



Carleton
UNIVERSITY



Millennium Institute
Foundational
Research on Data

Explanations in AI and Explainable AI: (Some of) The Classic and The New

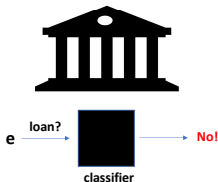
Prof. Leopoldo Bertossi

Explanations in Machine Learning

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

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

- e requests a loan from a bank, which uses a classifier



- The client asks *Why?*
- What kind of *explanation?*
How?
From what?

Explanations (in AI)

- Users and those affected by results from AI systems, the stakeholders, request explanations

Assessments (e.g. a credit score), classifications (good/bad client), decisions (approve/reject loan), etc.

- A whole new area of AI has emerged: *Explainable AI* (XAI)

A whole discipline has emerged: *Ethical AI*

- It touches Law, Sociology, Philosophy, ...
- Motivated by the need for more *transparent, trustable, fair, unbiased, ...* and *interpretable* AI systems



It may really be a “black box”!

- New legislation forces AI systems affecting users to provide explanations and guarantee all the above

- Search for explanations belongs to the nature of human beings
- The quest has been around since the inception of humans
- Ancient Greeks already concerned with *causes* (and effects)
- Studied as such by Philosophers, Logicians, Physicists, ...
- Are explanations a new subject in AI?
- Yes and No
- Explanations have been studied in AI for some decades by now, and in related disciplines, such as Logic, Statistics

Some forms of explanations are new in AI

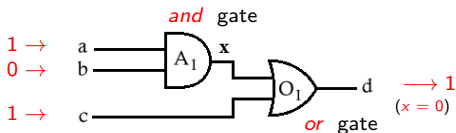
Others have roots in already existing ones

Model-Based Diagnosis

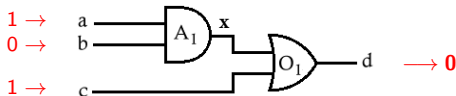
- MBD has been an area of AI for some time
- It is about doing a *diagnosis* of a system (exhibiting some unexpected behavior) using a model of the system (and possibly a bit more)

Example: A very simple Boolean circuit (a classifier?)

It should be:



However:

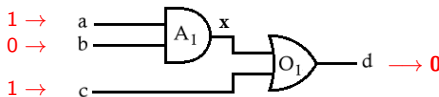


Why? What's wrong?

A diagnosis?

- What is a diagnosis?

We need a characterization ...



- A logical model of the *ideal circuit*:

$$\{(x \longleftrightarrow (a \wedge b)), (d \longleftrightarrow (x \vee c))\}$$

- The *observation* *Obs*: $a \wedge \neg b \wedge c \wedge \neg d$

- What can be get from the combination?

Logically?

Since the combination is inconsistent, **everything!**

Trivial, irrelevant, **useless conclusion** ...

- Need flexible **model that allows failures**: (a “weak model of failure”, specifying things under normality)

$$\mathcal{M} = \{\neg AbA \longrightarrow (x \leftrightarrow (a \wedge b)), \neg AbO \longrightarrow (d \leftrightarrow (x \vee c))\}$$

“when A is not abnormal, it works as an and gate”, etc.

Now gates could be abnormal (faulty)

- Now, $Obs \cup \mathcal{M}$ is consistent, but, as before:

$$Obs \cup \mathcal{M} \cup \{\neg AbA, \neg AbO\} \text{ is inconsistent} \quad (*)$$

- So, something has to be abnormal ...
- $D = \{abO\}$ is a diagnosis, because making gate O abnormal restores consistency

$$Obs \cup \mathcal{M} \cup \{\neg AbA, AbO\} \text{ is consistent} \quad (**)$$

Abnormality of gate O is an explanation for the malfunction of the circuit

- This can be seen as a binary classification problem:

$$\begin{array}{ccc}
 \mathbf{e} = \langle abA, abO, a, b, c, d \rangle & \longrightarrow & \begin{array}{c} SAT? \\ \boxed{\begin{array}{l} \neg abA \longrightarrow (x \leftrightarrow (a \wedge b)) \\ \neg abO \longrightarrow (d \leftrightarrow (x \vee c)) \\ + \text{ SAT solver} \end{array}} \longrightarrow yes/no \\
 \text{(propositional feature vector)} & &
 \end{array}$$

Hypothetical input changes to change the classification!

