CS 329T: Trustworthy Machine Learning

Lab 2

Outline

- HW 1 common queries (Due tomorrow at 11:59 pm)
- Final Report overview
- LIME and SHAP colab
- Explanations Survey Paper

HW 1 Queries?

Final report overview

- Individual project
- Pick a model, dataset or application that interests you!
- Try picking a fairly large and complex model like DNNs, large Gradient Boosted trees etc.
- Evaluate the trustworthiness of your model from explainability, fairness, privacy and robustness POV!
- The idea is to apply your learnings over the weeks on a single model and present your findings.
- There won’t be mid-point check-ins/milestones.

Final report overview (contd)

Example questions to ask your model regarding explainability:

- evaluating global drivers of model feature importance
- evaluating local drivers of individual decisions
- comparing different evaluation techniques on the model and seeing whether explanations differ?

LIME: Local Interpretable Model-agnostic Explanations

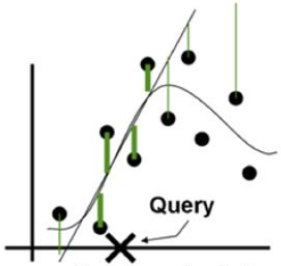
The key intuition behind LIME is that it is much easier to approximate a black-box model by a simple model *locally* (in the neighborhood of the prediction we want to explain), as opposed to trying to approximate a model globally.

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(tree frog) = 0.54

Perturbed Instances	P(tree frog)
	0.85
	0.00001
	0.52



Locally weighted regression

Explanation

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

$$g(\mathbf{z}') = \phi_0 + \sum_{i=1}^{M} \phi_i z_i'$$

LIME and SHAP Colab

Colab Link
SHAP GitHub