Classification with Imprecise Probability

Cassio P. de Campos, Alessandro Antonucci, Giorgio Corani
{cassio,alessandro,giorgio}@idsia.ch

IDSIA - Switzerland

SIPTA School – Durham 2010

Exercise 1

Instance	*Bag (A)*	*Marble (C)*
1	1	*blue (0)*
2	1	*blue (0)*
3	1	*blue (0)*
4	2	*blue (0)*
5	2	*red (1)*

- Use the local IDM (with $s = 2$) to compute lower and upper probs for *(red|bag1)* considering the first 0, 1, 2 and 3 instances.
- Do the same with the precise Laplace prior and compare the results.
- Use the global IDM (with $s = 1$) to compute the extreme mass functions of the joint $(Bag, Marble)$.

Exercise 1

Local IDM:

$$\underline{P}(c_j|b_i,\mathbf{n}) = \inf_{0<t<1} \frac{n_{ji}+t}{n_i+s} = \frac{n_{ji}}{n_i+s}$$

$$\overline{P}(c_j|b_i,\mathbf{n}) = \sup_{0<t<1} \frac{n_{ji}+t}{n_i+s} = \frac{n_{ji}+s}{n_i+s}$$

Global IDM:

$$\underline{P}(c_j,b_i|\mathbf{n}) = \inf_{0<t<1} \frac{n_{ji}+t}{n+s} = \frac{n_{ji}}{n+s}$$

$$\overline{P}(c_j,b_i|\mathbf{n}) = \sup_{0<t<1} \frac{n_{ji}+t}{n+s} = \frac{n_{ji}+s}{n+s}$$

Exercise 1

1. $p(C = 1|A = 1, \mathbf{d})$? IDM$(s = 2)$ on the sub-sample with $A = 1$:

$$p(C = 1|A = 1, \mathbf{d}) \in \left[\frac{n_{A=1,C=1}}{n_{A=1} + 2}, \frac{n_{A=1,C=1} + 2}{n_{A=1} + 2}\right]$$

- For each data set: $n_{A=1,C=1} = 0$, $n_{A=1} = \#$, $p \in [0, \frac{2}{2+\#}]$

empty sample	first observation	first two observations	whole sample
$[0, 1]$	$[0, 2/3]$	$[0, 2/4]$	$[0, 2/5]$

Exercise 1

2. $\mathrm{IDM}(s=2) \rightarrow \mathrm{PDM}(s=2, t=\frac{1}{2})$

$$p(C=1|A=1,\mathbf{d}) = \frac{n_{A=1,C=1}+1}{n_{A=1}+2}$$

empty sample	first observation	first two observations	whol samp
1/2	1/3	1/4	1/5

Exercise 1

Instance	*Bag (A)*	*Marble (C)*
1	1	*blue*
2	1	*blue*
3	1	*blue*
4	2	*blue*
5	2	*red*

(bag1,blue)	*(bag1,red)*	*(bag2,blue)*	*(bag2,red)*
$\frac{3+1}{5+1}$	$\frac{0}{5+1}$	$\frac{1}{5+1}$	$\frac{1}{5+1}$
$\frac{3}{5+1}$	$\frac{0+1}{5+1}$	$\frac{1}{5+1}$	$\frac{1}{5+1}$
$\frac{3}{5+1}$	$\frac{0}{5+1}$	$\frac{1+1}{5+1}$	$\frac{1}{5+1}$
$\frac{3}{5+1}$	$\frac{0}{5+1}$	$\frac{1}{5+1}$	$\frac{1+1}{5+1}$

Exercise 2

Instance	*Bag (A)*	*Marble (C)*
1	1	*blue (0)*
2	1	*blue (0)*
3	1	*blue (0)*
4	2	*blue (0)*
5	2	*red (1)*

- Using the local IDM (with $s = 2$), compute the non-dominated states of $Marble|bag1$ considering only the first 0, 1, 2 and 3 instances.
- Do the same with the precise Laplace prior and compare the results.

Exercise 2

- Class c' dominates c'' iff:

$$P(c'|\mathbf{a}, \mathbf{n}, \mathbf{t}) > P(c''|\mathbf{a}, \mathbf{n}, \mathbf{t})$$

$$\forall\ \mathbf{t} \in T$$

Exercise 3

day	outlook	temp	humidity	wind	play
1	sunny	hot	high	weak	**no**
2	sunny	hot	high	strong	**no**
3	overcast	hot	high	weak	**yes**
4	rainy	mild	high	weak	**yes**
5	rainy	cool	normal	weak	**yes**
6	rainy	cool	normal	strong	**no**
7	overcast	cool	normal	strong	**yes**
8	sunny	mild	high	weak	**no**
9	sunny	cool	normal	weak	**yes**
10	rainy	mild	normal	weak	**yes**
11	sunny	mild	normal	strong	**yes**
12	overcast	mild	high	strong	**yes**
13	overcast	hot	normal	weak	**yes**
14	rainy	mild	high	strong	**no**

- Discover my orientation to play tennis if *outlook* is **overcast**, *temperature* is **cool**, *humidity* is **high**, *wind* is **strong**, using NCC with $s = 1$.
- Try also when *outlook* is **sunny**.

