Attribution Scores in Explainable AI

Leopoldo Bertossi
FIAD, Santiago, Chile

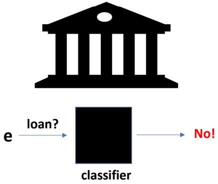
www.scs.carleton.ca/∼bertossi

Explanations in Machine Learning

- Bank client $\mathbf{e} = \langle \text{john}, 18, \text{plumber}, \text{70K}, \text{harlem}, \ldots \rangle$

  As an entity represented as a record of values for features Name, Age, Activity, Income, ...

- $\mathbf{e}$ requests a loan from a bank that uses a classifier



- The client asks *Why?*

- What kind of *explanation?*

  How?

  From what?

- Some of them are *causal explanations*, some are *explanation scores* a.k.a. *attribution scores*
- They quantify the relevance of each feature value in **e** for the assigned label
- Here two of them:
  - *Shap* (based on Shapley value of Coalition Game Theory)
  - *Resp* (Responsibility, based on Actual Causality)
- We will consider only binary features and a binary classifier

  Entity population $\mathcal{E} = \{0, 1\}^N$

  Classifier $L\colon\ \mathcal{E} \rightarrow \{0, 1\}$

Shap **Score**

- Set of players $\mathcal{F}$ contain features, relative to classified entity $\mathbf{e}$
- An appropriate $\mathbf{e}$-dependent game function (shared wealth-function) mapping subsets of players to real numbers
- For $S \subseteq \mathcal{F}$, and $\mathbf{e}_S$ the projection of $\mathbf{e}$ on S:

$$\mathcal{G}_{\mathbf{e}}(S) \ := \ \mathbb{E}(\ L(\mathbf{e}'\ ) \ \mid \ \mathbf{e}' \in \mathcal{E} \ \text{ and } \ \mathbf{e}'_S = \mathbf{e}_S)$$

- For a feature $F^\star \in \mathcal{F}$, compute: $\mathit{Shap}(\mathcal{F}, \mathcal{G}_{\mathbf{e}}, F^\star)$

$$\sum_{S \subseteq \mathcal{F} \setminus \{F^\star\}} \frac{|S|!(|\mathcal{F}|-|S|-1)!}{|\mathcal{F}|!} [\underbrace{\mathbb{E}(L(\mathbf{e}'|\mathbf{e}'_{S\cup\{F^\star\}} = \mathbf{e}_{S\cup\{F^\star\}})}_{\mathcal{G}_{\mathbf{e}}(S\cup\{F^\star\})} - \underbrace{\mathbb{E}(L(\mathbf{e}')|\mathbf{e}'_S = \mathbf{e}_S)}_{\mathcal{G}_{\mathbf{e}}(S)}]$$

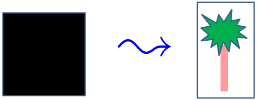
(Lee & Lundberg, 2017)

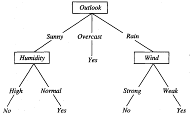
- Assumes a probability distribution on entity population $\mathcal{E}$

- *Shap*: Exponentially many subsets of players, and multiple passes through a possibly black-box classifier

  Shap computation is #P-hard in general

- Can we do better with an open-box classifier?





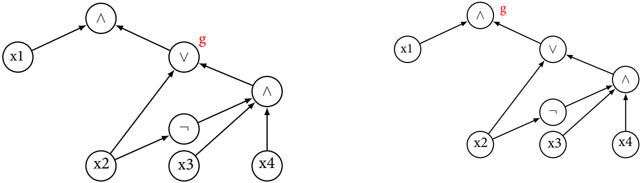
  Exploiting its elements and internal structure?

- A decision tree, or a random forest, or a Boolean circuit?

- Can we compute *Shap* in polynomial time?

