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

Assume we had a scenario where an unemployed young wanted to get a loan from the bank to start a business. The bank manager served him politely and then entered his information into the banking system, but unfortunately, the system rejected his request. Of course, the young man desired to know why his loan was rejected, and the bank was obligated to provide justifications. Hopefully, the bank manager would receive some explanations from the system telling that the bank was not willing to take risks offering a loan because of his unemployment and no assert mortgage. However, the manager only obtained the final decision made by the banking algorithms rather than the explanations for this decision. If so, how can we trust banking algorithms without explanations and how can we ensure there are no mistakes while making the decisions since the model is a complete "black box" to customers. And that is why we would like to explore variable influence to assist decision-makers to explain model decisions.

In this chapter, I will provide an overview of how to interpret the results generated by black box models, how this topic caught our attention, what is the purpose of this master thesis and finally how to construct the model interpretation framework. This chapter shall give readers a complete plan for this thesis.

1.1 Background

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 boosting computational power and the availability of "Big data", a myriad of classification or regression tasks could be solved by applying machine learning algorithms. For a simple classification scenario, like predicting the house prices based on the historical data, a traditional regression model might be adequate, like logistic regression. However, for tackling complex problems like language translation, more complicated models are required, such as neural networks.

When evaluating machine learning models, people have a tendency to focus more on

the performance by observing metrics like accuracy, precision, recall and etc., which are of course very fundamental to assess the model. Nevertheless, they neglect the importance of interpretability to the model, which shows "the degree for a human to understand model decisions and the ability to consistently predict the results"[1]. For those models with high interpretability, it is rather clear for human to understand the decisions, while for those are not easily interpretable by human, we should utilize interpretation methods to facilitate us to explain the outcomes. Therefore, to have a better understanding of the decisions made by the model, it is of paramount importance to investigate model interpretability.

From model interpretability perspective, models could be classified as white box models and black box models. Roughly to say, white box models have simple structures, a limited number of coefficients and can be understood by human, such as linear regression, logistic regression, and decision trees, since the prediction results could be interpreted by exploring into the model parameters. On the contrary, black box models usually have more complex structures and a substantial number of parameters which are not understandable. For example, ensemble models or neural networks could be regarded as black box models for the reason that decisions made by black box models cannot be understood by looking at their parameters, which is a major disadvantage for complex models. Typically, those complicated models could achieve better performance for the sake of 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 output of automated algorithms was stated by General Data Protection Regulation(GDPR), which requires businesses to provide understandable justifications to their users [2], just like the scenario formerly described that the bank manager should be able to clarify the reason of loan rejection.

Except for the models with interpretable parameters that could be used directly to explain the decision, more model explanation methods should be explored to support the explanation for black box models as well. As defined in [3], an explanation lies the connections between the input feature values and its model output in a human-understandable way. Generally, two broad genres of explanation methods are often mentioned, one is the global interpretation methods and the other refers to local interpretation methods. As the name suggests, the global interpretation focus on the global view of the input variables, more specifically, it points out the most significant features that can affect predictions of the entire data set. After exploring the importance ranking of input features, we could at least obtain a general overview of variable influence. In contrast, our attentions are more inclined to local interpretation methods, which are more compatible with the scenario previously described, targeting at the instance level explanation. In this case, each instance should be supplied with a corresponding explanation specifying the cause to prediction results.

Though the above-mentioned methods could indeed give explanations to some degree, the global methods give a too broad interpretation view while local interpretation may become too sensitive to reveal the underlying cause due to the particularity of that instance, causing either inaccurate or unreliable explanations. Thus, explor-

ing the patterns of explanations could be a good amendment that comes into our mind, which aims to discover subgroups which share interchangeable explanations by applying pattern mining technique.

In this thesis, the huge effort would be made to investigate this novel technique that combines model interpretation methods and pattern mining technique.

1.1.1 Purpose of this thesis

Given the urgent need to obtain decent justifications for every decision made by the algorithms as well as the enforcement by GDPR, a feasible approach is to build a framework to provide explanations for each model regardless of model types. Undoubtedly, many interpretation methods have already come to the surface to facilitate model explanation, but they are merely limited to a single instance explanation, which might be unreliable and causing misunderstanding due to the excessive interpretation of that specific instance. Therefore, it is wiser to not only study the instance explanation but also investigate the groups of instances that have similar explanations using subgroup discovery technique. And we noticed that we could generate more robust explanations by combining those two approaches.

Therefore, in this thesis, I would like to construct a robust framework combing the benefits of model interpretation methods and pattern mining technique to furnish us with good model interpretation.

1.1.2 Thesis Structure

The remainder of this thesis is structured as following.

