Sélection de variables avec CART

Marie Sauvé & Christine Tuleau

Laboratoire de Mathématique – U.M.R. C 8628, "Probabilités, Statistique et Modélisation", Université Paris-Sud, Bât. 425, 91405 Orsay Cedex, France
marie.sauve@math.u-psud.fr & christine.tuleau@math.u-psud.fr

Abstract
Our theorical work deals with the problem of variables selection with the CART algorithm in the gaussian regression and classification frameworks. This algorithm provides a "good" piecewise constant predictor, but does not really give the most influential explanatory variables. To solve this issue, we applied the algorithm to a family of subsets of variables. Then we considered model selection via penalization and found a penalty function well adapted to our problem.

Key words: CART, model selection, variables selection, penalization, regression, classification

Résumé
Notre travail, de nature théorique, traite du problème de la sélection de variables au moyen de l'algorithme CART dans le cadre de la régression gaussienne et de la classification binaire. Cet algorithme fournit un "bon"prédicteur constant par morceaux, sans réellement donner les variables explicatives les plus influentes. Afin de résoudre ce problème, nous avons appliqué l'algorithme à une famille de paquets de variables explicatives, puis nous avons opté pour le point de vue de la sélection de modèle par minimisation de contrastes pénalisés, et déterminé une forme de pénalité adaptée à notre problème.

Mots clés: CART, sélection de modèle, sélection de variables, pénalisation, régression, classification

1 Introduction

Notre travail, de nature théorique, porte sur le problème de la sélection de variables. De nombreux travaux ont été menés dans ce sens, comme le prouve le livre de Hastie *et al.* (2001) qui expose différentes méthodes de sélection de variables, tant dans le cas linéaire (LDA, LARS) que dans celui non-linéaire. Nous avons envisagé le problème sous ce dernier angle, en considérant l'algorithme CART (Classification And Regression Trees), proposé par Breiman *et al.* (1984), comme support de notre étude.
Cet algorithme construit un arbre et un prédicteur par histogrammes à partir d'un échantillon de couples dont le premier terme est un vecteur de variables explicatives et le deuxième une variable réponse. En lui-même, cet algorithme ne procède pas à une réelle sélection de variables, même s'il existe des outils pour y remédier en partie, comme l'importance des variables, proposée par Breiman *et al.* et étudiée par Ghattas (1999).

Nous avons alors envisagé d'appliquer l'algorithme CART à une famille de paquets de variables explicatives et d'adopter un point de vue sélection de modèle, en minimisant un critère pénalisé. Tant dans le cadre de la classification que de la régression gaussienne, nous avons déterminé, par le biais d'inégalités oracles, des pénalités adaptées à notre problème. Les résultats étant similaires, nous présenterons uniquement le cas de la régression.
En recourant à CART et en transformant notre problème de sélection de variables en un problème de sélection de modèle, notre travail est ainsi à rapprocher de ceux menés par Birgé et Massart (2001), Gey et Nédélec, et Massart et Nédélec.

Après un petit rappel sur CART, nous expliquerons plus précisément notre procédure et notre objectif, puis nous présenterons certains des résultats obtenus.

2 Un bref aperçu de CART

CART est un outil rapide, facilement interprétable et fréquemment utilisé dans la résolution de problèmes de classification et de régression. Son principe est de diviser l'espace des observations de façon récursive et de définir un prédicteur constant par morceaux sur la partition ainsi obtenue, à partir d'un échantillon d'apprentissage. Pour ce faire, CART procède en trois étapes successives que sont la construction d'un arbre maximal, son élagage et une phase de sélection finale.
La première étape consiste à diviser l'espace des observations $\mathcal{X}$ au moyen de découpes successives parallèles aux axes des variables explicatives, de manière à regrouper au "mieux" les observations de même réponse à chacune des divisions, jusqu'à n'avoir qu'une seule observation par région d'espace. Une représentation de cette succession de découpes s'effectue à l'aide d'un arbre de profondeur maximale, encore appelé arbre maximal.
La phase d'élagage sert à extraire une suite de sous-arbres emboîtés correspondant aux "meilleurs" sous-arbres de quelques unes des tailles intermédiaires. Ceci est réalisé à l'aide d'un critère pénalisé qui effectue un compromis entre l'adéquation aux données et la taille du modèle.
Enfin, la dernière étape vise à sélectionner un arbre parmi la sous-suite obtenue précédemment, soit par échantillon test, soit par validation croisée.

