Aspects of Transparency in Machine Learning

Doctoral Consortium

Martin Strobel
National University of Singapore
mstrobel@comp.nus.edu.sg

ABSTRACT

The fact that machine learning is growing more and more entrenched in almost every aspect of society, combined with the opacity of various of its algorithms has induced the relatively young research area of transparent machine learning. The aim of this domain is to provide explanations for automated decisions to increase public trust. In my thesis, I am going to consider certain problems that arise from this research agenda. Particularly, I so far considered, the dilemma of conflicting explanations and the issue of privacy concerns arising from transparency.

KEYWORDS

Cooperative games: theory & analysis;Social choice theory; Game theory for practical applications; Values in MAS (privacy, safety, security, transparency, . . .)

ACM Reference Format:
Martin Strobel. 2019. Aspects of Transparency in Machine Learning. In *Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019,* IFAAMAS, 3 pages.

1 INTRODUCTION

Recent years have seen the widespread implementation of data-driven algorithms making decisions in increasingly high-stakes domains, such as finance, healthcare, transportation, and public safety. Using novel ML techniques, these algorithms are capable of processing massive amounts of data and produce highly accurate predictions; however, their inherent complexity makes it increasingly difficult for humans to understand certain decisions. Indeed, these algorithms are *black-box decision makers*: their underlying decision processes are either hidden from human scrutiny by proprietary law or (as is often the case) their inner workings are so complicated that even their designers will be hard-pressed to explain the reasoning behind the algorithms decision-making processes. By obfuscating their function, data-driven classifiers run the prospect of exposing human stakeholders to risks. These may include incorrect decisions (e.g. a loan application that was wrongly rejected due to system error), information leaks (e.g. an algorithm inadvertently uses information it should not have used), or discrimination (e.g. biased decisions against certain ethnic or gender groups). Government bodies and regulatory authorities have recently begun calling for *algorithmic transparency*: providing human-interpretable explanations of the underlying reasoning behind large-scale decision-making algorithms. My thesis research will be concerned with issues that arise from this research agenda. Especially, I'm interested in how to decide which explanation of a decision to trust given that there are many, potentially conflicting, possible explanations for any given decision. I'm also interested in how transparency is in conflict with other desirable objectives for machine learning an initial example is the privacy of the training data

1.1 Monotone Influence Measures

In our initial work [16], we investigated *influence measures*: these are functions that, given a dataset, assign a value to every feature; this value should roughly correspond to the feature's importance in affecting the classification outcome for individual data points. We identified a set of axioms that any reasonable influence measure should satisfy. Given the space constraints, here only a very brief overview of what these axioms look like: some are concerned with geometric manipulation of the data set i.e. behavior of the measure under rotation or shifting of the data; we also considered axioms concerning continuity and a form of monotonicity. From these axioms, we derived a class of influence measures, dubbed *monotone influence measures* (MIM), which uniquely satisfied these axioms. Moreover, we showed that MIM is interpretable as the optimal solution to a natural optimization problem. Unlike most influence measures in the literature, we assumed neither knowledge of the underlying decision-making algorithm nor of its behavior on points outside the dataset. Indeed, some methodologies are heavily reliant on access to counterfactual information: what would the classifier have done if some features were changed? This may be a strong assumption in some cases, as it assumes not only access to the classifier but also the potential ability to use it on nonsensical data.[1] Further, we conducted an analysis of several existing measures based on our axioms, showing which of the axioms are satisfied by existing measures and how they could be improved accordingly. Finally, we showed that despite our rather limiting conceptual framework, MIM does surprisingly well on a sparse image dataset, and provides an interesting analysis of a recidivism dataset. We showed that the outputs

Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), N. Agmon, M. E. Taylor, E. Elkind, M. Veloso (eds.), May 13–17, 2019, Montreal, Canada.

[1] For example, if the dataset consists of medical records of men and women, the classifier might need to answer how it would handle pregnant men

of MIM are comparable to those of other measures, and provide interpretable results.

