Explaining robust classification through prime implicants

Hénoïk Willot, Sébastien Destercke & Khaled Belahcene

15th international conference on Scalable Uncertainty Management

- Introduction
  - Classification
  - Prime implicant
- Naive Credal Classifier [3]
  - General case
  - Prime implicants formulation
  - Computation
- Conclusion

Introduction

Classification problem

Recommend : class $\mathbf{y} \in \mathscr{Y} = \{y_1, \ldots y_m\}$

Features : $\mathscr{X}^N = \prod_{i=1}^{n} \mathscr{X}_i$

Discrete domains : $\mathscr{X}_i = \{x_i^1, \ldots, x_i^{k_i}\}$

Observation : $\mathbf{x}^o \in \mathscr{X}^N$

Introduction

Classification problem

Recommend : class $\mathbf{y} \in \mathscr{Y} = \{y_1, \ldots y_m\}$

Features : $\mathscr{X}^N = \prod_{i=1}^{n} \mathscr{X}_i$

Discrete domains : $\mathscr{X}_i = \{x_i^1, \ldots, x_i^{k_i}\}$

Observation : $\mathbf{x}^o \in \mathscr{X}^N$

Crisp case :

One probability distribution p

$$\mathbf{y} \succeq_p \mathbf{y}' \text{ if } p(\mathbf{y}|\mathbf{x}^o) \geq p(\mathbf{y}'|\mathbf{x}^o)$$

$\Rightarrow$ Explanations by prime implicants are known

Introduction
Classification problem

Credal case :

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

Introduction
Classification problem

Credal case :
Probability distribution p replaced by convex sets of probabilities $\mathscr{P}$

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

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

$\Rightarrow$ What happens to prime implicants in this case ?

Introduction

Running example [1]

Objective : predict an animal in $\mathscr{Y} = \{$🐶, 🐱, 🐴, 🐰$\}$

Features : 3 lengths :

- $\mathscr{X}_1$: ears
- $\mathscr{X}_2$: tail
- $\mathscr{X}_3$: hair

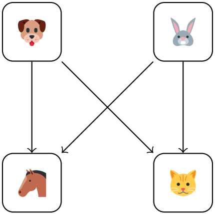
Domains : $\mathscr{X}_i = \{Long, Medium, Short\}$

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

Introduction
Running example [1]

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

- $p(\text{🐶}|\mathbf{x}^o) \in [0.30, 0.42]$
- $p(\text{🐱}|\mathbf{x}^o) \in [0.03, 0.15]$
- $p(\text{🐴}|\mathbf{x}^o) \in [0.06, 0.18]$
- $p(\text{🐰}|\mathbf{x}^o) \in [0.18, 0.42]$



$\text{🐶} \succeq_{\mathscr{P}} \text{🐴}$ because $\inf_{p \in \mathscr{P}} \frac{p(\text{🐶}|\mathbf{x}^o)}{p(\text{🐴}|\mathbf{x}^o)} = \frac{0.30}{0.18} \geq 1$

🐶 and 🐰 indifferent :

$$\inf_{p \in \mathscr{P}} \frac{p(\text{🐶}|\mathbf{x}^o)}{p(\text{🐰}|\mathbf{x}^o)} = \frac{0.30}{0.42} < 1 \text{ et } \inf_{p \in \mathscr{P}} \frac{p(\text{🐰}|\mathbf{x}^o)}{p(\text{🐶}|\mathbf{x}^o)} = \frac{0.18}{0.42} < 1$$

Introduction
Implicant

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

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

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
  - Classification
  - Prime implicant
- Naive Credal Classifier [3]
  - General case
  - Prime implicants formulation
  - Computation
- Conclusion

Definition

Bayes theorem :

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

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

Definition

Bayes theorem :

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

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

We can rewrite $\phi(E)$:

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

Convex probabilities

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

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

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

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

To avoid null probabilities, we add a small regularization

Running example [3]

🐶	🐱	🐴	🐰
[25, 26]	[29, 31]	[20, 22]	[25, 26]

$p(x_1 \mid \mathbf{y})$	🐶	🐱	🐴	🐰
L	[33,40]	[2,8]	[10,19]	[58,65]
M	[30,37]	[55,61]	[66,75]	[26,33]
S	[30,37]	[37,43]	[15,23]	[9,16]

