

Explaining robust classification through prime implicants

Hénoïk Willot, Sébastien Destercke & Khaled Belahcene

15th international conference on Scalable Uncertainty Management



SUM22 - October 17-19th 2022



• Introduction

- o Classification
- o Prime implicant

• Naive Credal Classifier [3]

- o General case
- Prime implicants formulation
- o Computation

Conclusion





Recommend : class $\mathbf{y} \in \mathscr{Y} = \{y_1, \dots, y_m\}$ Features : $\mathscr{X}^N = \prod_{i=1}^n \mathscr{X}_i$ Discrete domains : $\mathscr{X}_i = \{x_i^1, \dots, x_i^{k_i}\}$ Observation : $\mathbf{x}^o \in \mathscr{X}^N$





Recommend : class $\mathbf{y} \in \mathscr{Y} = \{y_1, \dots, y_m\}$ Features : $\mathscr{X}^N = \prod_{i=1}^n \mathscr{X}_i$ Discrete domains : $\mathscr{X}_i = \{x_i^1, \dots, x_i^{k_i}\}$ Observation : $\mathbf{x}^o \in \mathscr{X}^N$

Crisp case : One probability distribution *p*

$$\mathbf{y} \succeq_{\rho} \mathbf{y}' \text{ if } \rho(\mathbf{y} | \mathbf{x}^{o}) \ge \rho(\mathbf{y}' | \mathbf{x}^{o})$$

 \Rightarrow Explanations by prime implicants are known





Credal case :

Probability distribution p replaced by convex sets of probabilities \mathcal{P}





Credal case :

Probability distribution p replaced by convex sets of probabilities \mathcal{P}

Robust classification :

Necessary recommendation $\mathbf{y} \succeq_{\mathscr{P}} \mathbf{y}'$,

$$\mathbf{y} \succeq_{\mathscr{P}} \mathbf{y}' \Leftrightarrow \forall p \in \mathscr{P}, \ p(\mathbf{y}|\mathbf{x}^{o}) \ge p(\mathbf{y}'|\mathbf{x}^{o}) \Leftrightarrow \inf_{p \in \mathscr{P}} \frac{p(\mathbf{y}|\mathbf{x}^{o})}{p(\mathbf{y}'|\mathbf{x}^{o})} \ge 1$$

 \Rightarrow What happens to prime implicants in this case?





Introduction Running example [1]

Objective : predict an animal in $\mathscr{Y} = \{ \mathfrak{P}, \mathfrak{W}, \mathfrak{M}, \mathfrak{M} \}$ Features : 3 lengths :

- \mathscr{X}_1 : ears
- \mathscr{X}_2 : tail
- *X*₃ : hair

Domains : $\mathscr{X}_i = \{Long, Medium, Short\}$ Observation : $\mathbf{x}^o = (Long, Short, Long)$

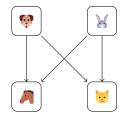




Introduction Running example [1]

Observation : $\mathbf{x}^{o} = (Long, Short, Long)$ Modèle :

- $p(\mathfrak{P}|\mathbf{x}^{o}) \in [0.30, 0.42]$
- $p(\forall | \mathbf{x}^{o}) \in [0.03, 0.15]$
- *p*(*▶*|**x**^o) ∈ [0.06, 0.18]
- $p(\forall | \mathbf{x}^{o}) \in [0.18, 0.42]$



$$\begin{array}{l} \textcircled{P} \succeq \mathscr{P} & \fbox{p(\textcircled{P} | \mathbf{x}^{o})} \\ \end{array} & \texttt{because inf}_{p \in \mathscr{P}} \frac{p(\textcircled{P} | \mathbf{x}^{o})}{p(\textcircled{P} | \mathbf{x}^{o})} = \frac{0.30}{0.18} \ge 1 \\ \\ \textcircled{P} & \texttt{and} & \bigstar \texttt{indifferent} : \\ \texttt{inf}_{p \in \mathscr{P}} \frac{p(\textcircled{P} | \mathbf{x}^{o})}{p(\textcircled{P} | \mathbf{x}^{o})} = \frac{0.30}{0.42} < 1 \ \texttt{et inf}_{p \in \mathscr{P}} \frac{p(\textcircled{P} | \mathbf{x}^{o})}{p(\textcircled{P} | \mathbf{x}^{o})} = \frac{0.18}{0.42} < 1 \end{aligned}$$





