Score-Based Explanations in Data Management and Machine Learning

Leopoldo Bertossi

Universidad Adolfo Ibáñez
Faculty of Engineering and Sciences
Santiago, Chile
&
IMFD (Chile)

Tutorial at SUM 2020

CONTENTS:

- ▷ Explanations and Causality in Databases
- ▷ The Causal-Effect Score in DBs
- ▷ The Shapley-Value as Explanation Score in DBs
- ▷ Score-Based Explanations for Classification
- ▷ The SHAP-Score (based on Shapley-Value)
- ▷ The RESP-Score (based on Causal Responsibility)

This is not an exhaustive or broad survey

This tutorial is largely influenced by my own research in these areas

Explanations in Databases

- In data management (DM), we need to understand ***why*** certain results are obtained or not

  And characterize and compute "reasons" therefor

  E.g. for query answers, violation of semantic conditions, ...

- A DB system should provide *explanations*

  In our case, causality-based explanations (Halpern and Pearl, 2001)

- There are other (related) approaches, e.g. *lineage*, *provenance*

- We have tried to understand causality in DM from different perspectives

  Taking advantage of the connections

Causality in DBs

Example: DB D as below

Boolean conjunctive query (BCQ):

$\mathcal{Q}$: $\exists x \exists y (S(x) \wedge R(x,y) \wedge S(y))$

$D \models \mathcal{Q}$ Causes?

R	A	B
	a	b
	c	d
	b	b

S	A
	a
	c
	b

(Meliou, Gatterbauer, Moore, Suciu; 2010)

- Tuple $\tau \in D$ is counterfactual cause for $\mathcal{Q}$ if $D \models \mathcal{Q}$ and $D \setminus \{\tau\} \not\models \mathcal{Q}$

  $S(b)$ is counterfactual cause for $\mathcal{Q}$: if $S(b)$ is removed from D, $\mathcal{Q}$ is not true anymore

- Tuple $\tau \in D$ is actual cause for $\mathcal{Q}$ if there is a contingency set $\Gamma \subseteq D$, such that τ is a counterfactual cause for $\mathcal{Q}$ in $D \setminus \Gamma$

  $R(a,b)$ is an actual cause for $\mathcal{Q}$ with contingency set $\{R(b,b)\}$: if $R(a,b)$ is removed from D, $\mathcal{Q}$ is still true, but further removing $R(b,b)$ makes $\mathcal{Q}$ false

- How strong are these as causes? (Chockler and Halpern, 2004)

  - The responsibility of an actual cause τ for $\mathcal{Q}$:

    $$\rho_D(\tau) \;:=\; \frac{1}{|\Gamma| \,+\, 1}$$

    $|\Gamma|$ = size of smallest contingency set for τ (0 otherwise)

    Responsibility of $R(a,b)$ is $\frac{1}{2} = \frac{1}{1+1}$ (its several smallest contingency sets have all size 1)

    $R(b,b)$ and $S(a)$ are also actual causes with responsibility $\frac{1}{2}$

    $S(b)$ is actual (counterfactual) cause with responsibility $1 = \frac{1}{1+0}$

- High responsibility tuples provide more interesting explanations
- Responsibilities can be seen as "scores" assigned to tuples
- Causality can be extended to attribute-value level (Bertossi, Salimi; TOCS 2017)
- Causality under ICs (Bertossi, Salimi; IJAR, 2017)

Causality Connections

- There are mutual reductions with repairs of DBs wrt. ICs
- The same with consistency-based diagnosis and abductive diagnosis
- This led to new complexity and algorithmic results for causality and responsibility

(Bertossi, Salimi; TOCS, IJAR, 2017)

- Causality can be seen as a way to provide diagnosis

  So, as with model-based diagnosis above

- Pearl's causality: Perform counterfactual *interventions* on a structural, logico/probabilistic model

  What would happen if we change ...?

