



Faithful Perturbations and Evaluations for Post-Hoc Local Explanation Methods

Iain Smith and Osmar Zaiane

Department of Computing Science, University of Alberta

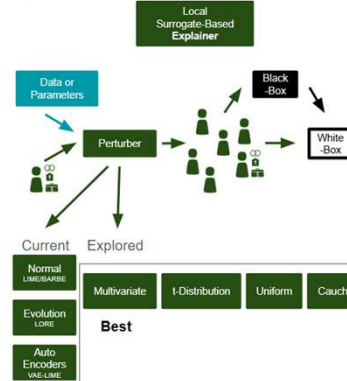


Post-Hoc Explainers

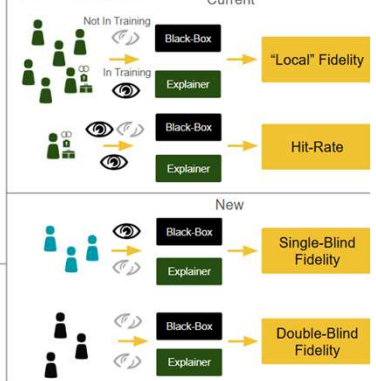
- Explain a **pre-trained black-box**
- **No access to original training data** is guaranteed
- Only get a **label from the black-box**

- **Produce data** using a strategy or distribution **labeled by black-box**
- **Other methods** require access to data
- Other methods prefer **probability outputs** as labels to learn on

a) Architecture

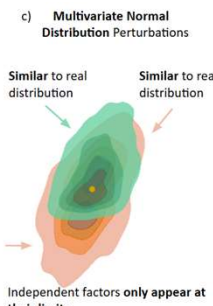
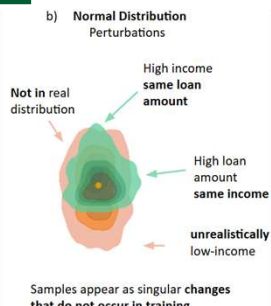
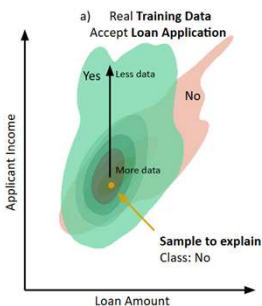


b) Evaluation



Perturbation

Evaluation



Data	Eval.	LIME+P	BARBE+P	LIME	BARBE	LORE	VAE-LIME
Iris	Pert. Fid.	0.87 ± 0.06	0.84 ± 0.07	0.87 ± 0.04	0.83 ± 0.03	0.99 ± 0.01	0.97 ± 0.06
	SB Fid.	0.83 ± 0.10	0.73 ± 0.14	0.76 ± 0.16	0.78 ± 0.14	0.73 ± 0.15	0.58 ± 0.19
	DB Fid.	0.79 ± 0.17	0.69 ± 0.16	0.74 ± 0.16	0.74 ± 0.19	0.72 ± 0.16	0.49 ± 0.20
	Hit Rate	0.65	1	0.633	1	1	0.533
BC	Pert. Fid.	0.92 ± 0.05	0.99 ± 0.01	0.86 ± 0.07	0.99 ± 0.02	0.98 ± 0.02	0.97 ± 0.02
	SB Fid.	0.92 ± 0.08	0.89 ± 0.08	0.60 ± 0.27	0.67 ± 0.20	0.74 ± 0.17	0.87 ± 0.09
	DB Fid.	0.93 ± 0.09	0.89 ± 0.01	0.60 ± 0.27	0.67 ± 0.20	0.74 ± 0.17	0.86 ± 0.11
	Hit Rate	0.7	1	0.72	1	1	0.61
Loan	Pert. Fid.	0.72 ± 0.05	0.87 ± 0.04	0.73 ± 0.05	0.83 ± 0.04	0.97 ± 0.08	0.85 ± 0.06
	SB Fid.	0.61 ± 0.20	0.81 ± 0.10	0.64 ± 0.19	0.76 ± 0.19	0.78 ± 0.07	0.76 ± 0.11
	DB Fid.	0.65 ± 0.23	0.81 ± 0.13	0.68 ± 0.20	0.76 ± 0.20	0.79 ± 0.16	0.80 ± 0.13
	Hit Rate	0.68	0.98	0.69	0.94	0.95	0.64
Libras	Pert. Fid.	0.58 ± 0.09	0.79 ± 0.04	0.70 ± 0.11	0.99 ± 0.00	0.93 ± 0.02	0.84 ± 0.09
	1-SB Fid.	0.48 ± 0.22	0.43 ± 0.16	0.38 ± 0.18	0.23 ± 0.11	0.41 ± 0.19	0.39 ± 0.20
	1-DB Fid.	0.48 ± 0.30	0.42 ± 0.22	0.29 ± 0.06	0.18 ± 0.15	0.42 ± 0.24	0.47 ± 0.24
	Hit Rate	0.5	1	0.417	1	0.8	0.417
Avg.	Pert. Fid.	0.773	0.873	0.790	0.910	0.968	0.908
	1-SB Fid.	0.710	0.715	0.595	0.610	0.665	0.650
	1-DB Fid.	0.713	0.703	0.578	0.588	0.668	0.655
	Hit Rate	0.633	0.995	0.615	0.985	0.938	0.550

- **Perturbation:** use multivariate distributions for **correlated features**
- **Evaluation:** fidelity on **real data** weighted by distance to the perturbed sample

Our Changes

Results

- Averaged results using a **Random Forest and Neural Network** black-box
- Tested on **six variation of explainer**
- **Two with our perturbations**

- Found that fidelity on **perturbed data is not the same as on real data**
- Found that **changes improved LIME**
- Even greater **improvements on other methods like BARBE**