MAP553 - Ecole Polytechnique Année 2014-2015
Apprentissage statistique
Rapports à envoyer à guillaume.lecue@cmap.polytechnique.fr
au plus tard le 10/12/2014

Optimalité de la méthode de minimisation du risque empirique pour le problème d'agrégation convexe.

1. Introduction au problème d'agrégation convexe

L'objectif de ce Projet Individuel est de proposer une introduction à l'apprentissage statistique, aux méthodes de processus empiriques et à la méthode de Maurey par le biais du probléme d'agrégation convexe.

On se propose d'étudier des données de type entrée/sortie. L'objectif étant d'inférer ou prédire une sortie associée à une nouvelle entrèe en fonction des données précédemment observées. On dispose de n données $(X_i, Y_i)_{i=1}^n$ où X_i est une donnée d'entrée à valeurs dans un espace mesurable quelconque $\mathcal{X}$ et Y_i est un "label" ou sortie associée à l'entrée X_i à valeurs dans un intervalle borné $[-b, b]$ pour un certain $b > 0$. On reçoit une nouvelle entrée X et on souhaite prédire la sortie Y la plus naturellement associée à X en restant en accord avec ce qui a été observé avant. Ce problème a de multiple applications concrétes.

On modèlise ce problème de la manière suivante : les données (X_i, Y_i) pour $i = 1, \dots, n$ et le "nouveau" couple entrée/sortie (X, Y) sont supposés indépendants et identiquement distribués. On souhaite construire des fonctions qui dépendent des données $\mathcal{D} := \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ et de la nouvelle entrée X pour prédire au mieux la sortie Y. De telles procédures sont appelées procèdures d'apprentissage, estimateurs ou statistiques. On va donc s'intérésser aux fonctions mesurables $\hat{f}_n : (\mathcal{X} \times [-b, b])^n \times \mathcal{X} \longmapsto \mathbb{R}$ telles que la distance moyenne entre $\hat{f}_n(\mathcal{D}, X)$ (la prédiction qui est faite à partir des données $\mathcal{D}$ pour l'entrée X) et Y (en quelque sorte la "vraie" sortie) soient la plus petite possible. On considére la distance L_2 (mais d'autres distances sont aussi envisageables). On définit alors le risque quadratique de $\hat{f}_n$ par

$$\begin{aligned} R(\hat{f}_n) &= \mathbb{E}_{(X,Y)}(\hat{f}_n(\mathcal{D}, X) - Y)^2 = \mathbb{E}\big[(\hat{f}_n(\mathcal{D}, X) - Y)^2 | \mathcal{D}\big] \\ &= \int_{\mathcal{X}\times\mathbb{R}} (\hat{f}_n(\mathcal{D}, x) - y)^2 d\mathbb{P}_{(X,Y)}(x, y) \end{aligned} \tag{1}$$

où $\mathbb{E}_{(X,Y)}$ est l'espérance par rapport à (X, Y) et $\mathbb{P}_{(X,Y)}$ est la mesure de probabilité de (X, Y). Pour simplifier l'écriture, on ne présisera plus que les statistiques

$\hat{f}_n$ dépendent des données $\mathcal{D}$. Suivant cette convention, le risque quadratique d'un estimateur $\hat{f}_n$ s'écrit $R(\hat{f}_n) = \mathbb{E}\big[(\hat{f}_n(X) - Y)^2 | \mathcal{D}\big]$. On cherche donc à constuire des estimateurs $\hat{f}_n$ ayant le plus petit risque quadratique $R(\hat{f}_n)$.

