No Explanation without Inference

David S. Watson[1]

Abstract. Complex algorithms are increasingly used to automate high-stakes decisions in sensitive areas like healthcare and finance. However, the opacity of such models raises problems of intelligibility and trust. Researchers in interpretable machine learning (iML) have proposed a number of solutions, including local linear approximations, rule lists, and counterfactuals. I argue that all three methods share the same fundamental flaw – namely, a disregard for *severe testing*. Techniques for quantifying uncertainty and error are central to scientific explanation, yet iML has largely ignored this methodological imperative. I consider examples that illustrate the dangers of such negligence, with an emphasis on issues of scoping and confounding. Drawing on recent work in philosophy of science, I conclude that there can be no explanation – algorithmic or otherwise – without inference. I propose several ways to severely test existing iML methods and evaluate the resulting trade-offs.

1 Introduction

Machine learning (ML) is increasingly ubiquitous in modern society. Complex algorithms are widely deployed in private industries like finance [3], as well as public services such as healthcare [20]. Their prevalence is driven by results. ML models outperform humans not just at strategy games like chess [17], but at important scientific tasks like antibiotic discovery [19] and tumor diagnosis [10].

High-performance algorithms are often opaque, in the sense that it is difficult for humans to understand the internal logic behind individual predictions. This raises fundamental issues of trust. How can we be sure a model is right when we have no idea why it predicts particular values? While model interpretation is by no means a new concern in statistics, it is only in the last few years that a dedicated subfield has emerged to address the issues surrounding algorithmic opacity.

Interpretable machine learning (iML) comprises a diverse collection of technical approaches intended to render statistical predictions more intelligible to humans [11]. My focus here is on model-agnostic, post-hoc local methods, which explain the individual predictions of some target model without making any assumptions about its form. Prominent examples include local linear approximators (e.g., SHAP [6]), which produce feature attributions that sum to the explanandum; rule lists (e.g., Anchors [15]), which provide explanations via sequences of if-then statements; and counterfactuals (e.g., MACE [4]), which identify one or several nearest neighbors on the opposite side of a decision boundary. Despite their merits, all three approaches fail to meet the severity criteria outlined in Sect. 2. I illustrate the issues with this failure in Sect. 3, and propose some directions for improvement. I conclude in Sect. 4 with a reflection on the trade-offs implied by this analysis.

[1] University College London, London, United Kingdom. Email: david.watson@ucl.ac.uk

2 Severe Testing

Mayo [9, 8, 7] argues that the problem of induction is defeasibly resolved by severe testing. The basis for this resolution is her severity principle, which states that "We have evidence for a claim C just to the extent it survives a stringent scrutiny. If C passes a test that was highly capable of finding flaws or discrepancies from C, and yet none or few are found, then the passing result, x, is evidence for C" [7, p. 14]. On Mayo's view, the justification for believing a given hypothesis is a function not of the hypothesis itself or the data it purportedly explains, so much as the tests it has passed. When tests are sufficiently sensitive (i.e., likely to detect true effects) and specific (i.e., likely to reject false effects), then we say they are *severe*.

To make matters concrete, consider a single parameter location test. Let Θ denote the parameter space, and let T be a test that decides between $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_1$, where Θ_0 and Θ_1 are some partition of Θ. We observe sample data x and compute sufficient statistic $d(x)$, which measures the disagreement between x and H_0. Test T rejects H_0 when $d(x)$ meets or exceeds the critical value c_α. We say that H_0 passes an (α, β)-severe test T with data x if:

(S1) $d(x) < c_\alpha$; and
(S2) with probability at least $1 - \beta$, if H_1 were true, then we would observe some sufficient statistic $d(x')$ such that $d(x') \geq c_\alpha$.

Readers well-versed in frequentist inference will recognize some familiar concepts here. The critical value is indexed by the type I error rate α, such that, under H_0, the rejection region of statistics greater than or equal to c_α integrates to α. Under H_1, the rejection region of statistics less than c_α integrates to the type II error rate, β. The complement of this value, $1 - \beta$, denotes the power of the test. A test with small α is said to be specific, since it only accepts hypotheses that are likely to be true; a test with small β is said to be sensitive, since it is able to detect even slight deviations from the null.

While this explication is faithful to the frequentist framework that Mayo favors, the severity criteria are in fact very general, and have been reformulated along Bayesian lines [2]. ML is not inherently aligned with any particular interpretation of probability, and nothing in the proceeding argument depends upon one's preferred method of inference. The epistemological upshot of Mayo's analysis is that science advances knowledge not just by falsifying theories, as Popper would have it [12], but by subjecting hypotheses to increasingly severe tests. Hypotheses earn their warrant by passing such tests, thereby providing positive justification for successful theories.

