Learning Beyond Black Box: Symbolic Interpretability for Classification Models

Avinash Kori
koriavinash1@gmail.com

Abstract

The black-box nature of deep learning models prevents them from being deployed or completely trusted, especially in biomedicine domains. Most explainability techniques try to capture the importance of feature or feature interactions. Current Symbolic methods are used in the context of regression models; in this work, we aim to use symbolic methods to understand any classification black-box models. Here, we use a modified projection pursuit algorithm and optimize for both projection plane and projection function. The family of projection functions used in this work is the Meijer G-functions, which are generalized hyperbolic functions and help capture the model's continuous global mathematical trend. Optimizing the hyperparameters of Meijer G-functions allows us to capture the trend for most well-known functions and doesn't bind us from fixing on polynomial, trigonometric, or splines. We outline a modified projection function formulation for 1-D and 2-D data along with constrained optimization function. Later, we demonstrate the results by interpreting MLPs and CNNs trained on multiple UCI datasets along with MNIST. Due to time constraints, we evaluated single term symbolic interpreters, which seem to be less accurate but still comparable with other interpretability methods, indicating the scope for further in-depth analysis.

1 Introduction

Current deep learning models are black boxes. As they are integrated with real-life applications and especially in medical diagnosis, it becomes necessary to understand and demystify these models' workings. Clinicians prefer upfront information about the global properties, such as models known strengths and limitations [1]. The explainability of deep learning models can be studied in two different ways, i.e., post-hoc and ante-hoc methods. Post-hoc methods tend to analyze the trained deep learning models and extract hidden meaning and representations learned by a network [2, 3, 4], In contrast, ante-hoc methods try to build more transparent and explainable model while training itself [5, 6]. The majority of current explainability research is in post-hoc methods, which mainly focus on visualizing network attributions or illustrative samples in the input space [7, 8] or analyzing feature importance and feature interactions [9, 10].

Interpretability in the sense of assembling analytical functions is known as Symbolic interpretability. In symbolic interpretability, the superposition of known functional forms is used to approximate the behavior of black-box models. In most previous works, this method was explored in the sense of regression models, where the series of analytical functions were used to learn a given dataset, [11, 12] works used evolutionary methods to estimate hyperparameters. Later, in [13] authors proposed meta-model based on Meijer G-functions [14] to model black-box functions. Recently, in [15], authors extended the work of [13] so that the resulting analytical expression is both accurate and parsimonious and capture a global trend.

This work is the direct extension of [15]; here, we try to extend the method for classification data by formulating a constrained optimization function. Here, the primary assumption is that the decision

Preprint. Under review.

boundaries can be approximated as a continuous function. On a very broad perspective, the problem formulated here is using multiple Meijer G-functions with the constraints on projection vector (to enforce perpendicularity on different classes). The method is outlined in next session 2, section 3 describes the results on synthetic and couple of real-world datasets, section 4 discusses the limitations of the proposed method along with the possible way to address them, and concludes the work.

2 Materials and Methods

2.1 Projection-Pursuit

In case of 1-D data feature vector $\mathbb{X} \in [0,1]^d$, where d is the dimensionality of the vector, The projection pursuit algorithm [16] for regresion is formulated as $\hat{f}(x) = \sum_{k=1}^{K} g_k(v_k^T x)$. For classification task, as described in [17], instead of using trees structure we are using similar concepts of incorporating multiple projection vectors with joint optimization. Joint optimization helps in learning class dependent importance vectors and feature correlation. The proposed projection function $\hat{f}$ is described in the equation (1), where $x \in \mathbb{X}$, c indicates class index, v^c corresponds to class specific feature importance vector, $v_k^c \in \mathbb{R}^d$, and n correspond to total number of classes. Projection vectors $\{v^c\}$, are constrained to be perpendicular to projection vectors of other classes, so in general there are $\binom{n}{2}$ constrain equations. The final softmax function ensures that each class outputs are squeezed between $(0,1)$ along with their sum to be 1.

$$\hat{f}(x) = \text{softmax}\left(\left[\sum_{k=1}^{K} g_k^c(v_k^{cT} x)\right]_{c=1,2,3,\ldots,n}\right) \tag{1}$$

