Robust and Stable Black Box Explanations

Hima Lakkaraju
Harvard University

Nino Arsov
Macedonian Academy of Arts & Sciences

Osbert Bastani
University of Pennsylvania

Motivation

- ML models are increasingly proprietary and complex, and are therefore not interpretable

- Several post hoc explanation techniques proposed in recent literature
  - E.g., LIME, SHAP, MUSE, Anchors, MAPLE

Motivation

- However, post hoc explanations have been shown to be unstable and unreliable
  - Small perturbations to input can substantially change the explanations; running same algorithm multiple times results in different explanations (Ghorbani et. al.)
  - High-fidelity explanations with very different covariates than black box (Lakkaraju & Bastani)
  - Also, they are not robust to distribution shifts

Why can explanations be unstable?

- Distribution $p(x_1, x_2)$ where x_1 and x_2 are perfectly correlated
- Blackbox $B^*(x_1, x_2) = \mathrm{I}(x_1 \geq 0)$
- Explanation $\hat{E}(x_1, x_2) = \mathrm{I}(x_2 \geq 0)$
- $\hat{E}$ has perfect fidelity, but is completely different from B^*!
  - If $p(x_1, x_2)$ shifts, $\hat{E}$ may no longer have high fidelity

Why do we care?

- Domain experts rely on explanations to validate properties of the black box model
  - Check if model uses spurious or sensitive attributes [Caruana 2015, Bastani 2017, Rudin 2019]

- Poor explanations may mislead experts into drawing incorrect conclusions

Our Contributions: ROPE

- We propose ROPE (RObust Post hoc Explanations)
  - Framework for generating stable and robust explanations
  - It is flexible, e.g., it can be instantiated for local vs. global explanations as well as linear vs. rule based explanations
  - First approach to generating explanations robust to distribution shifts
  - Our experiments show that ROPE significantly improves robustness on real-world distribution shifts

Robust Learning Objective

- ROPE ensures robustness via a minimax objective:

$$\hat{E} = \arg\min_{E \in \mathcal{E}} \max_{\delta \in \Delta} \mathbb{E}_{p_\delta(x)}[\ell(E(x), B^*(x))].$$

worst-case over distribution shifts

standard supervised learning loss for $p_\delta(x)$

- The maximum in the objective is over possible distribution shifts $p_\delta(x) = p(x - \delta)$

- Ensures $\hat{E}$ has high fidelity for all distributions $p_\delta(x)$

Robust Learning Objective

- We can upper bound the objective as follows:

$$\max_{\delta \in \Delta} \mathbb{E}_{p_\delta(x)}[\ell(E(x), B^*(x))]$$
$$\leq \mathbb{E}_{p(x)} \left[\max_{\delta \in \Delta} \ell(E(x+\delta), B^*(x+\delta)) \right].$$

- Thus, we can approximate $\hat{E}$ as follows:

$$\hat{E} = \arg\min_{E \in \mathcal{E}} \mathbb{E}_{p(x)} \left[\max_{\delta \in \Delta} \ell(E(x+\delta), B^*(x+\delta)) \right].$$

Class of Distribution Shifts

- **Key question:** How to choose Δ?
  - Determines distributions p_δ to which $\hat{E}$ is robust

- **Our choice**

$$\Delta(s_0, \delta_{\max}) = \{\delta \in \mathbb{R}^n \mid \|\delta\|_1 \leq s_0 \wedge \|\delta\|_\infty \leq \delta_{\max}\}$$

  - L_1 constraint induces sparsity, i.e., only a few covariates are perturbed
  - L_∞ constraint bounds the magnitude of the perturbation, i.e., covariates do not change too much

Robust Linear Explanations

- Use adversarial training, i.e., approximate stochastic gradient descent on the objective

$$\nabla_\theta \max_{\delta \in \Delta} \ell(E_\theta(x+\delta), B^*(x+\delta)) \approx \nabla_\theta \ell(E_\theta(x+\delta^*), B^*(x+\delta^*)),$$

where $\delta^* = \arg\max_{\delta \in \Delta} \ell(E_\theta(x+\delta), B^*(x+\delta)).$

- Can approximate δ^* using a linear program

Robust Rule Based Explanations

- Approximate the objective using sampling

$$\hat{E} = \arg\min_{E \in \mathcal{E}} \mathbb{E}_{p(x)} \left[\max_{\delta^j \sim p_0(\delta)} \ell(E(x + \delta^j), B^*(x + \delta^j)) \right].$$

Distribution over shifts $\delta \in \Delta$

- Adjust learning algorithm to handle maximum over finite set
  - For rule lists and decision sets, only count a point $\left(x, \hat{E}(x)\right)$ as correct if $\hat{E}(x) = B^*\left(x + \delta^j\right)$ for all of the possible perturbations δ^j

Experimental Evaluation

- Real-world distribution shifts

Dataset	# of Cases	Attributes	Outcomes
Bail	31K defendants (2 courts)	Criminal History, Demographic Attributes, Current Offenses	Bail (Yes/No)
Healthcare	22K patients (2 hospitals)	Symptoms, Demographic Attributes, Current & Past Conditions	Diabetes (Yes/No)
Academic	19K students (2 schools)	Grades, Absence Rates, Suspensions, Tardiness Scores	Graduated High School on Time (Yes/No)

