Combinaison d'estimateurs en apprentissage statistique : quel impact respectif des proximités en entrée et sortie ?

Aurélie Fischer & Mathilde Mougeot

Laboratoire de Probabilités et Modèles Aléatoires,
Université Paris Diderot
75013 Paris, France
aurelie.fischer@univ-paris-diderot.fr
mathilde.mougeot@univ-paris-diderot.fr

Résumé. Nous introduisons une nouvelle stratégie en apprentissage statistique basée sur une idée de Mojirsheibani (1999, 2000, 2002a, 2002b): cet auteur a proposé une méthode pour combiner plusieurs classifieurs, reposant sur une notion de consensus, qui a été étendue récemment au contexte de la régression dans Biau et al. (2015).

Dans ces approches, une certaine condition de consensus entre estimateurs doit être satisfaite pour tous les estimateurs individuels, ce qui pourrait poser problème s'il existe un mauvais estimateur initial. En pratique, quelques désaccords sont autorisés ; pour l'obtention des résultats théoriques, il est demandé que la proportion d'estimateurs satisfaisant la condition de consensus tende vers 1.

Ici, nous proposons une procédure modifiée mélangeant ces idées de consensus avec une méthode plus classique basée sur la distance euclidienne entre les entrées. Cette stratégie peut être vue comme une approche alternative pour réduire l'effet d'un éventuel mauvais estimateur dans la liste initiale.

Nous démontrons la convergence de cette méthode et proposons quelques expériences numériques.

Mots-clés. Classification, régression, estimation, consistance, agrégation.

Abstract. We introduce a new learning strategy based on an idea of Mojirsheibani (1999, 2000, 2002a, 2002b): this author proposed a method for combining several classifiers, relying on a consensus notion, which has been recently extended to the context of regression in Biau et al. (2015).

In these approaches, some agreement condition between estimators has to be satisfied for all individual estimators, which could lead to problems if there is a bad initial estimator. In practice, a few disagreements are allowed ; for establishing the theoretical results, the proportion of estimators satisfying the agreement condition is required to tend to 1.

Here, we propose a modified procedure mixing these consensus ideas with a more classical method based on the Euclidean distance between entries. This may be seen as

an alternative approach allowing to minder the effect of a possibly bad estimator in the initial list.

We show the consistency of this new strategy and provide some numerical experiments.

Keywords. Classification, regression, estimation, consistency, aggregation.

1 Introduction

Nous proposons d'étudier une stratégie de combinaison d'estimateurs en apprentissage statistique. Notre point de vue est inspiré d'une idée de Mojirsheibani (1999, 2000, 2002a, 2002b), qui suggère d'utiliser une notion de consensus afin de combiner plusieurs classifieurs.

Dans les travaux de Mojirsheibani, tout comme dans une extension récente au contexte de la régression proposée par Biau et al. (2015), une certaine condition de proximité intervenant dans la définition de l'estimateur combiné doit être satisfaite pour tous les estimateurs initiaux, ce qui peut poser problème, notamment en présence d'un estimateur se comportant plutôt mal par rapport aux autres. Pour remédier en pratique à ce problème, la condition de proximité est requise seulement pour une certaine proportion d'estimateurs. Pour l'obtention de résultats théoriques, cette proportion d'estimateurs doit tendre vers 1.

Ici, nous proposons une modification de la procédure, en associant à l'idée de consensus, qui s'est avérée globalement très performante, l'information fournie par les distances entre les entrées. Cette nouvelle version de la méthode peut être vue comme une manière alternative de réduire l'effet d'un éventuel mauvais estimateur dans la liste d'estimateurs initiaux.

2 Notations, définition de l'estimateur

Soit (X, Y) un couple de variables aléatoires à valeurs dans $\mathbb{R}^d \times \mathcal{Y}$, où X a pour distribution μ. Nous nous intéressons à deux situations : la classification binaire, où $\mathcal{Y} = \{0, 1\}$, et la régression bornée, où $\mathcal{Y} = [0, 1]$. Soit η la fonction de régression, définie par $\eta(x) = \mathbb{E}[Y|X = x]$. Notons que $\eta(x) = \mathbb{P}(Y = 1|X = x)$ dans le contexte de la classification binaire.

