Sur le pouvoir explicatif des arbres de décision

Gilles Audemard*, Steve Bellart*, Louenas Bounia*
Frédéric Koriche*, Jean-Marie Lagniez*, Pierre Marquis* **

Univ. Artois, CNRS, CRIL, F-62300 Lens*
Institut universitaire de France**
nom@cril.fr,
http://www.cril.univ-artois.fr/

Résumé. Les arbres de décision constituent un modèle d'apprentissage adapté aux applications pour lesquelles l'interprétabilité des décisions est d'une importance primordiale. Nous examinons ici la capacité des arbres de décision binaires à extraire, minimiser et compter des explications abductives / contrastives. Nous montrons que l'ensemble de toutes les explications abductives irredondantes (ou raisons suffisantes) d'une instance peut être de taille exponentielle. Aussi, générer l'intégralité de cet ensemble peut se révéler hors de portée. De plus, deux raisons suffisantes d'une même instance peuvent différer sur tous leurs attributs. Ainsi, le calcul d'une seule raison suffisante ne donne qu'une vision parcellaire des explications abductives possibles. Nous introduisons les notions d'attribut nécessaire / pertinent pour l'explication et la notion d'importance explicative d'un attribut et nous montrons que ces notions peuvent être utiles pour dériver une vue synthétique des raisons suffisantes d'une instance.

1 Introduction

Expliquer une décision à une personne, c'est donner les détails ou les raisons qui aident cette personne à comprendre pourquoi une telle décision a été prise. Lorsque les décisions sont prises par des modèles d'apprentissage automatique (ML) dits opaques, comme les forêts aléatoires, les SVM et les réseaux de neurones, la génération d'explications est une tâche complexe. Pour autant, avec le nombre croissant d'applications qui reposent sur des techniques d'apprentissage automatique, les recherches sur l'IA explicable (XAI) sont devenues essentielles. Elles visent à développer des méthodes et des approches efficaces pour interpréter les modèles d'apprentissage et expliquer les décisions prises (Frosst et Hinton, 2017; Guidotti et al., 2019; Hooker et al., 2019; Huysmans et al., 2011; Ignatiev et al., 2019; Lundberg et Lee, 2017; Miller, 2019; Molnar, 2019; Ribeiro et al., 2016; Shih et al., 2019).

Dans cet article, nous nous intéressons aux classeurs booléens où seules deux décisions (classements) sont possibles, 1 pour les instances positives, et 0 pour les instances négatives. Peu importe que l'instance considérée $\boldsymbol{x}$ soit positive ou pas, la recherche d'explications de son classement est une question importante dans bien des cas (Miller, 2019). D'une part, les explications « abductives » visent à expliquer pourquoi $\boldsymbol{x}$ est classée comme elle a été classée