Partition BIC optimale de l'espace des prédicteurs

Gilbert Ritschard

*Département d'économétrie, Université de Genève
gilbert.ritschard@themes.unige.ch

Résumé. Cet article traite du partitionnement optimal de l'espace de prédicteurs catégoriels dans le but de prédire la distribution a posteriori d'une variable réponse elle-même catégorielle. Cette partition optimale doit répondre à un double critère d'ajustement et de simplicité que prennent précisément en compte les critères d'information d'Akaike (AIC) ou bayésien (BIC). Après avoir montré comment ces critères s'appliquent dans notre contexte, on s'intéresse à la recherche de la partition qui minimise le critère retenu. L'article propose une heuristique rudimentaire et démontre son efficacité par une série de simulations qui comparent le quasi optimum trouvé au vrai optimum. Plus que pour la partition elle-même, la connaissance de cet optimum s'avère précieuse pour juger du potentiel d'amélioration d'une partition, notamment celle fournie par un algorithme d'induction d'arbre. Un exemple sur données réelles illustre ce dernier point.

1 Introduction

En apprentissage supervisé, des techniques comme l'analyse discriminante, la régression logistique multinomiale, les modèles bayésiens ou les arbres de décisions induits de données (arbres d'induction) apprennent la distribution a posteriori de la variable à prédire, l'objectif étant d'affecter un cas avec profil $\mathbf{x}$ en termes de prédicteurs à la classe y_i ayant la plus forte probabilité a posteriori $p(Y = y_i|\mathbf{x})$. La régression logistique, l'analyse discriminante et les modèles bayésiens par exemple, modélisent la distribution a posteriori sous forme d'une fonction vectorielle continue de $\mathbf{x}$. Par contraste, les arbres d'induction conduisent à un ensemble fini de distributions, chaque distribution étant associée à une classe d'une partition « apprise » de l'ensemble des profils $\mathbf{x}$ admissibles. Nous nous plaçons dans ce dernier contexte et nous intéressons à la détermination de la partition optimale. Nous examinons tout d'abord les critères d'optimalité qui peuvent s'avérer pertinents. Parmi ceux-ci, nous porterons un intérêt particulier aux critères d'information du type AIC et BIC qui permettent d'arbitrer entre qualité d'ajustement et complexité. Le calcul des AIC et BIC pour une partition quelconque se fait par simple adaptation du principe de l'arbre étendu introduit dans Ritschard et Zighed (2003, 2002) pour le cas des arbres.

La recherche de la partition optimale par exploration exhaustive des partitions étant de complexité non polynomiale, il convient de recourir à des heuristiques. Pour cette première approche du problème, on envisage ici une procédure ascendante dont on examine les performances par une analyse de simulations. L'heuristique est rudi-