Bornes de généralisation

Thierry Bazier-Matte

11 février 2017

1 Garanties statistiques

1.1 Bornes de généralisation

Exposition du problème Soit $\boldsymbol{Q}$ un espace de Hilbert à noyau reproduisant induit par κ et soit un ensemble d'entraînement $\mathcal{S}_n = \{(x_i, y_i)\}_{i=1}^n \sim M^n$ échantilloné à partir de la distribution de marché. Alors on peut définir *l'algorithme de décision* $\mathcal{Q} : \boldsymbol{M}^n \to \boldsymbol{Q}$ par

$$\mathcal{Q}(\mathcal{S}_n) = \arg\max_{q \in \boldsymbol{Q}} \left\{ \widehat{\boldsymbol{EU}}(\mathcal{S}_n, q) - \lambda \|q\|^2 \right\}. \tag{1}$$

Comme on l'a vu, résoudre (1) est aussi équivalent à

$$\underset{\alpha \in \mathscr{R}^n}{\text{maximiser}} \quad n^{-1} \sum_{i=1}^{n} u(r_i \, \alpha^T \phi(x_i)) - \lambda \|\alpha\|_K^2, \tag{2}$$

où $\phi : \mathscr{R}^p \to \mathscr{R}^n$ le vecteur d'application induit par la matrice d'information Ξ. La relation $q = \alpha^T \phi$ permet de passer d'une représentation à l'autre.

La question qui se pose naturellement est de savoir dans quelle mesure une fonction de décision $\hat{q} = \mathcal{Q}(\mathcal{S}_n)$ est capable d'offrir à un investisseur une utilité espérée comparable à celle qu'il aurait observée au sein de l'ensemble d'entraînement. Il serait aussi souhaitable qu'une telle garantie soit indépendante de l'ensemble d'entraînement $\mathcal{S}_n$. Autrement dit, on cherche à déterminer une borne probabiliste Ω sur l'erreur de généralisation de $\hat{q} = \mathcal{Q}(\mathcal{S}_n)$ valide pour tout $\mathcal{S}_n \sim M^n$:

$$\hat{\zeta}(\mathcal{S}_n) \leq \Omega(n, \ldots), \tag{3}$$

où

$$\hat{\zeta}(\mathcal{S}_n) = \widehat{\boldsymbol{EU}}(\mathcal{S}_n, \mathcal{Q}(\mathcal{S}_n)) - \boldsymbol{EU}(\mathcal{Q}(\mathcal{S}_n)) \tag{4}$$

représente l'erreur de généralisation. On peut en fait montrer que $\Omega \to 0$ à mesure que $n \to \infty$. En fait, on peut démontrer que $\Omega = O(n^{-1/2})$, ce qui permet de quantifier la "vitesse" à laquelle la convergence à lieu.

Démonstration Considérons deux ensembles d'entraînement : $\mathcal{S}_n \sim M^n$ et $\mathcal{S}'_n$, où $\mathcal{S}'_n$ ne diffère de $\mathcal{S}_n$ que par un seul point (par exemple le j-ème point serait rééchantilloné de la distribution de marché M). De l'algorithme $\mathcal{Q}$ on dérivera alors deux décisions : $\hat{q}$ et $\hat{q}'$. Pour n suffisament grand, on peut alors s'attendre à ce que l'utilité dérivée de ces deux décisions soit relativement proche, et ce, pour toute observation. On aurait alors une borne $\beta(n)$ telle que pour tout $(x, r) \sim M$,

$$|u(r\,\hat{q}(x)) - u(r\,\hat{q}'(x))| \leq \beta. \tag{5}$$

C'est ce qu'on appelle dans la littérature la *stabilité algorithmique*. La plupart des algorithmes régularisés classiques disposent par ailleurs d'une telle stabilité. En particulier, le terme de régularisation $\lambda\|q\|^2$, combiné à la continuité Lipschitz de u font en sorte que $\beta = O(n^{-1})$.