  Identify causes through changes after intervention in the model

- In the case of DBs the underlying logical model is *query lineage*
- Counterfactual approaches are much newer in "explainable AI"

  Use to provide explanations, possibly in the absence of a model

  Scores can be derived from counterfactuals
- Explanation scores have become popular in ML

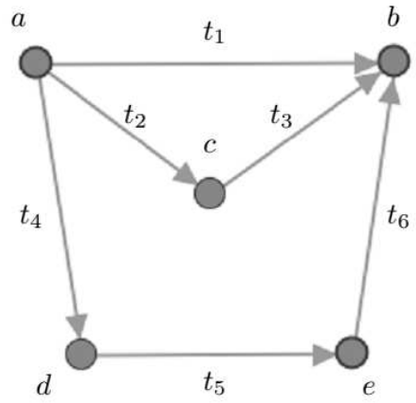
  Usually with a counterfactual component or flavor
- Responsibility can be seen as one of them ...

- Not the only one; not even in DBs ...

The Causal-Effect Score

Example: Boolean Datalog query Π becomes true on E if there is a path between a and b

E	X	Y
t_1	a	b
t_2	a	c
t_3	c	b
t_4	a	d
t_5	d	e
t_6	e	b



$$yes \leftarrow P(a,b)$$
$$P(x,y) \leftarrow E(x,y)$$
$$P(x,y) \leftarrow P(x,z), E(z,y)$$

$E \cup \Pi \models yes$

All tuples are actual causes: every tuple appears in a path from a to b

All the tuples have the same causal responsibility: $\frac{1}{3}$

Maybe counterintuitive: t_1 provides a direct path from a to b

- Alternative notion to responsibility: *causal-effect* (Salimi et al., TaPP'16)
- Retake question: How answer to $\mathcal{Q}$ changes if τ deleted from D?

  An *intervention* on a *structural causal model*

  In this case provided by the the *lineage of the query*

Example: Database D

R	A	B
	a	b
	a	c
	c	b

S	B
	b
	c

BCQ $\mathcal{Q}: \exists x \exists y (R(x,y) \wedge S(y))$ True in D

Query lineage instantiated on D given by propositional formula:

$$\Phi_{\mathcal{Q}}(D) = (X_{R(a,b)} \wedge X_{S(b)}) \vee (X_{R(a,c)} \wedge X_{S(c)}) \vee (X_{R(c,b)} \wedge X_{S(b)}) \qquad (*)$$

X_τ: propositional variable that is true iff $\tau \in D$

$\Phi_{\mathcal{Q}}(D)$ takes value 1 in D

- Want to quantify contribution of a tuple to a query answer, say, $S(b)$

  Assign probabilities, uniformly and independently, to the tuples in D

- A probabilistic database D^p (tuples outside D get probability 0)

R^p	A	B	prob
	a	b	$\frac{1}{2}$
	a	c	$\frac{1}{2}$
	c	b	$\frac{1}{2}$

S^p	B	prob
	b	$\frac{1}{2}$
	c	$\frac{1}{2}$

- The X_τ's become independent, identically distributed random variables; and $\mathcal{Q}$ is Bernoulli random variable

  What's the probability that $\mathcal{Q}$ takes truth value 1 (or 0) when an intervention is done on D?

- Interventions of the form $do(X = x), \quad x \in \{0, 1\}$

  In the lineage $(*)$ make X take value x: $P(\mathcal{Q} = y \mid do(X_\tau = x))$? $\quad y \in \{0, 1\}$

  E.g. with $do(X_{S(b)} = 0)$: $\quad \Phi_{\mathcal{Q}}(D)\frac{X_{S(b)}}{0} := (X_{R(a,c)} \wedge X_{S(c)})$

- The *causal-effect* of τ: $\quad \mathcal{CE}^{D,\mathcal{Q}}(\tau) \;:=\; \mathbb{E}(\mathcal{Q} \mid do(X_\tau = 1)) - \mathbb{E}(\mathcal{Q} \mid do(X_\tau = 0))$

Example: (cont.)