Introduction Implicant

 $E \subseteq N$, as a subset of feature indices, is an implicant of decision $\mathbf{y} \succeq_{\mathscr{P}} \mathbf{y}'$:

$$\phi(E) = \inf_{\substack{p \in \mathscr{P} \\ x_{-E} \in \mathscr{X}^{-E}}} \frac{p(\mathbf{y} | \mathbf{x}_{E}^{o}, x_{-E})}{p(\mathbf{y}' | \mathbf{x}_{E}^{o}, x_{-E})} \ge 1$$

i.e. observe \mathbf{x}_{E}^{o} is sufficient to conclude $\mathbf{y} \succeq_{\mathscr{P}} \mathbf{y}'$ no matter the values on other attributes $x_{-E} \in \mathscr{X}^{-E}$





Introduction Prime implicant

 $E \subseteq N$ is a *prime* implicant if

```
\forall i \in E, \ \phi(E \setminus \{i\}) < 1
```

i.e. E is minimal

For one decision, it might exists different prime implicants with different cardinals !





• Introduction

- o Classification
- o Prime implicant

• Naive Credal Classifier [3]

- o General case
- o Prime implicants formulation
- o Computation

Conclusion



Definition

Bayes theorem :

$$\rho(\mathbf{y}|\mathbf{x}^{o}) = \frac{\rho(\mathbf{x}^{o}|\mathbf{y}) \times \rho_{\mathscr{Y}}(\mathbf{y})}{\rho(\mathbf{x}^{o})}$$

Independence hypothesis (Naive Bayes) :

$$p(\mathbf{y}|\mathbf{x}^{o}) = \frac{\prod_{i=1}^{n} p_{i}(\mathbf{x}_{i}^{o}|\mathbf{y}) \times p_{\mathscr{Y}}(\mathbf{y})}{p(\mathbf{x}^{o})}$$

Features are independent, given the class



heudiasyc

Definition

Bayes theorem :

$$\rho(\mathbf{y}|\mathbf{x}^{o}) = \frac{\rho(\mathbf{x}^{o}|\mathbf{y}) \times \rho_{\mathscr{Y}}(\mathbf{y})}{\rho(\mathbf{x}^{o})}$$

Independence hypothesis (Naive Bayes) :

$$\rho(\mathbf{y}|\mathbf{x}^{o}) = \frac{\prod_{i=1}^{n} \rho_{i}(\mathbf{x}_{i}^{o}|\mathbf{y}) \times \rho_{\mathscr{Y}}(\mathbf{y})}{\rho(\mathbf{x}^{o})}$$

Features are independent, given the class We can rewrite $\phi(E)$:

$$\phi(E) = \inf_{\substack{\mathbf{x}_{-E} \in \mathscr{X}^{-E} \\ p_{\mathscr{Y}} \in \mathscr{P}_{\mathscr{X}_{i}}}} \frac{p_{\mathscr{Y}}(\mathbf{y})}{p_{\mathscr{Y}}(\mathbf{y}')} \underbrace{\prod_{i \in E} \frac{p_{i}(\mathbf{x}_{i}^{o}|\mathbf{y})}{p_{i}(\mathbf{x}_{i}^{o}|\mathbf{y}')}}_{\text{Implicant part}} \underbrace{\prod_{i \in -E} \frac{p_{i}(x_{i}|\mathbf{y})}{p_{i}(x_{i}|\mathbf{y}')}}_{\text{Adversarial part}}$$

heudiasyc



Convex probabilities

As $\mathcal{P}_{\mathscr{Y}}$ and $\mathcal{P}_{\mathscr{X}_i}$ are convex and $p_i(\cdot|\mathbf{y}')$ independent of $p_j(\cdot|\mathbf{y})$ if $\mathbf{y} \neq \mathbf{y}'$ or $i \neq j$:

$$\phi(E) = \inf_{x_{-E} \in \mathscr{X}^{-E}} \frac{\underline{p}_{\mathscr{Y}}(\mathbf{y})}{\overline{p}_{\mathscr{Y}}(\mathbf{y}')} \underbrace{\prod_{i \in E} \frac{\underline{p}_{i}(\mathbf{x}_{i}^{o}|\mathbf{y})}{\overline{p}_{i}(\mathbf{x}_{i}^{o}|\mathbf{y}')}}_{\text{Implicant part}} \underbrace{\prod_{i \in -E} \frac{\underline{p}_{i}(x_{i}|\mathbf{y})}{\overline{p}_{i}(x_{i}|\mathbf{y}')}}_{\text{Adversarial part}}$$
(1)

with \underline{p} and \overline{p} lower and upper bounds of $p \in \mathscr{P}$