The joint optimization is performed to estimate $(\{g_j^c\}, \{v_j^c\})$. As the defined method is iterative in nature, the approximation of $\hat{f}_j$ is obtained by previous $j-1$ terms. To capture as terms as possible, the optimization is performed on residuals, similar to approach defined in [15]. The residual term is defined as $r_j = f - \hat{f}_j$, the optimization function is defined by minimizing the log probabilities of prediction probabilities as described in equation (2), in which the first term $-\sum_c f(x)\log(\hat{f}(x))$ corresponds to cross entropy loss, while the second term $\sum_c \hat{f}_j(x)\log(\hat{f}(x))$ helps in reducing variance in learning curve (or ensures smooth learning) and helps in generating produce parsimonious expressions.

$$\begin{aligned}
(\{g_j^c\}, \{v_j^c\}) &= \arg\min_{G, \mathbb{R}^{c\times d}} \quad -\sum_c \left(r_j \log(\hat{f})\right) \\
\Rightarrow (\{g_j^c\}, \{v_j^c\}) &= \arg\min_{G, \mathbb{R}^{c\times d}} \quad -\sum_c \left((f(x) - \hat{f}_j(x))\log(\hat{f}(x))\right) \\
&\text{s.t.} \quad v_k^{pT} v_k^q = 0 \\
&\qquad \forall \quad p \in \{1,2,3,...,n\} \\
&\qquad \forall \quad q \in \{1,2,3,...,n\} \setminus p
\end{aligned} \tag{2}$$

In the case of 2-D data, $\mathbb{X} \in [0,1]^{(m\times n)}$, where $m \times n$ is the dimensionality of the input vector. The function argument is modified by including v, u projection vectors, which ensures projections in both the dimensions of the data X. These projections vectors u, v can be considered as row importance vector and column importance vector respectively, equation (3) describes the same. By singular decomposition of X, It can be easily shown that v, u correspond to projection vectors for left and right singular vectors of matrix $X \in \mathbb{X}$, which is described in equation (4), where V is left singular matrix, U is a right singular matrix, and D is diagonal Eigenvalue matrix.

$$\hat{f}(X) = \text{softmax}\left(\left[\sum_{k=1}^{K} g_k^c(v_k^{cT} X u_k^c)\right]_{c=1,2,3,\ldots,n}\right) \tag{3}$$

$$\hat{f}(X) = \text{softmax}\left(\left[\sum_{k=1}^{K} g_k^c((v_k^c V)^T D(U u_k^c))\right]_{c=1,2,3,\ldots,n}\right) \tag{4}$$

The learning objective in 2-D case is quite similar to that in 1-D case with incorporation of additional projection vector $\{u^c\}$ in an objective function, which is described in 5.

$$\begin{aligned}(\{g_j^c\},\{v_j^c\},\{u_j^c\}) = \arg\min_{G,\mathbb{R}^{c\times m},\mathbb{R}^{c\times n}} \quad & -\sum_c \left(r_j \log(\hat{f})\right) \\ \text{s.t.} \quad & v_k^{pT} v_k^q = 0 \\ & \forall \quad p \in \{1,2,3,...,n\} \\ & \forall \quad q \in \{1,2,3,...,n\} \setminus p \\ \text{and} \quad & u_k^{pT} u_k^q = 0 \\ & \forall \quad p \in \{1,2,3,...,n\} \\ & \forall \quad q \in \{1,2,3,...,n\} \setminus p\end{aligned} \tag{5}$$

2.2 Meijer G-functions

The Meijer G-functions [14] is a generalized hyperbolic function, which consists of most of the special functions as a particular case. These functions are determined by finitely many indices capturing any special functions in a global sense. G-functions are defined along a path in a line integral $\mathcal{L}$ which can be in an entire complex plane, which is described in equation (6), where $m, n., q, p \in \mathbb{N}$, $a_i, b_j \in \mathbb{R}$, with additional constraints as $m \leq q$, $n \leq p$, and Γ is gamma function. In our experiments, we consider the behavior of real z in domain $(-1, 0) \cup (0, 1)$.

