Exercises

Exercise 1 – IDM

- Consider the following data set (assume Boolean variables)

#	C	A
1	0	1
2	0	1
3	0	1

- Use the IDM(s=2) sequentially to
  - compute the lower and upper probabilities of (C=1|A=1) for the observations 0, 1, 2, 3
  - classify the same observations
- Repeat the above using the precise Dirichlet model with Bayes-Laplace's uniform prior, and compare the results

Exercise 2 – NCC, complete data

- Consider the weather data set

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
d1	sunny	hot	high	weak	no
d2	sunny	hot	high	strong	no
d3	overcast	hot	high	weak	yes
d4	rainy	mild	high	weak	yes
d5	rainy	cool	normal	weak	yes
d6	rainy	cool	normal	strong	no
d7	overcast	cool	normal	strong	yes
d8	sunny	mild	high	weak	no
d9	sunny	cool	normal	weak	yes
d10	rainy	mild	normal	weak	yes
d11	sunny	mild	normal	strong	yes
d12	overcast	mild	high	strong	yes
d13	overcast	hot	normal	weak	yes
d14	rainy	mild	high	strong	no

- Use the NCC(s=1) to classify the instance (overcast, cool, high, strong)
- What if Outlook = sunny?

Exercise 3 – NCC, incomplete data

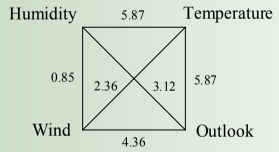
- Consider the incomplete weather data set

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
d1	sunny	hot	high	weak	no
d2	sunny	hot	high	strong	no
d3	overcast	hot	high	*	yes
d4	rainy	mild	high	weak	yes
d5	rainy	cool	normal	weak	yes
d6	rainy	cool	normal	strong	no
d7	overcast	cool	normal	strong	yes
d8	sunny	mild	high	weak	no
d9	sunny	cool	normal	weak	yes
d10	rainy	mild	normal	weak	yes
d11	*	mild	normal	strong	yes
d12	overcast	mild	high	strong	yes
d13	overcast	hot	normal	weak	yes
d14	rainy	mild	high	strong	no

- Use the NCC(s=1) to test whether or not 'no' dominates 'yes' with the instance (sunny, *, high, *)
  - Assume that 'Temperature' is missing at random, while 'Outlook' and 'Wind' are missing in non-ignorable way
  - Tip: minimum of function to optimize in $t_{yes} = 1$

Exercise 4 – TANC

- Consider the (complete) weather data set
- Use the TANC(s=1) to classify the instance (sunny, mild, high, strong)
  - Assuming the following values of conditional mutual information



  - and taking Humidity as root of the tree

Exercise 5 – mixed rules

- Sketch the theoretical derivation of the mixed rule in the learning setting
  - Stating new assumptions if needed

Appendixes

The multinomial case

- Traditional choice for $p(\theta)$ is Dirichlet(s,$\boldsymbol{t}$): $p(\theta) \propto \prod_{d\in\mathcal{D}} \theta_d^{st_d-1}$
  - s and $\boldsymbol{t} = (t_d)_{d\in\mathcal{D}}$ are hyperparameters
  - t_d = expectation of θ_d w.r.t. the prior
  - *Informal* view (prior is conjugate)
    - $s > 0$ as the number of additional (fractional) *virtual* observations
    - t_d as the proportion of observations equal to d in the virtual sample
    - $0 < t_d < 1, \sum_{d\in\mathcal{D}} t_d = 1$
  - e.g., uniform prior: $s = |\mathcal{D}|, t_d = 1/|\mathcal{D}|$; Perks prior: $s = 1, t_d = 1/|\mathcal{D}|$
- Posterior is Dirichlet(N+s,$\boldsymbol{t}^*$): $p(\theta|\boldsymbol{d}) \propto \prod_{d\in\mathcal{D}} \theta_d^{n_d+st_d-1}$
  - $\boldsymbol{t}^* = (t_d^*)_{d\in\mathcal{D}}, t_d^* = (n_d + st_d)/(N+s)$
  - t_d^* = expectation of θ_d w.r.t. the posterior
- The needed probability is then

$$p(c,x|\boldsymbol{d}) = E[\theta_{c,x}|\boldsymbol{d}] = \frac{n_{c,x}+st_{c,x}}{N+s}$$

