Apprentissage statistique et grande dimension

M. Hoffmann. M1 Paris Dauphine 2014-2015

Chapitre 1 : Expérience statistique et grande dimension (Séances 1-2)

1. Exemples de modèles statistiques non-paramétriques
   - Estimation d'une densité de probabilité (cadre paramétrique, non-paramétrique)
   - Régression non-paramétrique (cadre paramétrique, cadre grande dimension)
   - Exemples de problèmes statistiques en grande dimension.
2. L'heuristique du modèle de suite gaussienne, notion d'expériences statistiques (asymptotiquement) équivalentes.
   - Le cas du modèle de densité
   - La cas du modèle de régression
   - Définition du modèle de suite gaussienne

Chapitre 2 : Estimateurs linéaires et régression non-paramétrique (Séances 2-3-4)

1. Situation : modèle de régression, définition : on observe

   $$Y_i = f(X_i) + \xi_i, \dots, i = 1, \dots, n$$

   où $f : [0,1] \to \mathbb{R}$ est la fonction inconnue, les X_i sont le design i.i.d. et les ξ_i sont des bruits i.i.d. de variance finie et centrés.

   Notation vectorielle : $\boldsymbol{y} = f + \boldsymbol{\xi}$. Lorsqu'il n'y a pas (trop) d'ambiguité, on identifie la fonction f et le vecteur colonne $\big(f(X_1) \cdots f(X_n)\big)^T$. Notion de dictionnaire de taille M et modèle paramétrique associé. Définition de l'estimateur des moindres carrés non-paramétrique. L'EMC non-paramétrique est un estimateur linéaire si $\boldsymbol{X}^T\boldsymbol{X} > 0$. Le terme $\boldsymbol{\xi}$ est un bruit blanc fort de variance σ^2.
2. Propriétés statistiques des estimateurs linéaires en régression. Matrice de lissage S. Risque quadratique

   $$\mathcal{R}(\widehat{f}, f) = \mathbb{E}[\|\widehat{f} - f\|_n^2]$$

où, pour $g : [0,1] \to \mathbb{R}$, on note $\|g\|_n^2 = n^{-1} \sum_{i=1}^n g(X_i)^2$ la semi-norme L^2-empirique associée au design $\{X_i, i = 1, \ldots, n\}$. Décomposition biais-variance du risque quadratique des estimateurs linéaires : si $\widehat{f} = S\boldsymbol{y}$, alors

$$\mathcal{R}(\widehat{f}, f) = \mathbb{E}[\|Sf - f\|_n^2] + \frac{\sigma^2}{n} \mathrm{Trace}(S^T S).$$

Si l'on travaille conditionnellement au design $\{X_i = x_i, i = 1, \ldots, n\}$ (par exemple pour le design déterministe uniforme $X_i = i/n$) alors la décomposition précédente devient

$$\mathcal{R}(\widehat{f}, f) = \|Sf - f\|_n^2 + \tfrac{\sigma^2}{n} \mathrm{Trace}(S^T S).$$

Corollaire : si $\widehat{f}^{\mathrm{MC}}$ désigne l'EMC non-paramétrique, alors, conditionnellement au design $\{X_i = x_i, i = 1, \ldots, n\}$, on a

$$\mathcal{R}(\widehat{f}^{\mathrm{MC}}, f) \leq \min_{\theta \in \mathbb{R}^M} \|f_\theta - f\|_n^2 + \frac{\sigma^2}{n} M \wedge n, \qquad (\star)$$

où f_θ est la paramétrisation de f via un dictionnaire fini.

3. Vitesses de convergence de l'EMC non-paramétrique. Définition des espaces (boules) de Sobolev $W(\beta, L)$ périodiques en dimension 1, traduction séquentielle via la base de Fourier. Si $f \in W(\beta, L)$, alors

$$\inf_{\theta \in \mathbb{R}^M} \mathbb{E}[\|f_\theta - f\|_n^2] \leq L^2 M^{-2\beta}$$

pour un design aléatoire uniforme. Extension pour le risque conditionnel au design déterministe uniforme $X_i = i/n$. Bornes supérieures d'estimation : pour un choix de taille de dictionnaire $M_n \approx n^{1/(2\beta+1)}$ dans la base de Fourier, l'EMC non-paramétrique atteint la vitesse $n^{-2\beta/(2\beta+1)}$ pour le risque $\mathcal{R}$. Optimalité et notion de vitesse minimax.

