Probabilistic Post-hoc Explainable AI methods

Aditya Saini
Indraprastha Institute of Information Technology, Delhi
New Delhi, India
aditya18125@iiitd.ac.in

Ranjitha Prasad
Indraprastha Institute of Information Technology, Delhi
New Delhi, India
ranjitha@iiitd.ac.in

ABSTRACT

Traditional Post-hoc Explainable AI techniques such as LIME and SHAP are widely used for providing simple but quantitative explanations for the instance of interest. However, they suffer from multiple drawbacks attributed primarily to their method of generating surrogate samples, which renders such techniques to be unreliable in safety-critical domains such as healthcare and robotics, where the notion of trustworthiness and consistency are of the utmost importance. In this abstract, we highlight the open challenges emerging in post-hoc explainable models and motivate two novel post-hoc explanation techniques that propose probabilistically relevant approaches to generating surrogate samples, hence solving several issues encountered in traditional methods.

KEYWORDS

Explainable AI, Gaussian Process, Active Learning, Mixture of Gaussians, Variational Inference

1 INTRODUCTION

With data pouring from several applications coupled with the increased computational capability of modern systems, the field of Artificial Intelligence (AI) is the cornerstone of research and development. Recent research has highlighted that the complexity of an AI model is directly proportional to the quality of its results, be it generative or predictive. But the complex decisions taken by these models are too complicated for human minds to understand, thus making the end-user doubtful of the models' fairness and trustworthiness. Therefore explainable AI (or XAI) methods are of interest.

1.1 Traditional Explainers

Explaining a prediction of an existing black-box model involves presenting textual or visual artifacts that provide a qualitative understanding of the relationship between the instance's features and the model's prediction [7]. Gradient-based approaches such as DeepLIFT[9] and DeepSHAP[3] are model-aware techniques that explain gradient-based models such as neural networks. Mimic models such as decision trees attempt to mimic a predictive model's decisions iteratively. Post-hoc, perturbation-based methods such as LIME (Local Interpretable Model-Agnostic Explanations) [7] and SHAP[5] rely on feature permutation for their explanations. The model-agnostic property of these techniques plays an essential role in their popularity, i.e., they provide explanations independent of the training data modality and architecture of the prediction model. By producing weighted perturbations (surrogate data) in the neighborhood of the instance of interest, these techniques employ locally weighted regressors to obtain per-feature importance weights.

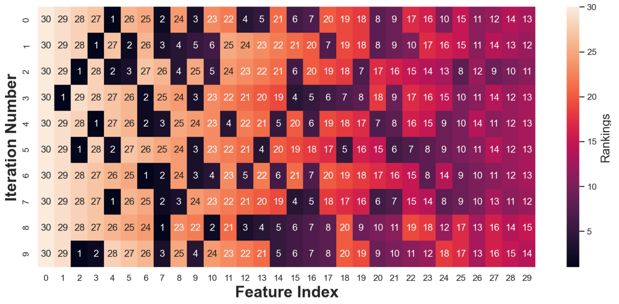

Figure 1: Variation of Importance scores across iterations

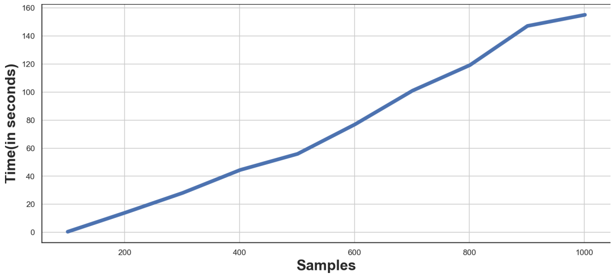

Figure 2: Sample/Time Complexity

1.2 Open Challenges

Despite the widespread usage of LIME and SHAP, subsequent works have pointed out various issues:

1.2.1 Inconsistency in repeated explanations: Owing to the technique employed for generating surrogate samples, LIME leads to inconsistent explanations [11, 12] when invoked multiple times. We depict this effect by first training a Support Vector Classifier on the Breast Cancer classification dataset [2] in Fig. 2.a) and using LIME to generate explanations for a randomly selected sample 10 consecutive times. As shown from the figure, LIME produces different rankings for many features when called consecutively, albeit same initial settings.

1.2.2 Sample Inefficiency: Both LIME and SHAP require a large number of samples in their surrogate datasets due to the low inductive bias of the linear explainer. This essentially implies that the number of samples in the surrogate dataset scales with the number of features. This interplay between the sample/time complexity and computational complexity of the explainer models is crucial [10, 12]. We depict this issue in Fig. 2.b), where we measured the time required to generate explanations for a pre-trained ResNet-152 model[4] for a randomly selected image from the Imagenet[8] dataset.

2 CURRENT WORK

The lack of structure in the sampling process hampers the quality of surrogate data [6]. This issue was addressed in BayesLIME and BayesSHAP strategies [10], which select informative samples from the surrogate dataset. However, sampling in BayesLIME and BayesSHAP is heuristic, and one wonders if a more principled probabilistic approach is feasible. A non-linear explainer overcomes the shortcomings of linear explainer models with respect to the bias-variance trade-off by avoiding a linear locality. Furthermore, using Bayesian methods for explaining ensures that fewer surrogate samples are required for providing explanations.