1.2 Transparency in Conflict with Privacy

[14] have shown that predictions of machine learning models can be exploited to infer if a certain data point was used to train the model. This privacy violation can have serious consequences. In ongoing work, we asked if commonly used transparency measures leak similar information about the training data. Our analysis indicates that this is indeed the case in for specific explanations in certain situations.

1.3 Related Work

Algorithmic transparency has been debated and called for by government bodies [8, 17], the legal community [12, 19], and the media [2, 7]. The AI and ML research community is part of the conversation: several ongoing research efforts are informing the design of explainable AI systems (e.g. [9, 22]), as well as tools that explain the behavior of existing black-box systems (see [20] for an overview); our initial work focuses on the latter.

Existing results closely related to our initial work are from Datta et al.. They axiomatically characterize an influence measure for datasets; however, in their work influence is interpreted as a global measure (e.g., what importance had 'age' for all decisions as a whole); we focused on feature importance for individual data points. Further, it has been shown by Datta et al. that the measure proposed by Datta et al. outputs undesirable values (e.g. zero influence) in many real instances; this is due to the fact that the Datta et al. measure relies on the existence of potentially counterfactual data: data points that differ from one another by only a single feature. This becomes especially problematic in situations with many features or sparse data. A data-based influence measure relying on a potential like approach has been proposed by Baehrens et al.. However, we could demonstrate that their approach fails to satisfy reasonable properties even on basic datasets.

Another stream of research assumes access to the classifier, which allows querying classifications for additional data points. Datta et al. use an axiomatically justified approach based on an economic paradigm of fairness to measure influence, called QII; briefly, QII perturbs feature values and observes the effect this has on the classification outcome. Another line of work using black-box access [11] uses queries to the classifier in a local region near the point of interest in order to measure influence. Adler et al. equate the influence of a given feature i with the ability to infer i's value from the rest of features, after it has been obscured; this idea is the basis for a framework for auditing black-box models based on statistical analysis. However, this approach assumes that one can make predictions on a dataset with some features removed. Finally, Sundararajan et al. provide a framework for explaining the behavior of black-box systems using a notion of economic fair allocation; however, their analysis assumes that the underlying classifier is a neural network. MIM assumes neither a specific algorithmic framework nor access to counterfactual data. This results in a more generic, albeit less powerful, explanatory framework.

The privacy in machine learning has gotten some attention. [10] use model explanations to efficiently reconstruct a target model, the explanations considered is the gradient with respect to the input and no conclusions are drawn about the privacy of the training set. Our attack falls into the general area of membership inference attacks. [15] train shadow models to infer membership of data points to the training set given query access to basic predictions of models. [21] analyze the effect of over-fitting and influence on the ability to infer membership to the training set.

2 PLANS FOR THE FUTURE

Both aspects touched on in this abstract need further exploration. Axiomatic approaches to influence measurement are common in economic domains. Of particular note are axiomatic approaches in cooperative game theory [4, 13]; we have started exploring the relation of MIM to game-theoretic influence, but there is much more potential in applying game-theoretic concepts in this new domain.

Further, we currently only consider binary classifications, a generalization into a multi-class our even regression domain is desirable and far from trivial. Besides the generalization of our axioms, it also requires a discussion of what 'closeness' means in those situations and what accounts as a positive or negative influence. Another major limitation of our current work is that it only focuses on single feature influence and largely ignores synergistic effects between features. Here existing work on coalition formation in cooperative game theory might help us to obtain further insights. Nevertheless, to axiomatize the pairwise interactions between features would be a major theoretical challenge.

Finally, these potential new measures would surely be more involved, and so harder to understand for humans. The study of this trade-off between understandability and explanatory power is another question we want to further analyze.

For the conflict between privacy and transparency, besides concluding our current work with formal results, it would be of interest to consider countermeasures or to focus on the existing measures which are robust against our attacks, showing formal privacy guarantees.

