CS6462

Probabilistic and Explainable AI

Lesson 27

Transparent Machine Learning

by Emil Vassev

April 7, 2025

Transparency in Machine Learning

Machine learning needs to be transparent:

- machine learning takes rational decision making to a whole new level
- rigorous mathematical models empower machine learning algorithms - often way more complicated than any human mind can comprehend
- transparency:
  - the most apparent challenge of making critical decisions with machine learning
  - critical decisions should not be made by and within a black box ML
  - without transparency, ML decisions cannot be justified along with their ultimate consequences on trust, fairness, privacy, and security

What is transparency in ML?

- clear distinction between “know how” (understand an action) and “know that” (understand a concept)
- a learning model is truly transparent only when we know both “how” and “that”

Model Transparency – "Know How"

The problem of "know how":

- searches for a set of decision objects that is cognitively fluent for human to follow
- set of decision objects serves as a transparent substitute for the original complex and possibly black-box decision model

Extensive research:

- increase the interpretability of different types of classification models
- use decision lists to simplify a high-dimensional, multivariate feature space
- assessment of the performance of classification models from a user perspective in terms of accuracy, comprehensibility, and justifiability
- examine different types of explanations for improving transparency of rule-based systems
- sparse linear model (LIME) for local exploration—providing interpretable representation that is locally faithful to the classifier
- Quantitative Input Influence (QII) – measures the influence of the inputs of a decision-making model on its output

Model Transparency – "Know How" (cont.)

Mathematical model of "know how":

- $H \; = < D, X, Y >$ – decision model built on D; takes inputs X to produce decision outputs Y
- D – data collection
- X – set of inputs (random variables) accepted by H
- Y – set of decision outputs (observables) produced by H
- $T \; =< H, R >$ – transparent model built from H by considering a set of R
- R – interpretable decision-making objects (e.g., decision rules, local linear models, etc.)

requirements for T:

- consistency: the set of decision objects R must be consistent with all outputs Y of H
- coverage: all inputs X of H must be covered by T:
  - an input $x \in X$ is covered by T if there is at least one $r \in R$ that can be applied to x
  - a transparent model is said to have a p-good coverage for any given x iff x is covered by T with a probability of $p(x)$ – the greater probability, the greater the coverage

Model Transparency – “Know How” (cont.)

Mathematical model of “know how” (cont.):

- $R = \{r_1, r_2, \dots, r_n\}$ – given a transparent model $T =< H, R >$
- given an input $x \in X$, the decision output $y \in Y$ by H can be mapped to a subset of R with a probability $p(x)$
- $F_R(R, x)$ - function that returns a Boolean vector that indicates which decision objects in R can be applied to a given input x and $0 \le F_R\ (R_s, x) \le 1$ provides the probability $p(x)$
  - if $R_s \subseteq R$, $F_R(R_s, x) = 1$ (True) then x is fully covered by R_s
  - else x is not fully covered by R_s
- for $y_i \in Y, x_i \in X,\ y_i = h(x_i)$ (the decision made by H for the input x, the general optimization problem is:

$$\min_{R_s \subseteq R} = \sum_{i=1}^{n} \mathcal{L}(T(R_s, x_i), y_i)$$

$$subject\ to: \sum_{i=1}^{n} F_R(R_s, x_i) \ge \lambda n$$

$\mathcal{L}$ - the loss function that measures the inconsistency between T and H for a given x

λn – minimum acceptable probability

Model Transparency – "Know That"

The problem of "know that":

- know-that is concerned with gaining more in-depth understanding of the internal justification of the decisions through external constraints on privacy, reliability, and fairness
- an esoteric exercise – may require long-term training
- a learning model can provide the insights into "know that" by the justification of decisions that can be gauged externally

Privacy constraints:

- model transparency can be both beneficial and pernicious: greater transparency in the decision processes:
  - can help users better understand how decisions are reached inside a ML model
  - may introduce biases and privacy/security risks

Model Transparency – “Know That” (cont.)