From $\Phi_{\mathcal{Q}}(D)\frac{X_{S(b)}}{0} := (X_{R(a,c)} \wedge X_{S(c)})$:

$$P(\mathcal{Q} = 1 \mid do(X_{S(b)} = 0)) = P(X_{R(a,c)} = 1) \times P(X_{S(c)} = 1) = \tfrac{1}{4}$$

From $\Phi_{\mathcal{Q}}(D)\frac{X_{S(b)}}{1} := X_{R(a,b)} \vee (X_{R(a,c)} \wedge X_{S(c)}) \vee X_{R(c,b)}$:

$$\begin{aligned} P(\mathcal{Q} = 1 \mid do(X_{S(b)} = 1)) &= P(X_{R(a,b)} \vee (X_{R(a,c)} \wedge X_{S(c)}) \vee X_{R(c,b)} = 1) \\ &= \cdots = \tfrac{13}{16} \end{aligned}$$

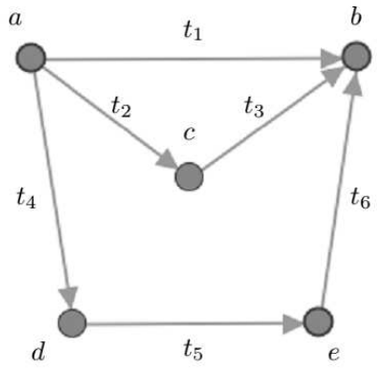
Then: $\mathbb{E}(\mathcal{Q} \mid do(X_{S(b)} = 0)) = P(\mathcal{Q} = 1 \mid do(X_{S(b)} = 0)) = \tfrac{1}{4}$

$$\mathbb{E}(\mathcal{Q} \mid do(X_{S(b)} = 1)) = P(\mathcal{Q} = 1 \mid do(X_{S(b)} = 1)) = \tfrac{13}{16}$$

Causal effect for $S(b)$:

$$\mathcal{CE}^{D,\mathcal{Q}}(S(b)) = \tfrac{13}{16} - \tfrac{1}{4} = \tfrac{9}{16} > 0$$

Example: (cont.) The Datalog query (here as a union of BCQs) has the lineage:



$$\Phi_{\mathcal{Q}}(D) = X_{t_1} \vee (X_{t_2} \wedge X_{t_3}) \vee (X_{t_4} \wedge X_{t_5} \wedge X_{t_6})$$

$$\mathcal{CE}^{D,\mathcal{Q}}(t_1) = 0.65625$$

$$\mathcal{CE}^{D,\mathcal{Q}}(t_2) = \mathcal{CE}^{D,\mathcal{Q}}(t_3) = 0.21875$$

$$\mathcal{CE}^{D,\mathcal{Q}}(t_4) = \mathcal{CE}^{D,\mathcal{Q}}(t_5) = \mathcal{CE}^{D,\mathcal{Q}}(t_6) = 0.09375$$

The causal-effects are different for different tuples!

More intuitive result than responsibility

- Rather *ad hoc* or arbitrary? (we'll be back ...)

Scores and Coalition Games

- Contribution of a DB tuple to a query answer?
- *Several tuples together* are necessary to violate an IC or produce a query result

  Like players in a coalition game, some may contribute more than others
- The Shapley-value is firmly established and widely used in Game Theory
- Shapley value is the only function that satisfies certain natural conditions
- Apply Shapley-value to QA in DBs (Livshits et al.; ICDT'20)

  The Shapley-value of a tuple will be a score for its contribution

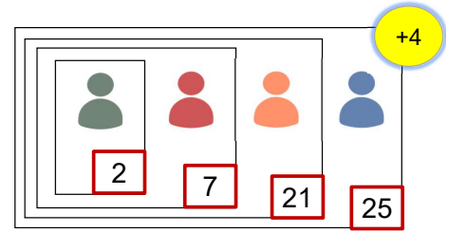
The Shapley Value

- Consider a set of players D, and a wealth-distribution (game) function $\mathcal{G}: \mathcal{P}(D) \longrightarrow \mathbb{R}$ ($\mathcal{P}(D)$ the power set of D)
- The Shapley value of player p among a set of players D:

$$Shapley(D, \mathcal{G}, p) := \sum_{S \subseteq D \setminus \{p\}} \frac{|S|!(|D| - |S| - 1)!}{|D|!} (\mathcal{G}(S \cup \{p\}) - \mathcal{G}(S))$$