Chapitre 3 : Représentation parcimonieuse (sparse) en régression (Séances 5-6)

1. Situation. Ecriture séquentielle de $(\star)$:

$$\mathbb{E}[n^{-1} \|X(\theta^{\mathrm{MC}} - \theta^\star)\|^2] \leq \sigma^2 \frac{M \wedge n}{n}$$

si $f = f_{\theta^\star}$ est représentés sans erreur via un dictionnaire de taille M. Problématique M "grand". Notion de vecteur ε-sparse (le nombre $M(\theta)$ de composantes non-nulles est majoré par εM). Recherche d'un estimateur $\widehat{\theta}$ de sorte que

$$\mathbb{E}[n^{-1} \|X(\widetilde{\theta} - \theta^\star)\|^2] \leq \sigma^2 \frac{M(\theta^\star)}{n} \log M. \qquad (\star\star)$$

2. Seuillage dans le modèle de suite gaussienne. Hypothèse restrictive ORT: $n^{-1}\boldsymbol{X}^T\boldsymbol{X} = \mathrm{Id}_{\mathbb{R}^n}$. Lemme de concentration gaussienne :

$$\mathbb{P}\big(\max_{1\leq j\leq M}|\eta_j| > t\sqrt{\log M}\big) \leq M^{1-t^2/2}$$

si $M \geq 2$, $t \geq \sqrt{2}$ et les η_j sont des gaussiennes standard. Notion de "hard-thresholding" et interprétation (à la Hodge-Lehmann). Théorème : l'estimateur par seuillage réalise (dans le modèle de suite gaussienne en ORT) le programme $(\star\star)$ en espérance et en probabilité. Il peut de plus estimer l'ensemble d'indices révélateurs $J(\theta) = \{j, \theta_j \neq 0\}$ sous une condition apparentée à la séparation d'hypothèses en théorie des tests.

3. Remarques et développements : soft-thresholding, formulations variationelles exactes dans le cadre ORT. Au delà, lien entre "hard-thresholding" et estimateur BIC, et "soft-thresholding" et estimateur LASSO. Extension du cadre ORT : hypothèse de cohérence.

Chapitre 4 : Classification et apprentissage statistique (séances 7-8-9)

1. Exemples de problèmes de classification

2. Formulation non-paramétrique de la classification (binaire): notion de classifieur $h : \mathcal{X} \to \{0, 1\}$ où $\mathcal{X}$ désigne l'espace des covariables, erreur/risque de classification $R(h)$.

3. Le classifieur bayésien. Le classifieur $h^\star(x) = \mathbf{1}_{\{\eta(x)>1/2\}}$ minimise l'erreur de classification, où $\eta(x) = \mathbb{P}(h(X) \neq Y)$. Pour tout classifieur h, on a

$$R(h) - R(h^\star) = \int_{x,h(x)\neq h^\star(x)} |2\eta(x) - 1|\mathbb{P}_X(dx)$$

où $\mathbb{P}_X$ désigne la loi du design. Limites de l'estimation directe du classifieur bayésien par estimation de la fonction de régression η.

4. Apprentissage et minimisation du risque empirique. Notion d'excès de risque.

   - Minimisation du risque empirique $R_n(h) = n^{-1}\sum_{i=1}^n \mathbf{1}_{\{Y_i\neq h(X_i)\}}$ où

   $$\mathcal{D}_n = \{(X_i, Y_i), i = 1, \ldots, n\}$$

   est un échantillon d'apprentissage. Pour un dictionnaire $\mathcal{H}$ défini *a priori*, le classifieur minimisant le risque empirique est défini comme $\widehat{h}_n^{\mathrm{erm}}$ vérifiant $R_n(\widehat{h}_n^{\mathrm{erm}}) = \min_{h\in\mathcal{H}} R_n(h)$. Notion de sur-apprentissage et de décomposition erreur stochastique + erreur d'approximation.

