Prime implicants as a versatile tool to explain robust classification

Hénoïk Willot, Sébastien Destercke & Khaled Belahcene

13th International Symposium on Imprecise Probabilities: Theories and Applications

Our team

Our team

CID Team

Our team

Prime implicants as a versatile tool to explain robust classification

Hénoïk Willot, Sébastien Destercke & Khaled Belahcene

13th International Symposium on Imprecise Probabilities: Theories and Applications

Introduction

Classification problem

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

Introduction

Classification problem

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

Crisp case :

One probability distribution p

$$\mathbf{y} \succeq_{p,(\mathbf{x})} \mathbf{y}' \text{ if } p(\mathbf{y}|\mathbf{x}) \geq p(\mathbf{y}'|\mathbf{x})$$

$\Rightarrow$ 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 $\mathbf{y} \succ_{\mathscr{P},(\mathbf{x})} \mathbf{y}'$,

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

Introduction

Classification problem

Credal case :

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

Robust classification :

Incomparability $\mathbf{y} \succ\prec_{\mathscr{P},(\mathbf{x})} \mathbf{y}'$,

$$\exists p \in \mathscr{P} \ \frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y}'|\mathbf{x})} < 1 \ \textit{and} \ \exists p' \in \mathscr{P} \frac{p'(\mathbf{y}'|\mathbf{x})}{p'(\mathbf{y}|\mathbf{x})} < 1$$

Introduction
Classification problem

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

Robust classification :
Incomparability $\mathbf{y} \succ\!\prec_{\mathscr{P},(\mathbf{x})} \mathbf{y}'$,

$$\exists p \in \mathscr{P} \ \frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y}'|\mathbf{x})} < 1 \ \textit{and} \ \exists p' \in \mathscr{P} \frac{p'(\mathbf{y}'|\mathbf{x})}{p'(\mathbf{y}|\mathbf{x})} < 1$$
$$\Leftrightarrow \inf_{p \in \mathscr{P}} \frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y}'|\mathbf{x})} < 1 \ \textit{and} \ \inf_{p' \in \mathscr{P}} \frac{p'(\mathbf{y}'|\mathbf{x})}{p'(\mathbf{y}|\mathbf{x})} < 1$$

Introduction
Running example

Data : ZOO dataset from UCI repository
Objective : predict **y** in $\mathscr{Y} = \{$ Mammal (M), Bird (B), Reptile (R), Fish (F) or Invertebrate (I) $\}$

Introduction
Running example

Data : ZOO dataset from UCI repository

Objective : predict **y** in $\mathscr{Y} = \{$ Mammal (M), Bird (B), Reptile (R), Fish (F) or Invertebrate (I) $\}$

Features :

- feathers : { , }
- eggs : { , }
- aquatic : { , }
- toothed : { , }
- backbone : { , }
- breathes with lungs : { , }
- venomous : { , }
- fins : { , }
- legs : { ,2 $*$,4 $*$,5 $*$,6 $*$,8 $*$ }
- tail : { , }

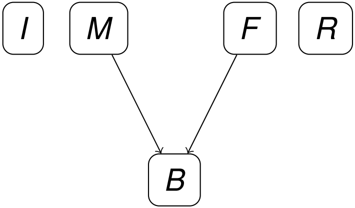
Introduction

Running example

Observation : 🦭 = (~~🪶~~, ~~🥚~~, 🌊, 🦷, 🦴, 🫁, ~~☣️~~, 🐟, ~~🦵~~, ~~🦎~~)

Model :

- $\log p(M|\text{🦭}) \in [-1.6, -0.012]$
- $\log p(B|\text{🦭}) \in [-8.36, -3]$
- $\log p(R|\text{🦭}) \in [-3.08, -0.016]$
- $\log p(F|\text{🦭}) \in [-2.59, -0.029]$
- $\log p(I|\text{🦭}) \in [-4.82, -0.45]$



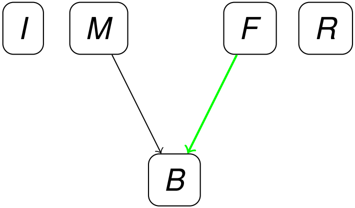
Introduction

Running example

Observation : 🦭 = (~~🪶~~, ~~🥚~~, 🌊, 🦷, 🦴, 🫁, ~~☣️~~, 🐟, ~~🦵~~, ~~🐍~~)

Model :

- $\log p(M|🦭) \in [-1.6, -0.012]$
- $\log p(B|🦭) \in [-8.36, -3]$
- $\log p(R|🦭) \in [-3.08, -0.016]$
- $\log p(F|🦭) \in [-2.59, -0.029]$
- $\log p(I|🦭) \in [-4.82, -0.45]$