Running example [2] Data

Data are obtained with the *Imprecise Dirichlet Model* [1] Idea : build a cautious interval around *p* using a number of fictive observations *s*

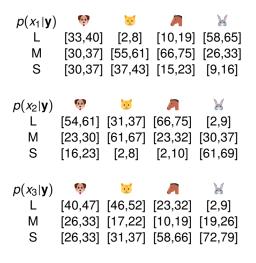
$$p(x) = \frac{n_x}{N} \stackrel{IDM}{\Rightarrow} p(x) \in \left[\frac{n_x}{N+s}, \frac{n_x+s}{N+s}\right]$$

To avoid null probabilities, we add a small regularization











ıtc

Recherche



 $\mathbf{x}^{o} = (Long, Short, Long)$

$p(x_1 \mathbf{y})$	Q	V	A	Ж
Ĺ	[33,40]	[2,8]	[10,19]	[58,65]
Μ	[30,37]	[55,61]	[66,75]	[26,33]
S	[30,37]	[37,43]	[15,23]	[9,16]
(-		*	
$p(x_2 \mathbf{y})$			2	
L	[54,61]	[31,37]	[66,75]	[2,9]
М	[23,30]	[61,67]	[23,32]	[30,37]
S	[16,23]	[2,8]	[2,10]	[61,69]
$p(x_3 \mathbf{y})$	Ģ	W	A	Ж
L	[40,47]	[46,52]	[23,32]	[2,9]
М	[26,33]	[17,22]	[10,19]	[19,26]
S		[31,37]		[72,79]





 $\mathbf{x}^{o} = (Long, Short, Long)$ $\mathfrak{P} \geq_{\mathscr{P}} \mathfrak{A}?$

$p(x_1 \mathbf{y})$	()	W	A	Ж
Ĺ	[33,40]	[2,8]	[10,19]	[58,65]
Μ	[30,37]	[55,61]	[66,75]	[26,33]
S	[30,37]	[37,43]	[15,23]	[9,16]
$p(x_2 \mathbf{y})$	(i)	V		Ж
L			[66,75]	
М		[61,67]		[30,37]
S		[2,8]		[61,69]
$p(x \mathbf{n})$	60	W		Ж
$p(x_3 \mathbf{y})$				
L			[23,32]	
М			[10,19]	
S	[26,33]	[31,37]	[58,66]	[72,79]





$$\mathbf{x}^{o} = (Long, Short, Long)$$
$$\mathfrak{P} \succeq_{\mathscr{P}} \mathfrak{A}?$$

$p(x_1 \mathbf{y})$	Ģ	W	A	X
L	[33,40]	[2,8]	[10,19]	[58,65]
М	[30,37]	[55,61]	[66,75]	[26,33]
S	[30,37]	[37,43]	[15,23]	[9,16]
$p(x_2 \mathbf{y})$	Ф.		2	Ж
L	[54,61]	[31,37]	[66,75]	[2,9]
Μ	[23,30]	[61,67]	[23,32]	[30,37]
S	[16,23]	[2,8]	[2,10]	[61,69]
$p(x_3 \mathbf{y})$	Ģ	W	2	Ж
L	[40,47]	[46,52]	[23,32]	[2,9]
Μ	[26,33]	[17,22]	[10,19]	[19,26]
S	[26,33]	[31,37]	[58,66]	[72,79]





$$\mathbf{x}^{o} = (Long, Short, Long)$$
$$\mathfrak{P} \succeq_{\mathcal{P}} \mathfrak{A}?$$