Soit ψ^* le classifieur de Bayes, donné par

$$\psi^*(x) = \begin{cases} 1 & \text{if } \eta(x) > 1/2 \\ 0 & \text{otherwise.} \end{cases}$$

Ce classifieur ψ^* peut être utilisé comme référence puisqu'il minimise l'erreur de classification $\mathbb{P}(\psi(X) \neq Y)$.

On suppose que l'on dispose d'un échantillon $\mathcal{D}_n = \{(X_1, Y_1) \ldots, (X_n, Y_n)\}$ du couple (X, Y).

En régression, notre but est d'estimer η en utilisant $\mathcal{D}_n$. En classification, il s'agit de construire un classifieur basé sur $\mathcal{D}_n$, dont l'erreur approche celle du classifieur de Bayes.

Pour $k \geq 1$, notons $B_k(x, r)$ la boule fermée centrée en $x \in \mathbb{R}^k$, de rayon $r > 0$. Soit $K : \mathbb{R}^{d+p} \mapsto \mathbb{R}_+$ un noyau, c'est-à-dire une fonction positive décroissante, radiale. On suppose que K est un noyau régulier (voir Devroye *et al*, 1996), c'est-à-dire

- Pour tout z, $K(z) \geq c\mathbf{1}_{B_{d+p}(0,r)}(z)$.
- $\int \sup_{t \in B_{d+p}(z,r)} K(t)dz < \infty$.

Nous proposons de combiner les prédictions de p estimateurs initiaux, notés $c_1, \ldots, c_p$ dans le contexte de la classification, et $r_1, \ldots, r_p$ en régression. Pour $x \in \mathbb{R}^d$, soit $\mathbf{c}(x) = (c_1(x), \ldots, c_p(x))$, $\mathbf{r}(x) = (r_1(x), \ldots, r_p(x))$. Soit $g : \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}_+$ une fonction telle que $g(v_1, v_2) = K(v)$, où $v = (v_1, v_2) \in \mathbb{R}^{d+p}$.

Definition 2.1 *Pour un ensemble d'estimateurs initiaux de la régression $r_1, \ldots, r_p$, l'estimateur combiné $\mathcal{T}_n$ est défini par*

$$\begin{aligned}\mathcal{T}_n(x) &= \frac{\sum_{i=1}^n Y_i K(\frac{Z_i^r - z^r}{\alpha})}{\sum_{i=1}^n K(\frac{Z_i^r - z^r}{\alpha})} \\ &= \frac{\sum_{i=1}^n Y_i g\left(\frac{X_i - x}{\alpha}, \frac{\mathbf{r}(X_i) - \mathbf{r}(x)}{\beta}\right)}{\sum_{i=1}^n g\left(\frac{X_i - x}{\alpha}, \frac{\mathbf{r}(X_i) - \mathbf{r}(x)}{\beta}\right)},\end{aligned}$$

où $Z_i^r = (X_{i1}, \ldots, X_{id}, \frac{\alpha}{\beta} r_1(X_i), \ldots, \frac{\alpha}{\beta} r_p(X_i))$, $i = 1, \ldots, n$, $z^r = (x_1, \ldots, x_d, \frac{\alpha}{\beta} r_1(x) \ldots, \frac{\alpha}{\beta} r_p(x))$.

Soit

$$\mathcal{T}_n^\star = \frac{\sum_{i=1}^n Y_i g\left(\frac{X_i - x}{\alpha}, \frac{\mathbf{r}(X_i) - \mathbf{r}(x)}{\beta}\right)}{n\mathbb{E}\left[g\left(\frac{X - x}{\alpha}, \frac{\mathbf{r}(X) - \mathbf{r}(x)}{\beta}\right)\right]}.$$

Pour alléger les équations, introduisons les notations $x \mapsto K_\alpha(x)$ pour $x \mapsto K(\frac{x}{\alpha})$ et $(v_1, v_2) \mapsto g_{\alpha,\beta}(v_1, v_2)$ pour $(v_1, v_2) \mapsto g(\frac{v_1}{\alpha}, \frac{v_2}{\beta})$.