$$G_{p,q}^{m,n}\big(\{a_1,..,a_p\},\{b_1,..,b_q\} \mid z\big) = \frac{1}{2\pi i}\int_{\mathcal{L}} \frac{\prod_{j=1}^{m}\Gamma(b_j - s)\prod_{j=1}^{n}\Gamma(1-a_j+s)}{\prod_{j=m+1}^{q}\Gamma(1-b_j+s)\prod_{j=n+1}^{p}\Gamma(a_j-s)} z^s ds \tag{6}$$

The selection of path $\mathcal{L}$ is task dependent, in this work we directly use the family of functions $\mathbb{G} \in \{\text{id}, \sin, \cos, \sinh, \cosh, \exp, \log, J_v, Y_v, \Gamma\}$ as described in [15].

2.3 Method

Now as we have all the required mathematical tools, the symbolic pursuit model works by considering all $g_1, g_2, ...g_k$ defined in equation (1) and 3 from the set $\mathbb{G}$. As described in 2.2, all the input to function $g(.)$, To constrain the input between $(-1, 0) \cup (0, 1)$, we use similar formulation as described in paper [15], by considering Cauchy-Schwartz inequality, which guarantees that $|v_k^{cT} x| \leq ||v_k^c||||x||$, using this inequality we can easily scale input vectors in a required range as described in (7). Algorithm 1, describes the followed approach in detail.

$$v_k^{cT} x \to \left(\frac{v_k^{cT} x}{||v_k^c||\sqrt{d}}\right) \tag{7}$$

In case of 2-D inputs, the above equation (7) takes a different form (8), as Cauchy-Schwartz inequality changes to $|v_k^{cT} X u_k^c| \leq ||v_k^c||||X||||u_k^c||$, where the norm of a matrix can be derived as maximum Eigen value λ_{max} of a matrix.

$$v_k^{cT} X u_k^c \to \left(\frac{v_k^{cT} X u_k^c}{||v_k^c||\lambda_{max}||u_k^c||}\right) \tag{8}$$

Algorithm 1 Proposed Method for Symbolic Interpretability of Classification Models

Input Black-box function f; Training Set X; Number of classes n
Output Interpretable function $\hat{f}$
Initialize $r_0 \in (0,1)^c$
while $\frac{||r_{k+1}||}{||r_k||} < tolerance$ **do**

$$(\{g_k^c\},\{v_k^c\},\{w_k^c\}) \leftarrow \arg\min_{G,\mathbb{R}^{c\times d},\mathbb{R}^c} \quad \sum_i \left[-\sum_c \left(r_k^c(x_i)\log\left(w^c g_k^c\left(\frac{v_k^{cT}x_i}{||v_k^{cT}||||x_i||}\right)\right)\right)\right];$$

$$\text{s.t.} \quad v_k^{pT} v_k^q = 0$$
$$\forall \quad p \in \{1,2,3,...,n\}$$
$$\forall \quad q \in \{1,2,3,...,n\} \setminus p$$

$$r_{k+1} \leftarrow r_k - \left[w^c g_k^c\left(\frac{v_k^{cT}x_i}{||v_k^{cT}||||x_i||}\right)\right]_{(c=\{1,2,..,n\}),(i=\{1,2,3...N\})};$$

for $l = 1,2,3,...k-1$ **do**
$(m,n,p,q) \leftarrow hyperparameter(g_l)$

$$(\{g_l^c\},\{v_l^c\},\{w_l^c\}) \leftarrow \arg\min_{G_{p,q}^{m,n},\mathbb{R}^{c\times d},\mathbb{R}^c} \quad \sum_i \left[-\sum_c \left(r_{k,l}^c(x_i)\log\left(w^c g^c\left(\frac{v^{cT}x_i}{||v^{cT}||||x_i||}\right)\right)\right)\right];$$

$$\text{s.t.} \quad v_k^{pT} v_k^q = 0$$
$$\forall \quad p \in \{1,2,3,...,n\}$$
$$\forall \quad q \in \{1,2,3,...,n\} \setminus p$$

$$r_{k+1} \leftarrow r_k - \left[w^c g_k^c\left(\frac{v_k^{cT}x_i}{||v_k^{cT}||||x_i||}\right)\right]_{(c=\{1,2,..,n\}),(i=\{1,2,3...N\})};$$

end for
$k \leftarrow k+1$
end While

$$\hat{f}(x) = \left[\sum_k w_k g_k\left(\frac{v_k^{cT}x}{||v_k^{cT}||||x||}\right)\right]_{c=\{1,2,3...,n\}}$$

