Probabilistic Post-hoc Explainable AI methods

Aditya Saini, Ranjitha Prasad

Department of Electronics & Communications Engineering, IIITD

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
DELHI

IEEE Signal Processing Society
DELHI CHAPTER

Contents

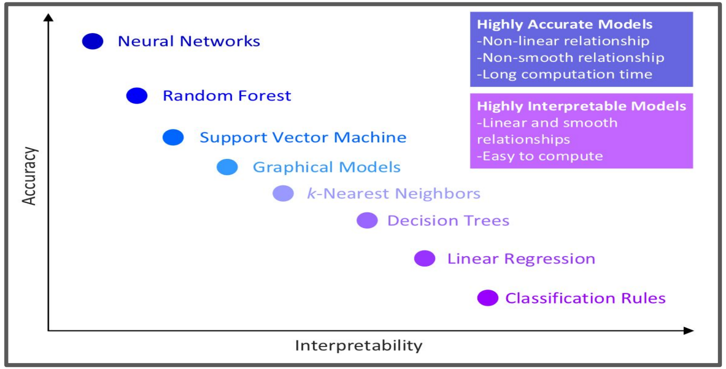
- **Introduction**
  - Need for Explainable AI(XAI)
  - XAI approaches
- **New Methods**
  - Motivation
  - **Method 1: UnRAvEL**
    - Methodology
    - Results
  - **Method 2: BGMLIME**
    - Methodology
    - Results
- **Future work**

Introduction

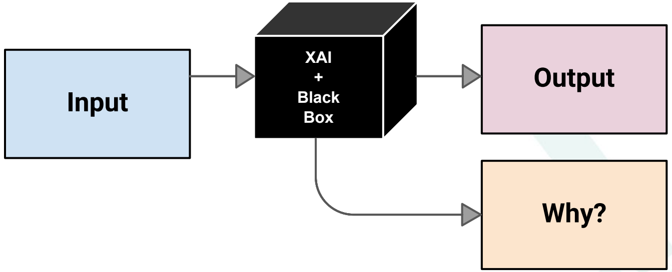
- Real world problems require high capacity AI models which though performant, possess decision making paradigms which are tough to *explain*.

- Issues:
    - lack of trustworthiness
    - inhibited use in safety critical domains
    - Decreased productivity in automated systems

- Goal of XAI(or Explainable AI) - **Transparent decision making process**



Accuracy-Interpretability Tradeoff for Popular Machine Learning Models[1]



XAI pipeline enables transparency in decision making process

1. Morocho-Cayamcela, M. E., Lee, H., & Lim, W. (2019). Machine learning for 5G/B5G mobile and wireless communications: Potential, limitations, and future directions. IEEE access, 7, 137184-137206.

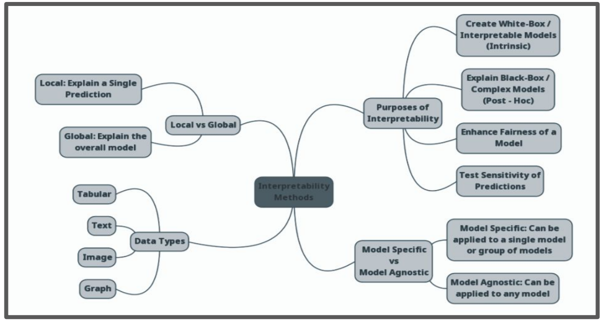
Introduction - Post hoc Perturbation XAI methods

- Post hoc Perturbation XAI methods assume:
  - A pre trained black box model, and
  - A sample of interest

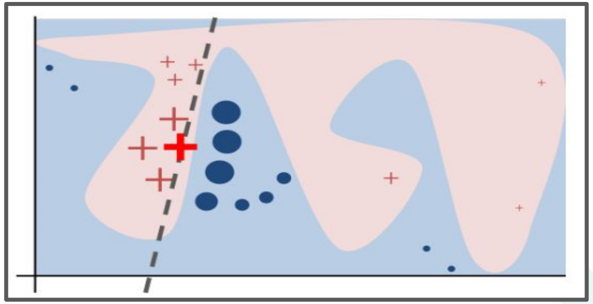
- Post hoc perturbation approach - Explainability through feature attribution scores. Popular approaches:
  - LIME[3]
  - KernelSHAP[4]

