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

Marco Tulio Ribeiro
University of Washington
Seattle, WA 98105, USA
marcotcr@cs.uw.edu

Sameer Singh
University of Washington
Seattle, WA 98105, USA
sameer@cs.uw.edu

Carlos Guestrin
University of Washington
Seattle, WA 98105, USA
guestrin@cs.uw.edu

KDD 2017

Presenter: Jack Lanchantin

Local Interpretable Model-agnostic Explanations (LIME)

- Goal: identify an **interpretable model** over the **interpretable representation** that is **locally faithful** to the classifier
- Finds a representation that is understandable to humans, regardless of the actual features used by the model.
- $x \in \mathbb{R}^d$: original representation of an instance being explained
- $x' \in \{0, 1\}^{d'}$: binary vector for its interpretable representation

Example: Interpretable Representation for Text Classification

- Binary vector indicating the presence or absence of a word
  - Independent of what the classifier uses (e.g. word embeddings)

Example: Interpretable Representation for Image Classification

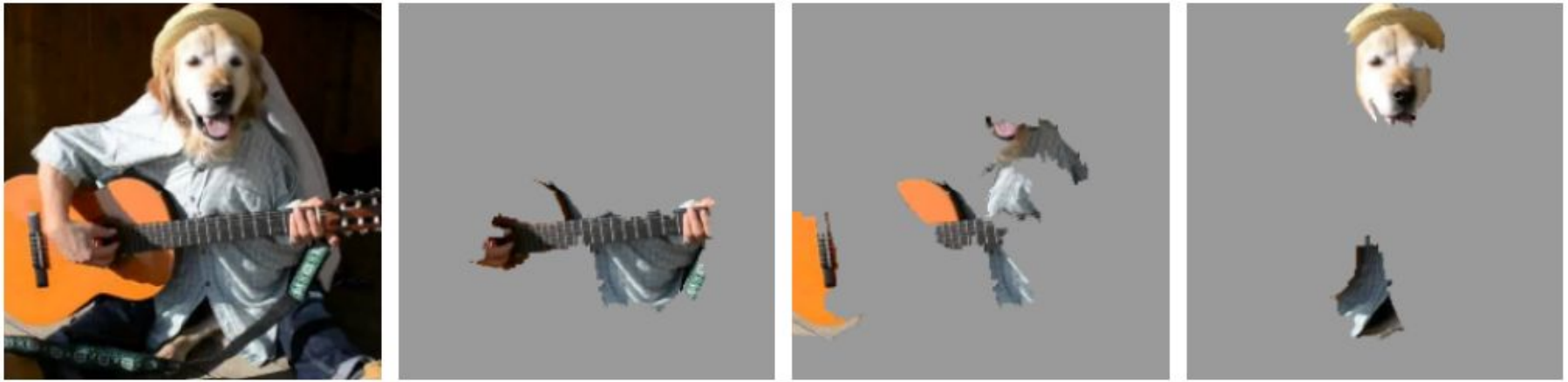
- Binary vector indicating the “presence” or “absence” of a contiguous patch of similar pixels (a super-pixel), while the classifier may represent the image as a tensor with three color channels

Original Image

Interpretable Components

Example: Interpretable Representation for Image Classification

- Binary vector indicating the “presence” or “absence” of a contiguous patch of similar pixels (a super-pixel), while the classifier may represent the image as a tensor with three color channels



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

Explanation Models

- Original model $f : \mathbb{R}^d \rightarrow \mathbb{R}$
- Interpretable model $g \in G: \mathbb{R}^{d'} \rightarrow \mathbb{R}$
- Domain of g is $\{0,1\}^{d'}$ where d' is the number of interpretable components
  - I.e. *g* acts over absence/presence of the interpretable components.

Explanation Models

- Original model $f : \mathbb{R}^d \rightarrow \mathbb{R}$
- Interpretable model $g \in G: \mathbb{R}^{d'} \rightarrow \mathbb{R}$
- Domain of g is $\{0,1\}^{d'}$ where d' is the number of interpretable components
  - I.e. *g* acts over absence/presence of the interpretable components.

- $\pi_x(z)$: proximity between an instance *z* to x, so as to define locality around *x*
- $L(f, g, \pi_x)$: measure of how unfaithful *g* is in approximating *f* in the locality defined by π_x
- $\Omega(g)$: measure of complexity of the explanation *g* (tradeoff of interpretability)

LIME

- 2 objectives:
  - **Local fidelity:** minimize L(*f, g*, π_x)
  - **Interpretability:** minimize Ω(g)
- The explanation produced by LIME is obtained by the following:

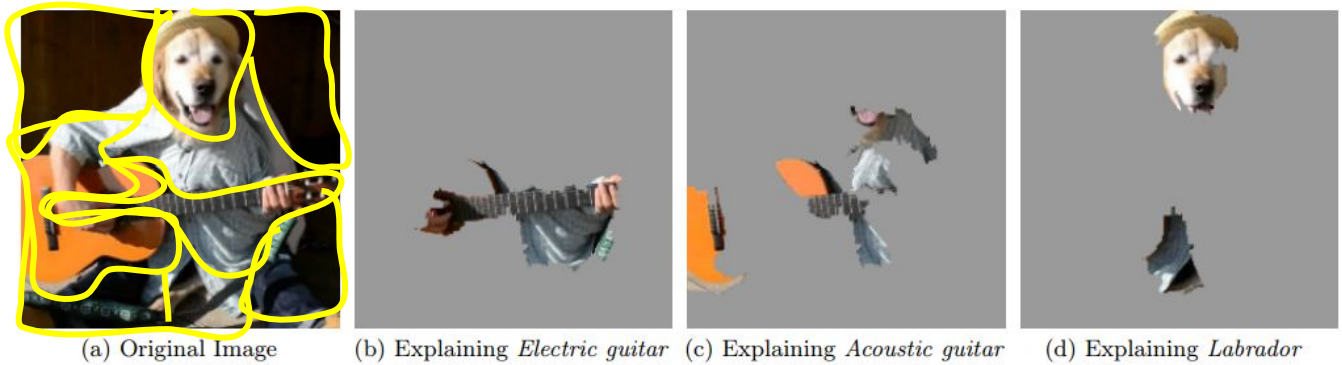
$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \quad \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

Sampling for Local Exploration

- Approximate $L(f, g, \pi_x)$ by drawing samples, weighted by π_x
- Sample instances around x' by drawing nonzero elements of x' uniformly at random (number of such draws is also uniformly sampled)

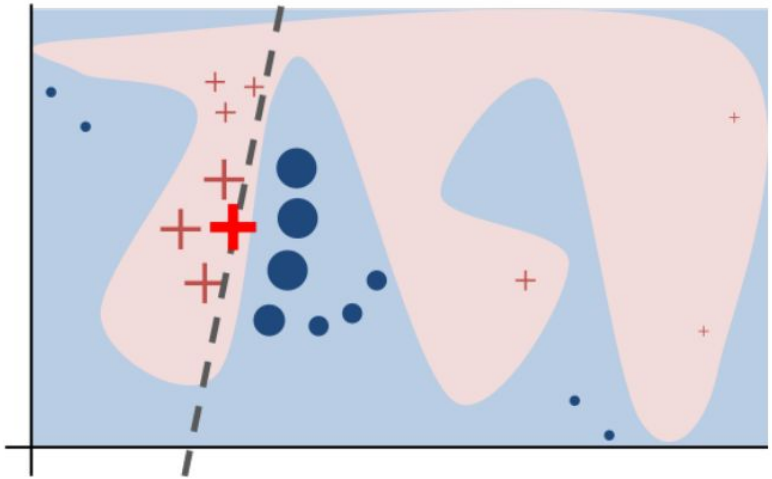
Sampling for Local Exploration

- Approximate L(*f, g*, π_x) by drawing samples, weighted by π_x
- Sample instances around x' by drawing nonzero elements of x' uniformly at random (number of such draws is also uniformly sampled)
- Given a perturbed sample $z' \in \{0,1\}^{d'}$ (which contains a fraction of the nonzero elements of x'), we recover the sample in the original representation $z \in \mathbb{R}^d$ and obtain f(z), which is used as a label for the explanation model.



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

Sampling for Local Exploration



Sampling for Local Exploration

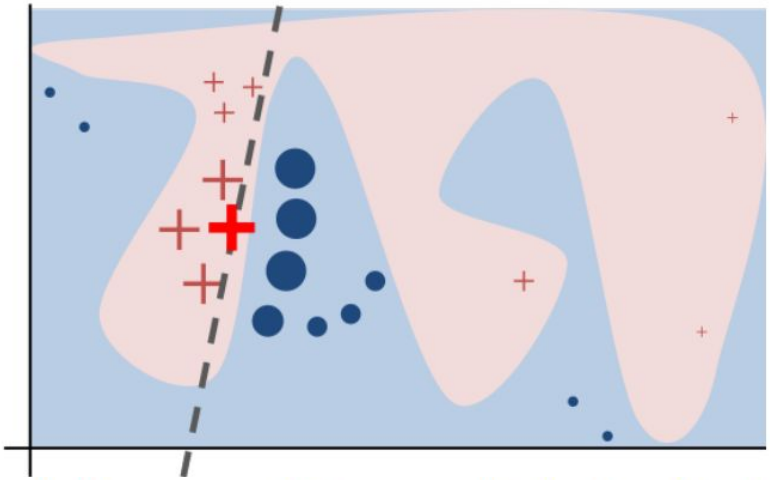


Figure 3: Toy example to present intuition for LIME. The black-box model's complex decision function f (unknown to LIME) is represented by the blue/pink background, which cannot be approximated well by a linear model. The bold red cross is the instance being explained. LIME samples instances, gets predictions using f, 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.