Doté de cette stabilité de $\mathcal{Q}$, on peut alors borner la différence dans l'erreur de généralisation de $\mathcal{S}_n$ et $\mathcal{S}'_n$:

$$\begin{aligned}
|\hat{\zeta}(\mathcal{S}_n) - \hat{\zeta}(\mathcal{S}'_n)| &= |\boldsymbol{EU}(\hat{q}) - \boldsymbol{EU}(\hat{q}') + \widehat{\boldsymbol{EU}}(\mathcal{S}_n, \hat{q}) - \widehat{\boldsymbol{EU}}(\mathcal{S}'_n, \hat{q}')| && (6)\\
&\leq |\boldsymbol{EU}(\hat{q}) - \boldsymbol{EU}(\hat{q}')| + |\widehat{\boldsymbol{EU}}(\mathcal{S}_n, \hat{q}) - \widehat{\boldsymbol{EU}}(\mathcal{S}'_n, \hat{q}')|. && (7)
\end{aligned}$$

Or, par le théorème de Jensen appliqué à la fonction valeur absolue, on obtient du premier terme que

$$\begin{aligned}
|\boldsymbol{EU}(\hat{q}) - \boldsymbol{EU}(\hat{q}')| &= |\boldsymbol{E}(u(R \cdot \hat{q}(X)) - u(R \cdot \hat{q}'(X)))| && (8)\\
&\leq \boldsymbol{E}(|u(R \cdot \hat{q}(X)) - u(R \cdot \hat{q}'(X)|) && (9)\\
&\leq \beta, && (10)
\end{aligned}$$

par définiton de la stabilité. Quant au deuxième terme de (7) on peut le borner de la même façon :

$$\begin{aligned}
&|\widehat{\boldsymbol{EU}}(\mathcal{S}_n, \hat{q}) - \widehat{\boldsymbol{EU}}(\mathcal{S}'_n, \hat{q}')| && (11)\\
&\quad = n^{-1} \left| \sum_{i=1}^{n} \mathbb{I}_{i\neq j} u(r_i\,\hat{q}(x_i)) + u(r_j\,\hat{q}(x_j)) - \sum_{i=1}^{n} \mathbb{I}_{i\neq j} u(r_i\,\hat{q}'(x_i)) - u(r'_j\,\hat{q}'(x'_j)) \right| && (12)\\
&\quad \leq n^{-1} \left(|u(r_j\,\hat{q}(x_j) - u(r'_j\,\hat{q}'(x'_j))| + \sum_{i=1}^{n} \mathbb{I}_{i\neq j} \big| u(r_i\,\hat{q}(x_i)) - u(r_i\,\hat{q}'(x_i)) \big| \right) && (13)\\
&\quad \leq n^{-1} \left(|u(r_j\,\hat{q}(x_j) - u(r'_j\,\hat{q}'(x'_j))| + (n-1)\beta \right). && (14)
\end{aligned}$$

Considérons le premier terme. Par le lemme **[Todo:]**, On sait que $\hat{q}(x) \leq (2\lambda)^{-1}\bar{r}\xi^2$ et que $|R| \leq \bar{r}$. On peut donc borner cette différence par la différence dans l'utilité dérivée par la meilleure décision d'investissement sur le meilleur rendement et sur le pire rendement. Par hypothèse Lipschitz et de sous-gradient de 1 à $r = 0$, on sait que pour $r > 0, u(r) < r$ et que pour $r < 0$, $\gamma r \leq u(r)$. On peut donc conclure que

$$|u(r_j\,\hat{q}(x_j) - u(r'_j\,\hat{q}'(x'_j)| \leq u((2\lambda)^{-1}\bar{r}^2\xi^2) - u(-(2\lambda)^{-1}\bar{r}^2\xi^2) \tag{15}$$

$$\leq (2\lambda)^{-1}(\gamma+1)\bar{r}^2\xi^2. \tag{16}$$

Ce qui entraîne donc que

$$|\widehat{\boldsymbol{EU}}(\mathcal{S}_n,\hat{q}) - \widehat{\boldsymbol{EU}}(\mathcal{S}'_n,\hat{q}')| \leq \frac{\gamma+1}{2\lambda n}\bar{r}^2\xi^2 + \frac{n-1}{n}\beta \tag{17}$$

$$\leq \beta + \frac{\gamma+1}{2\lambda n}\bar{r}^2\xi^2. \tag{18}$$

Next McDiarmid et autre. Easy stuff.

Équivalent certain Puis inverser pour obtenir l'équivalent certain.