- $D' = \{abO, abA\}$ is a diagnosis, because making every gate abnormal restores consistency
 $Obs \cup \mathcal{M} \cup \{AbA, AbO\}$ is consistent (***)
- D is “better” than D' : fewer assumptions, narrower, more focused and informative
- This is *Consistency-Based Diagnosis* (CBD, Ray Reiter, 1987)
- Can we assign scores to diagnoses? (coming)

Causality and Responsibility as Explanations

- **Example:** Could be the initial circuit (malfunctioning or not), or any binary classifier for entities with binary feature values

- Feature values in $\mathbf{e} = \langle 1, 0, 0, 1 \rangle$ get *responsibility scores*, quantifying *how much* they contribute to the label

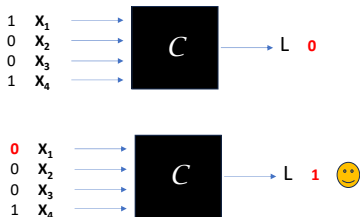


- These methods can be applied without necessarily knowing “the internals” of the classifier

The latter is (or is treated as) a “black box”

Only input/output relation is needed

- Given an entity $\mathbf{e}$ that gets a (binary) label, say 0
- Do changes of (interventions on) feature values change the label from 0 to 1?



- Change of $x_1 = 1$ changes label

$x_1 = 1$ is a **counterfactual cause** for $L = 0$

(J. Halpern & J. Pearl, 2001)

- Its **responsibility** is $Resp(x_1) := 1$

A change of it alone changes the label

The maximum responsibility

(Chokler & Halpern, 2004)

- **Counterfactuals**: hypothetical changes of (interventions on) variables' values, to see if something else changes

- Concentrate on x_2 :



Not a counterfactual cause

- But now:



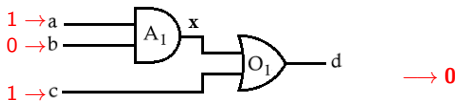
- Changes on x_3, x_4 do not change label

Change on x_2 *plus* changes on x_3, x_4 does change the label

- $x_2 = 0$ is an **actual cause** for $L = 0$ with $\Gamma = \{x_3, x_4\}$ as a **contingency set** (CS) for x_2
- If Γ is **minimum-size CS** for x_2 : $\text{Resp}(x_2) := \frac{1}{1+|\Gamma|} = \frac{1}{3}$
- We call $\langle 1, 1, 1, 0 \rangle$ a counterfactual (version) of original entity

Back to the Circuit

- We can see similarities between CBD and Actual Causality
Diagnosis problem can be cast as actual causality problem



Model: $\mathcal{M} = \{\neg AbA \rightarrow (x \leftrightarrow (a \wedge b)), \neg AbO \rightarrow (d \leftrightarrow (x \vee c))\}$

- Meta-Observation:

$\{a, \neg b, c, \neg d\} \cup \mathcal{M} \cup \{\neg AbA, \neg AbO\}$ is inconsistent

- $\{a, \neg b, c, \neg d\} \cup \mathcal{M} \cup \{\neg AbA, \underbrace{AbO}_{\text{counterfactual change}}\}$ is consistent

Meta-Observation changed: abO is *counterfactual cause* for meta-observation

- However: $\{a, \neg b, c, \neg d\} \cup \mathcal{M} \cup \{\underbrace{AbA}_{\text{changed}}, \neg AbO\}$ is inconsistent

Meta-Observation not changed: AbA is not counterfactual cause

- Extra, **contingent interventions** (changes) may be necessary:

$$\{a, \neg b, c\} \cup \mathcal{M} \cup \{\underbrace{AbA}_{\text{changed}}, \underbrace{AbO}_{\text{contingent change}}\} \text{ is consistent}$$

- AbA is neither counterfactual nor actual cause

By definition, because the possible “contingent change” AbO is already a counterfactual cause

