Explainability methods for machine learning models in static and evolving data

Artificial intelligence and, in particular, machine learning are subfields of computer science that have been developing very intensively in the last years. This is related to both the proposals of new algorithms and access to massive and more complex data, the emergence of specialised software libraries and the easier availability of more powerful computing hardware. This has led to the success of intelligent systems in areas such as, image recognition, natural language processing, speech recognition, robotics, control of autonomous cars, medical diagnostics and mining of various types of data.

In some areas, data has been generated automatically in the form of data streams. Such data sets arrive continuously, are massive and require sufficiently fast processing and response from the intelligent system. Furthermore, due to the non-stationary characteristics of the data source, there are changes in the probability distributions of examples in the stream, i.e. so-called concept drift. As the predictive model has been learned on historical data, it loses its predictive ability when the drift occurs. Therefore, many methods for detecting the occurrence of drift and specialized incremental algorithms for learning models that adapt to it have been introduced so far. However, the models themselves are quite complex, such as ensembles of classifiers, and are difficult to be analysed and understood especially in the context of changes.

It should be noted that modern learning systems are complex models, so called black boxes, that are not transparent and easy to analyze even for specialists. At the same time, in many applications, e.g., in critical systems or in human decision support, the way of obtaining their predictions or recommendations is expected to be explained. In response to these needs, the field of explainable artificial intelligence (XAI) has been developing quite fast. It proposes methods that provide information about the inner workings of a system or a local explanation of the reasons for making a prediction for a given instance. XAI has so far been developed mainly for static data.
However both fields, i.e., machine learning explainability and learning models from data streams, are developing independently of each other and there is a lack of research on adaptation of explainability methods for evolving data streams. This project aims to fill this gap for selected methods. In the proposed research project, we intend to carry out the following research tasks: (1) Examining the properties and safeness of counterfactual explanations; (2) Multi-criteria analysis of a set of explanations; (3) Investigating the inter-relations of data stream changes, models and explanations within a three layer framework; (4) Incremental generalisation of methods for generating counterfactual and prototypical explanations for data streams; (5) Use of changes in explanations of machine learning models to describe concept drift; (6) Investigation of the sequence of changing explanations and their use to improve models in the stream.

The expected results of this project should provide new methodological insights into the application of XAI to concept drifting data streams and be useful for improving the performance of models, correcting possible errors and increasing people's trust in such systems.