Apprentissage de la fonction quantile conditionnelle par partitionnement 2D

Carine Hue & Marc Boullé

France Télécom R & D; 2, avenue Pierre Marzin; 22307 Lannion cedex

Résumé : Dans cet article, on propose un estimateur de la fonction quantile conditionnelle dans le cas où la variable à expliquer est ordinale. Pour cela, une famille de modèles de partitionnement 2D de l'espace (variable à expliquer x variable explicative) est définie et munie d'une distribution *a priori* et d'une fonction de vraisemblance. Le nombre et la taille de chaque partie sont optimisés conjointement pour chacune des variables selon une approche MAP. On déduit de la partition optimale un estimateur constant par morceaux de la fonction quantile conditionnelle.
Summary : In this article, we propose a conditional quantile function estimator for an ordinal response. For that, a 2D partitioning model family is defined and provided with a prior law and a likelihood function. The number and the size of the intervals of each variable are optimized with a MAP approach. A piecewise constant estimator of the conditional quantile function is deduced from the optimal partition.

Mots-clés : apprentissage supervisé, régression ordinale, quantiles conditionels, sélection de modèles
Keywords : supervised learning, ordinal regression, conditional quantiles, model selection

1 Introduction

On considère la tâche d'apprentissage supervisé connue sous le terme de *régression ordinale*, lorsque la variable à prédire est ordinale. Dans la communauté statistique les approches utilisent généralement le modèle linéaire généralisé et notamment le modèle cumulatif [7] qui fait l'hypothèse d'une relation d'ordre stochastique sur l'espace des prédicteurs. En apprentissage automatique, plusieurs techniques employées en classification supervisée ou en régression métrique ont été appliquées à la régression ordinale (cf [3] pour un état de l'art). Les problèmes considérés comprennent cependant une échelle de rangs fixée au préalable et relativement restreinte (de l'ordre de 5 ou 10).
Afin de mieux rendre compte de la loi conditionnelle qu'avec des modèles ponctuels, on s'intéresse à des modèles prédictifs probabilistes tels que ceux obtenus par des techniques d'estimation de densité ou d'estimation de quantiles. Pour α réél dans $[0, 1]$, le quantile conditionnel d'ordre α noté $q_\alpha(x)$ est défini comme le réél le plus petit tel que la fonction de répartition conditionnelle soit supérieure à α. Reformulé comme la minimisation d'une fonction de coût adéquate, l'estimation des quantiles est obtenue par l'utilisation de splines dans [6] et de fonctions à noyaux dans [9]. Un partitionnement en arbre de

l'espace des prédicteurs est proposé dans [2] et utilisé dans des forêts aléatoires dans [8]. Toutefois, en régression quantile, les quantiles que l'on souhaite estimer sont généralement régulièrement espacés (q-quantiles avec q connu) et les performances sont évaluées pour chaque quantile.
En estimation de densité, les méthodes dites à noyau définissent le voisinage de chaque point en convoluant la loi empirique des données par une densité à noyau centrée en ce point. La forme du noyau et la largeur de la fenêtre sont des paramètres à régler. Une fois la notion de voisinage définie, les techniques diffèrent selon la famille d'estimateurs visée. Cette démarche d'estimation de la loi complète a déjà été adoptée en régression ordinale dans [3] en utilisant des processus Gaussiens dans un cadre Bayésien.
Nous proposons ici une approche de sélection de partitions 2D pour l'estimation de la fonction quantile conditionnelle. Notre approche utilise la statistique d'ordre en amont du processus d'apprentissage. La manipulation exclusive des rangs au détriment des valeurs rend notre estimateur invariant par toute transformation monotone des données et peu sensible aux valeurs atypiques. Notre méthode effectue un partitionnement 2D optimal qui utilise l'information d'une variable pour mettre en évidence des plages de rangs de la variable à expliquer dont le nombre n'est pas fixé à l'avance. Disposant d'un échantillon de données de taille finie N et ne souhaitant pas émettre d'hypothèse supplémentaire sur la forme de la densité prédictive, nous nous restreignons à des densités conditionnelles sur les rangs constantes sur chaque cellule. Notre estimateur se ramène donc à un vecteur d'estimateurs de quantiles de la loi conditionnelle sur les rangs. A la différence de la régression quantile, les quantiles estimés ne sont pas régulièrement espacés, le choix de leur ordre n'est pas décidé au préalable mais est donné par la partition retenue.

