AdLearn: Un algorithme d'apprentissage interprétable

Vincent Margot [1,2]

[1] *LSTA (Bureau 15-25 205), Université Pierre et Marie Curie 4 place Jussieu, 75005 Paris, FRANCE*

[2] *Advestis, 8 rue Vernier 75017 Paris, FRANCE,*
vincent.margot@advestis-conseil.com

Résumé. L'algorithme que nous présentons est un algorithme d'apprentissage développé pour être facilement interprétable. Il est basé sur des objets de la forme "*If ... Then ...*", appelés experts. L'algorithme identifie un ensemble d'experts par la méthode du minimum de contraste. Cet ensemble est ensuite agrégé via l'utilisation des méthodes d'agrégation d'experts, nous fournissant ainsi un prédicteur dont les performances sont comparables à celle de la meilleure combinaison convexe. La construction de ce prédicteur nous permet aussi de l'exprimer comme un estimateur de la fonction de régression.

Mots-clés. Apprentissage et classification, Modèles non-paramétriques, Statistique mathématique

Abstract. This presentation introduces a learning algorithm developed in a way that is easily interpretable. This one is based on objects of the form "*If ... Then ...*", called experts. The algorithm identifies a set of experts by the minimum contrast method. This set is then aggregated by the use of experts aggregation methods, giving us a predictor whose performance is comparable to that of the best convex combination. The construction of this predictor also allows us to express it as an estimator of the regression function.

Keywords. Machine learning and classification, Nonparametric models, Mathematical statistics

1 Introduction

L'algorithme que nous présentons est né d'une problématique d'entreprise. Comment créer un algorithme d'apprentissage adapté à tous types de données (qualitatives et quantitatives), déterministe et surtout interprétable par tous? L'interprétabilité désigne ici la facilité de comprendre la logique conduisant à la prédiction. Dans cette optique nous avons choisi de baser notre algorithme sur un ensemble de prédicteurs simples, nommés experts, de la forme "*If ... Then ...*".

Cette idée n'est pas nouvelle, on la retrouve par exemple dans les algorithmes RuleFit de Friedman et Popescu (2008), RegEnder de Dembczyński, Kotłowski et Słowiński (2008) ou plus communément dans les modèles CART Breiman et Friedman (1984).

L'originalité repose sur la méthode d'identification de ces experts et sur la façon de les combiner entre eux pour obtenir un prédicteur unique. De plus l'algorithme nous permet de travailler à la fois dans un problème de prédiction et dans un problème d'estimation. Nous présentons, dans l'exposé, des résultats encourageants sur des données simulées.

1.1 Cadre théorique

Soient $\mathcal{X}$ un espace de dimension $d \geq 1$ et (X, Y) un couple de variables aléatoires à valeur dans $\mathcal{X} \times \mathbb{R}$ de loi P inconnue. On observe un échantillon $D_T = ((X_1, Y_1), \ldots, (X_T, Y_T))$, où les couples (X_s, Y_s) sont supposés i.i.d et l'on veut prédire Y conditionnellement à X. Pour cela nous construisons une fonction $\hat{g}$, appelée prédicteur, telle que

$$\begin{array}{ccccccc} \hat{g} & : & (\mathcal{X} \times \mathbb{R})^T & \times & \mathcal{X} & \to & \mathbb{R} \\ & & D_T & ; & X & \mapsto & \hat{g}(D_T, X). \end{array} \tag{1}$$

Sauf ambiguïté possible, nous noterons $\hat{g}_T(X) := \hat{g}(D_T, X)$. On note, également, $\mathbb{G} = \{f \mid f : \mathcal{X} \to \mathbb{R}\}$.

Afin de pouvoir évaluer la proximité de $\hat{g}_T(X)$ par rapport à Y nous introduisons une **fonction de perte** notée Ł, comme présentée dans (Arlot et Celisse (2010)). C'est une fonction de $\mathbb{G}$ dans $\mathbb{R}$ définie par:

$$Ł\left(\hat{g}_T\right) = \mathbb{E}_{(X,Y)\sim P}\left[\gamma\left(\hat{g}_T; (X, Y)\right)\right], \tag{2}$$

avec γ une **fonction de contraste** telle que $\gamma : \mathbb{G} \times (\mathcal{X} \times \mathbb{R}) \to \mathbb{R}_+$. Nous voulons que $\hat{g}_T$ ait une perte la plus petite possible.

