User-Based Experiment Guidelines for Measuring Interpretability in Machine Learning

Adrien Bibal, Bruno Dumas and Benoît Frénay

PReCISE Research Center, Faculty of Computer Science, University of Namur

January 22, 2019

Overview

Interpretability

Interpretability

What is Interpretability?

- Ill-defined concept
- Basically: the level of model understandability
- Many questions around interpretability, such as:
  - How to evaluate the interpretability of models of different types?
  - How to deal with semantics?

Interpretability

Lots of success, lately

- Annual workshops at NeurIPS and ICML
- Other punctual workshops (e.g., EGC and ESANN)
- Often boosted by deep neural networks

Model-oriented

- Often concerned with developing interpretable models
- Rare focus on interpretability evaluation

Evaluation of Interpretability: in the Literature

Evaluation of Interpretability: in the Literature

Different types of evaluation

- Application-grounded metrics: real task
- Human-grounded metrics: simplified task (e.g. comparison)
- Functionally-grounded metrics: heuristics (e.g. complexity)

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

Evaluation of Interpretability: in the Literature

Different types of simplified tasks

- Classify
- Explain
- Validate
- Discover
- Rate
- Compare

Piltaver, R., M. Luštrek, M. Gams, and S. Martinčić-Ipšić (2014). Comprehensibility of classification trees - survey design. In Proceedings of the International multiconference Information Society, pp. 70–73.

Guidelines (from HCI) on Questions to Answer

Guidelines (from HCI) on Questions to Answer

What do you want to measure?

- Getting qualitative insights on model interpretability
  $\rightarrow$ 5 users for 85% of the usability

  Nielsen, J., & Landauer, T. K. (1993). A mathematical model of the finding of usability problems. In Proceedings of the INTERACT and CHI conference on Human factors in computing systems (pp. 206-213).

  $\rightarrow$ Observing user manipulation
- Evaluating something specific related to interpretability
  $\rightarrow$ Experiment must be designed according to the real task
  - Focus directly on the real task (Application-grounded metrics)
  - Find an adapted simplified task (Human-grounded metrics)

  $\rightarrow$ As many users as necessary for statistical significance

Guidelines (from HCI) on Questions to Answer

Who are your users?

- Identify the real user profile related to the real task
  $\rightarrow$ Should match as much as possible the work domain expert profile
- In practice, users with the exact profile are hard to gather
  $\rightarrow$ Find the closest profile
  $\rightarrow$ But students can be OK too... Because e.g.:
  - Homogeneity of the user pool
  - Control of user expertise

Carver, J. C., Jaccheri, L., Morasca, S., & Shull, F. (2010). A checklist for integrating student empirical studies with research and teaching goals. Empirical Software Engineering, 15(1), 35-59.

Guidelines (from HCI) on Questions to Answer

Which type of metric can you use?

- Three typical (and non-exclusive) ways to measure:
  - Measuring the user's errors (e.g. classify)
  - Measuring the time (e.g. time needed to classify)
    $\rightarrow$ Also useful when measuring errors is difficult (e.g. unsupervised learning)
  - Gather the user's opinion
    $\rightarrow$ Experimental survey

Conclusion

Conclusion

- As Doshi-Velez & Kim presented:
  Need for a rigorous science of interpretability
- Guidelines from HCI
  - What do you want to measure?
  - Who are your users?
  - Which type of metric can you use?
- Future work: link between real task and Piltaver's tasks