Solutions

Marco Zaffalon: Knowledge discovery from data sets under tenable assumptions. School on Imprecise Probabilities, Lugano (CH), 27-31 July 2004.

Exercise 1

- IDM expression for $p(C=1|A=1,\boldsymbol{d})$
  - Apply IDM to the sub-sample where A=1
  - Whence $p(C=1|A=1,\boldsymbol{d}) \in [\frac{n_{11}}{n_{\cdot 1}+2}, \frac{n_{11}+2}{n_{\cdot 1}+2}]$
- Precise DM expression for $p(C=1|A=1,\boldsymbol{d})$
  - Apply PDM to the sub-sample where A=1
    - Bayes-Laplace: $s = 2$, $t_j = 1/2$
  - Whence $p(C=1|A=1,\boldsymbol{d}) = \frac{n_{11}+1}{n_{\cdot 1}+2}$
- Solution

#	IDM prob.	PDM prob.	IDM class.	PDM class.
0	[0,1]	0.5	{0,1}	{0,1}
1	[0,0.67]	0.33	{0,1}	0
2	[0,0.5]	0.25	{0,1}	0
3	[0,0.4]	0.2	0	0

Exercise 2 – question 1

- $c' > c'' \iff \inf_{t_{c'},t_{c''}} \left[\left(\frac{n_{c''}+st_{c''}}{n_{c'}+st_{c'}} \right)^{m-1} \prod_{j=1}^{m} \frac{n_{c',a_j}}{n_{c'',a_j}+st_{c''}} \right] > 1$
- Notation: t_{yes} → y, t_{no} → n
- Case $c' = \text{no}, c'' = \text{yes}$
  - Note that $n_{\text{no,overcast}} = 0 \Longrightarrow \inf = 0 \Longrightarrow \text{no} \not> \text{yes}$
- Case $c' = \text{yes}, c'' = \text{no}$
  - Obj. function $= f = \left(\frac{5+n}{10-n}\right)^3 \cdot \frac{4}{n} \cdot \frac{3}{1+n} \cdot \frac{3}{4+n} \cdot \frac{3}{3+n}$
  - $f(1) = 4/5 \Longrightarrow \text{yes} \not> \text{no}$
- Solution: (overcast, cool, high, strong) → {yes,no}

Exercise 2 – question 2

- Case $c' = \text{no}, c'' = \text{yes}$
  - $f = \left(\frac{9+y}{6-y}\right)^3 \cdot \frac{3}{2+y} \cdot \frac{1}{3+y} \cdot \frac{4}{3+y} \cdot \frac{3}{3+y} = \frac{36(9+y)^3}{(2+y)(6-y)^3(3+y)^3}$
  - $\frac{\mathrm{d}}{\mathrm{d}y} \ln f = \frac{3}{9+y} - \frac{1}{2+y} + \frac{3}{6-y} - \frac{3}{3+y} < 0$
    - $\min f$ in $y = 1$
    - $f(1) = 3/2 \Longrightarrow \text{no} > \text{yes}$
- Solution: (sunny, cool, high, strong) → no

Exercise 3

- ‘Temperature’ can be discarded (i.e., drop column T from table)
- Case $c' = \text{no}, c'' = \text{yes}$
- Basic formula:

$$c' > c'' \iff \inf_{0<y<1} \left[\left(\frac{n_{\text{yes}} + y}{n_{\text{no}} + 1 - y} \right)^2 \cdot \frac{n_{\text{no,sunny}}}{n_{\text{yes,sunny}} + y} \cdot \frac{n_{\text{no,high}}}{n_{\text{yes,high}} + y} \cdot \min_{w \in \{\text{strong,weak}\}} \frac{n_{\text{no},w}}{n_{\text{yes},w} + y} \right] > 1$$

  - min in w = weak for each y

- $f = \left(\frac{9+y}{6-y}\right)^2 \cdot \frac{3}{2+y} \cdot \frac{4}{3+y} \cdot \frac{2}{6+y}$
- $f(1) = 8/7 \implies$ no > yes
- Solution: (sunny, *, high, *) → no

Exercise 4

- Optimal tree:

$$\begin{tikzcd} \text{Humidity} \arrow[r] & \text{Temperature} \arrow[d] \\ \text{Wind} & \text{Outlook} \arrow[l] \end{tikzcd}$$