Reliability constraints:

- concerned with the robustness of a ML model when it faces adversarial attacks
- standard ML techniques are susceptible to adversarial attacks
- *evasion attack*: one of the important lines of attack against standard ML techniques
    - consists of carefully perturbing the input samples at test time to have them misclassified

Fairness constraints:

- concerned with whether a transparent model is “fair” for a protected or sensitive group
- two questions to consider:
    1) Is the bias in the original decision model transferable to its transparent counterpart?
    2) Is there a trade-off between transparency and fairness?

Transparent Models

Linear\Logistic Regression models:

- class of models used for predicting continuous and categorical targets under the assumption that these targets are a linear combination of the predictor variables
- allow us to view models as a transparent method
- explainability depends on the simplicity of the regression model – complexity is exponential to the number of variables and inter-variable relationships
- usually satisfy the transparency criteria
- may benefit from post-hoc explainability approaches (e.g., as visualization) – e.g., when a non-expert audience needs to get a better understanding of the models' intrinsic reasoning
- have been largely applied within Social Sciences
- to maintain transparency, model size (number of variables and complexity of their relationships) must be limited, and the variables must be understandable

Transparent Models

Decision Trees:

- contain a set of conditional control statements
- nodes are arranged in a hierarchical manner:
  - intermediate nodes represent decisions
  - leaf nodes can be either class labels (for classification problems) or continuous quantities (for regression problems)
- fall into the level of *simulatability transparency* models: iff have a small number of features - the number is not that long, so it can be processed by a human
- if the models' number of features does not allow simulating, but the features are still understandable by a human user: the models are not *simulatability transparency* models, but *decomposable* models
- fall into the level of *algorithmic transparency* models: if the models utilize complex feature relationships as part of their optimization algorithms

Transparent Models

K-Nearest Neighbors (KNN):

- deal with classification problems in a simple and straightforward way - predict the class of a new data point by inspecting the classes of its K nearest neighbors

- neighborhood relation is induced by a measure of the distance between data points

- capable of satisfying any level of transparency, which depends heavily on:
  - the distance function that is employed
  - the model’s size
  - the features’ complexity

Transparent Models

Rule-Based Learning:

- built on the intuitive basis of producing rules to describe how a model generates its outputs
- complexity of rules ranges from simple “if-else” expressions to fuzzy rules, or propositional rules encoding complex relationships between variables
- level of transparency depends on some designing aspects, such as the level of coverage and the specificity of the generated rules
- systems with a very large number of rules are infeasible to be simulated by humans
- rules may contain an unacceptable number of antecedents or consequents, including cumbersome features in the rules

Transparent Models

Generalized Additive Models (GAMs):

- a class of linear models where the outcome is a linear combination of some functions of the input features
- goal: infer the form of the unknown linear functions – this form may belong to a parametric family (e.g., as polynomials), or it could be defined non-parametrical
- allow for a large degree of flexibility where some applications might:
  - take the form of a simple function
  - be handcrafted to represent background knowledge
  - be specified by simple properties (e.g., being smooth)
- satisfy the requirements for being algorithmic transparent
- some could also be considered as of being at the *simulatability* level of transparency: in applications with low dimension of the problem of optimization

Transparent Models

Bayesian networks (recall):

- refer to a designing approach where the probabilistic relationships between variables are explicitly represented using a DAG
- explainability features:
  - clear characterization of the connections among the variables
  - graphical criteria that examines probabilistic relationships by only inspecting the graphs topology
  - probabilistic semantics, which allows conditioning and interventions, allows researchers to look for ways of augmenting directed graphical models
- used extensively in a wide range of applications
- might be used to construct explanatory arguments from Bayesian models, where explanations are produced in order to assess the trustworthiness of a model

Summary

Transparent Machine Learning

Transparency in Machine Learning

Model Transparency "Know How"

Model Transparency "Know That"

Transparent Models

Next Lesson:

- Post-Hoc Explainability in Shallow ML Models

Thank You!

Questions?