- The workflow involves:
  - Generation of surrogate data
  - Optimization of locality inducing loss function of the form

$$L(f_p, f_e, \pi_{\mathbf{x}}) = \sum_{\mathbf{x}_0, \mathbf{x} \in \mathcal{X}} \pi_{\mathbf{x}}(\mathbf{x}_0)\left(f_p(\mathbf{x}_0) - f_e(\mathbf{x})\right)^2$$



A basic taxonomy of XAI approaches[2]

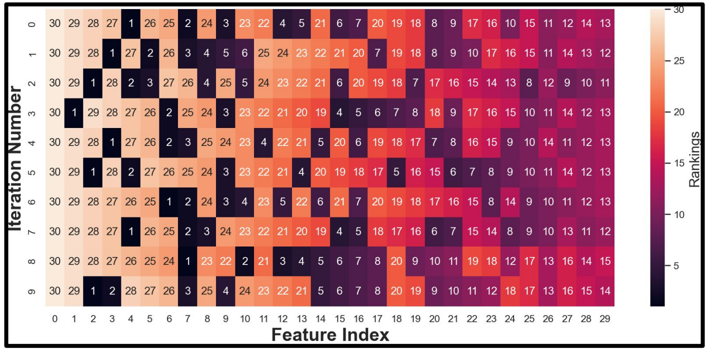


LIME samples instances, gets predictions using original predictive function, and weighs them by the proximity to the instance being explained (represented here by size). The dashed line is the learned explanation that is locally (but not globally) faithful.[3]

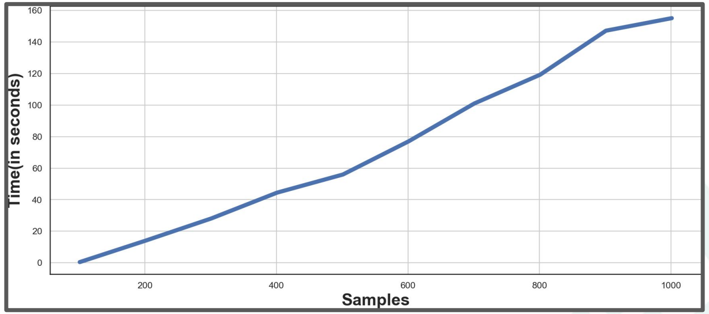
2. Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable ai: A review of machine learning interpretability methods. Entropy, 23(1), 18.
3. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).
4. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.

Issues in Existing Approaches

- **Inconsistent and unreliable explanations.**
  - Variation in ranking
- **Highly sample inefficient**
  - Sample size is positively correlated with fidelity
- **Prone to adversarial attacks**
- **Low Fidelity Explanations**
  - Reliance on low capacity models



LIME rankings for a sample taken from the UCI[5] Breast cancer dataset. Black box = Support Vector Classifier, RoC = 0.98



Sample size vs Time for a single sample from the Imagenet-1000 dataset[6]. Black box = Pretrained ResNet-18

5. Asuncion, A., & Newman, D. (2007). UCI machine learning repository.
6. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. International journal of computer vision, 115(3), 211-252.

Post-hoc Explainers

Technique name	Strategy used	Issues
DLIME[9]	Uses a deterministic clustering algorithm for creating surrogate dataset	In presence of less training points, the model gives bad approximation of the underlying function
BayLIME[10]	By using a Bayesian modification of LIME, it incorporates prior knowledge of a given sample to remove inconsistency for similar samples	Finding useful priors is nuanced and difficult for each unique problem
ALIME[11]	Uses an auto encoder based approach for weighing the generated samples to get better accuracy	The complex structure counters itself as explaining ALIME's decision becomes another XAI task
BayesLIME/BayesSHAP[12]	Uses focussed sampling for producing high information surrogate dataset	Uses the same low capacity linear model, which can make it difficult to produce high fidelity results