Dans ce projet individuel, on s'intéressera à un certain type d'estimateur : ceux qui peuvent s'écrire comme combinaison convexe d'éléments d'un ensemble fini de fonctions de $\mathcal{X}$ dans $[-b, b]$. Un tel ensemble s'appelle un *dictionnaire*. Un dictionnaire peut se construire à partir d'éléments d'une base qu'on pense particulièrement bien adaptée au probléme traité, ou d'un grand nombre de fonctions simples comme des indicatrices de demi-espace ou encore, si on dispose d'autres données, on peut aussi construire une multitude d'estimateurs (possiblement non-adaptatifs) et en faire un dictionnaire, etc.. Qu'importe la manière dont a été construit ce dictionnaire, pour notre problème d'agrégation, on notera ses éléments par $f_1, \ldots, f_M$. Les f_j sont donc des fonctions de $\mathcal{X}$ à valeurs dans $[-b, b]$. On s'intéressera alors à des estimateurs de la forme

$$\hat{f}_n = \sum_{j=1}^{M} w_j f_j \tag{2}$$

où les poids w_j sont positifs et de somme égale à 1 (de telle sorte que $\hat{f}_n$ est bien une combinaison convexe d'éléments du dictionnaire). Un tel estimateur est appelé *méthode d'agrégation.* On souhaite choisir les poids w_j de telle sorte que (2) fasse aussi bien que la meilleure combinaison convexe d'élément dans $F = \{f_1, \ldots, f_M\}$, le dictionnaire. Les poids w_j devront donc être choisis à l'aide des données $\mathcal{D}$.

D'un point de vue mathématique, ce problème d'optimalité (i.e. "faire mieux que la meilleure combinaison convexe dans F) se traduit par une *inégalité oracle* : construire $\hat{f}_n$ telle que "avec grande probabilité (vis-à-vis des données)", on a

$$R(\hat{f}_n) \leqslant \inf_{f \in \mathrm{conv}(F)} R(f) + r(n, M) \tag{3}$$

où $\mathrm{conv}(F)$ est l'enveloppe convexe de F définie par

$$\mathrm{conv}(F) = \Big\{ \sum_{j=1}^{M} \lambda_j f_j : \lambda_j \geqslant 0, \sum_j \lambda_j = 1 \Big\}$$

et $r(n, M)$ est le terme résiduel qu'on souhaite aussi petit que possible. On sera aussi intéressé par des résultats en espérance, càd des inégalités oracle du type :

$$\mathbb{E}R(\hat{f}_n) \leqslant \inf_{f \in \mathrm{conv}(F)} R(f) + r(n, M) \tag{4}$$

où l'espérance $\mathbb{E}$ est prise par rapport aux données $\mathcal{D}$. La construction d'estimateur tels que (3) et/ou (4) sont satisfaites avec un terme résidual $r(n, M)$ aussi petit que possible s'appele le probléme d'agrégation convexe. Il existe d'autres problèmes d'agrégation : faire mieux que le meilleur élément dans F, faire mieux

que le meilleur élément dans l'espace linéaire engendré par F etc.. Pour ce Projet Individuel, on s'interessera d'abord au problème d'agrégation convexe.

2. Vitesse optimale d'agrégation et minimiseur du risque empirique

Un exemple de méthode d'agrégation est le *minimiseur du risque empirique* défini par :

$$\hat{f}_n^{ERM} \in \operatorname*{argmin}_{f \in \mathrm{conv}(F)} R_n(f) \tag{5}$$

où $R_n(f)$ est le risque empirique de f défini par

$$R_n(f) = \frac{1}{n}\sum_{i=1}^{n}(f(X_i) - Y_i)^2. \tag{6}$$

L'objectif de ce PI est de montrer que cette méthode est optimale pour le problème d'agrégation convexe. On doit d'abord définir ce qui est entendu par optimal. On introduit ici une définition d'optimalité pour ce problème.

Definition 2.1. *Soit n (nombre d'observations) et M (nombre d'éléments dans le dictionnaire) deux entiers. On dit que $\hat{f}_n$ est une* ***procédure optimale d'agrégation convexe*** *et que $r(n, M)$ est une* ***vitesse optimale d'agrégation*** *quand il existe deux constantes absolues $c_0 > 0$ et $c_1 > 0$ telles que les deux points suivants sont vérifiés :*

- *Pour tout dictionnaire $F = \{f_1, \dots, f_M\}$ de cardinal M et tout couple (X, Y) de variables aléatoires tels que $|Y| \leqslant b$ et $|f_j(X)| \leqslant b, \forall j = 1, \dots, M$ p.s., on a*

$$\mathbb{E}R(\hat{f}_n) \leqslant \min_{f \in \mathrm{conv}(F)} R(f) + c_0 r(n, M).$$

