Rensselaer

why not change the world?®

Explaining the Predictions of Any Classifier

Lucky Yerimah | Chemical Engineering

April 26, 2022

Presentation Outline

- Introduction
- Problem motivation and theory.
- Learning algorithm
- Experiments
- Results
- Some limitations and future directions.
- Conclusion
- References.

Introduction

- Understanding how and why a machine learning model makes predictions can be as important as the prediction itself.
- Interpreting a prediction model's output:
  - Engenders appropriate user trust
  - Provides insight into how the model may be improved
  - Reinforces the understanding of the process being modeled.
- Various methods have been proposed to interpret a model's prediction including:
  - LIME: The use of locally faithful linear models
  - SHAP: The use of Sharpley values for additive feature explanation
  - ILIME: Improves on the performance of LIME for explaining Gaussian process models.

Problem Motivation and Theory

- We are interested in investigating the Gaussian process constrained linear model for both local and possibly global interpretation of machine learning models.
- We modify the original LIME algorithm by including an additional constraint on the weights to be Gaussian distributed.

$x \in \mathbb{R}^d$ denotes the original representation while $x' \in \{0,1\}^{d'}$ denotes the interpretable representation.

Let an explanation model $g \in \mathbb{G}$. Where $\mathbb{G}$ is a class of potential interpretable models. The domain of $g(z)$ is $\{0,1\}^{d'}$

$$J(\mathbf{w}, \Delta) = \mathcal{L}(f, g, \pi_x) + \lambda||\mathbf{w}||_1 + \gamma \sum_{i=1}^{d} \log p(w_i|0,\ ZDZ^T)$$

$\mathcal{L}(f, g, \pi_x)$ is a measure of how unfaithful g is in approximating f in the locality defined by π_x, $||\boldsymbol{w}||_1$ enforces sparsity on $\boldsymbol{w}$ and the last term on the right represents the additional Gaussian distribution constraint.

λ and γ are tunable hyperparameters to control each constraint and $D = diag(\Delta^2) \in \mathbb{R}^{d \times d}$

- We hope that the additional constraint will produce better explanation for prediction instances.

Learning Algorithm

Algorithm 1 : GPLIME Training Algorithm

Require: Classifier f, Number of samples N, Epoch
Require: Instance x and its interpretable version x'
Require: Explainer g, Similarity kernel π_x, Length of explanation K

$\mathcal{Z} \leftarrow \{\}$

for $i \in \{1, 2, 3, \dots, N\}$ **do**

$z'_i \leftarrow$ sample around (x')

$\mathcal{Z} \leftarrow \mathcal{Z} \cup \langle z'_i, f(z_i), \pi_x(z_i) \rangle$,

$\hat{y} = g(z_i, \theta)$, $y = f(z_i)$

end for

for i in range Epoch

$\mathcal{L}(\hat{\mathbf{y}}, \mathbf{y}, \theta) = \frac{1}{2N} \sum_{i=1}^{N} \|\hat{\mathbf{y}}^{(i)} - \mathbf{y}^{(i)}\|_2^2 + \lambda|\theta| - \frac{\gamma}{2}\left[d \log \det(zDz^T) - \sum_{i=1}^{d} c_i^T (zDz^T)^{-1} c_i\right]$

$\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}(\hat{\mathbf{y}}, \mathbf{y}, \theta)$

end for

Select K features from θ

$w \leftarrow$ K-Ridge $(\mathcal{Z}, K) \triangleright$ with z'_i as features, $f(z)$ as target

