Analyzing Explainer Robustness via Probabilistic Lipschitzness of Prediction Functions

Zulqarnain Khan*, Davin Hill*, Aria Masoomi, Josh Bone, Jennifer Dy

Department of Electrical and Computer Engineering, Northeastern University, Boston, MA

Northeastern University

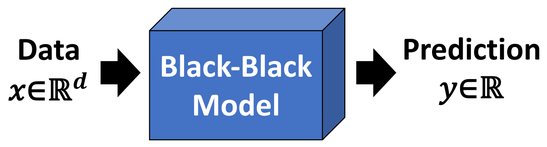
The Institute for Experiential AI
Northeastern University

TL; DR

- We formalize the **explainer astuteness** metric, which captures how well a given explainer provides similar explanations to similar points.
- We prove theorems which imply **locally smooth prediction functions lend themselves to locally robust explanations**.

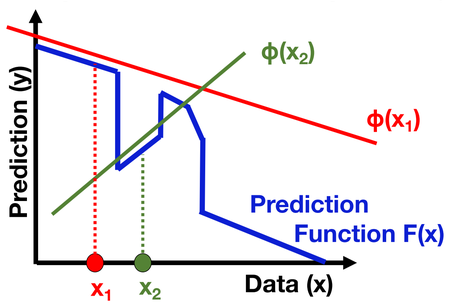
Background and Motivation

- Local attribution explainers generate feature importance values for a given data sample and black-box model.



- Removal-Based Explainers [4] are a class of explainer that evaluate change in model prediction when "removing" features

Example: Local Linear Explainer ϕ

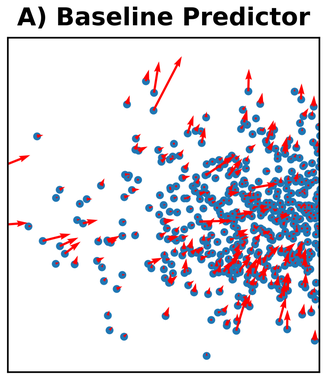


Explainer Robustness:

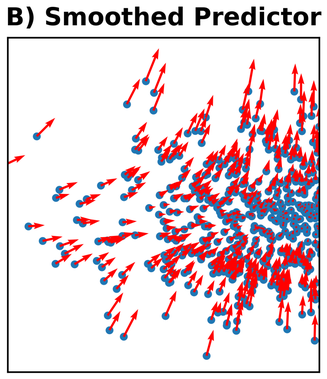
Similar data samples x,x' should have similar explanations $\phi(x)$,$\phi(x')$

- We quantify explainer robustness and establish a theoretical connection with black-box model smoothness.

A) Baseline Predictor



B) Smoothed Predictor

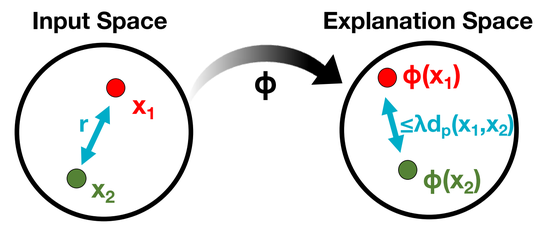


Theoretical Results

Astuteness: A Metric for Explainer Robustness

Intuition: The probability of two data samples x,x' within distance r of each other will also have explanations that are within some bounded distance of each other

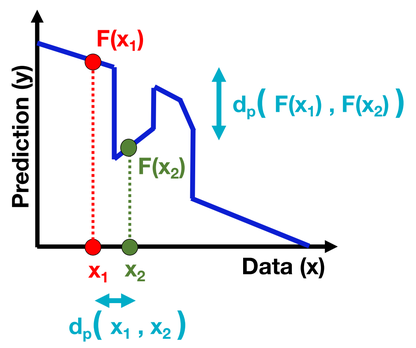
$$A_{r,\lambda}(E,\mathcal{D}) = \mathbb{P}_{x,x' \sim \mathcal{D}}[d_p(\phi(x),\phi(x')) \leq \lambda \cdot d_p(x,x') \mid d_p(x,x') \leq r]$$



Astuteness and Model Smoothness

Definition: Probabilistic Lipschitzness [2]

Given $\alpha \in [0,1], r \geq 0$, a function $f: \mathbb{R}^d \to \mathbb{R}$ is probabilistic Lipschitz with constant $L \geq 0$ if $P_{x,x' \sim \mathcal{D}}[d_p(f(x),f(x')) \leq L d_p(x,x') \mid d_p(x,x') \leq r] \geq 1 - \alpha$



Theorems: Lower Bounds for Explainer Astuteness

Consider a given $r \geq 0$ and $\alpha \in [0,1]$ and a trained predictive function that is probabilistic Lipschitz with constant L, radius r measured with and $d_p(\cdot,\cdot)$ and with probability at least $1 - \alpha$.

