

Inférence Adaptative, Inductive et Transductive, pour l'Estimation de la Régression et de la Densité

Pierre Alquier

Soutenance de thèse, Université Paris 6

08/12/2006

Présentation et notations

Régression: Contextes Inductif et Transductif
Modèle, fonction de perte et risques
Estimation de la densité

Régression PAC-Bayésienne

Présentation de la méthode PAC-Bayésienne
Principaux résultats
Algorithmes

Sélection itérative de variables pour les problèmes d'estimation avec perte quadratique

Contexte général
Bornes PAC et intervalles de confiance
Algorithme de sélection
Résultats asymptotiques

Présentation et notations

Régression: Contextes Inductif et Transductif
Modèle, fonction de perte et risques
Estimation de la densité

Régression PAC-Bayésienne

Présentation de la méthode PAC-Bayésienne
Principaux résultats
Algorithmes

Sélection itérative de variables pour les problèmes d'estimation avec perte quadratique

Contexte général
Bornes PAC et intervalles de confiance
Algorithme de sélection
Résultats asymptotiques

Soient $(\mathcal{X}, \mathcal{B})$ un espace mesurable et P une loi sur $(\mathcal{X} \times \mathbb{R}, \mathcal{B} \otimes \mathcal{B}_{\mathbb{R}})$, et $N \in \mathbb{N}$.

Regression dans le contexte "transductif" (Vapnik)

Soit $k \in \mathbb{N}^*$. Un échantillon:

$$\left((X_1, Y_1), \dots, (X_{(k+1)N}, Y_{(k+1)N}) \right)$$

est tiré suivant la loi $P_{(k+1)N}$ loi échangeable. On observe "l'échantillon d'apprentissage" (éch. 1) et les X_i sur "l'échantillon de test" (2):

$$(X_1, Y_1), \dots, (X_N, Y_N) \quad \text{et} \quad X_{N+1}, \dots, X_{(k+1)N},$$

et on veut estimer:

$$Y_{N+1}, \dots, Y_{(k+1)N}.$$

Rmq: cas particulier $k = 1$.

Regression dans le contexte "inductif"

On suppose que $P(|X|) < +\infty$.

Un échantillon:

$$\left((X_1, Y_1), \dots, (X_N, Y_N) \right)$$

est tiré suivant la loi $P_N = P^{\otimes N}$. On observe cet échantillon et on veut estimer la fonction:

$$f : \mathcal{X} \rightarrow \mathbb{R}$$

$$x \mapsto f(x) = P(Y|X = x).$$

Modèle

Soit un ensemble de fonctions:

$$\mathcal{M} = \left\{ f_{\theta} : \mathcal{X} \rightarrow \mathbb{R}; \quad \theta \in \Theta \right\}.$$

Idée: approcher f par une fonction f_{θ} en inductif.

En transductif: approcher $(Y_{N+1}, \dots, Y_{(k+1)N})$ par

$$\left(f_{\theta}(X_{N+1}), \dots, f_{\theta}(X_{(k+1)N}) \right).$$

Fonction de perte et risque

Soit ψ une fonction symétrique $\mathbb{R}^2 \rightarrow \mathbb{R}_+$ (fonction de perte).

$$r_1(\psi, \theta) = \frac{1}{N} \sum_{i=1}^N \psi \left[Y_i, f_\theta(X_i) \right] \quad \text{observable,}$$

$$r_2(\psi, \theta) = \frac{1}{kN} \sum_{i=N+1}^{(k+1)N} \psi \left[Y_i, f_\theta(X_i) \right] \quad \text{non observable,}$$

$$R(\psi, \theta) = P \left\{ \psi \left[Y, f_\theta(X) \right] \right\} \quad \text{non observable.}$$

Estimation de la densité

On observe seulement (X_1, \dots, X_N) de loi $P^{\otimes N}$ sur $(\mathcal{X}, \mathcal{B})$, on suppose qu'il y a une mesure connue μ telle que $P \ll \mu$ et:

$$\frac{dP}{d\mu} = g$$

et on veut estimer g .

Présentation et notations

Régression: Contextes Inductif et Transductif

Modèle, fonction de perte et risques

Estimation de la densité

Régression PAC-Bayésienne

Présentation de la méthode PAC-Bayésienne

Principaux résultats

Algorithmes

Sélection itérative de variables pour les problèmes d'estimation
avec perte quadratique

Contexte général

Bornes PAC et intervalles de confiance

Algorithme de sélection

Résultats asymptotiques

Changements de variables

Idée de base: appliquer une inégalité de concentration sur:

$$\psi[f_\theta(X_i), Y_i]$$

(Hoeffding, Bernstein, ...).

Changement de variable si ψ bornée par C :

$$\Phi_{\lambda/N}(\psi[f_\theta(X_i), Y_i])$$

où $\lambda < (N/C)$, pour tout $\alpha \in \mathbb{R}_+^*$:

$$\Phi_\alpha(x) = \frac{-\log(1 - \alpha x)}{\alpha}.$$

Résultat de base

Alors, pour θ fixé:

$$\begin{aligned} P^{\otimes N} \exp \left\{ \lambda \Phi_{\frac{\lambda}{N}} [R(\psi, \theta)] - \frac{\lambda}{N} \sum_{i=1}^N \Phi_{\frac{\lambda}{N}} (\psi_i(\theta)) \right\} \\ = \prod_{i=1}^N P \left[\frac{1 - \frac{\lambda}{N} \psi_i(\theta)}{1 - \frac{\lambda}{N} R(\psi, \theta)} \right] = 1. \end{aligned}$$

Conséquence sur le risque:

Fixons $\theta \in \Theta$. Pour tout $\varepsilon > 0$, pour tout $\lambda \leq N/(2C)$, avec P_N -probabilité au moins $1 - \varepsilon$:

$$\begin{aligned} R(\psi, \theta) &\leq \Phi_{\frac{\lambda}{N}}^{-1} \left\{ r_1 \left(\Phi_{\frac{\lambda}{N}} \circ \psi, \theta \right) + \frac{\log \frac{1}{\varepsilon}}{\lambda} \right\} \\ &\leq r_1(\psi, \theta) + \frac{C^2 \lambda}{2N} + \frac{\log \frac{1}{\varepsilon}}{\lambda}. \end{aligned}$$

Pour λ optimal on obtient:

$$R(\psi, \theta) \leq r_1(\psi, \theta) + C \sqrt{\frac{2 \log \frac{1}{\varepsilon}}{N}}.$$

Nouveau changement de variable

Si ψ n'est pas bornée, on utilise le même changement de variable avec un seuillage:

$$\Phi_{\lambda/N} \left(\psi[f_{\theta}(X_i), Y_i] \wedge \frac{N}{\lambda} \right).$$

On obtient une borne sur:

$$R \left(\psi \wedge \frac{N}{\lambda}, \theta \right).$$

Pour borner:

$$R(\psi, \theta) - R \left(\psi \wedge \frac{N}{\lambda}, \theta \right),$$

formuler des hypothèses de moments sur ψ .

Principe de la méthode PAC-Bayésienne (McAllester, Catoni)

Contrôler (en espérance, ou avec grande probabilité):

$$\rho[R(\psi, \cdot)] = \int_{\Theta} R(\psi, \theta) d\rho(\theta), \quad \text{ou} \quad \rho[r_2(\psi, \cdot)],$$

pour n'importe quelle mesure $\rho \in \mathcal{M}_+^1(\Theta)$.

Le contrôle fera intervenir le terme empirique:

$$\rho[r_1(\psi, \cdot)]$$

et un "terme de complexité" qui mesure la distance entre ρ et une mesure fixée "a priori", $\pi \in \mathcal{M}_+^1(\Theta)$.

Divergence de Kullback

Définition:

$$\mathcal{K}(\rho, \pi) = \begin{cases} \rho \left[\log \left(\frac{d\rho}{d\pi} \right) \right] & \text{si } \rho \ll \pi, \\ +\infty & \text{sinon.} \end{cases}$$

Présentation et notations

Régression: Contextes Inductif et Transductif

Modèle, fonction de perte et risques

Estimation de la densité

Régression PAC-Bayésienne

Présentation de la méthode PAC-Bayésienne

Principaux résultats

Algorithmes

Sélection itérative de variables pour les problèmes d'estimation
avec perte quadratique

Contexte général

Bornes PAC et intervalles de confiance

Algorithme de sélection

Résultats asymptotiques

Un premier théorème

Pour tout $\varepsilon > 0$, pour tout $\lambda \in \mathbb{R}_+^*$, avec P_N -probabilité au moins $1 - \varepsilon$, pour toute mesure de probabilité $\rho \in \mathcal{M}_+^1(\Theta)$:

$$\begin{aligned} & \rho \left[R \left(\psi \wedge \frac{N}{\lambda}, \cdot \right) \right] \\ & \leq \Phi_{\frac{\lambda}{N}}^{-1} \left\{ \rho \left\{ r_1 \left[\Phi_{\frac{\lambda}{N}} \circ \left(\psi \wedge \frac{N}{\lambda} \right), \cdot \right] \right\} \right. \\ & \quad \left. + \frac{\mathcal{K}(\rho, \pi) + \log \frac{1}{\varepsilon}}{\lambda} \right\}, \end{aligned}$$

Lien avec la minimisation du risque empirique ou structurel

Supposons que $\psi(.,.) \leq C$. Le théorème devient: pour tout $\lambda \leq N/(2C)$, avec P_N -probabilité au moins $1 - \varepsilon$, pour toute mesure de probabilité $\rho \in \mathcal{M}_+^1(\Theta)$:

$$\rho[R(\psi, .)] \leq \rho[r_1(\psi, .)] + \frac{C^2 \lambda}{2N} + \frac{\mathcal{K}(\rho, \pi) + \log \frac{1}{\varepsilon}}{\lambda}.$$

Estimateur:

$$\hat{\rho} = \arg \min_{\rho} \left\{ \rho[r_1(\psi, .)] + \frac{\mathcal{K}(\rho, \pi)}{\lambda} \right\}.$$

En fait $\hat{\rho}$ est connu sous forme explicite (Catoni 2003), c'est une loi de Gibbs:

$$\frac{d\hat{\rho}}{d\pi}(\theta) = \frac{\exp[-\lambda r_1(\psi, \theta)]}{\pi \{\exp[-\lambda r_1(\psi, \cdot)]\}},$$

on notera:

$$\hat{\rho} = \pi_{\exp[-\lambda r_1(\psi, \cdot)]}.$$

Exemple: Θ fini

Avec $\pi = \frac{1}{|\Theta|} \sum_{\theta \in \Theta} \delta_{\theta}$, ρ restreint à l'ensemble des δ_{θ} et $\lambda = (1/C) \sqrt{2N \log(|\Theta|/\varepsilon)}$ le théorème devient: avec P_N -probabilité au moins $1 - \varepsilon$, pour tout $\theta \in \Theta$:

$$R(\psi, \theta) \leq r_1(\psi, \theta) + C \sqrt{\frac{2 \log \frac{|\Theta|}{\varepsilon}}{N}}.$$

