CS 329T: Trustworthy Machine Learning

Lab 2: April 7, 2022
Outline

- HW 1 common queries (Due tomorrow at 11:59 pm)
- Final Project Deadlines
- LIME Revision
- SHAP Revision
- QII Revision
- LIME and SHAP Exercise
- NLP Trulens primer
Follow-ups on HW1?

- One vs Rest
- Evaluation, tests
Final Project Deadlines

- Project proposal submission (1 page) - Apr 25, 2022
- Project milestone submission (1 page) - May 9, 2022
- Project presentations - May 24-May 26, 2022
- Project report submission (~2-3 pages) - May 30, 2022
LIME: Local Interpretable Model-agnostic Explanations

Introduction: The key intuition behind LIME is that it is much easier to approximate a black-box model by a simple model locally (in the neighborhood of the prediction we want to explain), as opposed to trying to approximate a model globally.
LIME: Methodology

- Given an observation, permute it to create replicated feature data with slight value modifications.
- Compute similarity distance measure between original observation and permuted observations.
- Apply selected machine learning model to predict outcomes of permuted data.
- Select m number of features to best describe predicted outcomes.
- Fit a simple model to the permuted data, explaining the complex model outcome with m features from the permuted data weighted by its similarity to the original observation.
- Use the resulting feature weights to explain local behavior.
Original Image

\[ P(\text{tree frog}) = 0.54 \]

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

**Locally weighted regression**

**Explanation**
Advantages, disadvantages:

● Model agnostic, abstraction
● Predictions vs reasons
● What exactly is local?
● Comparable examples
SHAP: SHapley Additive exPlanations

- Shapley values and SHAP
- An intuitive way to understand the Shapley value is the following illustration: The feature values enter a room in random order. All feature values in the room participate in the game (= contribute to the prediction). The Shapley value of a feature value is the average change in the prediction that the coalition already in the room receives when the feature value joins them.
- In particular, we want this explanation model to be simple like our linear regression

\[ g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i \]
QII: Quantitative Input Influence

- Query definition
- Output comparison
- Summarization

Individual + Aggregate model prediction analysis
Exercises

LIME and SHAP: Colab notebook

NLP with Trulens: Colab notebook