Formal Explanations of Black-Box Ranking Functions

Francesco Chiariello [1] Joao Marques-Silva [2]

[1]RWTH Aachen University, Germany
francesco.chiariello@ml.rwth-aachen.de

[2]ICREA, University of Lleida, Spain
jpms@icrea.cat

Formal Explanations of Black-Box Ranking Functions

Ranking

- **Ranking** is the task of arranging items according to some criterion.
- It is important across several domains;
  - **Search and Information Retrieval**: documents
  - **E-commerce and Recommendations**: products, media
  - **Professional and Academic**: job candidates, college applicants
  - **Medical Diagnosis**: patients
- It helps improve planning, scheduling, and decision-making.
  - In **healthcare scheduling**, ranking patients by disease probability allows prioritizing examinations and interventions based on urgency and severity.

The Need for FXAI

- Given the impact of ranking on our life it is important to have explanations that ensure transparency, understanding, and trust.
  - This is even more pressing, if considering that rankings are generated by **Machine Learning** algorithms.
- We leverage **Formal eXplainable AI (FXAI)**, adapting ideas for classifiers to the case of rankings.

Features

- **Feature Set**: A set of features $\mathcal{F} = \{1, \ldots, m\}$.
  - Each feature $i \in \mathcal{F}$ has an associated domain D_i.
  - Domains can be either categorical or numerical.
- **Feature Space**: The space of all possible **feature vectors**, defined as

$$\mathbb{F} = \prod_{i=1}^{m} D_i.$$

- Given $\mathcal{S} \subseteq \mathcal{F}$, two vectors $\mathbf{x}, \mathbf{v} \in \mathbb{F}$ **agree** on $\mathcal{S}$

$$\mathbf{x} \sim_{\mathcal{S}} \mathbf{v} \overset{\text{def}}{\iff} \forall i \in \mathcal{S}, x_i = v_i$$

- We also define

$$[\mathbf{v}]_{\mathcal{S}} := [\mathbf{v}]_{\sim_{\mathcal{S}}} = \{\mathbf{x} \in \mathbb{F} : \mathbf{x} \sim_{\mathcal{S}} \mathbf{v}\}$$

Classifiers

- **Classifier**: Given a set of classes $\mathcal{K} = \{c_1, \ldots, c_k\}$, a classifier is a function
$$\kappa : \mathbb{F} \to \mathcal{K}$$
that assigns each feature vector $\mathbf{x} \in \mathbb{F}$ to a class $c \in \mathcal{K}$.
- **Classification Problem**: Learn the classifier κ from training examples $(\mathbf{x}, c)$.
- **Explanation problem**: given the classifier κ and a $\mathbf{v} \in \mathbb{F}$, **why** κ predict $\kappa(\mathbf{v})$ on $\mathbf{v}$?

Abductive Explanations

- A set $\mathcal{S} \subseteq \mathcal{F}$ is a **Weak Abductive Explanation (WeakAXp)** for $(\kappa, \mathbf{v})$ if
$$\forall \mathbf{x} \in [\mathbf{v}]_{\mathcal{S}}, \kappa(\mathbf{x}) = \kappa(\mathbf{v})$$
i.e., if the classifier predicts the same class for all $\mathbf{x}$ that agree with $\mathbf{v}$ on $\mathcal{S}$.
- A set $\mathcal{S} \subseteq \mathcal{F}$ is an **Abductive Explanation (AXp)** if:
  - $\mathit{WeakAXp}(\mathcal{S})$
  - $\mathcal{S}' \subset \mathcal{S} \implies \neg \mathit{WeakAXp}(S')$

  In other words, AXps are subset-minimal WeakAXps.

Rankings and Ranking Functions

- Given a finite set A, a **ranking** $\preceq$ on A is a binary relation that is:
  - **Reflexive**: $\forall a \in A, a \preceq a$.
  - **Transitive**: $\forall a, b, c \in A, a \preceq b \wedge b \preceq c \implies a \preceq c$.
  - **Strongly connected**: $\forall a, b \in A,\ a \preceq b \vee b \preceq a$.
- A **ranking function** (or ranker) on A is a function $f : A \to \mathbb{R}$.
  - The value $f(a) \in \mathbb{R}$ represents the *score* assigned to $a \in A$.
  - The ranker f on A induce a ranking $\preceq_f$ on A, defined by

$$a \preceq_f b \iff f(a) \leq f(b)$$

- **Note 1**: Rankings (ranking functions) correspond to preferences (resp. utility functions).
- **Note 2**: Rankings are more general than linear orders as they allow for ties.