The imprecise Dirichlet model (IDM)

- An imprecise probability method for multinomial sampling
  - Generalization of Bayesian approach
    - Model prior ignorance by the set of all the Dirichlet(s,$\boldsymbol{t}$) priors with s fixed
    - s is a *degree of caution*
  - From this, a set of posterior densities

$$p(\theta) \propto \prod_{d\in\mathcal{D}} \theta_d^{st_d-1}$$
$$\sum_{d\in\mathcal{D}} t_d = 1$$
$$0 < t_d < 1 \;\forall d \in \mathcal{D}$$

  - Lower and upper probabilities

$$p(c,x|\boldsymbol{d},s,\boldsymbol{t}) = \frac{n_{c,x}+st_{c,x}}{N+s}$$
$$\Downarrow$$
$$p(c,x|\boldsymbol{d},s) \in \left[\frac{n_{c,x}}{N+s}, \frac{n_{c,x}+s}{N+s}\right]$$

  - Imprecision (upper – lower) is $\frac{s}{N+s}$

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

Credal classification

- There is generally no single *optimal* class with imprecise models
  - e.g., $p(c'|x) \in [0.5, 0.8], p(c''|x) \in [0.3, 0.6]$
    - No class *dominates* the other (they are *incomparable*); both are plausible options
  - Opposite approach: rather than looking for the optimal class, the focus in on **discarding the dominated classes**
- *Credal dominance* (*strict preference*)
  - $c' > c''$ iff $p(c'|\boldsymbol{d},x,s,\boldsymbol{t}) > p(c''|\boldsymbol{d},x,s,\boldsymbol{t})$ for all $\boldsymbol{t}$ in the IDM
    iff $\inf_{\boldsymbol{t}} p(c',x|\boldsymbol{d},s,\boldsymbol{t})/p(c'',x|\boldsymbol{d},s,\boldsymbol{t}) > 1$
- Classification
  - Compare all the classes and output the undominated ones
    - The output is a set of classes! The set shrinks with sample size; **reliability**
- *Credal classification*

A *credal classifier* is a function $f : \mathcal{A}_1 \times \cdots \times \mathcal{A}_m \to \wp(\mathcal{C})$

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

The naive credal classifier (NCC)

- Recall that $p(c,x|\boldsymbol{d},s,\boldsymbol{t}) = \frac{n_c+st_c}{N+s}\prod_{j=1}^m \frac{n_{c,a_j}+st_{c,a_j}}{n_c+st_c}$
- The optimization problem for testing credal dominance is

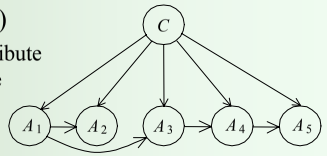
$$\inf_{\boldsymbol{t}} \frac{p(c',x|\boldsymbol{d},s,\boldsymbol{t})}{p(c'',x|\boldsymbol{d},s,\boldsymbol{t})} = \inf\left[\left(\frac{n_{c'}+st_{c'}}{n_{c'}+st_{c'}}\right)^{m-1}\prod_{j=1}^m \frac{n_{c',a_j}+st_{c',a_j}}{n_{c'',a_j}+st_{c'',a_j}}\right]$$
$$\text{s.t. } \sum_{c\in\mathcal{C}} t_c = 1$$
$$0 < t_c < 1 \quad \forall c \in \mathcal{C}$$
$$0 < t_{c,a_j} < t_c \quad \forall c \in \mathcal{C}, j \in \{1,\ldots,m\}$$
$$= \inf\left[\left(\frac{n_{c'}+st_{c'}}{n_{c'}+st_{c'}}\right)^{m-1}\prod_{j=1}^m \frac{n_{c',a_j}}{n_{c'',a_j}+st_{c''}}\right]$$
$$\text{s.t. } t_{c'}+t_{c''} = 1$$
$$t_{c'}, t_{c''} > 0$$