Donc minimisation du risque empirique:

$$\hat{\theta} = \arg \min_{\theta \in \Theta} r_1(\psi, \theta).$$

Choix de λ en général

On peut faire une borne d'union sur plusieurs valeurs de λ et puis optimiser en λ . Soit m une mesure sur \mathbb{R}_+^* à support fini. Pour tout $\varepsilon > 0$, avec P_N -probabilité au moins $1 - \varepsilon$, pour toute mesure de probabilité $\rho \in \mathcal{M}_+^1(\Theta)$, pour tout $\lambda \in \mathbb{R}_+^*$:

$$\begin{aligned} & \rho \left[R \left(\psi \wedge \frac{N}{\lambda}, \cdot \right) \right] \\ & \leq \Phi_{\frac{\lambda}{N}}^{-1} \left\{ \rho \left\{ r_1 \left[\Phi_{\frac{\lambda}{N}} \circ \left(\psi \wedge \frac{N}{\lambda} \right), \cdot \right] \right\} \right. \\ & \quad \left. + \frac{\mathcal{K}(\rho, \pi) + \log \frac{m(\lambda)}{\varepsilon}}{\lambda} \right\}, \end{aligned}$$

Hypothèse de moment

Nouveau changement de variable:

$$\Phi_{\lambda/N} \left\{ \alpha \left[\psi[f_{\theta}(X_i), Y_i] - \frac{1}{s} \left(\frac{(s-1)\lambda}{sN} \right)^{s-1} |\psi|^s[f_{\theta}(X_i), Y_i] \right] \right\}$$

Supposons que pour un certain $s > 1$, pour $\theta \in \Theta$, $R(|\psi|^s, \theta) < +\infty$. Alors pour tout $\alpha \in (0, 1)$, pour tout $\varepsilon > 0$, pour tout $\lambda \in \mathbb{R}_+^*$, avec P_N -probabilité au moins $1 - \varepsilon$, pour toute mesure $\rho \in \mathcal{M}_+^1(\Theta)$:

$$\begin{aligned} & \rho[R(\psi, \cdot)] \\ & \leq \frac{1}{\alpha} \Phi_{\frac{\lambda}{N}}^{-1} \left\{ \rho r_1 \left[\Phi_{\frac{\lambda}{N}} \circ (\alpha\psi), \cdot \right] + \frac{\mathcal{K}(\rho, \pi) + \log \frac{1}{\varepsilon}}{\lambda} \right\} \\ & \quad + \frac{1}{s} \left(\frac{(s-1)\lambda}{sN} \right)^{s-1} \rho[R(|\psi|^s, \cdot)]. \end{aligned}$$

Bornes relatives

Soit $\mu \in \mathcal{M}_+^1(\Theta^2)$. Supposons $\psi(.,.) \leq C$. Pour tout $\varepsilon > 0$, pour tout $\lambda \leq N/C$, avec P_N -probabilité au moins $1 - \varepsilon$, pour tout $\nu \in \mathcal{M}_+^1(\Theta^2)$:

$$\begin{aligned} & \nu \left[R(\psi, \theta) - R(\psi, \theta') \right] \\ & \leq \Phi_{\frac{\lambda}{N}}^{-1} \left\{ \nu \left\{ \frac{1}{N} \sum_{i=1}^N \Phi_{\frac{\lambda}{N}} \left[\psi[f_\theta(X_i, Y_i)] - \psi[f_{\theta'}(X_i, Y_i)] \right] \right\} \right. \\ & \quad \left. + \frac{\mathcal{K}(\nu, \mu) + \log \frac{1}{\varepsilon}}{\lambda} \right\}. \end{aligned}$$

Calculs en espérance dans le cas linéaire

Modèle linéaire, borné, de dimension d , perte quadratique.
Soit:

$$\bar{\theta} = \arg \min_{\theta \in \Theta} R(\psi, \theta).$$

- ▶ Le premier théorème (avec un certain choix pour le paramètre λ) conduit à une borne en:

$$P^{\otimes N} \left\{ \hat{\rho}[R(\psi, \cdot)] \right\} \leq R(\psi, \bar{\theta}) + c \cdot \sqrt{\frac{d}{N} \log N}.$$

- ▶ Les bornes relatives conduisent à une borne en:

$$P^{\otimes N} \left\{ \hat{\rho}[R(\psi, \cdot)] \right\} \leq R(\psi, \bar{\theta}) + c' \cdot \frac{d \log N}{N}.$$

- ▶ En remplaçant π par la loi de Gibbs $\pi_{\exp[-\beta R(\psi, \cdot)]}$ (pour un certain $\beta \in \mathbb{R}_+^*$) on obtient la vitesse de convergence optimale:

$$P^{\otimes N} \left\{ \hat{\rho}[R(\psi, \cdot)] \right\} \leq R(\psi, \bar{\theta}) + c'' \cdot \frac{d}{N}.$$

Technique dite "localisation".

Localisation (Catoni 2003)

Remarquons que, pour toute loi π déterministe:

$$P_N \left[\mathcal{K}(\rho, \pi) \right] = P_N \left[\mathcal{K}(\rho, P_N(\rho)) \right] + \mathcal{K}(P_N(\rho), \pi).$$

Puisque le ρ optimal de la forme (pour un $\beta > 0$):

$$\pi_{\exp[-\beta r_1(\psi, \cdot)]}$$

incitation à remplacer π par:

$$\pi_{\exp[-\beta R(\psi, \cdot)]}.$$

PROBLEME:

$$\mathcal{K} \left(\rho, \pi_{\exp[-\beta R(\psi, \cdot)]} \right)$$

non observable.