Note bibliographique La théorie de la stabilité algorithmique remonte en fait aux années 70 avec les travaux de Luc Devroye appliqués à l'algorithme des k plus proches voisins[Citation needed]. Jusqu'alors, les bornes de généralisation étaient présentées pour toute décision $q \in \boldsymbol{Q}$ (ie Vapnik). Bousquet[Citation needed]a été le premier a présenter des résultats dans des espaces de Hilbert à noyau reproduisant. La démonstration est fortement inspirée de l'excellente référence Mohri[Citation needed]. La démonstration de la borne de la décision bornée est un résultat inédit, dûe à Delage dans le cas linéaire.

1.2 Bornes de sous optimalité

Exposition du problème Jusqu'ici, les efforts théoriques ont été déployés pour déterminer comment se comportait la fonction de décision $\hat{q} = \mathcal{Q}(\mathcal{S}_n)$ dans un univers probabiliste par rapport à l'univers statistique dans lequel elle avait été construite. Notre attention va maintenant se tourner vers la performance de $\hat{q}$ dans l'univers probabiliste par rapport à la meilleure décision disponible, c'est à dire la solution $q^\star$ de

$$\underset{q\in\boldsymbol{Q}}{\text{maximiser}} \quad \boldsymbol{EU}(R\cdot q(X)). \tag{19}$$

Il convient cependant de réaliser que l'existence d'une borne sur $q^\star$ n'est pas assurée. En effet, supposons d'une part que l'on dispose d'une utilité neutre au risque, telle que $u(r) = r$, et d'autre part que $\boldsymbol{E}R = 0$. Soit $\alpha > 0$. On pourrait alors définir la fonction suivante :

$$q = \alpha\boldsymbol{E}(R\,\kappa(X,\cdot)) \tag{20}$$

On aurait alors

$$\boldsymbol{EU}(q) = \boldsymbol{E}(Rq(X)) = \boldsymbol{E}(R\boldsymbol{E}(R\,\kappa(X,X))) \tag{21}$$

$$= \boldsymbol{E}(R^2\,\kappa(X,X)) \geq 0, \tag{22}$$

On peut alors obtenir une utilité espérée non bornée à mesure que $\alpha \to \infty$. Par ailleurs, ainsi défini, q représente effectivement la covariance entre R et la projection de X dans l'espace dual de $\boldsymbol{Q}$. Puisque l'utilité est neutre, on sait qu'en espérance l'application

de q à X variera de la même façon que celle de R et donc qu'on aura une utilité infinie. On verra plus loin au cours d'une démonstration la motivation derrière cette hypothèse supplémentaire :

Hypothèse 1. *L'utilité croît sous-linéairement, ie.* $u(r) = o(r)$.

Une autre hypothèse est maintenant nécessaire pour s'assurer que $q^\star$ soit borné : l'efficience des marchés. Dans notre cadre théorique, ceci se traduit par l'absence de l'existence d'une fonction $q \in \boldsymbol{Q}$ telle que

$$\boldsymbol{P}\{R \cdot q(X) > 0\} = 1. \tag{23}$$

D'un point de vue strictement financier, cela fait certainement du sens en vertu de l'efficience des marchés, version semi-forte[Citation needed]. D'un point de vue théorique, ceci exige en fait qu'il n'y ait pas de région dans $\boldsymbol{X}$ telle que tous les rendements s'y produisant soient nécessairement positifs ou négatifs.**[Todo:** Inserer image**]**.

Hypothèse 2. *Pour toute région* $\mathcal{R} \subseteq \boldsymbol{X}$*,*

$$\boldsymbol{P}\{R \geq 0 \mid X \in \mathcal{R}\} < 1, \tag{24}$$

et de la même façon avec l'évènement $\boldsymbol{P}\{R \leq 0\}$.

Borne On cherchera donc à établir une borne sur *l'erreur de sous-optimalité* de $\hat{q} \sim \mathcal{Q}(M^n)$.

