Explications de données et de classifieurs : quelques méthodes et risques notables

Marie-Jeanne Lesot*

* Sorbonne Université,
Laboratoire d'Informatique de Paris 6, LIP6
Bureau 26-00/510
DAPA, LIP6
BP 169, 4 place Jussieu, 75005 Paris
France
https://webia.lip6.fr/~lesot/
Marie-Jeanne.Lesot@lip6.fr

Résumé

Au delà de la question de la performance des méthodes d'apprentissage automatique, il est devenu crucial d'augmenter la lisibilité des résultats obtenus, pour permettre aux utilisateurs de les comprendre et d'interpréter. Ces problématiques sont regroupées sous le terme "eXplainable Artificial Intelligence" (XAI). Au sein de ce vaste domaine, cet exposé aborde deux niveaux : le premier porte sur la compréhension des données elles-mêmes, dans un cadre d'analyse exploratoire et de description intelligible. L'objectif est de permettre à un utilisateur de comprendre le contenu des données en les résumant par le biais de formulations linguistiques, ce qui soulève en particulier les problèmes de choix des mots et de cohérence des résumés.

Le second niveau considéré est celui de tâches de classification, dans le cadre classique de l'interprétation locale, post-hoc et agnostique de la prédiction d'une classe pour une donnée. Les questions soulevées sont celles des risques liés à la définition de la localité et à la génération d'explications non justifiées.