Résumé de "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead"

Auteur: Alassane Watt -- alassane.watt@student.ecp.fr

Cet article présente les raisons pour lesquelles les modèles intrinsèquement interprétables devraient être préférés aux modèles explicatifs de boîtes noires dans des décisions à enjeux élevés. Un modèle de boîte noire est une formule qui est soit très compliquée à comprendre ou propriétaire. L'auteur identifie les défis liés à l'apprentissage automatique interprétable, et fournit plusieurs exemples d'applications où des modèles interprétables pourraient potentiellement remplacer les modèles de boîte noire en justice pénale, santé et vision par ordinateur.

Utiliser des boîtes noires ne permet pas de rendre compte des décisions prises par ces modèles. L'incompréhension de ces décisions peuvent être à l'origine de conséquences négatives tel que l'illustre l'exemple d'un détenu qui s'est vu refuser la libération conditionnelle à cause d'un score «COMPAS» mal calculé. Comment une telle décision a été prise? C'est ce que se propose d'expliquer l'XAI à travers les modèles intrinsèquement interprétables et les modèles explicatifs.

Selon l'auteur, un modèle d'apprentissage automatique interprétable est contraint sous forme de modèle afin qu'il soit soit utile à quelqu'un, soit obéit à une connaissance structurelle du domaine, telle que la monotonie, la causalité, les contraintes structurelles (génératives), l'additivité ou les contraintes physiques qui proviennent de la connaissance du domaine. Cependant, les modèles interprétables ont quelques défauts qui font qu'ils ne sont pas faciles à implémenter. Leurs principaux défauts sont le fait qu'ils nécessitent de l'expertise dans le domaine d'intérêt et puis leur complexité de calcul peut être élevée.

De l'autre côté, les modèles explicatifs sont des modèles à part qui ont tendance à expliquer les comportements des boîtes noires. Néanmoins, ce type de modèles souffre de défauts qui font que certains, dont l'auteur de l'article, le défavorisent au profit des modèles intrinsèquement interprétables. Le principal élément qui questionne la pertinence des modèles explicatifs est le fait qu'ils fournissent des explications qui ne sont pas tout le temps complètement fidèles aux comportements du modèle d'origine. En effet, si l'explication était complètement fidèle à ce que le modèle d'origine calcule, l'explication serait égale au modèle d'origine, et on n'aurait pas besoin du modèle d'origine en premier lieu. Ensuite, les explications peuvent ne pas avoir de sens ou fournir assez de détails pour une bonne compréhension d'une boîte noire. L'auteur évoque un exemple (la carte de saillance) qui illustre cet argument, cependant cela ne signifie pas que tous les modèles explicatifs sont vagues. De plus, l'auteur souligne que la croyance relative au compromis entre la précision d'un modèle et son interprétabilité est un mythe. Les modèles les plus complexes ne sont pas toujours les plus précis. La différence de performance des classificateurs complexes par rapport aux classificateurs plus simples est pratiquement insignifiante s'il y a une structure dans les données avec une bonne représentation des caractéristiques.

Tous ces éléments font que les modèles interprétables sont préférés aux modèles explicatifs la plupart du temps dans des décisions à enjeux élevés. Je suis assez favorable avec l'opinion de l'auteur car si on peut produire un modèle interprétable, pourquoi expliquer les boîtes noires? Un bel exemple de réseaux de neurones interprétables est présenté par l'article "Concept Whitening for interpretable Image Recognition" qui montre le processus de classification d'une image par les couches internes d'un réseau de neurones convolutifs. L'article démontre qu'il n'ya pas de perte de précision avec l'interprétabilité, ce qui appuie les propos de l'auteur.