($|S|!(|D| - |S| - 1)!$ is number of permutations of D with all players in S coming first, then p, and then all the others)

Expected contribution of player p under all possible additions of p to a partial random sequence of players followed by a random sequence of the rest of the players



- Shapley difficult to compute; provably #P-hard in general
- Counterfactual flavor: What happens having p vs. not having p?

Shapley as Score for QA

- Back to QA in DBs, players are tuples in DB D

  Boolean query $\mathcal{Q}$ becomes game function: for $S \subseteq D$

$$\mathcal{Q}(S) = \begin{cases} 1 & \text{if } S \models \mathcal{Q} \\ 0 & \text{if } S \not\models \mathcal{Q} \end{cases}$$

- Concentrated on BCQs (and some aggregation on CQs)

$$Shapley(D, \mathcal{Q}, \tau) := \sum_{S \subseteq D \setminus \{\tau\}} \frac{|S|!(|D|-|S|-1)!}{|D|!} (\mathcal{Q}(S \cup \{\tau\}) - \mathcal{Q}(S))$$

  Quantifies the contribution of tuple τ to query result (Livshits et al.; ICDT'20)

- Dichotomy Theorem: $\mathcal{Q}$ BCQ without self-joins

  If $\mathcal{Q}$ hierarchical, then $Shapley(D, \mathcal{Q}, \tau)$ can be computed in PTIME

  Otherwise, the problem is $FP^{\#P}$-complete

- $\mathcal{Q}$ is hierarchical if for every two existential variables x and y:
  - $Atoms(x) \subseteq Atoms(y)$, or
  - $Atoms(y) \subseteq Atoms(x)$, or
  - $Atoms(x) \cap Atoms(y) = \emptyset$

Example: $\mathcal{Q}: \ \exists x \exists y \exists z (R(x, y) \wedge S(x, z))$

$Atoms(x) = \{R(x, y), \ S(x, z)\}, \ Atoms(y) = \{R(x, y)\}, \ Atoms(z) = \{S(x, z)\}$ Hierarchical!

Example: $\mathcal{Q}^{nh}: \ \exists x \exists y (R(x) \wedge S(x, y) \wedge T(y))$

$Atoms(x) = \{R(x), \ S(x, y)\}, \ Atoms(y) = \{S(x, y), T(y)\}$ Not hierarchical!

- Same criteria as for QA over probabilistic DBs (Dalvi & Suciu; 2004)
- Positive case: A dynamic programming approach works
- Negative case: requires a fresh approach (not from probabilistic DBs)

  Use query $\mathcal{Q}^{nh}$ above

  Reduction from counting independent sets in a bipartite graph
- What to do in hard cases?

  For every fixed BCQ $\mathcal{Q}$, there is a multiplicative fully-polynomial randomized approximation scheme (FPRAS)

- Also investigated related score: the Bahnzhaf Power Index (order ignored)

  $$Banzhaf(D, \mathcal{Q}, \tau) := \frac{1}{2^{|D|-1}} \cdot \sum_{S \subseteq (D \setminus \{\tau\})} (\mathcal{Q}(S \cup \{\tau\}) - \mathcal{Q}(S))$$

  Bahnzhaf also difficult to compute; provably #P-hard in general

- We proved Causal-Effect coincides with the Banzhaf Index (op. cit.)