Notations

Posons:

$$\begin{aligned} & v_{\psi, \frac{\lambda}{N}}(\theta, \theta') \\ &= \frac{2N}{\lambda} \left\{ \Phi_{\frac{\lambda}{N}} \left[r_1(\psi, \theta) - r_1(\psi, \theta') \right] - \left[r_1(\psi, \theta) - r_1(\psi, \theta') \right] \right\}. \end{aligned}$$

Supposons que $\psi(., .) \leq C = 1$.

Majoration de la divergence de Kullback

Pour tout $\varepsilon > 0$, pour tout $(\gamma, \beta) \in (0, N)^2$ tel que $\beta < \gamma$, avec P_N -probabilité au moins $1 - \varepsilon$, pour tout $\rho \in \mathcal{M}_+^1(\Theta)$,

$$\begin{aligned} \mathcal{K}(\rho, \pi_{\exp[-\beta R(\psi, \cdot)]}) &\leq \left(1 - \frac{\beta}{\gamma}\right)^{-1} \left\{ \mathcal{K}(\rho, \pi_{\exp[-\beta r_1(\psi, \cdot)]}) \right. \\ &\quad \left. - \frac{\beta}{\gamma} \log \varepsilon + \log \left[\pi_{\exp[-\beta r(\psi, d\theta')]} \exp \left(\frac{\beta\gamma}{2N} \rho(d\theta) \mathbf{v}_{\psi, \frac{\lambda}{N}}(\theta, \theta') \right) \right] \right\} \\ &= BK_{\gamma, \beta, \varepsilon}(\rho, \pi). \end{aligned}$$

Localisation pour la comparaison de deux estimateurs

Choisissons deux lois a priori π^1 et π^2 .

Choisissons $\varepsilon > 0$, et $(\lambda, \beta_1, \beta_2, \gamma_1, \gamma_2) \in (0, N)^5$ tels que $\beta_1 < \gamma_1$ et $\beta_2 < \gamma_2$.

Avec P_N -probabilité au moins $1 - \varepsilon$, pour tout couple $(\rho^1, \rho^2) \in [\mathcal{M}_+^1(\Theta)]^2$:

$$\begin{aligned} & \Phi_{\frac{\lambda}{N}} \left\{ \rho^1 [R(\psi, \cdot)] - \rho^2 [R(\psi, \cdot)] \right\} \\ & \leq \rho^1 [r_1(\psi, \cdot)] - \rho^2 [r_1(\psi, \cdot)] + \frac{\lambda}{2N} \rho^1 \otimes \rho^2 [v_{\psi, \frac{\lambda}{N}}(\cdot, \cdot)] \\ & \quad + \frac{1}{\lambda} \left\{ \log \frac{3}{\varepsilon} + \sum_{i=1}^2 BK_{\gamma_i, \beta_i, \frac{\varepsilon}{3}}(\rho^i, \pi^i) \right\}. \end{aligned}$$

Comparaison à une distribution de Gibbs

Pour tout $\varepsilon > 0$, pour tout $(\lambda, \beta, \gamma) \in (0, N)^3$ tel que $\beta < \gamma < \lambda$, avec P_N -probabilité au moins $1 - \varepsilon$, pour tout $\rho \in \mathcal{M}_+^1(\Theta)$:

$$\begin{aligned} & \left(\lambda \Phi_{\frac{\lambda}{N}} - \beta Id \right) [\rho R(\psi, \cdot) - \pi_{\exp[-\beta R(\psi, \cdot)]} R(\psi, \cdot)] \\ & \leq (\lambda - \gamma) [\rho r_1(\psi, \cdot) - \pi_{\exp[-\gamma r_1(\psi, \cdot)]} r_1(\psi, \cdot)] \\ & + \frac{\beta(\lambda - \gamma)}{\lambda(\gamma - \beta)} \log \pi_{\exp[-\gamma r_1(\psi, d\theta')]} \exp \left[\frac{\lambda^2}{2N} \pi_{\exp[-\gamma r_1(\psi, d\theta)]} \mathbf{v}_{\psi, \frac{\lambda}{N}}(\theta, \theta') \right] \\ & + \log \pi_{\exp[-\gamma r_1(\psi, d\theta')]} \exp \left[\frac{\lambda^2}{2N} \rho(d\theta) \mathbf{v}_{\psi, \frac{\lambda}{N}}(\theta, \theta') \right] + \mathcal{K}(\rho, \pi_{\exp[-\gamma r_1(\psi, \cdot)]}) \\ & \qquad \qquad \qquad + \log \frac{1}{\varepsilon}. \end{aligned}$$

Présentation et notations

Régression: Contextes Inductif et Transductif

Modèle, fonction de perte et risques

Estimation de la densité

Régression PAC-Bayésienne

Présentation de la méthode PAC-Bayésienne

Principaux résultats

Algorithmes

Sélection itérative de variables pour les problèmes d'estimation
avec perte quadratique

Contexte général

Bornes PAC et intervalles de confiance

Algorithme de sélection

Résultats asymptotiques

Comparaison à une distribution de Gibbs

Adaptation d'une proposition de Catoni dans le contexte de la classification.

La borne précédente permet de se référer à une échelle de comparaison non observable:

$$\forall \beta \in \Lambda, \quad \rho R(\psi, \cdot) \leq \pi_{\exp[-\beta R(\psi, \cdot)]} R(\psi, \cdot) + \mathcal{B}(\rho, \beta).$$

Rmq: $\beta \mapsto \pi_{\exp[-\beta R(\psi, \cdot)]} R(\psi, \cdot)$ décroît avec β vers $R(\psi, \bar{\theta})$.