Explanation Problem for Ranking Functions

We aim to address the following question:

- Given a ranker $f : \mathbb{F} \to \mathbb{R}$ and a pair of vectors $\mathbf{v}, \mathbf{v}' \in \mathbb{F}$ such that $\mathbf{v} \preceq_f \mathbf{v}'$,

Why is $\mathbf{v}'$ ranked at least as highly as $\mathbf{v}$?

Reduction to Classification

- Consider the binary classifier $\kappa_f : \mathbb{F}^2 \to \{0, 1\}$, defined by

$$\kappa_f(\mathbf{x}, \mathbf{x}') = \begin{cases} 1, & \text{if } \mathbf{x} \preceq_f \mathbf{x}' \\ 0, & \text{otherwise.} \end{cases}$$

- One can then apply FXAI for classifiers to κ_f.

Reduction to Classification

- Consider the binary classifier $\kappa_f : \mathbb{F}^2 \to \{0, 1\}$, defined by

$$\kappa_f(\mathbf{x}, \mathbf{x}') = \begin{cases} 1, & \text{if } \mathbf{x} \preceq_f \mathbf{x}' \\ 0, & \text{otherwise.} \end{cases}$$

- One can then apply FXAI for classifiers to κ_f.

Issues

- each vector has its own copy of the features,
- each feature is treated independently,
- explanations are defined over the new feature set $\mathcal{F} \cup \mathcal{F}'$ obtained by adding a primed copy for each feature.

Abductive Explanations for Rankings

- A set $\mathcal{S} \subseteq \mathcal{F}$ is a **WeakAX** for $(f; \mathbf{v}, \mathbf{v}')$ if

$$\forall (\mathbf{x}, \mathbf{x}') \in [\mathbf{v}]_{\mathcal{S}} \times [\mathbf{v}']_{\mathcal{S}}, \mathbf{x} \preceq_f \mathbf{x}'.$$

- **Note**:
  - features $i \in \mathcal{S}$ are fixed for both vectors $\mathbf{x}, \mathbf{x}'$
  - explanations are defined over the original feature set $\mathcal{F}$.

Theorem (Monotonicity)

If $\mathcal{S}$ is a WeakAXp, then any superset $S' \supseteq S$ is also a WeakAXp.

- A set $\mathcal{S} \subseteq \mathcal{F}$ is an **Abductive Explanation** if:
  1. $WeakAXp(\mathcal{S})$
  2. $\mathcal{S}' \subset \mathcal{S} \implies \neg WeakAXp(S')$

Which Explanation to Prefer?

Problem:

- AXps are not unique.
- Several cardinality-minimal AXps may exist.
- $\Rightarrow$ Which explanation should we prefer?

Solution:

- **Score function:**

$$\mathit{score}(\mathcal{S}) = \min_{(\mathbf{x},\mathbf{x}')\in[\mathbf{v}]_{\mathcal{S}}\times[\mathbf{v}']_{\mathcal{S}}} (f(\mathbf{x}') - f(\mathbf{x}))$$

- **Key property:**

$$\mathrm{WeakAXp}(\mathcal{S}) \iff \mathit{score}(\mathcal{S}) \geq 0$$

- **Preference relation:**

$$\mathcal{S}_1 \preceq \mathcal{S}_2 \quad \iff \quad \mathit{score}(\mathcal{S}_1) \leq \mathit{score}(\mathcal{S}_2)$$

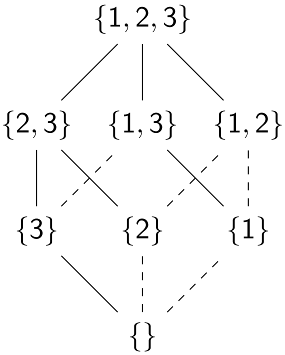
The score is especially meaningful when f has an intrinsic interpretation.