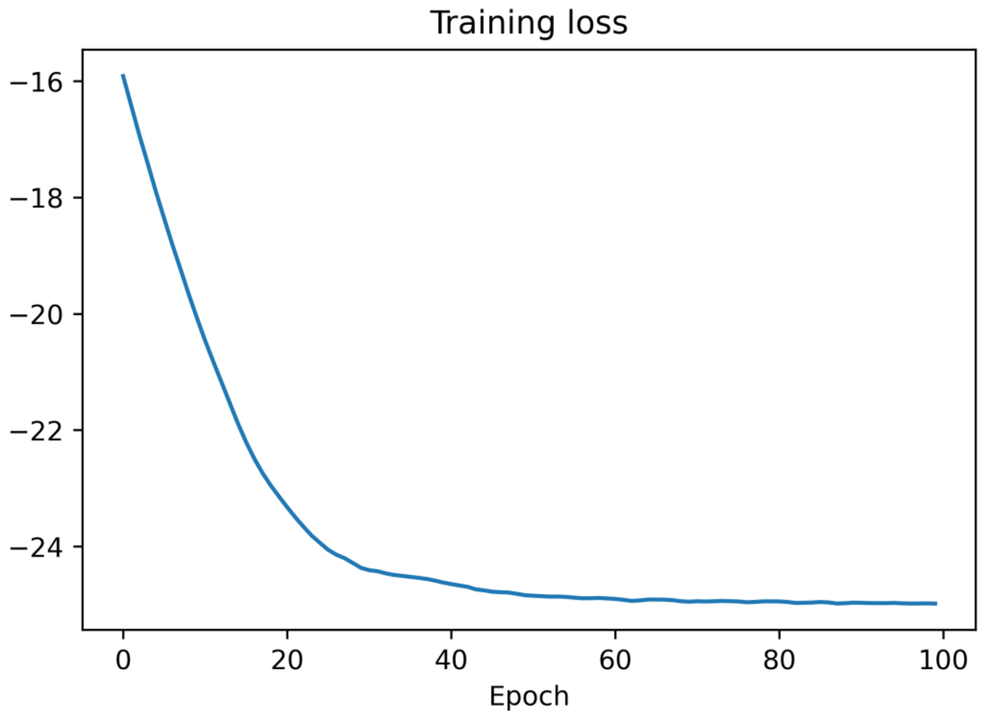
return w

$$g(z') = w_g \cdot z'$$
$$\pi_x(z) = \exp(-M(x, z)^2/\sigma^2$$

Where M is the distance function with width σ

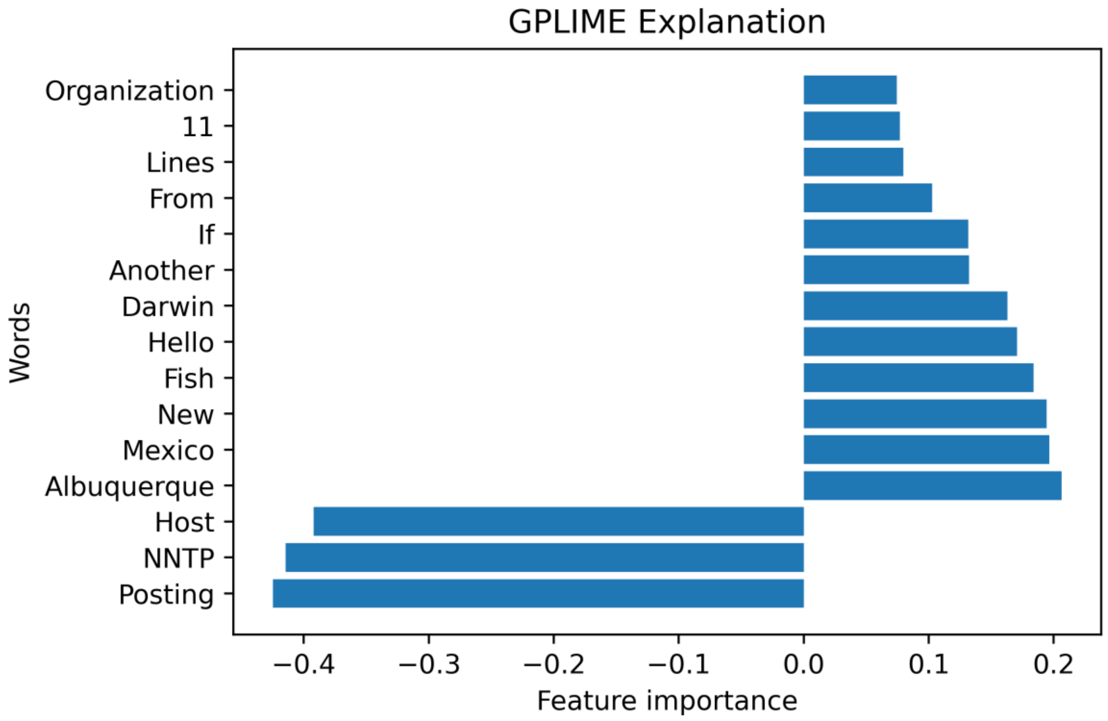
Experiments

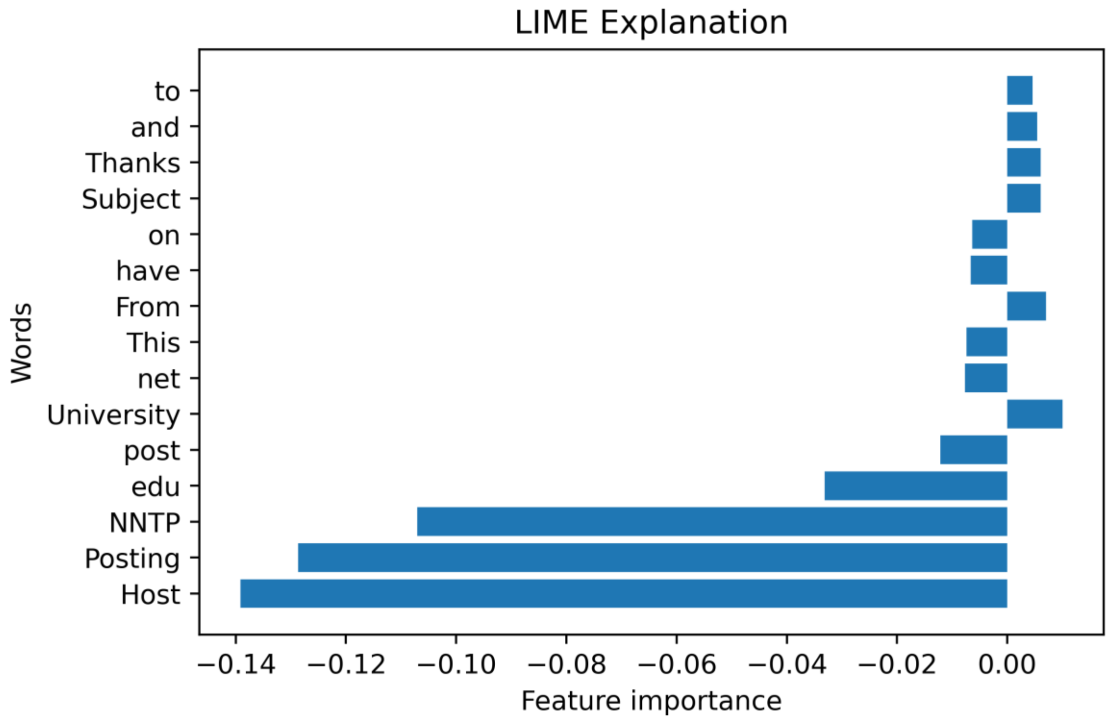
- Experiments are designed to explain prediction instances of a random forest classifier.
- LIME is used the baseline for comparison
- The 20 newsgroups text dataset with two classes (Christianity and atheism) is used.