Table 1: Table shows the result of various methods on synthetic data, Feature Importance columns basically indicates the mean and variance of a ratio of an estimated importance vectors over 50 datapoints

Method	Feature Importance
LIME	-0.016 $\pm$ 0.63
Proposed	0.997 $\pm$ **0.0017**

3 Experiments

3.1 Synthetic Data

In this experiment, we constructed a dummy task of classifying points in and out of a unit circle. Figure 1 describes the data used in this experiment; the red dots indicate class 1 and blue indicates class 0. The data generating process for the same is described in equation (9). We consider this function as an oracle and try to estimate circle coefficients using the proposed method. We consider the variance of the importance vector for 50 test points as a measurable metric to compare the method with other standard techniques. Later we run the same experiment with LIME to compare the performance; these results are described in the table 1.

$$f(x,y) = \begin{cases} 1, & \text{if } (x_0-0.5)^2 + (x_1-0.5)^2 \leq 0.25 \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

In this experiment, after inspecting each vector v_k^c, which easily shows the perpendicularity behavior, at the same time, the components of v_k describe the importance of feature x_0, x_1 in input data. The

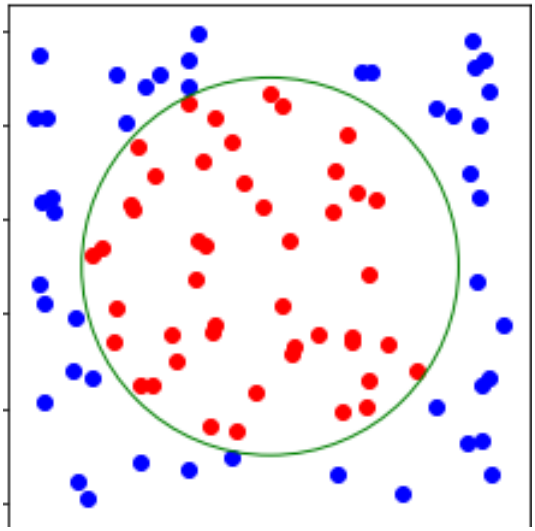

Figure 1: Synthetic Data

Table 2: Table shows the importance vectors for each class

Class	$x_0(sepal \quad length)$	$x_1(sepal \quad width)$	$x_2(petal \quad length)$	$x_3(petal \quad width)$
setosa	0.722	-0.021	0.578	0.378
versicolor	-0.474	0.433	0.337	0.687
virginica	0.319	0.825	-0.394	-0.246

approximate decision boundary obtained by Taylor expansion of the final expression is $-0.22x_0 - 0.53x_1 + 0.27(0.42x_0 + x_1)^2 + 0.64$. The limitation observed was that the operating range of final predictions is kind of skewed between (0.4, 0.6), which needs further exploration.

3.2 Real World Data

We further evaluate the proposed method on multiple real-world datasets, including both 1-D and 2-D input data. For 1-D, we considered 3 UCI datasets [18], namely IRIS, WINE, and Breast-Cancer. In the case of 2-D, we tested the method on MNIST [19] datasets. For evaluation of method on multiple black-box models, we make use of MLP and CNN. MLP and CNNs are implemented using Keras. All the experiments are made opensource on Github at `https://github.com/koriavinash1/Symbolic-Pursuit`.

We split the data into 80-20 splits for training and testing in all the experiments, respectively. In the case of CNN's to train a symbolic interpreter, we used a subset of dataset (as the proposed method is time consuming). To evaluate the method, we compute classification accuracy along with confidence intervals for all the methods. Table 3 describes the performance of all the methods on various datasets, the confidence intervals described in the table are the result of the bootstrapping method, rather than running multiple experiments.

As described previously, the importance vectors for all four features in the IRIS dataset are described in table 2. This suggests that the major contributing factors to classify a data point in class *setosa* are *sepal length*, *petal length*, and *petal width*. Similarly, to classify the datapoint in class *versicolor* features required are *sepal width*, *petal length*, and *petal width*, similar coefficient trends can be observed by fitting logistic regression.

