Can Local Additive Explanations explain Linear Additive Models?

Amir Rahnama[1], Judith Bütepage[1], Pierre Geurts[2], Henrik Boström[1]

1: KTH Royal Institute of Technology, Sweden, 2: University of Liege, Belgium

Introduction

- Local additive model-agnostic explanation techniques decompose the predicted output to an additive sum of feature importance scores [5].
- These techniques can provide inaccurate explanations. Some argue against using them in high-stake decision-making domain domains [4].
- The real problem is that *we do not have reliable evaluation measures.*
- Why? Because extracting ground truth importance scores from black-boxes is challenging.

Evaluation Methods for Local Explanations

- **Robustness measures** [3]
  - Limitation: The model can provide wrongful predictions.
- **Ground Truth from synthetic datasets** [2]**:**
  - Limitation: Explanations are not supposed to be faithful to the data
- **Ground truth from Interpretable Models** [1] is arguably most principled way to evaluate local explanations, but there is a problem ...
  - We do not have good baselines in this category of methods.

Model Intrinsic Additive Scores

We *propose* **MIAS** to extract ground truth from interpretable model. Here is how it works:

1. We extracts additive scores from linear additive prediction functions.
2. We directly measure the similarity of the additive score to the feature importance score from the explanation using Spearman's rank correlation.

- **MIAS** can be used to evaluate local explanations of:
  - Linear and Logistic Regression
  - Gaussian Naive Bayes
- Example (Logistic Regression):

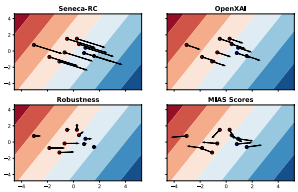
$$\log \frac{P(y_n = c|x_n, w)}{P(y_n = \neg c|x_n, w)} = \sum_{m=0}^{M} w^m x_n^m \quad (1)$$

$$f(x) \leftarrow \log \frac{P(y_n = c|x_n, w)}{P(y_n = \neg c|x_n, w)} \quad (2)$$

$$\lambda_n^m \leftarrow w^m x_n^m \quad (3)$$

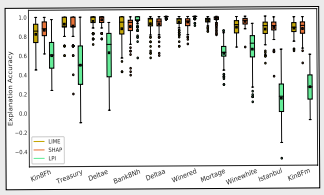
$$\text{Accuracy of Explanations} = \rho(\lambda_n, \phi_n) \quad (4)$$

Motivating Example

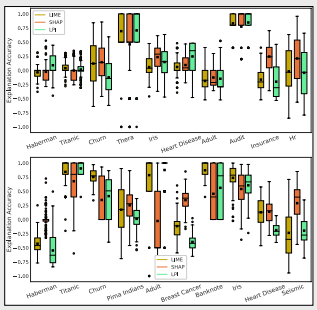


Experiments:

- The explanation accuracy of LIME, SHAP and Local Permutation Importance (LPI) on 20 datasets:



Linear Regression



Logistic Regression (Top), Naive Bayes (Bottom)

Key Findings:

- LIME and SHAP are relatively accurate when explaining Linear Regression, but they are significantly inaccurate when explaining Logistic Regression and Naiva Bayes models in many datasets.
- Using **MIAS**, we measure the effect of seven factors on local explanation accuracy:
  - Model generalisation (model related)
  - Number of numerical, categorical, correlated features and the preprocessing technique (data related)
  - The sample size of LIME and SHAP (related to the explanation techniques)
  - Similarity metric (evaluation method related)

Reference:

1. Agarwal, Chirag, et al. "Openxai: Towards a transparent evaluation of model explanations." *Advances in Neural Information Processing Systems* 35 (2022): 15784-15799.
2. Guidotti, Riccardo. "Evaluating local explanation methods on ground truth." *Artificial Intelligence* 291 (2021): 103428.
3. Fong, Ruth C., and Andrea Vedaldi. "Interpretable explanations of black boxes by meaningful perturbation." *Proceedings of the IEEE international conference on computer vision*. 2017.
4. Rudin, Cynthia. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead." *Nature machine intelligence* 1.5 (2019): 206-215
5. Guidotti, Riccardo, et al. "A survey of methods for explaining black box models." *ACM computing surveys (CSUR)* 51.5 (2018): 1-42.