Idée:

$$B(\zeta, \beta) = \mathcal{B}(\pi_{\exp[-\zeta r_1(\psi, \cdot)]}, \beta),$$

$$\hat{\beta}(\zeta) = \sup \{ \beta \in \Lambda : B(\zeta, \beta) \leq 0 \},$$

$$\hat{\zeta} \in \arg \max_{\zeta > 0} \hat{\beta}(\zeta).$$

Une borne sous-additive pour la comparaison

Soit $B(\rho_1, \rho_2)$ la borne pour la comparaison de ρ_1 et ρ_2 après borne d'union et optimisation par rapport aux différents paramètres:

$$P \left[\rho_1 R(\psi, \cdot) - \rho_2 R(\psi, \cdot) \leq B(\rho_1, \rho_2) \right] \geq 1 - \varepsilon.$$

Soient ρ^1, \dots, ρ^M des lois chargeant éventuellement des sous-modèles différents.

On définit:

$$\tilde{B}(\rho, \rho') = \inf \left\{ \sum_{k=1}^h B(\rho^{k-1}, \rho^k), \right. \\ \left. h \geq 1, (\rho^0, \dots, \rho^h) \in \mathcal{P}^{h+1}, \rho^0 = \rho, \rho^h = \rho' \right\}$$

Un ordre sur les estimateurs

On définit (par exemple à l'aide du terme de complexité dans le théorème précédent) une fonction qui mesure la "complexité" d'un estimateur:

$$\mathcal{C} : \mathcal{M}_+^1(\Theta) \rightarrow \mathbb{R}_+.$$

Quitte à changer l'indexation des ρ^k on suppose:

$$\mathcal{C}(\rho^1) \leq \dots \leq \mathcal{C}(\rho^M).$$

Algorithme

On pose, pour tout $k \in \{1, \dots, M\}$:

$$t(k) = \max \left\{ j \in \{1, \dots, M\}, \quad \forall \ell \in \{1, \dots, j\}, \tilde{B}(\rho^k, \rho^\ell) \leq 0 \right\}.$$

On sélectionne alors l'estimateur $\rho_{\hat{k}}$ donné par:

$$\hat{k} = \min(\arg \max t).$$

Présentation et notations

- Régression: Contextes Inductif et Transductif
- Modèle, fonction de perte et risques
- Estimation de la densité

Régression PAC-Bayésienne

- Présentation de la méthode PAC-Bayésienne
- Principaux résultats
- Algorithmes

Sélection itérative de variables pour les problèmes d'estimation avec perte quadratique

- Contexte général
- Bornes PAC et intervalles de confiance
- Algorithme de sélection
- Résultats asymptotiques

Fonction de perte quadratique

A partir de maintenant: $\psi(y, y') = (y - y')^2$. Pour simplifier on notera:

$$r_1(\theta) = r_1(\psi, \theta),$$

$$r_2(\theta) = r_2(\psi, \theta) \quad \text{en transductif,}$$

$$R(\theta) = R(\psi, \theta) \quad \text{en inductif.}$$

Rms: pour simplifier la présentation, $k = 1$, les deux échantillons sont de même taille.

Modèle linéaire

Le modèle \mathcal{M} est tel que:

- ▶ Θ un espace vectoriel;
- ▶ $\theta \mapsto f_\theta(\cdot)$ est une application linéaire.

Fonctions de base

Soit $m \in \mathbb{N}$ et $(\theta_1, \dots, \theta_m) \in \Theta^m$ avec éventuellement:

$$(\theta_1, \dots, \theta_m) = F\left(\left\{X_1, \dots, X_{2N}\right\}\right)$$

On va utiliser les "fonctions de base" $f_{\theta_1}, \dots, f_{\theta_m}$ pour définir:

$$\hat{f}(\cdot) = \sum_{i=1}^m \hat{\alpha}_i f_{\theta_i}(\cdot) = f_{\sum_{i=1}^m \hat{\alpha}_i \theta_i}(\cdot) = f_{\hat{\theta}}(\cdot).$$

Premier exemple

Estimation "par projections", Θ est un espace de fonctions, $f_{\theta}(x) = \theta(x)$ et $(\theta_1, \dots, \theta_m)$ début d'une base $(\theta_i, i \in \mathbb{N})$ de Θ (ne dépend pas des observations).

$$f_{\hat{\theta}}(.) = \sum_{i=1}^m \hat{\alpha}_i f_{\theta_i}(.).$$

Problème: choix de m ?

Deuxième exemple ("SVM")

Estimation à noyau, soit $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ un espace de Hilbert, $\Psi : \mathcal{X} \rightarrow \mathcal{H}$, $\Theta = \Psi(\mathcal{X})$, $\{\theta_1, \dots, \theta_m\} = \{\Psi(X_1), \dots, \Psi(X_{2N})\}$ (donc $m \leq 2N$) et $f_{\theta}(x) = \langle \theta, \Psi(x) \rangle$. Ici:

$$f_{\hat{\theta}}(\cdot) = \sum_{i=1}^{2N} \hat{\alpha}_i \langle \Psi(X_i), \Psi(x) \rangle = \sum_{i=1}^{2N} \hat{\alpha}_i K(X_i, x)$$

où K est le noyau de Mercer défini par:

$$K(x, x') = \langle \Psi(x), \Psi(x') \rangle.$$

Troisième exemple ("SVM à plusieurs noyaux")

On prend $\Theta = \mathbb{R}^{\mathcal{X}}$, avec $f_{\theta}(x) = \theta(x)$. On prend:

$$\{\theta_1, \dots, \theta_m\} = \{K_1(X_1, \cdot), \dots, K_1(X_{2N}, \cdot), \dots, K_J(X_1, \cdot), \dots, K_J(X_{2N}, \cdot)\}$$

avec $m = 2NJ$ et K_1, \dots, K_J fonctions $\mathcal{X}^2 \rightarrow \mathbb{R}$.

