Sujet Impact : mélange de régressions logistiques

Adrien Ehrhardt, Christophe Biernacki, Philippe Heinrich, Vincent Vandewalle
Inria, Crédit Agricole Consumer Finance
first.last@inria.fr

Ce sujet s'inscrit dans le cadre d'une thèse CIFRE (ex-DAD) entre Inria et Crédit Agricole Consumer Finance.

Classification supervisée avec données mixtes En matières de crédit à la consommation, les instituts financiers cherchent à automatiser la décision de financement tout en ne sélectionnant que les clients susceptibles de rembourser ledit crédit. Depuis une quarantaine d'années, le *Credit Scoring* consiste à construire des modèles de classification supervisée $p_{\boldsymbol{\theta}}$ à partir des données demandées au clients $\boldsymbol{x} = (x_j)_1^d$ et de l'observation du remboursement des clients passés $y \in \{0, 1\}$. Historiquement, des scores différents sont développés sur des marchés (e.g. grande distribution, électroménager, ...) et/ou des produits (e.g. renouvelable, amortissable, ...) et/ou des partenaires et/ou des profils clients différents dans l'esprit de la figure 1. Ce découpage est historique et relève d'un a priori. On cherche ici à rationnaliser cette pratique en considérant le *cluster* d'appartenance du client comme un paramètre à optimiser. Si l'on note K le nombre de scores à construire (inconnu) et $c = 1..K$ chaque score, correspondant à un *cluster* de clients, le mélange de régressions logistiques s'écrit :

$$p(y|\boldsymbol{x}) = \sum_{c=1}^{K} p_{\boldsymbol{\theta}_c}(y|\boldsymbol{x}, c)p(c|\boldsymbol{x}),$$

où l'on restreint $p(c|\boldsymbol{x})$ à prendre la forme de la figure 1, de telle sorte que le mélange n'est pas "flou" comme pour un modèle de mélange classique où la contribution de chaque classe est pondérée par sa probabilité. La difficulté d'une approche directe réside dans cette contrainte discrète.

Solutions existantes À notre connaissance, il existe trois approches : les algorithmes LOTUS [Chan and Loh, 2004], LMT [Landwehr et al., 2005] et MOB [Zeileis et al., 2008] qui ont tous trois des avantages et des inconvénients ; en particulier, aucune de ces méthodes ne maximise un critère de choix de modèle explicite sur $p(y|\boldsymbol{x})$.

Proposition d'un modèle On peut voir c comme une variable latente, ce qui, dans la littérature statistique, fait immédiatement penser à la mise en oeuvre d'un algorithme de type EM. En particulier, on peut tirer de bons candidats en remarquant que :

$$p(c|\boldsymbol{x}; y) \propto p_{\boldsymbol{\theta}_c}(y|\boldsymbol{x}; c)p(c|\boldsymbol{x}).$$

Ce tirage a été mis en oeuvre dans une preuve de concept sur données simulées et fonctionne bien.

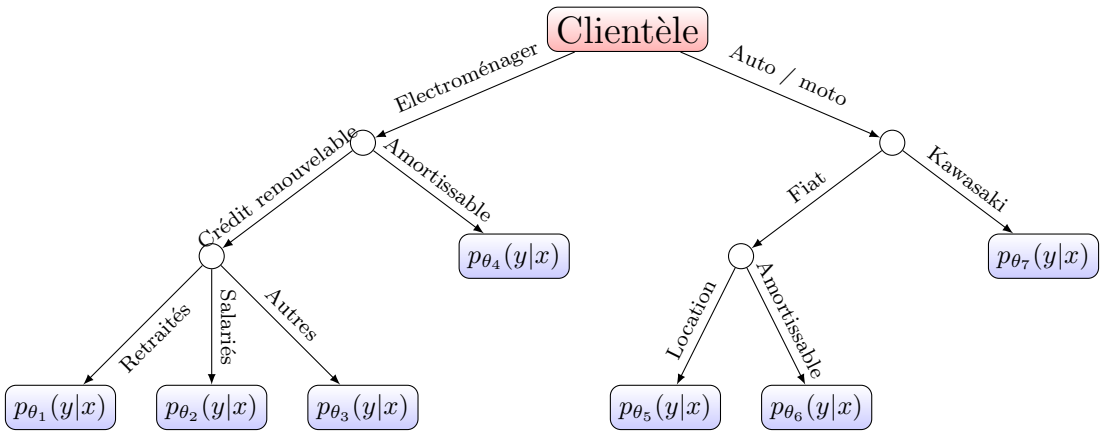

FIGURE 1 – Cartographie simplifiée de la chaîne de construction des scores.

Contributions de l'Impact Deux directions d'étude peuvent être envisagés : la première consiste à étendre la preuve de concept en intégrant par exemple la sélection de variables, voire la discrétisation (voir autre sujet Impact proposé) sur chacun des scores de l'arbre et en testant la tractabilité de cette approche sur des données réelles. La seconde direction possible est le test des trois algorithmes susmentionnés : on commencera par exemple par l'algorithme MOB dont l'implémentation R semble simple à mettre en oeuvre et peut permettre l'intégration "facile" de modèles $p_{\boldsymbol{\theta}_c}$ à la main de l'utilisateur (pour faire de la sélection de variables par exemple).

Références

[Chan and Loh, 2004] Chan, K.-Y. and Loh, W.-Y. (2004). Lotus : An algorithm for building accurate and comprehensible logistic regression trees. *Journal of Computational and Graphical Statistics*, 13(4) :826–852.

[Landwehr et al., 2005] Landwehr, N., Hall, M., and Frank, E. (2005). Logistic model trees. *Machine learning*, 59(1-2) :161–205.

[Zeileis et al., 2008] Zeileis, A., Hothorn, T., and Hornik, K. (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, 17(2) :492–514.