$p(x_2 \mid \mathbf{y})$	🐶	🐱	🐴	🐰
L	[54,61]	[31,37]	[66,75]	[2,9]
M	[23,30]	[61,67]	[23,32]	[30,37]
S	[16,23]	[2,8]	[2,10]	[61,69]

$p(x_3 \mid \mathbf{y})$	🐶	🐱	🐴	🐰
L	[40,47]	[46,52]	[23,32]	[2,9]
M	[26,33]	[17,22]	[10,19]	[19,26]
S	[26,33]	[31,37]	[58,66]	[72,79]

Running example [3]

🐶	🐱	🐴	🐰
$[25,26]$	$[29,31]$	$[20,22]$	$[25,26]$

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

$p(x_1\|\mathbf{y})$	🐶	🐱	🐴	🐰
L	[33,40]	[2,8]	[10,19]	[58,65]
M	[30,37]	[55,61]	[66,75]	[26,33]
S	[30,37]	[37,43]	[15,23]	[9,16]

$p(x_2\|\mathbf{y})$	🐶	🐱	🐴	🐰
L	[54,61]	[31,37]	[66,75]	[2,9]
M	[23,30]	[61,67]	[23,32]	[30,37]
S	[16,23]	[2,8]	[2,10]	[61,69]

$p(x_3\|\mathbf{y})$	🐶	🐱	🐴	🐰
L	[40,47]	[46,52]	[23,32]	[2,9]
M	[26,33]	[17,22]	[10,19]	[19,26]
S	[26,33]	[31,37]	[58,66]	[72,79]

Running example [3]

🐶	🐱	🐴	🐰
$[25,26]$	$[29,31]$	$[20,22]$	$[25,26]$

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

🐶 $\succeq_{\mathscr{P}}$ 🐴 ?

$p(x_1\|\mathbf{y})$	🐶	🐱	🐴	🐰
L	[33,40]	[2,8]	[10,19]	[58,65]
M	[30,37]	[55,61]	[66,75]	[26,33]
S	[30,37]	[37,43]	[15,23]	[9,16]

$p(x_2\|\mathbf{y})$	🐶	🐱	🐴	🐰
L	[54,61]	[31,37]	[66,75]	[2,9]
M	[23,30]	[61,67]	[23,32]	[30,37]
S	[16,23]	[2,8]	[2,10]	[61,69]

$p(x_3\|\mathbf{y})$	🐶	🐱	🐴	🐰
L	[40,47]	[46,52]	[23,32]	[2,9]
M	[26,33]	[17,22]	[10,19]	[19,26]
S	[26,33]	[31,37]	[58,66]	[72,79]

Running example [3]

🐶	🐱	🐴	🐰
[25,26]	[29,31]	[20,22]	[25,26]

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

🐶 $\succeq_{\mathscr{P}}$ 🐴 ?

$$\phi(N) = \frac{\underline{p}_{\mathscr{Y}}(\mathbf{y})}{\overline{p}_{\mathscr{Y}}(\mathbf{y}')} \times \frac{\underline{p}_1(\mathbf{x}_1^o|\mathbf{y})}{\overline{p}_1(\mathbf{x}_1^o|\mathbf{y}')} \times \frac{\underline{p}_2(\mathbf{x}_2^o|\mathbf{y})}{\overline{p}_2(\mathbf{x}_2^o|\mathbf{y}')} \times \frac{\underline{p}_3(\mathbf{x}_3^o|\mathbf{y})}{\overline{p}_3(\mathbf{x}_3^o|\mathbf{y}')}$$

$p(x_1\|\mathbf{y})$	🐶	🐱	🐴	🐰
L	[33,40]	[2,8]	[10,19]	[58,65]
M	[30,37]	[55,61]	[66,75]	[26,33]
S	[30,37]	[37,43]	[15,23]	[9,16]

$p(x_2\|\mathbf{y})$	🐶	🐱	🐴	🐰
L	[54,61]	[31,37]	[66,75]	[2,9]
M	[23,30]	[61,67]	[23,32]	[30,37]
S	[16,23]	[2,8]	[2,10]	[61,69]

