Politecnico di Torino

COMPUTER AND CONTROL ENGINEERING

DAUIN - Safe and trustworthy AI

Funded By

Dipartimento DAUIN

Supervisor

BARALIS ELENA MARIA - elena.baralis@polito.it

Contact

PASTOR ELIANA - eliana.pastor@polito.it
CERQUITELLI TANIA - tania.cerquitelli@polito.it
BARALIS ELENA MARIA - elena.baralis@polito.it

Context of the research activity

The wide adoption of machine learning (ML) models requires both establishing trust in a model decision and providing a ML pipeline robust to errors and attacks. Safety and robustness come into play along the entire pipeline of ML-based systems and in a wide range of application domains (e.g., automotive, healthcare).

The main goal of this research is the study of methods for the deployment of safe ML pipelines in real-life settings. It addresses the many facets of safety, ranging from explaining the different steps of the ML pipeline to detecting concept drift that may raise safety issues.

Ensuring safety of ML based pipelines is fundamental to allow their acceptance in a wide range of critical application domains. Different techniques are usually needed to account for different data types (e.g., images, structured data, time series). All the different steps in ML-based development pipelines should be addressed: requirement definition, data preparation and model selection, training, evaluation and testing, monitoring.

The research activity will consider industrial domains (e.g., critical infrastructures, aerospace, automotive, manufacturing) in which safety plays a key role. The algorithms and methods will target different data types, possibly considered jointly in multimodal applications. The following different facets of trustworthy AI will be addressed.

Model understanding. The research work will address local analysis of individual predictions. These techniques will allow the inspection of the local behavior of different classifiers and the analysis of the knowledge different classifiers are exploiting for their prediction. The final aim is to support human-in-the-loop inspection of the reasons behind model predictions.

Model trust and robustness. Insights into how machine learning models make their decision allow evaluating if the model may be trusted. Methods to evaluate the reliability of different models will be proposed. In case of negative outcomes, techniques to suggest enhancements of the model to

Objectives	cope with wrong behaviors and improve the trustworthiness of the model will be studied. Robustness is the ability of a ML algorithm or pipeline to cope with errors during execution or with erroneous inputs. Criteria to evaluate model robustness and resiliency will be studied. Model debugging and improvement. The evaluation of classification models generally focuses on their overall performance, which is estimated over all the available test data. An interesting research line is the exploration of differences in the model behavior, which may characterize different data subsets, thus allowing the identification of problematic data subsets, which may cause anomalous behaviors in the ML model. YEAR I: state-of-the-art survey for safe AI methods, performance analysis and preliminary proposals of improvements over state-of-the-art algorithms, exploratory analysis of novel, creative solutions for trustworthy AI; assessment of main explanation issues in 1-2 specific industrial case studies. YEAR 2: new algorithm design and development, experimental evaluation on a subset of application domains; deployment of algorithms in selected industrial contexts. YEAR 3: algorithm improvements, both in design and development, experimental evaluation in new application domains. During the second-third year, the candidate will have the opportunity to spend a period of time abroad in a leading research center. List of possible venues for publications IEEE TKDE (Trans. on Knowledge and Data Engineering) ACM TKDD (Trans. on Knowledge Discovery in Data) IEEE TNNLS (Trans. On Neural Networks and Learning Systems) ACM TOIS (Trans. on Information Systems) Information sciences (Elsevier) Expert systems with Applications (Elsevier) Machine Learning with Applications (Elsevier) Engineering Applications of Artificial Intelligence (Elsevier) Journal of Big Data (Springer) IEEE/ACM International Conferences

Skills and competencies for the development of the activity	Strong background in data science and related topics such as machine learning, deep learning, artificial intelligence, and data management.