Politecnico di Torino

COMPUTER AND CONTROL ENGINEERING

Explaining AI (XAI) models for sequential data

Funded By	Dipartimento DAUIN Centro Interdipartimentale SmartData@PoliTO
Supervisor	BARALIS ELENA MARIA - elena.baralis@polito.it
Contact	MELLIA MARCO - marco.mellia@polito.it BARALIS ELENA MARIA - elena.baralis@polito.it
Context of the research activity	Machine learning models are increasingly adopted to assist human experts in decision-making. Especially in critical tasks, understanding the reasons behind model predictions is essential for trusting the model itself. Investigating model behavior can provide actionable insights. For example, experts can detect model wrong behaviors and actively work on model debugging and improvement. Unfortunately, most high-performance ML models lack interpretability. Time series data allow an effective representation of many interesting phenomena, e.g., sensor reading in many different application domains ranging from predictive maintenance to automated factory floor management. Current state-of-the-art techniques for time series prediction models are based on deep learning techniques (e.g., RNN, but also CNN). These techniques provide so called black-box models, i.e., models that do not expose the motivations for their predictions. The main goal of this research activity is the study of methods to allow human-in-the-loop inspection of classifier reasons behind predictions for time series data. Explanations can help data scientists and domain experts to understand and interactively investigate individual decisions made by black-box models.
	Exploring and understanding the motivations behind black-box model predictions is becoming essential in many different applications. Different techniques are usually needed to account for different data types (e.g., images, structured data, time series). The research activity will consider industrial domains (e.g., the construction and spatial domains) in which the availability of understandable explanations is particularly relevant for explaining anomalous behaviors. The explanation algorithms will target both structured data and time series. The following different facets of XAI (Explainable AI) will be addressed. Model understanding. The research work will address local analysis of

Objectives	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. Insights into how machine learning models arrive at 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 cope with wrong behaviors and improve the trustworthiness of the model 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 potential sources of bias in the data.

Skills and competencies for the development of the activity	- Good knowledge of Machine learning and Deep learning algorithms - Good programming skill - (optional) Knowledge of Big data frameworks (Spark, Hadoop)