Formal Explanations

From Classifiers to Rankers

Francesco Chiariello

Deep Learning Revolution

Milestones in Deep Learning

- **2012:** the CNN **AlexNet** wins the ImageNet Challenge, showcasing the power of DL techniques
- **2013-2014:** **VAE** (Variational Autoencoder) and **GANs** (Generative Adversarial Networks) are introduced, marking the first major success of **Generative AI**
- **2013-2015:** **DQNs** (Deep Q-Networks) achieve human-level performance on Atari games
- **2016:** **AlphaGo** defeats the world Go champion
- **2017:** **Transformer** architecture revolutionizes sequence modeling
- **2022:** **ChatGPT** popularizes large-scale language models

Deep Learning Applications

As deep learning performance continues to improve, its range of applications continues to expand, including

- **High-risk:**
  - Critical infrastructure
  - Creditworthiness
  - Law enforcement
  - Biometric data
- **Safety-critical:**
  - Self-driving cars
  - Unmanned aerial vehicles
  - ...

eXplainable Artificial Intelligence (XAI)

- While models become larger, more complex, more powerful, and widespread, they remain **opaque**.
- There is, therefore, an increasing need to **explain** them.
- **XAI** is dedicated to helping human decision-makers understand the decisions made by ML systems, to deliver **Trustworthy AI**.

XAI Approaches

Popular XAI approaches include:

- **LIME** (Local Interpretable Model-agnostic Explanations) Ribeiro et al., 2016
    - Produces interpretable models that locally approximate the behavior of the original model around a specific prediction.
- **SHAP** (SHapley Additive exPlanations) Lundberg and Lee, 2017
    - Assigns feature importance based on Shapley values Shapley, 1953.
- **Anchors** Ribeiro et al., 2018
    - Identifies a set of features that, with high precision, "anchor" a prediction.

However, these approaches are based on heuristic methods and provide **no formal guarantees** of rigour.

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

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

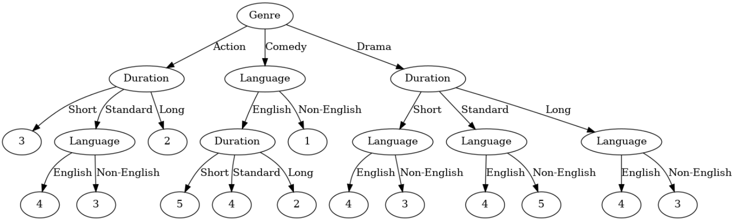
Classifiers

- **Classifier**: Given a set of classes $\mathcal{K} = \{c_1, \ldots, c_k\}$, a classifier is a function
$$\kappa : \mathbb{F} \rightarrow \mathcal{K}$$
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 $(\mathbf{x}, c)$.
- In what follows, we assume the classifier is given
- **Explanation problem**: given the classifier κ and a $\mathbf{v} \in \mathbb{F}$, **why** κ predict $\kappa(\mathbf{v})$ on $\mathbf{v}$?

Running example: Classifier

- $\mathcal{F} = \{\text{Genre}, \text{Dur.}, \text{Lang.}\}$
- $\mathcal{K} = \{1, 2, 3, 4, 5\}$

- $D_{\text{Genre}} = \{\text{Action}, \text{Comedy}, \text{Drama}\}$
- $D_{\text{Dur.}} = \{\text{Short}, \text{Standard}, \text{Long}\}$
- $D_{\text{Lang.}} = \{\text{English}, \text{Non-English}\}$



- $\mathbf{v} = \langle \text{Comedy, Long, Non-English} \rangle \mapsto 1$
- $\mathbf{v}' = \langle \text{Action, Standard, English} \rangle \mapsto 4$

Weak Abductive Explanation (WeakAXp)

- A set $\mathcal{S} \subseteq \mathcal{F}$ is a **Weak Abductive Explanation** 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 $\mathcal{S}$.

Theorem (Monotonicity)

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

Abductive Explanation (AXp)

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

  In other words, AXps are subset-minimal WeakAXps.