9. Zafar, M. R., & Khan, N. M. (2019). DLIME: A deterministic local interpretable model-agnostic explanations approach for computer-aided diagnosis systems. arXiv preprint arXiv:1906.10263.
10. Zhao, X., Huang, W., Huang, X., Robu, V., & Flynn, D. (2021, December). Baylime: Bayesian local interpretable model-agnostic explanations. In Uncertainty in Artificial Intelligence (pp. 887-896). PMLR.
11. Shankaranarayana, S. M., & Runje, D. (2019, November). ALIME: Autoencoder based approach for local interpretability. In International conference on intelligent data engineering and automated learning (pp. 454-463). Springer, Cham.
12. Slack, D., Hilgard, A., Singh, S., & Lakkaraju, H. (2021). Reliable post hoc explanations: Modeling uncertainty in explainability. Advances in Neural Information Processing Systems, 34, 9391-9404.

Method: Motivation

- Perturbation based techniques consists of two main parts:
  a. **Local sampler:** For generation of surrogate dataset
  b. **Sparse linear explainer module:** For providing feature importance scores

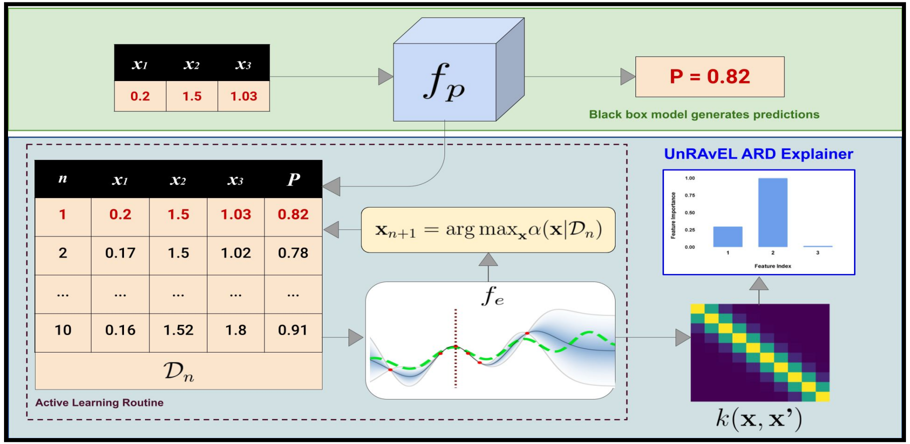
- **Both these modules are strongly intertwined, and are yet modelled independently!**
  - Intuitively, it makes sense to have a connection between the sampler and the explainer model

- Bayesian methods have many advantages
  - Gaussian process models can alter their linearity using kernel settings
  - Active Learning driven acquisition function based strategy can be used for sampling on top of GP models

New Method 1: UnRAvEL

Accepted in 5th AAAI/ACM Conference on AI, Ethics, and Society 2022 main track titled "Select Wisely and Explain: Active Learning and Probabilistic Local Post-hoc Explainability"

- **UnRAvEL = Uncertainty driven Robust Active learning based locally faithful Explanations**
- Novel perturbation based XAI technique
  - Uses Gaussian process with ARD kernel as explainer model
  - Consists of a novel acquisition function FUR(Faithful Uncertainty Reduction)
    - Uses uncertainty-driven sampling based on the posterior distribution on the probabilistic locality using Gaussian process regression
  - Differs from other existing models as **sampler and explainer are jointly designed**



UnRAvEL: Active Learning Routine

- GP models can be used to develop an exploration-exploitation based strategy for inducing locality into the model.

- We want an acquisition function - samples in the vicinity of a given sample by trading-off
  - information gain and
  - local fidelity.