1.3 Lemmes

Stabilité On montre ici que

$$\beta \leq \frac{(\gamma \bar{r} \xi)^2}{2\lambda n}. \tag{25}$$

Borne sur la décision algorithmique On va ici démontrer que la décision $\hat{q}(x)$ est bornée, et ce, pour tout $x \in \boldsymbol{X}$ et pour toute solution $\hat{q}$ de

$$\underset{q \in \boldsymbol{Q}}{\text{maximiser}} \quad n^{-1} \sum_{i=1}^{n} u(r_i \, q(x_i)) - \lambda \|q\|^2. \tag{26}$$

Pour ce faire, on va mettre à profit la propriété reproductive de $\boldsymbol{Q}$ induite par κ. En effet, celle-ci stipule que

$$q(x) = \langle q, \kappa(x, \cdot) \rangle_{\boldsymbol{Q}} \leq \|q\| \sqrt{\kappa(x, x)}, \tag{27}$$

où l'inégalité découle de l'inégalité Cauchy-Schwartz appliquée au produit interne de $\boldsymbol{Q}$. On rappelle que, par hypothèse, $\forall x \in \boldsymbol{X}, \kappa(x, x) \leq \xi^2$; il suffit donc de borner $\|q\|_{\boldsymbol{Q}}$. Or, puisque $u(r) \leq r$, on remarque que

$$n^{-1} \sum_{i=1}^{n} u(r_i \, q(x_i)) \leq n^{-1} \sum_{i=1}^{n} r_i \, q(x_i) \tag{28}$$

$$\leq n^{-1} \sum_{i=1}^{n} r_i \sqrt{\kappa(x_i, x_i)} \|q\| \tag{29}$$

$$\leq \bar{r}\xi \|q\|. \tag{30}$$

Puisque l'expression $\bar{r}\xi\|q\| - \lambda\|q\|^2$ est quadratique et atteint son maximum lorsque

$$\|q\| = \frac{\bar{r}\xi}{2\lambda}, \tag{31}$$

on en conclut que $\|\hat{q}\| \leq (2\lambda)^{-1}\bar{r}\xi$ et donc que

$$\hat{q}(x) \leq \frac{\bar{r}\xi^2}{2\lambda}. \tag{32}$$

[Todo: Montrer qu'on peut effectivement montrer que la borne de l'expression est plus grande que celle du problème initial... Pour ce faire, utiliser a) le sous gradient de u et b) la dominance de q' sur q.**]**

Forte concavité L'objectif est fortement concave, que ce soit sous sa version statistique $\widehat{\boldsymbol{EU}}_\lambda$ ou probabiliste $\boldsymbol{EU}_\lambda$. Autrement dit, pour tout $\alpha \in [0, 1]$, on a

$$\boldsymbol{EU}_\lambda(tq_1+(1-\alpha)q_2) \geq \alpha \boldsymbol{EU}_\lambda(q_1)+(1-\alpha)\boldsymbol{EU}_\lambda(q_2)+\lambda\alpha(1-\alpha)\|q_1-q_2\|^2, \tag{33}$$

et de même pour $\widehat{\boldsymbol{EU}}_\lambda$. Effectivement, puisque u est concave et $\|\cdot\|^2$ est convexe, on a successivement :

$$\boldsymbol{EU}_\lambda(\alpha q_1 + (1-\alpha)q_2) \tag{34}$$

$$= \boldsymbol{E}u(R \cdot (\alpha q_1 + (1-\alpha)q_2)(X)) - \lambda\|\alpha q_1 + (1-\alpha)q_2\|^2 \tag{35}$$

$$= \boldsymbol{E}u(\alpha(R \cdot q_1(X)) + (1-\alpha)(R \cdot q_2(X))) - \lambda\|\alpha q_1 + (1-\alpha)q_2\|^2 \tag{36}$$

$$\geq \boldsymbol{E}(\alpha u(R \cdot q_1(X)) + (1-\alpha)u(R \cdot q_2(X))) - \lambda\|\alpha q_1 + (1-\alpha)q_2\|^2 \tag{37}$$

$$= \alpha \boldsymbol{EU}(q_1) + (1-\alpha)\boldsymbol{EU}(q_2) - \lambda\|\alpha q_1 + (1-\alpha)q_2\|^2 \tag{38}$$

$$= \alpha \boldsymbol{EU}_\lambda(q_1) + (1-\alpha)\boldsymbol{EU}_\lambda(q_2) - \lambda(\|\alpha q_1 + (1-\alpha)q_2\|^2 - \alpha\|q_1\|^2 - (1-\alpha)\|q_2\|^2) \tag{39}$$

$$\geq \alpha \boldsymbol{EU}_\lambda(q_1) + (1-\alpha)\boldsymbol{EU}_\lambda(q_2) - \lambda(\alpha\|q_1\|^2 + (1-\alpha)\|q_2\|^2 - \alpha\|q_1\|^2 - (1-\alpha)\|q_2\|^2) \tag{40}$$

$$= \alpha \boldsymbol{EU}_\lambda(q_1) + (1-\alpha)\boldsymbol{EU}_\lambda(q_2). \tag{41}$$

La preuve est la même lorsqu'on considère $\widehat{\boldsymbol{EU}}_\lambda$.

