phg foundation
making science work for health

Why explainable machine learning matters for health

Discussion paper

UNIVERSITY OF CAMBRIDGE

Authors

Johan Ordish and Alison Hall

Acknowledgements

This work was supported by the Wellcome Trust, grant number: 213623/Z/18/Z

URLs in this paper were correct as of November 2019

Published by PHG Foundation
2 Worts Causeway
Cambridge
CB1 8RN
UK
+44 (0)1223 761900

May 2019

Correspondence to:
intelligence@phgfoundation.org

How to reference this publication
Why explainable machine learning matters for health
PHG Foundation (2019)

Why explainable machine learning matters for health

Machine learning promises to change the way we diagnose and ultimately treat patients. However, some argue that machine learning for health also threatens to usher in an age of black box medicine, where 'opaque computational models make decisions related to healthcare.'[1] This paper explores the use of machine learning for health, stating to what extent and why machine learning models might be opaque, noting why human interpretability of machine learning matters, and outlining the different ways in which machine learning models can be interpretable to humans. We conclude that while not all machine learning models are black boxes, interpretability of machine learning models will often be important when providing proper assurances that a model is safe and effective.

Summary

- Machine learning models vary in the extent to which they are interpretable - ranging from those that are intrinsically human interpretable to black boxes that are not intrinsically interpretable to humans
- Black box models may be made somewhat human interpretable through the use of post hoc explainers that explain a particular decision of the model (local interpretability) and/or how the model functions generally (global interpretability)
- Post hoc explainers have weaknesses and are ultimately only an estimation of an underlying black box model. Given this, it is unclear where and when we should demand the use of intrinsically interpretable machine learning

A call for interpretability - two examples to demonstrate why interpretability is important:

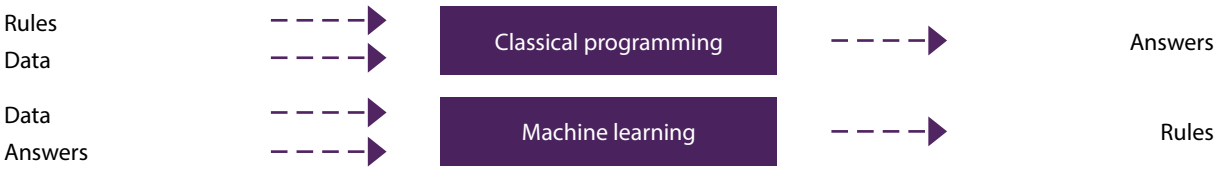
Caruana *et al* (2015) describe a series of models to predict the probability of death for patients with pneumonia.[2] The group found that neural networks produced the most accurate models. However, when the group trained in parallel a less accurate but interpretable rule-based model, the group found that this model learned the following rule: 'HasAsthma(x) ⇒LowerRisk(x).' Consequently, it was shown that a confounding variable influenced the neural networks, the models correctly identifying that those with asthma were less likely to die but only because as a group they were more likely to receive treatment.

Zech *et al* (2015) trained a convolutional neural network to screen for pneumonia using x-rays.[3] Subsequent manual image review noticed that the model was able to differentiate between those x-rays taken by portable scanner (identified by the word 'portable' and inversion of colour in the x-ray) and those by static scanner, the model finding this distinction significant, portable scanners being used in the emergency department but not for inpatient units. Consequently, when the model found the word 'portable' significant it introduced a potentially confounding factor into the screening process.

These two examples illustrate that the mere fact of accuracy may be insufficient; that it is important to know why machine learning models are accurate.

What is machine learning?

Machine learning describes an approach to programming that typically produces algorithms with bounded, task-specific intelligence. In a phrase, machine learning algorithms are narrowly intelligent (they do one thing well), but broadly unintelligent (lacking broad capacity to reason). How does machine learning differ from classical programming? Classical programming combines rules and data to provide answers. Machine learning combines data and answers to provide the rules (see diagram below). Machine learning models are trained with many examples (data) relevant to the task, the system finding structure in these examples to provide rules to automate the task.[4]



A potential disadvantage of using these tools is that some machine learning models may be black boxes.

Black boxes and human interpretability

Black box models are models 'whose internal workings are either unknown to the observer or known but uninterpretable to humans.'[5] In short, because the model has been trained rather than explicitly programmed it may be difficult to explain why a machine learning model generated a certain output or to understand what the model finds significant.

Core to the definition of 'black box' is interpretability. Interpretability has different definitions across different domains. In the machine learning context, a useful definition comes from Miller (2017):

> '*Interpretability* is the degree to which a human can understand the cause of a decision.'[6]

Why is interpretability important? How does the problem of interpretability arise?

Interpretability and incompleteness

Arguably, the problem of interpretability arises because many machine learning problems are (often necessarily) incomplete.[7] In this sense, an incomplete problem leaves a gap which creates the need for a machine learning model to be interpretable.