2.1 UnRAvEL: Uncertainty driven Robust Active learning-based Explanations

UnRAvEL is a novel explainer where we propose a novel acquisition function called "Faithful Uncertainty Reduction(FUR)" along with a Gaussian Process (GP) based explainer for probabilistic local sampling and explanations by trading-off information gain and local fidelity. FUR is given by

$$\mathbf{x}_n = \arg\max_{\mathbf{x}} - \underbrace{\left\| \left(\mathbf{x} - \mathbf{x}_0 - \frac{\overline{\sigma}\epsilon}{\log(n)} \right) \right\|_2}_{\text{T1}} + \underbrace{\sigma_n(\mathbf{x})}_{\text{T2}}, \qquad (1)$$

where $\overline{\sigma}$ is the empirical mean of the standard deviation of individual features in training data, $\epsilon \sim \mathcal{N}(0,1)$, $\mathbf{x}_0$ is the index sample, and $\sigma_n(\mathbf{x}_n)$ is the standard deviation of f_e obtained until the n-th sample $\mathbf{x}_n$. Here the term 'T1' helps in maintaining local fidelity and the term 'T2' helps in trading off information gain through uncertainty. The GP based explainer module provides explanations using the inverse length-scale hyperparameters of the ARD(Automatic Relevance Determination) kernel.

2.2 GLIME: Gaussian mixture based Local Interpretable Model agnostic Explanations

To counter the instability of LIME[7] in repeated iterations, DLIME (Deterministic LIME)[11] employed an agglomerative clustering-based approach instead of random sampling. The central idea is to cluster the entire dataset and recognize the cluster in which the instance of interest lies so that the neighborhood points are sampled as surrogate data. The biggest drawback of DLIME is that it employs training data for clustering. Further, since DLIME uses label-based clustering, some clusters may contain very few members due to the sparsity of labels in imbalanced datasets. In summary, DLIME counters stability but fails to generate high-quality explanations.

GLIME is a probabilistic sampling alternative that exploits the generative and regularizing properties of Gaussian mixture models (GMM)[1]. We assume that we have information regarding the Gaussian cluster mean and variances that fit the actual data distribution. This can be obtained from the training data and must be made available during the explanation. While the availability of the cluster centers and variance allows us to provide soft assignments to the instance of interest, the generative nature of GMMs enables us to generate surrogate samples using popular sampling techniques like MCMC(Markov Chain Monte Carlo). Incorporating a prior in GMM ensures that overfitting is decreased even in highly imbalanced data. Furthermore, since only the cluster centers and variances need to be made available during the explanation phase, it saves on storage requirements compared to DLIME and addresses privacy issues, especially in decentralized training methodologies such as federated learning. Since choosing the optimal number of clusters is crucial, we proposed a *Bayesian* Gaussian Mixture model, which uses ARD for cluster parameter selection. [1].

3 FUTURE WORK

In this extended abstract, we provide a brief introduction to two novel XAI techniques, namely GLIME and UnRAvEL. The core theme of our research is to provide XAI solutions in use cases where the concepts of trustworthiness and consistency are of the utmost importance. Currently, we are working on the following:

- **Global Explainer based on UnRAvEl:** GPs are computationally complex, which does not affect the local setting, but makes it hard to be employed in the global setting. We are working on building a global explainer based out of sparse approximations of Gaussian Processes.
- **Multimodal joint explanations:** The kernel used in the Gaussian Process explainer can be utilized in many domain-specific applications. Following that direction, we're working on building a novel explainer module that can consider ML models of different modalities.
- **GLIME using Bayesian Optimization:** To make GLIME hyperparameter free, we are working on Bayesian Optimization based pre-processing module for choosing the optimal hyper priors used in the GMM module.

REFERENCES

[1] Christopher M Bishop. 2006. Pattern recognition. *Machine learning* 128, 9 (2006).

[2] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

[3] Zeon Trevor Fernando, Jaspreet Singh, and Avishek Anand. 2019. A study on the Interpretability of Neural Retrieval Models using DeepSHAP. *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (2019).

[4] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), 770–778.

[5] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. *NeurIPS* 30 (2017), 4765–4774.

[6] Brent Daniel Mittelstadt, Chris Russell, and Sandra Wachter. 2019. Explaining Explanations in AI. *Proceedings of the Conference on Fairness, Accountability, and Transparency* (2019).

[7] Marco T Ribeiro, S Singh, and C Guestrin. 2016. Why should i trust you? Explaining the predictions of any classifier. In *Proceedings of ACM SIGKDD*. 1135–1144.

[8] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* 115, 3 (2015), 211–252. https://doi.org/10.1007/s11263-015-0816-y

[9] A Shrikumar, P Greenside, and A Kundaje. 2017. Learning important features through propagating activation differences. In *Proceedings of ICML*. PMLR, 3145–3153.

[10] Dylan Slack, Sophie Hilgard, Sameer Singh, and Himabindu Lakkaraju. 2021. Reliable Post hoc Explanations Modeling Uncertainty in Explainability. In *Neural Information Processing Systems (NeurIPS)*.

[11] Muhammad R. Zafar and N. M. Khan. 2019. DLIME: a deterministic local interpretable model-agnostic explanations approach for computer-aided diagnosis systems. *arXiv:1906.10263* (2019).

[12] Xingyu Zhao, Xiaowei Huang, Valentin Robu, and David Flynn. 2020. BayLIME: Bayesian Local Interpretable Model-Agnostic Explanations. *Proc. of UAI* (2020).