Sujet PLDAC 2023-2024
XAI : explications contrastives bi-factuelles

Isabelle Bloch
Marie-Jeanne Lesot

Les utilisateurs d'une méthode ou d'un algorithme d'intelligence artificielle demandent fréquemment des explications à propos des résultats qu'ils obtiennent, en particulier dans des situations inattendues. Leurs questions peuvent par exemple prendre la forme : pourquoi la décision est-elle P dans un certain contexte et Q dans un autre ?

Ces situations peuvent correspondre à deux modèles dont l'un a été mis à jour par rapport à l'autre, ou à deux cas qui semblent similaires et pour lesquels on se serait attendu à des résultats similaires également. On parle dans ce cas d'*explications contrastives bi-factuelles*, qui ont été étudiées en particulier par T. Miller grâce à un formalisme associant logique et graphes de causalité. Le principe proposé par T. Miller consiste à identifier quels changements de valeurs de variables peuvent constituer des causes contrastives.

L'objectif du projet est de mettre en œuvre cette approche et de l'illustrer sur des exemples simples.

Quelques références

Miller, T. (2019). Explanation in artificial intelligence : insights from the social sciences. *Artificial Intelligence*, *267*, 1–38.

Miller, T. (2021). Contrastive explanation : a structural-model approach. *The Knowledge Engineering Review*, *36*.