Some Title

Xiaoqi Ma

RWTH Aachen University
xiaoqi.ma@rwth-aachen.de

Abstract. Understanding the decisions made by machine learning models is crucial for decision-makers and end-users. Enforced by GDPR, right to explanation demands businesses to provide understandable justifications to their users for the decision. Thus, it is of paramount importance to elucidate the model decision, which could be measured by interpretability, the degree to which a human can understand the cause of a decision. In order to interpret black box models, model-agnostic approaches could be applied, which provide flexibility in the choice of models, explanations and representation for models. From global interpretability viewpoint, feature importance and global surrogate are explored. We also investigate on the local model-agnostic methods, like LIME and Shapley value. After obtaining the future contribution for each instance, we could use the subgroup discovery technique to figure out interesting patterns. In this thesis, the aim is to build up a python package to provide a collection of tools to explain the black-box models.

Keywords: Gender detection · Textual · Visual · Behavioral

1 Introduction

Machine learning is a set of methods that are used to teach computers to perform different tasks without hard-coding instructions. Over the last decades, Machine learning area has gone through unprecedented growth. Due to the increasing computational power, a myriad of classification or regression tasks could be solved by applying machine learning algorithms. For a simple classification task, like predicting the house prices based on the historical data, a traditional regression model is adequate. However, for tackling complex problems like language translation, more complicated models are required.

When evaluating machine learning models, people have a tendency to focus on the performance by observing metrics like accuracy, precision, recall and etc., which are of course very fundamental. Nevertheless, they neglect the importance of interpretability for the model, which shows the degree for a human can consistently predict the models result[1]. As Albert Einstein once said, If you cant explain it simply, you dont understand it well enough. Therefore, it is of paramount importance to achieve high model interpretability as well to clearly understand the decisions made by the model.

For those models that can be easily explained are called Interpretable models, such as Linear regression, logistic regression, and decision trees, since the results

could be interpreted by exploring into the model parameters. On the contrary, ensemble models or neural networks could be regarded as black box models that decisions cannot be understood by looking at their parameters, which is a major disadvantage for a complex model. Typically, those complicated models could offer better performance while provides less interpretability. However, proper interpretability is crucial to explain the choice made by the model and especially important for decision makers. Besides, right to explanation meaning the right to be given an explanation for an algorithms output was stated by General Data Protection Regulation(GDPR), which requires businesses to provide understandable justifications to their users for decisions [2].

To understand model predictions, some explanation methods are necessary, which are algorithms to provides explanations. An explanation usually links the input feature values of an instance to its model prediction in a human understandable way [Christoph Molnar]. There are plenty of properties of explanation methods, and one of them is the Degree of Importance, which reflects the importance of features in the explanation[3]. Several approaches to calculate the feature importance score are available, and we could rank the score to obtain a general overview of the most dominating features in the black-box model. According to the contributions of the specific variable, we might go step further to find out more detailed explanations through subgroup discovery technique, which is a data mining technique to automatically discover similar patterns from data. For example, once we know the education level is a supreme feature in predicting the salary, we might want to dig some patterns that comply with this explanation or even disagree with this explanation.

Thus, this thesis is aimed to explore the effect of independent variables to facilitate understanding decisions.

2 Related Work

As stated before, there are three main method groups, each focus on one type of features. In this section, several related work reviews outlining each method group to detect gender on the web are discussed.

3 Research questions

The following research questions shall be covered in this thesis.

1. How can we identify the importance of a specific variable in a black box model?
2. How can we discover interesting subgroups under consideration that we constrain specific variables?
3. How can we avoid the redundancy in discovered subgroups? And how can we stabilize the results?
4. How can we make sure the explanation is desirable?

To comprehend and interpret the whole model, we need global interpretability, which means we can understand and explain the interactions between dependent variables and independent variables based on the complete dataset. Trying to figure out the feature importance is always a good step to have a good grasp of global interpretability. Nevertheless, the interactions between input features might have an impact on the analysis, like the existence of collinear variables in the same model, which has to be taken into account. Some methods for feature importance calculation will be mentioned in the later chapter.

After global interpretability exploration, we might be interested to understand Why did the model make specific decisions for a single instance?, which leads to local interpretations. In this case, we focus on each instance and try to understand the model decision for this specific instance based on this local region. Furthermore, we could discover subgroups of the dataset with statistically interesting decisions [Francisco]. In this thesis, we would adopt the Exceptional model mining framework to mine subgroups[]. Obviously, we would like to avoid redundancy in those mined subgroups, therefore, the choice of quality measures is of big concern. Another issue is the stability of our model interpretation, which requires further discussion in this thesis.

