Rethinking interpretability

Yaoding Chen

Introduction

Although neural networks have successfully revolutionised multiple domains, e.g., computer vision, they are still doubted in safety-critical applications, such as health diagnosis, credit allowance, or criminal justice, for they are still mostly human-intelligible (Molnar 2020; Camburu 2020). Two approaches are typically used to make these models more plausible, i.e., providing feature-based post-hoc explanations to a model and creating self-explanatory neural models that generate natural language explanations (Camburu 2020). However, Rudin (2019) argues against a post-hoc justification for a black box or a model that lacks a mechanistic basis. Instead, she postulates that there should be a so-called "Rashomon set," which are a set of similarly accurate models containing at least one accurate and plausible model. Yet, any notion of Rashomon sets should be questioned in two ways: 1) can the Rashomon Set be realistically found? 2) more importantly, can best-fitting and interpretable models authentically explain the world we are living in?

An unproven, empirical theory

Intuitively, Rudin (2019)'s notion of Rashomon sets is likely to be true. If the set is large enough, there is likely to be an accurate, plausible model, even if the possibility is low. However, it is not guaranteed that it is true in all cases. However, the theory does not provide any convincing mechanistic explanation within itself. The reason Rudin makes to support her idea is mostly empirical, i.e., it is always possible to find a both highly accurate and interpretable model in most scenarios she works with, ranging from New York City power grid to criminology data, so she postulates that it is possible that the theory is not true in certain scenarios. Also, it is technically difficult to evaluate all known models in a Rashomon set. Also, it is possible that the results that are true in a specific Rashomon set are not robust enough to be generalised. Besides, the theory is statically probable, which means it should not be expected to be true for all scenarios. An analogue here is that although the water on Earth is highly likely to be saline statically since saline water takes up over 97% of Earth's water, yet undoubtedly there still exists freshwater (Shiklomanov 1993), where the high likelihood is not useful. Similarly, the theory may not help in specific scenarios, for interpretability is domain-specific; thus, we are unable to tell whether the model is plausible without a specific context, which makes Rashomon sets less helpful (Martin 2019).

When human fails to interpret things

Practically, it is not always realistic to find a best-fitting interpretable model that outcompetes their black-box counterparts, due to our limitations in term of data availability. For example, in genomic medicine, it is not until in recent years, we have come to take more omics approaches into considerations due to the complex mechanism regulating phenotypic variation, which is beyond our expectation (Boyle et al. 2017). Yet, examining all epigenetic changes and other biological alterations is technically difficult, and even if possible, it will be too expensive, although it is helpful to build a biologically plausible model. Besides economic considerations, Li and Lehner (2020) prove that the phenotypic effect of two mutations combined is not accurately predictable, even when individual effects of a set of mutations and the mechanistic links are well-understood, due to biophysical ambiguity to be measured. Thus, chances are that not only do we fail to have enough data, but we may not fully realise

what data we need to have, which reflects limitations of human knowledge that hinder the establishment of human-intelligible models, especially in poorly understood domains.

Why the Rashomon set is still import?

In the Japanese movie *Rashomon*, a murder is described by different witnesses in four contradictory ways (Davenport 2010). Consider all characters in the movie as functions, which receives the murder as an input, and produce different outputs which all makes sense in a way. The outputs may be confusing or even contradict each other, although they are accurate in some ways. It is hard to guess the story without knowing the context related to each character. Similarly, in statistical modelling, Breiman (2001) coins the term “Rashomon effect,” where he postulates the multiplicity of good models, meaning that for a given dataset, it is possible to construct multiple well-performing models in term of their accuracy, despite their differences in internal structures and attendant explanations (Hancox-Li 2020). Furthermore, Rudin (2019) experimentally shows similar accuracy of popular machine learning models for a structured dataset, whereas complexity does not yield higher accuracy. Therefore, Rudin (2019) asserts that we should choose a more interpretable model, which makes the model more meaningful, reasonable and no less efficient.

Interpretability vs. Explainability: helpful in different ways

However, interpretable models are different from explainable models. Rudin (2019) believes that interpretable models must be inherently interpretable, with a transparent internal structure, while an explainable model, however, often contains two models, i.e., a black-box model and an explanatory model, which must contradict each other. Post-hoc explanations about these models must be at least partly wrong, or else the explanatory model should be used instead. Moreover, a meaningful model choice does not necessarily lead to an explainable model, if we don’t understand the meaning underlying the model. Often, our data often involve extraordinarily complex interactions between variables, which make simple, contrastive explanations that people would like to know highly unlikely. For example, in quantitative genetics, we could identify hundreds of loci responsible for a single trait but the contribution of even the most important loci is only moderate (Boyle et al. 2017), which makes the genetic explanation of a trait less intuitively straightforward. Therefore, Rudin also points out that this conclusion does not apply to complicated, less-structured, noisy data that may lack representativeness, such as dataset used by computer vision, where a lot of different highly accurate models are unusual.

Ultimate consideration: who choose what we believe?

Furthermore, Hancox-Li (2020) and others argue against the robustness of the Rashomon set. A concrete example is made by Dong and Rudin (2020) who construct a Rashomon set for the recidivism data set used in the COMPAS algorithm. They find that lower importance of race is associated with higher importance of age or prior criminal history and vice versa, in the almost-equally accurate predictive models given by the Rashomon set and that race is important only when age or prior criminal history are unimportant. However, although race is less important in some models in the set, it does not mean that it is objectively unimportant; Also, it is questionable how we should choose the explanations for the model. Although Rudin argues that a domain expert could help with model improvement, it is possible that a company deliberately hide all the other model choices, providing an explanation that is likely to be accepted, which may cause a moral crisis. Thus, a description of variable-variable interaction will therefore provide a fuller understanding of the real-world fact that is reflected by the Rashomon set (Hancox-Li 2020).

References

Boyle EA, Li YI, Pritchard JK (2017) An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell 169:1177–1186. https://doi.org/10.1016/j.cell.2017.05.038

Breiman L (2001) Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). Statistical Science 16:199–231. https://doi.org/10.1214/ss/1009213726

Camburu O-M (2020) Explaining Deep Neural Networks. arXiv:201001496 [cs]

Davenport C (2010) Media bias, perspective, and state repression : the Black Panther Party. Cambridge ; New York : Cambridge University Press

Dong J, Rudin C (2020) Variable Importance Clouds: A Way to Explore Variable Importance for the Set of Good Models. arXiv:190103209 [cs, stat]

Hancox-Li L (2020) Robustness in machine learning explanations: does it matter? In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. Association for Computing Machinery, New York, NY, USA, pp 640–647

Li X, Lehner B (2020) Biophysical ambiguities prevent accurate genetic prediction. Nature Communications 11:4923. https://doi.org/10.1038/s41467-020-18694-0

Martin T (2019) Interpretable Machine Learning. Masters Dissertation, University of Cambridge

Molnar C (2020) 2.1 Importance of Interpretability. In: Interpretable Machine Learning

Qiu J, Sun Y (2015) A Research on Machine Learning Methods for Big Data Processing. Atlantis Press, pp 920–928

Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence 1:206–215. https://doi.org/10.1038/s42256-019-0048-x

Shiklomanov I (1993) The distribution of water on, in, and above the Earth. In: Gleick, Peter H. (ed) Water in Crisis: A Guide to the World’s Fresh Water Resources