Chapter 2 focuses on previous work on related fields, such as Model Interpretation methods and Subgroup Discovery field. In particular, it is dedicated to review the existing global interpretation methods and local interpretation methods. In addition, the fundamentals of subgroup discovery are discussed as well like the selection of interestingness measure.

Chapter 3 is concerned with the theoretical knowledge of local interpretation methods to explain black box models. It begins with simple approaches on specific scenarios, for example, to inspect the influence of binary feature using the binary flip approach. Then the methods becomes more general which can be applied on any type of features, like Shapley values. Finally, more attention is laid on the novel technique which combines the approach of model agnostic local interpretation and subgroup discovery.

Chapter 4 presents the detailed description of datasets that are collected and the set up of experiments. In experiments, the comparison of several local interpretation methods is covered. Additionally, we demonstrate case studies on specific datasets.

Finally, Chapter 5 concludes the work with a summary of results and ideas on future

work.

Chapter 2

Related Work

In this chapter, related work... [4]

2.1 Related work and Applications

2.1.1 Model interpretation methods

Global interpretation

Default feature importance mechanism The most common mechanism to compute feature importances, and the one used in scikit-learn's RandomForestClassifier and RandomForestRegressor, is the mean decrease in impurity (or gini importance) mechanism (check out the Stack Overflow conversation). The mean decrease in impurity importance of a feature is computed by measuring how effective the feature is at reducing uncertainty (classifiers) or variance (regressors) when creating decision trees within RFs. The problem is that this mechanism, while fast, does not always give an accurate picture of importance. Breiman and Cutler, the inventors of RFs, indicate that this method of "adding up the gini decreases for each individual variable over all trees in the forest gives a fast variable importance that is often very consistent with the permutation importance measure." (Emphasis ours and we'll get to permutation importance shortly.)

Permutation importance

- Feature importance
- Partial dependence plot

Local Interpretation

2.1.2 Subgroup discovery overview

2.1.3 Applications

Chapter 3

Approach

In this chapter, the details of approaches will be discussed. It starts with an overview of local interpretation methods and several variants of them are introduced respectively. Firstly, we present the binary feature flip idea which aims to characterize the impact of binary features by flipping the feature values. After that, we are interested in the effect of numeric features in the model by inspecting the outcome change of the model when the numeric feature values are perturbed to generate noises. Despite the "variable-specific" methods, we also focus on local interpretable model-agnostic explanations (LIME) which is able to explain individual predictions for any types of features and models. However, no theory can support why LIME can fit linear behavior locally on black box models. Therefore, we continue exploring Shapley value, which is a reasonable explanation method with well founded theory. In addition, the appealing approach assigns a contribution score for each feature value to smooth the path of interpreting the final prediction of individual instances by calculating the Shapley value.

Then the following section describes the novel technique which combines the local interpretation methods and pattern mining technique. Since the target concept during subgroup discovery in our situation is either prediction change or feature influence score, therefore, the focus shall attribute to numeric target. Later, the standard approaches to measure the interestingness of subgroups are discussed. Furthermore, methods to avoid redundancy in subgroups are explored.

3.1 Local interpretation methods

In comparison to Global interpretable methods which are dedicated to explain the global model output by comprehending the entire datasets, it is more interesting to examine the model prediction for an individual instance. Besides, it could be observed that the global interpretation methods are less sensitive to noises if we make some perturbations on feature values, however, it could lead to tremendous changes in the prediction for an instance. Therefore, the local explanations shall preserve high accuracy than global explanations. In the following, few local interpretation methods will be covered in detail.

3.1.1 Binary feature value flip

Binary feature implies that the feature only contains two unique values. In another words, if it is encoded as discrete numeric number, the feature value should be either 1 or 0. Thus, to flip binary feature value means to convert from 1 to 0 or the other way around. In practice, we could also use XOR operation to map from 1 to 0. For instance, gender is regarded as a binary feature which only holds value "male" and "female".

As mentioned previously, the assumption is that we hold the dataset and the corresponding model trained on that dataset. Initially, we could obtain the prediction from the model for a specific instance. Then, a binary value is flipped on a chosen feature and afterwards a new prediction is generated by applying the model to the modified instance. Therefore, as a simple measurement, the effect of this binary feature could be estimated by the difference of two outputs.

In practice, there are two variants to assess the variable influence. One way is to calculate the absolute difference of two predictions, and in this way we could ignore the bias of this binary feature on the original dataset. Literally to say, the binary feature is more influential when the difference becomes larger. In contrast, we could compute the difference for a defined direction, for example, we just care about the effect of gender changing from male to female. In this case, not only the magnitude of the effect is obtained, but also the positive or negative sign towards the prediction.