$p(x_3\|\mathbf{y})$	🐶	🐱	🐴	🐰
L	[40,47]	[46,52]	[23,32]	[2,9]
M	[26,33]	[17,22]	[10,19]	[19,26]
S	[26,33]	[31,37]	[58,66]	[72,79]

Running example [3]

🐶	🐱	🐴	🐰
[25,26]	[29,31]	[20,22]	[25,26]

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

🐶 $\succeq_{\mathscr{P}}$ 🐴 ?

$$\phi(N) = \frac{\underline{p}_{\mathscr{Y}}(\mathbf{y})}{\overline{p}_{\mathscr{Y}}(\mathbf{y}')} \times \frac{\underline{p}_1(\mathbf{x}_1^o|\mathbf{y})}{\overline{p}_1(\mathbf{x}_1^o|\mathbf{y}')} \times \frac{\underline{p}_2(\mathbf{x}_2^o|\mathbf{y})}{\overline{p}_2(\mathbf{x}_2^o|\mathbf{y}')} \times \frac{\underline{p}_3(\mathbf{x}_3^o|\mathbf{y})}{\overline{p}_3(\mathbf{x}_3^o|\mathbf{y}')}$$

$$= \frac{0.25}{0.22}$$

$p(x_1\|\mathbf{y})$	🐶	🐱	🐴	🐰
L	[33,40]	[2,8]	[10,19]	[58,65]
M	[30,37]	[55,61]	[66,75]	[26,33]
S	[30,37]	[37,43]	[15,23]	[9,16]

$p(x_2\|\mathbf{y})$	🐶	🐱	🐴	🐰
L	[54,61]	[31,37]	[66,75]	[2,9]
M	[23,30]	[61,67]	[23,32]	[30,37]
S	[16,23]	[2,8]	[2,10]	[61,69]

$p(x_3\|\mathbf{y})$	🐶	🐱	🐴	🐰
L	[40,47]	[46,52]	[23,32]	[2,9]
M	[26,33]	[17,22]	[10,19]	[19,26]
S	[26,33]	[31,37]	[58,66]	[72,79]

Running example [3]

🐶	🐱	🐴	🐰
[25,26]	[29,31]	[20,22]	[25,26]

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

🐶 $\succeq_{\mathscr{P}}$ 🐴 ?

$$\phi(N) = \frac{\underline{p}_{\mathscr{Y}}(\mathbf{y})}{\overline{p}_{\mathscr{Y}}(\mathbf{y}')} \times \frac{\underline{p}_1(\mathbf{x}_1^o|\mathbf{y})}{\overline{p}_1(\mathbf{x}_1^o|\mathbf{y}')} \times \frac{\underline{p}_2(\mathbf{x}_2^o|\mathbf{y})}{\overline{p}_2(\mathbf{x}_2^o|\mathbf{y}')} \times \frac{\underline{p}_3(\mathbf{x}_3^o|\mathbf{y})}{\overline{p}_3(\mathbf{x}_3^o|\mathbf{y}')}$$

$$= \frac{0.25}{0.22} \times \frac{0.33}{0.19}$$

$p(x_1\|\mathbf{y})$	🐶	🐱	🐴	🐰
L	[33,40]	[2,8]	[10,19]	[58,65]
M	[30,37]	[55,61]	[66,75]	[26,33]
S	[30,37]	[37,43]	[15,23]	[9,16]

$p(x_2\|\mathbf{y})$	🐶	🐱	🐴	🐰
L	[54,61]	[31,37]	[66,75]	[2,9]
M	[23,30]	[61,67]	[23,32]	[30,37]
S	[16,23]	[2,8]	[2,10]	[61,69]

$p(x_3\|\mathbf{y})$	🐶	🐱	🐴	🐰
L	[40,47]	[46,52]	[23,32]	[2,9]
M	[26,33]	[17,22]	[10,19]	[19,26]
S	[26,33]	[31,37]	[58,66]	[72,79]

Running example [3]

🐶	🐱	🐴	🐰
$[25,26]$	$[29,31]$	$[20,22]$	$[25,26]$

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

🐶 $\succeq_{\mathscr{P}}$ 🐴 ?