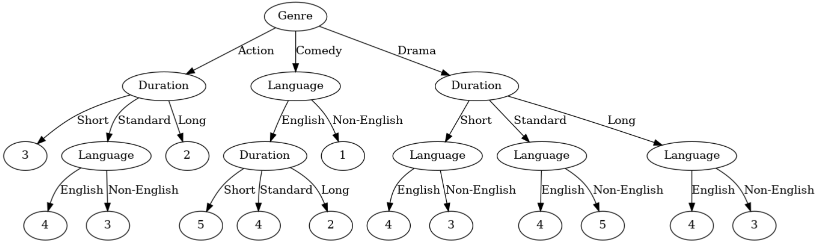
Observation

To verify condition (2), it is sufficient to consider only the maximal proper subsets of $\mathcal{S}$.

Property (2) can then be rewritten as follows:

$$\forall i \in \mathcal{S} : \neg \mathit{WeakAXp}(\mathcal{S} \setminus \{i\})$$

Running Example: Explanations



- $\mathcal{F}$ = {Genre, Duration, Language}
- $\mathbf{v} = \langle$Comedy, Long, Non-English$\rangle \mapsto 1$
  - **AXps:** {Genre, Language}
- $\mathbf{v}' = \langle$Action, Standard, English$\rangle \mapsto 4$
  - **AXps:** {Duration,Language}

Contrastive Explanation (CXp)

- A set $\mathcal{S} \subseteq \mathcal{F}$ is a **Weak Contrastive Explanation** (WeakCXp) if

$$\exists \mathbf{x} \in [\mathbf{v}]_{\mathcal{F} \setminus \mathcal{S}}, \kappa(\mathbf{x}) \neq \kappa(\mathbf{v})$$

  i.e., even by fixing all the features not in $\mathcal{S}$, the prediction still change.
- A Contrastive Explanation is a subset-minimal WeakCXp.

AXps and CXps

- **AXp**: subset-minimal set of features to ensure the predictions
- **CXp**: subset-minimal set of features to change the predictions
- **Duality**: AXps are **Minimal Hitting Sets** of CXps and vice-versa

Rankings and Preorders

- Given a set S, a **preorder** $\preceq$ on S is a binary relation on S that is both
  - **Reflexive**: $\forall a \in S, a \preceq a$.
  - **Transitive**: $\forall a, b, c \in S, a \preceq b \wedge b \preceq c \implies a \preceq c$.
- A **ranking** $\preceq$ is a preorder which is also
  - **Strongly connected**: $\forall a, b \in S, a \preceq b \vee b \preceq a$.

Orders

- An order $\preceq$ is a preorder that is also
  - **Antisymmetric**: $\forall a, b \in S, a \preceq b \wedge b \preceq a \implies a = b$.
- We call **linear order** an order that is also strongly connected.
- Preorders are more general than orders in that they admit ties.

Ranking Functions (or Rankers)

- A **ranking function** on S is a function $f : S \to \mathbb{R}$.
  - The value $f(a) \in \mathbb{R}$ represents the *score* assigned to $a \in S$.
- The ranker f on S induce a ranking $\preceq_f$ on S, defined by

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

- Conversely, given a ranking $\preceq$ on S there exists a ranking function f on S, such that $\preceq = \preceq_f$.
- **Note:** Rankings and ranking functions are also referred to as *preferences* and *utility functions* in microeconomic theory.

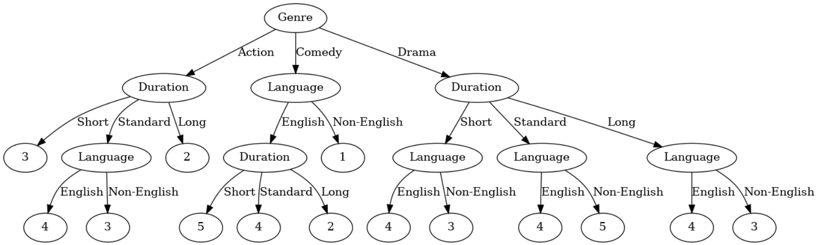
Classifiers as Rankers

- Let $\kappa : \mathbb{F} \to \mathcal{K}$ a classifier with $\mathcal{K} = \{c_1, \ldots, c_k\}$ linearly ordered, i.e., $c_i \preceq_{\mathcal{K}} c_{i+1}$. Such a classifier induces a ranking $\preceq_\kappa$ defined by

$$\mathbf{x} \preceq_\kappa \mathbf{x}' \iff \kappa(\mathbf{x}) \preceq_{\mathcal{K}} \kappa(\mathbf{x}').$$