$$\begin{split} \phi(\mathbf{N}) &= \\ \underline{P}_{\mathscr{Y}}(\mathbf{y}) &= \\ \overline{p}_{\mathscr{Y}}(\mathbf{y}') \times \frac{\underline{p}_{1}(\mathbf{x}_{1}^{o}|\mathbf{y})}{\overline{p}_{1}(\mathbf{x}_{1}^{o}|\mathbf{y}')} \times \frac{\underline{p}_{2}(\mathbf{x}_{2}^{o}|\mathbf{y})}{\overline{p}_{2}(\mathbf{x}_{2}^{o}|\mathbf{y}')} \times \frac{\underline{p}_{3}(\mathbf{x}_{3}^{o}|\mathbf{y})}{\overline{p}_{3}(\mathbf{x}_{3}^{o}|\mathbf{y}')} \\ &= \frac{0.25}{0.22} \end{split}$$

p(x ₁ y) L M S	[33,40] [30,37]	[2,8] [55,61]	(10,19] [66,75]	[58,65] [26,33]
s p(x ₂ y) L M S	? [54,61] [23,30]	<mark>\</mark> [31,37]	[15,23] [66,75] [23,32] [2,10]	¥ [2,9] [30,37]
p(x ₃ y) L M S	? [40,47] [26,33]	<mark>\</mark> [46,52]	(23,32) [10,19]	





$$\mathbf{x}^{o} = (Long, Short, Long)$$
$$\mathfrak{P} \succeq_{\mathcal{P}} \mathfrak{A}?$$

$$\begin{split} \phi(\mathbf{N}) &= \\ \frac{p_{\mathcal{Y}}(\mathbf{y})}{\overline{p}_{\mathcal{Y}}(\mathbf{y}')} \times \frac{p_{1}(\mathbf{x}_{1}^{o}|\mathbf{y})}{\overline{p}_{1}(\mathbf{x}_{1}^{o}|\mathbf{y}')} \times \frac{p_{2}(\mathbf{x}_{2}^{o}|\mathbf{y})}{\overline{p}_{2}(\mathbf{x}_{2}^{o}|\mathbf{y}')} \times \frac{p_{3}(\mathbf{x}_{3}^{o}|\mathbf{y})}{\overline{p}_{3}(\mathbf{x}_{3}^{o}|\mathbf{y}')} \\ &= \frac{0.25}{0.22} \times \frac{0.33}{0.19} \end{split}$$

p(x ₁ y) L M S	[<mark>33</mark> ,40]	[2,8] [55,61]	(10,19) [66,75] [15,23]	[58,65]
p(x ₂ y) L M S	? [54,61] [23,30]	; [31,37]	(66,75] [23,32]	¥ [2,9]
p(x ₃ y) L M S	[40,47] [26,33]		[23,32] [10,19]	





$$\mathbf{x}^{o} = (Long, Short, Long)$$
$$\mathfrak{P} \succeq_{\mathcal{P}} \mathfrak{A}?$$

$$\begin{split} \phi(\mathbf{N}) &= \\ \frac{p_{\mathcal{Y}}(\mathbf{y})}{\overline{p}_{\mathcal{Y}}(\mathbf{y}')} \times \frac{p_{1}(\mathbf{x}_{1}^{o}|\mathbf{y})}{\overline{p}_{1}(\mathbf{x}_{1}^{o}|\mathbf{y}')} \times \frac{p_{2}(\mathbf{x}_{2}^{o}|\mathbf{y})}{\overline{p}_{2}(\mathbf{x}_{2}^{o}|\mathbf{y}')} \times \frac{p_{3}(\mathbf{x}_{3}^{o}|\mathbf{y})}{\overline{p}_{3}(\mathbf{x}_{3}^{o}|\mathbf{y}')} \\ &= \frac{0.25}{0.22} \times \frac{0.33}{0.19} \times \frac{0.16}{0.10} \end{split}$$

p(x ₁ y) L M S	[33,40] [30,37]	[2,8] [55,61]	/ [10,19] [66,75] [15,23]	[58,65] [26,33]
p(x ₂ y) L M S	[54,61] [23,30]	[31,37]	[66,75] [23,32]	
p(x ₃ y) L M S	[40,47] [26,33]	[17,22]	7 [23,32] [10,19] [58,66]	[19,26]





$$\mathbf{x}^{o} = (Long, Short, Long)$$
$$\mathfrak{P} \succeq_{\mathcal{P}} \mathfrak{A}?$$