Définition 1.1 *On dira que $g^* \in \mathbb{G}$ est un prédicteur optimal s'il vérifie:*

$$L(g^*) = \inf_{g \in \mathbb{G}} L\left(g\right). \tag{3}$$

2 AdLearn

A partir du résultat montrant qu'en utilisant la fonction de contraste quadratique, $\gamma_q\left(\hat{g}_T; (x, y)\right) = (\hat{g}_T(x) - y)^2$, le prédicteur optimal est la fonction de régression, $g^*(x) = \mathbb{E}\left[Y|X = x\right]$ (exhibant ainsi la correspondance entre prédiction et estimation), l'objectif est de construire un estimateur consistant de g^*.

2.1 Les définitions

Définition 2.1 *Soit $k_i \subset \mathcal{X}$. Un expert f_i est une fonction définie sur $k_i \times (\mathcal{X} \times \mathbb{R})^T$, constante en son premier argument et qui vaut l'espérance conditionnelle empirique de Y sachant $X \in k_i$, c'est-à-dire*

$$f_i(x, D_T) = \frac{\sum_{s=1}^{T} y_s \mathbf{1}_{x_s \in k_i}}{\sum_{s=1}^{T} \mathbf{1}_{x_s \in k_i}}.$$

Notation 1
$\forall x \in k_i$, $\mu_i := f_i(x, D_T)$.
On note $\mathcal{F}$ l'ensemble de tous les experts au sens de la définition 2.1.
L'événement $\{X \in k_i\}$ est appelé la ***condition d'activation*** *de l'expert f_i.*

On définit alors deux opérations sur les experts.

Définition 2.2 *Soient deux experts f_i et f_j.*

- *L'intersection de f_i et f_j, notée $f_i \wedge f_j$, est l'expert défini sur $k_i \cap k_j$.*
- *L'union de f_i et f_j, notée $f_i \vee f_j$ est l'expert défini par*

$$\begin{array}{rcccccc} f_i \vee f_j & : & k_i \cup k_j & \times & (\mathcal{X} \times \mathbb{R})^t & \to & \mathbb{R} \\ & & x & ; & D_T & \mapsto & \dfrac{\pi_i \mu_i \mathbf{1}_{\{x \in k_i\}} + \pi_j \mu_j \mathbf{1}_{\{x \in k_j\}}}{\pi_i \mathbf{1}_{\{x \in k_i\}} + \pi_j \mathbf{1}_{\{x \in k_j\}}}, \end{array} \tag{4}$$

avec π_i est le poids associé à l'expert f_i. On a $\pi_k > 0, k \in \{i, j\}$ et $\pi_i + \pi_j = 1$

L'union est un peu plus délicate puisque nos experts peuvent avoir des ensembles de définition d'intersection non vide. Pour une union de M experts, on obtient l'expression suivante:

$$\vee_{i=1}^{M} f_i(x, D_T) = \sum_{i=1}^{M} \frac{\mathbf{1}_{\{x \in k_i\}} \pi_i \mu_i}{\sum_{j=1}^{M} \mathbf{1}_{\{x \in k_j\}} \pi_j}, \tag{5}$$

avec $\sum_{i=1}^{M} \pi_i = 1$ et $\pi_i > 0$, $\forall i \in \{1, \ldots, M\}$.

2.2 Les différentes formes du prédicteur

Dans cette représentation de l'union (5) nous introduisons la notion d'agrégation d'experts via les π_i. En effet, ces poids peuvent être calculés via une stratégie d'agrégation $\mathcal{S}$ (voir Cesa-Bianchi et Lugosi (2006) ou Stoltz (2010) pour une présentation des stratégies usuelles). Cela nous permet de construire un estimateur comme l'union d'un ensemble de

M experts. Soient $x \in \mathcal{X}$, $\mathcal{S}$ une stratégie d'agrégation et $\mathbf{f}^* \subset \mathcal{F}$ tel que $\mathbf{f}^* = \{f_i, i \in \{1, \ldots, M\}\}$ nous avons alors:

$$\hat{g}_T^{\{\mathbf{f}^*, \mathcal{S}\}}(x) = \vee_{i=1}^{M} f_i(x, D_T). \tag{6}$$