We should distinguish between the following:

- *Uncertainty*, meaning quantified variance that can be formalised

  For example, false positive rates, confidence intervals

- *Incompleteness*, meaning the problem includes elements that have not been formalised and quantified

  For example, scientific discovery and diagnosis make causative inferences, but causation cannot be formalised or quantified, so scientific discovery and diagnosis remain an incomplete problem for machine learning to address

Many useful machine learning models in health tackle problems that are incomplete and will remain incomplete for the foreseeable future. As a consequence, machine learning for health will have a gap which interpretability may have to bridge. Furthermore, even where a machine learning problem is virtually complete, its implementation as a device or service may still require interpretability. For instance, concepts like 'safety' cannot fully be quantified and so if a model requires some form of human input or the results need to be put into further context by a human, the interpretability of the model may remain important.

When to demand interpretable machine learning

When should we require that a machine learning model be human interpretable? Doshi-Velez *et al* (2017) suggest that explanation of an incomplete model may be unnecessary where either a) 'there is no significant consequences for unacceptable results' or b) 'the problem is sufficiently well-studied and validated in real applications that we trust the system's decision, even if the system is not perfect.'[8] In the health sector, this might mean that interpretability may be less important for machine learning models that are 'lifestyle/wellbeing devices' (devices that do not have a medical purpose and pose little to no risk to the user).[9] Further, the better studied a problem is, the more confidence we might have in the model and our ability to check for any artefacts or confounding factors, and so the less we might lean on its interpretability. In this way, the closer machine learning models for health get to models that are well-understood and tested (e.g. aircraft avoidance systems), the more content we might be with the accuracy and general evidence base making up for their lack of interpretability.

One concern with Doshi-Velez's account of incompleteness and interpretability is that the account may not fully capture the need to make machine learning models interpretable for other reasons such as fairness, privacy, and other related rights. In this way, we can imagine a fully complete model that produces perfect results but is totally opaque. Arguably, in this situation, an ethical (and possibly legal) obligation to render the model interpretable may still exist. After all, the model may still process sensitive data that may significantly impact the data subject in question.

Dimensions of interpretability

There are different ways in which a machine learning model can be human interpretable or made interpretable. Methods used to make machine learning interpretable can be categorised using a number of different criteria[10]:

- *Intrinsic interpretability or post hoc interpretability* - Is the model intrinsically interpretable due to its simple structure or is a post hoc ('after the fact') method to render the model interpretable necessary?
- *Model-specific or model-agnostic* - Is the explanation specific to the model in question (as with intrinsic interpretability) or is the explanation tool model-agnostic, meaning it can be applied

(theoretically) to explain any machine learning model?

- *Global or local* - Does the interpretability method explain how the model functions in general (global), just an individual decision (local), or a mixture of both?

These dimensions of interpretability as well as their various weaknesses and strengths are outlined below.

Intrinsically interpretable models

Not all machine learning models are black boxes - some techniques are relatively simple and susceptible to human interpretation. The challenge with these models often concerns the communication and visualisation of their decision processes. For instance, decision trees, rules, and linear models are generally recognised as being easily understandable and interpretable for humans[11]. So long as the decision process is accessible to the user, the way the model functions and how a particular decision was arrived at will be human interpretable.

Why not always use intrinsically interpretable models?

The most common argument against insisting upon only intrinsically interpretable models is the supposed trade-off between accuracy and interpretability[12]. Some argue that there is an inverse relationship between the accuracy of a machine learning model and the interpretability of that model. However, this relationship is contentious, some noting that this general proposition remains unevidenced.[13] Nevertheless, it is true that the computational goal of building the most accurate model is not exactly the same as building the most interpretable model.[14] While there might not be an inverse relationship between accuracy and interpretability – in fact, the two concepts can often operate in tandem – there may be a tradeoff to be made between the two at some point.

Post hoc interpretability

If a model is not intrinsically interpretable it is a black box model. These models can be rendered somewhat human interpretable by using methods such as post hoc explainers. These post hoc explainers can explain the overall model (global) or the specific decisions of that model (local) or a combination of both. They may be specific to a particular machine learning model or be model-agnostic, being able to be bolted onto many different models (see table below).

	Global	**Local**
Model-agnostic	Post hoc explanations explain the general function of any given machine learning model	Post hoc explanations explain the specific decisions of any given machine learning model
Model-specific	Post hoc explanations explain the function of the model but are specific to only this machine learning model	Post hoc explanations explain the specific decision of the model but are specific only to this machine learning model

Generally, model-agnostic explainers work by treating the underlying machine learning model as a black box, testing the relationship between inputs and outputs to approximate a view of what the model finds significant generally or in relation to a particular decision. Model-agnostic methods are particularly powerful as they can often be bolted on to elucidate the inner workings of what would otherwise remain an opaque machine learning model.

