Interpretability of Machine Learning algorithms: how these techniques can correctly guess the physical laws?

Marzio De Corato, Alfio Ferrara and Silvia Salini

Abstract The Machine learning algorithms (MLA) provide a formidable tool for making progress among different sciences [1]. Among them, remarkable results were obtained for physical sciences [2]; however, despite the high accuracy in predictions that can be obtained with these algorithms, using them for base scientific research also requires to have an interpretation of their machinery. Furthermore, it is worth mentioning that, apart from being a requirement for scientific purposes [2], interpretability is a requirement imposed on algorithms by the GDPR [3]. Moreover, as shown by Miller in [4], the interpretability of a MLA is strictly connected to finding the causal connection between the features analysed: therefore, if one is interested in going beyond the statistical correlation, he/she has to face how to make the MLA used interpretable [5]. While for some MLA, the interpretation is straightforward, for instance, in the case of linear regression, for others, like the neural networks and the support vector machines, such insight seems less evident. The interpretability issue was faced previously by a restricted set of authors ([3, 4, 6] and Ref. therein) with respect to the community that uses the MLA algorithm. In this study, we propose a systematic investigation of how a selected set of MLA algorithms can capture the generating laws for an input dataset. For this purpose, we started with datasets generated by a physical law or from real data (both taken from astronomy). While for the first case, the public datasets were considered, such as the NASA dataset of exoplanets [7] as well the hazardous asteroids [8], for the second case, the data were generated starting, for instance, from the gravitational law.

Marzio De Corato
Università degli Studi di Milano - Dipartimento di Informatica, Via Giovanni Celoria, 18, 20133 Milano MI, e-mail: marzio.decorato@unimi.it

Silvia Salini
Università degli Studi di Milano - Dipartimento di Economia, Management e Metodi Quantitativi, Via Conservatorio, 7 20122 MILANO (MI) e-mail: silvia.salini@unimi.it

Alfio Ferrara
Università degli Studi di Milano - Dipartimento di Informatica, Via Giovanni Celoria, 18, 20133 Milano MI, e-mail: alfio.ferrara@unimi.it

In this last case, other features were considered: in particular, these were generated with a different type of noise added to the correct input features. In the end, for these cases, we have datasets for which the underlying generating laws are known. Once prepared these datasets, an output variable was considered based on the known laws. After these steps, the following MLA algorithms were considered for the analysis: Neural networks (with different architectures), Support Vector Machines, Logistic Regression, Quadratic Discriminant Analysis, Random Forest [9], and graphical models [10]. After the mentioned algorithms were trained and tested, we considered the standard interpretation techniques [11] such as the Partial Dependence Plots, as implemented in the *iml* R package [12] to get an insight into the machinery of the algorithms considered. This outcome was compared with the prior knowledge about the generating law of the datasets. In this way, one obtains an assessment of the algorithms' accuracy and how well these approximate the underlying generating law. Given such validation on how the MLA correctly guess the physics of the input dataset, one can consider moving more safely on a real dataset in which the underlying laws are less known.

Key words: Algorithm intepretability, Machine Learning algorithms, ML algorithms for physics

References

1. Eric Mjolsness and Dennis DeCoste. Machine learning for science: state of the art and future prospects. *science*, 293(5537):2051–2055, 2001.
2. Giuseppe Carleo, Ignacio Cirac, Kyle Cranmer, Laurent Daudet, Maria Schuld, Naftali Tishby, Leslie Vogt-Maranto, and Lenka Zdeborová. Machine learning and the physical sciences. *Reviews of Modern Physics*, 91(4):045002, 2019.
3. Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.
4. Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019.
5. Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.
6. W James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080, 2019.
7. `https://exoplanetarchive.ipac.caltech.edu/applications/DocSet/index.html?doctree=/docs/docmenu.xml&startdoc=1`.
8. `https://cneos.jpl.nasa.gov/about/neo_groups.html`.
9. Max Kuhn. Caret: classification and regression training. *Astrophysics Source Code Library*, pages ascl–1505, 2015.
10. Jonas Haslbeck and Lourens J Waldorp. mgm: Estimating time-varying mixed graphical models in high-dimensional data. *arXiv preprint arXiv:1510.06871*, 2015.
11. Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.
12. Christoph Molnar, Bernd Bischl, and Giuseppe Casalicchio. iml: An r package for interpretable machine learning. *JOSS*, 3(26):786, 2018.