2 Une méthode de partitionnement 2D optimale pour la régression quantile

Nous décrivons ici une méthode de partitionnement 2D récemment présentée à la communauté francophone d'Extraction de Connaissances [5].
Formellement, on caractérise un modèle de discrétisation 2D par les paramètres $\left\{I, J, \{N_{i.}\}_{1\leq i\leq I}, \{N_{ij}\}_{1\leq i\leq I; 1\leq j\leq J}\right\}$ où I est le nombre d'intervalles de la variable explicative, J le nombre d'intervalles de la variable à expliquer, $N_{i.}$ le nombre d'individus pour lesquels la variable explicative est dans l'intervalle i et N_{ij} le nombre d'individus pour lesquels la variable explicative est dans l'intervalle i et la variable à expliquer dans l'intervalle j. Le nombre d'individus dans chaque intervalle cible est noté $N_{.j}$ et se déduit par sommation des nombres d'individus des cellules d'un même intervalle de rang.
Un tel partitionnement permet de décrire la distribution des rangs de la variable à expliquer étant donné le rang de la variable explicative. Le partitionnement recherché est celui qui présente le meilleur compromis entre la qualité de l'information de corrélation détectée

entre les deux variables et la capacité de généralisation de la grille. On adopte une approche de sélection de modèles avec la distribution *a priori* suivante pour les paramètres du modèle:

1. les nombres d'intervalles I et J sont indépendents et distribués uniformément entre 1 et N;

2. pour un nombre d'intervalles I donné, toutes les partitions en intervalles de rangs de la variable explicative sont équiprobables;

3. pour un intervalle de la variable explicative donné, toutes les distributions des exemples sur les intervalles à expliquer sont équiprobables;

4. les distributions des exemples sur les intervalles à expliquer pour chaque intervalle de la variable explicatiave sont indépendantes les unes des autres;

5. pour un intervalle à expliquer donné, toutes les distributions des rangs sont équiprobables.

La vraisemblance des données conditionnellement à une partition se décompose également de manière hiérachique en:

1. la probabilité que le rang d'un individu soit dans un intervalle donné;

2. la probabilité du rang local de l'individu dans l'intervalle considéré.

En utilisant le modèle de discrétisation 2D et les lois décrites ci-dessus, on peut écrire de manière exacte le logarithme négatif du produit $p(\mathrm{M}) \times p(\text{données}|M)$ sous la forme du critère (1) pour un modèle de discrétisation M:

$$\begin{aligned} c_{reg}(M) = 2\log(N) + \log\left(\begin{array}{c} N+I-1 \\ I-1 \end{array}\right) + \sum_{i=1}^{I} \log\left(\begin{array}{c} N_i+J-1 \\ J-1 \end{array}\right) \\ + \sum_{i=1}^{I} \log \frac{N_i!}{N_{i,1}!N_{i,2}!\ldots N_{i,J}!} + \sum_{j=1}^{J} \log N_{\cdot j}! \end{aligned} \tag{1}$$

Le partitionnement 2D obtenu nous permet de construire un estimateur de la fonction de répartitionnelle conditionnelle sur les rangs de la façon suivante : les J quantiles d'ordre $N_{\cdot j}$ où $j = 1, \ldots, J$ sont estimés à l'aide des probabilités empiriques $P(rg(Y) \in T_j | rg(X) \in P_i) = N_{ij}/N_{i\cdot}$ où T_j et P_i sont les intervalles de rang auxquelles appartiennent $rg(Y)$ et $rg(X)$. L'estimateur est de plus supposé constant entre deux quantiles conditionnels successifs.

