Interpretability

Advanced Machine Learning for NLP
Jordan Boyd-Graber

NEED FOR INTERPRETABILITY

Slides adapted from Marco Tulio Ribeiro
Classification

• But representation is often less important (means to end)
• We really care about end result
• And not doing simple things like decision trees / linear classifier
Classification

- But representation is often less important (means to end)
- We really care about end result
- And not doing simple things like decision trees / linear classifier
- That’s why we’re making complicated algorithms

Can explain this mess 😊
Locally-faithful simple decision boundary

Good explanation for prediction

What’s an Explanation

From: Keith Richards
Subject: Christianity is the answer
NTTP-Posting-Host: x.x.com

I think Christianity is the one true religion.
If you’d like to know more, send me a note
What’s an Explanation

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

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

\[ P(\text{cat}) = 0.21 \]
What makes good Explanation?

- Interpretable: Humans can Understand
- Faithful: Describes Model
- Model Agnostic: Generalize to Many Models
Method

- Sample points around $x_i$
- Use complex model to predict labels for each sample
- Weigh samples according to distance to $x_i$
- Learn new simple model on weighted samples
- Use simple model to explain
Perturbing an Example

\[ x \text{ (3 color channels / pixel)} \]

\[ x' \text{ (contiguous superpixels)} \]

Model

Human
Perturbing an Example

\[ x \text{ (embeddings)} \]

\[
\begin{array}{ccccc}
0.5 & 0.3 & 1.3 & 4.4 & 1.1 \\
\end{array}
\]

... 

\[ x' \text{ (words)} \]

This is a horrible movie.

Model

Human
Image Example

Original Image

P(labrador) = 0.21

Perturbed Instances    P(Labrador)

<table>
<thead>
<tr>
<th>Perturbed Instances</th>
<th>P(Labrador)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td>0.34</td>
</tr>
</tbody>
</table>

Locally weighted regression

Explanation
Is this a good Classifier?

- Predicted: wolf
  True: wolf

- Predicted: husky
  True: husky

- Predicted: wolf
  True: wolf

- Predicted: husky
  True: husky

- Predicted: husky
  True: husky

- Predicted: wolf
  True: wolf
Is this a good Classifier?

Didn't trust the model: 60% Before explanations, 40% After explanations.

"Snow insight": 90% Before explanations, 100% After explanations.

Interpretability | 9 of 1
Improving ML Algorithms

20 newsgroups - Train

Hidden religion dataset

Evaluate

20 newsgroups - test

Turkers don’t know about this dataset

Explain

REPEAT
Example #5 of 10

True Class: Atheism

Words that the algorithm considers important:

- Host
- Posting
- NNTP
- to
- New
- Thanks
- anyone
- email
- not
- has

Bar length indicates importance, and color indicates to which topic: Christianity (green) or Atheism (Pink).

Document:

From: johnchad@triton.unm.edu (jchadwick)
Subject: Another request for Darwin Fish
Organization: University of New Mexico, Albuquerque
Lines: 11
NNTP-Posting-Host: triton.unm.edu

Hello Gang,

There have been some notes recently asking where to obtain the DARWIN fish.
This is the same question I have and I have not seen an answer on the net. If anyone has a contact please post on the net or email me.

Thanks,

john chadwick
johnchad@triton.unm.edu
or
Improving ML Algorithms

Accuracy on hidden set

Train on 20 newsgroups
Train on hand-cleaned 20 newsgroups
Train on 20 newsgroups turkers clean data