Exercise 2

1. $p(C = 1|A = 1, \mathbf{d})$? IDM($s = 2$) on the sub-sample with $A = 1$:

$$p(C = 1|A = 1, \mathbf{d}) \in \left[\frac{n_{A=1,C=1}}{n_{A=1} + 2}, \frac{n_{A=1,C=1} + 2}{n_{A=1} + 2}\right]$$

- For each data set: $n_{A=1,C=1} = 0$, $n_{A=1} = \#$, $p \in [0, \frac{2}{2+\#}]$

empty sample	first observation	first two observations	whole sample
$[0, 1]$	$[0, 2/3]$	$[0, 2/4]$	$[0, 2/5]$
INDECISION	INDECISION	INDECISION	DOMINANCE

Exercise 2

2. $\mathrm{IDM}(s=2) \rightarrow \mathrm{PDM}(s=2, t=\frac{1}{2})$

$$p(C=1|A=1,\mathbf{d}) = \frac{n_{A=1,C=1}+1}{n_{A=1}+2}$$

empty sample	first observation	first two observations	whole sample
1/2	1/3	1/4	1/5
INDIFFERENCE	DOMINANCE	DOMINANCE	DOMINANCE

Exercise 2

#	IDM	class	PDM	class
0	[0,1]	{0, 1}	.5	{0, 1}
1	[0,.67]	{0, 1}	.33	{0}
2	[0,.5]	{0, 1}	.25	{0}
3	[0,.4]	{0}	.20	{0}

Exercise 3

- Comparing c' and c'' through maximality requires to solve:

$$\min_{\mathbf{t} \in T} \frac{P(c'|\mathbf{a}, \mathbf{n}, \mathbf{t})}{P(c''|\mathbf{a}, \mathbf{n}, \mathbf{t})} > 1$$

- The value of the features is specified as $\mathbf{a} = (a_i, \ldots, a_k)$.

$$P(c, \mathbf{a}|\mathbf{n}, \mathbf{t}) = E\left[\theta_{c,\mathbf{a}}|\mathbf{n}, \mathbf{t}\right] = P(c|\mathbf{n}, \mathbf{t}) \prod_{i=1}^{k} P(a_i|c, \mathbf{n}, \mathbf{t})$$

where

$$P(c|\mathbf{n}, \mathbf{t}) = \frac{n(c)+st(c)}{n+s}$$
$$P(a_i|c, \mathbf{n}, \mathbf{t}) = \frac{n(a_i,c)+st(a_i,c)}{n(c)+st(c)}.$$

- $\sum_{a_i,c} t(a_i, c) = 1$, $\sum_c t(c) = 1$, and $0 < t(a_i, c) \leq t(c)$.

Exercise 3

- NCC-based test of credal dominance $\Leftrightarrow$ optimization problem
  - NCC: $c' \succ c'' \iff \inf_{x\in[0,s]} \left[\left(\frac{n_{c''}+x}{n_{c'}+s-x} \right)^{m-1} \prod_{j=1}^{m} \frac{n_{c',a_j}}{n_{c'',a_j}+x} \right] > 1$
    - $s = 1$, $m = 4$, $a_1 =$ overcast , $a_2 =$ cool , $a_3 =$ high , $a_4 =$ strong
  - Case $c' =$ no , $c'' =$ yes
    - $n_{c',\text{overcast}} = 0 \Rightarrow$ target function null $\Rightarrow$ no$\not\succ$yes
  - Case $c' =$ yes , $c'' =$ no
    - Function to minimize $f(x) = \left(\frac{5+x}{10-x} \right)^3 \cdot \frac{4}{x} \cdot \frac{3}{1+x} \cdot \frac{3}{4+x} \cdot \frac{3}{3+x}$
    - We simply note $f(1) = 4/5 < 1 \Rightarrow$ yes $\not\succ$ no
  - **Indecision** between play and not play tennis
- Second case: if the outlook is sunny, I don't go to play tennis

