Generative Local Interpretable Model-Agnostic Explanations

Mohammad Nagahisarchoghaei, Mirhossein Mousavi Karimi, Shahram Rahimi, Logan Cummins, Ghodsieh Ghanbari
Computer Science and Engineering, Mississippi State University
Mississippi State, MS 39762
mn890@msstate.edu, mm4949@msstate.edu, rahimi@cse.msstate.edu,
nlc123@cavs.msstate.edu, gg646@msstate.edu

Abstract

The use of AI and machine learning models in the industry is rapidly increasing. Despite their success, when used for decision-making, AI solutions have a significant drawback: transparency. To address this issue, algorithms such as LIME and SHAP (Kernel SHAP) have been introduced. These algorithms aim to explain AI models by generating data samples around an intended test instance by perturbing the various features. This process has the drawback of potentially generating invalid data points outside of the data domain. In this paper, we aim to improve LIME and SHAP by using a pretrained Variational AutoEncoder (VAE) on the training dataset to generate realistic data around the test instance. We also employ a sensitivity feature importance with Boltzmann distribution to aid in explaining the behavior of the black-box model surrounding the intended test instance.

Introduction

Machine learning and AI models are playing a significant role in today's society, however, AI models' undesirable behavior may frustrate and confuse users, and AI decisions may have a direct negative impact on human life, particularly in vital decision-making domains such as financial loans, insurance quotes, disease diagnostics (Khojaste 2022), recidivism risk assessment (Soares 2019), game theory-based predictive analytics (Mousavi et al. 2021 and 2022), and self-driving cars (Zablocki 2021) among others. In recent years, the field of explainability has expanded exponentially . One of the major methodologies in explainability is prediction-based explainability. The prediction-based explainability methods provide a point-wise explanation. In other words, a single attribution or saliency map might not explain the whole picture and result in a wrong interpretation. One idea to solve this issue would be examining the behavior of the model in a vicinity of the place of interest not just single points.

LIME (Ribeiro, Singh, and Guestrin 2016) and later SHAP (Lundberg and Lee 2017) respond to this issue by using the local perturbations-based methodology with simple formulation and fast data generation. The perturbation-based method depends on querying the black-box model in vicinity of the prediction of interest to infer relevance of input features towards the output. Alvarez-Melis and Jaakkola evaluated and questioned the robustness of the perturbation-based methods like LIME and Kernel SHAP (Alvarez-Melis and Jaakkola 2018).

Continuing their evaluation, the purpose of this work is to further investigate LIME and SHAP's flaws related to using perturbation and to propose a new methodology in order to remedy the flaws in these methods.

Methodology

This section describes the proposed model-agnostic architecture in detail. First, the data generation using VAE is explained. Then, a feature importance measure is formulated to provide global explainability for local data points.

For the data generation local sampling around test instance, we utilize a generative model. After training VAE, the test instance i is passed to the encoder and corresponding μ_i and $logvar_i$ are produced. Then we use normal sampling to generate 5000 new $logvar$s around $logvar_i$. We sample these new $logvar$s based on normalized euclidean distance from $logvar_i$. Samples can be taken from the $N(\mu_i, N(logvar_i, 0.3))$ and fed to the decoder to generate similar data points to input data around the test instance.

The model reliance is a model-agnostic version of feature importance Breiman (2001) and Fisher, Rudin, and Dominici (2019). In the model reliance implementation, we permute feature i of the input data, feed the permuted data to the prediction model, calculate the model's prediction for this permuted data, and measure estimated error on the prediction. The sensitivity analysis (permuted feature importance) can summarize the features' local sensitivities in the vicinity of the test instances.

The black-box model will be fed with the permuted data X^i and the originaly generated data X to achieve the model sensitivity to feature i:

$$S_i = \frac{1}{n}\sum_{k=1}^{n} g(x_k) - g(x_k^i) \tag{1}$$

where $k = 1, 2, ..., n$ is the i^{th} generated data, function g is inference of the black box model, n is the number of all generated data (n=5000), and x_k^i and x_k are the permuted data at feature i and the original ones, respectively.