Le sous-ensemble d'experts $\mathbf{f}^*$ est identifié par minimum de contraste (Vapnik et Kotz (1982)), c'est-à-dire qu'il vérifie:

$$\sum_{s=1}^{T} \gamma_q \left(\hat{g}_T^{\{\mathbf{f}^*, \mathcal{S}\}}; (x_s, y_s) \right) := \min_{\mathbf{f} \subset \mathcal{F}} \frac{1}{t} \sum_{s=1}^{T} \gamma_q \left(\hat{g}_T^{\{\mathbf{f}, \mathcal{S}\}}; (x_s, y_s) \right), \tag{7}$$

en prenant le contraste quadratique.

A partir de (6) nous pouvons écrire notre prédicteur $\hat{g}_T^{\{\mathbf{f}^*, \mathcal{S}\}}$, sous la forme d'un prédicteur issu d'une agrégation d'experts via une stratégie $\mathcal{S}$.

$$\hat{g}_T^{\{\mathbf{f}^*, \mathcal{S}\}}(x) = \sum_{i=1}^{M} \pi_i(x) f_i(x, D_T), \tag{8}$$

avec $\pi_i(x) = \dfrac{\mathbf{1}_{\{x \in k_i\}} \pi_i}{\sum_{j=1}^{M} \mathbf{1}_{\{x \in k_j\}} \pi_j}$.

Le predicteur (8) peut aussi s'écrire sous la forme d'une combinaison linéaire des Y_s. En effet en remplaçant les $f_i(x, D_T)$ par leurs expressions (def 2.1) et en réarrangeant les termes nous pouvons l'écrire sous la forme d'un estimateur linéaire de la fonction de régression.

$$\hat{g}_T^{\{\mathbf{f}^*, \mathcal{S}\}}(x) = \sum_{i=1}^{M} \pi_i(x) \sum_{s=1}^{T} \frac{y_s \mathbf{1}_{x_s \in k_i}}{\sum_{u=1}^{T} \mathbf{1}_{x_u \in k_i}} = \sum_{s=1}^{T} w_s(x) y_s, \tag{9}$$

avec w_s défini par:

$$w_s(x) = \sum_{j=1}^{M} \pi_j(x) \frac{\mathbf{1}_{x_s \in k_j}}{\sum_{u=1}^{T} \mathbf{1}_{x_u \in k_j}}.$$

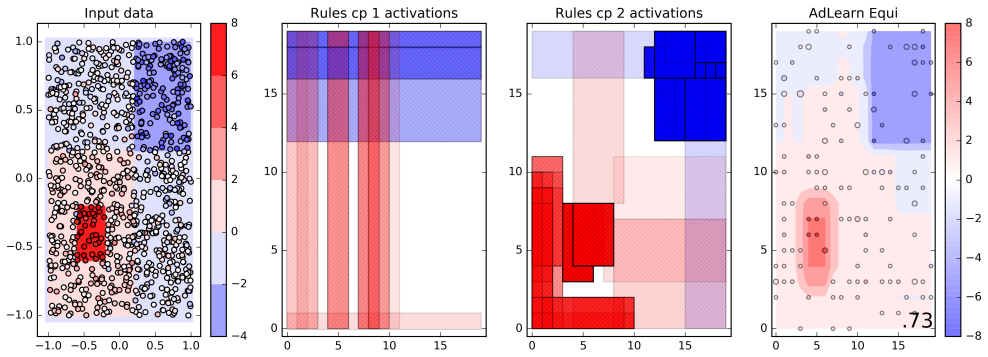
Les deux formes (8) et (9) pour $\hat{g}_T^{\{\mathbf{f}^*, \mathcal{S}\}}$, nous permettent d'évoluer dans le domaine de l'agrégation d'experts et dans celui de l'estimation non paramétrique respectivement. L'objectif est donc d'obtenir des résultats théoriques dans ces deux domaines pour $\hat{g}_T^{\{\mathbf{f}^*, \mathcal{S}\}}$. Nous aurons alors à la fois un estimateur de la fonction de régression ainsi qu'un prédicteur avec des performances comparables à celles de la meilleure combinaison convexe des experts de $\mathbf{f}^*$.

