“Why Should I Trust You?”

Early Explainability: A Review of LIME

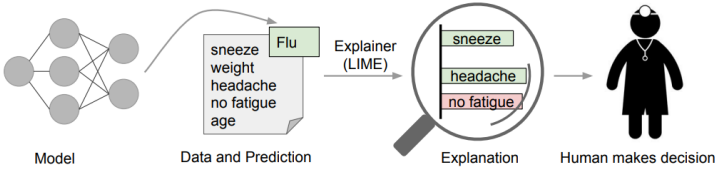
Steven Morse
William & Mary, Data Science
DATA 691, Graph Learning

April 2, 2025

Overview

Idea

The surface of a complex model is locally simple $\rightarrow$ we can approximate it with a simple model $\rightarrow$ simple models are interpretable.



"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 (2016).

Related work

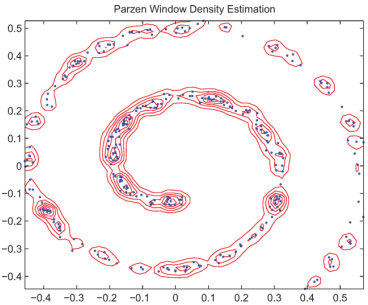
At the time (2016), explainability methods (to explain black box models) fell into generally two approaches:

- Local approximation methods. Ex: Parzen window estimators: estimate a density at x with summed kernel functions.

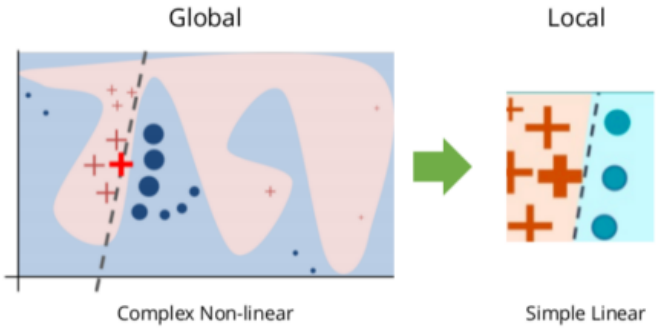
$$\hat{p}(x) = \frac{1}{n} \sum_i K(x - x_i; h)$$

- Gradient- and relevance-based methods. Ex: saliency maps, Layer-wise Relevance Propagation (LRP) (similar to GradCAM, which came out 2017).

Note: there are many other ways to taxonomize (see a review like Linardatos 2020).



LIME conceptual intro



LIME formalities

- Given a model $f : \mathcal{X} \to T$,
- Choose an "interpretable" model $g \in G$ with measure of interpretability $\Omega(g)$,
- Define a proximity measure $\pi : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$.

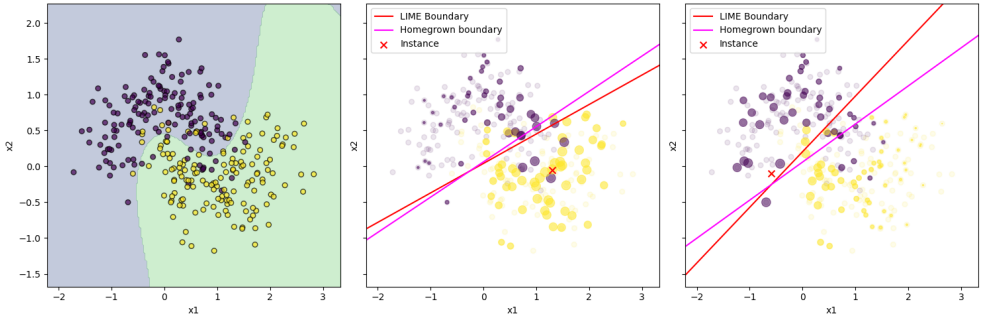
LIME Definition

The LIME explainer of x is defined as

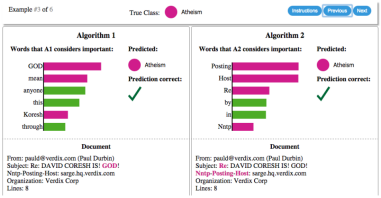
$$\xi(x) = \arg\min_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g) \tag{1}$$

The paper focuses on linear models g, RBF $\pi_x(z) = \exp(-D(x,z)^2/\sigma^2)$, weighted $L2$ loss $\mathcal{L}(f, g, \pi) = \sum_z \pi_x(z)(f(z) - g(z))^2$, and standard regularization like LASSO.

Toy example (re-implementation)



Examples

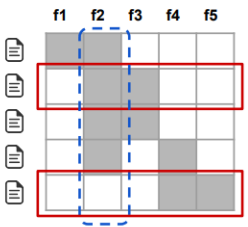


(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

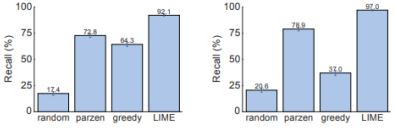
Figure 4: Explaining an image classification prediction made by Google's Inception neural network. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)

Submodular Pick

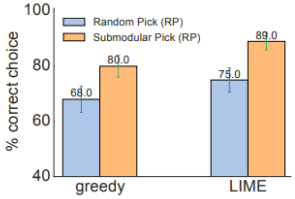
- So far we've only explained a single prediction
- We can apply this model to many samples and use the resulting explanations to get:
  - Most explanatory features
  - Most explanatory set of datapoints
- More formally, given n samples and their $g_i = \xi(x_i)$ explainers corresponding to weights w_i, we compute a $\mathcal{W}$ with $\mathcal{W}_{ij} = w_i^{(j)}$.
- Picking explanatory features is easy. (e.g. $\max_j \sum_i w_i^{(j)}$)
- Picking explanatory datapoints is combinatorially hard, but since the score function is submodular (diminishing gains from adding an example), a greedy algorithm has a constant-factor approximation guarantee.

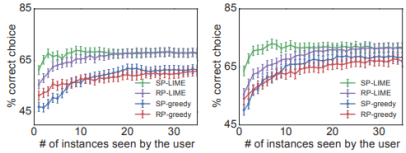


User Experiments

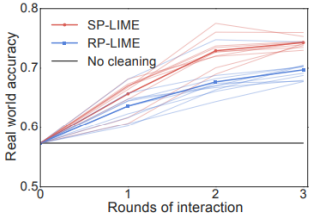


(a) Sparse LR (b) Decision Tree





(a) Books dataset (b) DVDs dataset



Subsequent work

- Local linear approximation approaches: Kernel SHAP, MAPLE (local linear + random forest)
- Gradient-based: GradCAM (2017), Integrated Gradients, SmoothGrad
- Contribution propagation: DeepLIFT
- Game theory: SHAP

Figure: Google trends for "explainable ai"

Questions?