- We have two popular acquisition strategies in literature already:
  - UCB(Upper Confidence Bound): Used for finding global optimum

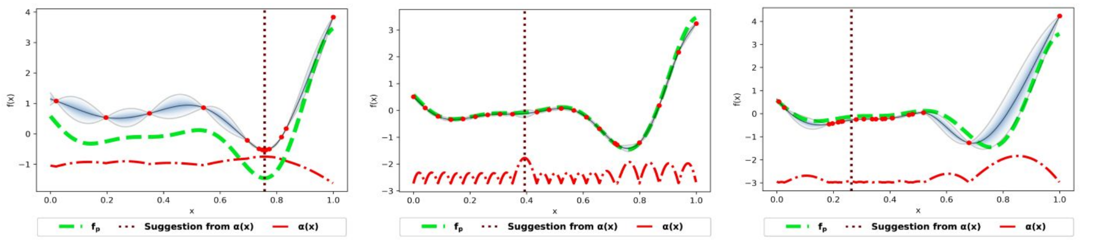
$$\boxed{\mathbf{x}_n = \arg\max_{\mathbf{x}} \mu_{n-1}(\mathbf{x}) + \sqrt{\beta_n}\sigma_{n-1}(\mathbf{x})}$$

  - UR(Uncertainty Reduction): Used for efficiently traversing residual space

$$\boxed{\mathbf{x}_n = \arg\max_{\mathbf{x}} \sigma_{n-1}(\mathbf{x})}$$

UnRAvEL: Active Learning Routine

- We developed an acquisition function that is exactly in the between the previous two.
  - Faithful Uncertainty Reduction(FUR)



(a) UCB(Upper confidence bound) (b) UR(Uncertainty Reduction) (c) FUR(Faithful Uncertainty Reduction)

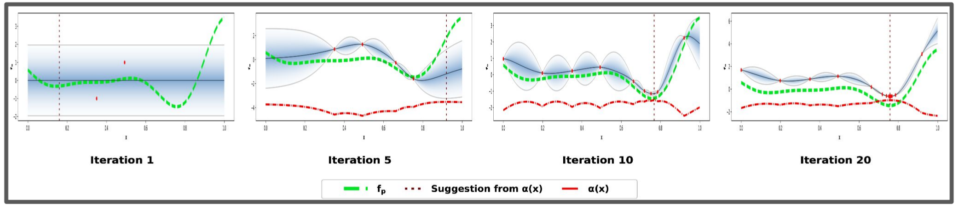
- This function

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

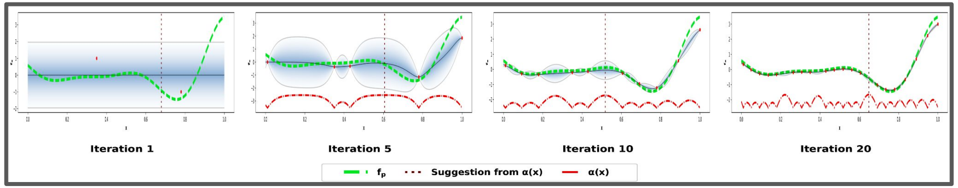
is able to selectively choose high information samples as the term
  - T1 controls the local fidelity, and
  - T2 depends on the information gain.

UnRAvEL: Active Learning Routine

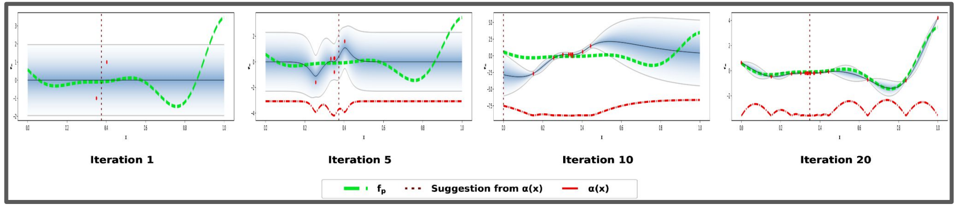
- We have two popular acquisition strategies in literature already:
  - UCB(Upper Confidence Bound): Used for finding global optimum



  - UR(Uncertainty Reduction): Used for efficiently traversing residual space



  - FUR(Faithful Uncertainty Reduction): Can be used for generating localized surrogate dataset



Experiments: Stability

- **Goals**
  - We wanted to evaluate how UnRAvEL would perform against LIME and BayLIME
- **Setup**
  - Metric: Used Jaccard distance over the rankings collected for 10 randomly selected test samples generated in 10 consecutive runs of an XAI module

$$J(X_i, X_j) = 1 - \frac{|X_i \cap X_j|}{|X_i \cup X_j|}$$

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 is able to produce more consistent and reliable explanations as compared to baselines