- Approach
  - Generate explanation on one distribution (e.g., first court)
  - Evaluate fidelity on shifted distribution (e.g., second court)

Experimental Evaluation

- Baselines
  - LIME, SHAP, MUSE
  - All state-of-the-art post hoc explanation tools

- Instantiations of ROPE
  - Linear models (comparison to LIME and SHAP)
  - Decision sets (comparison to MUSE)
  - Focus on global explanations

Robustness to Real Distribution Shifts

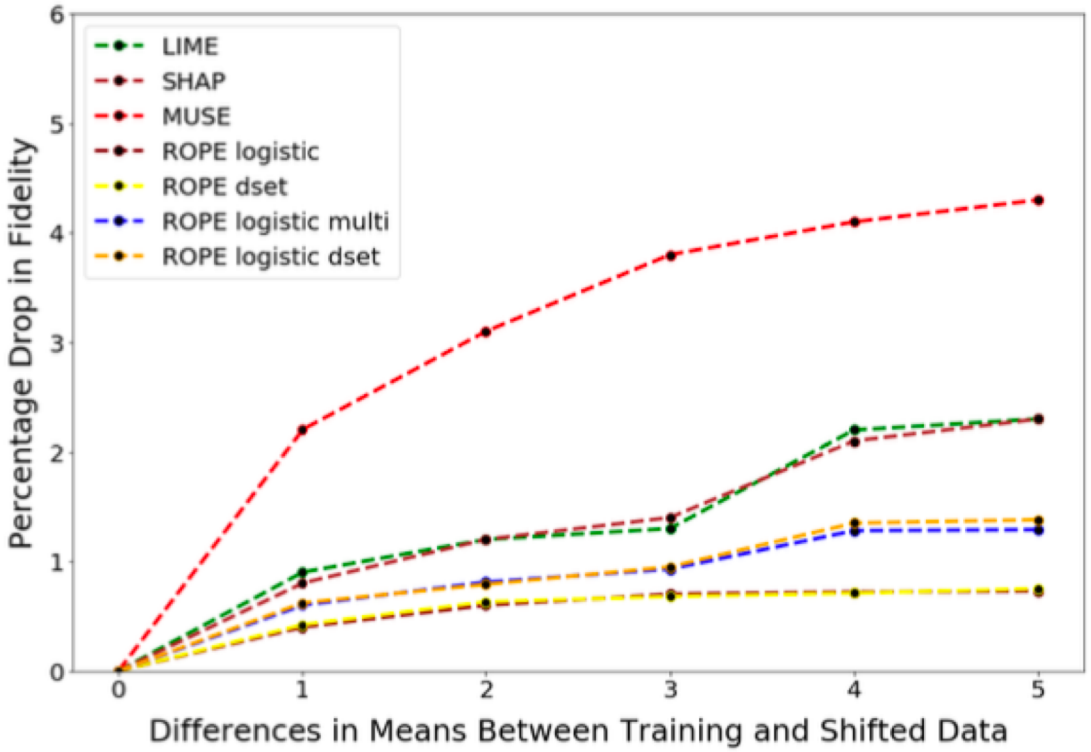
- Report fidelity on both original and shifted distributions, as well as percentage drop in fidelity

Algorithms	Bail			Academic			Health		
	Train	Shift	% Drop	Train	Shift	% Drop	Train	Shift	% Drop
LIME	0.79	0.64	18.99%	0.68	0.57	16.18%	0.81	0.69	14.81%
SHAP	0.76	0.66	13.16%	0.67	0.59	11.94%	0.83	0.68	18.07%
MUSE	0.75	0.59	21.33%	0.66	0.51	22.73%	0.79	0.61	22.78%
ROPE logistic	0.79	0.74	6.33%	0.70	0.69	1.43%	0.82	0.76	7.32%
ROPE dset	**0.82**	**0.77**	**6.1%**	**0.73**	**0.71**	**2.74%**	**0.84**	**0.78**	**7.14%**

- ROPE is substantially more robust without sacrificing fidelity on original distribution

Percentage Drop in Fidelity vs. Size of Distribution Shift

- Use synthetic data and vary size of shift
- Report percentage drop in fidelity



Structural Match with the Black Box

- Choose “black box” from the same model class as explanation (e.g., linear or decision set)
- Report match between explanation and black box

Algorithms	**Black Boxes**					
	LR	**Multiple LR**	**DS**		**Multiple DS**	
	Coefficient Mismatch	Coefficient Mismatch	Rule Match	Feature Match	Rule Match	Feature Match
LIME	4.37	5.01	–	–	–	–
SHAP	4.28	4.96	–	–	–	–
MUSE	–	–	4.39	11.81	4.42	9.23
ROPE logistic	**2.70**	**2.93**	–	–	–	–
ROPE dset	–	–	**6.25**	**16.18**	**7.09**	**16.78**

- ROPE explanations match black box substantially better

Conclusions

- We have proposed the first framework for generating stable and robust explanations
- Our approach significantly improves explanation robustness to real-world distribution shifts