Inférence statistique des relations de causalité. Vers des algorithmes d'apprentissage éthiques ?

Frédéric Pennerath

Contexte et enjeux

L'analyse statistique des données et l'apprentissage automatique offrent de nombreux modèles pour représenter les distributions jointes entre variables aléatoires. Ces modèles permettent ainsi d'identifier les dépendances ou "corrélations" entre variables, ou a contrario, les relations d'indépendance. Les logiciels du machine learning permettent aujourd'hui d'automatiser entièrement le processus d'apprentissage de ces modèles à partir de données et ce faisant, permettent au plus grand nombre de mettre au point des modèles prédictifs sans compétence statistique particulière. Ce recours généralisé, s'il se fait de manière aveugle, peut rapidement aboutir à des systèmes agissant de façon contraire à nos principes éthiques. Aux Etats-Unis, plusieurs polémiques ont éclaté faisant état de systèmes d'IA aux comportements manifestement racistes ou sexistes. Plusieurs causes peuvent être à l'origine de ces déviances. Il peut d'abord s'agir de problèmes de biais dans les données. Par exemple, un logiciel de recrutement ne sélectionnera pas de femmes si les données d'apprentissage ne contenaient presque que des hommes. Parfois le problème est plus sournois, en lien avec la notion de causalité : si les minorités ethniques sont surreprésentées dans les prisons américaines et si on suppose que la police et la justice sont neutres, faut-il en déduire comme le ferait un logiciel d'IA que parce qu'une personne est issue d'une minorité, elle aura plus tendance à sombrer dans la criminalité ? La réponse est bien évidemment non. Il s'agit en réalité d'un problème de *causalité* vis-à-vis de variables latentes dites *confondantes* : ainsi, si on conditionne l'origine ethnique et la criminalité à la catégorie sociale, les niveaux d'éducation et de revenu, etc, ces deux variables totalement indépendantes. Dit autrement, la cause première de la criminalité n'est pas l'appartenance à une minorité mais le niveau social, le niveau social n'étant par ailleurs toujours pas indépendant de l'origine ethnique. Cet exemple illustre pourquoi les statisticiens n'ont de cesse de nous mettre en garde en répétant que « les corrélations ne sont pas des relations de causalité ».

Comme le montre le paradoxe de Simpson revisité par Pearl (cf [3], p.24), la mauvaise nouvelle est qu'on ne peut espérer extraire de relations de causalité à partir de données brutes sans disposer d'informations contextuelles. C'est pourquoi les statisticiens ont pendant longtemps abordé le problème de l'identification des causes à travers uniquement des plans d'expérience très contrôlés (études pharmaceutiques, sondages politiques, etc). Depuis les travaux fondateurs de J. Pearl [1], différents modèles ont été proposés pour formaliser la notion de causalité en s'appuyant sur les *réseaux causaux* [3] et l'*analyse contrefactuelle* [4]. De nombreux travaux comme [6, 2] ont été récemment proposés pour apprendre automatiquement, sous certaines hypothèses, des relations causales à partir de données en s'appuyant parfois sur les techniques les plus récentes d'apprentissage automatique. On pourra se référer à l'état de l'art [5] pour découvrir les développements les plus récents.

Travail demandé

L'étude consistera à analyser et faire la synthèse des différentes définitions formelles données à la notion de causalité, aux notions théoriques utiles dans ce cadre ainsi qu'aux algorithmes capables d'estimer des relations causales à partir de données. On cherchera aussi à tester et évaluer les algorithmes de l'état de l'art qui sont disponibles. L'ensemble des résultats de l'étude donnera lieu à la rédaction d'un rapport écrit en Latex.

Références

[1] Judea Pearl. *Causality, Models, Reasoning and Inference*. Cambridge University Press, 2nd edition, 2009.

[2] Alexander Marx and Jilles Vreeken. Testing Conditional Independence on Discrete Data using Stochastic Complexity. In *AISTATS*, volume 89, pages 496–505. PMLR, April 2019.

[3] Judea Pearl, Glymour, Madelyn, and Jewell, Nicholas P. *Causal Inference in Statistics. A Primer*. Wiley, 2016.

[4] Donald B Rubin. Causal Inference Using Potential Outcomes : Design, Modeling, Decisions. *Journal of the American Statistical Association*, 100(469) :322–331, March 2005.

[5] Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. A Survey on Causal Inference. *arXiv :2002.02770 [cs, stat]*, February 2020. arXiv : 2002.02770.

[6] Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. Representation Learning for Treatment Effect Estimation from Observational Data. In *NIPS*, pages 2633–2643, 2018.