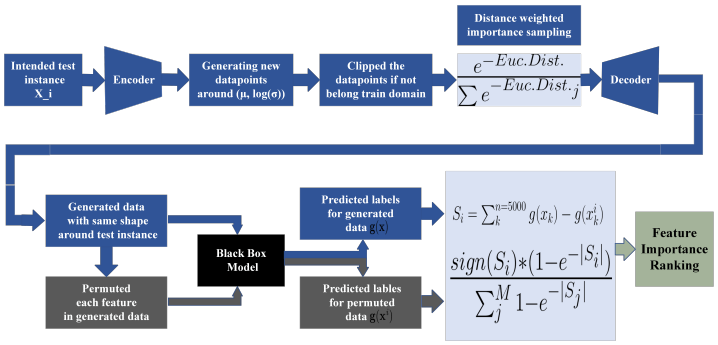

Figure 1: Process block diagram for local explainability method

Local generated data will have less sensitivity to permuting feature j, and consequently, less importance if differences between permuted data and generated data (S_i) are close to zero. Let us consider $1 - e^{-|S_i|}$ as feature i's sensitivity, which maps the output to [0-1]. It will be zero when S_i approaches zero, indicating that the black-box model is less sensitive to feature i.

Next, the feature importance can be obtained by normalizing feature sensitivity after applying Boltzmann distribution to S_i. Partition function will be the sum of all feature sensitivity:

$$FI_i = \frac{sign(S_i) * (1 - e^{-|S_i|})}{\sum_{j=1}^{M} 1 - e^{-|S_j|}} \tag{2}$$

where $j = 1, 2, ..., M$ is used to refer the j^{th} feature number, and M indicates the number of features. FI_i is in the interval $[0, 1]$ and represents the feature importance level calculated for feature i Figure 1.

Results

In this section, we provide the explanation obtained by the presented method and compare it with the results of Linear LIME and Kernel SHAP methods on Boston Housing Dataset. Lundberg and Lee (2017) illustrated that coefficient of a local surrogate linear regression model with appropriate weighting kernel can approximate the Shapley values. Therefore, due to similarity in methodology of both Linear LIME and Kernel SHAP, we use both of them for further comparison. We selected a randomly selected data point (41) in the test data set. The best way to compare the result of all models is visualizing the scatter plot of selected features distribution and median house value (target feature). When looking at LIME, SHAP, and the presented method results in Figure 2, LIME and SHAP put the most feature importance on a low RM value. the presented method denotes a low DIS value as the most import feature. The impacts that these features and more should have on the target value, Median Value of a house, can be determined using the distributions in Figure 3. Both LIME and SHAP determine that RM should negatively affect the Median Value which should not be true given the distribution. They also determine that LSTAT should positively affect the Median Value which follows the same explanation. Our model is able to properly indicate both of these factors as important, and our model is able to properly indicate the correct correlations. As seen in Figure 2, LIME and SHAP put the most feature importance on RM and LSTAT. The presented method notes DIS as the most important feature followed by RM and LSTAT as second and third, respectively. All three methods agree on the importance of RM and LSTAT based on their magnitudes which are supported by the clear linear trends surrounding our point of interest (red) in Figure 3b and Figure 3f. These linear trends confirm that both features are good predictors for Median House Value. With all of the methods agreeing on these features, we proceed to looking at the features ranked differently and their disagreements. The presented method disagrees with SHAP and LIME on the importance of DIS, CRIM, and TAX. In the DIS distribution, a high-density of points surround our point of interest in red in Figure 3a. These points show a linear trend between DIS and Median House Value in the locality of our point of interest. We can conclude that DIS is a good predictor of Median House Value in the vicinity of the point of interest, but LIME and SHAP put low importance on DIS. Both LIME and SHAP determine that CRIM is an important feature; however, the distribution, shown in Figure 3c, illustrates that there is no evident trend in the vicinity of the point of interest. The presented methods ranking of CRIM as the 9th most important feature is more supported by the uncorrelated distribution between CRIM and Median House Value around the point of interest.

Conclusion

Slack et al. (2020) outlined that both LIME and SHAP may be unreliable due to relying on the input perturbations to learn local area around an intended data point. To address this drawback, we argued that the proposed method can be effective in generating realistic data around the intended data point. This was done by using VAE to stricken the level of perturbing around a data point, so the generated data points could more accurately represent the original data point. Besides, the results show that the proposed method can give priority to the feature with higher predictive power in the vicinity of the intended data point, which is an important benefit.

References

ALVAREZ-MELIS, D., AND JAAKKOLA, T. S. On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049* (2018).

BREIMAN, L. Random forests. *Machine learning 45*, 1 (2001), 5–32.

FISHER, A., RUDIN, C., AND DOMINICI, F. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *J. Mach. Learn. Res. 20*, 177 (2019), 1–81.

KARIMI, M. M., AND RAHIMI, S. A two-dimensional model for game theory based predictive analytics. In *2021*