Table 3: Table shows the results for various methods on various datasets, by fitting a single term interpreter model due to time constrain.

Models	Datasets	Black-box Acc	Symbolic Vs Black-box	Symbolic Acc
MLP	IRIS	0.956 ± 0.054	0.941 ± 0.014	0.933 ± 0.021
	WINE	0.963 ± 0.041	0.651 ± 0.023	0.752 ± 0.031
	Breast-Cancer	0.971 ± 0.012	0.683 ± 0.084	0.681 ± 0.077
CNN	MNIST	0.984 ± 0.017	0.553 ± 0.011	0.551 ± 0.014

4 Conclusion and Future work

The current work is a preliminary analysis of symbolic models for classification tasks. The current approach is hardly scalable to larger datasets, which is crucial to make generalized claims on model behavior. One way to explore along these lines would be to formulate this optimization as a mini-batch problem. Another interesting area of exploration would be formulating margin-based decision boundaries to address the operating range problem discussed in section 3.1. In the case of projection functions, the optimal operating range needs to be experimented with, either by considering different transformation functions or consider a different operating range for Meijer G-function.

In conclusion, this work proposes an algorithm for a symbolic interpretation of classification models in both 1-D and 2-D data with their respective projection functions. The final symbolic expression is the sum of multiple Meijer G-functions, rather than limiting the function scope to trigonometric or spline functions. This work shows the preliminary results on multiple classification datasets and lists the possible loopholes and the ways to improve in future approaches. Apart from obtaining a symbolic function for any given black box, it's essential to ensure that these functions provide meaningful representations. For example, in the medical domain, interpretability should provide clear representations that can be verified by domain experts, which also will be explored in future work.

References

[1] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. " hello ai": Uncovering the onboarding needs of medical practitioners for human-ai collaborative decision-making. *Proceedings of the ACM on Human-computer Interaction*, 3(CSCW):1–24, 2019.

[2] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[3] Berk Ustun and Cynthia Rudin. Methods and models for interpretable linear classification. *arXiv preprint arXiv:1405.4047*, 2014.

[4] Aravindh Mahendran and Andrea Vedaldi. Visualizing deep convolutional neural networks using natural pre-images. *International Journal of Computer Vision*, 120(3):233–255, 2016.

[5] Andreas Holzinger, Markus Plass, Katharina Holzinger, Gloria Cerasela Crisan, Camelia-M Pintea, and Vasile Palade. A glass-box interactive machine learning approach for solving np-hard problems with the human-in-the-loop. *arXiv preprint arXiv:1708.01104*, 2017.

[6] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1721–1730, 2015.

[7] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[8] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2017. https://distill.pub/2017/feature-visualization.

[9] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144, 2016.

[10] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, pages 4765–4774, 2017.

[11] Patryk Orzechowski, William La Cava, and Jason H Moore. Where are we now? a large benchmark study of recent symbolic regression methods. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 1183–1190, 2018.

[12] Telmo Menezes and Camille Roth. Symbolic regression of generative network models. *Scientific reports*, 4:6284, 2014.

[13] Ahmed M Alaa and Mihaela van der Schaar. Demystifying black-box models with symbolic metamodels. In *Advances in Neural Information Processing Systems*, pages 11304–11314, 2019.

[14] Richard Beals and Jacek Szmigielski. Meijer g-functions: a gentle introduction. *Notices of the AMS*, 60(7):866–872, 2013.

[15] Jonathan Crabbe, Yao Zhang, William Zame, and Mihaela van der Schaar. Learning outside the black-box: The pursuit of interpretable models. *Advances in Neural Information Processing Systems*, 33, 2020.

[16] Jerome H Friedman and Werner Stuetzle. Projection pursuit regression. *Journal of the American statistical Association*, 76(376):817–823, 1981.

[17] Yoon Dong Lee, Dianne Cook, Ji-won Park, Eun-Kyung Lee, et al. Pptree: Projection pursuit classification tree. *Electronic Journal of Statistics*, 7:1369–1386, 2013.

[18] Dheeru Dua and Casey Graff. Uci machine learning repository. 2017.

[19] Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010.