Exemplar based approaches for explaining predictions of machine learning models

Objective: Research and develop a new method for explaining the predictions of machine learning models based on *exemplar instances* from the training set, i.e., instances that can be seen as representatives of a prediction and/or the instance from which the prediction originated.

Supervisors: Hugo Proença, Peter van der Putten, Matthijs van Leeuwen

A definition of interpretability is naturally subjective but is well captured by the following quote: "interpretability is the degree to which a human can consistently predict the model's result" [1]. Interpretability gathered considerable momentum in recent years due to the dissemination of machine learning throughout industry and society. This is especially true for applications domains where decision making is crucial and requires transparency, such as health care and societal problems. From the existing approaches that aim to make machine learning understandable for humans, two clearly stand out: interpretable machine learning and explainable machine learning. The first focuses on the use of models that are simpler and that are naturally understandable by humans, such as linear regression, decision trees, and rule lists. This approach offers a global explanation of the whole dataset or problem at hand. The second approach is concerned with explaining the predictions of models with more complicated internal structures, such as neural networks and random forests. These are commonly referred to as "black-box" models. Explanations for such models can come in several forms and can make use of the models internal parameters, feature importance, feature combination, etc.

This project focuses on the second type of approaches: explainable machine learning. This area has seen a recent increase in relevance due to the introduction of model-agnostic explanations [2], i.e., a set of methods that can be applied to any kind of model and produce reasonable and intuitive explanations to why the model makes its predictions. In general these explanations are of a local nature, i.e., given one model and an instance, the methods explain why this instance was predicted in this way by the model. While most model agnostic techniques have focused on explaining the role of the *variables* used, we propose to instead look at the *instances* in the training set that led to the model making these particular predictions.

These instance based methods are called example-based explanations and can be divided in four types [3]: counterfactual explanations, adversarial examples, prototypes and criticisms, and influential instances. The goal of this project is to combine the ideas behind model-agnostic explanations, based on features, with existing literature of example-based explanations, to develop a new method that uses the best of both. The envisioned benefit of this approach is the intuitive nature of giving several examples to a human, in order to explain why something was decided in a certain way. This new approach can be seen as an independent explanation method or as an added value to existing methods-based on features, as it can give insight about why certain predictions occurred based on the existence of certain samples.

Possible steps for the project

1. Literature overview of example-based explanations and model agnostic methods
2. Enumerate the possible ways of making a model-agnostic framework for example-based explanations.
3. (extra) Derive theoretical bounds to the functionality of the method developed.
4. Apply the methods to case studies that exemplify its usefulness
5. Write down the results found.

Requirements/good to have: skilled in Python programming, good knowledge of machine learning/data mining

References

[1] "Examples are not enough, learn to criticize! Criticism for interpretability" http://papers.nips.cc/paper/6300-examples-are-not-enough-learn-to-criticize-criticism-for-interpretability

[2] "Why Should I Trust You?" Explaining the Predictions of Any Classifier (2016): Model-agnostic explanations for classifiers https://www.kdd.org/kdd2016/papers/files/rfp0573-ribeiroA.pdf

[3] "Interpretable Machine Learning" book: Chapter 6: Example-based Explanations https://christophm.github.io/interpretable-ml-book/