The weaknesses of black box models and black box explainers

Post hoc explainers to interpret black box models are promising but have limitations. There are three main issues with using post hoc explainers to explain black box models:

1. **Fidelity**. post hoc explainers often approximate the underlying machine learning model to explain its contents. Since these explainers estimate the underlying model they may provide inaccurate answers, especially if these explainers are highly localised and taken outside their local context[15]
2. **Partial explanations.** Even if the post hoc explanation generated is correct, it may be incomplete and (potentially) instil a false sense of confidence.[16] For example, saliency maps provide a heatmap overlay of an image, demonstrating what part of the image the model found relevant. However, knowing where the model is looking does not tell us what the model is doing with that part of the image
3. **Calibration of machine learning models.** If the underlying machine learning model is a black box, it is difficult to calibrate the model in light of external information not input into the model.[17] If contextual information informs the data underpinning the model, it is often not possible to manually calibrate models that use convolutional neural networks to take account of this discrepancy. For instance, suppose we know that our dataset has a racial bias. If the model trained using this data is a black box it is difficult to manually adjust for this discrepancy without removing data points

Following these three weaknesses, authors like Rudin (2019) emphasise that the gains in interpretability by using intrinsically interpretable machine learning often exceeds the cost of reduced accuracy.[18] That is, while the accuracy loss in choosing an intrinsically interpretable model is low, the gain that interpretability brings usually outweighs this loss. This underlines the point that post hoc explanations for black boxes are not a shortcut to interpretability - they are imperfect and inappropriate in some circumstances.

Why interpretability matters for health

There are special reasons to make machine learning for health human interpretable as a) many of the machine learning problems in the sector will be incomplete, and b) many machine learning applications risk serious consequences if unacceptable results are returned. There are strong practical reasons to provide interpretable models to assure users, regulators, and commissioners that the model is safe and effective. Apart from this, many machine learning models will also process health (or health-related, biometric, or genetic) data, meaning that they will process sensitive personal data to draw their conclusions. Given this, and the importance of the decision at stake, there may be a strong ethical (and possibly legal) imperative to provide an explanation to the user, whether that be a clinician or a patient. In summary, there are often strong practical, ethical, and legal reasons to explain machine learning models.

Further questions

- When, if at all, should we demand the use of intrinsically interpretable machine learning models in health?
- Are post hoc methods to interpret a black box model appropriate for machine learning models that might have serious implications?
- Which of these explanations might satisfy the GDPR's right to explanation? See *A right to explanation* for more information
- Which types of explanation might be most appropriate for patients, clinicians, or consumers?
- Which types of explanation might be most appropriate to generate the trust and confidence of health care professionals who might rely on machine learning applications?

References

1. Price W.N. Black-Box Medicine. Harvard Journal of Law & Technology. 2015; 28(2): 420-467.
2. Caruana R. *et al.* Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30-day Readmission. KDD 2015. 2015; 1721-1730.
3. Zech JR. *et al.* Confounding variables can degrade generalization performance of radiological deep learning models. PLoS Medicine. 2015; 1-15.
4. Chollet F. Deep Learning with Python. Manning Publications. 2017; 2-3.
5. Guidotti R. *et al.* A Survey of Methods for Explaining Black Box Models. ACM Computer Survey. 2019; 51(5): 1-42.
6. Miller T. Explanation in Artificial Intelligence: Insights from the Social Sciences. Artificial Intelligence. 2017; 14.
7. Doshi-Velez F. *et al.* Towards A Rigorous Science of Interpretable Machine Learning. 2017; 1-13.
8. Ibid.
9. Regulation (EU) 2017/745 on medical devices. Recital 19.
10. Molnar C. Interpretable Machine Learning: A Guide for Making Black Box Models Interpretable. 2019; 26-27.
11. Guidotti R. A Survey of Methods for Explaining Black Box Models.
12. DARPA. Broad Agency Announcement: Explainable Artificial Intelligence. DARPA Information Innovation Office. 2016; p14.
13. Rudin C. Please Stop Explaining Black Box Models for High Stakes Decisions. NIPS 2018. arXIV:1811. 10154: 14.
14. Dwork C, *et al.* The Algorithmic Foundations of Differential Privacy. Foundations and Trends in Theoretical Computer Science. 2014; 9(3): 211.
15. Guidotti R. A Survey of Methods for Explaining Black Box Models.
16. Rudin. Please Stop Explaining Black Box Models for High Stakes Decisions. NIPS 2018. arXIV:1811. 10154: 4.
17. Ibid. p5.
18. Ibid. p3.

PHG Foundation
2 Worts Causeway
Cambridge
CB1 8RN
+44 (0) 1223 761900

@phgfoundation
www.phgfoundation.org