Physikalisch-Technische Bundesanstalt
Braunschweig und Berlin
Nationales Metrologieinstitut

Advancing the theory and practice of machine learning model explanations in biomedicine

In healthcare, as in other safety-critical domains, there is a strong desire – driven by clinical and scientific but also ethical and legal considerations – to understand how a given machine learning model arrives at a certain prediction for a particular data instance (e.g., a new patient), which motivates the field of explainable or interpretable artificial intelligence (xAI, e.g., Montavon et al., 2018). This problem is inherently unsupervised, which means that the ground-truth cannot, even retrospectively, be obtained in practice. Validation of unsupervised methods is a prerequisite for applying such methods in clinical contexts. This principle must also hold for xAI methods. However, due to the lack of ground-truth information in real data, the vast literature on xAI resorts to subjective qualitative assessments or surrogate metrics, such as relative prediction accuracy, to demonstrate the "plausibility" of the provided explanations. Moreover, there is no universally accepted definition of the "importance" of a feature that could be utilized to construct synthetic ground-truth data. Features are often considered important if their omission leads to a degradation of prediction performance. However, it has been pointed out that this definition is flawed, as it applies to noise features lacking any statistical relation to the prediction target (Haufe et al., 2014). While we have provided a simple remedy for linear learning problems, this problem is expected to be aggravated in use cases requiring non-linear prediction models, such as the classification or segmentation of radiological images using convolutive neural networks, for which no comparable remedy exists yet. Novel, theoretically founded, definitions of explainability along with appropriately designed synthetic ground-truth data are needed in order to benchmark existing xAI approaches as well as to drive the development of improved methods.

This project aims to advance both the theoretical foundation of xAI and the practical, in particular, clinical, utility of explanation methods. As such, it will extend prior and ongoing work in the group of Dr. Haufe at Charité Berlin (novel appointment as professor for machine learning and inverse problems at TU Berlin ongoing), and benefit from close interactions with domain experts at TU Berlin and the BIFOLD research center. We will develop novel, useful, definitions of feature importance that can be leveraged to generate synthetic ground-truth data. These data will be used to quantitatively assess the "explanation performance" of existing xAI methods such as layerwise relevance propagation (Bach et al., 2015), local surrogates (Ribeiro et al., 2016), PatternNet (Kindermans et al., 2017), SHAP scores (Lundberg and Lee, 2017). To this end, novel performance metrics will be developed.

We will create a benchmark suite of non-linear prediction problems where the set of important features is known a-priori. These problems will range from simple toy examples involving few variables to realistic settings mimicking clinical use cases such as image classification or segmentation tasks. Based on the results of our objective quantitative assessments, improved explanation methods should be developed for particular tasks as well as classes of machine learning approaches. The developed reference data and tools should be disseminated and further developed (e.g. by organizing symposia, public data analysis challenges) in a community-driven effort.

Principal Investigator

- Stefan Haufe (PTB 8.44, TU Berlin and Charité)

PhD student

- Benedict Clark (University supervisor: Prof. Stefan Haufe, TU Berlin)

References

Bach S, Binder A, Montavon G, Klauschen F, Müller KR, Samek W. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. PLoS One. 2015 Jul 10;10(7):e0130140.

Haufe S, Meinecke F, Görgen K, Dähne S, Haynes JD, Blankertz B, Bießmann F. On the interpretation of weight vectors of linear models in multivariate neuroimaging. Neuroimage. 2014 Feb 15;87:96-110.

Kindermans PJ, Schütt KT, Alber M, Müller KR, Erhan D, Kim B, Dähne S. (2017). Learning how to explain neural networks: PatternNet and PatternAttribution. International Conference on Learning Representations. 2018; arXiv preprint arXiv:1705.05598.

Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In Advances in neural information processing systems. 2017.

Montavon G, Samek W, Müller KR. Methods for interpreting and understanding deep neural networks. Digital Signal Processing. 2018; 73:1-15.

Ribeiro MT, Singh S, Guestrin C. "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.