Des algorithmes pour expliquer les algorithmes

Christine Solnon

Des algorithmes pour expliquer les algorithmes

Christine Solnon, CITI, Inria, INSA Lyon, F-69621 Villeurbanne

Chronique parue dans La recherche N° 573, avril-juin 2023

De nombreuses décisions sont prises ou recommandées par des algorithmes : choisir des vidéos, affecter des élèves à des établissements d'enseignement, accorder un prêt ou une remise de peine, ... Obtenir des explications sur ces décisions est fondamental afin d'avoir confiance et mieux accepter ces systèmes. C'est aussi un droit affirmé dans la loi pour une République numérique de 2016, et dans le Règlement général sur la protection des données de 2018. Mais si ces lois imposent d'expliquer, elles ne précisent pas comment le faire, ni ne définissent des critères pour juger de leur qualité.

Comment, dès lors, évaluer la qualité d'une explication ? Cela ne peut se faire qu'au regard de l'objectif de la personne à qui elle s'adresse : s'agit-il de comprendre la logique globale du système avant de l'utiliser ? De vérifier qu'une décision donnée est équitable et n'a pas exploité des informations non pertinentes (comme la couleur de la peau pour accorder un prêt, par exemple) ? De guider la prise de décision dans le cas d'une recommandation (une application proposant un diagnostic médical doit donner les raisons de ce diagnostic, afin de convaincre le médecin qu'il est correct) ?

Certains algorithmes sont naturellement explicables. Par exemple, les décisions d'un système à base de règles peuvent être justifiées en listant les règles utilisées. Mais de nombreux systèmes d'IA ne peuvent être expliqués en observant leur code, soit parce qu'il n'est pas accessible (pour des raisons de propriété intellectuelle, par exemple), soit parce qu'il ne permet pas de construire des explications intelligibles par un être humain. C'est le cas, notamment, pour les réseaux de neurones profonds : ils possèdent un très grand nombre de paramètres (dont les valeurs sont fixées progressivement afin de faire correspondre au mieux le modèle à des données d'entraînement), et la connaissance des valeurs de ces paramètres ne permet pas de fournir des explications intelligibles. Ces IA fonctionnant en boîtes noires peuvent construire des explications en même temps que la décision. On peut, par exemple, entraîner un réseau de neurones à simultanément faire

une prédiction et fournir une justification intelligible de cette prédiction sous la forme d'un ensemble de caractéristiques ayant influencé la décision. Mais comment vérifier leur sincérité ? Une possibilité est de concevoir des algorithmes en charge de construire des explications uniquement en observant les décisions prises. Prenons l'exemple d'une application qui affecte des élèves à des établissements d'enseignement supérieur, chaque élève étant décrit par un ensemble de caractéristiques (notes, vœux, genre). Si un élève x souhaite avoir une justification de son affectation à un établissement Y, nous pouvons chercher les caractéristiques communes avec d'autres élèves que l'algorithme a affectés à Y. Nous pourrons justifier l'affectation de x à Y, en expliquant que c'est aussi le cas de tous les élèves ayant classé Y parmi leurs trois premiers voeux et ayant une moyenne en musique supérieure à 14/20. Si x souhaite contester cette affectation car il préfère un autre établissement Z, nous pouvons rechercher une explication *contrefactuelle* en identifiant ce qu'il faudrait changer dans les données décrivant x pour que l'algorithme l'affecte à Z plutôt qu'à Y : un élève dont la seule différence avec x est le genre et ayant été affecté à Z fournit une bonne raison pour contester l'affectation de x à Y.

La confrontation des explications fournies par l'IA avec celles construites par un système extérieur (observant les décisions prises) sert à vérifier que l'IA ne masque pas délibérément des explications contestables (à caractère discriminatoire, entre autres), ou que les considérations éthiques revendiquées par le concepteur d'un système sont effectivement mises en pratique dans le système. Ainsi, des chercheurs ont montré que les explications fournies par Facebook sur le choix des publicités montrées étaient souvent incomplètes et parfois trompeuses [1]. La connaissance de divergences entre comportement déclaré et comportement effectif est fondamentale pour permettre aux utilisateurs d'adapter leurs usages. Force est de constater qu'elle est rarement disponible et il est donc très important de continuer les recherches sur l'IA explicable.

1. Voir A. Andreou et al., doi : 10.14722/ndss.2018.23191, 2018