An axiomatic approach to explain Computer generated decisions (Extended Abstract)

Martin Strobel
National University of Singapore

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 able to process massive amounts of data and make highly accurate predictions; however, their inherent complexity makes it increasingly difficult for humans to understand *why* certain decisions were made. 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 own designers will be hard-pressed to explain the underlying reasoning behind their decision making processes. By obfuscating their function, data-driven classifiers run the risk 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 an axiomatic analysis of automatically generated explanations of such classifiers. 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.

Work done so far

In our initial work (J. Sliwinski 2017), 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 datapoints. We identified a set of axioms that any reasonable influence measure should satisfy. Given the space constraints, here only a very brief overview of the what these axioms look like. Some where concerned with geometric manipulation of the data set i.e. behaviour of the measure under rotation or shifting of the data points, but 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. I significantly contributed to this part of our work. Moreover, we showed that MIM can be interpreted as the optimal solution to a natural optimization problem. Unlike most existing 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 having 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 points[1]. Further, I conducted an initial analysis of some 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 of MIM are comparable to those of other measures, and provide interpretable results.

Related Work

Algorithmic transparency has been debated and called for by government bodies (Hollande 2016; Smith, Patil, and Muoz 2016), the legal community (Roggensack and Abrahamson 2016; Suzor 2015), and the media (Hofman, Sharma, and Watts 2017; Angwin et al. 2016). 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. (Kroll et al. 2017; Zeng, Ustun, and Rudin 2017)), as well as tools that explain the behavior of existing black-box systems (see (Weller 2017) 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 in-

[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

terpreted as a global measure (e.g., what importance had age for all decisions as a whole); we focused on feature importance for individual datapoints. Further, it has been shown by Datta, Sen, and Zick 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: datapoints 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 to query classifiactions for additional datapoints. Datta, Sen, and Zick 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 (Ribeiro, Singh, and Guestrin 2016) 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, Taly, and Yan 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.

Plans for the future

Clearly this is just a initial starting point on which I want to build my Ph.D. research. We are planning on pursuing the following directions. First, axiomatic approaches for influence measurement are common in economic domains. Of particular note are axiomatic approaches in cooperative game theory (Shapley 1953; Banzhaf 1965); 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 what 'closeness' means in those situations and what accounts as positive or negative influence. Another major limitation of our current work is that it only focuses on on single feature influence and largely ignores synergistic effects between features. Here existing works on coalition formation in cooperative game theory might help us to obtain further insights. Nevertheless, to axiomatize the pairwise interactions between features would be major theoretical challenge.

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

References

Adler, P.; Falk, C.; Friedler, S. A.; Rybeck, G.; Scheidegger, C.; Smith, B.; and Venkatasubramanian, S. 2016. Auditing black-box models for indirect influence. In *Proceedings of the 16th IEEE International Conference on Data Mining (ICDM)*, 1–10.

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine bias theres software used across the country to predict future criminals. and it's biased against blacks. *ProPublica*.

Baehrens, D.; Schroeter, T.; Harmeling, S.; Kawanabe, M.; Hansen, K.; and Müller, K.-R. 2010. How to explain individual classification decisions. *Journal of Machine Learning Research* 11:1803–1831.

Banzhaf, J. 1965. Weighted voting doesn't work: a mathematical analysis. *Rutgers Law Review* 19:317–343.

Datta, A.; Datta, A.; Procaccia, A. D.; and Zick, Y. 2015. Influence in classification via cooperative game theory. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI)*.

Datta, A.; Sen, S.; and Zick, Y. 2016. Algorithmic transparency via quantitative input influence. In *Proceedings of the 37th IEEE Conference on Security and Privacy (Oakland)*.

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

Hollande, F. 2016. Pour une république numérique (1). LOI n 2016-1321 NOR: ECFI1524250L.

J. Sliwinski, M. Strobel, Y. 2017. A characterization of monotone influence measures for data classification. *In Proceedings of the IJCAI 2017 Workshop on Explainable Artificial Intelligence (XAI)*.

Kroll, J.; Huey, J.; Barocas, S.; Felten, E.; Reidenberg, J.; Robinson, D.; ; and Yu, H. 2017. Accountable algorithms. *University of Pennsylvania Law Review* 165.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 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 (KDD)*, 1513–1522.

Roggensack, C. J., and Abrahamson, J. 2016. Wisconsin v. Loomis. Case No.: 2015AP157 - CR.

Shapley, L. 1953. A value for n-person games. In *Contributions to the Theory of Games, vol. 2*, Annals of Mathematics Studies, no. 28. Princeton University Press. 307–317.

Smith, M.; Patil, D.; and Muoz, C. 2016. Big data: A report on algorithmic systems, opportunity, and civil rights. White House Report.

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic attribution for deep networks. *arXiv preprint arXiv:1703.01365*.

Suzor, N. 2015. Google defamation case highlights complex jurisdiction problem. *The Conversation*.

Weller, A. 2017. Challenges for transparency. *CoRR* abs/1708.01870.

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