User-Based Experiment Guidelines for Measuring Interpretability in Machine Learning

Bibal, Adrien; Dumas, Bruno; Frenay, Benoît

Published in:
EGC Workshop on Advances in Interpretable Machine Learning and Artificial Intelligence

Publication date:
2019

Document Version
Publisher's PDF, also known as Version of record

User-Based Experiment Guidelines for Measuring Interpretability in Machine Learning

Adrien Bibal, Bruno Dumas, Benoît Frénay

PReCISE - Faculty of Computer Science - NaDI - University of Namur
Rue Grandgagnage 21, 5000 Namur, Belgium
{adrien.bibal, bruno.dumas, benoit.frenay}@unamur.be

Abstract. With the advent of high-performance black-box models, interpretability is becoming a hot topic today in machine learning. While a lot of research is done on interpretability, machine learning researchers do not have precise guidelines for setting up user-based experiments. This paper provides well-established guidelines from the human-computer interaction community.

1 Introduction

Interpretability is a major concern nowadays in machine learning (Bibal and Frénay, 2016; Lipton, 2016). In several applications, such as credit scoring (Martens et al., 2011), machine learning models need to be interpretable in order to be accepted and used. However, despite being a natural way of evaluating interpretability (Doshi-Velez and Kim, 2017), user-based experiments are not widespread in the machine learning literature (Bibal and Frénay, 2016). This may be due to a lack of time or other resources, but also to a lack of guidelines on how to set up such experiments. Inspired by the human-computer interaction (HCI) literature, this paper provides guidelines on what to consider in order to set up user-based experiments.

2 User-Based Experiments on Interpretability in ML

As interpretability is about user comprehensibility of models, it may seem natural that machine learning experiments assessing interpretability involve users. Doshi-Velez and Kim (2017) stress the need to answer several questions when evaluating interpretability. One of the most important questions is how we should set up experiments involving users.

Doshi-Velez and Kim (2017) consider three experimental setups for answering this question. The first experimental setup concerns application-grounded metrics, in which the real task is sought to be evaluated. This kind of setup requires gathering users in order to evaluate the real performance of users on a real task. Second, human-grounded metrics consider experiments in which real task metrics are replaced by simplified tasks for measuring interpretability. For instance, asking users to compare two models may not be the real task, but the comparison makes it possible to get insights on interpretability. Finally, functionally-grounded metrics involve heuristics used to measure interpretability without the need to gather users. These are not user-based experiments, but may be considered when gathering users is too complex or if the resources needed for user-based experiments are not available for the researcher.

Several simplified tasks for the human-grounded metrics are listed by Piltaver et al. (2014a): "classify", "explain", "validate", "discover", "rate" and "compare". For instance, the model interpretability can be measured by asking users to manually classify an instance using the model. This "classify" metric provides an accuracy error representing the agreement between the classification manually made by the user and the one automatically made by the machine using the same model. Another example is "compare", for which two or more models are proposed to users, who are asked to choose the more interpretable among them. The authors evaluated the interpretability of decision trees based on their tasks in (Piltaver et al., 2014b).

Most user-based experiments on interpretability in the machine learning literature can be characterized given the Piltaver's categorization. Allahyari and Lavesson (2011) use a "compare" task for measuring the interpretability of decision trees and rules obtained by various algorithms. Huysmans et al. (2011) use a "classify" task by measuring accuracy, answer time and confidence of users. Other examples can be found in (Poursabzi-Sangdeh et al., 2018).

Despite these works on the classification of user-based experiments and user-based experimental tasks, no precise guidelines are provided to the machine learning researchers for setting up user-based experiments. The following section builds on guidelines established in the human-computer interaction (HCI) community in order to set up such kind of experiments.

3 Guidelines on User-Based Experiments

The guidelines proposed in this paper can be decomposed into three questions: "what do you want to measure" (Section 3.1), "who are your users" (Section 3.2) and "which type of metric can you use" (Section 3.3). Answering these questions may allow machine learning practitioners to better frame how to conduct a user-based experiment.

3.1 What do you Want to Measure?

As outlined in Doshi-Velez and Kim (2017)'s conclusion, it is important to note that "the claim of the research should match the type of the evaluation." This means that the research questions must be clearly stated before establishing the evaluation type.

On the one hand, one may want to get qualitative insights on the overall interpretability of a particular model. In this case, Nielsen and Landauer (1993) demonstrated that even just 5 users can identify 85% of usability problems, including most of the severe problems. The usual approach involves observing and taking notes of how the 5 users manipulate the model during the experiment. This can reveal a large part of the possible answers to questions such as "is the depth of my decision tree important regarding the interpretability", "does the balance of the tree play a role at all", etc.

On the other hand, if something specific, related to interpretability, is to be assessed, then a more specific experiment needs to be set up. First, the research questions must be clearly stated to allow the identification of the real task. Identifying the real task is important for designing an experiment that focuses on this real task (Doshi-Velez and Kim (2017)'s application-grounded metrics) or on the right simplified tasks (Doshi-Velez and Kim (2017)'s human-grounded metrics). Then, as a precise research question needs to be answered, as many users as needed for statistical significance have to be gathered. Finally, after the experiment is over, statistical tools can be used to analyze the results.