3 Severity and iML

An algorithmic explanation is an empirical claim relating certain factors in the input data to the resulting prediction. Since empirical claims are typically the realm of science, we may justifiably wonder

whether Mayo's severity criteria can be fruitfully applied in this setting. I argue that they can and should. I highlight two ways that algorithmic explanations mislead when severity criteria are not taken into account: through ambiguity of scope and sensitivity to confounding.

Local explanations are constructed to apply only in some fixed region of the feature space. Yet iML methods do not generally provide information about the bounds of a given explanation or goodness of fit within the target region. For illustration, I will focus on linear approximators, but the point applies more broadly.

If you zoom in far enough to any point on a continuous function, you will eventually find a linear tangent. This is the intuition behind methods like LIME [14] and SHAP [6]. However, when the regression surface or decision boundary around the target point is extremely nonlinear, the linear region tends to be very small and the estimated coefficients highly unstable. In this case, feature attributions are acutely sensitive to regional bounds. In a simple two-dimensional example, Wachter et al. [21] visually demonstrate how a linear explanation for the same model prediction may assign positive, negative, or zero weight to a feature depending on the scope of the linear window (see Fig. 1).

The most obvious statistical solution here would be to augment iML outputs with information regarding the scope and fit of the approximation. It is common, for instance, in linear regression to compute the significance and standard error of model coefficients. This would satisfy (S1). Power analysis typically requires parametric assumptions or data simulations, which could be used to satisfy (S2). Unfortunately, these strategies are not readily available to algorithms like LIME and SHAP, which use unconventional sampling techniques, kernel weights, and regularization penalties that preclude easy analytic solutions for calculating expected error rates. Nonparametric resampling methods could help but at major computational cost. The problem becomes especially acute as the number of explananda increases.

Another challenge for iML arises when features are highly dependent. The issue can be especially nefarious when auditing for algorithmic bias. If a sensitive attribute is associated with a permissible variable (e.g., if race is well predicted by zip code) then the latter can serve as a proxy for the former. This allows bad actors to get away with discrimination, so long as they can fool an auditor into believing they were using the permissible variable rather than the sensitive one. The concern is not merely speculative. Authors have exploited these vulnerabilities to make discriminatory models pass algorithmic audits [18] and appear fair in user studies [13].

Severe testing cannot, on its own, prevent bad actors from engaging in discriminatory behavior. However, it can make it harder for them to get away with it by elucidating the uncertainty associated with algorithmic explanations under confounding. Just as standard errors for regression coefficients are inflated by collinear predictors, the severity of particular explanations will tend to decrease with strongly correlated features. Reporting the error rates of given outputs will provide much-needed context for users and regulators alike.

Algorithmic fairness is a complex and contested topic. Dozens of statistical fairness criteria have been proposed [1], while impossibility theorems have shown that most popular definitions are mutually incompatible except in trivial cases [5]. No matter which criteria one adopts for a given application, almost all may be expressed in terms of marginal or conditional independences, which means that classical tests can be used for auditing purposes. Severity therefore has a central role to play in holding people and institutions accountable for their algorithmically mediated decisions.

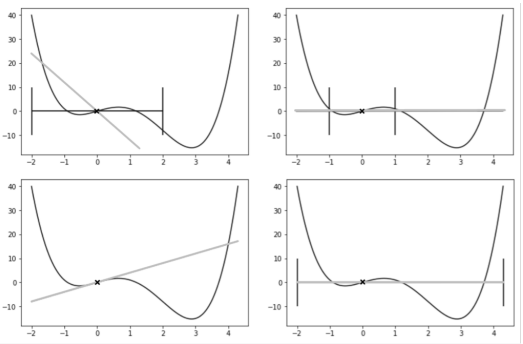

Figure 1: Unstable linear approximations. The grey line in each panel shows a local approximation of the same function centered at the same location. The varying range is indicated by the black bars, leading to vastly different linear explanations. From [21, p. 885].

4 Discussion

Many authors motivate the iML project with appeals to trust. "Why should I trust you?" reads the title of Ribeiro et al.'s paper introducing LIME [14]. "Building trust is essential to increase societal acceptance of algorithmic decision-making," [21, p. 843] write Wachter et al. in their paper on counterfactual explanations. So long as complex algorithms remain opaque, users will harbor suspicions about their reliability in particular cases. That is why we seek transparent explanations that can assuage concerns about unfair or unreasonable model predictions.

But do iML algorithms really settle matters, or merely push the problem one rung up the ladder? After all, why should we trust their outputs? Presumably the target function at least has the advantage of performing well on some test dataset. Can we say the same of algorithms like SHAP, Anchors, or MACE? Their outputs are readily intelligible, and that is clearly a start. But does that necessarily mean that their explanations should all be given equal weight, or are some more reliable than others? How can we be sure that they have not produced unstable estimates or selected the wrong features? Are there principled methods for critically evaluating individual explanations, much like we can critically evaluate individual predictions?

