Voir le processus de création d'exemples contrefactuels comme une source de connaissance - Application au classifieur Naïf de Bayes

Vincent Lemaire*, Nathan Le Boudec*,**, Françoise Fessant*, Victor Guyomard*

* Orange Innovation, Lannion, France
** Université de technologie de Compiègne, France

Résumé. Il existe aujourd'hui de nombreux algorithmes de compréhension des décisions d'un algorithme d'apprentissage automatique. Parmi ceux-ci, on trouve ceux basés sur la génération d'exemples contrefactuels. Cet article propose de considérer ce processus de génération comme une source de connaissance qui peut être stockée puis utilisée de différentes manières. Ce processus est illustré dans le cas des modèles additifs et en particulier dans le cas du classifieur naïf de Bayes, dont on montre des propriétés intéressantes pour ce faire.

1 Introduction

L'apprentissage automatique, l'une des branches de l'intelligence artificielle, a connu de nombreux succès ces dernières années. Les décisions prises par ces modèles sont de plus en plus précises, mais aussi de plus en plus complexes. Il apparaît néanmoins que certains de ces modèles sont apparentés à des boîtes noires : leurs décisions sont difficiles, voire impossibles, à expliquer (Bodria et al., 2023). Ce manque d'explicabilité peut entraîner un certain nombre de conséquences indésirables : manque de confiance de l'utilisateur, réduction de l'utilisabilité des modèles, présence de biais, etc. C'est à partir de ces besoins qu'est né le domaine de la XAI (eXplainable AI). Le XAI (Saeed et Omlin, 2023; Allen et al., 2023) est une branche de l'intelligence artificielle qui vise à rendre les décisions prises par les modèles d'apprentissage automatique intelligibles pour les utilisateurs.

Parmi les méthodes XAI, le raisonnement contrefactuel est un concept issu de la psychologie et des sciences sociales (Miller, 2019). Il consiste à examiner les alternatives possibles aux événements passés (Stepin et al., 2021). Les humains utilisent souvent le raisonnement contrefactuel en imaginant ce qui se passerait si un événement ne s'était pas produit, et c'est ce qu'est exactement le raisonnement contrefactuel. Appliquée à l'intelligence artificielle, la question est, par exemple, "Pourquoi le modèle a-t-il pris cette décision plutôt qu'une autre ?" ou "En quoi la décision aurait-elle été différente si une certaine condition avait été modifiée ?

Dans le cadre du raisonnement contrefactuel, cet article propose de considérer ce processus de génération comme une source de connaissances qui peut être stockée puis exploitée de différentes manières. Ce processus est illustré dans le cas des modèles additifs et en particulier dans le cas du classificateur de Bayes naïf, dont on montrera des propriétés intéressantes pour ce faire.