3.2 Who are your Users?

Echoing the "what do you want to measure" question, the question "who are your users" needs to be answered. Indeed, the real task is never realized in a vacuum, and users performing the task, in a real setting, have a particular profile. The goal of this question is to identify the user profile related to the task at hand. This identification is mandatory as the pool of users considered for the experiment should match as much as possible the work domain expert profile. This is a point considered by Doshi-Velez and Kim (2017) when they mention the nature of user expertise. Crowdsourcing platforms, such as Amazon Mechanical Turk [1] or CrowdCrafting [2], are valuable resources to gather users as long as they match the target profile.

In practice, users with the targeted profile may be hard to gather, especially when the required expertise is high and/or rare. This explains why students are often used in user-based experiments. For instance, in the examples considered in Section 2, Piltaver et al. (2014b), Allahyari and Lavesson (2011), and Huysmans et al. (2011) all enrolled students in their experiments. It has been shown that in certain cases, considering students in the evaluation, more than a choice by default, is in fact a good choice (Carver et al., 2010), as long as threats to validity are carefully addressed. One reason is the homogeneity of the student pool, limiting the difference between each profile and focusing the experiment on variables that are specific to the task. It also makes it easier to control the expertise background, as the same courses on the domain expertise have been taught to the student pool.

3.3 Which Type of Metric can you use?

The last question is about the different ways interpretability can be measured. Three non-exclusive possibilities can be mentioned: measuring users' errors, time and users' opinions.

First, the errors made by users can be measured. The error assessment can take several forms, such as the tasks identified in Piltaver et al. (2014a)'s design. For instance, the classify task can be used to assess if users can accurately use the model for prediction.

The second possibility is to consider the time taken by users to answer specific questions or the number of tasks performed in a given time. As an example, the time taken by users to classify a set of instances using two different models can be used to compare the interpretability of these two models (for a more extended discussion on the use of Piltaver's tasks for error and time measurement, see Piltaver et al. (2014a)). The duration can also be useful when an error measure is hard to define. For instance, for measuring the interpretability of an unsupervised model, it is not always possible to know what is a correct user answer. Instead, measuring the time needed for the user to grasp a clustering model may be more appropriate.

The third possibility is to consider users' opinions. This option can be combined with the others, and often takes the form of an experimental survey. After having measured the errors or the time taken by the users, questions can be asked about the interpretability of the model.

4 Conclusion

Based on the human-computer interaction (HCI) literature and by referring to the work of Doshi-Velez and Kim (2017) and of Piltaver et al. (2014a), this paper presents guidelines that

1. www.mturk.com
2. www.crowdcrafting.org

can be used by machine learning researchers interested in setting up user-based experiments to measure interpretability. These guidelines correspond to the minimal set of questions typically addressed in the HCI community. The three questions of this minimal set are: "what do you want to measure", "who are your users", and "which type of metric can you use." Through these questions, researchers can align themselves with the experimental settings that are standard in user-centric communities. Future works include finding how to choose between Piltaver's tasks regarding the questions presented in this paper.

References

Allahyari, H. and N. Lavesson (2011). User-oriented assessment of classification model understandability. In *Proc. SCAI*, pp. 11–19.

Bibal, A. and B. Frénay (2016). Interpretability of machine learning models and representations: an introduction. In *Proc. ESANN*, pp. 77–82.

Carver, J. C., L. Jaccheri, S. Morasca, and F. Shull (2010). A checklist for integrating student empirical studies with research and teaching goals. *ESEJ 15*(1), 35–59.

Doshi-Velez, F. and B. Kim (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Huysmans, J., K. Dejaeger, C. Mues, J. Vanthienen, and B. Baesens (2011). An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. *Decision Support Systems 51*(1), 141–154.

Lipton, Z. C. (2016). The mythos of model interpretability. In *Proc. ICML Workshop on Human Interpretability in Machine Learning*.

Martens, D., J. Vanthienen, W. Verbeke, and B. Baesens (2011). Performance of classification models from a user perspective. *Decision Support Systems 51*(4), 782–793.

Nielsen, J. and T. K. Landauer (1993). A mathematical model of the finding of usability problems. In *Proc. INTERACT and CHI*, pp. 206–213.

Piltaver, R., M. Luštrek, M. Gams, and S. Martinčić-Ipšić (2014a). Comprehensibility of classification trees - survey design. In *Proc. IS*, pp. 70–73.

Piltaver, R., M. Luštrek, M. Gams, and S. Martinčić-Ipšić (2014b). Comprehensibility of classification trees - survey design validation. In *Proc. ITIS*, pp. 5–7.

Poursabzi-Sangdeh, F., D. G. Goldstein, J. M. Hofman, J. W. Vaughan, and H. Wallach (2018). Manipulating and measuring model interpretability. In *Proc. NIPS WiML Workshop*.

Résumé

Avec l'avancée des modèles "boîtes noires" hautement performants, l'interprétabilité est devenu un sujet de recherche majeur aujourd'hui. Alors que de plus en plus de recherches en apprentissage automatique portent sur l'interprétabilité, les chercheurs en apprentissage automatique n'ont pas de directives précises pour mettre en place des expériences utilisateurs. Cet article fournit une suite de directives à suivre, provenant de la communauté de l'interaction homme-machine, afin de mettre en place ce type d'expériences.