Select Wisely and Explain: Active Learning and Probabilistic Local Post-hoc Explainability

- Aditya Saini, Dr. Ranjitha Prasad in AIES'22

Trustworthy AI Systems
Instructor: Dr. C. Anantaram

Sumedha Chugh
PhD21123

INDRAPRASTHA INSTITUTE *of* INFORMATION TECHNOLOGY **DELHI**

Outline

- Introduction
- Existing Post-hoc Methods
- Motivation
- Basics
  - Active Learning
  - Gaussian Process
- Methodology
  - Sampler
  - Explainer
- Experiments
  - Stability
  - Uncertainty
  - Ability on Image Dataset
- Conclusion & Future Scope

Introduction

Explain it to me like I am 5
- Users to ML System Designer
  - Michael Scott

- Local Post-hoc Explainability Method
- Based on Active Learning
- Introduces a novel locally faithful acquisition function
- Proposes a Gaussian Process Regressor based uncertainty driven sampling method

Existing Post-hoc Explainability Methods

Method	Technique	Cons
LIME	Shadow-model method – construct an interpretable model that replicates the original	Low consistency and robustness due to random-perturbation based surrogate dataset
KernelSHAP	Makes use of SHAP values to determine feature importance	Requires complete training data, takes a long time to compute, explains correlation not causation
BayesLIME & BayesSHAP	Models uncertainty of local explanations to sample better from randomly sampled dataset	Low Fidelity
ALIME	Uses auto-encoder based approach to generate samples	Highly Complex
DLIME	Uses Clustering Algorithm to create surrogate dataset	When training data is lesser, gives bad output, poor fidelity due to uneven distribution of points across clusters
BayeLIME	Incorporates weighed sum of prior knowledge, creating Bayesian version of LIME	Need to find a unique prior for each problem, employs hyper-parameter tuning

Motivation

Previous Works

- Inconsistent or Unreliable Explanations
- No Guidance for choosing number of perturbations
- Sampler and Explainer related but modelled independently

Goal

- Introduce Information Theory driven sampling procedure that chooses accurate number of perturbations
- Jointly design sampler and explainer

Basics

Active Learning

- We have one sample and its corresponding output
- Active learning minimizes cost of sampling next point by uncertainty reduction
- Variance acts as a measure of uncertainty, next point x_n is chosen where acquisition function α is maximum

$$\mathbf{x}_n = \arg\max_{\mathbf{x}} \alpha(\mathbf{x}|\mathcal{D}_{n-1}).$$

- Most popular acquisition functions include Upper confidence bound (UCB), Uncertainty reduction (UR)

$$\text{UCB} : \mathbf{x}_n = \arg\max_{\mathbf{x}} \mu_{n-1}(\mathbf{x}) + \sqrt{\beta_n}\sigma_{n-1}(\mathbf{x})$$
$$\text{UR} : \mathbf{x}_n = \arg\max_{\mathbf{x}} \sigma_{n-1}(\mathbf{x}).$$

