CS6462

Probabilistic and Explainable AI

Lesson 29

Post-Hoc Explainability in Shallow ML Models and Deep Learning

by Emil Vassev

April 7, 2025

Shallow ML

Specifics:

- feature extraction in Shallow ML is a manual process that requires domain knowledge of the data that we are learning from
- we learn from data described by pre-defined features
- covers a diversity of supervised learning models:
  - group A: strictly interpretable (transparent) approaches (e.g. KNN and Decision Trees)
  - group B: shallow ML models that rely on more sophisticated learning algorithms - require additional layers of explanation
- popular shallow ML models are Tree Ensembles and Support Vector Machines:
  - require the adoption of post-hoc explainability techniques for explaining their decisions
  - notable performance in predictive tasks

Post-Hoc Explainability in Shallow ML

Tree Ensembles:

- among the most accurate ML models in use
- means to improve the generalization capability of single decision trees - usually prone to overfitting
- combine different trees to obtain an aggregated prediction/regression:
  - effective against overfitting
  - the combination of models makes the interpretation of the overall ensemble more complex than each of its compounding tree learners – requires post-hoc explainability techniques
  - explainability techniques - explanation by simplification and feature relevance techniques
- post-hoc explainability goal: simplify tree ensembles while maintaining part of the accuracy accounted for the added complexity

Post-Hoc Explainability in Shallow ML (cont.)

Tree Ensembles (cont.):

- simplification approaches:
    1) train a single albeit less complex model from a set of random samples from the data (ideally following the real data distribution) labeled by the ensemble model
    2) create a Simplified Tree Ensemble Learner (STEL)
    3) use two models: simple and complex, where the former oversees the interpretation, and the *latter* takes care of the prediction by means of expectation-maximization
    4) Feature Relevance techniques - analyze the variable importance within Random Forests
    5) crosswise technique - proposes a framework that possesses recommendations that, if taken, would convert an example from one class to another
    6) stacking classifiers - compounding learner of the ensemble produces a specific prediction on a given data, and contributes to the output of the ensemble

Post-Hoc Explainability in Shallow ML (cont.)

Support Vector Machines:

- among the most used ML models due to their excellent prediction and generalization capabilities
- more complex than Tree Ensembles, with a much opaquer structure
- construct a hyper-plane or set of hyper-planes in a high or infinite-dimensional space, which can be used for classification, regression, or other tasks such as outlier detection
  - goal is to find the decision boundary to separate different classes and maximize the margin
  - a good separation is achieved by the hyperplane that has the largest distance (or functional margin) to the nearest training-data point of any class: the larger the margin, the lower the generalization error of the classifier
- post-hoc explainability techniques:
  - relate what is mathematically described internally in SVMs
  - cover explanation by simplification, local explanations, visualizations and explanations by example

Post-Hoc Explainability in Shallow ML (cont.)

Support Vector Machines (cont.):

- explanation by simplification:
  - technique to build rule-based models only from the support vectors of a trained model - extract rules directly from the support vectors of a trained SVM using a modified sequential covering algorithm
  - four classes of simplification, differentiate by how deep they go into the algorithm's inner structure:
    - 1st class of simplification – generates fuzzy rules instead of classical propositional rules - long antecedents reduce comprehensibility, hence, a fuzzy approach allows for a more linguistically understandable result.
    - 2nd class of simplification – adds SVM's hyperplane, along with the support vectors, to the components in charge of creating the rules; relies on the creation of hyper-rectangles from the intersections between the support vectors and the hyper-plane
    - 3rd class of simplification – adds the training data as a component for building the rules; clustering methods are used to group prototype vectors for each class
    - 4th class of simplification – uses a growing support vector classification to give an interpretation to SVM decisions in terms of linear rules

Post-Hoc Explainability in Shallow ML (cont.)

Support Vector Machines (cont.):

- explanation by visualization – three approaches:
  - Approach 1: Support Vector Regression models are visualized as trained SVMs to extract the information content from the kernel matrix – focus is on input variables related with associated output data
  - Approach 2: a visual way combines the output of the SVMs with heatmaps to guide the input modifications; colors are assigned based on the weights of a trained linear SVM - allows for a more comprehensive way of debugging the training process
  - Approach 3: interpreting SVMs through weight vectors and statistical analysis that explicitly accounts for the SVM margin

Post-Hoc Explainability in Deep Learning

Multi-Layer Neural Networks:

- multi-layer neural networks (multi-layer perceptrons) – able to infer complex relations among variables
- with questionable explainability - neural networks are considered as black-box models
- explainability is needed, so the model can have a practical value
- multiple explainability techniques: *model simplification approaches*, *feature relevance estimators*, *text explanations*, *local explanations* and *model visualizations*
- simplification techniques for neural networks with one single hidden layer
- lack of simplification techniques for neural networks with multiple hidden layers - DeepRED algorithm extends the de-compositional approach to rule extraction for multi-layer neural network by adding more decision trees and rules

Post-Hoc Explainability in Deep Learning (cont.)

