PhD proposal
May 2022

Defining Differential Explanations:
Understanding the Dynamic of Changes in Machine Learning Models

AXA-Sorbonne University Trustworthy and Responsible AI Lab (TRAIL)

Supervision: Marie-Jeanne Lesot (Associate Professor, Sorbonne University), Christophe Marsala (Professor, Sorbonne University), Xavier Renard (Research Scientist Lead, AXA)

Keywords: Machine Learning, Explainability, Interpretability, XAI, Differential Explainability

More information & Application: xavier.renard@axa.com - trail.lip6.fr

1 Context

Organizations are adopting Machine Learning (ML) models to automate many of their tasks at an accelerating pace. This adoption brings important challenges relative to a responsible and trustworthy use of ML techniques. How to assess a ML model beyond the typical aggregated performance metrics, which depict a very incomplete picture of the model behaviour? How to inspect a ML model to look for potential flaws, such as fairness or robustness issues? How to support users with these models and avoid the black-box effect, to improve the decision making and increase trust in those systems? What are the consequences of the task automation on its environment? Explaining ML models and their predictions is instrumental to answer those questions and support model deployment.

The research dedicated to explaining ML models is usually called AI/ML interpretability, explainability or explainable, interpretable AI/ML and XAI. It is a cross-disciplinary field, across computer science, mathematics, human-computer interface and humanities (such as disciplines that study human understanding through explanations or, on a different note, ethics). A broad variety of explanations can be expected on the behaviour of a model, which influences the information that has to be extracted from it [8, 14]. But a large part of the literature in ML explainability is focused on explaining in a human-understandable way what patterns one single ML model has learnt from the data: either globally that is over the entire feature space [2], or locally usually at the scale of a prediction [9, 11, 7] (for surveys, see for instance [5, 1]).

2 Scientific Objectives

Instead of explaining a static situation (a trained model on a fixed dataset), this PhD aims to explore the novel concept of *differential explainability*. We define the concept of *differential explainability* as explaining the differences or the evolution between successive versions of a ML model devoted to the same task (*e.g.* classification with fraud detection, regression with GDP prediction).

When are successive versions of a model generated? Successive versions are needed to improve the model performances on its task in the same context or to preserve the performances with the adaptation of the model to an evolving context. In a first scenario, the context of the task remains the same and the dataset used for training the model, and its underlying distribution, is still relevant. Given the task under-specification [3] of most ML problems (*i.e.* sparse datasets), leading to issues like discrepancies in models [10], a model is likely to benefit from a re-training, especially when it is associated with a better model optimization or model selection, with the design of an improved ML pipeline, with the inclusion of novel modelling techniques or with the augmentation of the dataset. In a second scenario, more critical, the context is changing. These

steps are then required to adapt the model to the new context and counteract phenomenons such as for instance the concept drift [4], when the data distribution changes over time.

In both cases, a second improved model version is generated after an initial version. Different versions can follow each other this way. Besides an improved model optimization or an improved ML pipeline, the evolution of the dataset is a central factor that leads to new model versions. New data instances can be sampled, to complement the original dataset in order to improve the description of the task and reduce the dataset sparsity or to describe emerging patterns. These new instances can be sampled with a simple collection of new data points or by leveraging dedicated techniques such as active learning [12], online learning [6] or reinforcement learning [13]. The feature space itself can evolve for a better description of the task with additional sources of information (*e.g.* additional variables) or a degradation of the description with alteration or deletion of the information (*e.g.* if a data source is no longer available and a variable is removed or if a measurement performed differently). On a different note, transfer learning generates successive models for improved performances by leveraging the knowledge learnt by a first model on a task to create a second model specific to a different but related task [15].

In summary, in a model lifecycle, several model versions can be generated to preserve or improve the performances on the task at hand, in a stationary or evolving context. **The successive versions of the model are likely to have learnt different patterns: we seek to explain those differences and the corresponding evolution.** What are the new or different patterns learnt by a new model version in comparison with the previous version? Why is a prediction different from one model version to another? A model developer, a domain expert, a user or an auditor will benefit from explanations on the differences between model versions and the corresponding evolution. They would be able to use these explanations to assess the impact of changes in the modelling (*e.g.* changes in the dataset or the configuration of the ML pipeline); to debug a model; to assess or anticipate flaws; to accelerate the adoption of successive model versions by, for instance, only assessing the diverging parts of successive models especially when new versions of a model are frequently released; to explain users why his/her predictions are changing over time; to gain new knowledge if the latest model is more accurate or if the modeled phenomenon is evolving.

