39th Conference on Uncertainty in Artificial Intelligence
Pittsburgh, PA, USA
31st July – 4th August, 2023
uai2023

Approximating Probabilistic Explanations via Supermodular Minimization

Louenas Bounia [1], Frédéric Koriche [1]

[1]CRIL, Université d'Artois – CNRS, Lens, France

{name}@cril.fr

Motivations of the work

- Capacity Constraints: cognitive limitations refer to the finite capacity and processing limitations of the human mind [Miller 56]
- Abductive explanations are often too large to be interpretable and computing probabilistic explanations is an NP-hard problem

Notations and Problem Formulation

- **Error Function:** Given a classifier $h : \{0,1\}^d \longrightarrow \{0,1\}$, and an instance $x \in \{0,1\}^d$ for which the prediction $h(x)$ must be explained, let $\epsilon_{h,x} : 2^{[d]} \longrightarrow \mathbb{R}^+$ denote the error function given by:

$$\epsilon_{h,x}(S) = \frac{|\{z \in \{0,1\}^d : h(z) \neq h(x) \text{ and } z_S = x_S\}|}{|\{z \in \{0,1\}^d : z_S = x_S\}|} = \frac{\mu_{h,x}(S)}{2^{d-|S|}} \quad (1)$$

- **Probabilistic Explanation:** Given a precision parameter $\sigma \in [0,1)$. An explanation S is called $(1-\sigma)$-probable if S satisfies : $\epsilon_{h,x}(S) \leq 1-\sigma$.
- **Problem 1:** Given a classifier $h : \{0,1\}^d \to \{0,1\}$, an instance $x \in \{0,1\}^d$, a set $I \subseteq \{1,2,\ldots,d\}$ of features, and a size limit $k \leq |I|$, find a subset $S \subseteq I$ of size at most k such that $\epsilon_{h,x}(S)$ is minimized

Supermodular Minimization

- **Proposition 1:** Let $h : \{0,1\}^d \to \{0,1\}$ be a classifier, $x \in \{0,1\}^d$ an instance, and $I \subseteq [d]$. $\mu_{h,x}(.)$ is supermodular and non-increasing.

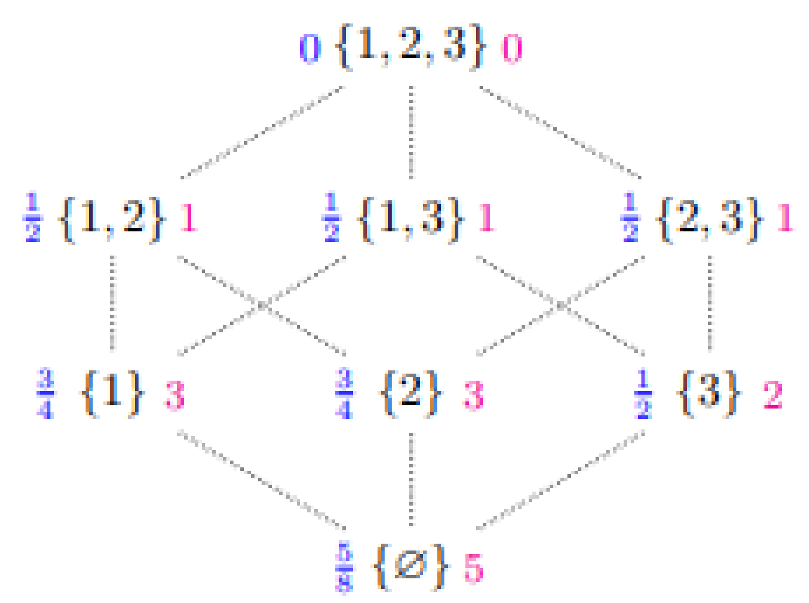
Algorithm 1: Greedy Descent (GD)