Dans le cas d'une variable explicative catégorielle, un partitionnement 2D peut également être obtenu par extension à la régression du groupage des valeurs proposé dans [1] en classification supervisée.

Dans le cas de plusieurs variables explicatives, un estimateur Bayésien naïf multivarié peut être construit à partir des estimateurs univariés, en supposant l'indépendance des variables explicatives conditionnellement à la variable à expliquer.

Chaque partitionnement fournit également un indice normalisé de compression $g(M) = 1 - \frac{c(M)}{c(M_\emptyset)}$, où $M_\emptyset$ est le modèle vide avec $I = J = 1$ qui nous permet d'évaluer de manière univariée chaque prédicteur puis de les classer par importance prédictive.

	Synthetic	SO2	Precip	Temp
Nbre de prédicteurs	1	27	106	106
Nbre d'ex en apprentissage	384	22956	10546	10675
Nbre d'ex en test	1024	7652	3517	3560

Table 1: Caractéristiques des jeux de données du Challenge Predictive Uncertainty in Environmental Modelling.

3 Evaluation expérimentale

Comme exposé précedemment, notre approche se distingue des approches habituellement proposées en régression quantile du fait qu'elle optimise le nombre de quantiles à estimer et leur espacement conjointement à leur estimation. Afin de positionner la méthode, nous avons choisi de la comparer en premier lieu à d'autres estimateurs de densité conditionnelle sur les valeurs. Nous décrivons ici les résultats obtenus pour les quatre jeux de données proposés lors du récent Challenge Predictive Uncertainty in Environmental Modelling organisé en 2006 [1] et décrits dans le tableau 1. Notre approche étant par nature régularisée et n'ayant aucun paramètre de réglage , nous prenons le parti d'utiliser les jeux de données d'apprentissage et de validation pour calculer les partitionnements 2D optimaux.

Pour le jeu de données Synthetic comportant une seule variable explicative, le diagramme de dispersion des données d'apprentissage ainsi que le partitionnement optimal obtenu sont reportés à gauche de la Fig. 1. La table des effectifs figure à droite.

Pour les jeux de données réels nous avons calculé l'estimateur multivarié Bayesien naïf utilisant l'ensemble des prédicteurs mais également l'estimateur univarié utilisant uniquement le prédicteur dont l'importance prédictive était la plus élevée ainsi que l'estimateur Bayesien naïf utilisant le couple de variables le plus informatif.

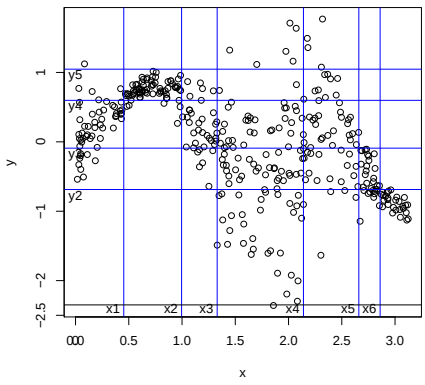
Les différents estimateurs obtenus sont comparées sur le critère du NLPD sur le jeu de données test, après projection des rangs sur les valeurs de l'échantillon d'apprentissage.

Nous avons tout d'abord constaté les mauvaises performances de l'estimateur Bayesien naïf utilisant l'ensemble des prédicteurs. Sur les trois jeux de données réels, le NLPD est en effet supérieur au NLPD pour la méthode dite de référence qui calcule à partir des données d'apprentissage l'estimateur empirique de la loi marginale $p(y)$. Lorsque l'hypothèse d'indépendance est trop forte, il est connu qu'elle dégrade fortement l'estimation des probabilités *a posteriori* [4]. Ces mauvaises performances sont donc certainement dues à des corrélations importantes entre les prédicteurs.

