Stage Explainable AI
JM. Loubes (IMT) & E. Pauwels (IRIT)

Setting : Machine learning algorithms build predictive models which are nowadays used for a large variety of tasks. They have become extremely popular in various applications such as finance, insurance risk, health-care, recommendation systems as well as industrial applications of all kinds including predictive maintenance, defect detection or industrial liability. Such algorithms are designed to assist human experts by giving access to valuable predictions and even tend to replace human decisions in many fields, achieving an extremely good performance. Over the last decades, the complexity of such algorithms has grown, going from simple and interpretable prediction models based on regression rules to very complex models such as random forest, gradient boosting and models using deep neural networks. Such models are designed to maximize the accuracy of their predictions at the expense of the interpretability of the decision rule. Little is also known about how the information is processed in order to obtain a prediction, which explains why such models are widely considered as black-box models.

Different methods have been proposed to make understandable the reasons leading to a prediction, each author using a different notion of explainability and interpretability of a decision rule. We mention early works by [12] for recommender systems, [5] for neural networks [9], or [16] for generalized additive models. Recently a special attention has also been given to deep neural systems. We refer for instance to [17], [20] and [19]. Understanding a black box requires being able to quantify the particular influence of the variables in the decision rule but also in the learning process. This point of view is close to sensivity analysis of computer code experiments. In this context, sensitivity analysis allows to rank the relative importance of the input variables involved in an abstract input-output relationship modeling the computer code under study.

Goal of internship: We propose to leverage sensitivity analysis and explainability techniques to propose methodologies to visualize and understand how modifications of the learning distribution impacts the decision rule. After a bibliographic study on explainability and entropic projections, we will develop a framework to implement deformations of input learning distributions by entropic projection [1]. This provides a natural tool to control stepwise influence of different types of deviations of the inputs from the original learning data. In a second step, we will investigate the evolution of the decision rule under specific data perturbations using continuity arguments. We will start with a review of related existing notions such as leverage [21].

Funding: This internship will be funded by IRT Saint Exupéry and co-supervised with University Toulouse 3, Paul Sabatier. Depending on the output of the internship, the succesful candidate could be proposed to continue with a Ph.D. (funding is available). The internship will be hosted by the DEEL team (DEpendable and Explainable Learning), a joint project team between Toulouse and Montréal, in building B612, IRT Saint Exupéry.

1 Bachoc et al (2018) Entropic Variable Boosting for Explainability & Interpretability in Machine Learning.

5 Mark Craven and Jude W. Shavlik. Extracting tree-structured representations of trained net- works. In Advances in Neural Information Pro- cessing Systems 8, NIPS, Denver, CO, Novem- ber 27-30, 1995 , pages 24?30, 1995.

9 Mary T Dzindolet, Scott A Peterson, Regina A Pomranky, Linda G Pierce, and Hall P Beck. The role of trust in automation reliance. International journal of human-computer studies, 58(6):697?718, 2003

12 Jonathan L Herlocker, Joseph A Konstan, and John Riedl. Explaining collaborative filtering recommendations. In Proceedings of the 2000 ACM conference on Computer supported coop- erative work , pages 241?250. ACM, 2000

19 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 Confer- ence on Knowledge Discovery and Data Mining , pages 1135?1144. ACM, 2016

20 Ramprasaath R Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that ? arXiv preprint arXiv:1611.07450, 2016.

21 S. Chatterjee and A.S. Hadi. Influential observations, high leverage points, and outliers in linear regression. Statistical Science, 1(3):379-393, 1986.