$$\phi(N) = \frac{\underline{p}_{\mathscr{Y}}(\mathbf{y})}{\overline{p}_{\mathscr{Y}}(\mathbf{y}')} \times \frac{\underline{p}_1(\mathbf{x}_1^o|\mathbf{y})}{\overline{p}_1(\mathbf{x}_1^o|\mathbf{y}')} \times \frac{\underline{p}_2(\mathbf{x}_2^o|\mathbf{y})}{\overline{p}_2(\mathbf{x}_2^o|\mathbf{y}')} \times \frac{\underline{p}_3(\mathbf{x}_3^o|\mathbf{y})}{\overline{p}_3(\mathbf{x}_3^o|\mathbf{y}')}$$

$$= \frac{0.25}{0.22} \times \frac{0.33}{0.19} \times \frac{0.16}{0.10}$$

$p(x_1\|\mathbf{y})$	🐶	🐱	🐴	🐰
L	[33,40]	[2,8]	[10,19]	[58,65]
M	[30,37]	[55,61]	[66,75]	[26,33]
S	[30,37]	[37,43]	[15,23]	[9,16]

$p(x_2\|\mathbf{y})$	🐶	🐱	🐴	🐰
L	[54,61]	[31,37]	[66,75]	[2,9]
M	[23,30]	[61,67]	[23,32]	[30,37]
S	[16,23]	[2,8]	[2,10]	[61,69]

$p(x_3\|\mathbf{y})$	🐶	🐱	🐴	🐰
L	[40,47]	[46,52]	[23,32]	[2,9]
M	[26,33]	[17,22]	[10,19]	[19,26]
S	[26,33]	[31,37]	[58,66]	[72,79]

Running example [3]

🐶	🐱	🐴	🐰
[25,26]	[29,31]	[20,22]	[25,26]

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

🐶 $\succeq_{\mathscr{P}}$ 🐴 ?

$$\phi(N) =$$
$$\frac{\underline{p}_{\mathscr{Y}}(\mathbf{y})}{\overline{p}_{\mathscr{Y}}(\mathbf{y}')} \times \frac{\underline{p}_1(\mathbf{x}_1^o|\mathbf{y})}{\overline{p}_1(\mathbf{x}_1^o|\mathbf{y}')} \times \frac{\underline{p}_2(\mathbf{x}_2^o|\mathbf{y})}{\overline{p}_2(\mathbf{x}_2^o|\mathbf{y}')} \times \frac{\underline{p}_3(\mathbf{x}_3^o|\mathbf{y})}{\overline{p}_3(\mathbf{x}_3^o|\mathbf{y}')}$$
$$= \frac{0.25}{0.22} \times \frac{0.33}{0.19} \times \frac{0.16}{0.10} \times \frac{0.40}{0.32}$$

$p(x_1\|\mathbf{y})$	🐶	🐱	🐴	🐰
L	[33,40]	[2,8]	[10,19]	[58,65]
M	[30,37]	[55,61]	[66,75]	[26,33]
S	[30,37]	[37,43]	[15,23]	[9,16]

$p(x_2\|\mathbf{y})$	🐶	🐱	🐴	🐰
L	[54,61]	[31,37]	[66,75]	[2,9]
M	[23,30]	[61,67]	[23,32]	[30,37]
S	[16,23]	[2,8]	[2,10]	[61,69]

$p(x_3\|\mathbf{y})$	🐶	🐱	🐴	🐰
L	[40,47]	[46,52]	[23,32]	[2,9]
M	[26,33]	[17,22]	[10,19]	[19,26]
S	[26,33]	[31,37]	[58,66]	[72,79]

Running example [3]

🐶	🐱	🐴	🐰
[25,26]	[29,31]	[20,22]	[25,26]

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

🐶 $\succeq_{\mathscr{P}}$ 🐴 ?

$$\phi(N) = \frac{\underline{p}_{\mathscr{Y}}(\mathbf{y})}{\overline{p}_{\mathscr{Y}}(\mathbf{y}')} \times \frac{\underline{p}_1(\mathbf{x}_1^o|\mathbf{y})}{\overline{p}_1(\mathbf{x}_1^o|\mathbf{y}')} \times \frac{\underline{p}_2(\mathbf{x}_2^o|\mathbf{y})}{\overline{p}_2(\mathbf{x}_2^o|\mathbf{y}')} \times \frac{\underline{p}_3(\mathbf{x}_3^o|\mathbf{y})}{\overline{p}_3(\mathbf{x}_3^o|\mathbf{y}')}$$

$$= \frac{0.25}{0.22} \times \frac{0.33}{0.19} \times \frac{0.16}{0.10} \times \frac{0.40}{0.32} > 1$$