Input: classifier h, instance $\boldsymbol{x}$, feature set I, integer k
Set $S_n = I$, where $n = |I|$
For $j = n$ **downto** 1 **do**
  Let $i^* \in \text{Argmin}_{i \in S_j} \mu_{h,\boldsymbol{x}}(S_j \setminus \{i\})$
  Set $S_{j-1} = S_j \setminus \{i^*\}$
Let $S_{\text{GD}} \in \text{Argmin}_{S \in \{S_0, S_1, \cdots, S_k\}} \epsilon_{h,\boldsymbol{x}}(S)$
Return S_{GD}

Algorithm 2: Greedy Ascent (GA)

Input: classifier h, instance $\boldsymbol{x}$, feature set I, integer k
Let c be the curvature of $\mu_{h,\boldsymbol{x}}(\cdot)$ over 2^I
Set $j = 0$, $S_0 = \varnothing$ and $\gamma = \max\{\frac{1}{e}, c\}$
Repeat
  Let $i^* \in \text{Argmin}_{i \in I \setminus S_{j-1}} \mu_{h,\boldsymbol{x}}(S_{j-1} \cup \{i\})$
  Set $S_j = S_{j-1} \cup \{i^*\}$
Until $j = k \left\lceil \ln\left(\frac{\mu_{h,\boldsymbol{x}}(\varnothing)}{\gamma \cdot \mu_{h,\boldsymbol{x}}(S_j)}\right)\right\rceil$
Let $S_{\text{GA}} \in \text{Argmin}_{S \in \{S_0, S_1, \cdots, S_j\}} \epsilon_{h,\boldsymbol{x}}(S)$
Return S_{GA}

Example 1: Consider the classifier $h : \{0,1\}^3 \longrightarrow \{0,1\}$ specified by the function: $h(x) = 1 \iff x_1x_2x_3 + x_1x_2 - x_1 - x_2 \geq 0$



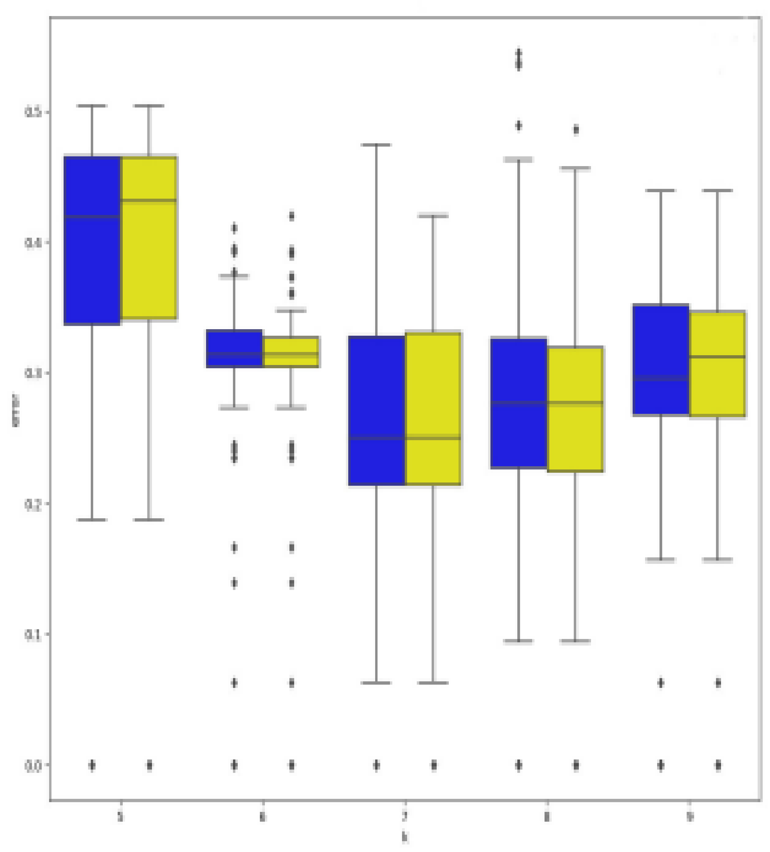
- **Proposition 2:** Let S^* be an optimal solution of problem 1, let c be the curvature of $\mu_{h,x}(.)$ over 2^I, and assume that I is an abductive explanation for h and x. Then, the solution S_{GD} and S_{GA} returned by GD and GD satisfies:

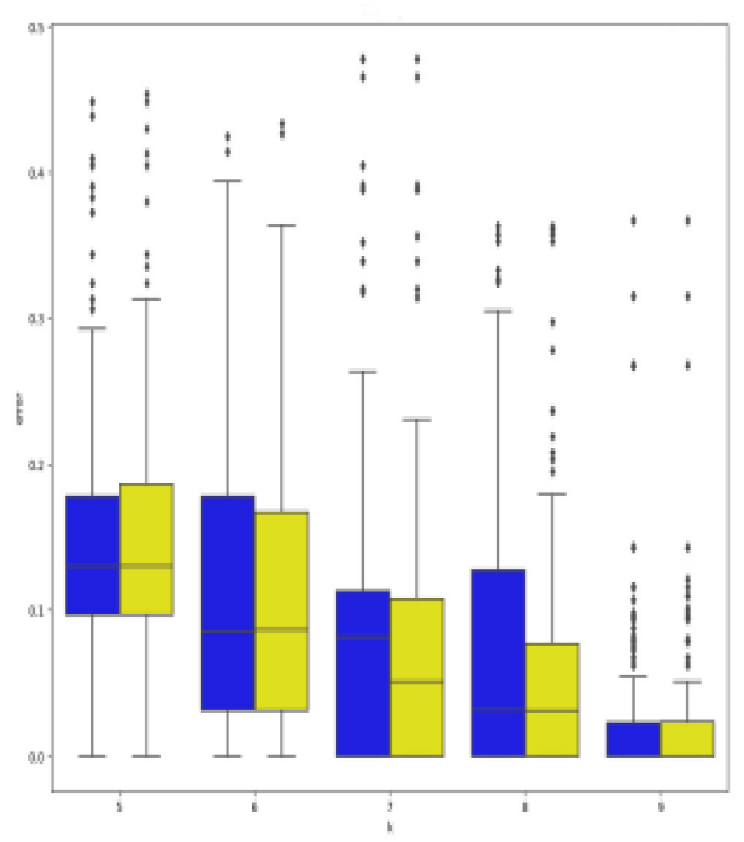
$$\epsilon_{h,x}(S_{GD}) \leq \left(\frac{e^p-1}{p}\right)\epsilon_{h,x}(S^*) \text{ and } \epsilon_{h,x}(S_{GA}) \leq \left(\frac{1}{1-c}\right)\epsilon_{h,x}(S^*)$$
where $p = \frac{c}{1-c} < 1$

The figure shows the error $\epsilon_{h,x}(.)$ (blue), the number of mistakes $\mu_{h,x}(S)$ (magenta). Given the instance $x = (1,1,1)$ for which we need to explain $h(x) = 1$, and using the Hasse diagram in figure, we observe that $\{x_1, x_2, x_3\}$ is abductive explanation. However, $\{x_1, x_2\}$ and $\{x_3\}$ are subset-minimal explanations with a probability at least $\frac{1}{2}$ for h and x.