Remarque 1:
Le critère pénalisé, intervenant dans la phase d'élagage et proposé par Breiman *et al.*, se décompose en deux parties :

- un contraste empirique qui permet une quantification de l'ajustement aux données,

- une pénalité proportionnelle à la complexité du modèle mesurée par le nombre de feuilles de l'arbre servant de support au modèle. Ainsi, si T est un arbre et $\mathcal{S}_T$ le modèle associé, à savoir le sous-espace linéaire de $\mathbb{L}^2(\mathcal{X})$ formé des fonctions constantes par morceaux définies sur les feuilles de T, la pénalité est $\alpha\frac{|T|}{n}$ où α est une constante positive, n le nombre d'observations et $|T|$ le nombre de feuilles de T.

Ce terme de pénalité a été justifié par les récents travaux de Gey et Nédélec pour la régression gaussienne ou bornée dans deux situations précises, que nous avons reprises au cours de notre étude. Il s'agit des situations (M1) et (M2) suivantes:

- (M1) : l'échantillon d'apprentissage est scindé en 3 sous-échantillons indépendants $\mathcal{L}_1$, $\mathcal{L}_2$ et $\mathcal{L}_3$ de cardinal respectif n_1, n_2, n_3, où $\mathcal{L}_1$ sert à la construction de l'arbre maximal, $\mathcal{L}_2$ à son élagage et $\mathcal{L}_3$ à la sélection finale du sous-arbre,

- (M2) : l'échantillon d'apprentissage est scindé en 2 sous-échantillons indépendants $\mathcal{L}_1 = \mathcal{L}_2$ et $\mathcal{L}_3$, où le même échantillon $\mathcal{L}_1 = \mathcal{L}_2$ sert à la construction de l'arbre maximal et à son élagage et $\mathcal{L}_3$ à la sélection finale du sous-arbre.

3 La procédure

On observe n copies indépendantes $(X_1, Y_1), ..., (X_n, Y_n)$ d'un couple de variables aléatoires (X, Y) suivant le modèle :

$$Y_i = s(X_i) + \varepsilon_i, \quad 1 \leq i \leq n$$

où $\varepsilon_1, ..., \varepsilon_n$ sont indépendants et pour tout $i \in \{1, ..., n\}$, ε_i suit une loi $\mathcal{N}(0, \sigma^2)$ conditionnellement à X_i, s est la fonction de régression à estimer, et $X = (X^1, .., X^p)$.

Disposant de données d'apprentissage, nous considérons l'une des situations (M1) ou (M2) et nous appliquons l'algorithme CART à tous les paquets de variables explicatives, autrement dit à tout $M \in \mathcal{P}(\Lambda) = \mathcal{P}(\{X^1, ..., X^p\})$.
Ainsi, pour tout $M \in \mathcal{P}(\Lambda)$, on construit un arbre maximal $T_{max}^{(M)}$, en n'autorisant des découpes que sur les variables explicatives du paquet M.
Pour tout sous-arbre T de $T_{max}^{(M)}$, ce que l'on note encore $T \preceq T_{max}^{(M)}$, on définit $\mathcal{S}_{M,T}$ comme étant le sous-espace de $\mathbb{L}^2(\mathbb{R}^p)$ constitué des fonctions constantes sur chaque feuille de T, et $\hat{s}_{M,T} = \underset{u \in \mathcal{S}_{M,T}}{argmin}\ \gamma_{n_2}(u)$, où γ_{n_2} est le contraste empirique :

$$\gamma_{n_2}(u) = \frac{1}{n_2} \sum_{(X_i, Y_i) \in \mathcal{L}_2} (Y_i - u(X_i))^2$$