$$\begin{split} \phi(\mathbf{N}) &= \\ \underline{p}_{\mathcal{Y}}(\mathbf{y}) &= \\ \overline{p}_{\mathcal{Y}}(\mathbf{y}') &\times \frac{\underline{p}_{1}(\mathbf{x}_{1}^{o}|\mathbf{y})}{\overline{p}_{1}(\mathbf{x}_{1}^{o}|\mathbf{y}')} \times \frac{\underline{p}_{2}(\mathbf{x}_{2}^{o}|\mathbf{y})}{\overline{p}_{2}(\mathbf{x}_{2}^{o}|\mathbf{y}')} \times \frac{\underline{p}_{3}(\mathbf{x}_{3}^{o}|\mathbf{y})}{\overline{p}_{3}(\mathbf{x}_{3}^{o}|\mathbf{y}')} \\ &= \frac{0.25}{0.22} \times \frac{0.33}{0.19} \times \frac{0.16}{0.10} \times \frac{0.40}{0.32} \end{split}$$

p(x ₁ y) L M S	[33,40] [30,37]	<mark>\</mark> [2,8] [55,61] [37,43]	[10,19] [66,75]	[58,65]
p(x ₂ y) L M S	[54,61] [23,30]	[31,37] [61,67] [2,8]	[66,75] [23,32]	
p(x ₃ y) L M S	[<mark>40</mark> ,47] [26,33]	₩ [46,52] [17,22] [31,37]	[23, <mark>32</mark>] [10,19]	[19,26]





$$\mathbf{x}^{o} = (Long, Short, Long)$$
$$\mathfrak{P} \succeq_{\mathcal{P}} \mathfrak{A}?$$

$$\begin{split} \phi(\mathbf{N}) &= \\ \frac{p_{\mathcal{Y}}(\mathbf{y})}{\overline{p}_{\mathcal{Y}}(\mathbf{y}')} \times \frac{\underline{p}_{1}(\mathbf{x}_{1}^{o}|\mathbf{y})}{\overline{p}_{1}(\mathbf{x}_{1}^{o}|\mathbf{y}')} \times \frac{\underline{p}_{2}(\mathbf{x}_{2}^{o}|\mathbf{y})}{\overline{p}_{2}(\mathbf{x}_{2}^{o}|\mathbf{y}')} \times \frac{\underline{p}_{3}(\mathbf{x}_{3}^{o}|\mathbf{y})}{\overline{p}_{3}(\mathbf{x}_{3}^{o}|\mathbf{y}')} \\ &= \frac{0.25}{0.22} \times \frac{0.33}{0.19} \times \frac{0.16}{0.10} \times \frac{0.40}{0.32} > 1 \end{split}$$

$p(x_1 \mathbf{y})$	ŵ	V	2	Х
L	[33,40]	[2,8]	[10,19]	[58,65]
М	[30,37]	[55,61]	[66,75]	[26,33]
S	[30,37]	[37,43]	[15,23]	[9,16]
$p(x_2 \mathbf{y})$	Ø	V	8	Ж
$P(\lambda_2 \mathbf{y})$			[66,75]	
M			[23,32]	
S				
5	[16,23]	[2,8]	[2,10]	[61,69]
$p(x_3 \mathbf{y})$	Ģ	V	7	Ж
L	[40,47]	[46,52]	[23,32]	[2,9]
Μ	[26,33]	[17,22]	[10,19]	[19,26]
S	[26,33]	[31,37]	[58,66]	[72,79]





Independence of the adversary

Equation (1) defines the adversarial part x_{-E} :

$$\phi(E) = \inf_{\substack{x_{-E} \in \mathscr{X}^{-E}}} \frac{\underline{p}_{\mathscr{Y}}(\mathbf{y})}{\overline{p}_{\mathscr{Y}}(\mathbf{y}')} \underbrace{\prod_{i \in E} \frac{\underline{p}_{i}(\mathbf{x}_{i}^{O}|\mathbf{y})}{\overline{p}_{i}(\mathbf{x}_{i}^{O}|\mathbf{y}')}}_{\text{Implicant part}} \underbrace{\prod_{i \in -E} \frac{\underline{p}_{i}(x_{i}|\mathbf{y})}{\overline{p}_{i}(x_{i}|\mathbf{y}')}}_{\text{Adversarial part}}$$





Independence of the adversary

Each $x_i^a \in x_{-E}$ is :

- 1. independent of \mathbf{x}_i^o
- 2. independent of every $j \in N \setminus \{i\}$