$p(x_1\|\mathbf{y})$	🐶	🐱	🐴	🐰
L	[33,40]	[2,8]	[10,19]	[58,65]
M	[30,37]	[55,61]	[66,75]	[26,33]
S	[30,37]	[37,43]	[15,23]	[9,16]

$p(x_2\|\mathbf{y})$	🐶	🐱	🐴	🐰
L	[54,61]	[31,37]	[66,75]	[2,9]
M	[23,30]	[61,67]	[23,32]	[30,37]
S	[16,23]	[2,8]	[2,10]	[61,69]

$p(x_3\|\mathbf{y})$	🐶	🐱	🐴	🐰
L	[40,47]	[46,52]	[23,32]	[2,9]
M	[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} :

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

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 $\mathbf{y} \succeq_{\mathscr{P}} \mathbf{y}'$:

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

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 $\mathbf{y} \succeq_{\mathscr{P}} \mathbf{y}'$:

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

Let

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

Running example [4]

🐶	🐱	🐴	🐰
$[25,26]$	$[29,31]$	$[20,22]$	$[25,26]$

🐶 $\succeq_{\mathscr{P}}$ 🐴

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

$p(x_1\|\mathbf{y})$	🐶	🐱	🐴	🐰
L	[33,40]	[2,8]	[10,19]	[58,65]
M	[30,37]	[55,61]	[66,75]	[26,33]
S	[30,37]	[37,43]	[15,23]	[9,16]

$p(x_2\|\mathbf{y})$	🐶	🐱	🐴	🐰
L	[54,61]	[31,37]	[66,75]	[2,9]
M	[23,30]	[61,67]	[23,32]	[30,37]
S	[16,23]	[2,8]	[2,10]	[61,69]

$p(x_3\|\mathbf{y})$	🐶	🐱	🐴	🐰
L	[40,47]	[46,52]	[23,32]	[2,9]
M	[26,33]	[17,22]	[10,19]	[19,26]
S	[26,33]	[31,37]	[58,66]	[72,79]

Running example [4]

🐶	🐱	🐴	🐰
$[25,26]$	$[29,31]$	$[20,22]$	$[25,26]$

🐶 $\succeq_{\mathscr{P}}$ 🐴

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

$p(x_1 \mid \mathbf{y})$	🐶	🐱	🐴	🐰
L	[33,40]	[2,8]	[10,19]	[58,65]
M	[30,37]	[55,61]	[66,75]	[26,33]
S	[30,37]	[37,43]	[15,23]	[9,16]

$p(x_2 \mid \mathbf{y})$	🐶	🐱	🐴	🐰
L	[54,61]	[31,37]	[66,75]	[2,9]
M	[23,30]	[61,67]	[23,32]	[30,37]
S	[16,23]	[2,8]	[2,10]	[61,69]

$p(x_3 \mid \mathbf{y})$	🐶	🐱	🐴	🐰
L	[40,47]	[46,52]	[23,32]	[2,9]
M	[26,33]	[17,22]	[10,19]	[19,26]
S	[26,33]	[31,37]	[58,66]	[72,79]

Running example [4]

🐶	🐱	🐴	🐰
[25,26]	[29,31]	[20,22]	[25,26]

🐶 $\succeq_{\mathscr{P}}$ 🐴

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

$$x_2^a = \arg\min\{\frac{0.54}{0.75}, \frac{0.23}{0.32}, \frac{0.16}{0.10}\} = M$$

$$x_3^a = \arg\min\{\frac{0.40}{0.32}, \frac{0.26}{0.19}, \frac{0.26}{0.66}\} = S$$

$p(x_1 \vert \mathbf{y})$	🐶	🐱	🐴	🐰
L	[33,40]	[2,8]	[10,19]	[58,65]
M	[30,37]	[55,61]	[66,75]	[26,33]
S	[30,37]	[37,43]	[15,23]	[9,16]