Exercise 4

day	outlook	temp	humidity	wind	play
1	sunny	hot	high	weak	**no**
2	sunny	hot	high	strong	**no**
3	overcast	hot	high	*	**yes**
4	rainy	mild	high	weak	**yes**
5	rainy	cool	normal	weak	**yes**
6	rainy	cool	normal	strong	**no**
7	overcast	cool	normal	strong	**yes**
8	sunny	mild	high	weak	**no**
9	sunny	cool	normal	weak	**yes**
10	rainy	mild	normal	weak	**yes**
11	*	mild	normal	strong	**yes**
12	overcast	mild	high	strong	**yes**
13	overcast	hot	normal	weak	**yes**
14	rainy	mild	high	strong	**no**

- Decide whether or not I play tennis if the *outlook* is **sunny** and the *humidity* is **high** using NCC with $s = 1$ where the missing observations about *wind* are not assumed MAR, while *outlook* and *temp* are assumed to be MAR.

Exercise 4

- Comparing c' and c'' through maximality requires to solve:

$$\min_{\mathbf{t}\in T} \min_{\mathbf{n}\in N} \frac{P(c'|\mathbf{a},\mathbf{n},\mathbf{t})}{P(c''|\mathbf{a},\mathbf{n},\mathbf{t})} > 1$$

  where the $\min_{\mathbf{t}\in T}$ is processed as before, while $\min_{\mathbf{n}\in N}$ is over all possible completions of the dataset where MAR was not assumed.

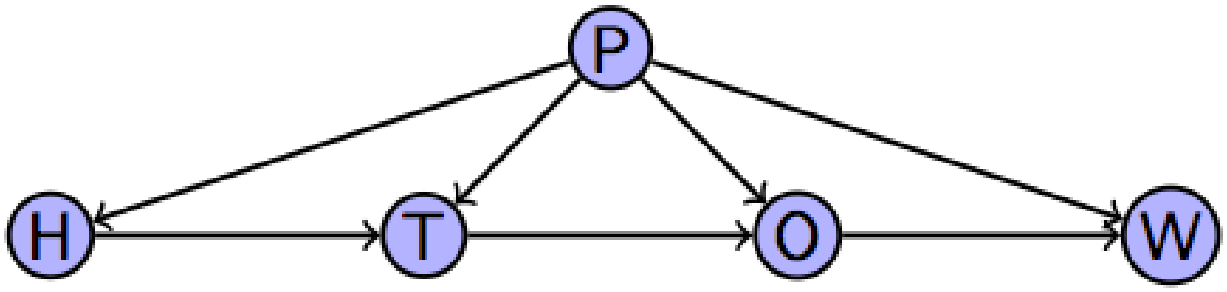
Exercise 4

- Temperature is *missing at random* (MAR) ⇒ ignore the attribute
  drop the relative column from the sample
- Case $c' = \text{no}$, $c'' = \text{yes}$:
  - no $>$ yes if and only if

$$\inf_{0<x<1} \left[\left(\frac{n_{\text{yes}} + x}{n_{\text{no}} + 1 - x} \right)^2 \cdot \frac{\beta_{\text{no,sunny}}}{\bar{n}_{\text{yes,sunny}} + x} \cdot \frac{\beta_{\text{no,high}}}{\bar{n}_{\text{yes,high}} + x} \cdot \min_{w \in \{\text{strong,weak}\}} \frac{\beta_{\text{no},w}}{\bar{n}_{\text{yes},w} + x} \right]$$

  - The minimum over w(ind) is reached in $w =$ **weak** for each x
  - $f(1) = 8/7$ no $>$ yes
- I don't go to play tennis

Exercise 5



- Use the TANC (graph of dependencies in the picture) with $s = 1$ and local IDM to classify whether or not I play tennis given **sunny, mild, high, strong** and the complete weather dataset of Ex.3.

Exercise 5

- Play tennis
  $p(\text{yes}|\mathbf{d}) \in [\frac{9}{15}, \frac{10}{15}]$
  $p(\text{no}|\mathbf{d}) \in [\frac{5}{15}, \frac{6}{15}]$
- Humidity
  $p(\text{high}|\text{yes}, \mathbf{d}) \in [\frac{3}{10}, \frac{4}{10}]$
  $p(\text{high}|\text{no}, \mathbf{d}) \in [\frac{4}{6}, \frac{5}{6}]$
- Temperature
  $p(\text{mild}|\text{high,yes}, \mathbf{d}) \in [\frac{2}{4}, \frac{3}{4}]$
  $p(\text{mild}|\text{high,no}, \mathbf{d}) \in [\frac{2}{5}, \frac{3}{5}]$
- Outlook
  $p(\text{sunny}|\text{mild,yes}, \mathbf{d}) \in [\frac{1}{5}, \frac{2}{5}]$
  $p(\text{sunny}|\text{mild,no}, \mathbf{d}) \in [\frac{1}{3}, \frac{2}{3}]$
- Wind
  $p(\text{strong}|\text{sunny,yes}, \mathbf{d}) \in [\frac{1}{3}, \frac{2}{3}]$
  $p(\text{strong}|\text{sunny,no}, \mathbf{d}) \in [\frac{1}{4}, \frac{2}{4}]$