- *Pour toute statistique $\tilde{f}_n$, il existe un dictionnaire $F = \{f_1, \dots, f_M\}$ et un couple (X, Y) de variables aléatoires tels que $|Y| \leqslant b$ et $|f_j(X)| \leqslant b, \forall j = 1, \dots, M$ p.s. et*

$$\mathbb{E}R(\tilde{f}_n) \geqslant \min_{f \in \mathrm{conv}(F)} R(f) + c_1 r(n, M).$$

On remarque que la vitesse optimale d'agrégation convexe est définie ici à une constante absolue prés. La théorie minimax nous apprend que la vitesse minimale d'agrégation convexe est donnée par

$$\psi_{n,M}^{(C)} = \begin{cases} \frac{M}{n} & \text{quand } M \leqslant \sqrt{n} \\ \sqrt{\frac{1}{n}\log\left(\frac{eM}{\sqrt{n}}\right)} & \text{sinon.} \end{cases} \tag{7}$$

L'objectif de ce PI est de démontrer que le minimiseur du risque empirique défini en (5) atteint cette vitesse. C'est-à-dire que $\hat{f}_n^{ERM}$ vérfie une inégalité oracle comme (4) où le terme résiduel est proportionnel à $\psi_{n,M}^{(C)}$.

Theorem 2.2. *Il existe une constante absolue $c_0 > 0$ telle que pour tout $n \geqslant 1$ et $M \geqslant 1$, ce qui suit est vérifié. Pour tout dictionnaire F de cardinal M et tout couple (X, Y) de variables aléatoires tels que $|Y| \leqslant b$ p.s. et $|f(X)| \leqslant b, \forall f \in F$ p.s., on a pour $\hat{f}_n^{ERM}$ le minimiseur du risque empirique sur* $\mathrm{conv}(F)$,

$$\mathbb{E}R(\hat{f}_n^{ERM}) \leqslant \min_{f \in F} R(f) + c_0 b^2 \psi_{n,M}^{(C)}.$$

On ne prouvera ce résultat que dans le cas (le plus intéressant) $M \geqslant \sqrt{n}$. Pour cela, on aura recours à un résultat sur les processus empiriques qu'on pourra admettre dans une première lecture et à la méthode de Maurey. C'est cette méthode qu'on introduit en premier lieu.

3. La méthode empirique de Maurey

La méthode empirique de Maurey a été introduite pour le calcul de l'entropie de la boule unité B_1^d par rapport à la métrique euclidienne de $\mathbb{R}^d$. On rappelle ici ce calcul.

On commence par quelques notations. Les boules unités pour les normes ℓ_1^d et ℓ_2^d sont

$$B_1^d = \Big\{x \in \mathbb{R}^d : \sum_{j=1}^M |x_j| \leqslant 1\Big\} \text{ et } B_2^d = \Big\{x \in \mathbb{R}^d : \sum_{j=1}^M x_j^2 \leqslant 1\Big\}.$$

Pour tout ensemble $T \subset \mathbb{R}^d$, on note par $N(T, \varepsilon B_2^d)$ le plus petit nombre de translatés de εB_2^d nécessaires pour recouvrir entièrement T. L'entropie de T par rapport à ℓ_2^d est la fonction $\varepsilon \longmapsto \log N(T, \varepsilon B_2^d) := \mathcal{N}(T, \varepsilon, \ell_2^d)$. On va démontrer la proposition suivante (qui est optimale à des constantes absolues prés) par la méthode empirique de Maurey.

Proposition 3.1. *Il existe une constante absolue $c_0 > 0$ telle que pour tout $\varepsilon > 0$,*

$$\log N(B_1^d, \varepsilon B_2^d) \leqslant c_0 \begin{cases} 0 & si\ \varepsilon \geqslant 1, \\ \frac{1}{\varepsilon^2} \log(d\varepsilon^2) & si\ d^{-1/2} \leqslant \varepsilon \leqslant 1, \\ d \log\left(\frac{e}{d\varepsilon^2}\right) & si\ \varepsilon \leqslant d^{-1/2}. \end{cases}$$

Q1.1 Montrer le cas $\varepsilon \geqslant 1$.