- The classifier κ itself can be identified with the ranking function $f : \mathbb{F} \to \{1, \ldots, k\}$ by identifying $c_i = i$, for $i = 1, \ldots, k$.

Running Example: Classifier as ranker



Given the two points

- $\mathbf{v} = \langle \text{Comedy, Long, Non-English} \rangle \mapsto 1$
- $\mathbf{v}' = \langle \text{Action, Standard, English} \rangle \mapsto 4$

the decision tree classifier defines the rank $\mathbf{v} \preceq \mathbf{v}'$.

Explanation Problem

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 $\kappa : \mathbb{F}^2 \rightarrow \{0, 1\}$, defined by

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

- One can then apply FXAI for classifiers to κ

Reduction to Classification

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

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

- One can then apply FXAI for classifiers to κ

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.

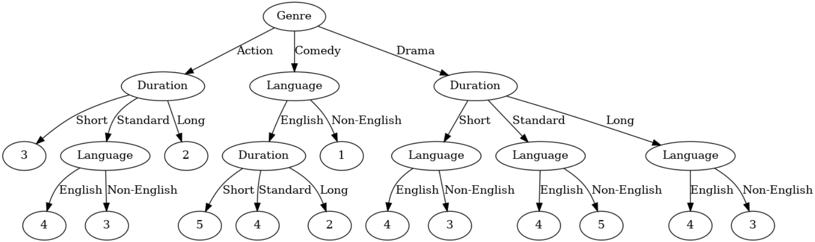
Abductive Explanations

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

Running Example



Given the two points

- $\mathbf{v} = \langle \text{Comedy, Long, Non-English} \rangle$
- $\mathbf{v}' = \langle \text{Action, Standard, English} \rangle$

AXps for why $\mathbf{v} \preceq \mathbf{v}'$ are the following:
{Duration, Language}, {Genre, Language}, {Genre, Duration}.

Properties

Theorem (Monotonicity)

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

Properties

Theorem (Monotonicity)

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

Theorem (Granularity)

If $\forall \mathbf{x}, \mathbf{x}' \in \mathbb{F} : (\mathbf{x} \preceq_1 \mathbf{x}' \implies \mathbf{x} \preceq_2 \mathbf{x}')$ *then every WeakAXp of* $\preceq_1$ *is also a WeakAXp of* $\preceq_2$.

Which Explanation to Prefer?

- AXps are not unique.
- Multiple cardinality-minimal AXps may exist.
- This raises the question: which explanation should be preferred?
- We address this by defining a preference relation over sets of features of the same size.

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

Preference Relation: $$\mathcal{S}_1 \preceq \mathcal{S}_2 \iff \mathit{score}(\mathcal{S}_1) \leq \mathit{score}(\mathcal{S}_2)$$

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

The score is particularly important when f has an intrinsic meaning.

Comparing Multiple Vectors

- So far, we have only considered pairwise comparisons.
- We now address full rankings:

$$\mathbf{v}^{(1)} \preceq_f \cdots \preceq_f \mathbf{v}^{(n)}$$

- A set $\mathcal{S} \subseteq \mathcal{F}$ is a **WeakAXp** if:

$$\forall\, (\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}) \in [\mathbf{v}^{(1)}]_{\mathcal{S}} \times \cdots \times [\mathbf{v}^{(n)}]_{\mathcal{S}},\ \mathbf{x}^{(1)} \preceq_f \cdots \preceq_f \mathbf{x}^{(n)}$$

Algorithms for Model-agnostic Explanations

- In the following, we shall see how to compute an AXp.
- The proposed approach is model-agnostic, requiring only black-box access to the model.
- We then test our approach on a neural network model that estimates the probability of breast cancer recurrence.