- Contrôle de l'erreur stochastique

$$R_{\mathcal{D}_n}(\widehat{h}_n^{\text{erm}}) - \inf_{h\in\mathcal{H}} R(h) \leq 2 \sup_{h\in\mathcal{H}} |R_n(h) - R(h)|$$

où $R_{\mathcal{D}_n}(h) = \mathbb{P}(h(X, \mathcal{D}_n) \neq Y | \mathcal{D}_n)$ est l'erreur de classification conditionnelle à l'échantillon d'apprentissage pour un classifieur $h(x) = h(x, \mathcal{D}_n)$ construit à l'aide d'un échantillon d'apprentissage $\mathcal{D}_n$.

- Classification pour un dictionnaire fini. Si $\mathcal{H}$ est fini et de taille M, alors

$$R_{\mathcal{D}_n}(\widehat{h}_n^{\text{erm}}) \leq \min_{h\in\mathcal{H}} R(h) + \sqrt{\frac{2}{n} \log \frac{2M}{\delta}}$$

avec probabilité plus grande que $1 - \delta$, et

$$\mathbb{E}\big[R_{\mathcal{D}_n}(\widehat{h}_n^{\text{erm}})\big] \leq \min_{h\in\mathcal{H}} R(h) + \sqrt{\frac{\log 2M}{2n}}.$$

Preuves via Chernoff.

Chapitre 5. Introduction à la théorie de Vapnik-Chervonenkis (séance 10-11)

1. Si $\mu_n = n^{-1} \sum_{i=1}^n \delta_{X_i}$ et $\mu = \mathbb{P}(X \in \dot{})$, alors, en identifiant un classifieur $h = \mathbf{1}_A$ avec un ensemble mesurable $A \subset \mathcal{X}$, on a

$$\sup_{h\in\mathcal{X}} |R_n(h) - R(h)| = \sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)|.$$

Symétrisation du risque :

$$\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)| = \sup_{A\in\mathcal{A}} |\mu_n(A) - \mu_n'(A)|$$

où μ_n' est obtenue via une copie indépendante de $(X_1, \ldots, X_n)$.

2. Complexité de Rademacher : pour $B \subset \mathbb{R}^n$ borné, elle est définie comme

$$\mathcal{R}_n(B) = \mathbb{E}\big[\sup_{(b_1,\ldots,b_n)\in B} \big|n^{-1} \sum_{i=1}^n \sigma_i b_i\big|\big],$$

où les σ_i sont des variables de Rademacher indépendantes :

$$\mathbb{P}(\sigma_i = 1) = 1 - \mathbb{P}(\sigma_i = -1) = \tfrac{1}{2}.$$

Notion d'emprunte binaire d'un vecteur $(x_1, \ldots, x_n) \in \mathcal{X}^n$ sur une classe $\mathcal{A}$ d'ensembles mesurables de $\mathcal{X}$:

$$\mathcal{A}(x_1, \ldots, x_n) = \big\{(b_1, \ldots, b_n) \in \{0,1\}^n, b_i = \mathbf{1}_{\{x_i \in A\}}, A \in \mathcal{A}\big\}.$$

Alors

$$\begin{aligned}\mathbb{E}\big[\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu'_n(A)|\big] &\leq 2\mathbb{E}\big[\sup_{A\in\mathcal{A}} n^{-1}\big|\sum_{i=1}^n \sigma_i \mathbf{1}_{\{X_i\in A\}}\big|\big] \\ &\leq 2\mathbb{E}\big[\mathcal{R}_n\big(\mathcal{A}(X_1, \ldots, X_n)\big)\big].\end{aligned}$$

3. Contrôle de la complexité de Rademacher. Si B est un sous-ensemble fini de $\mathbb{R}^n$, alors
$$\mathcal{R}_n(B) \leq \max_{b\in B} \|b\|_{\ell^2} \frac{\sqrt{2\log(2\,\mathrm{Card} B)}}{n}$$
et conclusion :
$$\mathbb{E}\big[\sup_{A\in\mathcal{A}} \big|\mu_n(A) - \mu(A)\big|\big] \leq 2\sqrt{\frac{2\log(2\,\mathbb{S}_{\mathcal{A}}(n))}{n}},$$
où $\mathbb{S}_{\mathcal{A}}(n) = \max_{(x_1,\ldots,x_n)\in\mathbb{R}^n} \mathrm{Card}\mathcal{A}(x_1, \ldots, x_n)$ est le coefficient d'éclatement de la classe $\mathcal{A}$. Ce résultat ne dépend plus de $\mathcal{L}(X, Y)$.