$F \succ_{\mathscr{P},(🦭)} B$ because $-2.59 + 3 \geq 0$

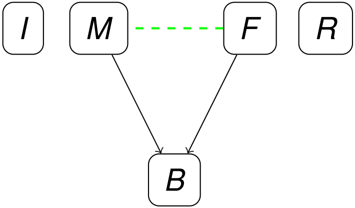
Introduction

Running example

Observation : 🦭 = (~~🪶~~, ~~🥚~~, 🌊, 🦷, 🦴, 🫁, ~~☣️~~, 🐟, ~~🦵~~, ~~🦎~~)

Model :

- $\log p(M|\text{🦭}) \in [-1.6, -0.012]$
- $\log p(B|\text{🦭}) \in [-8.36, -3]$
- $\log p(R|\text{🦭}) \in [-3.08, -0.016]$
- $\log p(F|\text{🦭}) \in [-2.59, -0.029]$
- $\log p(I|\text{🦭}) \in [-4.82, -0.45]$



$F \succ_{\mathscr{P},(\text{🦭})} B$ because $-2.59 + 3 \geq 0$

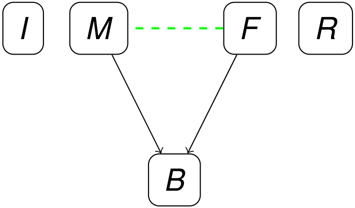
$M \succ\!\prec_{\mathscr{P},(\text{🦭})} F$ because $-1.6 + 0.029 < 0$

Introduction
Running example

Observation : 🦭 = (~~🪶~~, ~~🥚~~, 🌊, 🦷, 🦴, 🫁, ~~☣️~~, 🐬, ~~🦵~~, ~~🦎~~)

Model :

- $\log p(M|\text{🦭}) \in [-1.6, -0.012]$
- $\log p(B|\text{🦭}) \in [-8.36, -3]$
- $\log p(R|\text{🦭}) \in [-3.08, -0.016]$
- $\log p(F|\text{🦭}) \in [-2.59, -0.029]$
- $\log p(I|\text{🦭}) \in [-4.82, -0.45]$



$F \succ_{\mathscr{P},(\text{🦭})} B$ because $-2.59 + 3 \geq 0$

$M \succ\prec_{\mathscr{P},(\text{🦭})} F$ because $-1.6 + 0.029 < 0$ and $-2.59 + 0.012 < 0$

... validatory

$E \subseteq N$, as a subset of feature indices, is a validatory implicant of decision $\mathbf{y} \succ_{\mathscr{P},(\mathbf{x})} \mathbf{y}'$ if :

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

... validatory

$E \subseteq N$, as a subset of feature indices, is a validatory implicant of decision $\mathbf{y} \succ_{\mathscr{P},(\mathbf{x})} \mathbf{y}'$ if :

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

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

... validatory

$E \subseteq N$, as a subset of feature indices, is a validatory implicant of decision $\mathbf{y} \succ_{\mathscr{P},(\mathbf{x})} \mathbf{y}'$ if :

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

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

$E \subseteq N$ is a *prime* implicant if $\forall i \in E$, the inequality does not hold, *i.e.* E is subset-minimal

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

...validatory

i.e. observing $\mathbf{x}_E$ is sufficient to conclude no matter the values on other attributes of $\mathscr{X}^{-E}$

$E = \{'feathers', 'eggs', 'toothed'\}$ is a prime implicant of $M \succ_{\mathscr{P},(\text{🦭})} B$

Observation :
🦭 = (❌🪶, ❌🥚, 🌊, 🦷, 🦴, 🫁, ❌☣️, 🐟, ❌🦵, ❌🦎) $\Longrightarrow M \succ_{\mathscr{P},(\text{🦭})} B$

Observation :
$(\text{🦭}_E, \text{❓}_{-E})$ = ((❌🪶, ❌🥚, 🦷), (❓🌊, ❓🦴, ❓🫁, ❓☣️, ❓🐟, ❓🦵, ❓🦎))
$\Longrightarrow M \succ_{\mathscr{P},(\text{🦭}_E, \text{❓}_{-E})} B$

... contrastive and doubt ?

Prime implicant can also be used for constrastive explanation

... contrastive and doubt ?

Prime implicant can also be used for constrastive explanation

But also to explain incomparability !

... contrastive and doubt ?

Prime implicant can also be used for constrastive explanation

But also to explain incomparability !

To know more come visit my poster 😉

Thank you for your attention !