Exercise 5 and 6

- $c' > c''$ iff

$$\frac{\underline{p}(c'|\mathbf{d}) \cdot \underline{p}(\text{high}|c', \mathbf{d}) \cdot \underline{p}(\text{mild}|\text{high}, c', \mathbf{d}) \cdot \underline{p}(\text{sunny}|\text{mild}, c', \mathbf{d}) \cdot \underline{p}(\text{wind}|\text{sunny}, c', \mathbf{d})}{\overline{p}(c''|\mathbf{d}) \cdot \overline{p}(\text{high}|c'', \mathbf{d}) \cdot \overline{p}(\text{mild}|\text{high}, c'', \mathbf{d}) \cdot \overline{p}(\text{sunny}|\text{mild}, c'', \mathbf{d}) \cdot \overline{p}(\text{wind}|\text{sunny}, c'', \mathbf{d})} > 1$$

- Exercise 6 has many ways to complete the data. Is this the moment of using computers? :)

Exercise 6

day	outlook	temp	humidity	wind	play
1	sunny	hot	high	weak	no
2	sunny	hot	high	strong	no
3	overcast	hot	high	weak	yes
4	rainy	mild	high	weak	yes
5	rainy	cool	normal	weak	yes
6	rainy	cool	normal	strong	no
7	overcast	cool	normal	strong	yes
8	sunny	mild	high	weak	no
9	sunny	cool	normal	weak	yes
10	rainy	mild	normal	weak	yes
11	sunny	mild	normal	strong	yes
12	overcast	mild	high	strong	yes
13	overcast	hot	normal	weak	yes
14	rainy	mild	high	strong	no

- Use the TANC (dependencies defined in Ex.5) with $s = 1$ and extreme global IDM to classify whether or not I play tennis given **sunny, mild, high, strong** and the complete weather dataset.

Exercise 6

- $c' > c''$ iff

$$\frac{\underline{p}(c'|\mathbf{d}) \cdot \underline{p}(\text{high}|c', \mathbf{d}) \cdot \underline{p}(\text{mild}|\text{high}, c', \mathbf{d}) \cdot \underline{p}(\text{sunny}|\text{mild}, c', \mathbf{d}) \cdot \underline{p}(\text{wind}|\text{sunny}, c', \mathbf{d})}{\overline{p}(c''|\mathbf{d}) \cdot \overline{p}(\text{high}|c'', \mathbf{d}) \cdot \overline{p}(\text{mild}|\text{high}, c'', \mathbf{d}) \cdot \overline{p}(\text{sunny}|\text{mild}, c'', \mathbf{d}) \cdot \overline{p}(\text{wind}|\text{sunny}, c'', \mathbf{d})} >$$

- yes $\not>$ no, because

$$\frac{\frac{9}{15} \cdot \frac{3}{10} \cdot \frac{2}{4} \cdot \frac{1}{5} \cdot \frac{1}{3}}{\frac{6}{15} \cdot \frac{5}{6} \cdot \frac{3}{5} \cdot \frac{2}{3} \cdot \frac{2}{4}} = \frac{9}{100} < 1$$

- no $\not>$ yes, because

$$\frac{\frac{5}{15} \cdot \frac{4}{6} \cdot \frac{2}{5} \cdot \frac{1}{3} \cdot \frac{1}{4}}{\frac{10}{15} \cdot \frac{4}{10} \cdot \frac{3}{4} \cdot \frac{2}{5} \cdot \frac{2}{3}} = \frac{5}{36} < 1$$

- **Indecision**

Exercise 7

day	outlook	temp	humidity	wind	play
1	sunny	hot	high	weak	no
2	sunny	hot	high	strong	no
3	overcast	hot	high	weak	yes
4	rainy	mild	high	weak	yes
5	rainy	cool	normal	weak	yes
6	rainy	cool	normal	strong	no
7	overcast	cool	normal	strong	yes
8	sunny	mild	high	weak	no
9	sunny	cool	normal	weak	yes
10	rainy	mild	normal	weak	yes
11	sunny	mild	normal	strong	yes
12	overcast	mild	high	strong	yes
13	overcast	hot	normal	weak	yes
14	rainy	mild	high	strong	no

- Use the credal ID3 to classify whether or not I play tennis given **high, strong** and the complete weather dataset (assume *outlook* and *temp* can be ignored).