MASTER THESIS PROPOSAL

Something interesting and meaningful

submitted by

XIAOQI MA

Submitted to the

Chair of Computational Social Sciences and Humanities

within the

Faculty of Mathematics, Computer Science and Natural Sciences

at RWTH Aachen University

May 19, 2019

Advisor:

Dr. Florian Lemmerich

First Supervisor

Dr. Florian Lemmerich

Second Supervisor

Prof. Dr. Markus Strohmaier

Declaration of Authorship

I, Xiaoqi Ma, hereby declare in lieu of an oath that the present Bachelor's thesis titled, "Something interesting and meaningful" and the work presented in it are my own. I confirm that:

- I have completed this thesis indepentently and without illigimimate assistance from third parties.
- I have used no other than the specified sources and aids.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- The written and electronic versions which were submitted are fully identical.
- The thesis has not been submitted to any examination body in this, or similar, form.
- I have read and understood the official notification concerning §156 StGB and §161 StGB.

City, Date:

_____

Signature:

_____

Contents

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.*

0.1 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 model's result[1]. As Albert Einstein once said, "If you can't explain it simply, you don't 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 algorithm's output was stated by General Data Protection Regulation(GDPR), which requires businesses to provide understandable justifications to their users for decisions [3].

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 [2]. 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.

0.2 Section

0.3 Some drawing and tables