ACKNOWLEDGMENTS

The author would like to thank his supervisor Prof. Yair Zick and Prof. Reza Shokri for their constant support and guidance.

REFERENCES

[1] P Adler, C Falk, S A Friedler, G Rybeck, C Scheidegger, B Smith, and S Venkatasubramanian. 2018. Auditing black-box models for indirect influence. In *Knowledge and Information Systems*.

[2] J Angwin, J Larson, S Mattu, and L Kirchner. 2016. Machine Bias There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica* (may 2016). https://goo.gl/sRA6bg

[3] David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Mueller. 2009. How to Explain Individual Classification Decisions. *Journal of Machine Learning Research* 11 (2009). arXiv:0912.1128 http://arxiv.org/abs/0912.1128

[4] J F Banzhaf. 1965. Weighted Voting Doesn't Work: a Mathematical Analysis. *Rutgers Law Review* 19 (1965), 317–343.

[5] A Datta, A Datta, A D Procaccia, and Y Zick. 2015. Influence in Classification via Cooperative Game Theory. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI)*.

[6] A Datta, S Sen, and Y Zick. 2016. Algorithmic Transparency via Quantitative Input Influence. In *Proceedings of the 37th IEEE Conference on Security and Privacy (Oakland)*.

[7] J M Hofman, A Sharma, and D J Watts. 2017. Prediction and explanation in social systems. *Science* 355, 6324 (2017), 486–488.

[8] F Hollande. 2016. Pour une République numérique (1). (oct 2016). https://goo.gl/jnE1tQ

[9] J A Kroll, J Huey, S Barocas, E Felten, J R Reidenberg, D G Robinson, and and H. Yu. 2017. Accountable Algorithms. *University of Pennsylvania Law Review* 165, 3 (2017).

[10] Smitha Milli, Ludwig Schmidt, Anca D. Dragan, and Moritz Hardt. 2018. Model Reconstruction from Model Explanations. *arXiv preprint arXiv:1807.05185* (2018). arXiv:1807.05185 http://arxiv.org/abs/1807.05185

[11] M T Ribeiro, S Singh, and C Guestrin. 2016. Why Should I Trust You?: Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd International Conference on Knowledge Discovery and Data mining*.

[12] C J Roggensack and J Abrahamson. 2016. Wisconsin v. Loomis. (2016). https://goo.gl/S35mD6

[13] L S Shapley. 1953. A Value for n-Person Games. In *Contributions to the Theory of Games, vol. 2*. Princeton University Press, 307–317.

[14] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership Inference Attacks Against Machine Learning Models. *Proceedings - IEEE Symposium on Security and Privacy* (2017), 3–18. https://doi.org/10.1109/SP.2017.41 arXiv:1610.05820

[15] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership Inference Attacks Against Machine Learning Models. In *Proceedings of the 38th IEEE Conference on Security and Privacy (Oakland)*. https://doi.org/10.1109/SP.2017.41 arXiv:1610.05820

[16] J Sliwinski, M Strobel, and Y Zick. 2019. Axiomatic Characterization of Data-Driven Influence Measures for Classification. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI)*. arXiv:1708.02153 http://arxiv.org/abs/1708.02153

[17] M Smith, D Patil, and Muñoz C. 2016. Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights. White House Report. https://goo.gl/WbgMnK

[18] M Sundararajan, A Taly, and Q Yan. 2017. Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*.

[19] N Suzor. 2015. Google defamation case highlights complex jurisdiction problem. *The Conversation* (oct 2015). https://goo.gl/XS1Z6y

[20] A Weller. 2017. Challenges for Transparency. *CoRR* abs/1708.0 (2017).

[21] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. 2017. Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting. *arXiv preprint arXiv:1709.01604* (2017). arXiv:1709.01604 http://arxiv.org/abs/1709.01604

[22] J Zeng, B Ustun, and C Rudin. 2017. Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 180, 3 (2017), 689–722.