Approches formelles pour l'explicabilité de l'IA

Aurélie Hurault - IRIT - Toulouse

Sujet

L'intelligence artificielle est présente un peu partout dans l'industrie, et se fraie une place dans des applications où ses décisions peuvent avoir des conséquences dramatiques (véhicules autonomes, décisions financières, etc.). La question se pose alors : quelle confiance accorder à une intelligence artificielle ? Comment justifier cette confiance ?

Pour gagner en confiance, il est possible de s'intéresser à la façon dont le modèle est généré ou de s'intéresser à des propriétés sur le modèle indépendamment de sa construction. Les propriétés d'intérêt sont par exemple la robustesse, l'équité ou l'explicabilité. C'est la seconde approche que nous explorons dans nos travaux et nous nous intéressons plus particulièrement à l'explicabilité.

L'explicabilité des modèles issus de l'apprentissage automatique vise à aider les humains à comprendre les décisions prises par ces modèles complexes. Le besoin d'explicabilité est un des défis pour l'utilisation des techniques d'apprentissage automatique, dans les systèmes critiques mais également pour des problèmes d'éthiques (aide à la décision pour des banques ou la médecine par exemple).

L'objectif de ce stage est de développer des outils qui permettent un raisonnement efficace et rigoureux sur les modèles issu de l'apprentissage automatique, en particulier sur les problématiques d'explicabilité. Nous ne cherchons pas à avoir des garanties formelles sur le modèle lui-même. Nous souhaitons que les algorithmes et outils qui raisonnent sur les modèles nous fournissent des réponses fiables. Par exemple si nous demandons à un modèle d'aide à la décision pour l'octroiement d'un prêt, pourquoi un prêt a été refusé, nous voulons être sûre que cette réponse est correcte (algorithme juste et implantation de l'algorithme non buggée). Un certain nombre d'algorithmes d'explicabilité existent dans la littérature, nous nous intéresserons à leur correction.

Ce stage s'inscrit dans le cadre d'une collaboration avec Joao Silva-Marques, titulaire d'une chaire de l'institut ANITI (https://aniti.univ-toulouse.fr/).

État de l'art

Explicabilité

- "Logic-Based Explainability in Machine Learning". João Marques-Silva. Reasoning Web 2022: 24-104
- "Explanations for Monotonic Classifiers", João Marques-Silva, Thomas Gerspacher, Martin C. Cooper, Alexey Ignatiev, Nina Narodytska, ICML 2021: 7469-7479
- "On explaining random forests with SAT", Y. Izza and J. Marques-Silva, IJCAI 2021, 2584–2591

Preuves d'algorithmes

- "Certified Logic-Based Explainable AI - The Case of Monotonic Classifiers", Aurélie Hurault, João Marques-Silva. TAP 2023: 51-67
- "Formally verifying the solution to the boolean pythagorean triples problem", L. Cruz-Filipe, J. Marques-Silva, and P. Schneider-Kamp, J. Autom. Reason., 63(3):695–722, 2019.

Lieu et contact

Laboratoire IRIT, site ENSEEIHT, Toulouse.

Équipe ACADIE http://www.irit.fr/-Equipe-ACADIE-

Contact : Aurélie Hurault hurault@enseeiht.fr