Tractability for BC-Classifiers

- Theorem: *Shap* can be computed in polynomial time for dDBCs under the uniform distribution[1]



Deterministic and Decomposable Boolean Circuit

- Can be extended to a product distribution on $\mathcal{E} = \{0, 1\}^N$
- They (and related models) are relevant in *Knowledge Compilation*

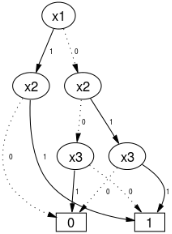
[1] Arenas, Bertossi, Barcelo, Monet; AAAI'21; JMLR'23

- Corollary: Via polynomial time transformations, under the uniform and product distributions, *Shap* can be computed in polynomial time for
  - Decision trees (and random forests)
  - Ordered binary decision diagrams (OBDDs)

    $(\neg x_1 \wedge \neg x_2 \wedge \neg x_3) \vee (x_1 \wedge x_2) \vee (x_2 \wedge x_3)$

    Compatible variable orders along full paths

    Compact representation of Boolean formulas

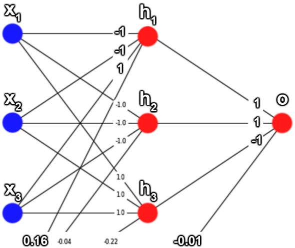


  - Sentential decision diagrams (SDDs)
    Generalization of OBDDs
  - Deterministic-decomposable negation normal-form (dDNNFs)
    As dDBC, with negations affecting only input variables
- An optimized efficient algorithm for *Shap* computation can be applied to all of these

Shap on Neural Networks

- Binary Neural Networks (BNNs) are commonly considered black-box models
- We experimented with *Shap* computation with a black-box BNN and with its compilation into a dDBC[2]
- Even if the compilation is not entirely of polynomial time, it may be worth performing this one-time computation
- Particularly if the target dDBC will be used multiple times, as is the case for explanations

[2] Bertossi, Leon; JELIA'23



$$\phi_g(\bar{i}) = sp(\bar{w}_g \bullet \bar{i} + b_g) := \begin{cases} 1 & \text{if } \bar{w}_g \bullet \bar{i} + b_g \geq 0, \\ -1 & \text{otherwise,} \end{cases}$$

$$\begin{aligned} o \longleftrightarrow\ & (-[(x_3 \wedge (x_2 \vee x_1)) \vee (x_2 \wedge x_1)] \wedge \\ & ([(-x_3 \wedge (-x_2 \vee -x_1)) \vee (-x_2 \wedge -x_1)] \vee \\ & [(x_3 \wedge (-x_2 \vee -x_1)) \vee (-x_2 \wedge -x_1)])) \vee \\ & ([(-x_3 \wedge (-x_2 \vee -x_1)) \vee (-x_2 \wedge -x_1)] \wedge \\ & [(x_3 \wedge (-x_2 \vee -x_1)) \vee (-x_2 \wedge -x_1)]). \end{aligned}$$

- BNN described by a propositional formula, which is further transformed into an optimized CNF
- Actually, done using always CNFs and keeping them "short" ...
  (room for optimizations)
- In CNF: $o \longleftrightarrow (-x_1 \vee -x_2) \wedge (-x_1 \vee -x_3) \wedge (-x_2 \vee -x_3)$

- The CNF is transformed into an SDD

  Succinctly representing the CNF

- The expensive compilation step

  But upper-bounded by an exponential only in the tree-width of the CNF

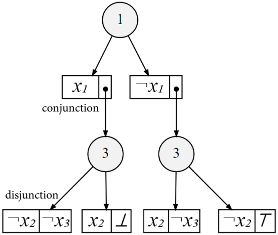
  TW of the associated undirected graph: an edge between variables if together in a clause

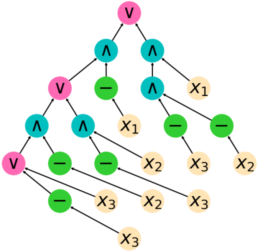
  (In example, graph is clique, TW is #vars -1 =2)

