Introduction to Model Interpretability

Jonathan Mak

May 27th, 2020

CS 224U
Table of Contents

Approach
- What is model interpretability?
- Why is it important?

Model Agnostic Tactics
- LIME
- SHAP
A Typical Machine Learning Example

- I have data, and I want to solve a problem. (How do I diagnose Disease X?) So, just deploy a model!
A Typical Machine Learning Example

- I have data, and I want to solve a problem. (How do I diagnose Disease X?)
  So, just deploy a model!
- But in real life, things are much more complicated.
A Typical Machine Learning Example

- I have data, and I want to solve a problem. (How do I diagnose Disease X?) So, just deploy a model!
- But in real life, things are much more complicated.
- You have various parties: ML Scientists, Product Managers, End Users.
A Typical Machine Learning Example

- I have data, and I want to solve a problem. (How do I diagnose Disease X?) So, just deploy a model!
- But in real life, things are much more complicated.
- You have various parties: ML Scientists, Product Managers, End Users.
  - ML Scientist: Which model features should I use? Does my model perform well?
A Typical Machine Learning Example

- I have data, and I want to solve a problem. (How do I diagnose Disease X?)
  So, just deploy a model!
- But in real life, things are much more complicated.
- You have various parties: ML Scientists, Product Managers, End Users.
  - ML Scientist: Which model features should I use? Does my model perform well?
  - Product Managers: Can I trust/deploy this model? Is it fair for all parties?
A Typical Machine Learning Example

- I have data, and I want to solve a problem. (How do I diagnose Disease X?)
  So, just deploy a model!
- But in real life, things are much more complicated.
- You have various parties: ML Scientists, Product Managers, End Users.
  - ML Scientist: Which model features should I use? Does my model perform well?
  - Product Managers: Can I trust/deploy this model? Is it fair for all parties?
  - End User: Why did it give me this prediction?
What is Interpretability?
What is Interpretability?

- Interpretability is **NOT**...
What is Interpretability?

- Interpretability is **NOT**...
- About making **all** models interpretable
What is Interpretability?

- Interpretability is **NOT**...
  - About making **all** models interpretable
  - About understanding **every single bit** about the model
What is Interpretability?

- Interpretability is **NOT**...
  - About making **all** models interpretable
  - About understanding **every single bit** about the model
  - Against developing highly **complex** models
What is Interpretability?

- Interpretability is **NOT**...
  - About making **all** models interpretable
  - About understanding **every single bit** about the model
  - Against developing highly **complex** models
  - About **only** gaining user trust or fairness.
What is Interpretability?

- Interpretability is **NOT**...
  - About making **all** models interpretable
  - About understanding **every single bit** about the model
  - Against developing highly **complex** models
  - About **only** gaining user trust or fairness.

- **Interpretability is the ability to understand the overall consequences of the model and ensuring the things we predict are accurate knowledge aligned with our initial research goal.**
Why is it Important?
Why is it Important?

- Correlation often does not equal causality, so a solid model understanding is needed when it comes to making decisions and explaining them.
Why is it Important?

- Correlation often does not equal causality, so a solid model understanding is needed when it comes to making decisions and explaining them.
- Helps us identify and mitigate bias, account for context, improve generalization and performance, and is also there for ethical and legal reasons.
Why is it Important?

- Correlation often does not equal causality, so a solid model understanding is needed when it comes to making decisions and explaining them.
- Helps us identify and mitigate bias, account for context, improve generalization and performance, and is also there for ethical and legal reasons.
- Don’t treat the model as a black box!
Model Agnostic Tactics

HOW THEY HELP
Model Agnostic
Model Agnostic

- Ability to compare any two models to each other
Model Agnostic

- Ability to compare any two models to each other
- Ignore internal structure
Model Agnostic

- Ability to compare any two models to each other
- Ignore internal structure
- Adapt explanation to target user
Local Interpretable Model-Agnostic Explanations (LIME)
Local Interpretable Model-Agnostic Explanations (LIME)

- Global explanations can be too complicated.
Local Interpretable Model-Agnostic Explanations (LIME)

- Global explanations can be too complicated.
- Zoom in to examine local interpretability
Local Interpretable Model-Agnostic Explanations (LIME)

- Global explanations can be too complicated.
- Zoom in to examine local interpretability
- Summary:
Local Interpretable Model-Agnostic Explanations (LIME)

- Global explanations can be too complicated.

- Zoom in to examine local interpretability

- Summary:
  - Simplify a global model by perturbing input to see how predictions change
Local Interpretable Model-Agnostic Explanations (LIME)

- Global explanations can be too complicated.
- Zoom in to examine local interpretability
- Summary:
  - Simplify a global model by perturbing input to see how predictions change
  - Approximate underlying model learned on these perturbations
Local Interpretable Model-Agnostic Explanations (LIME)

- Steps:
Local Interpretable Model-Agnostic Explanations (LIME)

- Steps:
  - Sample points around $X$
Local Interpretable Model-Agnostic Explanations (LIME)

- Steps:
  - Sample points around X
  - Get predictions from our original model (complex)
Local Interpretable Model-Agnostic Explanations (LIME)

- Steps:
  - Sample points around $X$
  - Get predictions from our original model (complex)
  - Weight samples according to our distance from $x$ (cos for text, L2 for images)
Local Interpretable Model-Agnostic Explanations (LIME)

