Regularizing Black-box Models for Improved Interpretability

Gregory Plumb, Maruan Al-Shedivat, Eric Xing, and Ameet Talwalkar

CMU

Summary

- Most work on interpretable machine learning has focused on designing either inherently interpretable models, which typically trade-off accuracy for interpretability, or post-hoc explanation systems, which lack guarantees about their explanation quality.
- We propose an alternative to these approaches by directly regularizing a black-box model for interpretability at training time.
- By doing this, we find that we can substantially improve the explanation fidelity and stability metrics while slightly improving accuracy.

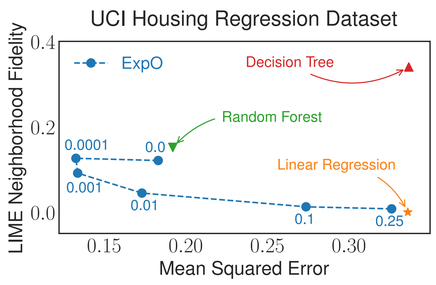

Figure 1: Neighborhood Fidelity of LIME-generated explanations (lower is better) vs. predictive error of several models. The values in blue denote regularization weight.

Standard Approaches for Producing Explanations

- **by-design** approaches modify the model structure to produce both predictions and explanations, but often sacrifice accuracy.
- **post-hoc** approaches produce explanations for a black-box pre-trained model, but cannot guarantee explanation quality.

We blend these two approaches, by regularizing a black-box model at training time to improve its post-hoc interpretability.

Why Local Explanations?

- They answer the question "What could I have done differently to get the desired result?"
- They can be quite flexible because we can choose what "local" means.

Local Explanations with Semantic Features

- Semantic features have an inherent meaning.
- Local explanations approximate the model, f, across some neighborhood, N_x, around a point, x, with an interpretable function, g.
- **Neighborhood Fidelity (NF) Metric**[3, 2]:

$$F(f, g, N_x) := \mathbb{E}_{x' \sim N_x}[(g(x') - f(x'))^2] \quad (1)$$

Local Explanations with Non-Semantic Features

- Here, we cannot assign meaning to the difference between x and x', so it does not make sense to explain the difference between $f(x)$ and $f(x')$.
- Local explanations identify which parts of the input are influential on a prediction [5]
- **Stability (S) Metric** [1]:

$$S(f, e, N_x) := \mathbb{E}_{x' \sim N_x}[||e(x, f) - e(x', f)||_2^2] \quad (2)$$

Regularizers

- Computing Eq. 1 or Eq. 2 is too expensive.
- Approximate them with either Alg. 1 or Alg. 2.

Algorithm 1 ExpO-Fidelity Regularizer

input f_θ, x, N_x^{reg}, m

1: Sample points: $x'_1, \ldots, x'_m \sim N_x^{\text{reg}}$

2: Compute predictions:

$$\hat{y}_j(\theta) = f_\theta(x'_j) \text{ for } j = 1, \ldots, m$$

3: Produce a local linear explanation:

$$\beta_x(\theta) = \arg\min_\beta \sum_{j=1}^m (\hat{y}_j(\theta) - \beta^\top x'_j)^2$$

output $\frac{1}{m}\sum_{j=1}^m (\hat{y}_j(\theta) - \beta_x(\theta)^\top x'_j)^2$

Algorithm 2 ExpO-Stability Regularizer

input f_θ, x, N_x^{reg}, m

1: Sample points: $x'_1, \ldots, x'_m \sim N_x^{\text{reg}}$

2: Compute predictions:

$$\hat{y}_j(\theta) = f_\theta(x'_j), \text{ for } j = 1, \ldots, m$$

output $\frac{1}{m}\sum_{j=1}^m (\hat{y}_j(\theta) - f_\theta(x))^2$

Experiments with Semantic Features

- We test ExpO-Fidelity on seven regression problems from the UCI collection as well as SUPPORT2, an in-hospital mortality classification problem.
- We use LIME [3] and MAPLE [2] as explainers.
- We set $N_x = N_x^{reg} = \mathcal{N}(x, 0.1)$ and standardize all features.

Across these datasets, ExpO improved the interpretability metrics by at least 25% and had a small positive impact on accuracy. A sample of the results are in Table 1.

Table 1: Unregularized model vs. the same model trained with ExpO-Fidelity on two of the UCI regression problems. Results are shown across 10 trials (standard error is in parenthesis).

Dataset	Regularizer	MSE	LIME-NF	LIME-S	MAPLE-NF	MAPLE-S
autompgs	None	0.14 (0.03)	0.041 (0.012)	0.0011 (0.0006)	0.0160 (0.0088)	0.0150 (0.0099)
	Fidelity	0.13 (0.02)	**0.011 (0.003)**	**0.0001 (0.0003)**	**0.0014 (0.0006)**	**0.0017 (0.0005)**
communities	None	0.49 (0.05)	0.110 (0.012)	0.022 (0.003)	0.16 (0.02)	1.2 (0.2)
	Fidelity	**0.46 (0.03)**	**0.079 (0.007)**	**0.005 (0.001)**	**0.13 (0.01)**	**0.8 (0.4)**

Experiments with Non-Semantic Features

- We test ExpO-Stability by creating saliency maps [4] on MNIST.
- **The stability metric decreased from 6.94 to 0.0008 with** $N_x = Uniform[x - 0.05, x + 0.05]$
- The saliency maps now focus on the presence or absence of certain pen strokes (Figure 2).

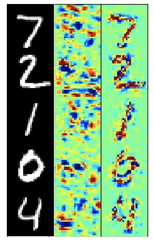

Figure 2: Original MNIST images (left) and saliency maps of an unregularized (middle) and regularized (right) models.

What Else is in the Paper?

- A comparison of sample-based local interpretability to classic approaches such as Taylor Approximations and the Lipschitz Constant
- A discussion of how N_x, N_x^{reg}, and the explainer's definition of "local" are connected.
- Theory demonstrating that the effects of this regularization generalize to unseen points.

References

[1] D. Alvarez-Melis and T. Jaakkola.
Towards robust interpretability with self-explaining neural networks.
In *Advances in Neural Information Processing Systems*, pages 7785–7794, 2018.

[2] G. Plumb, D. Molitor, and A. S. Talwalkar.
Model agnostic supervised local explanations.
In *Advances in Neural Information Processing Systems*, pages 2516–2525, 2018.

[3] M. T. Ribeiro, S. Singh, and C. 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*, pages 1135–1144. ACM, 2016.

[4] K. Simonyan, A. Vedaldi, and A. Zisserman.
Deep inside convolutional networks: Visualising image classification models and saliency maps.
arXiv preprint arXiv:1312.6034, 2013.

[5] M. Sundararajan, A. Taly, and Q. Yan.
Axiomatic attribution for deep networks.
arXiv preprint arXiv:1703.01365, 2017.