Q1.2 Soit $x \in B_1^d$ et $d^{-1/2} \leqslant \varepsilon \leqslant 1$. On veut montrer que x est proche (au sens ℓ_2^d) d'un sous-ensemble Λ de B_1^d dont le logarithme du cardinal est plus petit que $c_0 \varepsilon^{-2} \log(d\varepsilon^2)$. Pour cela, on utilise la méthode empirique de Maurey. On écrit $x = \sum_{j=1}^d x_j e_j$ où $(e_1, \ldots, e_M)$ est la base canonique de $\mathbb{R}^d$ et $\sum_j |x_j| \leqslant 1$. On considère une variable aléatoire Z à valeurs dans

$\{0, \pm e_1, \dots, \pm e_d\}$ telle que $\mathbb{P}[Z = 0] = 1 - \|x\|_1$ et $\mathbb{P}[Z = \mathrm{sign}(x_i)e_i] = |x_i|$. Montrer que $\mathbb{E}Z = x$.

Q1.3 Soit $Z_1, \dots, Z_p$ des variables aléatoires i.i.d. distribuées comme Z. Montrer que

$$\mathbb{E}\left\|x - \frac{1}{m}\sum_{i=1}^{m} Z_i\right\|_2^2 = \frac{\mathbb{E}\,\|Z - \mathbb{E}Z\|_2^2}{m} \leqslant \frac{4}{m}.$$

Q1.4 En déduire que pour m_ε le plus petit entier m tel que $4/m \leqslant \varepsilon^2$, l'ensemble

$$\Lambda := \left\{\frac{1}{m_\varepsilon}\sum_{i=1}^{m_\varepsilon} z_i : z_1, \dots, z_{m_\varepsilon} \in \{0, \pm e_1, \dots, \pm e_d\}\right\} \tag{8}$$

est un ε-réseau de B_1^d pour ℓ_2^d (c'est-à-dire que pour tout $x \in B_1^d$ il existe $y \in \Lambda$ tel que $\|x - y\|_2 \leqslant \varepsilon$).

Q1.5 Montrer que le cardinal de Λ est tel que

$$\log|\Lambda| \leqslant \frac{c_0}{\varepsilon^2}\log(d\varepsilon^2).$$

En déduire le cas $d^{-1/2} \leqslant \varepsilon \leqslant 1$ de Proposition 3.1.

Q1.6 Montrer que pour tout $\varepsilon, \eta > 0$, on a

$$\log N(B_1^d, \varepsilon B_2^d) \leqslant \log N(B_1^d, \eta B_2^d) + \log N(\eta B_2^d, \varepsilon B_2^d).$$

Q1.7 Par un argument volumique, montrer que

$$N(\eta B_2^d, \varepsilon B_2^d) \leqslant \left(1 + \frac{2\eta}{\varepsilon}\right)^d. \tag{9}$$

Q1.8 Déduire le troisième cas de Proposition 3.1 de Q1.7, Q1.6 et du deuxième cas.

4. Un résultat sur les processus empirique

On introduit quelques notations classique en apprentissage statistique. La fonction de perte quadratique d'une fonction $f : \mathcal{X} \longmapsto \mathbb{R}$ est donnée par,

$$\ell_f(x, y) = (y - f(x))^2, \quad \forall x \in \mathcal{X}, y \in \mathbb{R}.$$

Le risque quadratique d'une fonction f s'écrit alors $R(f) = \mathbb{E}\ell_f(X, Y)$.

Soit f, g deux fonctions. On note par $[f, g]$ le ségment de f à g.

Q2.1 Montrer que $R(\cdot)$ atteint son minimum sur $[f, g]$.

Soit $f^* \in \mathrm{argmin}_{h\in[f,g]} R(h)$. Pour tout $h \in [f, g]$, on note par

$$\mathcal{L}_h(x, y) = \ell_h(x, y) - \ell_{f^*}(x, y), \quad \forall x \in \mathcal{X}, y \in \mathbb{R}$$

la *fonction de perte en excés* de h. Par ailleurs, on note

$$P\mathcal{L}_h = \mathbb{E}\mathcal{L}_h(X, Y) \text{ et } P_n\mathcal{L}_h = \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_h(X_i, Y_i). \tag{10}$$

On admettra le résultat suivant.