Basics

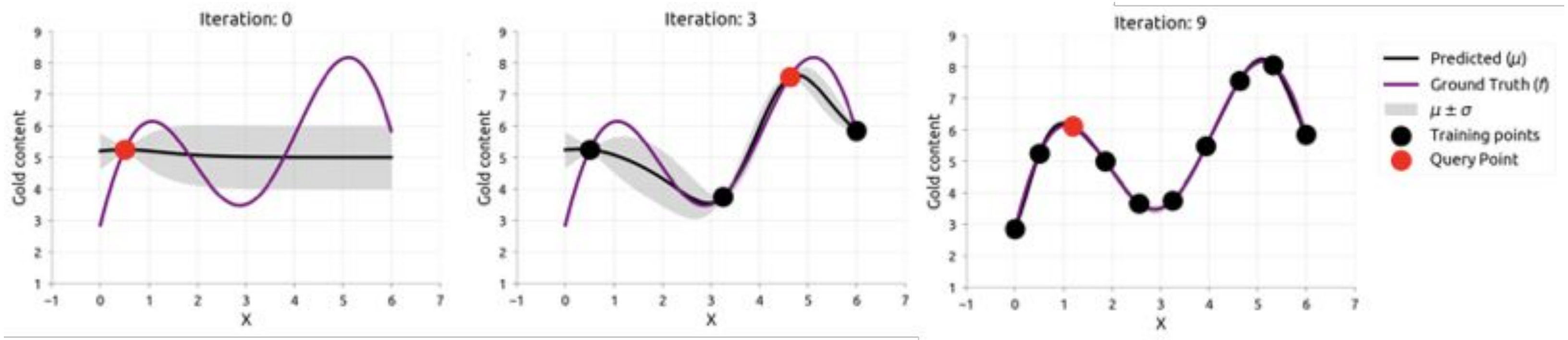
Active Learning

Algorithm:

1. Choose and add the point with the highest uncertainty to the training set (fits a Gaussian Process (GP) as surrogate model)

2. Train on the new training set

3. Go to #1 till convergence or budget elapsed



Basics

Gaussian Process

- A Gaussian process is a random process, where any point $x \in \mathbb{R}^d$ is assigned a random variable f(x) and where the joint distribution of a finite number of these variables p(f(x1),…,f(xN)) is itself Gaussian.

- Mathematically, $p(f|X)=N(f|\mu,K)$

- $f=(f(x_1),\ldots,f(x_N))$, $\mu=(m(x_1),\ldots,m(x_N))$ and $K_{ij}=\kappa(x_i,x_j)$.

- m is the mean function, κ is a positive definite kernel function or covariance function.

- Different Types of Gaussian Process Kernels are: White noise kernel, Gaussian kernel or radial basis function kernel, Rational quadratic kernel, Periodic kernel, Matern kernel etc.

Methodology

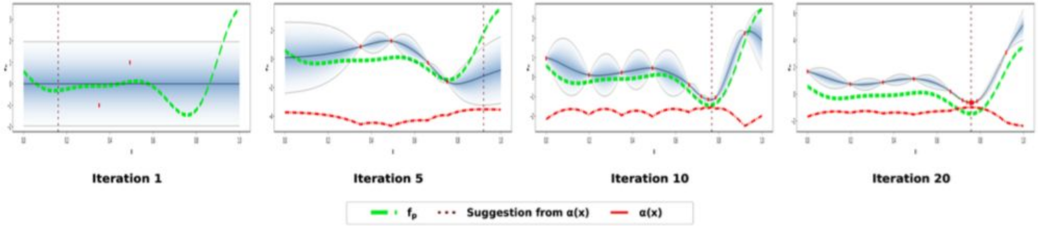
Sampler

- Previous acquisition functions not directly applicable for local explainability
- Proposed acquisition function: Faithful Uncertainty Reduction(FUR)

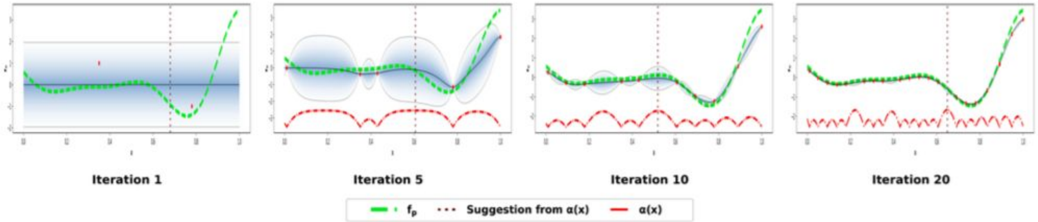
$$\mathbf{x}_n = \arg\max_{\mathbf{x}} \underbrace{-\left\| \left(\mathbf{x} - \mathbf{x}_0 - \frac{\bar{\sigma}\epsilon}{\log(n)} \right) \right\|_2}_{\mathrm{T1}} + \underbrace{\sigma_n(\mathbf{x})}_{\mathrm{T2}},$$