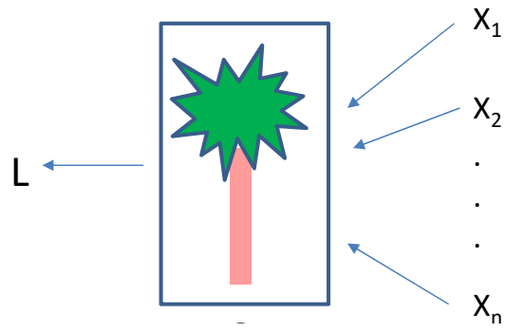
Score-Based Explanations for Classification

$\mathbf{e} = \langle x_1, \ldots, x_n \rangle$ entity requesting a loan

- Black-box binary classification model returns label $L(\mathbf{e}) = 1$, i.e. rejected

  Why???!!!

- Similarly if we had a model, e.g. a decision tree or a logistic regression model



- Which feature values x_i contribute the most?

  Assign numerical scores to feature values in $\mathbf{e}$

  Capturing the relevance of the feature value for the outcome

- In general they are (but not always) based on counterfactual interventions

- Some scores can be applied both with black-box and open models

  E.g. Shapley $\rightsquigarrow$ SHAP has become popular (Lee& Lundberg; 2017, 2020)

  - Players are features in $\mathcal{F}$
  - Game function determined by $\mathbf{e}$: $\mathcal{G}_{\mathbf{e}}(S) := \mathbb{E}(L(\mathbf{e}') \mid \mathbf{e}'_S = \mathbf{e}_S)$

    In this way features values for $\mathbf{e}$ are being assessed ($\mathbf{e}_S$: projection of $\mathbf{e}$ on S)
  - For a feature $F \in \mathcal{F}$, compute: $Shapley(\mathcal{F}, \mathcal{G}_{\mathbf{e}}, F)$
  - Assuming an underlying probability space of entities $\mathbf{e}'$
  - L acts as a Bernoulli random variable

- $Shapley(\mathcal{F}, \mathcal{G}_{\mathbf{e}}, F)$ requires computing

$$\sum_{S \subseteq \mathcal{F} \setminus \{F\}} \frac{|S|!(|D|-|S|-1)!}{|D|!} (\mathbb{E}(L(\mathbf{e}'|\mathbf{e}'_{S \cup \{F\}} = \mathbf{e}_{S \cup \{F\}}) - \mathbb{E}(L(\mathbf{e}')|\mathbf{e}'_S = \mathbf{e}_S))$$

- As already mentioned SHAP can be applied with black-box, and also with open, explicit models
- With black-box models, using the classifier many times
  - With the entire space, and a given underlying distribution

    Not very appealing ...
  - Using a sample of the population, and computing weighted averages

    More natural and realistic in practice
- With explicit, open models
  - As with black-box models
  - Using the given classification model, and computing the expectation

    For some models and population distributions, SHAP computation can be done exactly and efficiently

- Original paper on SHAP claims it can be computed in PTIME for decision-trees (actually, random forests)

  Actually, introduced, discussed and experimented in this context

- Recently proved:

  SHAP can be computed in PTIME on a series of Binary Decision Circuits as classifiers (Arenas, Barcelo, Bertossi, Monet)

  The result applies in particular to decision-trees

Yet Another Score: RESP

- Same classification setting, $F \in \mathcal{F}$ (Bertossi, Li, Schleich, Suciu, Vagena; DEEM@SIGMOD'20)
- COUNTER$(\mathbf{e}, F) := L(\mathbf{e}) - \mathbb{E}(L(\mathbf{e}') \mid \mathbf{e}'_{\mathcal{F} \smallsetminus \{F\}} = \mathbf{e}_{\mathcal{F} \smallsetminus \{F\}})$ (local interventions)