Multi-Layer Neural Networks (cont.):

- model simplification as a post-hoc explainability approach:
  - Approach 1: Interpretable Mimic Learning - simple distillation method that extracts an interpretable model by means of gradient boosting trees
  - Approach 2: hierarchical partitioning of the feature space that reveals the iterative rejection of unlikely class labels, until association is predicted
  - Approach 3: distillation of knowledge from an ensemble of models into a single model
  - Approach 4: explains multi-layer neural networks by feature relevance – tackles the problem that the simplification of multi-layer neural networks is more complex as the number of layers increases;
  - Approach 5: Deep Taylor Decomposition - decomposes a network classification decision into contributions of its input elements - neurons are considered as objects that can be decomposed and expanded, and then these decompositions are aggregated and back-propagated through the network
  - Approach 6: computes importance scores in a multi-layer neural network to compare the activation of a neuron to the reference activation

Post-Hoc Explainability in Deep Learning (cont.)

Convolutional Neural Networks:

- constitute the state-of-art models in all fundamental computer vision tasks, from image classification and object detection to instance segmentation
  - built as a sequence of convolutional layers and pooling layers to automatically learn increasingly higher-level features
  - at the end of the sequence, one or multiple fully connected layers are used to map the output features map into scores
- the CNN’s structure - extremely complex internal relations that are very difficult to explain
- explainability of CNNs divided into two broad categories:
  1) Category 1: understanding the decision process by mapping the output to the input space to see which parts of the input are discriminative for the output
  2) Category 2: delve inside the network and interpret how the intermediate layers see the external world, not necessarily related to any specific input

Post-Hoc Explainability in Deep Learning (cont.)

Convolutional Neural Networks (cont.):

- Deconvnet (explainable method of Category 1):
  - an input image runs feed-forward through a CNN, so each layer can output a number of feature maps with strong and soft activations
  - a feature map from a selected layer is used to reconstruct the maximum activations - can give an idea about the parts of the image that have been used to produce that effect
  - Occlusion Sensitivity Method is used to generate a salience map - consists of iteratively-forwarded image through the network to occlude a different region at every iteration
- Deep Generator Network (explainable method of Category 2):
  - analyzes the visual information contained inside a CNN
  - reconstructs an image from the CNN internal representations to show that several layers retain photographically accurate information about the image, with different degrees of geometric and photometric invariance
  - visualizes the notion of a class captured by a CNN via an image for a given output neuron in a CNN

Post-Hoc Explainability in Deep Learning (cont.)

Recurrent Neural Networks:

- used extensively for predictive problems defined over inherently-sequential data - natural language processing and time series analysis
- able to retrieve time-dependent relationships by formulating the retention of knowledge in the neuron as another parametric characteristic that can be learned from data
- explainability of RNN (two categories):
    - Category 1: explainability by understanding what an RNN model has learned (mainly via feature relevance methods)
    - Category 2: explainability by modifying RNN architectures to provide insights about the decisions they make (local explanations)

Post-Hoc Explainability in Deep Learning (cont.)

Recurrent Neural Networks (cont.):

- explainable method of Group 1:
  - extends the interpretable mimic learning distillation method used for CNN models to LSTM (Long Short Term Memory) networks - interpretable features are learned by fitting Gradient Boosting Trees to a trained LSTM network
  - specific propagation rule that works with multiplicative connections as those in LSTM units and Gated Recurrent Units (GRU)
  - visualization technique based on finite horizon n-grams that discriminates interpretable cells within LSTM and GRU networks
- RETAIN (REverse Time AttentIoN) - explainable method of Group 2:
  - detects influential past patterns by using a two-level neural attention model
  - creates an interpretable RNN based on SISTA (Sequential Iterative Soft-Thresholding Algorithm) to model a sequence of correlated observations with a sequence of sparse latent vectors – the network weights are interpreted as parameters of a principled statistical model

Post-Hoc Explainability in Deep Learning (cont.)

Hybrid Transparent Methods:

- data fusion approaches endowing deep learning models with explainability provided by other domain-information sources
- explainability is improved via background knowledge used in the form of logical statements or constraints provided by Knowledge Bases (KBs)
- provide a hybrid approach that provides robustness to the learning system when errors are present in the training data labels
- able to jointly learn and reason with both symbolic and sub-symbolic representations and inference
- allow for expressive probabilistic reasoning in an end-to-end fashion
- use case - dietary recommendations, where explanations are extracted from the reasoning behind, i.e., non-deep but KB-based models

Summary

Post-Hoc Explainability in Shallow ML Models and Deep Learning

Shallow ML

Post-Hoc Explainability in Shallow ML

- *Tree Ensembles*
- *Support Vector Machines*

Post-Hoc Explainability in Deep Learning

- *Multi-Layer Neural Networks*
- *Convolutional Neural Networks*
- *Recurrent Neural Networks*
- *Hybrid Transparent Methods*

Thank You!

Questions?