Summary

Machine learning models are increasingly adopted in a wide range of critical areas. However, most high-performing models lack interpretability. Especially in critical tasks as health care, criminal justice, and finance, understanding the model behavior is of fundamental importance.

The thesis addresses the problem of the lack of interpretability of classification models. We propose post-hoc techniques to analyze the behavior of classifiers, from the perspective of individual instance predictions and data subgroups. The proposed techniques build on the notion of patterns. Patterns are conjunctions of attribute-value pairs intrinsically interpretable. We leverage patterns to capture relevant associations of attribute values and to define subgroups in the attribute domain. The proposed techniques are model agnostic because they do not rely on the knowledge of the inner workings of any classification paradigm.

At the level of individual instance predictions, we consider the lack of understanding of the reasons behind individual predictions for black-box models. We propose a rule-based explanation method that explains the prediction of any classifier on a specific instance by analyzing the joint effect of feature subsets on the classifier prediction. The approach relies on a local rule-based model to identify the relevant patterns determining locally the prediction. The extracted local patterns provide a qualitative understanding of the reasons behind predictions. We provide a quantitative understanding through the notion of prediction difference. We exploit a removal-based technique to compute the influence on individual predictions of feature values and subsets of feature values derived from patterns. We then propose an interactive tool that leverages the rule-based explanation method for a human-in-the-loop inspection of the reasons behind model predictions.

From the subgroup perspective, we investigate the behavior of models on data subgroups. Specifically, we address the problem of identifying and characterizing data subgroups in which a classification model behaves differently. The identification of these critical data subgroups plays an important role in many applications, for example, model validation and testing, evaluation of model fairness, and identification of bias. We introduce the notion of divergence to capture the different behavior of the model on data subgroups with respect to the overall behavior. We characterize data subgroups via patterns and we leverage frequent pattern mining techniques for their automatic extraction. We use the notion of Shapley value to quantify the contribution to the pattern divergence of the attribute values identifying a data subgroup. We also introduce a generalization of the Shapley value to estimate the global contribution to the divergent model behavior. We then propose an interactive system for the exploration of divergent subgroups which supports drill-down operations and human-in-the-loop inspections of peculiar subgroup behaviors.

The work is supported by theoretical analysis and experimental evaluations, showing the effectiveness of the proposed approaches to reveal the behavior of the model at the individual instance and subgroup level.