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

Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin
University of Washington

Presenter: Liyuan Liu (Lucas)
What’s this paper about?

• Propose algorithms (LIME and SP-LIME) to Explain the predictions of Any Classifier

Lucas has the flu

Why should I trust you?
What’s this paper about?

- Propose algorithms (LIME and SP-LIME) to Explain the predictions of Any Classifier
What can it do for us?

• Trust
  • Does my model really work?
Does my model really work?

20 Newsgroups subset – Atheism vs Christianity

20 newsgroups - Train

Accuracy: 94%

20 Newsgroups subset – Atheism vs Christianity

Accuracy: 89%

20 newsgroups - test
Does my model really work?

20 Newsgroups subset – Atheism vs Christianity

Accuracy: 94%

Accuracy: 89%

Accuracy: 57%

Accuracy: 69%
Does my model really work?

20 newsgroups - Train

Accuracy:

89%

Accuracy:

94%

Hidden religion dataset - Accuracy:

69%

57%

20 Newsgroups subset – Atheism vs Christianity

Document

From: pauld@verdix.com (Paul Durbin)
Subject: Re: DAVID CORESH IS! GOD!
Nntp-Posting-Host: sarge.hq.verdix.com
Organization: Verdix Corp
Lines: 8

Document

From: pauld@verdix.com (Paul Durbin)
Subject: Re: DAVID CORESH IS! GOD!
Nntp-Posting-Host: sarge.hq.verdix.com
Organization: Verdix Corp
Lines: 8
What’s the benefit for us?

• Trust
  • Does my model really work?

• Insights
  • What is my model doing?

• Improve
  • How do I fix my model?
What is my model doing?

Google’s Inception NN

\[ P(\text{guitar}) = 0.32 \]
\[ P(\text{dog}) = 0.24 \]
\[ P(\text{puppy}) = 0.21 \]
What is my model doing?

\[ P(\text{guitar}) = 0.32 \]

\[ P(\text{guitar}) = 0.24 \]

\[ P(\text{dog}) = 0.21 \]

Google’s Inception NN
How do I fix my model?

**wolf v.s. husky**

![Images of wolves and huskies with predictions and true labels]

Only 1 mistake!!!
How do I fix my model?
How these models work?

• 1. LIME
  • Explain one prediction of one classifier

• 2. SP-LIME
  • Select a set of representative instances with explanations
Using LIME to explain a complex model’s prediction for input $x_i$

1. Sample points around $x_i$
2. Use complex model to predict labels for each sample
3. Weigh samples according to distance to $x_i$
4. Learn new simple model on weighted samples
5. Use simple model to explain
Interpretable representation: images

$x$ (3 color channels / pixel)

$x'$ (contiguous superpixels)
Sampling example - images

Original Image
$P(\text{labrador}) = 0.21$

<table>
<thead>
<tr>
<th>Perturbed Instances</th>
<th>$P(\text{Labrador})$</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="number" alt="Image 1" /></td>
<td>0.92</td>
</tr>
<tr>
<td><img src="number" alt="Image 2" /></td>
<td>0.001</td>
</tr>
<tr>
<td><img src="number" alt="Image 3" /></td>
<td>0.34</td>
</tr>
</tbody>
</table>

Locally weighted regression

Explanation
Explain Whole Models
SP-LIME

1 prediction -- local behavior

Pick k ‘typical’ predictions

avoid redundancy in explanations

Submodular function optimization

Explanation for 1 example

Whole space of data
Understanding Black-box Predictions via Influence Functions

Pang Wei Koh, Percy Liang
Stanford University

Presenter: Liyuan Liu (Lucas)
Basic Intuition

• How to interpret a model?
  • what inputs maximally activate these neurons?
  • can we represent this model with a simpler one?
  • which part of the input was most responsible for this prediction?

• This paper:
  • Treat model as a function of the training data
  • Explain prediction w.r.t. the training data ``most responsible`` for prediction
Technique Details

• Talk given by the author:

  • https://www.youtube.com/watch?v=0w9fLX_T6tY