"Why Should I Trust You?" Explaining the Predictions of Any Classifier

Yeojin Joo

November 8, 2018

Overview

1. Introduction
2. Goals for Explainer
3. LIME (Local Interpretable Model-agnostic Explanations)
4. SP (Submodular Pick) LIME
5. Simulation

Contents

Introduction

Definitions of Trust

- **Trust a prediction** sufficiently to take action based on it.
- **Trust a model** to behave in reasonable ways if deployed.

Solutions

- Providing explanations (**LIME**) for individual predictions
- Selecting multiple such predictions (**SP-LIME**)

Contents

Goals for Explainer

Interpretable

Provide qualitative understanding between the input and the response and easy to understand.

Local fidelity

How the model behaves in the vicinity of the instance being predicted.

Model-agnostic

Explain any model i.e. Treat the original model as a black box.

Global perspective

Select a few explanations to ascertain trust in the model.

Contents

LIME (Local Interpretable Model-agnostic Explanations)

- Let the model being explained
  $f : R^d \to R,\ f(x) = Pr(Y = k|X = x)$
- Define an explanation as a model $g \in G,\ g : \{0,1\}^{d'} \to R$
  · G = Class of intepretable models
  · $z \in \{0,1\}^{d'}$: binary vector for interpretable representation.
  ex) the presence of a word or a super-pixel
- $\pi_x(z)$ = proximity measure between an instance z to x
  ex) $\pi_x(z) = exp(-D(x,z)^2/\sigma^2)$

LIME (Local Interpretable Model-agnostic Explanations)

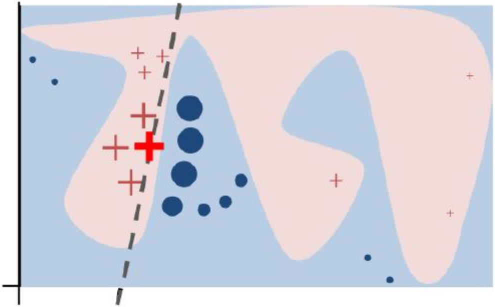
- The explanation produced by LIME:
$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} L(f, g, \pi_x) + \Omega(g)$$

  · $L(f, g, \pi_x)$ = measure of how unfaithful g in approximating f with locality π_x

  ex) $L(g, f, \pi_x) = \sum_{z, z' \in Z} \pi_x(z)(f(z) - g(z'))^2$

  · $\Omega(g)$ = measure of complexity

LIME (Local Interpretable Model-agnostic Explanations)



Blue/Pink = The black box model's complex decision function f
The bold red cross = the instance being explained
The dashed line = the learned explanation that is locally faithful

LIME (Local Interpretable Model-agnostic Explanations)

Algorithm 1 Sparse Linear Explanations using LIME

Require: Classifier f, Number of samples N
Require: Instance x, and its interpretable version x'
Require: Similarity kernel π_x, Length of explanation K

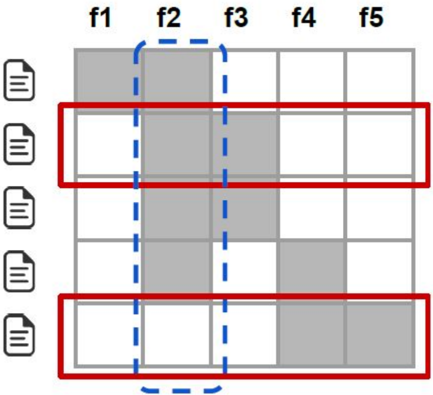
$Z \leftarrow \{\}$
for $i \in \{1, 2, 3, ..., N\}$ **do**
  $z'_i \leftarrow sample\_around(x')$
  $Z \leftarrow Z \cup \langle z_i, f(z_i), \pi_x(z_i) \rangle$
end for
$\omega \leftarrow$ K-Lasso(Z,K) $\triangleright$ with z_i as features, $f(z)$ as target
return ω