Proposition 4.1. *Il existe une constante absolue $c_0 > 0$ telle que ce quit suit a lieu. Pour tout $x > 0$, avec probabilité plus grande que $1 - 4\exp(-x)$, pour tout $h \in [f, g]$,*

$$\big|P\mathcal{L}_h - P_n\mathcal{L}_h\big| \leqslant \frac{1}{2}\max\big(P\mathcal{L}_h, \frac{c_0 x b^2}{n}\big).$$

5. Preuve du Théorème 2.2 pour le cas $M \geqslant \sqrt{n}$

On considère l'entier

$$m = \left\lceil \sqrt{\frac{n}{\log\big(eM/\sqrt{n}\big)}} \right\rceil$$

et le sous-ensemble $\mathcal{C}' \subset \mathcal{C} := \mathrm{conv}(F)$ défini par

$$\mathcal{C}' = \Big\{ \frac{1}{m}\sum_{j=1}^{m} h_j : h_1, \ldots, h_m \in F \Big\}$$

où, on rappelle que $F = \{f_1, \ldots, f_M\}$ est le dictionnaire.

Q3.1 Montrer en utilisant la méthode de Maurey que

$$\min_{f\in\mathcal{C}'} R(f) \leqslant \min_{f\in\mathcal{C}} R(f) + \frac{4b^2}{m}.$$

Pour cela, on pourra introduire $f^*_{\mathcal{C}} \in \mathrm{argmin}_{f\in\mathcal{C}} R(f)$.

Q3.2 En utilisant la Proposition 4.1 et une "union bound", prouver que pour $N = |\mathcal{C}'|$ et $\mathcal{C}' = \{g_1, \ldots, g_N\}$ tel que $R(g_1) = \min_{g\in\mathcal{C}'} R(g)$, on a pour tout $x > 0$, avec probabilité au moins $1 - 4\exp(-x)$, pour tout ségment $[g_1, g_j], j = 1, \ldots, N$,

$$\big|P\mathcal{L}_g^{1j} - P_n\mathcal{L}_g^{1j}\big| \leqslant (1/2)\max\big(P\mathcal{L}_g^{1j}, \gamma(x)\big), \quad \forall g \in [g_1, g_j] \tag{11}$$

où $\gamma(x) = c_0 b^2(x + \log N)/n$ et $\mathcal{L}_g^{1j}$ est la fonction d'excés de risque de g par rapport au ségment $[g_1, g_j]$ (càd si $g^*_{1j} \in \mathrm{argmin}_{g\in[g_1,g_j]} R(h)$ alors $\mathcal{L}_g^{1j} = \ell_g - \ell_{g^*_{1j}}$). On note par $\Omega(x)$ l'événement sur lequel (11) a lieu (pour tout j).

On fixe $X_1, \ldots, X_n$. On écrit $\hat{f}_n^{ERM} = \sum_{j=1}^M \beta_j f_j$ et on considère $\Theta : \Omega' \to F$ defini sur un autre espace de probabilité $(\Omega', \mathcal{A}', \mathbb{P}')$ tel que $\mathbb{P}'[\Theta = f_j] = \beta_j, \forall j = 1, \ldots, M$ et on prend m copies i.i.d. $\Theta_1, \ldots, \Theta_m$ de Θ. On note par $\mathbb{E}'_\Theta$ l'espérance par rapport à $\Theta_1, \ldots, \Theta_m$ et par $\mathbb{V}_\Theta$ la variance par rapport à Θ.