On dispose alors d'une collection d'estimateurs :

$$\{\hat{s}_{M,T};\ M \in \mathcal{P}(\Lambda)\ et\ T \preceq_{max}^{(M)}\}$$

Notre objectif consiste à trouver une fonction de pénalité $pen(M,T)$ telle que, si l'on définit :

- $\left(\hat{M}, \hat{T}_{\hat{M}}\right) = \underset{(M,T)}{argmin} \{\gamma_{n_2}(\hat{s}_{M,T}) + pen(M,T)\}$,

- $\tilde{s} = \hat{s}_{\hat{M},\hat{T}_{\hat{M}}}$,

le modèle $\left(\hat{M}, \hat{T}_{\hat{M}}\right)$ associé à $\tilde{s}$ est proche du modèle optimal, à savoir :

$$\mathbb{E}\left(\|s-\tilde{s}\|^2 \ |\mathcal{L}_1\right) \leq C \underset{(M,T)}{inf} \left\{\mathbb{E}\left(\|s-\hat{s}_{M,T}\|^2 \ |\mathcal{L}_1\right)\right\}$$

avec C le plus proche de 1 possible.

Remarque 2:
De part cette approche, on modifie le problème de sélection de variables en un problème de sélection de modèle dans lequel nous devons trouver une fonction de pénalité prenant en compte la complexité des modèles. (cf. Birgé et Massart).

4 Un résultat typique

Proposition 1: Dans le cas (M1):
Considérons une pénalité de la forme:

$\forall\ M \in \mathcal{P}(\Lambda)$ et $\forall\ T \preceq T_{max}^{(M)}$,

$$pen(M,T) = \alpha\frac{|T|}{n_2} + \beta\frac{|M|}{n_2}\left(1 + log\left(\frac{p}{|M|}\right)\right)$$

Si $\alpha > 2\sigma^2(1+\sqrt{2log(2)}+2log(2))$ et $\beta > 3\sigma^2$,
alors il existe 2 constantes positives C_1 et C_2 qui ne dépendent que de α et β telles que:

$$\mathbb{E}\left(\|\tilde{s}-s\|_{n_2}^2 \ |\mathcal{L}_1\right) \leq C_1 \underset{(M,T)}{inf}\left\{\underset{u\in S_{M,T}}{inf}\ \mathbb{E}\left(\|s-u\|_{n_2}^2 \ |\mathcal{L}_1\right) + pen(M,T)\right\} + C_2\frac{\sigma^2}{n_2}$$

où $\|\ .\ \|_{n_2}$ désigne la norme empirique sur $\{X_i;\ (X_i,Y_i)\in\mathcal{L}_2\}$.

Remarque 3:

- En observant la forme de la fonction de pénalité, on constate deux choses. La première partie de cette fonction, à savoir $\alpha\frac{|T|}{n_2}$ correspond au terme de pénalité proposé par Breiman *et al.* (1984) et justifié par Gey et Nédélec lorsque le paquet M est fixé. L'autre partie, quant à elle, provient de la sélection de variables. Elle sert à pénaliser les modèles qui se basent sur un paquet comprenant beaucoup de variables explicatives.

- La forme de la pénalité permet de décomposer le problème en deux étapes, la première étant le sélection de $\hat{T}_M$ par $\hat{T}_M = \underset{T \preceq T_{max}^{(M)}}{argmin} \left\{ \gamma_{n_2}(\hat{s}_{M,T}) + \alpha \frac{|T|}{n_2} \right\}$ et la deuxième la sélection du paquet $\hat{M}$ par $\hat{M} = \underset{M \in \mathcal{P}(\Lambda)}{argmin} \left\{ \gamma_{n_2}(\hat{s}_{M,\hat{T}_M}) + pen(M, \hat{T}_M) \right\}$

- Considérer la situation (M1) facilite la théorie puisque pour déterminer la fonction de pénalité, on peut travailler conditionnellement à la construction de l'arbre maximal, autrement dit, conditionnellement à $\mathcal{L}_1$. De ce fait, la collection de modèles considérée $\left(\{ \mathcal{S}_{M,T};\ M \in \mathcal{P}(\Lambda)\ et\ T \preceq_{max}^{(M)} \} \right)$ est déterministe.