- Steps:
  - Sample points around X
  - Get predictions from our original model (complex)
  - Weight samples according to our distance from x (cos for text, L2 for images)
  - Learn a simple model from our weighted samples
Local Interpretable Model-Agnostic Explanations (LIME)

- Steps:
  - Sample points around X
  - Get predictions from our original model (complex)
  - Weight samples according to our distance from x (cos for text, L2 for images)
  - Learn a simple model from our weighted samples
  - Utilize simple model for better interpretability!
LIME - Images

Original Image

Interpretable Components
LIME - Images

Original Image
P(tree frog) = 0.54

<table>
<thead>
<tr>
<th>Perturbed Instances</th>
<th>P(tree frog)</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Perturbed Image 1" /></td>
<td>0.85</td>
</tr>
<tr>
<td><img src="image2.png" alt="Perturbed Image 2" /></td>
<td>0.00001</td>
</tr>
<tr>
<td><img src="image3.png" alt="Perturbed Image 3" /></td>
<td>0.52</td>
</tr>
</tbody>
</table>

Locally weighted regression

Explanation
LIME – Text Classification

On 20 newsgroup dataset … what happened?

<table>
<thead>
<tr>
<th>Prediction probabilities</th>
<th>( \text{atheism} )</th>
<th>( \text{christian} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{atheism} )</td>
<td>0.58</td>
<td></td>
</tr>
<tr>
<td>( \text{christian} )</td>
<td>0.42</td>
<td></td>
</tr>
</tbody>
</table>

**Text with highlighted words**

From: johnchad@triton.unm.edu (jchadwic)
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.
LIME – Implementation (with simple Random Forest)

```python
from lime import lime_text
from lime.lime_text import LimeTextExplainer
from sklearn.pipeline import make_pipeline
c = make_pipeline(vectorizer, rf)
explainer = LimeTextExplainer(class_names=class_names)
exp = explainer.explain_instance(
    newsgroups_test.data[idx],
    c.predict_proba,
    num_features=6)
```

![Prediction probabilities]

<table>
<thead>
<tr>
<th>Prediction probabilities</th>
<th>atheism</th>
<th>christian</th>
</tr>
</thead>
<tbody>
<tr>
<td>atheism</td>
<td>0.59</td>
<td></td>
</tr>
<tr>
<td>christian</td>
<td>0.41</td>
<td></td>
</tr>
</tbody>
</table>

![Features]

<table>
<thead>
<tr>
<th>Posting</th>
<th>Host</th>
<th>NNTP</th>
<th>edu</th>
<th>have</th>
<th>There</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability</td>
<td>0.16</td>
<td>0.13</td>
<td>0.10</td>
<td>0.05</td>
<td>0.01</td>
</tr>
</tbody>
</table>

LIME – Drawbacks
LIME – Drawbacks

- Linear model approximating local behavior
LIME – Drawbacks

- Linear model approximating local behavior
- Perturbations can be very use case specific
LIME – Drawbacks

- Linear model approximating local behavior
- Perturbations can be very use case specific
- Ideally, drive perturbations by variation in dataset
LIME – Drawbacks

- Linear model approximating local behavior
- Perturbations can be very use case specific
- Ideally, drive perturbations by variation in dataset
- Labor/resource intensive when picking better models
Shapley Additive exPlanations (SHAP)
Shapley Additive exPlanations (SHAP)

- Explain output through optimal credit allocation using Shapley values
Shapley Additive exPlanations (SHAP)

- Explain output through optimal credit allocation using Shapley values
- Allow for both global interpretability (feature contribution) and local interpretability (observation contribution)
Shapley Additive exPlanations (SHAP) – Shapley Values
Shapley Additive exPlanations (SHAP) – Shapley Values

- Shapley value is the average marginal contribution of a feature value across all possible coalitions/orderings! Considers efficiency, symmetry, dummy, and additivity properties.

\[ g(x') = \phi_0 + \sum_{j=1}^{M} \phi_j \]
SHAP – Text Classification
import shap
explainer = shap.DeepExplainer(model, x_train[:100])
shap_values = explainer.shap_values(x_test[:10])
shap.initjs()

words = imdb.get_word_index()
num2word = {}
for w in words.keys():
    num2word[words[w]] = w
x_test_words = np.stack([np.array(list(map(lambda x: num2word.get(x, "NONE"), x_test[i]))) for i in range(10)])
shap.force_plot(explainer.expected_value[0], shap_values[0][0], x_test_words[0])
SHAP – Drawbacks
SHAP – Drawbacks

- Can be misinterpreted (don’t identify causality, and don’t break consistency!)
SHAP – Drawbacks

- Can be misinterpreted (don’t identify causality, and don’t break consistency!)
- Direct access to data is necessary
LIME vs SHAP

LIME: weight is local approximation

If you increase age, \( f(x) \) goes down
If you decrease it, \( f(x) \) goes up
Ergo, Age has negative weight

SHAP: weight is contribution w.r.t baseline

John’s age contributes positively towards \( f(x) \) w.r.t. the average.
Ergo, Age has positive weight
Thank you!

JMAK@STANFORD.EDU
Works Cited

- Intro to AI Interpretability + Model Agnostic Solutions (Marco Ribeiro)
- Interpretability for Everyone (Been Kim)
- Interpreting ML Models/LIME (Lars Hulstaert)
- SHAP - A unified approach to interpreting model predictions (Scott Lundberg)