Proceedings of the 1st International Workshop on Explainable and Interpretable Machine Learning (XI-ML)

(http://www.cslab.cc/xi-ml-2020/)

- Preface -

Recently, scientific discourse in artificial intelligence and data science has focused on explainable AI (XAI) with respect to algorithmic transparency, interpretability, accountability and finally explainability of algorithmic models and decisions. In machine learning, approaches can be classified as white-box and black-box. White-box approaches, such as rule learners and inductive programming, result in explicit models which are inherently interpretable (Rudin, 2019). On the other hand, black-box approaches, such as (deep) neural networks, result in opaque models. For this second type of models, over the last years, different approaches for ex-post explanation generation have been proposed.

In this workshop, we want to bring together research from interpretable and explanatory machine learning. Interpretable ML can profit from recently proposed explanation generation techniques to make complex learned models more comprehensible, especially to end-users (Muggleton, Schmid et al., 2018; Fürnkranz, Kliegr, Paulheim, 2020; Lonjarret et al., 2020). In particular, interpretable learning can be integrated into the construction of complex models, e.g., for guiding their construction (Atzmueller et al., 2017), as well as to refine the respective model (Weidner, Atzmueller, Seipel, 2019). Furthermore, it can provide rich rule-based techniques to generate interpretable surrogate models for black-box learners (Schmid, 2018). Such surrogate models can be global models generated by rule-extraction mechanisms (Hailesilassie, 2016) or local models which allow richer local explanations than simple linear rules as, for instance, proposed by LIME (Rabold et al., 2019). Also, a frontier direction is investigating psychological phenomena that can affect the understanding of machine learning models, such as cognitive biases and conversational maxims (Kliegr, Bahnik, Fürnkranz, 2018). This interdisciplinary inspiration, such as debiasing techniques long studied by psychologists, will hopefully contribute to a better comprehensibility of the results of models created by the next generation of machine learning algorithms.

XI-ML (Explainable and Interpretable Machine Learning) aims at bringing together research from interpretable and explainable machine learning. Hopefully, integrating both areas, allows new perspectives on questions on appropriate learning formalisms, interpretation and explanation techniques, their metrics, as well as the respective assessment options arise.

The first edition of the XI-ML (Explainable and Interpretable Machine Learning) workshop was held on September 21, 2020 at the 43rd German Conference on Artificial Intelligence, Bamberg, Germany. The workshop was devoted to the discussion of the topics mentioned above. It aimed to provide an interdisciplinary forum to investigate fundamental issues in explainable and interpretable machine learning as well as to discuss recent advances, trends, and challenges in these areas. From 8 submissions (6 full and two short papers), 5 full papers and the two short papers were accepted for presentation in a comprehensive review process.

The remaining part of the volume presents revised versions of papers that were discussed during the workshop. Boström et al. discuss in their paper about how to explain multivariate time series forecasting, in an application to predicting the Swedish GDP. Next, Mucha et al. present a position paper on how to construct participatory design spaces for the context of explainable AI interfaces in expert domains. After that, Volkert discussed how the application of the TED (Teaching Explanations for Decisions) explainable AI framework and the impact of class (im-)balance. Fleiss, Bäck and Thalmann present a short paper on empirical results in the context of recruiting - about explainability and the intention to use AI-based conversational agents. Potyka addresses foundational issues towards solving classification problems with quantitative abstract argumentation. Mollenhauer and Atzmueller present an approach for sequential exceptional pattern discovery using pattern-growth (SEPP) - as the basis of an extensible framework for interpretable machine learning on sequential data. Sun, Chakraborti and Noble discuss results of a comparative study of explainer modules in the context of automated skin lesion classification.
Finally, Marcin P. Joachimiak (Environmental Genomics and Systems Biology Division, Lawrence Berkeley Laboratory) kindly agreed to present a keynote entitled “How to teach a computer to learn about microbes with KG-COVID-19”. This talk introduced a new resource that amalgamates SARS-CoV-2 related biological knowledge from multiple specialized knowledge graphs and ontologies. With over 10 million nodes, it is one of the largest (if not the largest) resources of this kind. In his talk, Dr. Joachimiak demonstrated the utility of this resource for machine learning, emphasizing the need for explainable techniques.

References

Atzmueller, M., Hayat, N., Schmidt, A., & Klöpper, B. (2017). Explanation-aware feature selection using symbolic time series abstraction: approaches and experiences in a petro-chemical production context. In *IEEE International Conference on Industrial Informatics (INDIN)* (pp. 799-804). IEEE, Boston, MA, USA

Fürnkranz, J., Kliegr, T., & Paulheim, H. (2020). On cognitive preferences and the plausibility of rule-based models. *Machine Learning*, *109*(4), 853-898.

Hailesilassie, T. (2016). Rule extraction algorithm for deep neural networks: A review. *arXiv preprint arXiv:1610.05267*.

Kliegr, Tomáš, Štěpán Bahník, and Johannes Fürnkranz. "A review of possible effects of cognitive biases on interpretation of rule-based machine learning models." arXiv preprint arXiv:1804.02969 (2018).

Lonjarret, C., Robardet, C., Plantevit, M., Auburtin, R., & Atzmueller, M. (2020). Why Should I Trust This Item? Explaining the Recommendations of any Model. In *IEEE International Conference on Data Science and Analytics*. IEEE, Boston, MA, USA

Muggleton, S. H., Schmid, U., Zeller, C., Tamaddoni-Nezhad, A., & Besold, T. (2018). Ultra-Strong Machine Learning: comprehensibility of programs learned with ILP. *Machine Learning*, *107*(7), 1119-1140.

Rabold, J., Deininger, H., Siebers, M., & Schmid, U. (2019). Enriching visual with verbal explanations for relational concepts–combining LIME with Aleph. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 180-192). Springer, Cham.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206-215.

Schmid, U. (2018). Inductive Programming as Approach to Comprehensible Machine Learning. In *DKB/KIK@ KI* (pp. 4-12).

Weidner, D., Atzmueller, M., & Seipel, D. (2019). Finding Maximal Non-redundant Association Rules in Tennis Data. In *Declarative Programming and Knowledge Management* (pp. 59-78). Springer, Cham.

Editors

- Martin Atzmueller, Osnabrück University, Germany
- Tomáš Kliegr, University of Economics Prague, Czech Republic
- Ute Schmid, University of Bamberg, Germany

Program Committee of XI-ML 2020

- Klaus-Dieter Althoff, University of Hildesheim
- Maria Bielikova, Kempelen Institute of Intelligent Technologies, Slovakia
- Henrik Boström, KTH Royal Institute of Technology, Sweden
- Amit Dhurandhar, IBM TJ Watson Research Center, USA
- Johannes Fürnkranz, Johannes Kepler University, Linz
- Martin Holena, Czech Academy of Sciences
- Eyke Hüllermeier, University of Paderborn
- Kristian Kersting, TU Darmstadt, Germany
- Grzegorz Nalepa, Jagellonian University, Poland
- Mykola Pechenizkyi, TU Eindhoven
- Marc Plantevit, University Lyon
- Eric Postma, Tilburg University
- Celine Rouveirol, Université Sorbonne Paris Nord
- Stefano Teso, KU Leuven, Belgium