- The SDD is easily transformed into a dDBC

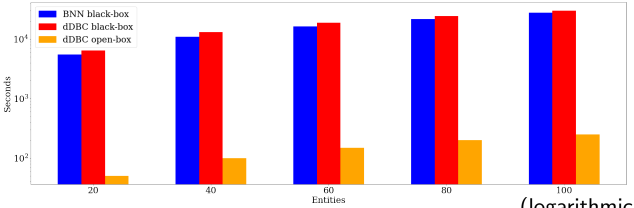
- *Shap* computed on it, possibly multiple times

- The uniform distribution was used





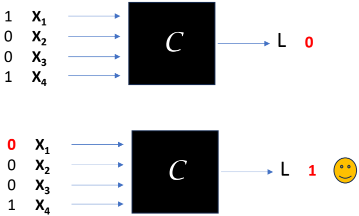
- In our experiments, we used a BNN with 14 gates
- Compiled into dDBC with 18,670 nodes (room for optimizations)
- A one-time computation that fully replaces the BNN
- Compared *Shap* computation time for: black-box BNN, open-box dDBC, and black-box dDBC
- Total time for computing *all Shap scores for all entities*, with increasing numbers of them



(logarithmic scale)

Resp: Causal Responsibility

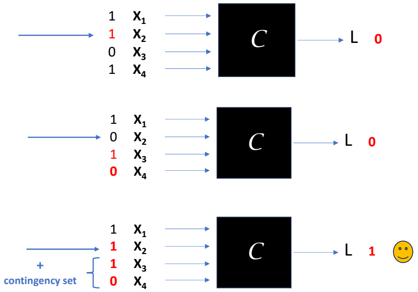
- Actual Causality is based on counterfactual interventions (Halpern & Pearl, 2001)
- Hypothetical changes of values in a causal model to detect other changes ... identifying then actual causes
- Do changes of feature values change the label from 0 to 1?
- A measure of causal contribution: Responsibility (Chockler & Halpern, 2004)



Counterfactual cause for $L = 0$: $\mathbf{x_1} = 1$

$\mathit{Resp}(\mathbf{x_1}) := 1$

maximum responsibility



concentrate on $\mathbf{x_2}$: not counterfactual cause

changes on $\mathbf{x_3}, \mathbf{x_4}$ do not change label

change on $\mathbf{x_2}$ *accompanied by* changes on $\mathbf{x_3}, \mathbf{x_4}$ does change label!

- $\Gamma = \{\mathbf{x_3}, \mathbf{x_4}\}$ is contingency set for $\mathbf{x_2}$
- $\mathbf{x_2}$ is actual cause for $L = 0$
- If Γ is minimum-size contingency set for $\mathbf{x_2}$:

$$Resp(\mathbf{x_2}) := \frac{1}{1+|\Gamma|} = \frac{1}{3}$$

- We call $\langle 1, 1, 1, 0 \rangle$ a counterfactual (version) of original entity

Final Remarks and Research Directions

- RESP has been generalized to deal with non-binary features

  Uses expected values for labels

- We have also defined and investigated the (probabilistic) Causal-Effect Score

  So far for Explainable Data Management

  We have uncovered and established categorical properties of the Causal-Effect Score

- Explanation scores commonly use the classifier plus a probability distribution over the underlying entity population

  Imposing or using explicit and additional domain semantics or domain knowledge is relevant to explore

  Can we modify *Shap*'s definition or computation accordingly?

  Or the probability distribution?

- The above results on *Shap* computation hold under the uniform and product distributions

  Other distributions?

  Do we still have efficient algorithms?

  Empirical and product-empirical distributions have been considered for *Shap* and other scores

- We have investigated the robustness of SHAP under distributional shifts (ECAI'24)

- Shapley values satisfy a categorical and desirable list of properties

  For the general context of coalition game theory

  Existing scores have been criticized or under-explored in terms of general properties

  Specific general and expected properties for Explanations Scores (in AI)?