Computing an AXp

- Instance of **Minimal Set over a Monotone Predicate** problem
  - Other examples: Minimal Unsatisfiable Subsets, Minimal Equivalent Subsets, Prime Implicates/Implicants
  - Use optimal algorithms from the literature
- Verify if a set of features is a weak abductive explanation
  - Exhaustive search for counterexample $(\mathbf{x}, \mathbf{x}')$ s.t. $\mathbf{x}$ ranks higher than $\mathbf{x}'$
  - If none found $\Rightarrow$ set is a WeakAXp
  - Works with **black-box models**, including large-scale or proprietary ones
- Use **deletion-based algorithm** to compute an AXp.

Deletion-based algorithm

Figure: Hasse diagram of the search space for $m = 3$ features. Dashed lines indicate child nodes skipped during traversal.



Algorithm 1: Deletion-based Computation of AXp.

Input: $\mathcal{S} \subseteq \mathcal{F}$
Output: AXp $\mathcal{S}' \subseteq \mathcal{S}$ or `None`

```
if not WeakAXp(S) then
    return None
S' ← S
for i ∈ S do
    if WeakAXp(S' \ {i}) then
        S' ← S' \ {i}
return S'
```

Case study: Breast Cancer

We consider the **Breast Cancer Dataset**[1] containing data about breast cancer recurrence within 5 years after surgery.

Characteristic	Value
#instances	286
#features	9
#classes	2
No recurrence	201
Recurrence	85
Recurrence rate	$\approx$ 30%

Feature	Name	$\lvert\mathbb{D}_i\rvert$
1	*age*	6
2	*menopause*	3
3	*tumor-size*	11
4	*inv-nodes*	7
5	*node-caps*	3
6	*deg-malig*	3
7	*breast*	2
8	*breast-quad*	6
9	*irradiat*	2

[1] https://archive.ics.uci.edu/dataset/14/breast+cancer

Dataset Preparation

- Cancer recurrence is denoted by 1, absence by 0.
- Categorical variables are one-hot encoded to make them compatible with the neural network.
- After encoding, the feature space has 43 dimensions, containing the $299,376$ distinct possible patient profiles.

Model

- **Architecture**: Feedforward Neural Network with 3 dense layers
- **Training**: We train the model using the Adam optimizer and binary cross-entropy as the loss function, allocating 80% of the dataset for training and 20% for testing
- **Results**: 72% accuracy, 53% F1 score. (as a comparator, the baseline model has 64% accuracy, 0% F1 score).

Layer type	Shape	Param #
Dense (ReLU)	(43, 64)	2816
Dense (ReLU)	(64, 32)	2080
Dense (sigmoid)	(32, 1)	33
Trainable params		4929
Optimizer params		9860
Total params		**14789**

Point-wise Learning to Rank

- We learn a ranking function using a **point-wise** approach:
  - Train a model for **binary classification**.
  - The model outputs the probability that a vector belongs to the positive class.
  - These probabilities are then used as ranking scores.

Experiments: multiple pairs

- We randomly sample the feature space to select 1000 pairs $\mathbf{v}, \mathbf{v}'$ such that $\mathbf{v} \preceq_f \mathbf{v}'$.
- For each pair, we then compute an AXp.

Exp. Size	Avg Time (s)	Std Dev (s)	Support
9	2.38	0.47	49
8	5.75	3.87	236
7	14.51	12.45	393
6	37.03	36.02	259
5	95.64	70.05	62
4	314.75	0.00	1
Overall	23.01	35.88	1000

Experiments: fixed pair

Feature Vectors and Abductive Explanations

$\mathcal{F}$	1	2	3	4	5	6	7	8	9
$\mathbf{v}$	2	2	3	0	1	1	1	3	0
$\mathbf{v}'$	4	0	3	3	2	2	0	2	1
$\mathcal{S}_1$	1	0	1	1	1	1	0	0	1
$\mathcal{S}_2$	1	0	1	0	1	1	1	0	1
$\mathcal{S}_3$	1	0	1	0	1	1	0	1	1

Scores:

- $\mathcal{S}_1 = 0.305$, $\mathcal{S}_2 = 0.002$, and $\mathcal{S}_3 = 0.292$.

Exp. Size	**Avg Time (s)**	**Support**
7	24.03	6
6	65.24	4
Overall	40.51	10

Summary and Future Work

Contributions

- First definitions of *abductive explanations* for ranking functions.
  - resembling those for classifiers but not reducible to them
- Proof-of-concept showing the practical feasibility of the approach.

Main Bottleneck: Scalability

- Use Automated Reasoning to efficiently verify *WeakAXp*.
- Probabilistic Formal Explanations.

Thank you for your attention!