Sparse Linear Explanations

- Let G be the class of linear models, such that $g(z') = w_g \cdot z'$
- Use the locally weighted square loss as L
  - let $\pi_x(z) = \exp(-D(x, z)^2 / \sigma^2)$ be an exponential kernel defined on some distance function D (e.g. cosine distance for text, L2 distance for images) with width σ.

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

Sparse Linear Explanations

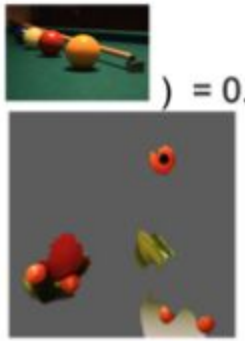
Algorithm 1 Sparse Linear Explanations using LIME

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

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

Sparse Linear Explanations

P() = 0.54



P( ) = 0.07

P() = 0.05

Are explanations faithful to the model?

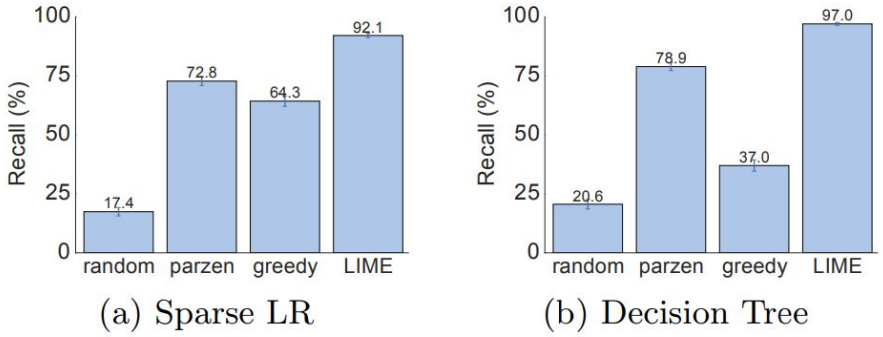


Figure 6: Recall on truly important features for two interpretable classifiers on the books dataset.

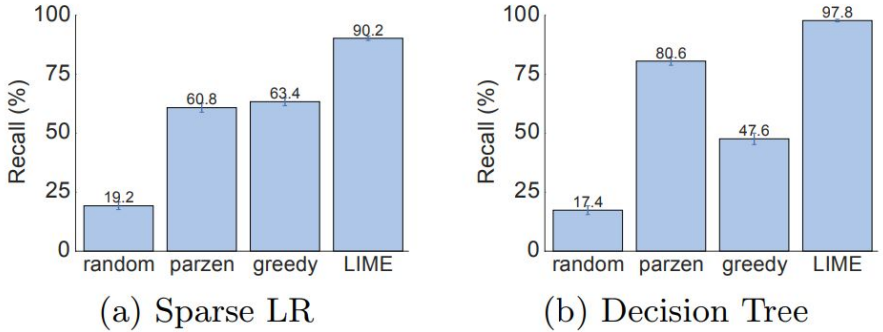
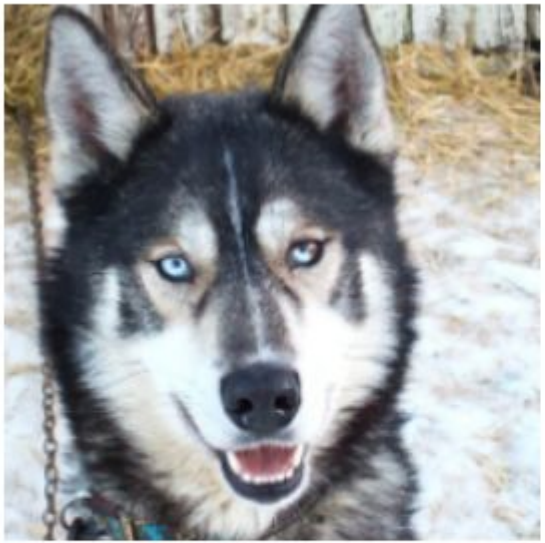
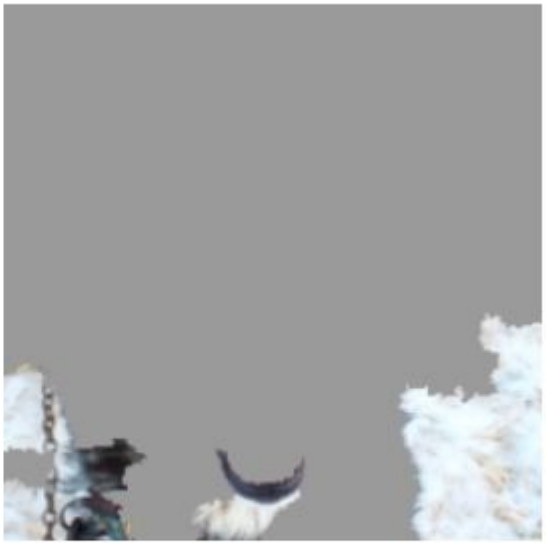


Figure 7: Recall on truly important features for two interpretable classifiers on the DVDs dataset.

Husky vs Wolf



(a) Husky classified as wolf



(b) Explanation

Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.