- Reasoning about scores, explanations and counterfactuals is what intelligent agents do

  We have done research on the use of Answer-Set Programming for automating this task

- Higher-level analytics should be characterized, formalized and automated:

- Learning from attribution scores?

  About the application domain and/or the ML system

- What can I learn through aggregation of attribution scores?

- Defining and aggregating at higher levels of abstraction

  Categorizing features at a higher level:

  "Your entire socio-economic situation is to be blamed for the rejection of your loan application"

EXTRA PAGES

The Need for Reasoning

- Logical specification of counterfactual interventions and *Resp*
- Logical reasoning for interaction with attribution-score spec/algorithm and classifier for further exploration
- Reason about interventions and counterfactuals
- Compute responsibility scores, and reason about them
- Impose semantic constraints on counterfactuals
- Counterfactuals can be queried or reasoned about
  - Specification of actionable counterfactuals?
  - Some actionable high-score feature value?
  - Specs of other counterfactuals of interest? Computing them?
  - What if I leave some feature values fixed?
  - Do I get same high-score feature with this "similar" entity?
  - Any high-score counterfactual version that changes this feature?

    Or never changes that one? ETC.

- Usually interested in maximum-responsibility feature values (associated to minimum-cardinality contingency sets)
- We have used Answer-Set Programming (ASP)
  - Declarative language, and reasoning via QA
  - Possibly several answer-sets (models)
  - Each counterfactual version leading to a new label corresponds to an answer set (model)
  - Minimality of answer-sets, and closed-world assumption
  - Non-monotonicity, and commonsense reasoning (persistence)
  - Program and semantic constraints (the latter on counterfactuals)
  - Required expressive power and computational complexity
  - Weak constraints (useful for specifying minimum cardinalities)
  - Set and numerical aggregations (useful for score computation)
  - Predicates for interaction with external classifiers
  - Reasoning is enabled by cautious and brave query answering

    True in all models vs. true in some model

The Generalized *Resp* Score

- For binary features the previous version of RESP works fine
- There could be many values that do not change the label, but one of them does
- Better consider all possible values, average labels, and contingencies
- $\mathbf{e}$ entity under classification, with $L(\mathbf{e}) = 1$, and $F^\star \in \mathcal{F}$
- Local *Resp*-score

$$\mathit{Resp}(\mathbf{e}, F^\star, \mathcal{F}, \Gamma, \bar{w}) := \frac{L(\mathbf{e}') - \mathbb{E}[L(\mathbf{e}'') \mid \mathbf{e}''_{\mathcal{F} \smallsetminus \{F^\star\}} = \mathbf{e}'_{\mathcal{F} \smallsetminus \{F^\star\}}]}{1 + |\Gamma|} \quad (*)$$

- $\Gamma \subseteq \mathcal{F} \smallsetminus \{F^\star\}$
- $\mathbf{e}' := \mathbf{e}[\Gamma := \bar{w}] \qquad L(\mathbf{e}') = L(\mathbf{e})$
- $\mathbf{e}'' := \mathbf{e}[\Gamma := \bar{w}, F^\star := v]$, with $v \in \mathit{dom}(F^\star)$
- When $F^\star(\mathbf{e}) \neq v$, $L(\mathbf{e}'') \neq L(\mathbf{e})$, $F^\star(\mathbf{e})$ is *actual causal explanation* for $L(\mathbf{e}) = 1$ with contingency $\langle \Gamma, \mathbf{e}_\Gamma \rangle$
- Globally: $\mathit{Resp}(\mathbf{e}, F^\star) := \max_{\substack{|\Gamma| \ \mathit{min.},\ (*) > 0 \\ \langle \Gamma, \bar{w} \rangle}} \mathit{Resp}(\mathbf{e}, F^\star, \mathcal{F}, \Gamma, \bar{w})$

- Requires underlying probability space on entity population
- No need to access the internals of the classification model