I argue that severe testing holds the key to securing the trustworthiness of algorithmic explanations. The goal of all iML algorithms is to produce claims relating inputs to outputs. Such claims can in principle be tested. That, for instance, is how we come to trust scientific theories – by repeatedly, mercilessly subjecting them to severe tests with quantifiable error rates. There is no good reason to hold iML to a lesser standard.

Concerns over feasibility are legitimate. Bootstrapping methods for evaluating the scope and stability of local explanations could be time consuming. Conditional independence testing, which may aid in fairness audits, is notoriously difficult in high-dimensional settings and provably hard for continuous conditioning events [16]. But if the stakes are sufficiently high that we need an algorithmic explanation in the first place – perhaps even a legally mandated one – then it is important that we get that explanation right.

Proponents of black box algorithms argue that results often matter above all else. Would we prefer a transparent model that diagnoses cancer with 90% accuracy or an opaque one that does so with 99% accuracy? By the same token, we cannot dismiss severe testing for iML merely due to concerns about the computational burden. When consequential decisions depend upon algorithmic explanations, we had better make sure they withstand a stringent scrutiny.

ACKNOWLEDGEMENTS

Thanks to Luciano Floridi and Carl Öhman for their valuable feedback. This research was partially supported by ONR grant N62909-19-1-2096.

REFERENCES

[1] Solon Barocas, Moritz Hardt, and Arvind Narayanan, *Fairness and Machine Learning*, fairmlbook.org, 2019.
[2] Andrew Gelman and Cosma Rohilla Shalizi, 'Philosophy and the practice of Bayesian statistics', *British Journal of Mathematical and Statistical Psychology*, **66**(1), 8–38, (2013).
[3] J B Heaton, N G Polson, and J H Witte, 'Deep learning for finance: deep portfolios', *Applied Stochastic Models in Business and Industry*, **33**(1), 3–12, (2017).
[4] Amir-Hossein Karimi, Gilles Barthe, Borja Balle, and Isabel Valera, 'Model-Agnostic Counterfactual Explanations for Consequential Decisions', in *The 23rd International Conference on Artificial Intelligence and Statistics*, volume 108, pp. 895–905, (2020).
[5] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan, 'Inherent Trade-Offs in the Fair Determination of Risk Scores', in *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, volume 67, pp. 43.1–43.23, (2017).
[6] Scott M Lundberg and Su-In Lee, 'A Unified Approach to Interpreting Model Predictions', in *Advances in Neural Information Processing Systems 30*, 4765–4774, (2017).
[7] Deborah Mayo, *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars*, Cambridge University Press, New York, 2018.
[8] *Error and Inference*, eds., Deborah Mayo and Aris Spanos, Cambridge University Press, New York, 2010.
[9] Deborah G Mayo, *Error and the Growth of Experimental Knowledge*, University of Chicago Press, Chicago, 1996.
[10] Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, and Shravya Ashrafian, Hutan...Shetty, 'International evaluation of an AI system for breast cancer screening', *Nature*, **577**(7788), 89–94, (2020).
[11] Christoph Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Interpretable*, christophm.github.io/interpretable-ml-book/, Münich, 2021.
[12] Karl Popper, *The Logic of Scientific Discovery*, Routledge, London, 1959.
[13] Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, and Zachary C Lipton, 'Learning to Deceive with Attention-Based Explanations', in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4782–4793, (2020).
[14] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin, '"Why Should I Trust You?": Explaining the Predictions of Any Classifier', in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 1135–1144. ACM, (2016).
[15] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin, 'Anchors: High-Precision Model-Agnostic Explanations', in *AAAI*, pp. 1527–1535, (2018).
[16] Rajen Shah and Jonas Peters, 'The Hardness of Conditional Independence Testing and the Generalised Covariance Measure', *Annals of Statistics*, **48**(3), 1514–1538, (2020).
[17] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, and Demis Guez, Arthur...Hassabis, 'A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play', *Science*, **362**(6419), 1140 LP – 1144, (2018).
[18] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju, 'Fooling LIME and SHAP: Adversarial Attacks on Post Hoc Explanation Methods', in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 180–186, (2020).
[19] Jonathan M Stokes, Kevin Yang, Kyle Swanson, Wengong Jin, Andres Cubillos-Ruiz, and James J Donghia, Nina M...Collins, 'A Deep Learning Approach to Antibiotic Discovery', *Cell*, **180**(4), 688–702.e13, (2020).
[20] Eric J Topol, *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*, Basic Books, New York, 2019.
[21] Sandra Wachter, Brent Mittelstadt, and Chris Russell, 'Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR', *Harvard Journal of Law and Technology*, **31**(2), 841–887, (2018).