Theorem 1: SHAP [3]

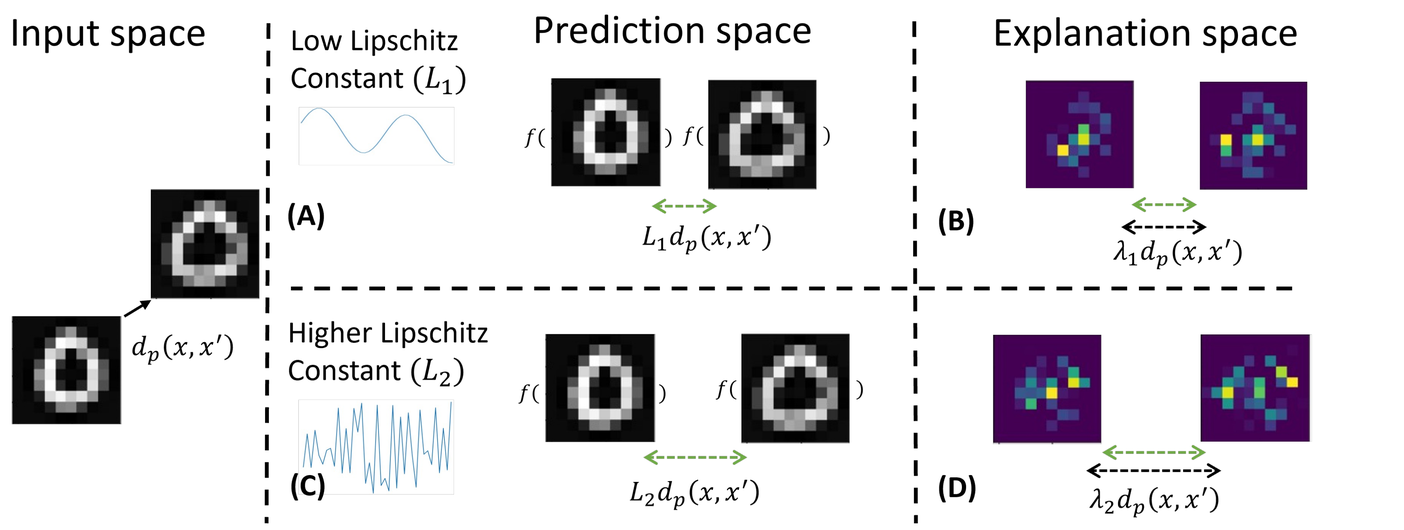
Astuteness $A_{r,\lambda} \geq 1 - \beta$ for $\lambda = 2\sqrt[p]{d}L$, where $\beta \geq \alpha$ and $\beta \to \alpha$

Theorem 2: Remove Individual [4]

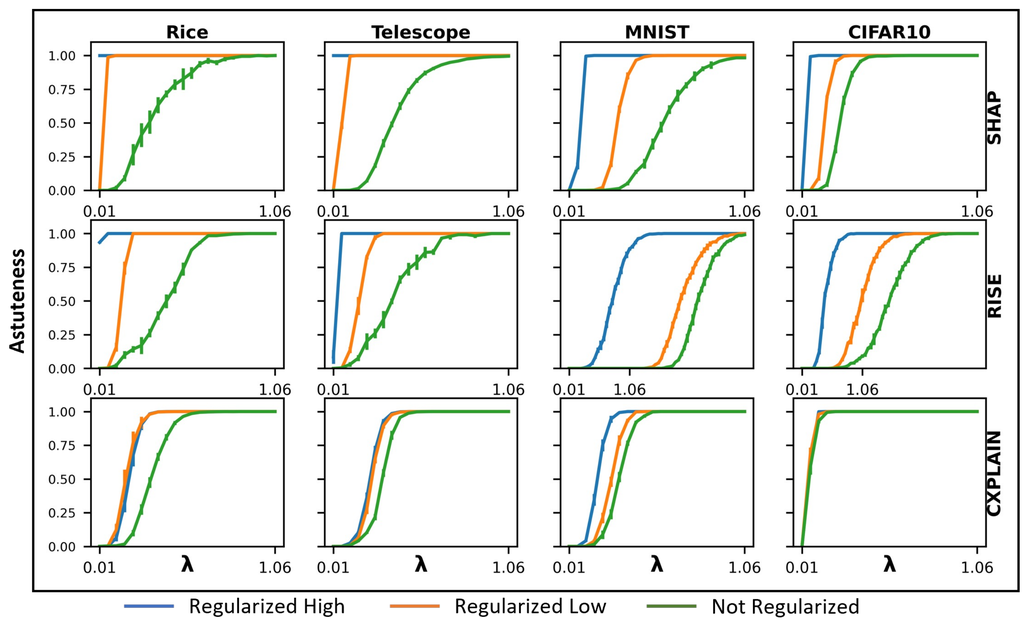
λ-astute for radius r and $\lambda = 2\sqrt[p]{d}L$