3 Illustration

Nous ne présentons ici qu'une seule illustration, mais d'autre seront présentées lors de la présentation. L'illustration suivante se compose de trois éléments[1].

- Le premier est la représentation graphique des données simulées, Y, en fonction des deux variables explicatives réparties uniformément sur $[-1, 1]$. Les données sont découpées de la façon suivante: 700 points pour l'apprentissage et 300 pour le test.
- Le deuxième élément (les deux graphiques du milieu) est la représention graphique des zones d'activations (i.e. les k_i) des experts dont les conditions portent sur une (cp 1) et deux (cp 2) variables explicatives.
- Enfin le dernier élément représente les prédictions de l'algorithme pour des données hors apprentissage. Le score (en bas à droite du dernier graphique) utilisé est le R^2 score.

Nous avons effectué plusieurs simulations avec $\mathcal{X} = [-1, 1]^2$ et Y suit une loi normale réduite et de moyenne -4 (zone bleue foncée), 0 (zone bleue), 1 (zone rouge claire) et 8 (zone rouge foncée). Pour éviter le sur-apprentissage et limiter le temps de calcul, les variables explicatives, $X1$ et $X2$, sont discrétisées en 20 modalités correspondantes à leurs 20-quantiles empiriques sur la période d'apprentissage.



Pour chaque $x \in \mathcal{X}$ nous connaissons les experts actifs (i.e leur condition d'activation est vérifiée), il suffit alors de combiner les prédictions de chaque experts actifs. Dans l'exemple la combinaison utilisée est simplement la moyenne. Pour la prediction nous avons utilisé une méthode equipondérée, c'est-à-dire que $\pi_i = 1/M$, pour tout $i \in \{1, \ldots, M\}$ dans (5).

Les premières conclusions sont qu'AdLearn identifie bien les différentes zones de concentration. De plus ses capacités prédictives sur ces échantillons simulés sont satisfaisantes.

[1]Sur les trois derniers graphiques les axes prennent pour valeurs les modalités de discrétisation de la variable.

3.1 Quid de l'interprétation

Chaque apprentissage donne en sortie un tableau d'experts, ce qui permet d'interpréter facilement les résultats.

Expert Name	Vars	Bmin	Bmax	Coverage	Prediction	MSE
Expert 1	X_2	0.20	1	0.40	-0.65	5.31
Expert 2	X_1	-0.24	-0.03	0.10	0.40	6.30
...						
Expert 20	X_2 & X_1	0.44 & 0.44	0.91 & 0.91	0.05	-1.63	5.49
Expert 21	X_1 & X_2	-0.68 & 0.68	-0.15 & -0.23	0.06	2.02	4.81

Exemple:
Soit $x = (x_1, x_2) \in \mathcal{X}$. Le premier expert nous informe que si $x_2 \in [0.20, 1]$, alors Y vaut en moyenne -0.65 sur l'ensemble d'apprentissage. Nous avons donc, $f_1(x) = -0.65$, si $x_2 \in [0.20, 1]$.

Bibliographie

[1] Breiman, L. et Friedman, J. et Olshen,R. et Stone, C. (1984), *Classification and Regression Trees.*, CRC press

[2] Hastie, T. et Friedman, J. et Tibshirani, R. (2003), *The elements of statistical learning*, Springer series in statistics Springer

[3] Cesa-Bianchi, N. et Lugosi, G. (2006), *Prediction, Learning and Games*, Cambridge university press

[4] Györfi, L. et Kohler, M. et Krzyzak, A. et Walk, H (2006), *A distribution-free theory of nonparametric regression*, Springer Science & Business Media

[5] Friedman, J. et Popescu, B. (2008), Predective learning via rule ensembles *The Annals of Applied Statistics*, 916–954.

[6] Dembczyński, K. and Kotłowski, W. and Słowiński, R. (2008), Solving Regression by Learning an Ensemble of Decision Rules, *International Conference on Artificial Intelligence and Soft Computing*, 533–544.

[7] Arlot, S. et Celisse, A. (2010), A survey of cross-validation procedures for model selection, *Statistics Surveys*, 4, 40–79.

[8] Stoltz, G. (2010), Agrégation séquentielle de prédicteurs: méthodologie générale et applications à la prévision de la qualité de l'air et celle de la consommation électrique, *Journal de la Société Française de Statistique.*, 151.2, 66-106.