Q3.3 Montrer que $\mathbb{E}'_\Theta\Theta_j = \hat{f}_n^{ERM}$ pour tout $j = 1, \ldots, m$ et en utilisant la méthode de Maurey que

$$\mathbb{E}'_\Theta R\Big(\frac{1}{m}\sum_{j=1}^{m}\Theta_j\Big) = R(\hat{f}_n^{ERM}) + \frac{\mathbb{E}\mathbb{V}'_\Theta(Y - \Theta(X))}{m}. \tag{12}$$

Montrer que la méthode de Maurey fournit une preuve que, pour le risque empirique, on a aussi

$$\mathbb{E}'_\Theta R_n\Big(\frac{1}{m}\sum_{j=1}^m \Theta_j\Big) = R_n(\hat{f}_n^{ERM}) + \frac{1}{m}\Big(\frac{1}{n}\sum_{i=1}^n \mathbb{V}'_\Theta(Y_i - \Theta(X_i))\Big). \tag{13}$$

On introduit la notation suivante :

$$g_\Theta = \frac{1}{m}\sum_{j=1}^m \Theta_j \text{ et } i_\Theta \in \{1,\dots,N\} \text{ tel que } g_{i_\Theta} = g_\Theta.$$

On remarque que g_Θ est un point aléatoire prenant ses valeurs dans $\mathcal{C}'$ (en tant que fonction measurable de Ω' dans $\mathcal{C}'$) et que sur l'événement $\Omega(x)$, on a la propriété d'isomorphie suivante sur le ségment $[g_1, g_\Theta]$:

$$\big|P_n\mathcal{L}_g^{1i_\Theta} - P\mathcal{L}_g^{1i_\Theta}\big| \leqslant (1/2)\max\big(P\mathcal{L}_g^{1i_\Theta}, \gamma(x)\big), \quad \forall g \in [g_1, g_{i_\Theta}]. \tag{14}$$

Q3.4 On fixe $\Theta_1,\dots,\Theta_m$. En introduisant le risque de $g^*_{1i_\Theta} \in \operatorname{argmin}_{g\in[g_{i_\Theta},g_1]} R(g)$, montrer que

$$R(\hat{f}_n^{ERM}) \leqslant \min_{f\in\mathcal{C}} R(f) + \frac{4b^2}{m} + P\mathcal{L}_{g_\Theta}^{1i_\Theta} + R(\hat{f}_n^{ERM}) - R(g_\Theta). \tag{15}$$

Q3.5 Montrer que sur l'événement $\Omega(x)$, on a pour tout $\Theta_1,\dots,\Theta_m$

$$R(\hat{f}_n^{ERM}) \leqslant \min_{f\in\mathcal{C}} R(f) + \frac{4b^2}{m} + \gamma(x) + 2\big(R_n(g_\Theta) - R_n(\hat{f}_n^{ERM})\big) + R(\hat{f}_n^{ERM}) - R(g_\Theta).$$

Q3.6 En prenant l'espérance par rapport à $\Theta_1,\dots,\Theta_m$ (defini sur Ω'), montrer que sur $\Omega(x)$,

$$R(\hat{f}_n^{ERM}) \leqslant \min_{f\in\mathcal{C}} R(f) + \frac{4b^2}{m} + \gamma(x) + 2\mathbb{E}'_\Theta\big(R_n(g_\Theta) - R_n(\hat{f}_n^{ERM})\big) + \mathbb{E}'_\Theta\Big(R(\hat{f}_n^{ERM}) - R(g_\Theta)\Big).$$

Q3.7 Démontrer le Théorème 2.2 dans le cas $M \geqslant \sqrt{n}$. On montrera d'abord un résultat en déviation : pour tout $x > 0$, avec probabilité plus grande que $1 - 4\exp(-x)$,

$$R(\hat{f}_n^{ERM}) \leqslant \min_{f\in\mathrm{conv}(F)} R(f) + c_0 b^2 \max\Big(\sqrt{\frac{1}{n}\log\Big(\frac{eM}{\sqrt{n}}\Big)}, \frac{x}{n}\Big).$$

On conclura par intégration de ce résultat pour obtenir un résultat en espérance.

Références

[1] Guillaume Lecué. Empirical risk minimization is optimal for the convex aggregation problem. To appear in Bernoulli journal, 2011.

[2] Alexandre Tsybakov. Optimal rate of aggregation. In *Computational Learning Theory and Kernel Machines (COLT-2003)*, volume 2777 of *Lecture Notes in Artificial Intelligence*, pages 303–313. Springer, Heidelberg, 2003.