CS6462

Probabilistic and Explainable AI

Lesson 28

Post-Hoc Explainability in ML Models

by Emil Vassev

April 7, 2025

Intrinsic vs. Post-Hoc XAI

Post-hoc model:

- creating ML models that are inherently interpretable – tedious and difficult task
- a second (post-hoc) model - created to explain the first black-box model

Intrinsic vs. post-hoc:

- one of the main criteria - used for distinguishing whether interpretability is achieved through constraints imposed on the complexity of the ML model (intrinsic) or by applying methods that analyze the model after training (post-hoc)

Post-hoc explanation:

- goal - make the user understand the predictions of ML models, which is achieved through explanations
- explanation method (explainer) - generates explanations (e.g, linear models, shallow decision trees, visualizations, etc.)
- model-agnostic in nature - not tied to a particular type of ML model and separates prediction from explanation

Approaches to Post-Hoc Explanations

Local explanations:

- focus on data and provide individual explanations
- provide trust to model outcomes
- methods:
  - feature importance
  - rule-based
  - salience maps
  - prototype-based
  - counterfactuals

Global explanations:

- focus on the model
- explain the decision process, i.e., the mechanism behind the ML model
- methods:
  - representation-based
  - model distillation
  - summaries of counterfactuals
  - collection of local explanations

Local Post-Hoc Explanations

Feature Importance methods:

- identify important dimensions and present their relative importance
- assign to each feature an importance value, which represents the importance of a particular feature for the prediction result
- Lime – Feature Importance method:
  - makes a mini-dataset of perturbations along with the effects each one has on the classification output
  - we can train a sparse linear regression on this dataset:
    - goal: determine the most important parts of the input that made the classifier produce its output
  - the less important areas are greyed out
  - the final output from the regressor is our explanation - it helps us to understand:
    - the features based on which the output is generated
    - if the output is based on wrong features

Local Post-Hoc Explanations (cont.)

Rule-Based methods:

- explicitly state the decision support system's decision boundary between the given and the opposite advices, which can be stated in a rule-based format, .e.g.:

  if … then … else ….

- find sufficient conditions for the prediction to stay intact - if those conditions stay the same, the output will remain intact
- Anchors – Rule-Based method:
  - provides rules on which the decision is based
  - allows us to understand what to do if we need to change the output
- LORE – Rule-Based method:
  - learns a local interpretable predictor on a synthetic neighborhood generated by a genetic algorithm
  - derives from the logic of the local interpretable predictor a meaningful explanation consisting of a decision rule and a set of counterfactual rules

Local Post-Hoc Explanations (cont.)

Salience Maps methods:

- explain what parts of the input are most relevant for the model's prediction
- generally used with image or video processing ML models
- Input Gradients – Salience Maps method:
  - does not need any instrumentation of the network
  - gradients can be computed using calls to the gradient operation
  - heatmaps are used to visualize gradients - can be visually noisy and difficult to interpret
  - types: Smooth Gradients - average input-gradients, Integrated Gradients - compute a path integral from a baseline all the way to the input that we want to explain
- Layer-Wise Relevance Propagation – Salience Maps method:
  - assumes that the classifier can be decomposed into several layers of computation
  - layers can be parts of the feature extraction from the image or parts of a classification algorithm run on the computed features

Local Post-Hoc Explanations (cont.)

Prototype-Based methods:

- a prototype is an explainer representing a set of similar records that the user can easily understand and appreciate the similarity to other validation methods
- explain a model with synthetic or natural input examples
- help to gain insights into the kind of input the model is most likely to misclassify, identify the input examples that are mislabeled, and the kind of input that activates an internal neuron
- Prototype Selection – Prototype-Based method:
  - a set of prototypes for a class is built to capture the full structure of the training examples of that class while taking into consideration the structure of other classes
- TracIn – Prototype-Based method:
  - computes the influence of a training example on a prediction made by the model by tracing how the loss on the test point changes during the training process whenever the training example of interest was utilized

Local Post-Hoc Explanations (cont.)

Counterfactuals methods:

- explanations that provide a link between what could have happened had the input to a model been changed in a particular way
- capture what features need to be changed (and by how much) so to flip a model's prediction, i.e., to reverse an unfavorable outcome
- Minimum Distance Counterfactuals – Counterfactuals method:
  - the choice of the distance metric dictates what kind of counterfactuals are to be chosen
- DiCE – Counterfactuals method:
  - generates and evaluates a diverse set of counterfactual explanations based on determinantal point processes
- FACE – Counterfactuals method:
  - generates counterfactuals that are coherent with the underlying data distribution and supported by the “feasible paths” of change, which are achievable and can be tailored to the problem at hand

Global Post-Hoc Explanations

Representation-Based methods:

- derive model understanding by analyzing intermediate representations and determine model's reliance on 'concepts' that are semantically meaningful to humans
- Network Dissection – Representation-Based method:
  - quantifies the interpretability of latent representations of CNNs by evaluating the alignment between individual hidden units and a set of semantic concepts
- Compositional Explanation – Representation-Based method:
  - automatically explains logical and perceptual abstractions encoded by individual neurons in deep networks
  - generates explanations by searching for logical forms defined by a set of composition operators over primitive concepts

Global Post-Hoc Explanations (cont.)

Model Distillations methods:

- leverage model distillations to learn feature shapes that describe the relationship between input features and model predictions
- LGAE – Model Distillations method:
  - leverages model distillation to learn global additive explanations that describe the relationship between input features and model predictions
  - global explanations take the form of feature shapes, which are more expressive than feature attributions
- Decision Trees – Model Distillations method:
  - generates new training data by actively sampling new inputs and labeling them using the complex model
  - nonparametric explainer

Global Post-Hoc Explanations (cont.)

Summaries of Counterfactuals methods:

- construct global counterfactuals explanations that provide an interpretable and accurate summary of resources for the entire population
- AReS – Summaries of Counterfactuals method:
  - constructs global counterfactual explanations which provide an interpretable and accurate summary of recourses for the entire population

Collection of Local Explanations methods:

- select a subset of **k** local explanations to constitute a global explanation after generating a local explanation for every data instance by using one of the Local Post-Hoc methods
- questions to answer:
  - Which local explanation method shall we use?
  - How shall we select a set of representative local explanations and how large is the set?
  - How shall we combine the local explanations to provide the global view of the model?

Summary

Post-Hoc Explainability in ML Models

Intrinsic vs. Post-Hoc XAI

Approaches to Post-Hoc Explanations

Local Post-Hoc Explanations

- feature importance
- rule-based
- salience maps
- prototype-based
- counterfactuals

Global Post-Hoc Explanations

- representation-based
- model distillation
- summaries of counterfactuals
- collection of local explanations

Next Lesson:

- Post-Hoc Explainability in Shallow ML Models and Deep Learning

Thank You!

Questions?