Explain variable influence in black box models through pattern mining

Xiaoqi Ma
xiaoqi.ma@rwth-aachen.de
Matriculation number: 383420

Supervisor: Prof. Dr. Markus Strohmaier
Second Examiner: Prof. Dr. Bastian Leibe
Advisor: Dr. Florian Lemmerich

Chair of Computational Social Sciences and Humanities
RWTH AachenFaculty of Mathematics, Computer Science and Natural Sciences
RWTH Aachen University

This thesis is submitted for the degree of
M.Sc. Media Informatics

Aachen, Germany
December 18, 2019

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. Thus, it is of paramount importance to elucidate the model decision, which could be measured by model 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 going to be explored. We also investigate the local model-agnostic methods, like LIME and Shapley value. After obtaining the designated feature contribution for each instance, we could use the subgroup discovery technique to figure out "interesting" patterns to provide more elaborate explanations. In this thesis, the aim is to build up a python package to provide a collection of tools to explain variable influence in black-box models through subgroup discovery.

Keywords: *Black box model interpretability; Model agnostic; Subgroup discovery*

Contents

Chapter 1

Introduction

Blockchain which was first presented in 2008 by Satoshi Nakamoto [?], is an emerging technology with a breakthrough potential. Since 2016, Blockchain have been listed in the Gartner's Technology Trends reports [?], and it is expected to revolutionize the IT, business, and society around the world [?].

In this chapter, I will present the overview of how industrial and academical field think of blockchain technology, how it caught our attention, what is the purpose of this master thesis and how to conduct the ideas. This chapter will give readers rather a complete plan of the thesis .

1.1 Background

1.1.1 Purpose of this thesis

1.1.2 Research questions

1.1.3 Thesis Structure

Chapter 2

Related Work

In this chapter, related work... [1]

2.1 Related work and Applications

2.1.1 Black box models interpretability

2.1.2 Subgroup discovery

2.1.3 Application

Chapter 3

Methods

In this chapter, methods...

3.1 Global interpretation methods

3.1.1 model feature importance

3.1.2 permutation feature importance

3.2 Local interpretation methods

3.2.1 model specific

DeepExplainer, TreeExplainer

3.2.2 model agnostic

KernelExplainer

3.3 Subgroup discovery

3.3.1 numeric target

3.3.2 complex target

3.3.3 redundancy avoidance

Chapter 4

Experiments

Experiments

4.1 Data

4.1.1 Data collection

4.2 Experiments

4.2.1 Experiment1: Binary feature flip

4.2.2 Experiment2: Numeric feature perturbation

4.2.3 Experiment3: Synthetic model evaluation

4.2.4 Experiment4: classification task vs. regression task

4.2.5 Experiment5: decision tree vs. subgroup discovery

Chapter 5

Results and Discussion

Results...

5.1 Results and Discussion

5.1.1 feature effect vs. shapley values

5.1.2 classification vs. regression

5.1.3 decision tree vs. subgroup discovery

5.1.4 tabular data vs. text data

Chapter 6

Conclusion and Future work

conclusion and future work...

6.1 Conclusion and Feature work

6.1.1 Factors to consider

6.1.2 Summary

6.1.3 Outlook

Bibliography

[1] K. L. Mikhail Korobov. (2019) Eli5: A library for debugging/inspecting machine learning classifiers and explaining their predictions. [Online]. Available: https://github.com/TeamHG-Memex/eli5