These questions, widely unexplored in the literature need to be precised, properly formalized and addressed. These questions also need to be confronted to the existing literature beyond the machine learning literature (*e.g.* statistical modelling).

3 PhD Schedule

The PhD is expected to last 3 years with the following *indicative* schedule:

- In a first phase, the PhD Candidate will review the scientific literature relevant to the topic of differential interpretability. It will obviously encompass the ML interpretability and explainability fields, as well as any other relevant ML domains with a particular focus on topics that cover model retraining, model evolution or context evolution such as the concept drift. Other scientific disciplines could be leveraged, when they deal with a multiplicity of model versions for a task, especially when this multiplicity must be assessed or explained (*e.g.* statistical modeling). During this phase, the PhD Candidate will also review and evaluate situations and use cases where differential interpretability would matter for users and practitioners, especially for the insurance domain and for economic analysis.

- In a second phase, the PhD Candidate will identify and focus on a subset of problems and challenges for differential interpretability. These problems will be formalized and propositions will be made to address them. During this phase, the propositions will be tested on some of the previously identified use cases.

- The last phase of the PhD will be dedicated to the PhD thesis.

4 Working Environment

This PhD is hosted by the joint research lab TRAIL (Trustworthy and Responsible AI Lab) between AXA and Sorbonne University, in Paris. As such, the PhD Candidate will be hired by Sorbonne University and supervised by TRAIL members from Sorbonne University, the LIP6 and the AI research team from AXA in Paris. The PhD Candidate will benefit from interactions with other researchers from the team in Responsible AI and ML Explainability in particular, with focuses on both the ML and Human-Computer Interface aspects.

References

[1] Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. "Machine learning interpretability: A survey on methods and metrics". In: *Electronics* 8.8 (2019), p. 832.

[2] Mark Craven and Jude Shavlik. "Extracting tree-structured representations of trained networks". In: *Advances in neural information processing systems* 8 (1995).

[3] Alexander D'Amour et al. *Underspecification Presents Challenges for Credibility in Modern Machine Learning*. 2020. DOI: 10.48550/ARXIV.2011.03395. URL: https://arxiv.org/abs/2011.03395.

[4] João Gama et al. "A survey on concept drift adaptation". In: *ACM computing surveys (CSUR)* 46.4 (2014), pp. 1–37.

[5] Riccardo Guidotti et al. "A survey of methods for explaining black box models". In: *ACM Computing Surveys (CSUR)* 51.5 (2018), pp. 1–42.

[6] Steven CH Hoi et al. "Online learning: A comprehensive survey". In: *Neurocomputing* 459 (2021), pp. 249–289.

[7] Thibault Laugel et al. "Comparison-based inverse classification for interpretability in machine learning". In: *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*. Springer. 2018, pp. 100–111.

[8] Q. Vera Liao, Daniel Gruen, and Sarah Miller. "Questioning the AI: Informing Design Practices for Explainable AI User Experiences". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, 2020, pp. 1–15. ISBN: 9781450367080. URL: https://doi.org/10.1145/3313831.3376590.

[9] Scott M Lundberg and Su-In Lee. "A unified approach to interpreting model predictions". In: *Advances in neural information processing systems* 30 (2017).

[10] Xavier Renard, Thibault Laugel, and Marcin Detyniecki. *Understanding Prediction Discrepancies in Machine Learning Classifiers*. 2021. DOI: 10.48550/ARXIV.2104.05467. URL: https://arxiv.org/abs/2104.05467.

[11] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier". In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, pp. 1135–1144.

[12] Burr Settles. "Active learning literature survey". In: (2009).

[13] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[14] Tom Vermeire et al. "How to Choose an Explainability Method? Towards a Methodical Implementation of XAI in Practice". In: *Machine Learning and Principles and Practice of Knowledge Discovery in Databases*. Cham: Springer International Publishing, 2021, pp. 521–533. ISBN: 978-3-030-93736-2.

[15] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. “A survey of transfer learning”. In: *Journal of Big data* 3.1 (2016), pp. 1–40.