Rmq: les fonctions ainsi définies ne sont pas orthogonales et forment un système redondant.

Modèles de dimension 1

Soit, pour chaque $i \in \{1, \dots, m\}$:

$$\bar{\alpha}_i = \arg \min_{\alpha \in \mathbb{R}} r_2(\alpha \theta_i) = \frac{\sum_{j=N+1}^{2N} f_{\theta_i}(X_j) Y_j}{\sum_{j=N+1}^{2N} f_{\theta_i}(X_j)^2}$$

que l'on va estimer par:

$$\tilde{\alpha}_i = \frac{\sum_{j=1}^N f_{\theta_i}(X_j) Y_j}{\sum_{j=N+1}^{2N} f_{\theta_i}(X_j)^2}$$

Présentation et notations

- Régression: Contextes Inductif et Transductif
- Modèle, fonction de perte et risques
- Estimation de la densité

Régression PAC-Bayésienne

- Présentation de la méthode PAC-Bayésienne
- Principaux résultats
- Algorithmes

Sélection itérative de variables pour les problèmes d'estimation avec perte quadratique

- Contexte général
- Bornes PAC et intervalles de confiance
- Algorithme de sélection
- Résultats asymptotiques

Théorème 1

Pour tout $\varepsilon > 0$, avec $P^{\otimes 2N}$ -probabilité au moins $1 - \varepsilon$, pour tout $i \in \{1, \dots, m\}$ on a:

$$r_2(\tilde{\alpha}_i; \theta_i) - r_2(\bar{\alpha}_i; \theta_i) \leq 4 \left[\frac{\sum_{j=1}^{2N} f_{\theta_i}(X_j)^2 Y_j^2}{\sum_{j=N+1}^{2N} f_{\theta_i}(X_j)^2} \right] \frac{\log \frac{2m}{\varepsilon}}{N}.$$

Remarques:

Théorème 1

Pour tout $\varepsilon > 0$, avec $P^{\otimes 2N}$ -probabilité au moins $1 - \varepsilon$, pour tout $i \in \{1, \dots, m\}$ on a:

$$r_2(\tilde{\alpha}_i; \theta_i) - r_2(\bar{\alpha}_i; \theta_i) \leq 4 \left[\frac{\sum_{j=1}^{2N} f_{\theta_i}(X_j)^2 Y_j^2}{\sum_{j=N+1}^{2N} f_{\theta_i}(X_j)^2} \right] \frac{\log \frac{2m}{\varepsilon}}{N}.$$

Remarques:

- ▶ Vitesse en N^{-1} pour un modèle de dimension 1.

Théorème 1

Pour tout $\varepsilon > 0$, avec $P^{\otimes 2N}$ -probabilité au moins $1 - \varepsilon$, pour tout $i \in \{1, \dots, m\}$ on a:

$$r_2(\tilde{\alpha}_i; \theta_i) - r_2(\bar{\alpha}_i; \theta_i) \leq 4 \left[\frac{\sum_{j=1}^{2N} f_{\theta_i}(X_j)^2 Y_j^2}{\sum_{j=N+1}^{2N} f_{\theta_i}(X_j)^2} \right] \frac{\log \frac{2m}{\varepsilon}}{N}.$$

Remarques:

- ▶ Vitesse en N^{-1} pour un modèle de dimension 1.
- ▶ La borne n'est pas observable, or on aura besoin de la calculer explicitement.

Rendre la borne observable: hypothèses nécessaires

Rendre la borne observable: hypothèses nécessaires

- ▶ 1er exemple: $P(|Y| < B) = 1$. Alors la borne devient observable:

$$4 \left[\frac{\sum_{j=1}^{2N} f_{\theta_i}(X_j)^2 Y_j^2}{\sum_{j=N+1}^{2N} f_{\theta_i}(X_j)^2} \right] \frac{\log \frac{2m}{\varepsilon}}{N} \\ \leq 4 \left[\frac{\sum_{j=1}^N f_{\theta_i}(X_j)^2 Y_j^2}{\sum_{j=N+1}^{2N} f_{\theta_i}(X_j)^2} + B^2 \right] \frac{\log \frac{2m}{\varepsilon}}{N}.$$

Rendre la borne observable: hypothèses nécessaires

- ▶ 1er exemple: $P(|Y| < B) = 1$. Alors la borne devient observable:

$$4 \left[\frac{\sum_{j=1}^{2N} f_{\theta_i}(X_j)^2 Y_j^2}{\sum_{j=N+1}^{2N} f_{\theta_i}(X_j)^2} \right] \frac{\log \frac{2m}{\varepsilon}}{N} \\ \leq 4 \left[\frac{\sum_{j=1}^N f_{\theta_i}(X_j)^2 Y_j^2}{\sum_{j=N+1}^{2N} f_{\theta_i}(X_j)^2} + B^2 \right] \frac{\log \frac{2m}{\varepsilon}}{N}.$$

- ▶ Existence de moments exponentiels de Y .