3.1.2 Numeric feature value perturbation

As the name suggests, this technique is applicable on features whose type is numeric. The idea is that we could apply binary operations to the input values to produce new values, which serves as injecting noises into the original dataset. In particular, only addition and subtraction are considered in this situation. For example, an instance includes a numeric feature called "age" and we could perturb this feature value by increasing or decreasing by a certain value to obtain the modified value.

The procedure of measuring the effect of a chosen input feature is similar to that in binary feature value flip approach. For classification or regression tasks, we could make predictions with the existing model on the instance we desire to explain. Afterwards, a new prediction is made on the adapted instance which is produced through perturbation on the selected numeric feature. And the impact of this numeric feature could be approximately evaluated by the absolute difference of two output predictions, which indicates that this particular feature plays an important role in this instance, causing unstable predictions. Roughly to say, larger prediction differences might imply the feature has stronger effect on the corresponding instance.

3.1.3 Local Surrogate: LIME

Various criteria can be used to classify types for machine learning interpretability. Intrinsic interpretability, for example, is one type of the interpretability methods, which refers to models that are intrinsic interpretable owing to their simple structures, such as linear models or decision trees. In contrast, post hoc interpretability is meant to analyze the model interpretability after model training. As introduced earlier, permutation feature importance is a post hoc interpretation method.

In this thesis, we would like to focus on post hoc interpretability, which indicates to explain model decisions after model has been trained. In particular, model agnostic interpretation methods, which extracts post hoc explanations by treating the original model as a black box, is highly valued. The model agnostic interpretation method is pretty flexible in terms of models, and it can work with any type of machine learning models, which provides a great advantage over model specific methods [5]. The principle behind is to learn an interpretable model on the decisions of the black box model and in return apply the interpretable model to those predictions that are expected to explain.

3.1.4 SHapley Additive exPlanations: SHAP

As we have seen, numerous approaches have been recently proposed to explain predictions for individual instances of black box models. As stated in [6], the presented approach is relied on the decomposition of a prediction for a single instance on individual contributions of each attribute, and the contribution for each feature value is measured as the difference between the output value and the average output over all perturbations of the corresponding feature. Nevertheless, this approach fails to work if the features are conditionally dependent.

Inspired by the coalitional game theory which instructs us to fairly distribute the "payout" among the "players", a general method for explaining black box models by taking into account interactions between features can be found in [7], whose fundamental concepts are borrowed to explain instance-level predictions with contributions of each feature values. Corresponding to the known concept in coalitional game theory, the contributions of individual feature values are called Shapley Value.

Despite from the abstract concept, an illustration taken from [3] might help us intuitively understand the Shapley value. Imagine there is a room and all feature values of a individual instance enter the room in a random order. All feature values, seen as players, need to collaborate with each other to participate the game, where each player contributes to receive the final prediction. And each order of feature values represents a coalition. Consequently, the Shapley value of a feature value corresponds to a difference in value of a coalition when the feature is added to it. In another words, the Shapley value is the average marginal contribution of a feature value across all possible coalitions.

Then, let us have a detailed look at the formal definition of Shapley value as shown

in equation 3.1, where S is the subset of the features in a individual instance, p is the number of features, and x is the vector of feature values of the instance to be interpreted. As for characteristic function val, it describes the contribution of feature j in each coalition.

$$\phi_j(val) = \sum_{S \subseteq \{x_1,\ldots,x_p\} \setminus \{x_j\}} \frac{|S|!(p-|S|-1)!}{p!} \left(\text{val}\left(S \cup \{x_j\}\right) - \text{val}(S)\right) \tag{3.1}$$

Referred to [8], the Shapley value can provide the unique solution that adheres to he desirable properties, which are Efficiency, Symmetry, Dummy, and Additivity.

Efficiency: denoted as 3.2, which requires that the sum of feature contributions must equal to the difference of the final prediction and the average prediction over all coalitions.

$$\sum_{j=1}^{p} \phi_j = \hat{f}(x) - E_X(\hat{f}(X)) \tag{3.2}$$

Symmetry: The contributions of two feature values j and k are the same, which means equation 3.3 should be satisfied.

$$\begin{gathered} if\ val\left(S \cup \{x_j\}\right) = val\left(S \cup \{x_k\}\right) \\ then\ \phi_j = \phi_k \end{gathered} \tag{3.3}$$

Dummy: The contribution of feature j is 0 if it does not change the predictions when it joins into any coalitions. This properties can be demonstrated in equation 3.4.

$$\begin{gathered} if\ val\left(S \cup \{x_j\}\right) = val\left(S\right) \\ then\ \phi_j = 0 \end{gathered} \tag{3.4}$$

