Model Agnostic Interpretability Techniques

Madhumita Sushil
Model Interpretability - What and why?

Understanding trained ML models and outputs for

- Error analysis
- Exploratory analysis
Interpretability vs. Accuracy

[Diagram showing a graph with axes for Interpretability and Accuracy, plotting different machine learning models such as Linear Regression, Decision Tree, K-Nearest Neighbors, Random Forest, Support Vector Machines, and Neural Nets.]

https://medium.com/ansaro-blog/interpreting-machine-learning-models-1234d735d6c9
Interpretability vs. Accuracy

https://medium.com/ansaro-blog/interpreting-machine-learning-models-1234d735d6c9
Making Neural Nets Interpretable

Model Agnostic Techniques

- Feature Deletion
- Sensitivity Analysis
- Relevance propagation
- Information Theory
- Explanation Model
Feature Deletion

Evaluating output change after masking a feature
Feature Deletion

Evaluating output change after masking a feature

Drawbacks:

- Model may learn completely different patterns
- Feature co-occurrence effects not modeled
Sensitivity Analysis

Gradient (absolute/squared) of output with respect to input
Sensitivity Analysis

Gradient (absolute/squared) of output with respect to input

**Drawback:** Accounts for variations in input instead of the actual input
Sensitivity Analysis

Gradient (absolute/squared) of output with respect to input

**Drawback:** Accounts for variations in input instead of the actual input

**Solution:** Gradient*input (saliency)
Layer-wise relevance propagation

Backpropagate output scores to input

- Deep Taylor decomposition
- Change in output w.r.t reference input
Layer-wise relevance propagation

Backpropagate output scores to input

- Deep Taylor decomposition
- Change in output w.r.t reference input (DeepLIFT)

**Drawback:** Calculation dependent on reference input value: 0, or user-selected.
Information Theory

Maximize mutual information between feature subset and output

**Drawback:** Feature subset size predetermined.

**Solution:** Tune the subset size
Explanation Model

Learning separate local explanation model

Linear model fit on feature subset to predict original model output

- LIME
- SHAP: combines the properties of LIME, relevance propagation and shapely value estimation
Explanation Model

Learning separate local explanation model

Linear model fit on feature subset to predict original model output

- LIME
- SHAP: combines the properties of LIME, relevance propagation and shapely value estimation

Drawbacks:

- Very expensive
- Does not estimate feature importance from the original model
Comparative performance

Sensitivity analysis and relevance propagation have comparative results

Information theoretic approach seems to work well

SHAP has better scores than LIME
Thank You!