Contents

SP (Submodular Pick) LIME

- LIME is not sufficient to evaluate and assess trust in the model as a whole.
- Give a global understanding of the model by explaining **a set of individual instances**.
- **B** = budget, the number of explanations
- **pick step** = the task of selecting B instances for the user to inspect.
  - pick a diverse, representative set of explanations

SP (Submodular Pick) LIME



Rows = instances
columns = features

SP (Submodular Pick) LIME

Algorithm 2 Submodular pick (SP) algorithm

Require: Instances X, Budget B
  for all $x_i \in X$ **do**
    $W_i \leftarrow$ **explain**(x_i, x_i') ▷ Using Algorithm 1
  end for
  for $j \in \{1, ..., d'\}$ **do**
    $I_j \leftarrow \sqrt{\sum_{i=1}^{n} |W_{ij}|}$ ▷ Compute feature importances
  end for
  $V \leftarrow \{\}$
  while $|V| < B$ **do**
    $V \leftarrow V \cup argmax_i c(V \cup \{i\}, W, I)$
  end while
  return V

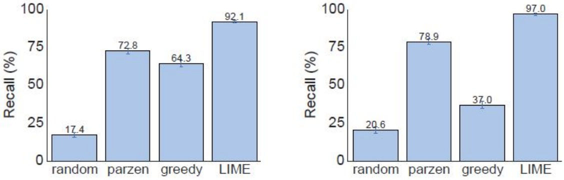
$$c(V, W, I) = \sum_{j=1}^{d'} 1_{[\exists i \in V : W_{ij} > 0]} I_j$$

Contents

Simulation

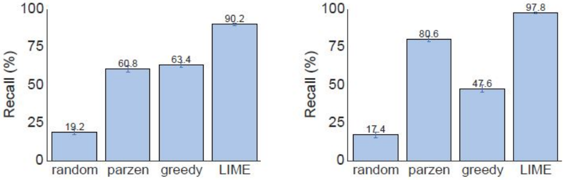
- Use two sentiment analysis datasets (books and DVDs).
- classify product reviews as positive or negative.
- Experiment Setup
  - · method = LIME, parzen, greedy, random procedure
  - · K=10
  - · train=1600, test=400

Simulation



(a) Sparse LR (b) Decision Tree

Figure 6: Recall on truly important features for two interpretable classifiers on the books dataset.



(a) Sparse LR (b) Decision Tree

Figure 7: Recall on truly important features for two interpretable classifiers on the DVDs dataset.

Simulation

Table 1: Average F1 of *trustworthiness* for different explainers on a collection of classifiers and datasets.

	Books				DVDs			
	LR	NN	RF	SVM	LR	NN	RF	SVM
Random	14.6	14.8	14.7	14.7	14.2	14.3	14.5	14.4
Parzen	84.0	87.6	94.3	92.3	87.0	81.7	94.2	87.3
Greedy	53.7	47.4	45.0	53.3	52.4	58.1	46.6	55.1
LIME	**96.6**	**94.5**	**96.2**	**96.7**	**96.6**	**91.8**	**96.1**	**95.6**

- untrustworthy = the prediction changes when untrustworthy features are removed.
- trustworthy = otherwise.

Simulation

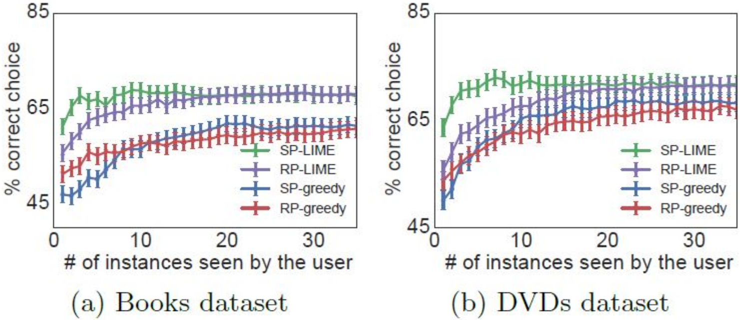


Figure 8: Choosing between two classifiers, as the number of instances shown to a simulated user is varied. Averages and standard errors from 800 runs.

The End