Borne sur la décision optimale On veut montrer que $\|q^\star\|$ est borné. Pour ce faire, on va tout d'abord décomposer $q = s\theta$, où on pose $\|\theta\| = 1$ et $s > 0$; ainsi on peut poser notre problème d'optimisation comme la recherche d'une 'direction' θ et d'une magnitude s dans $\boldsymbol{Q}$. De plus, puisque $\|q\| = s$, il suffit de montrer que $s^\star$ est borné.

Notons d'abord que l'hypothèse 2 entraîne en particulier qu'il existe $\delta > 0$ tel que

$$\boldsymbol{P}\{R \cdot \theta(X) \leq -\delta\} > \varrho \geq 0 \tag{42}$$

pour tout $\theta \in \boldsymbol{Q}$ tel que $\|\theta\| = 1$. Définissons maintenant une variable aléatoire à deux états : $B = -\delta$ avec probabilité ϱ et $B = \bar{r}\xi$ avec probabilité $1 - \varrho$. Puisque $R \cdot \theta(X) \leq \bar{r}\xi$, on a alors que, pour tout $r \in \boldsymbol{R}$,

$$\boldsymbol{P}\{B \geq r\} \geq \boldsymbol{P}\{R \cdot \theta(X) \geq r\} \tag{43}$$

[Todo: voir figure a produire.**]**

Puisque u est concave[Citation needed]et que B domine stochastiquement $R \cdot \theta(X)$, on a nécessairement que $\boldsymbol{E}u(sB) \geq \boldsymbol{E}u(R \cdot s\theta(X)$, pour tout $s > 0$. Or, par hypothèse de sous-linéarité on obtient que

$$\begin{aligned}
\lim_{s\to\infty} \boldsymbol{E}u(R \cdot s\theta(X)) &\leq \lim_{s\to\infty} u(sB) && (44)\\
&= \lim_{s\to\infty} (\varrho u(-s\delta) + (1-\varrho)u(s\bar{r}\xi)) && (45)\\
&\leq \lim_{s\to\infty} -\varrho s\delta + (1-\varrho)o(s) = -\infty, && (46)
\end{aligned}$$

ce qui démontre bien que s est borné.

Borne sur la solution utilitaire sur celle neutre au risque Soient $\hat{q}_u$ la solution de

$$\underset{q\in\boldsymbol{Q}}{\text{maximiser}} \quad \widehat{\boldsymbol{EU}}_\lambda(q) \tag{47}$$

et $\hat{q}_1$ la solution de

$$\underset{q\in\boldsymbol{Q}}{\text{maximiser}} \quad \widehat{\boldsymbol{E1}}_\lambda(q), \tag{48}$$

où $\widehat{\boldsymbol{E1}}(q) := n^{-1}\sum_{i=1}^n r_i\, q(x_i)$. On note tout d'abord avec l'inégalité de Jensen que $u(\widehat{\boldsymbol{E1}}(\hat{q}_u)) \geq \widehat{\boldsymbol{EU}}(\hat{q}_u) \geq \lambda\|\hat{q}_u\|^2 \geq 0$. Mais puisque u a un sur-gradient de 1 à 0, on déduit que $u(x) \geq 0$ entraîne $x \geq u(x)$. On a ainsi $\widehat{\boldsymbol{E1}}(\hat{q}_u) - \lambda\|\hat{q}_u\|^2 \geq 0$. Mais comme $\hat{q}_1$ maximise $\widehat{\boldsymbol{E1}}_\lambda$, on obtient

$$\widehat{\boldsymbol{E1}}(\hat{q}_1) - \lambda\|\hat{q}_1\|^2 \geq \widehat{\boldsymbol{E1}}(\hat{q}_u) - \lambda\|\hat{q}_u\|^2 \geq 0, \tag{49}$$

d'où on tire finalement $\|\hat{q}_u\| \leq \|\hat{q}_1\|$.