Therefore, \exists a **unique** "worst adversary" x^a for $\mathbf{y} \succeq_{\mathscr{P}} \mathbf{y}'$:

$$x^{a} \in \mathcal{X}^{N}$$
: $\forall i \in N \; x_{i}^{a} = \arg \min_{x_{i}^{k} \in \mathcal{X}_{i}} \frac{\underline{p}_{i}(x_{i}^{k}|\mathbf{y})}{\overline{p}_{i}(x_{i}^{k}|\mathbf{y}')}$





Independence of the adversary

Each $x_i^a \in x_{-E}$ is :

- 1. independent of \mathbf{x}_i^o
- 2. independent of every $j \in N \setminus \{i\}$

Therefore, \exists a **unique** "worst adversary" x^a for $\mathbf{y} \succeq_{\mathscr{P}} \mathbf{y}'$:

$$x^{a} \in \mathcal{X}^{N}$$
: $\forall i \in N \; x_{i}^{a} = \arg \min_{x_{i}^{k} \in \mathcal{X}_{i}} \frac{\underline{p}_{i}(x_{i}^{k}|\mathbf{y})}{\overline{p}_{i}(x_{i}^{k}|\mathbf{y}')}$

Let

$$C = \log \phi(\phi) = \log \left(\frac{\underline{\rho}_{\mathscr{Y}}(\mathbf{y})}{\overline{\rho}_{\mathscr{Y}}(\mathbf{y}')} \prod_{i \in N} \frac{\underline{\rho}_{i}(x_{i}^{a}|\mathbf{y})}{\overline{\rho}_{i}(x_{i}^{a}|\mathbf{y}')} \right)$$





Ģ	1		X
[25,26]	[29,31]	[20,22]	[25,26]

🦁 ≿*⊚ 🎮*

$$x_1^a = \arg\min\{\frac{0.33}{0.19}, \frac{0.30}{0.75}, \frac{0.30}{0.23}\} = M$$

p(x ₁ y) L M S	[<mark>33</mark> ,40] [<mark>30</mark> ,37]	[2,8]	(10,19) [66,75] [15,23]	[58,65] [26,33]
p(x ₂ y) L M S	[54,61] [23,30]	[31,37]	<pre>[66,75] [23,32] [2,10]</pre>	[2,9]
p(x ₃ y) L M S	[40,47] [26,33]	[46,52] [17,22]	(23,32) [10,19] [58,66]	[2,9] [19,26]





?? ≥*⊛* 🎘

$$\begin{aligned} x_1^a &= \arg\min\{\frac{0.33}{0.19}, \frac{0.30}{0.75}, \frac{0.30}{0.23}\} = M\\ x_2^a &= \arg\min\{\frac{0.54}{0.75}, \frac{0.23}{0.32}, \frac{0.16}{0.10}\} = M \end{aligned}$$

ρ(x ₁ y) L M S	[33,40] [30,37]	<mark>)</mark> [2,8] [55,61] [37,43]	[10,19]	[58,65] [26,33]
p(x ₂ y) L M S	[<mark>23</mark> ,30]	₩ [31,37] [61,67] [2,8]	[23, <mark>32</mark>]	
p(x ₃ y) L M S	[40,47] [26,33]	₩ [46,52] [17,22] [31,37]	[23,32] [10,19]	[2,9]





Ģ	U	A	X
[25,26]	[29,31]	[20,22]	[25,26]

💱 ≿*⊛ 🎮*

 $\begin{aligned} x_1^a &= \arg\min\{\frac{0.33}{0.19}, \frac{0.30}{0.75}, \frac{0.30}{0.23}\} = M\\ x_2^a &= \arg\min\{\frac{0.54}{0.75}, \frac{0.23}{0.32}, \frac{0.16}{0.10}\} = M\\ x_3^a &= \arg\min\{\frac{0.40}{0.32}, \frac{0.26}{0.19}, \frac{0.26}{0.66}\} = S \end{aligned}$

p(x ₁ y) L M S	[33,40]	[2,8] [55,61]	(10,19] [66,75] [15,23]	[58,65] [26,33]
p(x ₂ y) L M S	[54,61] [23,30]	[31,37]	<pre>[66,75] [23,32] [2,10]</pre>	[2,9]
p(x ₃ y) L M S	[40,47] [26,33]	[46,52] [17,22]	[23,32] [10,19] [58,66]	[2,9] [19,26]