  - T1 ensures local fidelity, sample efficiency
  - T2 ensures maximum Information gain

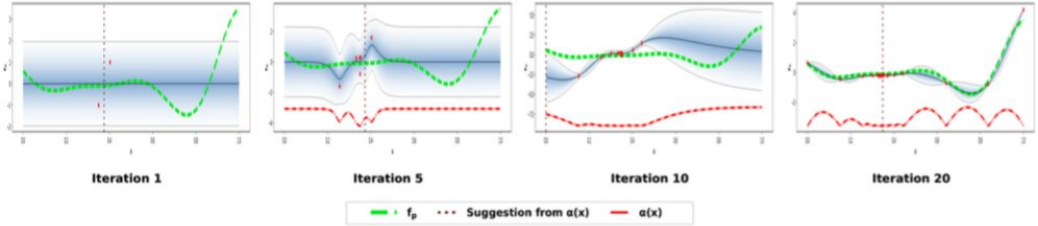
Methodology



(a) UCB(Upper confidence bound): $\arg\max_{\mathbf{x}} \alpha(\mathbf{x}|\mathcal{D}_n) = \arg\max_{\mathbf{x}} \mathbf{x} + k\sigma_n(\mathbf{x})$



(b) Uncertainty Reduction: $\arg\max_{\mathbf{x}} \alpha(\mathbf{x}|\mathcal{D}_n) = \arg\max_{\mathbf{x}} \sigma_n(\mathbf{x})$



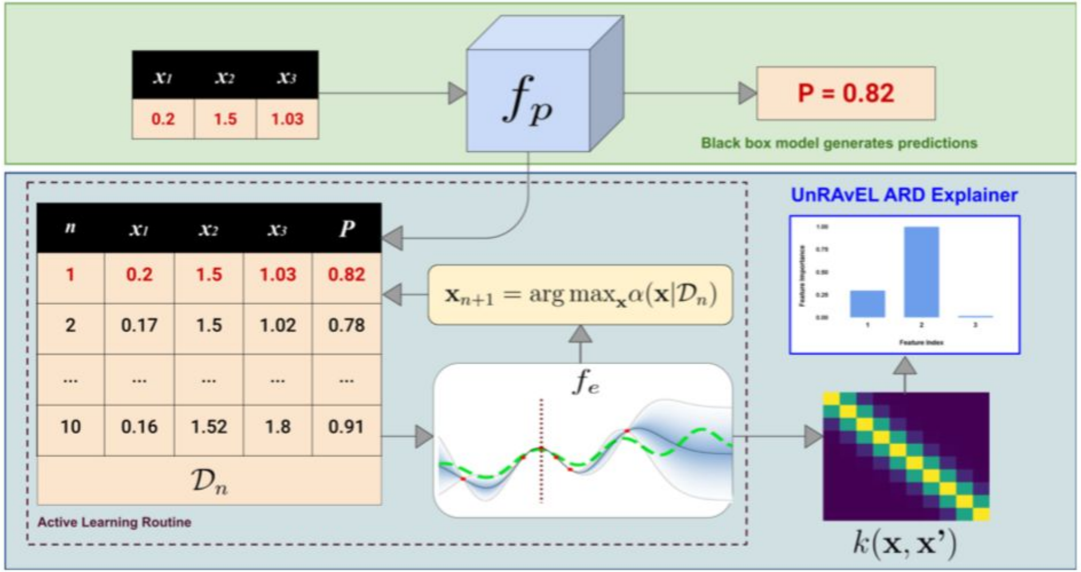
(c) Faithful Uncertainty Reduction: $\arg\max_{\mathbf{x}} \alpha(\mathbf{x}|\mathcal{D}_n) = \arg\max_{\mathbf{x}} -\left\|\left(\mathbf{x} - \mathbf{x}_0 - \frac{\overline{\sigma}\epsilon}{\log(n)}\right)\right\|_2 + \sigma_n(\mathbf{x})$