Theorem 3: RISE [5]

Astuteness $A\_(r,\lambda) \geq 1-\alpha$ for $\lambda = 2\sqrt[p]{d}L$

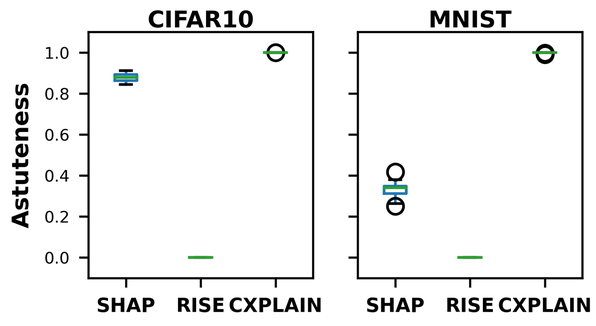


Experiments



Left: Plotting astuteness while varying the Lipschitz regularization on the prediction function shows that *smooth functions result in astute explanations.*

Right: Astuteness as a metric for comparing explainers.



Below: Difference between AUC predicted by our theorems and empirical AUC.

	2layer			4layer			linear			svm		
Datasets	SHAP	RISE	CXPlain	SHAP	RISE	CXPlain	SHAP	RISE	CXPlain	SHAP	RISE	CXPlain
OS	.585	.477	.551	.489	.415	.426	.043	.017	.043	.761	.628	.732
NA	.359	.289	.318	.285	.216	.244	.452	.391	.474	.742	.653	.708
Switch	.053	.053	.003	.086	.083	.039	.043	.028	.034	.557	.472	.524
Rice	.249	.142	.229	.292	.131	.252	.258	.165	.241	.426	.347	.413
Telescope	.324	.213	.317	.345	.244	.333	.223	.149	.211	.501	.439	.504
CIFAR10	.333	.238	.337	.350	.224	.356	.340	.235	.345	.336	.329	.255
MNIST	.441	.297	.435	.522	.357	.519	.455	.303	.448	.443	.379	.448

Below: Comparing astuteness AUC with other explainer metrics

	Neural Network						Logistic Regression					
	SHAP			LIME			SHAP			LIME		
Datasets	AUC ↑	PGI ↑	RIS ↓	AUC ↑	PGI ↑	RIS ↓	AUC ↑	PGI ↑	RIS ↓	AUC ↑	PGI ↑	RIS ↓
O-Synthetic	.198 ± .001	**.320**	5.96	**.379 ± .001**	.250	**5.53**	**.209 ± .000**	**.171**	**5.67**	.029 ± .000	.154	9.35
HELOC	.130 ± .000	**.260**	1.57	**.506 ± .002**	**.260**	**1.19**	**.177 ± .000**	.121	**1.46**	.045 ± .002	**.156**	4.53
COMPAS	.276 ± .001	**.274**	**4.24**	**.515 ± .000**	.232	4.49	.160 ± .000	**.128**	**3.12**	**.352 ± .001**	.108	4.24
Adult	.085 ± .000	.53	1.98	**.206 ± .000**	**.670**	**1.79**	**.114 ± .000**	.391	1.86	.042 ± .001	**.420**	**1.72**

Acknowledgements / References

[1] Alvarez-Melis & Jaakkola. *On the Robustness of Interpretability Methods.*
[2] Mangal et al. *Probabilistic lipschitz analysis of neural networks.*
[3] Lundberg & Lee. *A Unified Approach to Interpreting Model Predictions.*
[4] Covert et al. *Explaining by Removing: A Unified Framework for Model Explanation.*
[5] Petsiuk et al. *RISE: Randomized Input Sampling for Explanation of Black-box Models.*

This work was supported in part by grants NIH 2T32HL007427-41 and U01HL089856 from the National Heart, Lung, and Blood Institute and by Northeastern University's Institute for Experiential AI.