Théorème 2

Supposons que $P[\exp(bY)] \leq B < +\infty$. Alors pour tout $\varepsilon > 0$, avec $P^{\otimes 2N}$ -probabilité au moins $1 - \varepsilon$, pour tout $i \in \{1, \dots, m\}$ on a:

$$r_2(\tilde{\alpha}_i; \theta_i) - r_2(\bar{\alpha}_i; \theta_i) \leq 8 \left[\frac{\sum_{j=1}^N f_{\theta_i}(X_j)^2 Y_j^2}{\sum_{j=N+1}^{2N} f_{\theta_i}(X_j)^2} \right] \frac{\log \frac{4m}{\varepsilon}}{N} \\ + \frac{8}{b^2} \left(\frac{\log \frac{8m}{\varepsilon}}{N} \right)^{\frac{3}{2}} \left(\log \frac{2NB}{\varepsilon} \right)^2 \sqrt{\frac{1}{2N} \sum_{j=1}^{2N} f_{\theta_i}(X_j)^4}.$$

Présentation et notations

- Régression: Contextes Inductif et Transductif
- Modèle, fonction de perte et risques
- Estimation de la densité

Régression PAC-Bayésienne

- Présentation de la méthode PAC-Bayésienne
- Principaux résultats
- Algorithmes

Sélection itérative de variables pour les problèmes d'estimation avec perte quadratique

- Contexte général
- Bornes PAC et intervalles de confiance
- Algorithme de sélection
- Résultats asymptotiques

Interprétation du théorème

- Pour chaque i , le théorème donne un intervalle de confiance sur $\bar{\alpha}_i$:

$$r_2(\tilde{\alpha}_i; \theta_i) - r_2(\bar{\alpha}_i; \theta_i) \leq \mathcal{B}_{i,\varepsilon}$$

Interprétation du théorème

- ▶ Pour chaque i , le théorème donne un intervalle de confiance sur $\bar{\alpha}_i$:

$$r_2(\tilde{\alpha}_i; \theta_i) - r_2(\bar{\alpha}_i; \theta_i) \leq \mathcal{B}_{i, \varepsilon}$$



$$(\tilde{\alpha}_i - \bar{\alpha}_i)^2 \left[\frac{1}{N} \sum_{j=N+1}^{2N} f_{\theta_i}(X_j)^2 \right] \leq \mathcal{B}_{i, \varepsilon}$$

$$\bar{\alpha}_i \in \mathcal{IC}(i, \varepsilon)$$

Interprétation du théorème

- ▶ Pour chaque i , le théorème donne un intervalle de confiance sur $\bar{\alpha}_i$:

$$r_2(\tilde{\alpha}_i; \theta_i) - r_2(\bar{\alpha}_i; \theta_i) \leq \mathcal{B}_{i, \varepsilon}$$



$$(\tilde{\alpha}_i - \bar{\alpha}_i)^2 \left[\frac{1}{N} \sum_{j=N+1}^{2N} f_{\theta_i}(X_j)^2 \right] \leq \mathcal{B}_{i, \varepsilon}$$

$$\bar{\alpha}_i \in \mathcal{IC}(i, \varepsilon)$$

- ▶ Soit le produit scalaire sur Θ :

$$\langle \theta, \theta' \rangle_2 = \frac{1}{N} \sum_{j=N+1}^{2N} f_{\theta}(X_j) f_{\theta'}(X_j).$$

- ▶ Or $\bar{\alpha}_i \theta_i$ est la projection orthogonale de $\bar{\theta}$ sur $\{\alpha \theta_i, \alpha \in \mathbb{R}\}$, par rapport à $\langle \cdot, \cdot \rangle_2$.

- ▶ Or $\bar{\alpha}_i \theta_i$ est la projection orthogonale de $\bar{\theta}$ sur $\{\alpha \theta_i, \alpha \in \mathbb{R}\}$, par rapport à $\langle \cdot, \cdot \rangle_2$.
- ▶ Donc le théorème affirme que, pour chaque i , $\bar{\theta}$ appartient à la région qui se projette sur le bon intervalle de confiance:

$$\bar{\theta} \in \mathcal{RC}(i, \varepsilon) = \left\{ \theta \in \Theta : \left(\frac{\langle \theta - \tilde{\alpha}_i \theta_i, \theta_i \rangle_2}{\|\theta_i\|_2} \right)^2 \leq \mathcal{B}_{i, \varepsilon} \right\}.$$

Intéressant car $\mathcal{RC}(i, \varepsilon)$ ensemble convexe fermé qui contient $\bar{\theta}$. Notons: $\Pi_{i, \varepsilon}$ la projection orthogonale sur $\mathcal{RC}(i, \varepsilon)$ par rapport à $\langle \cdot, \cdot \rangle_2$.

$$\forall \theta \in \Theta, \|\Pi_{i, \varepsilon} \theta - \bar{\theta}\|_2^2 \leq \|\theta - \bar{\theta}\|_2^2,$$

$$\forall \theta \in \Theta, r_2(\Pi_{i, \varepsilon} \theta) - r_2(\bar{\theta}) \leq r_2(\theta) - r_2(\bar{\theta}).$$

Ou: quel que soit l'estimateur θ , $\Pi_{i, \varepsilon} \theta$ est un meilleur estimateur.

Algorithme 1

▶ $\theta(0) = 0;$

Algorithme 1

- ▶ $\theta(0) = 0$;
- ▶ $\theta(n + 1) = \prod_{m,\varepsilon} \dots \prod_{1,\varepsilon} \theta(n)$;

Algorithme 1

- ▶ $\theta(0) = 0$;
- ▶ $\theta(n + 1) = \prod_{m,\varepsilon} \dots \prod_{1,\varepsilon} \theta(n)$;
- ▶ arrêt lorsqu'on a atteint un point fixe ou lorsqu'un certain critère est satisfait.