  This score can be applied to same scenarios, it is easy to compute

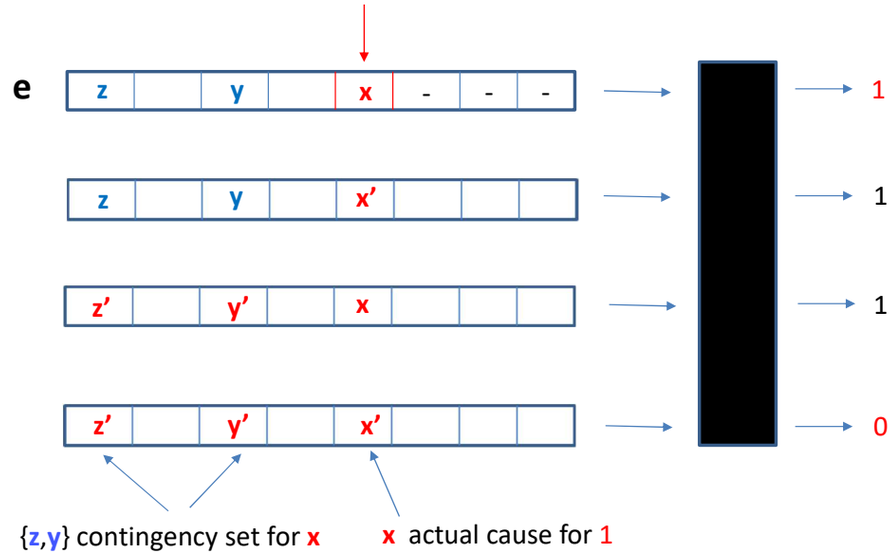
  Gives reasonable results, intuitively and in comparison to other scores
- So as SHAP, need an underlying probability space

  No need to access the internals of the classification model
- A problem: changing one value may not switch the label

  No explanations are obtained
- Extend this score bringing in contingency sets for features!

  The RESP-score (simplified version for binary features first)

- Want explanation for classification "1" for $\mathbf{e}$
- Through interventions, changes of feature values, try to change it to "0"
- Fix a feature value $\mathbf{x}= F(\mathbf{e})$



- $\mathbf{x}$ counterfactual explanation for $L(\mathbf{e}) = 1$ if $L(\mathbf{e}\frac{\mathbf{x}}{\mathbf{x}'}) = 0$, for $\mathbf{x}' \in dom(F)$
- $\mathbf{x}$ actual explanation for $L(\mathbf{e}) = 1$ if there is a set of values $\mathbf{Y}$ in $\mathbf{e}$, $\mathbf{x} \notin \mathbf{Y}$, and (all) new values $\mathbf{Y}' \cup \{\mathbf{x}'\}$:

  (a) $L(\mathbf{e}\frac{\mathbf{Y}}{\mathbf{Y}'}) = 1$ (b) $L(\mathbf{e}\frac{\mathbf{x}\mathbf{Y}}{\mathbf{x}'\mathbf{Y}'}) = 0$
- If $\mathbf{Y}$ is minimum in size, $\text{RESP}(\mathbf{x}) := \frac{1}{1+|\mathbf{Y}|}$

Example:

$\mathcal{C}$

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

▷ Due to $\mathbf{e}_7$, $F_2(\mathbf{e}_1)$ is counterfactual explanation; with $\text{RESP}(\mathbf{e}_1, F_2) = 1$

▷ Due to $\mathbf{e}_4$, $F_1(\mathbf{e}_1)$ is actual explanation; with $\{F_2(\mathbf{e}_1)\}$ as contingency set

And $\text{RESP}(\mathbf{e}_1, F_1) = \frac{1}{2}$

- For non-binary features, RESP can be expressed as an expected value

- Consider: $\mathbf{e}$ entity under classification, with $L(\mathbf{e}) = 1$, and $F \in \mathcal{F}$

Assume we have:

1. $\Gamma \subseteq \mathcal{F} \setminus \{F\}$, a set of features that may end up accompanying F
2. $\bar{w} = (w_F)_{F \in \Gamma}$, $w_F \in dom(F)$, $w_F \neq \mathbf{e}_F$, new values for features in Γ
3. $\mathbf{e}' := \mathbf{e}[\Gamma := \bar{w}]$, i.e. reset $\mathbf{e}$'s values for Γ as in $\bar{w}$
4. $L(\mathbf{e}') = L(\mathbf{e}) = 1$, no label change with $\bar{w}$, but maybe with extra change
5. Pick $v \in dom(F)$, $\mathbf{e}'' := \mathbf{e}[\Gamma := \bar{w}, F := v]$