Definition 2.2 *Pour un ensemble de classifieurs initiaux $c_1, \ldots, c_p$, le classifieur combiné est défini par*

$$\begin{aligned}\mathcal{C}_n(x) &= \begin{cases} 0 & si\ \frac{\sum_{i=1}^n Y_i K_\alpha(Z_i^c - z^c)}{n\mathbb{E}[K_\alpha(Z^c - z^c)]} \leq \frac{\sum_{i=1}^n (1-Y_i) K_\alpha(Z_i^c - z^c)}{n\mathbb{E}[K_\alpha(Z^c - z^c)]} \\ 1 & sinon \end{cases} \\ &= \begin{cases} 0 & si\ \frac{\sum_{i=1}^n Y_i g_{\alpha,\beta}(X_i - x, \mathbf{r}(X_i) - \mathbf{r}(x))}{n\mathbb{E}[g_{\alpha,\beta}(X - x, \mathbf{r}(X) - \mathbf{r}(x))]} \leq \frac{\sum_{i=1}^n (1-Y_i) g_{\alpha,\beta}(X_i - x, \mathbf{r}(X_i) - \mathbf{r}(x))}{n\mathbb{E}[g_{\alpha,\beta}(X - x, \mathbf{r}(X) - \mathbf{r}(x))]} \\ 1 & sinon, \end{cases}\end{aligned}$$

où $Z_i^c = (X_{i1}, \ldots, X_{id}, \frac{\alpha}{\beta} c_1(X_i), \ldots, \frac{\alpha}{\beta} c_p(X_i))$, $i = 1, \ldots, n$, $z^c = (x_1, \ldots, x_d, \frac{\alpha}{\beta} c_1(x) \ldots, \frac{\alpha}{\beta} c_p(x))$.

3 Résultats

La convergence de cette nouvelle stratégie peut être démontrée sous des hypothèses assez générales. En particulier, il n'est pas nécessaire que la liste initiale contienne un estimateur consistant.

Theorem 3.1 (Cas de la régression) *Si $\alpha \to 0$ et $n\alpha^d\beta^p \to \infty$ lorsque $n \to \infty$, alors, pour tout $\varepsilon > 0$, il existe n_0 tel que pour $n \geq n_0$:*

$$\mathbb{P}\left(\int |\eta(x) - \mathcal{T}_n(x)|\mu(dx) > \varepsilon\right) \leq 2\exp\left(-\frac{n\varepsilon^2}{32R^2}\right).$$

Soit L^* l'erreur de Bayes et L_n l'erreur de classification de $\mathcal{C}_n$.

Theorem 3.2 (Cas de la classification) *Si $\alpha \to 0$ et $n\alpha^d\beta^p \to \infty$ lorsque $n \to \infty$, alors, pour tout $\varepsilon > 0$, il existe n_0 tel que pour $n \geq n_0$:*

$$\mathbb{P}(L_n - L^* > \varepsilon) \leq 2\exp\left(-\frac{n\varepsilon^2}{32R^2}\right).$$

La méthode sera illustrée à l'aide de quelques expériences numériques, qui montrent son efficacité et sa flexibilité. Une caractéristique particulièrement appréciable de cette approche est que l'on peut tirer profit de la recherche d'un équilibre entre dimension (intrinsèque) des entrées et dimension des sorties (correspondant au nombre d'estimateurs à combiner).

Bibliographie

[1] Biau, G., Fischer, A., Guedj, B. et Malley, J. (2015). COBRA: A combined regression strategy, Journal of Multivariate Analysis, à paraître (disponible en ligne `http://www.sciencedirect.com/science/article/pii/S0047259X15000950`).

[2] Devroye, L., Györfi, L. et Lugosi, G. (1996) A Probabilistic Theory of Pattern Recognition. Applications of Mathematics. Springer, New York.

[3] Mojirsheibani, M. (1999). Combining classifiers via discretization, Journal of the American Statistical Association, 94, 600-609.

[4] Mojirsheibani, M. (2000). A kernel-based combined classification rule, Statistics & Probability Letters, 48, 411-419.

[5] Mojirsheibani, M. (2002). An almost surely optimal combined classification rule, Journal of Multivariate Analysis, 81, 28-46.

[6] Mojirsheibani, M. (2002). A comparison study of some combined classifiers, Communications in Statistics - Simulation and Computation, 31, 245-260.