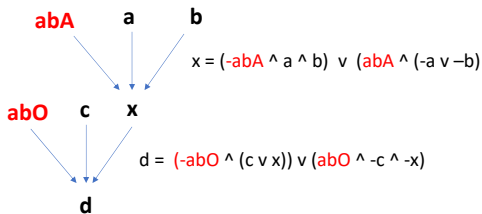
- **Responsibility Attribution Score** quantifies strength of a cause
- **Causal Responsibility:**

$$Resp(abO) := \frac{1}{1 + \min. \text{ cardinality of CS}} = \frac{1}{1+0} = 1 \quad (\text{max. responsibility})$$

$$Resp(abA) := 0 \quad (\text{no responsibility})$$

- We will concentrate on attribution scores as explanations

- There is a connection with causality based on **Causal Networks** and **Structural Models** used in AI
- Actual causality can be cast in those terms



- Here **abA**, **abO** are **endogenous variables**
 They can be subject to counterfactual intereventions
 The others are **exogenous variables**
- The CN above involves some **structural equations**

Back to the Bank



e = ⟨john, 18, plumber, 70K, harlem, ...⟩ **No**

- Counterfactual versions:

$e' = \langle \text{john}, 25, \text{plumber}, 70K, \text{harlem}, \dots \rangle$ Yes

Value for feature *Age* is counterfactual cause with explanatory responsibility $Resp(e, Age) = 1$

- $e'' = \langle \text{john}, 18, \text{plumber}, 80K, \text{brooklyn}, \dots \rangle$ Yes

Value for *Income* is actual cause with $Resp(e, Income) = \frac{1}{2}$

This one needs additional (contingent) changes ...

- Really? Inputs are non-binary now!

- For binary features the previous definition of responsibility works fine
- In the case of the classifier, possibly many new values for a feature do not change the label, and few of them do
- Then, the original value is not a great explanation
- Responsibility score has to be generalized (B. et al., Deem@SIGMOD20)
- Better consider contingent features and values for them, and average labels!
- We are considering binary classifiers, with labels 1 or 0
Assume label 1 is the one we want to explain
- *Resp* is a “local” explanation score: for a feature value in a particular entity

- $\mathbf{e}$ classified entity, $L(\mathbf{e}) = 1$, $F^* \in \mathcal{F}$ (set of features)
- “Local” *Resp*-score: for fixed contingent assignment $\Gamma := \bar{w}$
 $\Gamma \subseteq \mathcal{F} \setminus \{F^*\}$ (potential contingent set of features)
- $\mathbf{e}' := \mathbf{e}[\Gamma := \bar{w}]$ (potential contingent values), with $L(\mathbf{e}') = L(\mathbf{e})$

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

- $\mathbf{e}'' := \mathbf{e}[\Gamma := \bar{w}, F^* := v]$, with $v \in dom(F^*)$
- $\mathbf{e}_S$ is projection of $\mathbf{e}$ on $S \subseteq \mathcal{F}$
- When $(*) > 0$, $F^*(\mathbf{e})$ is *actual causal explanation* for $L(\mathbf{e}) = 1$ with contingency $\langle \Gamma, \mathbf{e}_\Gamma \rangle$
- Global score: $Resp(\mathbf{e}, F^*) := \max_{\langle \Gamma, \bar{w} \rangle, |\Gamma| \text{ min.}, (*) > 0} Resp(\mathbf{e}, F^*, \Gamma, \bar{w})$

- *Resp* requires multiple passes through the classifier ...
- *Resp* requires (assumes) a probability distribution on the entity population $\mathcal{E}$

Several probability distributions can be used

Quite a relevant issue ...

(B. et al., Deem@SIGMOD20)

- In our experiments, *Resp* score computed with empirical product distribution
- We are usually interested in max-*Resp* feature values

Associated to minimum (cardinality) contingency sets

Their computation is in some cases provably intractable

- *Resp* does not require the internals of a classifier

Can we compute it faster when we have access to the internals?

- Also relevant: doing something with a high-responsibility explanation

Some counterfactuals may not “make sense” or be “useful”

- In the example, changing the age (waiting for 7 years) may not be feasible

But maybe changing job and neighborhood could be done ...

