(x))^2 \, dx.$$

On suppose néanmoins, qu'au lieu d'obtenir des données d'entrées tirées aléatoirement, on obtient des observations bruitées Y_i selon l'équation (2) pour **des données d'entrée déterministes** de la forme $x_i = \frac{i}{n}$ pour $i \in \{1, \dots, n\}$, donc régulièrement espacées sur l'intervalle $[0,1]$. Le risque empirique est donc

$$\hat{\mathcal{R}}_n(f) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(\tfrac{i}{n}))^2.$$

Soit $\mathcal{F}_m$ l'ensemble des fonctions en escalier :

$$\mathcal{F}_m := \Big\{ f \in \mathbb{R}^{[0,1]} \mid f(x) = \sum_{j=1}^{m} c_j \, \mathbb{1}_{\{x \in I_{j,m}\}} \Big\} \quad \text{avec} \quad I_{j,m} = \big[\tfrac{j-1}{m}, \tfrac{j}{m}\big).$$

On notera $D_j = \big\{\frac{1}{n}, \dots, \frac{n-1}{n}, 1\big\} \cap I_{j,m}$ et $N_j = |D_j|$. On supposera que $n > m$ et on pourra utiliser que, par construction, $\frac{n}{m} - 1 \leqslant N_j < \frac{n}{m} + 1$.

On considère le minimiseur du risque empirique :

$$\hat{f} = \sum_{j=1}^{m} \hat{c}_j \mathbb{1}_{\{x \in I_{j,m}\}} = \operatorname{argmin}_{f \in \mathcal{F}_m} \hat{\mathcal{R}}_n(f).$$

(a) Explicitez le minimiseur du risque empirique $\hat{f}$.

(b) Soit $\bar{f}(x) = \sum_{j=1}^{m} \bar{c}_j \mathbb{1}_{\{x \in I_{j,m}\}} = \mathbb{E}[\hat{f}(x)]$. Montrez que $\mathcal{E}(\bar{f}) \leqslant \big(\frac{L}{m}\big)^2$.

(c) Proposez une borne supérieure à $\int_0^1 \operatorname{Var}(\hat{f}(x)) dx$ dépendant de σ^2, m et n.

(d) En déduire une borne supérieure pour l'espérance de l'excès de risque $\mathbb{E}[\mathcal{E}(\hat{f})]$. En supposant $n \gg \frac{L}{\sigma}$, quelle valeur m choisir pour minimiser approximativement cette borne supérieure ?

(e) On suppose désormais que $|\varepsilon_i| \leqslant b^2$ pour tout i. Montrer qu'avec probabilité au moins $1 - \frac{1}{2m}$, on a

$$\max_j |\hat{c}_j - \bar{c}_j| \leqslant 2b \sqrt{\frac{m \log(2m)}{n-m}}.$$

(f) En déduire que pour $n \geqslant m + m^3 \log(2m)$, on a avec probabilité $1 - \frac{1}{2m}$, que

$$\forall x \in [0,1], \qquad |\hat{f}(x) - \bar{f}(x)| \leqslant \frac{\kappa(b,L)}{m},$$

pour $\kappa(b,L)$ une constante dépendant de b et L. A n fixé, quel est le meilleur choix de m pour optimiser la convergence uniforme de $\hat{f}$ vers f^* ?