Can be seen that FUR maintains local faithfulness while UR and UCB can't

Methodology

Explainer

- ARD Explainer or UnRAvEL LIME



Algorithm 1 UnRAvEL: Uncertainty driven Robust Active learning based locally faithful Explanations

Require: Black-box model f_p, Instance $\mathbf{x}_0 \in \mathbb{R}^d$, $\bar{\sigma}$, $\boldsymbol{\sigma}_{\mathcal{D}} = [\sigma_1, \ldots, \sigma_n]$, Maximum iterations L, Acquisition function $\alpha(\cdot)$

Initialize $\mathcal{D}$ using $(\mathbf{x}_0, f_p(\mathbf{x}_0))$

Set exploration domain for $\alpha(\mathbf{x})$: $\mathbf{x} \in [\mathbf{x} - \boldsymbol{\sigma}_{\mathcal{D}}, \mathbf{x} + \boldsymbol{\sigma}_{\mathcal{D}}]$.

Initialize the GPR and the ARD kernel.

Active Learning Routine:

for $l = 1$ to L **do**

  Obtain $\mathbf{x}_{l+1}$ by optimizing $\alpha(\mathbf{x})$. $\mathcal{D} \leftarrow \mathcal{D} \cup (\mathbf{x}_{l+1}, f_p(\mathbf{x}_{l+1}))$.

  Train the GPR based f_e model on $\mathcal{D}$.

end for

return Surrogate data $\mathcal{D}$, Importance scores using ARD-explainer or UnRAvEL-LIME.

Experiments

Stability

Evaluation Metric : Jaccard Distance

Dataset:

Dataset	Task	p	n_{train}	n_{total}	R^2 score
Parkinson's	C	22	195	175	0.80
Cancer	C	30	512	569	0.98
Adult	C	14	30162	45222	0.84
Bodyfat	R	14	226	252	0.99
Boston	R	13	455	506	0.92

Dataset	LIME	BayLIME	UnRAvEL-L	UnRAvEL
Parkinson's	0.743	0.738	0.499	**0.146**
Cancer	0.826	0.824	0.655	**0.295**
Adult	0.520	0.524	0.402	**0.288**
Boston	0.664	0.668	**0.462**	0.539
Bodyfat	0.687	0.693	**0.503**	0.701

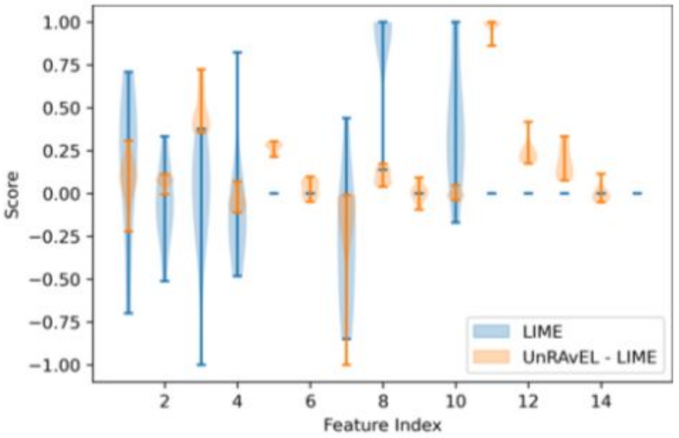
UnRAvEL outperforms both LIME and BayLIME, for both the regression datasets, UnRAvEL-L, i.e., UnRAvEL with a Linear kernel, outperforms the rest.