$p(x_2 \vert \mathbf{y})$	🐶	🐱	🐴	🐰
L	[54,61]	[31,37]	[66,75]	[2,9]
M	[23,30]	[61,67]	[23,32]	[30,37]
S	[16,23]	[2,8]	[2,10]	[61,69]

$p(x_3 \vert \mathbf{y})$	🐶	🐱	🐴	🐰
L	[40,47]	[46,52]	[23,32]	[2,9]
M	[26,33]	[17,22]	[10,19]	[19,26]
S	[26,33]	[31,37]	[58,66]	[72,79]

Running example [4]

🐶	🐱	🐴	🐰
$[25,26]$	$[29,31]$	$[20,22]$	$[25,26]$

🐶 $\succeq_{\mathscr{P}}$ 🐴

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

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

$p(x_1\|\mathbf{y})$	🐶	🐱	🐴	🐰
L	[33,40]	[2,8]	[10,19]	[58,65]
M	[30,37]	[55,61]	[66,75]	[26,33]
S	[30,37]	[37,43]	[15,23]	[9,16]

$p(x_2\|\mathbf{y})$	🐶	🐱	🐴	🐰
L	[54,61]	[31,37]	[66,75]	[2,9]
M	[23,30]	[61,67]	[23,32]	[30,37]
S	[16,23]	[2,8]	[2,10]	[61,69]

$p(x_3\|\mathbf{y})$	🐶	🐱	🐴	🐰
L	[40,47]	[46,52]	[23,32]	[2,9]
M	[26,33]	[17,22]	[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

$$
\begin{aligned}
G(i) &= \log \phi(E \cup \{i\}) - \log \phi(E) \\
&= \left(\log \underline{p}_i(\mathbf{x}_i^o|\mathbf{y}) - \log \overline{p}_i(\mathbf{x}_i^o|\mathbf{y}')\right) - \left(\log \underline{p}_i(x_i^a|\mathbf{y}) - \log \overline{p}_i(x_i^a|\mathbf{y}')\right)
\end{aligned}
$$

Contribution of feature i is independent of other features !

Running example [5]

🐶	🐱	🐴	🐰
$[25,26]$	$[29,31]$	$[20,22]$	$[25,26]$

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

🐶 $\succeq_{\mathscr{P}}$ 🐴

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

$p(x_1 \mid \mathbf{y})$	🐶	🐱	🐴	🐰
L	[33,40]	[2,8]	[10,19]	[58,65]
M	[30,37]	[55,61]	[66,75]	[26,33]
S	[30,37]	[37,43]	[15,23]	[9,16]

$p(x_2 \mid \mathbf{y})$	🐶	🐱	🐴	🐰
L	[54,61]	[31,37]	[66,75]	[2,9]
M	[23,30]	[61,67]	[23,32]	[30,37]
S	[16,23]	[2,8]	[2,10]	[61,69]

$p(x_3 \mid \mathbf{y})$	🐶	🐱	🐴	🐰
L	[40,47]	[46,52]	[23,32]	[2,9]
M	[26,33]	[17,22]	[10,19]	[19,26]
S	[26,33]	[31,37]	[58,66]	[72,79]

Running example [5]

🐶	🐱	🐴	🐰
$[25,26]$	$[29,31]$	$[20,22]$	$[25,26]$

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

🐶 $\succeq_{\mathscr{P}}$ 🐴

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

$$G(2) = (\log 0.16 - \log 0.10) - (\log 0.23 - \log 0.32) = 0.33$$

$p(x_1\|\mathbf{y})$	🐶	🐱	🐴	🐰
L	[33,40]	[2,8]	[10,19]	[58,65]
M	[30,37]	[55,61]	[66,75]	[26,33]
S	[30,37]	[37,43]	[15,23]	[9,16]

$p(x_2\|\mathbf{y})$	🐶	🐱	🐴	🐰
L	[54,61]	[31,37]	[66,75]	[2,9]
M	[23,30]	[61,67]	[23,32]	[30,37]
S	[16,23]	[2,8]	[2,10]	[61,69]

$p(x_3\|\mathbf{y})$	🐶	🐱	🐴	🐰
L	[40,47]	[46,52]	[23,32]	[2,9]
M	[26,33]	[17,22]	[10,19]	[19,26]
S	[26,33]	[31,37]	[58,66]	[72,79]