▷ When $v \neq F(\mathbf{e})$ and $L(\mathbf{e}) \neq L(\mathbf{e}'') = 0$, $F(\mathbf{e})$ is an *actual causal explanation* for $L(\mathbf{e}) = 1$ with contingency $\langle \Gamma, \mathbf{e}_\Gamma \rangle$

To define the "local" RESP-score make v vary randomly under conditions 1.-5.:

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

Globally: $\text{RESP}(\mathbf{e}, F) := \underset{\substack{|\Gamma| min., (*) > 0 \\ \langle \Gamma, \bar{w} \rangle \models 1.-4.}}{max_{\bar{w}}} \text{RESP}(\mathbf{e}, F, \Gamma, \bar{w})$

Final Remarks

- We compared COUNTER, RESP, SHAP, Banzhaf (c.f. paper)

  Kaggle loan data set, and XGBoost for classification (opaque enough)

- There is the Rudin's FICO-Score: open model, model dependent

  Uses internal outputs and coefficients of two nested logistic-regression models

  Model designed for FICO data; so we used FICO data, and made experimental comparisons

- Explainable AI (XAI) is an effervescent area of research

  Its relevance can only grow

  Legislation around explainability, transparency and fairness of AI/ML systems

- Different approaches and methodologies

  Causality, counterfactuals and scores have relevant roles to play

- Much research needed on the use of contextual, semantic and domain knowledge

  Some approaches are more appropriate, e.g. declarative (Bertossi; RuleML+RR'20)

- Still fundamental research is needed on what is a good explanation

  As needed in AI/ML

  And on desired properties of an explanation score

- Shapley-value originally emerged from a list of *desiderata*

  Could an explanation score emerge as the right one for certain explainability properties?

EXTRA SLIDES

More on Experiments

- We compared COUNTER, RESP, SHAP, Banzhaf

  Kaggle loan data set, and XGBoost with Python library for classification model (opaque enough)

- Also comparison with Rudin's FICO-Score: model dependent, open model

  Uses outputs and coefficients of two nested logistic-regression models

  Model designed for FICO data; so, we used FICO data

- Here we are interested more in the experimental setting than in results themselves

- RESP score: appealed to "product probability space": for n, say, binary features
  - $\Omega = \{0,1\}^n$, $T \subseteq \Omega$ a sample
  - $p_i = P(F_i = 1) \approx \frac{|\{\omega \in T \mid \omega_i = 1\}|}{|T|} =: \hat{p}_i$ (estimation of marginals)
  - Product distribution over Ω:

$$P(\omega) := \Pi_{\omega_i=1} \hat{p}_i \times \Pi_{\omega_j=0} (1 - \hat{p}_j), \quad \text{for } \omega \in \Omega$$

- Not very good at capturing feature correlations
- RESP score computation for $\mathbf{e} \in \Omega$:
  - Expectations relative to product probability space
  - Choose values for interventions from feature domains, as determined by T
  - Call the classifier
  - Restrict contingency sets to, say, two features

- SHAP score appealed to "empirical probability space"
- Computing it on the product probability space may be $\#P$-hard (c.f. paper)
- Use sample $T \subseteq \Omega$, test data

  Labels $L(\omega)$, $\omega \in T$, computed with learned classifier
- Empirical distribution: $P(\omega) := \begin{cases} \frac{1}{|T|} & \text{if } \omega \in T \\ 0 & \text{if } \omega \notin T \end{cases}$,for $\omega \in \Omega$
- SHAP value with expectations over this space, directly over data/labels in T
- The empirical distribution is not suitable for the RESP score (c.f. the paper)