Empirical results "Experimental results on 25 benchmarks for decision tree, using $k = 7 \pm 2$"

Benchmark				$\epsilon_{h,x}(S)$			$\|S\|$			Time (s)
name	acc	d	$\|I\|$	GA	GD	SAT	GA	GD	SAT	SAT
meta-data	87.42	44	5.09	0.08 (±0.11)	0.08 (±0.11)	0.08 (±0.11)	3.10	3.10	3.10	12.14
glass	78.46	31	5.38	0.26 (±0.11)	0.26 (±0.11)	0.26 (±0.11)	2.14	2.14	2.14	2.36
student perf.	91.79	30	5.41	0.26 (±0.11)	0.26 (±0.11)	0.26 (±0.11)	2.00	2.00	2.00	2.16
primary tumor	84.31	23	6.23	0.09 (±0.09)	0.09 (±0.09)	0.09 (±0.08)	4.22	4.22	4.22	3.58
liver disorders	75.96	58	6.38	0.18 (±0.09)	0.18 (±0.08)	0.18 (±0.08)	4.00	4.00	4.00	27.33
schizophrenia	80.39	33	6.39	0.37 (±0.24)	0.37 (±0.24)	0.37 (±0.24)	1.27	1.27	1.27	4.79
hungarian	62.92	13	6.65	0.12 (±0.12)	0.12 (±0.12)	0.11 (±0.10)	3.58	3.56	3.56	1.68
horse colic	75.68	40	6.73	0.14 (±0.07)	0.13 (±0.07)	0.13 (±0.07)	4.03	4.06	4.06	11.56
indian liver	64.57	84	8.21	0.10 (±0.09)	0.10 (±0.09)	0.16 (±0.12)	5.08	4.89	6.12	176.28
pima indians	75.32	97	8.30	0.15 (±0.14)	0.15 (±0.14)	0.16 (±0.12)	5.85	5.84	6.58	484.6
loan eligibility	74.31	68	8.47	0.19 (±0.13)	0.18 (±0.13)	0.20 (±0.14)	5.60	5.70	6.82	42.87
patient treat.	66.01	10	8.92	0.05 (±0.09)	0.03 (±0.06)	0.03 (±0.08)	5.63	5.94	5.94	24.08
wine	69.58	11	9.03	0.09 (±0.10)	0.09 (±0.09)	0.09 (±0.12)	5.59	5.64	5.62	36.32
employee attr.	82.45	63	10.56	0.06 (±0.09)	0.06 (±0.09)	0.20 (±0.11)	6.41	6.39	6.98	1017.24
contraceptive	51.36	90	10.84	0.06 (±0.08)	0.06 (±0.08)	0.39 (±0.17)	4.27	4.26	5.95	1096.07
compas	67.60	40	10.95	0.03 (±0.07)	0.04 (±0.08)	0.05 (±0.09)	5.68	5.83	6.78	1082.32
fetal health	91.85	93	11.33	0.12 (±0.06)	0.12 (±0.06)	0.23 (±0.11)	5.59	5.59	6.00	930.61
dorothea	91.88	10^5	12.90	0.25 (±0.10)	0.25 (±0.10)	–	6.70	6.70	–	–
bank market.	89.49	882	13.11	0.29 (±0.08)	0.29 (±0.07)	–	6.99	6.99	–	–
mnist49	95.99	784	15.57	0.37 (±0.14)	0.37 (±0.14)	–	6.97	6.89	–	–
spambase	92.11	236	16.09	0.24 (±0.11)	0.23 (±0.09)	–	6.87	6.87	–	–
mnist38	96.42	784	17.89	0.37 (±0.13)	0.38 (±0.14)	–	6.93	6.93	–	–
cnae	92.59	856	19.07	0.32 (±0.25)	0.32 (±0.25)	–	5.97	5.97	–	–
gisette	94.10	5000	21.42	0.32 (±0.11)	0.32 (±0.11)	–	6.88	6.88	–	–
farm ads	80.78	54877	23.15	0.13 (±0.17)	0.13 (±0.17)	–	6.31	6.31	–	–





Conclusion and future work

- The experimental results demonstrate that our greedy algorithms are highly effective for approximating a probabilistic explanation
- Approximating an abductive explanation of minimal size for a Boolean classifier through supermodular optimization
- Extending approximation algorithms to hypothesis classes for which the problem of evaluating $\mu_{h,x}(.)$ is intractable using sampling methods