- Needed probabilities
  - PlayTennis: $p(\text{yes}|\boldsymbol{d}) \in [\frac{9}{15}, \frac{10}{15}], p(\text{no}|\boldsymbol{d}) \in [\frac{5}{15}, \frac{6}{15}]$
  - Humidity: $p(\text{high}|\text{yes}, \boldsymbol{d}) \in [\frac{3}{10}, \frac{4}{10}], p(\text{high}|\text{no}, \boldsymbol{d}) \in [\frac{4}{6}, \frac{5}{6}]$
  - Temperature: $p(\text{mild}|\text{high}, \text{yes}, \boldsymbol{d}) \in [\frac{2}{4}, \frac{3}{4}], p(\text{mild}|\text{high}, \text{no}, \boldsymbol{d}) \in [\frac{2}{5}, \frac{3}{5}]$
  - Outlook: $p(\text{sunny}|\text{mild}, \text{yes}, \boldsymbol{d}) \in [\frac{1}{5}, \frac{2}{5}], p(\text{sunny}, |\text{mild}, \text{no}, \boldsymbol{d}) \in [\frac{1}{3}, \frac{2}{3}]$
  - Wind: $p(\text{strong}|\text{sunny}, \text{yes}, \boldsymbol{d}) \in [\frac{1}{3}, \frac{2}{3}], p(\text{strong}|\text{sunny}, \text{no}, \boldsymbol{d}) \in [\frac{1}{4}, \frac{2}{4}]$
- yes $\not>$ no: $[\frac{9}{15} \cdot \frac{3}{10} \cdot \frac{2}{4} \cdot \frac{1}{5} \cdot \frac{1}{3}]/[\frac{6}{15} \cdot \frac{5}{6} \cdot \frac{3}{5} \cdot \frac{2}{3} \cdot \frac{2}{4}] = \frac{9}{100}$
- no $\not>$ yes: $[\frac{5}{15} \cdot \frac{4}{6} \cdot \frac{2}{5} \cdot \frac{1}{3} \cdot \frac{1}{4}]/[\frac{10}{15} \cdot \frac{4}{10} \cdot \frac{3}{4} \cdot \frac{2}{5} \cdot \frac{2}{3}] = \frac{5}{36}$
- Solution: (sunny, mild, high, strong) → {yes,no}

Marco Zaffalon: Knowledge discovery from data sets under tenable assumptions. School on Imprecise Probabilities, Lugano (CH), 27-31 July 2004.

Exercise 5 – assumptions

- Overall model = joint density $p(T, \boldsymbol{D}, \hat{\boldsymbol{D}}, \boldsymbol{O}, \hat{\boldsymbol{O}})$
- Factorization
  - $p(T, \boldsymbol{D}, \hat{\boldsymbol{D}}, \boldsymbol{O}, \hat{\boldsymbol{O}}) = p(T)p(\boldsymbol{D}, \hat{\boldsymbol{D}}|T)p(\boldsymbol{O}|\boldsymbol{D})p(\hat{\boldsymbol{O}}|\hat{\boldsymbol{D}})$
- Accuracy (of mechanism)
  - $p(\boldsymbol{o}|\boldsymbol{d})p(\hat{\boldsymbol{o}}|\hat{\boldsymbol{d}}) = 0$ if $\boldsymbol{d} \notin \boldsymbol{o}$ or $\hat{\boldsymbol{d}} \notin \hat{\boldsymbol{o}}$
- Positivity
  - Ideal data: $p(\boldsymbol{D}, \hat{\boldsymbol{D}}) > 0$
  - Actual observation: $p(\boldsymbol{o}, \hat{\boldsymbol{o}}) > 0$
- CAR
  - $p(\hat{\boldsymbol{o}}|\hat{\boldsymbol{d}}) = \alpha \quad \forall \hat{\boldsymbol{d}} \in \hat{\boldsymbol{o}}$

Exercise 5 – sketch of proof

- $$\begin{aligned} E(f|\boldsymbol{o}, \hat{\boldsymbol{o}}) &= \frac{\sum_{\boldsymbol{d} \in \boldsymbol{o}} p(\boldsymbol{o}|\boldsymbol{d}) \int f(\theta)p(\theta) \sum_{\hat{\boldsymbol{d}} \in \hat{\boldsymbol{o}}} p(\boldsymbol{d}, \hat{\boldsymbol{d}}|\theta)p(\hat{\boldsymbol{o}}|\hat{\boldsymbol{d}})d\theta}{\sum_{\boldsymbol{d} \in \boldsymbol{o}} p(\boldsymbol{o}|\boldsymbol{d}) \int p(\theta) \sum_{\hat{\boldsymbol{d}} \in \hat{\boldsymbol{o}}} p(\boldsymbol{d}, \hat{\boldsymbol{d}}|\theta)p(\hat{\boldsymbol{o}}|\hat{\boldsymbol{d}})d\theta} \\ &= \frac{\sum_{\boldsymbol{d} \in \boldsymbol{o}} p(\boldsymbol{o}|\boldsymbol{d}) \int f(\theta)p(\theta)p(\boldsymbol{d}, \hat{\boldsymbol{d}} \in \hat{\boldsymbol{o}}|\theta)d\theta}{\sum_{\boldsymbol{d} \in \boldsymbol{o}} p(\boldsymbol{o}|\boldsymbol{d}) \int p(\theta)p(\boldsymbol{d}, \hat{\boldsymbol{d}} \in \hat{\boldsymbol{o}}|\theta)d\theta} \\ &= E(f|\boldsymbol{o}, \hat{\boldsymbol{d}} \in \hat{\boldsymbol{o}}) \end{aligned}$$