Verify a WeakAXp

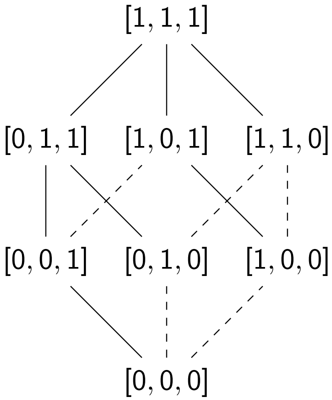
$$\text{WeakAXp}(\mathcal{S}) \iff \forall(\mathbf{x}, \mathbf{x}') \in [\mathbf{v}]_{\mathcal{S}} \times [\mathbf{v}']_{\mathcal{S}}, \mathbf{x} \preceq_f \mathbf{x}'$$

Input: $\mathcal{S} \subseteq \mathcal{F}$
Output: WeakAXp($\mathcal{S}$)

for $\mathbf{x} \in [\mathbf{v}]_{\mathcal{S}}$ **do**
  $f\_x \leftarrow f(\mathbf{x})$
  for $\mathbf{x}' \in [\mathbf{v}']_{\mathcal{S}}$ **do**
    $f\_x' \leftarrow f(\mathbf{x}')$
    if $f\_x > f\_x'$ **then**
      return false
    end if
  end for
end for
return true

Compute an AXp

The monotonicity of WeakAXps allows for efficient computation of an AXp.



Subset lattice for 3 features

Input: $\mathcal{S} \subseteq \mathcal{F}, start = 0$
Output: S

```
for i ← start to m − 1 do
    if S[i] = 1 then
        S[i] ← 0
        if WeakAXp(𝒮) then
            return DFS-AXp(S, i+1)
        end if
        S[i] ← 1 {Backtrack}
    end if
end for
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
With recurrence	85
Recurrence rate	$\approx$ 30%

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

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

Dataset Preparation

- We denote cancer recurrence with 1 and its absence with 0.
- To enable the neural network to handle categorical variables, we one-hot encode them.
- This results in a 43-dimensional feature space, representing 299376 distinct possible patients.

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**

Experiments: multiple pairs

- We randomly sample the feature space to select 500 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	2.49	0.65	27
8	6.55	4.18	104
7	19.67	16.90	212
6	42.02	39.08	123
5	129.37	78.33	32
4	353.58	14.11	2
Overall	29.87	46.70	500

Experiments: fixed pair

Feature Vectors and Abductive Explanations

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

Scores:

- $score(\mathcal{S}_1) = 0.056$; $score(\mathcal{S}_2) = 0.002$.

Exp. Size	Avg Time (s)	Support
7	67.04	3
6	74.72	3
5	157.33	4
Overall	105.46	10

Conclusions

In this talk, we have

- seen how to apply Formal Explainability to ranking functions
- implemented our approach and tested on real-world data on a real application, showing its feasibility

The bottleneck remains the scalability of the approach. To address this, we see two possibilities

- the use of a model-based approach that leverages Automated Reasoning tools.
- the use of probabilistic explanations.

Thank you for your attention!

Financé par
GOUVERNEMENT
Liberté
Égalité
Fraternité

FRANCE RELANCE

Financé par
l'Union européenne
NextGenerationEU

References I

- Shapley, L. S. (1953). A value for n-person games. In H. W. Kuhn & A. W. Tucker (Eds.), *Contributions to the theory of games II* (pp. 307–317). Princeton University Press.
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016)."why should I trust you?": Explaining the predictions of any classifier. *KDD*, 1135–1144.
- Lundberg, S. M., & Lee, S. (2017).A unified approach to interpreting model predictions. *NIPS*, 4765–4774.
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2018).Anchors: High-precision model-agnostic explanations. *AAAI*, 1527–1535.