2.8 Interpretability evaluation

Practical guidance – cross-domain

Authors: Rhys Ward

Evaluation of explanations

‘Unfortunately, there is little consensus on what interpretability in machine learning is and how to evaluate it for benchmarking' [2].

There are at least two different things which must be evaluated with regards to interpretability:

- how appropriate and satisfying to stakeholders are the produced explanations?
- how faithful are the explanations to the actual model workings?

Are explanations appropriate and satisfying to stakeholders?

There is some initial research on how to evaluate the explanations or interpretability of ML models. [4] outlines how levels of explainability can be measured with respect to different user groups. The measures include users' mental model, users' trust and reliance, users' satisfaction and understanding, human-machine task performance, and computational measures. Different measure are useful for describing the effectiveness of explanations to different groups. For example, lay users can be tested on their trust of the system whereas ML experts may seek transparent models. The main point is that different user-groups require different types of explanation.

[2] proposes an evidence-based taxonomy of evaluation approaches for interpretability:

- application-grounded
- human-grounded
- functionally-grounded

They give some examples of ways in which human-understanding can be tested to evaluate the quality of given explanations.

Application level evaluation (real task) puts the explanation into the product and have it tested by the end user. Imagine fracture detection software with a machine learning component that locates and marks fractures in X-rays. At the application level, radiologists would test the fracture detection software directly to evaluate the model. This requires a good experimental setup and an understanding of how to assess quality. A good baseline for this is always how good a human would be at explaining the same decision.

Human level evaluation (simple task) is a simplified application level evaluation. The difference is that these experiments are not carried out with the domain experts, but with laypersons. This makes experiments cheaper (especially if the domain experts are radiologists) and it is easier to find more testers. An example would be to show a user different explanations and the user would choose the best one.

Function level evaluation (proxy task) does not require humans. This works best when the class of model used has already been evaluated by someone else in a human level evaluation. For example, it might be known that the end users understand decision trees. In this case, a proxy for explanation quality may be the depth of the tree. Shorter trees would get a better explainability score. 'It would make sense to add the constraint that the predictive performance of the tree remains good and does not decrease too much compared to a larger tree' [5].

These are ways in which explanations can be evaluated with respect to how effective they are at convincing users. Whilst it is important that stakeholders are satisfied with explanations, these explanations also need to be an accurate depiction of how the system works.

Are explanations appropriate and satisfying to stakeholders?

'Explainable ML methods provide explanations that are not faithful to what the original model computes... This leads to the danger that the explanation method can be an inaccurate representation of the original model in parts of the feature space' [7]. Especially in safety-related systems, it is important that explanations of how a system work are not only convincing and satisfying but also reliably a faithful account of how the model is actually working. [6] presents a technical method for evaluating the faithfulness of a certain kind of local explanation technique. They present simulated user experiments to evaluate the utility of explanations in trust-related tasks. In particular, they address the following questions:

1. Are the explanations faithful to the model
2. Can the explanations aid users in ascertaining trust in predictions
3. Are the explanations useful for evaluating the model as a whole

They also do some evaluation with human subjects to ask:

1. Can users choose which of two classifiers generalises better
2. Based on the explanations, can users perform feature engineering to improve the model
3. Are users able to identify and describe classifier irregularities by looking at explanations.

These types of evaluation help users to understand how a model is genuinely working, even so far as the explanations can help users to gain enough insight to improve the model. [3] evaluates "fidelity" (faithfulness to the model) of explanations vs interpretability (how easy it is to understand) finding there are trade-offs between the two.

Model trade-offs

As previously mentioned, different models will be compatible with different types of explanations. Hence, during model selection the desired types of interpretability will have to be weighted and prioritised. Some models may allow for multiple types of interpretability. It is also commonly believed that trade-offs between interpretability and accuracy of the model will have to be taken into account ([1] and [8]), although [7] argues that, "It is a myth that there is necessarily a trade-off between accuracy and interpretability". There are also trade-offs between how easy explanations are to understand and how faithful they are to the model [3]. Hence, the following trade-offs will have to be considered:

- Between different types of interpretability
- Between interpretability and accuracy
- Between the fidelity of explanations and their comprehensibility

References

- [1] Amina Adadi and Mohammed Berrada. “Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)”. In: IEEE (2018).
- [2] Finale Doshi-Velez and Been Kim. “Towards A Rigorous Science of Interpretable Machine Learning”. In:arXiv:1702.08608v2 [stat.ML] 2Mar 2017(2017).
- [3] Himabindu Lakkaraju et al. “Interpretable Explorable Approximations of Black Box Models”. In: arXiv:1707.01154v1 [cs.AI] 4 Jul 2017.2017.
- [4] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. “A Survey of Evaluation Methods and Measures for Interpretable Machine Learning”. In: arXiv:1811.11839v2 [cs.HC] 4 Dec 2018(2018).
- [5] Christoph Molnar. “Interpretable Machine Learning A Guide for Making Black Box Models Explainable”. In: (2019).
- [6] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “Why Should I Trust You?” Explaining the Predictions of Any Classifier”. In: https://arxiv.org/abs/1602.04938(2016).
- [7] Cynthia Rudin. “Please Stop Explaining Black Box Models for High-Stakes Decisions”. In: arXiv:1811.10154v2 (2018).
- [8] S. Sarkar. “Accuracy and interpretability trade-offs in machine learning applied to safer gambling”. In: CEUR Workshop Proceedings.