Une approche non paramétrique Bayesienne pour l'estimation de densité conditionnelle sur les rangs

Carine Hue*, Marc Boullé*

*France Télécom R & D; 2, avenue Pierre Marzin; 22307 Lannion cedex
Carine.Hue@orange-ftgroup.com; Marc.Boulle@orange-ftgroup.com

Résumé. Nous nous intéressons à l'estimation de la distribution des rangs d'une variable cible numérique conditionnellement à un ensemble de prédicteurs numériques. Pour cela, nous proposons une nouvelle approche non paramétrique Bayesienne pour effectuer une partition rectangulaire optimale de chaque couple (cible,prédicteur) uniquement à partir des rangs des individus. Nous montrons ensuite comment les effectifs de ces grilles nous permettent de construire un estimateur univarié de la densité conditionnelle sur les rangs et un estimateur multivarié utilisant l'hypothèse Bayesienne naïve. Ces estimateurs sont comparés aux meilleures méthodes évaluées lors d'un récent Challenge sur l'estimation d'une densité prédictive. Si l'estimateur Bayésien naïf utilisant l'ensemble des prédicteurs se révèle peu performant, l'estimateur univarié et l'estimateur combinant deux prédicteurs donne de très bons résultats malgré leur simplicité.

1 Introduction

Dans cette introduction, nous décrivons tout d'abord une situation particulière de l'apprentissage supervisé où l'on s'intéresse à prédire le rang d'une cible plutôt que sa valeur. Nous exposons ensuite deux approches qui permettent de passer d'une prédiction ponctuelle en régression à une description plus fine de la loi prédictive. Nous présentons ensuite notre contribution qui vise à fournir une estimation de la densité conditionnelle complète du rang d'une cible par une approche Bayesienne non paramétrique.

1.1 Régression de valeur et régression de rang

En apprentissage supervisé on distingue généralement deux grands problèmes : la classification supervisée lorsque la variable à prédire est symbolique et la régression lorsqu'elle prend des valeurs numériques. Dans certains domaines tels que la recherche d'informations, l'intérêt réside cependant plus dans le rang d'un individu par rapport à une variable plutôt que dans la valeur de cette variable. Par exemple, la problématique initiale des moteurs de recherche est de classer les pages associées à une requête et la valeur intrinsèque du score n'est qu'un outil pour produire ce classement. Indépendamment de la nature du problème à traiter, utiliser les rangs plutôt que les valeurs est une pratique classique pour rendre les modèles plus robustes aux valeurs atypiques et à l'hétéroscédasticité. En régression linéaire par exemple, un estimateur utilisant les rangs centrés dans l'équation des moindres carrés à minimiser est proposé