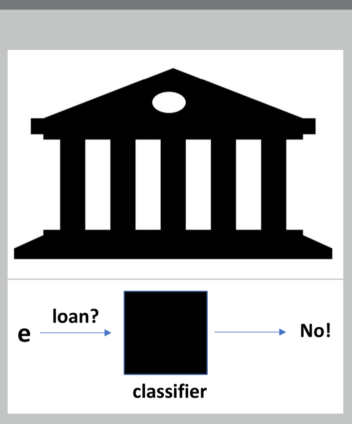
Answer-Set Programs for Reasoning about Counterfactual Interventions and Responsibility Scores for Classification

Leopoldo Bertossi and Gabriela Reyes

Adolfo Ibáñez University & Millennium Institute for Foundational Research on Data (IMFD), Chile

Explanations in Machine Learning

- Client requesting a loan from a bank
- Bank using a black-box classifier
- Entity represented as a record of values for features: Name, Age, Occupation, Income, ...



$$\mathbf{e} = \langle \mathit{John}, \mathbf{18}, \mathit{plumber}, \mathbf{70}K, \mathit{Harlem}\rangle$$

- Which are the feature values most relevant for the classification outcome, i.e. the label "No" ?

→ What is the contribution of each feature value to the outcome?

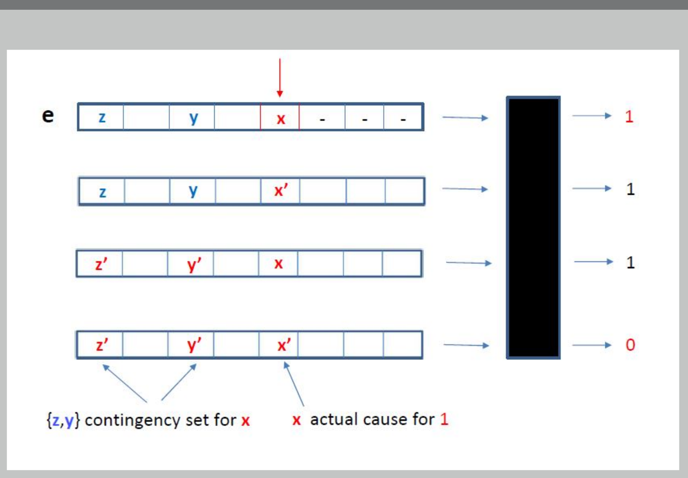
A Score-Based Approach: Responsibility

- Actual Causality based on Counterfactual Interventions

  (Halpern and Pearl, 2001)
- Hypothetical changes of values in a causal model to detect other changes: "What would happen if we change ..."?
  By so doing identify actual causes
- Do changes of feature values make the label change to "Yes"?
- Also the quantitative notion of Responsibility: a measure of causal contribution

  (Chockler and Halpern, 2004)
- We have investigated causality and responsibility in data management and classification
- Semantics, computational mechanisms, intrinsic complexity, logic-based specifications, reasoning, etc.

The Resp Score: Classification



- Want explanation for label "1"
- Through changes of feature values, try to get "0"
- Fix a feature value $x = e_F$
- x counterfactual explanation for $L(e) = 1$ if $L(e\frac{x}{x'}) = 0$, for $x' \in \mathit{Dom}(F)$
- x actual explanation for $L(e) = 1$ if there are values Y in e, $x \notin Y$, and new values $Y' \cup \{x'\}$:

$$\text{(a)}\, L(e\tfrac{Y}{Y'}) = 1 \qquad \text{(b)}\, L(e\tfrac{xY}{x'Y'}) = 0$$

- If Y is minimum in size: $x\text{-}\mathit{Resp}(x) := \frac{1}{1+|Y|}$

Example

- Due to $\mathbf{e}_7$, $F_2(\mathbf{e}_1)$ is counterfactual explanation, with $\Gamma = \emptyset$ and $\mathbf{Resp}(\mathbf{e}_1, F_2) = 1$
- Due to $\mathbf{e}_4$, $F_1(\mathbf{e}_1)$ is actual explanation; with $\Gamma = \{F_2(\mathbf{e}_1)\}$ as contingency set: $\mathbf{Resp}(\mathbf{e}_1, F_1) = \frac{1}{2}$

$\mathcal{C}$

entity (id)	F_1	F_2	F_3	L
$\mathbf{e}_1$	0	1	1	1
$\mathbf{e}_2$	1	1	1	1
$\mathbf{e}_3$	1	1	0	1
$\mathbf{e}_4$	1	0	1	0
$\mathbf{e}_5$	1	0	0	1
$\mathbf{e}_6$	0	1	0	1
$\mathbf{e}_7$	0	0	1	0
$\mathbf{e}_8$	0	0	0	0

