Josh Cubero
SEAS6414_HW7_Cubero

1. Why is interpretability important in machine learning models?

A machine learning model uses an objective function to minimize the distance between the actual and predicted outcomes. Because an objective function is a mathematical object, it is not able to capture the real-world impacts of the model's predictions. Many ML models can have ethical or fairness implications that are not easily understood by the objective functions. Thus, the ML community needs to implement interpretability when model metrics do not explain their outcomes. Therefore, interpretability enables the ML community to understand information about the ML model such as: what the model is learning, any additional information the model can provide, justifications about the outcomes, and to evaluate these outcomes with respect to the target problem.

2. What methods are used to make machine learning models interpretable?

Local Interpretable Modle-Agnostic Explanations (LIME)

LIME, an algorithm, seeks to fit a model that is interpretable and trusted, then compares that to the output of the original model. A LIME model is said to be understandable because is uses a separate local model to essentially cross-validate the production model by matching f(*x*) when its input is close to x. Additionally, LIME is model-agnostic, requiring only a black-box to query f(*x*). LIME functions by perturbing data with slight adjustments to features, then using those features to imitate the original model's behavior.

Shapley Additive exPlanations (SHAP)

According to the authors, SHAP is a natural extension of LIME and is based on a game theory that enables local and global interpretation of machine learning models. SHAP functions by calculating a SHAP value, or contribution feature among all relevant features. SHAP then calculates the difference between the model's predictions and the baseline prediction. Then, the model ensures fairness, employing the SHAP values derived from the game theory. Thus, SHAP ensures fairness through distributing feature importance and consistency by ensuring that feature importance remains consistent with the model's behavior. Lastly, SHAP explains the model's prediction by dividing it into intuitive explanations.

The Olah Method

The Olah Method primarily deals with neural networks and attempting to understand and explain the neural networks outcomes. There are several important aspects of the Olah Method: feature visualization, attribution, and dimensionality reduction. Olah uses feature visualization to observe the features detected by the neural network at various layers. Next, the method uses attribution to assign importance scores to the input features, aiding in understanding the most impactful inputs affecting the neural network's outputs. Furthermore, the Olah Method employs dimensionality reduction to reduce the model's complexity by projecting its data points from a higher dimension to a lower dimension. The Olah Method has a major drawback. However, the method focuses heavily on complex aspects of neurons and neural networks. The method's complexity makes it suboptimal for many ML practitioners.

3. How does interpretability help with building trust in machine learning models?

According to Huang, et. al, for humans to trust a model, the model must be interpretable. However, both trust and interpretability are difficult to define. Trust really boils down to the comfort level we have with taking the model and deploying that model to production. In doing, regarding trust, we'd naturally feel more comfortable with deploying an interpretable model, particularly when the model serves critical outcomes such as finance or medicine. Taking a more granular look at trust, the idea comes from the notion that if a human makes a mistake AND the AI makes a mistake, that's seemingly more acceptable. However, trust in the model breaks down when the AI makes a mistake that the human does not make. And lastly, humans expect that the model continues to perform well even though divergence between the training environment and the deployment environment exists.