4 Methods

4.1 Model-agnostic methods

Apart from the model-specific methods which are intrinsically interpretable, model-agnostic methods can be applied on any machine learning model, which provides a generic framework for interpretability that allows for flexibility in the choice of models. Desirable aspects of a model-agnostic explanation system are model flexibility, explanation flexibility, and representation flexibility as stated in [4]. The following methods which belong to model-agnostic methods will be potentially practiced to provide better explanations for model decisions.

Feature importance The feature importance is measured by the increase in the prediction error of the model after permuting the feature, which is known as permutation importance measurement. A feature is regarded as important if prediction error increases after shuffling feature values as the model depends on the feature for the prediction. Conversely, a feature is unimportant if prediction error seldom changes, which means the feature is hardly relied on for the model. Based on this idea, Fisher, Rudin, and Dominicіproposed a model-agnostic version of the feature importance and called it model reliance [5].

Local Surrogate To explain individual predictions of black box machine learning models, local surrogate models are a good choice. Instead of approximating the predictions of the underlying black box model like global surrogate models, the local surrogate models focus on explaining predictions for an individual instance. Local interpretable model-agnostic explanations (LIME), a well-received

implementation for the local surrogate, provides human-friendly interpretation and is currently available for application usage [lime]. The python package is accessible in [lime py], they claim to support explaining individual predictions for tabular data, texts, and images.

Shapley values The Shapley value, coined by Shapley, is a method for assigning payouts to players depending on their contribution to the total payout. Players cooperate in a coalition and receive a certain profit from this cooperation [6]. It can also be used as a locally accurate additive feature attribution method, and in this case, the shapley value is the average contribution of a feature value across all possible coalitions.

4.2 Subgroup discovery methods

5 Initial Experiments

6 Possible obstacles

First, from global interpretability perspective, we need to calculate the importance for each future. Meanwhile, we need to figure out an elegant method to cope with collinear variables. Besides, the interactions between input features should be considered as well. Through an initial experiment, we know that not all algorithms are suitable for mixed-type tabular data, which contains numeric and categorical features. Therefore, it deserves discussion about whether we modify categorical data by either label encoding or one-hot encoding or modify numerical data by discretization. One step further, while using subgroup discovery techniques, the choice of quality measures plays a vital role. Moreover, we not only need to keep an eye on the algorithm performance but also value on the stability. Even though more challenges might have emerged, we believe all obstacles shall be solved.

7 Potential outcomes

The aim is to build up a python package to provide a collection of tools to explain the black-box models. Ideally, it should at least support the following functions:

1. Provides data preprocessing methods, including feature encoding, data filtering and etc.
2. Calculate feature importance score using various methods
3. Identify feature contributions for a single instance
4. Discover subgroups relying on the feature effect
5. Visualize results in an elegant way

Task Completed	Due data
Topic Approved	1st June
Preliminary literature review	15th June (2 weeks)
Initial Implementation and experiment	29th June (2 weeks)
Prepare datasets	6th July (1 week)
Improve implementations	27th July (3 weeks)
Conduct experiment	10th August (2 weeks)
Implement extensions/ variations	31st August (3 weeks)
Conduct more experiment	14th September (2 weeks)
Writing thesis draft	26th October (6 weeks)
Revise thesis draft	16th November (3 weeks)
Final thesis draft submission	30th Nobember (2 weeks)

8 Time Schedule for Masters Thesis

References

1. B. Kim, R. Khanna, and O. O. Koyejo, "Examples are not enough, learn to criticize! criticism for interpretability," in *Advances in Neural Information Processing Systems*, 2016, pp. 2280–2288.
2. P. Voigt and A. Von dem Bussche, "The eu general data protection regulation (gdpr)," *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 2017.
3. M. Robnik-Šikonja and M. Bohanec, "Perturbation-based explanations of prediction models," in *Human and Machine Learning*. Springer, 2018, pp. 159–175.
4. M. T. Ribeiro, S. Singh, and C. Guestrin, "Model-agnostic interpretability of machine learning," *arXiv preprint arXiv:1606.05386*, 2016.
5. A. Fisher, C. Rudin, and F. Dominici, "Model class reliance: Variable importance measures for any machine learning model class, from the rashomon perspective," *arXiv preprint arXiv:1801.01489*, 2018.
6. L. Shapley, "A value for n-person games. contributions to the theory of games ii. ed. by hw kuhn and aw tucker," *Annals of Mathematics Studies*, vol. 28, 1953.