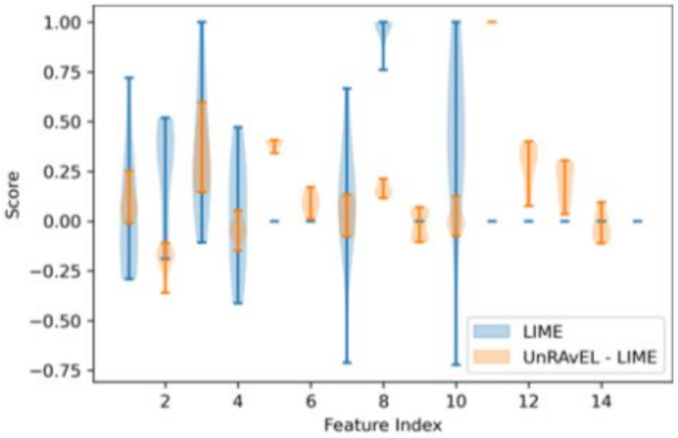
Experiments

Uncertainty

Evaluation Metric : Violin plot for variance in feature scores for 2 randomly selected test points

Dataset: Adult census Dataset





UnRAvEL-LIME has very low uncertainty as compared to the importance scores generated by LIME

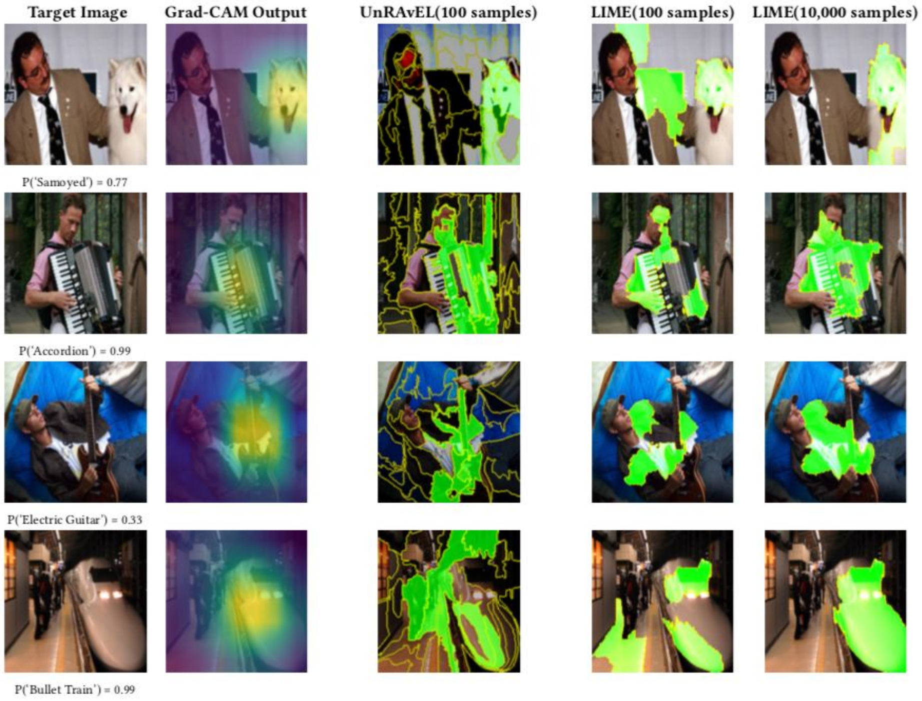
Experiments

Ability on Image Datasets

Evaluation Metric : Comparison with GradCAM

Dataset: Randomly selected images from Imagenet dataset

UnRAvEL at just 100 samples can produce explanations that are semantically accurate and are consistent with LIME at 10000 samples and Grad- CAM



Conclusion & Future Work

- A novel more stable and robust Local Post-hoc explainable AI method has been proposed
- It employs acquisition function based on Active Learning, followed by a Gaussian process regression model
- Future Research aims to make this work a global explanation module

Thank you !

Open to Feedback and Questions

sumedhac@iiitd.ac.in