Experiments: Fidelity

- **Goals**
  - We wanted to evaluate how UnRAvEL would perform even with low sample size against LIME at 100 and 10000 samples
  - We also plotted GradCam scores for the reader's reference
- **Setup**
  - Dataset: 4 randomly selected images from Imagenet dataset
  - Black box: Pretrained ResNet-18 model
- **Results**
  - UnRAvEL is able to produce high fidelity explanations even in low sample regimes

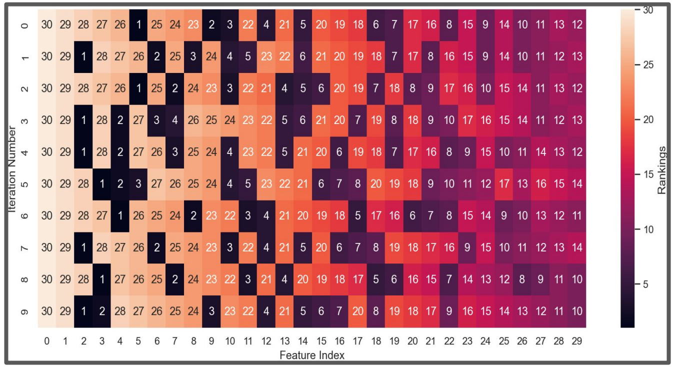
Top 5 features for some sample images from Imagenet dataset

New Method 2: BGMLIME

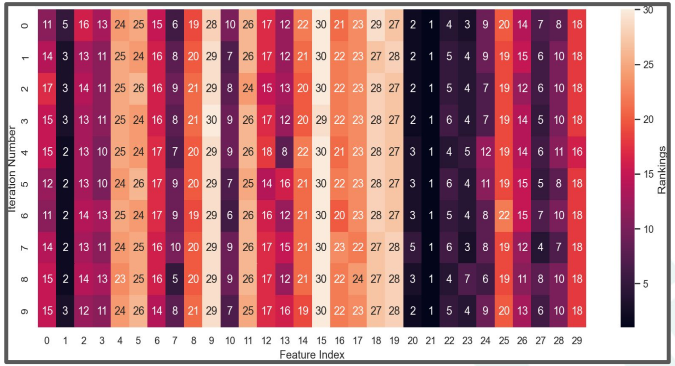
- **BGMLIME = Bayesian Gaussian Mixture based Local Interpretable Model Agnostic Explanations**

- Bayesian flavor of LIME
  - Uses probabilistic Gaussian mixture model based clustering to produce surrogate dataset
  - Uses same methodology in LIME as it only adds a Step-0 before the usual workflow

0. **Clustering the given dataset** into a mixture of Gaussian distributions and finding the mean and covariance matrix of each component.
1. **Sampling around a given feature set** using the respective mixture's means and covariances. After that **use the prediction model to get target values for the sampled sets.**
2. **Using a feature selection technique** like LASSO or forward selection on the newly created dataset to come up with the top K features.
3. Outputting the top K features in an ***interpretable*** and ***meaningful*** way.



LIME Rankings



BGMLIME Rankings

Future work

- **Global Explainer based on UnRAvEL:**
    - GPs are computationally complex!
    - Exploring sparse approximations of GPs for building global extension
- **Multimodal joint explanations:**
    - GP kernel can be utilized in many domain specific applications.
    - Working on building a novel explainer module that can consider ML models of different modalities.
- **BGMLIME using Bayesian Optimization:**
    - To make BGMLIME hyperparameter free, we are working on Bayesian Optimization based pre-processing module for choosing the optimal hyper priors used in the BGMM module.

Acknowledgements

- **iHub Anubhuti - IIITD Foundation**
  - Chanakya Undergraduate Fellowship

iHub Anubhuti-
IIITD Foundation

- **IntelliCom Lab, IIIT Delhi**

IntelliCom Lab

Thank you!

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
DELHI

IEEE Signal Processing Society
DELHI CHAPTER