Ģ			X
[25,26]	[29,31]	[20,22]	[25,26]

💱 ≿*⊛ 🎮*

 $\begin{aligned} x_1^a &= \arg\min\{\frac{0.33}{0.19}, \frac{0.30}{0.75}, \frac{0.30}{0.23}\} = M\\ x_2^a &= \arg\min\{\frac{0.54}{0.75}, \frac{0.23}{0.32}, \frac{0.16}{0.10}\} = M\\ x_3^a &= \arg\min\{\frac{0.40}{0.32}, \frac{0.26}{0.19}, \frac{0.26}{0.66}\} = S \end{aligned}$

$$C = \log\left(\frac{0.25}{0.22} \times \frac{0.30}{0.75} \times \frac{0.23}{0.32} \times \frac{0.26}{0.66}\right)$$
$$= -0.90$$

$p(x_1 \mathbf{y})$			/ [10,19]	
M S		[55,61]	[66,75]	[26,33]
$p(x_2 \mathbf{y})$. ,]		A	
L M	[54,61]	[31,37]	[66,75] [23,32]	[2,9]
S		[2,8]		[61,69]
$p(x_3 \mathbf{y})$	Ģ	W	2	Ж
L M	[26,33]	[17,22]	[23,32] [10,19]	[19,26]
S	[26,33]	[31,37]	[58,66]	[72,79]





Contribution of feature to the explanation

Let G(i) denote the contribution of feature *i* to function ϕ

$$G(i) = \log \phi(E \cup \{i\}) - \log \phi(E)$$

= $\left(\log \underline{p}_i(\mathbf{x}_i^o | \mathbf{y}) - \log \overline{p}_i(\mathbf{x}_i^o | \mathbf{y}')\right) - \left(\log \underline{p}_i(x_i^a | \mathbf{y}) - \log \overline{p}_i(x_i^a | \mathbf{y}')\right)$

Contribution of feature *i* is independent of other features !





 Image: Weight of the second second

 $\mathbf{x}^{o} = (Long, Short, Long)$ $\textcircled{P} \succeq_{\mathscr{P}} \blacksquare$

 $G(1) = (\log 0.33 - \log 0.19) - (\log 0.30 - \log 0.75) = 0.65$

Ð X $p(x_1|\mathbf{y})$ [33,40] [2,8] [10,19] [58,65] L Μ [<mark>30</mark>.37] [55,61] [66,75] [26,33] S [30,37] [37,43] [15,23] [9,16] Ģ 1.1 X $p(x_2|\mathbf{y})$ L [54,61] [31,37] [66,75] [2,9] Μ [23,30] [61,67] [23,32] [30,37] S [16,23] [2,8] [2,10] [61,69] 6 М $p(x_3|\mathbf{y})$ [40,47] [46,52] [23,32] [2,9] L Μ [26,33] [17,22] [10,19] [19,26] S [26,33] [31,37] [58,66] [72,79]





Ģ 🐱 🎢 👹 [25,26] [29,31] [20,22] [25,26]

 $\mathbf{x}^{o} = (Long, Short, Long)$ $\bigotimes \geq \varphi$

$$G(1) = (\log 0.33 - \log 0.19) - (\log 0.30 - \log 0.75) = 0.65$$
$$G(2) = (\log 0.16 - \log 0.10) - (\log 0.23 - \log 0.32) = 0.33$$

$p(x_1 \mathbf{y})$	Q	W	A	Х
L	[33,40]	[2,8]	[10,19]	[58,65]
Μ	[30,37]	[55,61]	[66,75]	[26,33]
S	[30,37]	[37,43]	[15,23]	[9,16]
$p(x_2 \mathbf{y})$	Ģ		2	Ж
L	[54,61]	[31,37]	[66,75]	[2,9]
Μ	[<mark>23</mark> ,30]	[61,67]	[23, <mark>32</mark>]	[30,37]
S	[<mark>16</mark> ,23]	[2,8]	[2, <mark>10</mark>]	[61,69]
$p(x_3 \mathbf{y})$	Ģ		2	Ж
L	[40,47]	[46,52]	[23,32]	[2,9]
Μ	[26,33]	[17,22]	[10,19]	[19,26]
S	[26,33]	[31,37]	[58,66]	[72,79]