4. Dimension de Vapnik-Chervonenkis, lemme de Sauer, exemples, interprétations et limites du résultat.

Chapitre 6 : Apprentissage et convexification du risque (séance 11-12)

1. On réécrit $Y \in \{0, 1\}$ sous la forme $Y' \in \{-1, 1\}$ via $Y = 2Y' - 1$. Sans changer de notation dans la suite, si $h : \mathcal{X} \to \{-1, 1\}$ est un classifieur,
$$\mathbb{P}(h(X) \neq Y) = \mathbb{E}[\varphi_0(-Yh(X))]$$
où $\varphi_0(z) = \mathbf{1}_{\{z\geq 0\}}$. Notion de substitut convexe : toute fonction φ t.q. $\varphi_0 \leq \varphi$. Hinger loss, fonction exponentielle; logistique.

2. Principe de convexification: si $\mathcal{F}$ est un ensemble convexe de fonctions de $\mathcal{X} \to \mathbb{R}$, on cherche
$$\widehat{f}_{n,\varphi} = \mathrm{argmin}_{f\in\mathcal{F}} R_{n,\varphi}(f),$$
où
$$R_{n,\varphi}(f) = n^{-1}\sum_{i=1}^n \varphi\big(-Y_i f(X_i)\big)$$
et on obtient un classifieur en posant $\widehat{h}_{n,\varphi}(x) = \mathrm{sign}\big(\widehat{f}_{n,\varphi}(x)\big)$. Choix de $\mathcal{F}$: enveloppe convexe ou boule ℓ^1 construite à partir d'un dictionnaire fini de classifieurs.

3. Minimisation du φ-risque empirique. Si φ est strictement convexe et différentiable, alors $f^\star = \mathrm{argmin}_f R_\varphi(f)$ est bien défini et $\mathrm{sign}(f^\star)$ est le classifieur bayésien.

4. Inégalité de Zhang : pour toute fonction $f : \mathcal{X} \to \mathbb{R}$ mesurable, il existe deux constantes $c > 0$ et $0 < \gamma \leq 1$ telles que

$$R\big(\mathrm{sign}(f)\big) - R^\star \leq 2c\big(R_\varphi(f) - R_\varphi^\star\big)^\gamma,$$

dès lors que $\varphi : \mathbb{R} \to \mathbb{R}$ est convexe et vérifie $\varphi(0) = 1$, $\varphi(z) \geq \varphi(-z)$ si $z \geq 0$ et $\tau(\eta) = \inf_\alpha \big(\eta\varphi(-\alpha) + (1-\eta)\varphi(\alpha)\big)$ vérifie

$$\big|\tfrac{1}{2} - \eta\big| \leq c\big(1 + \tau(\eta)\big)^\gamma.$$

5. Inégalité oracle pour le φ-risque. Si φ est L-Lipischitz , $\varphi(0) = 1$ et $\mathcal{F}$ est une boule ℓ^1 construite sur un dictionnaire de classifieurs de taille M, alors, sous les hypothèses de l'inégalité de Zhang,

$$\mathbb{E}\big[R_\varphi(\widehat{f}_{n,\varphi})\big] - \inf_{f\in\mathcal{F}} R_\varphi(f) \leq 8L\sqrt{\frac{2\log(2M)}{n}}$$

et

$$\mathbb{E}\big[R(\widehat{h}_{n,\varphi})\big] - R^\star \leq 2c\Big(8L\sqrt{\frac{2\log(2M)}{n}}\Big)^\gamma + 2c\big(\inf_{f\in\mathcal{F}} R_\varphi(f) - R_\varphi^\star\big)^\gamma.$$

Preuve à partir de l'inégalité de Zhang et du principe de contraction dù à Ledoux et Talagrand (admis).

Chapitre 7 : Introduction aux SVM (séance 13)