Results

- True class: Atheism
- Predicted class: Atheism

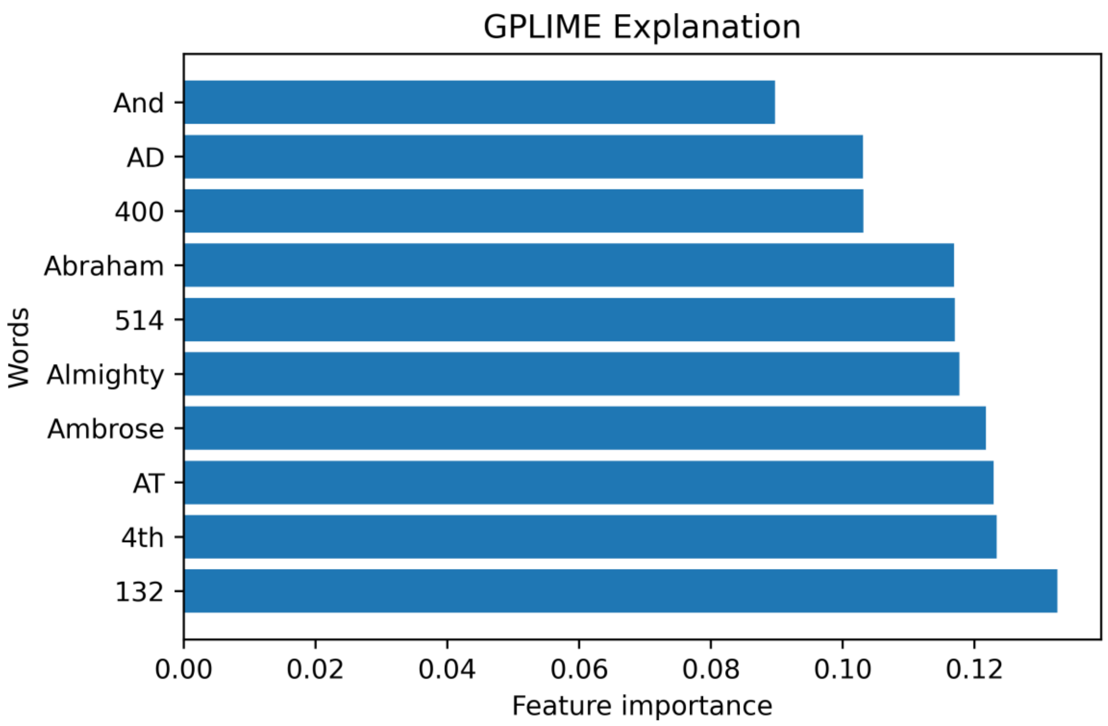


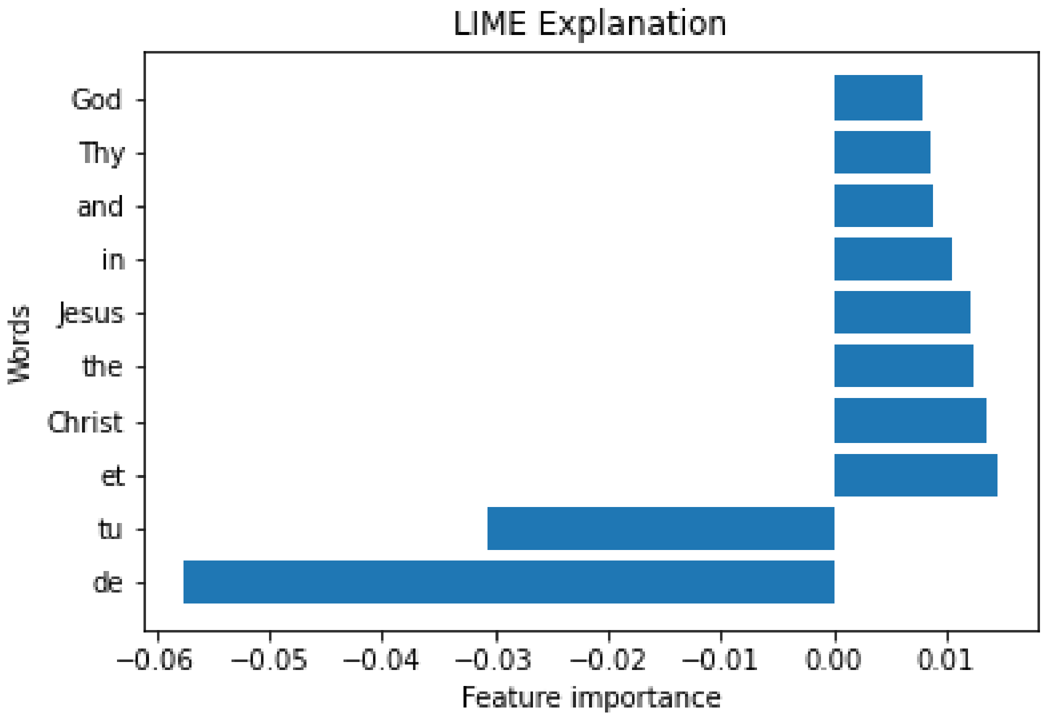


- “NNTP”, “Posting” and “Host” are the most importance features for this prediction instance.

