Construction de variables à l'aide de classifieurs comme aide à la régression : une évaluation empirique

Colin Troisemaine*, Vincent Lemaire*

*Orange Labs Lannion
colin.troisemenaine@orange.com

Résumé. Cet article propose une méthode de création automatique de variables (pour la régression) qui viennent compléter les informations contenues dans le vecteur initial des variables explicatives. Notre méthode fonctionne comme une étape de prétraitement dans laquelle les valeurs continues de la variable a régresser sont discrétisées en un ensemble d'intervalles ce qui permet de définir des seuils de valeurs. Ensuite, des classifieurs sont entraînés pour prédire si la valeur à régresser est inférieure ou égale à chacun de ces seuils. Les sorties des classifieurs sont ensuite concaténées sous la forme d'un vecteur additionnel qui vient enrichir le vecteur initial de variables explicatives natives du problème de régression. Le système implémenté peut donc être considéré comme un outil de prétraitement générique. Nous avons testé la méthode d'enrichissement proposée avec 5 types de régresseurs et l'avons évalué dans 33 jeux de données de régression. Nos résultats expérimentaux confirment l'intérêt de l'approche.

1 Introduction

Les techniques d'apprentissage peuvent se découper en deux grandes familles selon leur vocation principale : celles servant à décrire les données (méthodes descriptives) et celles permettant de prédire un phénomène (plus ou moins) observable (méthodes prédictives). Les méthodes prédictives permettent de prévoir et d'expliquer à partir d'un ensemble de données étiquetées un ou plusieurs phénomènes (plus ou moins) observables. Dans le cas de la régression il s'agira de prévoir la valeur une variable numérique (noté y), par exemple le montant d'une facture, à l'aide d'un ensemble de variables explicatives (un vecteur noté X).

Dans le cas de l'apprentissage automatique, on cherchera à apprendre une fonction f telle que $y = f(X)$ à l'aide d'un algorithme d'apprentissage automatique et d'un ensemble d'apprentissage, un ensemble de N couples entrée-sortie $(X_i, y_i), i = 1, ..., N$. Lors de cette étape de modélisation, il existe souvent le besoin de créer de nouvelles variables qui décrivent mieux le problème et permettent au modèle d'atteindre de meilleures performances. C'est ce qu'on appelle le "processus d'ingénierie de création de nouvelles variables explicatives" (Sondhi, 2009). Dans ce cas, on espère que les nouvelles variables (un vecteur qui sera ici noté X') apporteront une information additionnelle. L'automatisation de la génération de ces "nouvelles variables" permet d'extraire des informations plus utiles et significatives des données, dans un cadre qui peut être appliqué à n'importe quel problème. Ce qui permet à l'ingénieur en apprentissage automatique de consacrer plus de temps à des tâches plus utiles.