 Image: Weight of the second second

 $\mathbf{x}^{o} = (Long, Short, Long)$ $\textcircled{P} \geq_{\mathscr{P}} \overleftarrow{\mathbb{A}}$

$$G(1) = (\log 0.33 - \log 0.19) - (\log 0.30 - \log 0.75) = 0.65 G(2) = (\log 0.16 - \log 0.10) - (\log 0.23 - \log 0.32) = 0.33 G(3) = (\log 0.40 - \log 0.32) - (\log 0.26 - \log 0.66) = 0.50$$

Ģ X $p(x_1|\mathbf{y})$ 1 [33,40] [2,8] [10,19] [58,65] L Μ [30,37] [55,61] [66,75] [26,33] S [30,37] [37,43] [15,23] [9,16] Ģ 1.1 X $p(x_2|\mathbf{y})$ L [54,61] [31,37] [66,75] [2,9] Μ [23,30] [61,67] [23,32] [30,37] S [16,23] [2,8] [2,10] [61,69] 60 $p(x_3|\mathbf{y})$ М **[40**,47] **[46**,52] **[23**,**32**] [2,9] L Μ [26,33] [17,22] [10,19] [19,26] S [26,33] [31,37] [58,66] [72,79]



Building E

We want $E \subseteq N$ such that :

$$\phi(E) \ge 1 \Leftrightarrow \log \phi(E) \ge 0$$

As $\log \phi$ is additive we have :

$$\log \phi(E) = C + \sum_{i \in E} G(i) \ge 0$$

As G(i)'s are independent, finding the smallest prime implicant is polynomial [2]







Computing E

Algorithm 1: Compute first prime implicants explanation

Input: $C : log(\phi(\phi))$; G: Contributions of criteria; **Output:** $XpI = (E, \mathbf{x}_E^o)$: PI explanation and associated values Order G in decreasing order, with σ the associated permutation $i \leftarrow 1$

while
$$\phi(E) + C < 0$$
 do
 $i \leftarrow i + 1$
 $E \leftarrow E \cup \{\sigma^{-1}(i)\}$
 $\phi(E) \leftarrow \phi(E) + G_{\sigma(i)}$
 $\chi_{pl} \leftarrow (E, \mathbf{x}_{E}^{o})$
return (χ_{pl})





X

 $p(x_1|\mathbf{y})$ Ģ 1 Х [33,40] [2,8] [10,19] [58,65] L Μ [30,37] [55,61] [66,75] [26,33] S [30,37] [37,43] [15,23] [9,16] $p(x_2|\mathbf{y})$ Ģ 1 1 X [54,61] [31,37] [66,75] [2.9] L Μ [23,30] [61,67] [23,32] [30,37] S [16,23] [2,8] [2,10] [61,69] $p(x_3|\mathbf{y})$ Ģ 199 X 67 L [40,47] [46,52] [23,32] [2.9] Μ [26,33] [17,22] [10,19] [19,26] S [26,33] [31,37] [58,66] [72,79]



21



• Introduction

- o Classification
- o Prime implicant

• Naive Credal Classifier [3]

- o General case
- Prime implicants formulation
- o Computation

Conclusion





Conclusion

Summary :

- Prime implicants for robust preferences
- Application to the NCC with convex domains
- Polynomial calculation of implicants

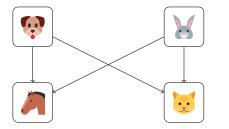
Perspectives :

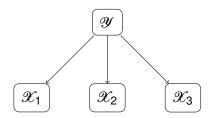
- Pairwise or holistic explanations?
- Implications on complexity to remove the independence hypothesis?
- Explanations for indifference?



Conclusion









SUM22 - October 17-19th 2022



References I

- Bernard, J.M. : An introduction to the imprecise dirichlet model for multinomial data. International Journal of Approximate Reasoning **39**(2-3), 123–150 (2005)
- Marques-Silva, J., Gerspacher, T., Cooper, M.C., Ignatiev, A., Narodytska, N. : Explaining Naive Bayes and Other Linear Classifiers with Polynomial Time and Delay. In : NeurIPS 2020, December 6-12, 2020, virtual (2020)
- Zaffalon, M. : The naive credal classifier. Journal of Statistical Planning and Inference **105**(1), 5–21 (2002)