Running example [5]

🐶	🐱	🐴	🐰
$[25,26]$	$[29,31]$	$[20,22]$	$[25,26]$

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

🐶 $\succeq_{\mathscr{P}}$ 🐴

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

$$G(2) = (\log 0.16 - \log 0.10) - (\log 0.23 - \log 0.32) = 0.33$$

$$G(3) = (\log 0.40 - \log 0.32) - (\log 0.26 - \log 0.66) = 0.50$$

$p(x_1 \mid \mathbf{y})$	🐶	🐱	🐴	🐰
L	[33,40]	[2,8]	[10,19]	[58,65]
M	[30,37]	[55,61]	[66,75]	[26,33]
S	[30,37]	[37,43]	[15,23]	[9,16]

$p(x_2 \mid \mathbf{y})$	🐶	🐱	🐴	🐰
L	[54,61]	[31,37]	[66,75]	[2,9]
M	[23,30]	[61,67]	[23,32]	[30,37]
S	[16,23]	[2,8]	[2,10]	[61,69]

$p(x_3 \mid \mathbf{y})$	🐶	🐱	🐴	🐰
L	[40,47]	[46,52]	[23,32]	[2,9]
M	[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 :

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

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

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

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

Computing E

Algorithm 1: Compute first prime implicants explanation

Input: $C : log(\phi(\emptyset))$; G : Contributions of criteria ;
Output: $Xpl = (E, \mathbf{x}^o_E)$: PI explanation and associated values
Order G in decreasing order, with σ the associated permutation
$i \leftarrow 1$
while $\phi(E) + C < 0$ **do**
  $i \leftarrow i + 1$
  $E \leftarrow E \cup \{\sigma^{-1}(i)\}$
  $\phi(E) \leftarrow \phi(E) + G_{\sigma(i)}$
$Xpl \leftarrow (E, \mathbf{x}^o_E)$
return (Xpl)

Running example [6]

🐶	🐱	🐴	🐰
$[25,26]$	$[29,31]$	$[20,22]$	$[25,26]$

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

🐶 $\succeq_{\mathscr{P}}$ 🐴

$C = -0.9$
$G(1) = 0.65$
$G(3) = 0.50$
$G(2) = 0.33$

$E_1 = \{Ears, Hair\}$
$E_2 = \{Ears, Tail\}$

$p(x_1 \mid \mathbf{y})$	🐶	🐱	🐴	🐰
L	[33,40]	[2,8]	[10,19]	[58,65]
M	[30,37]	[55,61]	[66,75]	[26,33]
S	[30,37]	[37,43]	[15,23]	[9,16]

$p(x_2 \mid \mathbf{y})$	🐶	🐱	🐴	🐰
L	[54,61]	[31,37]	[66,75]	[2,9]
M	[23,30]	[61,67]	[23,32]	[30,37]
S	[16,23]	[2,8]	[2,10]	[61,69]

$p(x_3 \mid \mathbf{y})$	🐶	🐱	🐴	🐰
L	[40,47]	[46,52]	[23,32]	[2,9]
M	[26,33]	[17,22]	[10,19]	[19,26]
S	[26,33]	[31,37]	[58,66]	[72,79]

- Introduction
  - Classification
  - Prime implicant
- Naive Credal Classifier [3]
  - General case
  - Prime implicants formulation
  - Computation
- Conclusion

Conclusion

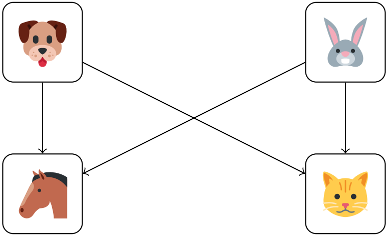
Summary :

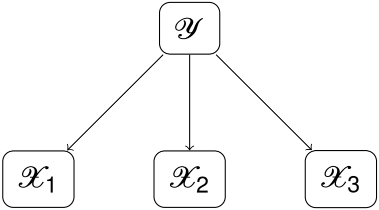
- Prime implicants for robust preferences
- Application to the NCC with convex domains
- Polynomial calculation of implicants

Perspectives :

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

Conclusion





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)