Results

- True class: Christianity
- Predicted class: Christianity

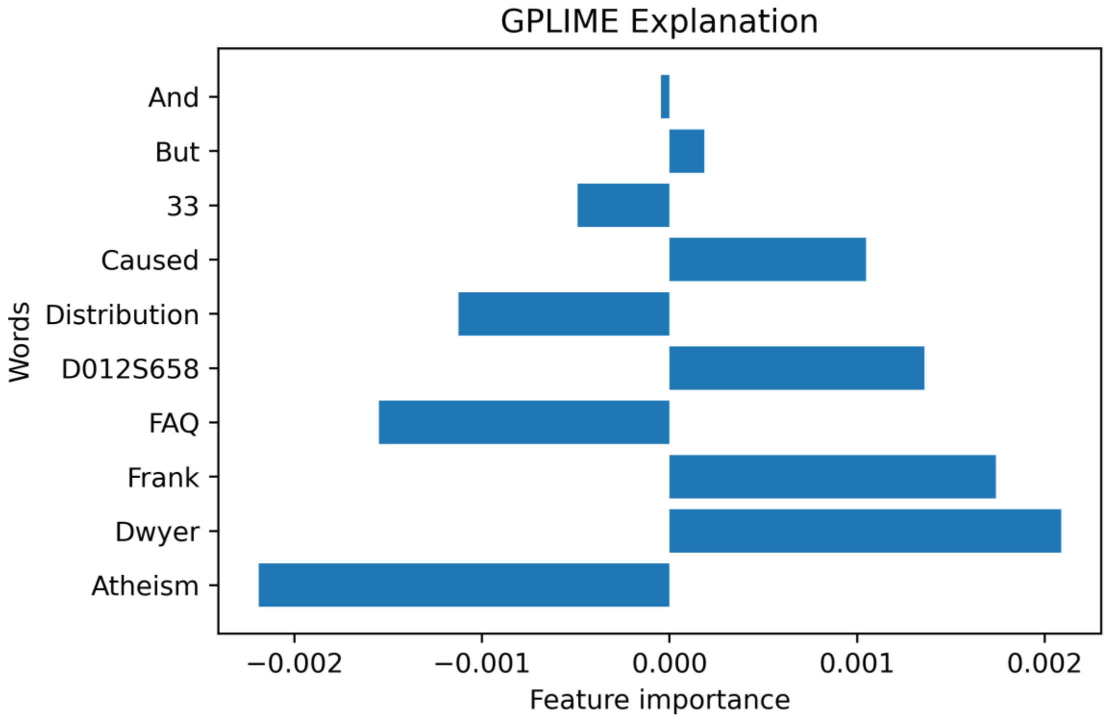


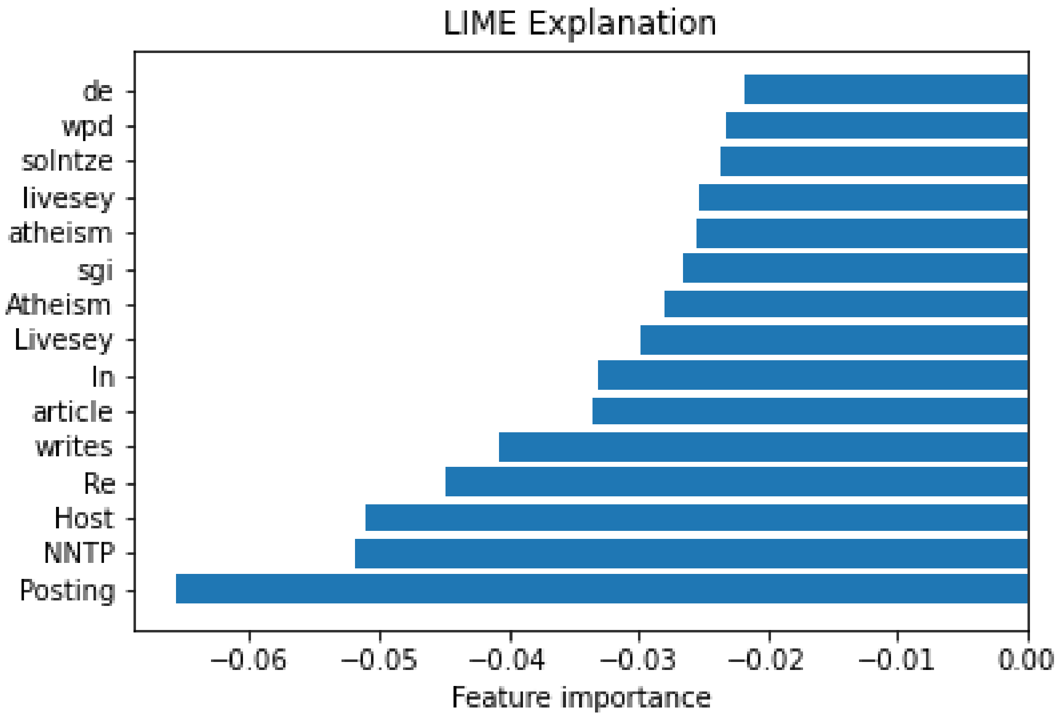


- Both explainers give different feature importance that raises some doubt about the trustworthiness of the explainers themselves.

Results

- True class: Atheism
- Predicted class: Atheism





Some limitations and future directions

- We noticed both models giving slightly different explanations for the same prediction instance.
- The differences in explanations raises concerns about trustworthiness of the explainers.
- Although LIME has been demonstrated for being faithful to a classifier, GPLIME showed better explanation of some instances.
- Faithfulness to a classifier can be investigated for GPLIME as a future direction.
- LIME and GPLIME are only locally faithful. How can we account for the classifier globally?
- Exploring other families of explanation models such as decision trees are possible future directions.

Conclusion

- We showed that GPLIME can produce explanations that are closely similar to LIME.
- GPLIME did not show clear superior performance compared to LIME hence we cannot conclude if the additional constraint on the weights have been helpful.
- Implementation codes and other results can be found on this repository: https://github.com/Eshemomoh/Trustworthy-ML-Project

References

- M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?' Explaining the Predictions of Any Classifier," NAACL-HLT 2016 - 2016 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. Proc. Demonstr. Sess., pp. 97–101, 2016, doi: 10.18653/v1/n16-3020.
- S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," Adv. Neural Inf. Process. Syst., vol. 2017-December, no. Section 2, pp. 4766–4775, 2017.
- N. Puri and P. Gupta, "MAGIX: Model Agnostic Globally Interpretable Explanations," 2017, [Online]. Available: https://arxiv.org/abs/1706.07160.
- Y. Yoshikawa and T. Iwata, "Gaussian Process Regression with Local Explanation," 2020, [Online]. Available: http://arxiv.org/abs/2007.01669.
- R. ElShawi, Y. Sherif, M. Al-Mallah, S. Sakr, "ILIME: Local and global interpretable model-agnostic explainer of black-box decision". In: Advances in Databases and Information Systems. ADBIS 2019. Lecture Notes in Computer Science, vol 11695, Springer.

Thank You