Proposition 2: Dans le cas (M2):
Considérons une pénalité de la forme:

$\forall\ M \in \mathcal{P}(\Lambda)$ et $\forall\ T \preceq T_{max}^{(M)}$,

$$pen(M,T) = \alpha L(n_1, M) \frac{|T|}{n_1} + \beta \frac{|M|}{n_1} \left(1 + log \left(\frac{p}{|M|} \right) \right)$$

où $L(n_1, M) = 1 + \sqrt{(|M|+1) \left(1 + log \left(\frac{n_1}{|M|+1} \right) \right)} + (|M|+1) \left(1 + log \left(\frac{n_1}{|M|+1} \right) \right)$

Si $\alpha > 2\sigma^2$ et $\beta > 3\sigma^2$, alors il existe 3 constantes positives C_1, C_2 et Σ qui ne dépendent que de α et β telles que:
$\forall\ \xi > 0$, avec probabilité $\geq 1 - e^{-\xi}\Sigma$,

$$\|\tilde{s} - s\|_{n_1}^2 \leq C_1 \underset{(M,T)}{inf} \left\{ d_1(s, S_{M,T})^2 + pen(M,T) \right\} + C_2 \frac{\sigma^2}{n_1} \xi$$

où $\|\ .\ \|_{n_1}$ désigne la norme empirique sur $\{X_i;\ (X_i, Y_i) \in \mathcal{L}_1\}$ et d_1 la distance correspondante.

Remarque 4:

- Dans la méthode (M2), il n'est plus possible de travailler conditionnellement à $\mathcal{L}_1$. Pour remédier au problème, on considère une sur-famille de modèles.

- Dans le cadre de la classification binaire, c'est à dire lorsque $Y \in \{0, 1\}$, on obtient des résultats analogues dans chacune des deux situations (M1) et (M2).

5 Conclusion

Si la procédure que nous venons de décrire se justifie théoriquement, elle est malheureusement impraticable par son aspect combinatoire. En effet, si l'on prend en compte p variables explicatives, il y a 2^p paquets sur lesquels nous devons appliquer l'algorithme CART.
Une solution consiste à appliquer la procédure à un sous-ensemble bien choisi de paquets, en s'inspirant de l'idée mise en œuvre, dans un contexte voisin, par Poggi et Tuleau (2004). Cela consiste à appliquer CART à l'ensemble des variables explicatives puis à calculer l'importance des variables relativement à l'arbre maximal ainsi construit. A partir de cet ordonnancement des variables explicatives, nous pouvons alors considérer le sous-ensemble de paquets formé en associant, pas à pas, les variables explicatives selon cet ordre prédéfini. Ceci permet de passer d'une famille à 2^p paquets à un ensemble n'en comptant plus que p.

En parallèle à ce travail appliqué, nous voulons aussi poursuivre le travail théorique.
En effet, pour commencer, nous souhaiterions étendre ce travail au cadre de la régression où l'on considère des erreurs à transformée de Laplace finie au voisinage de zéro.
Ensuite, au moyen de simulations (avec p petit), nous aimerions calibrer les constantes qui apparaissent dans la fonction de pénalisation.

Bibliographie

[1] Birgé, L. et Massart, P. (2001) A generalized C_p criterion for Gaussian model selection. *Prépublication 647, Universités de Paris 6 & Paris 7.*
[2] Breiman, L., Friedman, J., Olshen, R. et Stone, C. (1984) *Classification And Regression Trees*, Chapman et Hall.
[3] Gey, S. et Nédélec, E. (à paraître) Model Selection for CART Regression Trees. *IEEE Trans. on Inf. Theory.*
[4] Ghattas, B. (1999) Importance des variables dans les méthodes CART. *Revue de Modulad*, 24, 29–39.
[5] Hastie, T., Tibshirani, R. et Friedman, J. (2001) *The Elements of Statistical Learning*, Springer.
[6] Massart, P. et Nédélec, E. (soumis) Risk bounds for statistical learning.
[7] Poggi, J.M. et Tuleau, C. (2004) Classification supervisée en grande dimension, Application à l'agrément de conduite automobile *Résumé de la communication aux journées de la Statistique à Montpellier*