Comparaison des valeurs de Shapley et des valeurs du poids de l'évidence dans le cas du classifieur naïf de Bayes

Vincent Lemaire*, Fabrice Clérot*, Marc Boullé*

* Orange, Lannion, France

Résumé. La sélection de variables et/ou la mesure d'importance des variables en entrée d'un modèle de machine learning est (re)devenue le centre de nombreuses recherches du fait de la réglementation européenne sur la protection de la vie privée. Avoir un bon modèle ne suffit plus il faut aussi expliquer ses décisions. Il existe de ce fait aujourd'hui de nombreux algorithmes d'intelligibilité. Parmi ces derniers on trouve beaucoup , ces derniers temps, d'algorithmes d'estimation des valeurs de Shapley, une méthode d'intelligibilité reposant sur la théorie des jeux coopératifs. Cet article propose une comparaison des valeurs de Shapley dans le cas particulier du classifieur naïf de Bayes avec un autre indicateur fréquemment utilisé "le poids de l'évidence".

1 Introduction

Il existe de nombreux algorithmes d'intelligibilité, souvent empiriques et parfois sans justifications théoriques. C'est là l'une des raisons principales pour lesquelles la bibliothèque Python SHAP a été créée en 2017 par Scott Lundberg à la suite de sa publication (Lundberg et Lee, 2017), pour proposer des algorithmes d'estimation des valeurs de Shapley, une méthode d'intelligibilité reposant sur la théorie des jeux coopératifs. Depuis son lancement, cette bibliothèque connaît un succès grandissant, notamment grâce à de meilleures justifications théoriques et à des visualisations qualitatives.

On propose dans ce document une expression analytique des valeurs de Shapley dans le cas particulier du classifieur naïf de Bayes et une comparaison avec "le poids de l'évidence". Expliquer la prédiction à l'aide des valeurs de l'un de ces indicateurs consiste à attribuer à chaque variable d'entrée, plus précisément à chaque valeur décrivant l'individu considéré un coefficient réel. Chacun de ces coefficients indique comment cette valorisation a contribué à impacter la prévision.

2 Shapley pour le classifieur naïf de Bayes

A notre connaissance il n'existe pas dans la littérature de calcul "analytique" des valeurs de Shapley pour le classifieur naïf de Bayes. Cette première section est donc dédiée à une proposition de calcul de ces valeurs, en exploitant l'hypothèse d'indépendance conditionnelle des variables qui caractérise ce classifieur.