“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.
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.
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 $f : \mathbb{R}^d \rightarrow \mathbb{R}$
- Interpretative 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 $\pi_x$
- $\Omega(g)$: measure of complexity of the explanation $g$ (tradeoff of interpretability)
LIME

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

$$\xi(x) = \arg\min_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g)$$
Sampling for Local Exploration

- Approximate $L(f, g, \pi_x)$ by drawing samples, weighted by $\pi_x$
- 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 $L(f, g, \pi_x)$ by drawing samples, weighted by $\pi_x$
- 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 $\sigma$.

$$
\mathcal{L}(f, g, \pi_x) = \sum_{z, z' \in \mathcal{Z}} \pi_x(z) (f(z) - g(z'))^2
$$
**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 $\pi_x$, Length of explanation $K$

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

return $w$
Sparse Linear Explanations

\[ P(\text{fish}) = 0.54 \]

\[ P(\text{frog}) = 0.07 \]

\[ P(\text{balloon}) = 0.05 \]
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.
Figure 11: Raw data and explanation of a bad model’s prediction in the “Husky vs Wolf” task.