Interprétabilité en machine learning et interaction homme/machine

Stéphane Chrétien et Julien Velcin
Laboratoire ERIC - Université Lyon 2

1 Présentation du stage

Le machine learning est actuellement en plein essor, tant dans le monde industriel qu'académique. Les méthodes récentes, telles le *deep learning*, ont permis des avancées spectaculaires dans de nombreux domaines. L'horizon des applications potentielles est très large et on approche pas à pas d'un "plateau" où il semble que les progrès soient de plus en plus minces et que les techniques sont proches de réaliser tout leur potentiel.

Un aspect du machine learning, encore peu exploré mais tout aussi important pour l'acceptabilité du grand public, est celui de l'interprétabilité. Cet aspect s'inscrit dans l'idée d'outils d'une Intelligence Artificielle explicable (XAI). Cela concerne notamment les systèmes susceptibles de prendre des décisions automatiquement, tels les véhicules autonomes, les robots de chirurgie à distance, les systèmes d'analyse d'images de médicale ou les systèmes d'analyse textuelle, entre autres. Ces systèmes doivent être interprétables, en ce sens que l'on puisse comprendre sur quels éléments la décision s'appuie et donc pouvoir rendre des comptes. L'interprétabilité est un élément clé pour éviter que les systèmes artificiellement intelligents prennent des décisions non-transparentes, voire injustifiables, et ce quelque soient leurs performances.

1.1 But du stage

L'objet du stage est d'apprendre certaines des méthodes d'interprétabilité les plus en vogue parmi celles actuellement développées dans la communauté de l'apprentissage automatique. Un problème récurrent dans ce domaine est que l'architecture de nombreux systèmes de machine learning actuels rend souvent l'interpétabilité très délicate. Il existe heureusement des équipes de recherche dans le monde entier qui explorent ces questions avec beaucoup de créativité [1], [5], [2], etc. De nombreux résultats extrêmement intéressants ont été publiés récemment et un des délivrables du stage sera un tour d'horizon de ces nouvelles méthodes. Nous nous concentrerons en particulier sur les méthodes permettant de mettre l'humain dans la boucle [3], [4], etc afin de mieux comprendre comment l'apprentissage artificiel peut être efficacement implanté dans les projets

en relation avec les sciences sociales.

Les tâches à effectuer seront les suivantes :

- Prendre connaissance des méthodes exposées dans quelques articles récent de recherche de pointe dans le domaine de l'explicabilité
- Mettre en oeuvre une méthode sur ordinateur
- Rédiger un rapport faisant le point sur les méthodes étudiées et relatant les performances de la méthode mise en oeuvre.

References

[1] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[2] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE, 2018.

[3] Isaac Lage, Andrew Ross, Samuel J Gershman, Been Kim, and Finale Doshi-Velez. Human-in-the-loop interpretability prior. In *Advances in neural information processing systems*, pages 10159–10168, 2018.

[4] Teodora Popordanoska, Mohit Kumar, and Stefano Teso. Machine guides, human supervises: Interactive learning with global explanations. *arXiv preprint arXiv:2009.09723*, 2020.

[5] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*, 2017.