Le tableau 2 indique le NLPD sur le jeu de données test pour les estimateurs univarié et bivarié proposés ainsi que pour la meilleure méthode du Challenge [2] et pour la méthode

[1] http://theoval.cmp.uea.ac.uk/ gcc/competition/

[2] exceptée la soumission de l'organisateur



	P1	P2	P3	P4	P5	P6	P7	Total
T5	1	0	0	6	5	0	0	12
T4	4	68	5	2	9	0	0	88
T3	38	5	26	32	24	0	0	125
T2	9	0	6	33	15	21	0	84
T1	0	0	0	34	4	9	28	75
Total	52	73	37	107	57	30	28	384

Figure 1: Diagramme de dispersion, partitionnement 2D et effectifs de la grille 2D optimal pour le jeu de données Synthetic.

NLPD	Synthetic	SO2	Precip	Temp
Meilleure méthode soumise	0.386 (1er)	4.37 (2ème)	-0.279 (2ème)	0.034 (1er)
Bivarié	×	4.30 (1er)	-0.411 (1er)	0.25 (7ème)
Univarié	1.02(5ème)	4.33 (1er)	-0.361 (1er)	0.285 (7ème)
Référence	1.23 (6ème)	4.5 (3ème)	-0.177 (4ème)	1.30 (9ème)

Table 2: Valeur du NLPD (et classement) pour chacun des 4 jeux de données pour l'estimateur univarié utilisant le meilleur prédicteur (Univarié), l'estimateur Bayesien naïf utilisant le meilleur couple (Bivarié), la meilleure méthode soumise au Challenge et l'estimateur marginal de référence.

dite de référence. On observe tout d'abord que pour tous les jeux de données, les estimateurs MODL sont meilleurs que la méthode de référence, ce qui est loin d'être le cas pour toutes les méthodes soumises. On observe ensuite de bonnes performances des estimateurs proposés, notamment sur les jeux de données SO2 et Precip où les estimateurs univarié et bivarié se placent en tête. Les bonnes performances du prédicteur univarié montrent la qualité du partitionnement 2D obtenu malgré la manipulation exclusive des rangs et non des valeurs durant cette étape. D'autre part, l'estimateur bivarié est toujours meilleur que l'estimateur univarié. Cela indique la présence d'informations supplémentaires et nous encourage à améliorer l'estimateur Bayésien naïf par le biais d'une sélection ou d'un moyennage.

4 Conclusion

Nous avons proposé un estimateur de la densité conditionnelle pour la régression ordinale. Selon une approche de sélection de modèles, notre méthode recherche le partitionnement 2D de chaque couple (variable à expliquer, variable explicative) le plus vraisemblable a posteriori. Les effectifs de chaque partitionnement nous permettent d'estimer un vecteur de quantiles conditionnels dont les ordres sont également donnés par la partition. Les estimateurs obtenus donnent des résultats très encourageants sur quatre jeux de données proposés lors d'un récent challenge.

References

[1] M. Boullé. A Bayes optimal approach for partitioning the values of categorical attributes. *Journal of Machine Learning Research*, 2005.

[2] P. Chaudhuri and W.-Y. Loh. Nonparametric estimation of conditional quantiles using quantile regression trees. *Bernouilli*, 8, 2002.

[3] W. Chu and S. Keerthi. New approaches to support vector ordinal regression. In *ICML '05: Proceedings of the 22nd international conference on Machine Learning*, 2005.

[4] E. Frank, L. Trigg, G. Holmes, and I. Witten. Naive Bayes for regression, 1998. Working Paper 98/15. Hamilton, NZ: Waikato University, Department of Computer Science.

[5] C. Hue and M. Boullé. Une approche non paramétrique bayésienne pour l'estimation de densité conditionnelle sur les rangs. In *7èmes Journées Francophones "Extraction et Gestion de Connaissances" (EGC 2007)*, 2007.

[6] R. Koenker. *Quantile Regression*. Econometric Society Monograph Series. Cambridge University Press, 2005.

[7] P. McCullagh. Regression model for ordinal data (with discussion). *Journal of the Royal Statistical Society B*, 42:109–127, 1980.

[8] N. Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7:983–999, 2006.

[9] I. Takeuchi, Q.V. Le, T.D. Sears, and Smola A.J. Nonparametric quantile estimation. *Journal of Machine Learning Research*, 7:1231–1264, 2006.