Additivity: For any pair of games v, w, the combined payouts should equal to the sum of two individual payouts, as shown in equation 3.5. For example, if we trained a random forest and the additivity axiom guarantees that we can calculate the Shapley value for each tree respectively then average them to obtain the final Shapley value.

$$\begin{gathered} \phi_j(v+w) = \phi_j(v) + \phi_j(w) \\ where\ (v+w)(S) = v(S) + w(S) \end{gathered} \tag{3.5}$$

Though classical Shapley value leads to a potentially promising result, this approach is too computationally expensive owing to computations for the exponential number of possible coalitions. Feasibly, approximation algorithms could be used to reduce

the computational complexity, nevertheless, it inevitably will increase the variance for the calculation of Shapley value. What is worse, the explanation for the prediction of a model is just a simple value, rather than a explanation model like LIME, which fails to make judgments about the connections between input change and prediction change. To address those problems, Lundberg and Lee [9] proposed a unified framework for explaining predictions, which is based on the Shapley value, and they named it SHAP(SHapley Additive exPlanations). This novel approach unifies existing explanation methods and brings more clarity to the methods space. They introduced the explanation model by treating the explanation of a individual prediction as a model. Of course, the unique solution is guaranteed with the game theory. In addition, it provides a more human-understandable and intuitive explanation by user studies as they claimed.

In this case, SHAP values are introduced as a novel measure of feature contribution. Similar to classical Shapley value estimation methods, SHAP values provide the unique additive feature importance measure if the flowing properties are satisfied, which are Local accuracy, Missingness, and Consistency [9]. From another perspective, SHAP method transforms the Shapley value approach into an optimization problem by using kernel function to measure proximity of instances. Within this domain, the novel approximation model agnostic method is called kernel SHAP, which is a combination of LIME and Shapley value. In order to use linear explanation model to locally approximate predictions, we should minimize the following objective function 3.6.

$$\xi = \arg\min_{g \in \mathcal{G}} L\left(f, g, \pi_{x'}\right) + \Omega(g) \tag{3.6}$$

It is intended to obtain the unique solution of equation 3.6, which should also be in line with those three properties, the Shapley kernel is defined as [9]:

$$\begin{aligned} \Omega(g) &= 0 \\ \pi_{x'}\left(z'\right) &= \frac{(M-1)}{(M \text{ choose } |z'|)\,|z'|\,(M-|z'|)} \\ L\left(f, g, \pi_{x'}\right) &= \sum_{z' \in Z} \left[f\left(h_x\left(z'\right)\right) - g\left(z'\right)\right]^2 \pi_{x'}\left(z'\right) \end{aligned} \tag{3.7}$$

where $|z'|$ is the number of non-zero elements in z'

We are interested in this novel model explanation method, therefore, further experiments will be conducted to explore the usability of it.

3.2 Pattern mining with Local interpretation methods

3.2.1 Target Concept

Binary target

Nominal target

Numeric

3.2.2 Selection Criterion

Standard QF

Interestingness Measure

3.2.3 Redundancy avoidance

3.2.4 Decision trees with Local interpretation methods

Chapter 4

Experiments

Experiments

4.1 Datasets

4.1.1 Datasets description

4.1.2 Data preprocessing (experiment settings)

4.2 Experiments

4.2.1 Artificial Datasets

recover subgroups

4.2.2 Comparison of different local interpretation methods

binary flip: perturbation, LIME, SHAP

numeric perturbation: perturbation, LIME, SHAP

classification vs. regression

decision tree vs. subgroup discovery

4.2.3 Case Study

Chapter 5

Conclusion and Future work

conclusion and future work...

5.1 Conclusion and Feature work

5.1.1 Factors to consider

5.1.2 Summary

5.1.3 Outlook

Bibliography

[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] C. Molnar, *Interpretable Machine Learning*, 2019, https://christophm.github.io/interpretable-ml-book/.

[4] 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

[5] M. T. Ribeiro, S. Singh, and C. Guestrin, "Model-agnostic interpretability of machine learning," *arXiv preprint arXiv:1606.05386*, 2016.

[6] M. Robnik-Šikonja and I. Kononenko, "Explaining classifications for individual instances," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 5, pp. 589–600, 2008.

[7] I. Kononenko *et al.*, "An efficient explanation of individual classifications using game theory," *Journal of Machine Learning Research*, vol. 11, no. Jan, pp. 1–18, 2010.

[8] L. Shapley, "A value for n-person games," *Contributions to the Theory of Games*, pp. 31–40, 1953.

[9] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, 2017, pp. 4765–4774.