- We are usually interested in maximum-responsibility feature values
  Associated to minimum (cardinality) contingency sets of feature values
- Sometimes we may be interested in minimal contingency sets, under set-inclusion

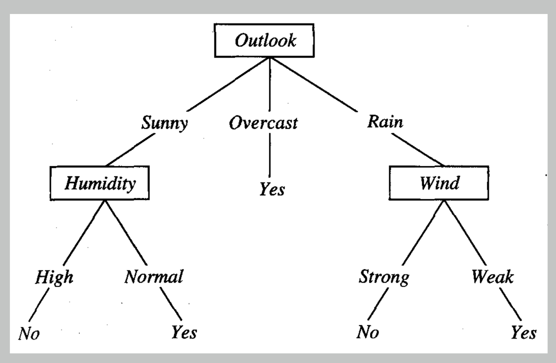
Objectives

- Obtaining responsibility scores
- Specify counterfactual interventions, preferably actionable ones
- Reason about them, and explanations
- Compute responsibility scores from the specifications

Reasoning about Counterfactual Interventions

- Given a classifier, one can reason in answer-set programming (ASP) about counterfactuals, lets say Mitchell' s Decision Tree:

Features $F = \{\mathbf{Outlook}, \mathbf{Humidity}, \mathbf{Wind}\}$
$\mathit{Dom}(\mathbf{Outlook}) = \{\mathbf{sunny}, \mathbf{overcast}, \mathbf{rain}\}$
$\mathit{Dom}(\mathbf{Humidity}) = \{\mathbf{high}, \mathbf{normal}\}$
$\mathit{Dom}(\mathbf{Wind}) = \{\mathbf{strong}, \mathbf{weak}\}$
Entity $\mathbf{e} = \mathbf{ent}(\mathbf{sunny}, \mathbf{normal}, \mathbf{weak})$ gets label Yes



- One can easily impose semantic constraints on counterfactuals
- Scores can be computed by means of set- and numerical aggregations
- Reasoning is enabled by cautious and brave query answering
- Explanations can be queried

ASPs for Counterfactual Interventions

- *Counterfactual Intervention Programs* (CIPs) specify counterfactual interventions on a given entity under classification
- We will use *DLV* and *DLV-Complex* notation
- So as with repair programs, we use annotation constants:

Annotation	Intended Meaning
o	original entity
do	do counterfactual intervention
tr	entity in transition
s	stop, label has changed (single change of feature value)

- Specifying domains, entity, classification tree, annotations:

```
% facts:
   dom1(sunny). dom1(overcast). dom1(rain). dom2(high). dom2(normal).
   dom3(strong). dom3(weak).
   ent(e,sunny,normal,weak,o).   % original entity at hand

% specification of the decision-tree classifier:
   cls(X,Y,Z,1) :- Y = normal, X = sunny, dom1(X), dom3(Z).
   cls(X,Y,Z,1) :- X = overcast, dom2(Y), dom3(Z).
   cls(X,Y,Z,1) :- Z = weak, X = rain, dom2(Y).
   cls(X,Y,Z,0) :- dom1(X), dom2(Y), dom3(Z), not cls(X,Y,Z,1).

% transition rules: the initial entity or one affected by a value change
   ent(E,X,Y,Z,tr) :- ent(E,X,Y,Z,o).
   ent(E,X,Y,Z,tr) :- ent(E,X,Y,Z,do).

% counterfactual rule: alternative single-value changes
   ent(E,Xp,Y,Z,do) v ent(E,X,Yp,Z,do) v ent(E,X,Y,Zp,do) :-
                     ent(E,X,Y,Z,tr), cls(X,Y,Z,1), dom1(Xp), dom2(Yp),
                     dom3(Zp), X != Xp, Y != Yp, Z!= Zp,
                     chosen1(X,Y,Z,Xp), chosen2(X,Y,Z,Yp),
                     chosen3(X,Y,Z,Zp).
```

- Classifier could be invoked as external predicate in Python
- The last is the counterfactual rule
- Only one disjunct in the head becomes true; one per feature
- It uses the non-deterministic choice predicate (choice makes the program non-stratified)
  Chooses a new value in last argument for each combination of the first three
- As long as the label does not depart from **1**, i.e. *yes*
- Non-stratified negation is what makes ASP necessary

Conclusions

- Addition of semantic and domain knowledge is important
  ASP-based approaches particulary appropriate
- Redefinition vs. hacked computation vs. change of distribution?
- Reasoning in general about explanations and counterfactuals is what intelligent agents do, score computation is not enough
- We should explore Resp, so as we did for SHAP, in the case of *deterministic and decomposable decision diagrams* (d-DDDs)
  Also with ASP-based specifications and computations
- Explanations are at the basis of fairness and bias analysis
- Understanding the decisions in relation to protected features becomes relevant
- Explaining *how* decisions are made