- We may want an *actionable* explanation

We may want the explanation to be a *resource*

- Specification of- and reasoning with counterfactuals can be *useful* (coming back)

Shapley Values: *Shap*

- Based on the general *Shapley value* of coalition game theory
- For each application of Shapley one needs an appropriate game function that maps (sub)sets of players to real numbers
- Our case: *Set of players $\mathcal{F}$ contain features*, but relative to $\mathbf{e}$
- Game function: For $S \subseteq \mathcal{F}$, and $\mathbf{e}_S$ the projection of $\mathbf{e}$ on S

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

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

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

- *Shap score* has become popular (Lee & Lundberg, 2017)
- Assumes a probability distribution on entity population
- Requires multiple passes through classifier ...

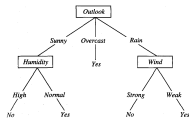
Tractability of SHAP on Open-Box Models

- SHAP has become the most popular explanation score in ML
Shapley value, in particular SHAP, is provably intractable

Can we do better at computing SHAP with an open classifier?

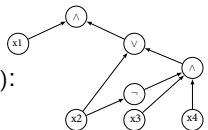
Having access to its structure and components?

- SHAP is tractable** for
Deterministic and Decomposable
Boolean Circuits (dDBCs)



(uniform or product distribution, to ease the analysis)

- We get tractability of SHAP for a vast collection of classifiers** (via Knowledge Compilation):
 - Decision Trees (even with non-binary)
 - Ordered Binary Decision Diagrams (OBDDs)
 - Sentential Decision Diagrams (SDDs)



(binary dDBC classifier)

(AAAI'21, JMLR 23)

SHAP on Binary Neural Networks

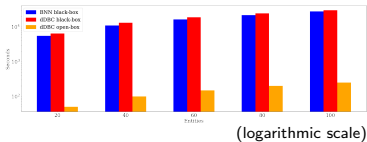
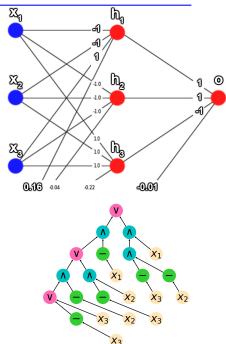
- NNs considered as black-boxes
- We experimented with SHAP computation on a BNN via compilation into a dDBC
 1. BNN $\mapsto$ CNF (parsimonious and optimized)
 2. CNF $\mapsto$ SDD (non-polynomial, but FPT)
 3. SDD $\mapsto$ dDBC (straightforward)

Still worth this one-time computation
(target dDBC may be used multiple times)

- Experiments: BNN with 14 gates, dDBC with 18,670 nodes

Compared SHAP computation for:
black-box BNN, open-box dDBC,
and black-box dDBC

- All SHAP scores for all entities,
with increasing numbers of them (JELIA'23)



The Need for Reasoning

- What can we do with attribution scores and counterfactual explanations? (apart from the obvious)
- We can **reason** about/with them, **analyze** them, **select** some of them, **aggregate** them, etc.

In **interaction** with both attribution-score model/algorithm or classifier, for further exploration

For **global understanding** of the classifier or application domain

- We need tools for **conveying or imposing domain knowledge** (domain semantics), e.g. an age never decreases

Only **some counterfactuals may make sense**

Some **combinations of feature values may not be allowed**

Some **changes may “trigger” other changes**

To impose **preferences** on counterfactuals

- We need tools for doing this kind of logical reasoning
- We need tools for posing and answering queries about explanations

Are there explanations with this particular property?

Or any two that differ by ...?

- Specification of high-score actionable explanations, and possibly computation of those only

Or others with a different preferred property

- On-the-fly interaction with different ML models and scores

Do I get same score with this different ML system?

Or this other attribution score (definition, algorithm or implementation)?

- Imposing conditions on feature values

What if I leave some feature values fixed?

Do I get same high-score feature with this “similar” entity?

Is there a high-score counterfactual version of the entity that changes this specific feature?

Or never changes that one?

- We have devised *declarative* (logic-based) methods to *reason with and about counterfactuals*, and compute *Resp* scores

In interaction with classifiers “called” from the program

We have used *Answer-Set Programming*, a form of logic programming

L. Bertossi. “Declarative Approaches to Counterfactual Explanations for Classification”. *Theory and Practice of Logic Programming*, 23 (3): 559–593, 2023.

Much in the direction of Neuro-Symbolic AI!