Theoretical foundations for explainability

Damien Garreau

Summary: The state-of-the-art algorithms for specific tasks such as object identification in images are better than humans in terms of accuracy in many cases. However, these algorithms are often perceived as ``black boxes,'' depending on millions of parameters. The goal of explainability is to provide tools helping us to understand how a given algorithm makes predictions. These tools come from traditional Statistics: linear model, statistical hypothesis testing, etc.
The goal of this internship is to study a recent method for explainability, LIME, and to provide theoretical guarantees in simple settings. Can we prove that the explanations given by LIME make sense? At the very least, we will show the limits of LIME on recent architectures. For instance, is it possible to ``forget'' important parameters in the produced explanation for some hyperparameters choices?

References: (a) online book on interpretability with a chapter on LIME: https://christophm.github.io/interpretable-ml-book/ (b) the original article: Why should i trust you?: Explaining the predictions of any classifier, Ribeiro, Singh, Guestrin (2016), SIGKDD, available here: https://arxiv.org/pdf/1602.04938.pdf

French

Titre : Fondements théoriques de l'interprétabilité

Description : Les algorithmes les plus performants aujourd'hui pour des tâches spécifiques telle que l'identification d'objets dans une image dépassent souvent les humains en terme de performance. Cependant, ces algorithmes sont souvent perçus comme des "boîtes noires" qui dépendent de millions de paramètres. Le but des méthodes d'interprétabilité est d'expliquer comment un algorithme réalise ses prédictions, souvent à l'aide d'outils venu des Statistiques traditionnelles : modèle linéaire, test d'hypothèse.
L'objectif de ce stage est d'étudier une méthode récente, LIME, et de fournir des garanties théoriques quant à son fonctionnement. En particulier, peut-on prouver dans des cas simples que les explications fournies par LIME sont bien fondées ? A défaut d'obtenir de tels résultats, on s'attachera à montrer expérimentalement les limite de cette méthode sur des architectures récentes. Par exemple, est-il possible pour certains choix de paramètres d'"oublier" des paramètres importants dans l'explication fournie ?

Références : - un livre sur l'interprétabilité : https://christophm.github.io/interpretable-ml-book/
- l'article original : Why should i trust you?: Explaining the predictions of any classifier, Rib