Quantification de l'amélioration

$$\forall \theta \in \Theta, \|\Pi_{i,\varepsilon}\theta - \bar{\theta}\|_2^2 \leq \|\theta - \bar{\theta}\|_2^2 - \|\Pi_{i,\varepsilon}\theta - \theta\|_2^2,$$

donc:

$$\forall \theta \in \Theta, r_2(\Pi_{i,\varepsilon}\theta) - r_2(\bar{\theta}) \leq r_2(\theta) - r_2(\bar{\theta}) - \|\Pi_{i,\varepsilon}\theta - \theta\|_2^2,$$

$$r_2(\Pi_{i,\varepsilon}\theta) \leq r_2(\theta) - \|\Pi_{i,\varepsilon}\theta - \theta\|_2^2.$$

Algorithmme 2

▶ $\theta(0) = 0;$

Algorithme 2

- ▶ $\theta(0) = 0$;
- ▶ $\theta(n + 1) = \Pi_{i(n), \varepsilon} \theta(n)$ où $i(n)$ maximise le critère d'amélioration: $\|\Pi_{i(n), \varepsilon} \theta(n) - \theta(n)\|_2^2$ (ou tout autre critère).

Algorithme 2

- ▶ $\theta(0) = 0$;
- ▶ $\theta(n + 1) = \Pi_{i(n), \varepsilon} \theta(n)$ où $i(n)$ maximise le critère d'amélioration: $\|\Pi_{i(n), \varepsilon} \theta(n) - \theta(n)\|_2^2$ (ou tout autre critère).
- ▶ arrêt lorsqu'on a atteint un point fixe ou lorsqu'un certain critère est satisfait.

Présentation et notations

- Régression: Contextes Inductif et Transductif
- Modèle, fonction de perte et risques
- Estimation de la densité

Régression PAC-Bayésienne

- Présentation de la méthode PAC-Bayésienne
- Principaux résultats
- Algorithmes

Sélection itérative de variables pour les problèmes d'estimation avec perte quadratique

- Contexte général
- Bornes PAC et intervalles de confiance
- Algorithme de sélection
- Résultats asymptotiques

Contexte "inductif"

Extension au cas inductif, le risque devient:

$$R(\theta) = P\left[\left(Y - f_{\theta}(X)\right)^2\right],$$

meilleures bornes mais nécessité de connaître la loi marginale de X : P_X , car le produit scalaire impliqué est:

$$\langle \theta, \theta' \rangle_P = P\left[f_{\theta}(X)f_{\theta'}(X)\right].$$

Hypothèse nécessaire: $P[(Y - f(x))^2|X = x] \leq \sigma^2 < +\infty$.

Estimation de la densité

On observe $(X_1, \dots, X_N) \in \mathcal{X}^N$ i.i.d. de distribution P qui a une densité par rapport à la mesure μ :

$$\frac{dP}{d\mu}(\cdot) = g(\cdot).$$

On essaie d'estimer g par les fonctions $\{f_\theta, \theta \in \Theta\}$. La même démarche reste valide en utilisant le produit scalaire:

$$\langle \theta, \theta' \rangle = \int_{\mathcal{Z}} f_\theta(z) f_{\theta'}(z) \mu(dz)$$

si la fonction de perte est:

$$L(\theta) = \int_{\mathcal{Z}} [f_\theta(x) - g(x)]^2 \mu(dz).$$

Autres problèmes

Autre problèmes avec perte quadratique:

- ▶ modèle de bruit blanc gaussien;
- ▶ modèle de régression avec design déterministe;
- ▶ ...

Vitesse de convergence pour le contexte inductif

Si $(f_{\theta_i})_{i=1}^{\infty}$ b. o. n. de $L^2([0, 1])$, $\mathcal{X} = [0, 1]$, et f de régularité β , l'algorithme est équivalent à une méthode de seuillage et permet de fabriquer un estimateur \hat{f} qui atteint la vitesse minimax (à un $\log N$ près).

Vitesse de convergence pour le contexte inductif

Si $(f_{\theta_i})_{i=1}^{\infty}$ b. o. n. de $L^2([0, 1])$, $\mathcal{X} = [0, 1]$, et f de régularité β , l'algorithme est équivalent à une méthode de seuillage et permet de fabriquer un estimateur \hat{f} qui atteint la vitesse minimax (à un $\log N$ près).

- ▶ Si f bornée, dans $W(\beta)$ espace de Sobolev et base de Fourier:

$$P^{\otimes N} \left\{ P \left[\left(\hat{f}(X) - f(X) \right)^2 \right] \right\} \leq C(f) \left(\frac{\log N}{N} \right)^{\frac{2\beta}{2\beta+1}}$$

Vitesse de convergence pour le contexte inductif

Si $(f_{\theta_i})_{i=1}^{\infty}$ b. o. n. de $L^2([0, 1])$, $\mathcal{X} = [0, 1]$, et f de régularité β , l'algorithme est équivalent à une méthode de seuillage et permet de fabriquer un estimateur \hat{f} qui atteint la vitesse minimax (à un $\log N$ près).

- ▶ Si f bornée, dans $W(\beta)$ espace de Sobolev et base de Fourier:

$$P^{\otimes N} \left\{ P \left[\left(\hat{f}(X) - f(X) \right)^2 \right] \right\} \leq C(f) \left(\frac{\log N}{N} \right)^{\frac{2\beta}{2\beta+1}}$$

- ▶ Si f bornée, dans $B_{s,p,+\infty}$ espace de Besov avec $(1/p) < s \leq R + 1$ et base d'ondelettes de régularité R , vitesse en:

$$C'(f) \left(\frac{\log N}{N} \right)^{\frac{2s}{2s+1}} \log N.$$

Présentation et notations

Régression: Contextes Inductif et Transductif
Modèle, fonction de perte et risques
Estimation de la densité

Régression PAC-Bayésienne

Présentation de la méthode PAC-Bayésienne
Principaux résultats
Algorithmes

Sélection itérative de variables pour les problèmes d'estimation avec perte quadratique

Contexte général
Bornes PAC et intervalles de confiance
Algorithme de sélection
Résultats asymptotiques