INTERPRÉTABILITÉ DES MODÈLES PRÉDICTIFS

Gilbert Saporta

Les succès du *Machine Learning* dans les problèmes de régression ou de classification supervisée ont conduit de nombreux auteurs à théoriser l'opposition entre comprendre et prédire, tant sur le plan conceptuel que pratique. Leo Breiman (2001) opposait ainsi ces deux cultures. Vladimir Vapnik (2006) renchérissait en montrant que « *Better models are sometimes obtained by deliberately avoiding to reproduce the true mechanisms* » .

Ce fut le triomphe des boîtes noires: peu importait que le modèle, ou plutôt l'algorithme, soit compréhensible du moment qu'il prédisait bien. La disponibilité de gigantesques bases de données, (*Big Data*), conduisit un peu vite Chris Anderson (2008) à prédire la fin de la démarche scientifique.

L'application à des décisions concernant la vie des individus finit par susciter des polémiques (Cathy O'Neil, 2016). Des codes d'éthique et des règlements, tel le RGPD européen (règlement général sur la protection des données, 2016) consacrèrent le droit à l'explication.

Dans cet exposé, on reviendra tout d'abord sur le dilemme entre **prédire** et **comprendre**. On distinguera ensuite les modèles explicables des modèles interprétables. Ces derniers incluent les modèles causaux , ce qui permettra de faire le lien avec l'ASI.

L'**explicabilité** des algorithmes (X-AI *explainable artificial intelligence*) est un domaine de recherche actif avec des approches globales ou locales, agnostiques ou spécifiques, pour mesurer l'importance des variables, utilisant souvent des modèles de substitution (Christoph Molnar, 2021).

Au-delà de la transparence nécessaire, le traitement **équitable (*fairness*)** ou non-discriminant par des algorithmes prédictifs est devenu une exigence éthique qui donne lieu à une abondante littérature non exempte de paradoxes sur les mesures d'équité (Mitchell *et al*., 2021). Ce que l'on nomme **biais** des algorithmes n'est la plupart du temps que la conséquence des biais des données exploitées.

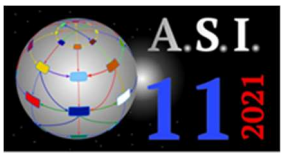
On se demandera en conclusion s'il faut ou non renoncer aux boîtes noires.

Références

[1] Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. *Wired magazine*, 16(7), 16-07

[2] Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, *16*(3), 199-231

[3] Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic Fairness: Choices, Assumptions, and Definitions. *Annual Review of Statistics and Its Application*, *8* (141-163)



[4] Molnar, C. (2021). *Interpretable machine learning , A Guide for Making Black Box Models Explainable*, https://christophm.github.io/interpretable-ml-book.

[5] O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

[6] Vapnik, V. (2006) *Estimation of Dependences Based on Empirical Data*, 2nd edition, Springer