- The latter is a convex optimization problem in a single variable!
  - *Global* optimum found rapidly via numerical method, e.g., Newton-Raphson

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

Relaxing the independence assumption: the TAN model

- *Tree augmented naive Bayes* (TAN)
  - The dependence structure between attribute variables, given class variable, is a tree



- Learning problem
  - Learning the tree structure + probabilities
  - Mix of *maximum likelihood* and Bayesian techniques
- Classification

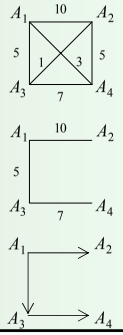
$$p(c,x|\boldsymbol{d}) = p(c|\boldsymbol{d})\prod_{j=1}^m p(a_j|\pi_{A_j},\boldsymbol{d})$$

Learning TAN structure

- Consider the *conditional mutual information* (MI)

$$\mathrm{MI}(A_j, A_k|C) = \sum_{c\in\mathcal{C}} \theta_c \sum_{(a_j,a_k)\in\mathcal{A}_j\times\mathcal{A}_k} \theta_{a_j,a_k|c} \log \frac{\theta_{a_j,a_k|c}}{\theta_{a_j|c}\theta_{a_k|c}}$$

- Procedure
  - Set up a fully connected undirected graph
    - Node ⇔ attribute variable
    - Weight each edge (A_j, A_k) by the *empirical* $\mathrm{MI}(A_j, A_k|C)$
      - Empirical MI = replace chances with relative frequencies
  - Compute the *maximum weight spanning tree*
  - Arbitrarily choose root node



- The TAN T constructed this way maximizes the likelihood $p(\boldsymbol{d}|T)$

Tree-augmented naive credal classifier (TANC)

- Straightforward extension of TAN to credal classification
- Structure learning as in the precise case
- IDM-based learning of probabilities
  - Replace $p(a_j|c,a_k,\boldsymbol{d},s,\boldsymbol{t}) = \dfrac{n_{a_j,c,a_k} + st_{a_j,c,a_k}}{n_{c,a_k}+s}$ with

$$p(a_j|c,a_k,\boldsymbol{d},s) \in \left[\frac{n_{a_j,c,a_k}}{n_{c,a_k}+s}, \frac{n_{a_j,c,a_k}+s}{n_{c,a_k}+s}\right]$$

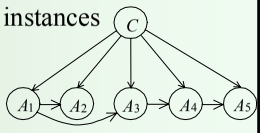
  that is, $p(a_j|\pi_{A_j}) \in \left[\underline{p}(a_j|\pi_{A_j}), \overline{p}(a_j|\pi_{A_j})\right]$
  - This way makes different credal sets be *separately specified*
    - (Conditional) Credal sets in the same node and in different nodes

TANC classification

- $\underline{p}(c, a_1, \ldots, a_m) = \underline{p}(c)\prod_{j=1}^m \underline{p}(a_j|\pi_{A_j})$
- Credal dominance test becomes (complete instance)

$$\frac{\underline{p}(c', a_1, \ldots, a_m)}{\overline{p}(c'', a_1, \ldots, a_m)} = \frac{\underline{p}(c')}{\overline{p}(c'')} \prod_{j=1}^m \frac{\underline{p}(a_j|\pi'_{A_j})}{\overline{p}(a_j|\pi''_{A_j})} > 1$$

  - Linear time
- More involved with incomplete (ignorable) instances
  - $\underline{p}(c, a_5)$?
  - $A_1, \ldots, A_4$ must be "marginalized out"
  - The expression to minimize involves a sum



- Exact minimization by propagating intervals over the tree
  - Linear time again

Learning – ideal variables

- Ideal learning data are regarded as an instance of the random matrix

$$\begin{bmatrix} C_1 & A_{11} & \cdots & A_{1j} & \cdots & A_{1k} \\ \vdots & \vdots & & \vdots & & \vdots \\ C_i & A_{i1} & \cdots & A_{ij} & \cdots & A_{ik} \\ \vdots & \vdots & & \vdots & & \vdots \\ C_N & A_{N1} & \cdots & A_{Nj} & \cdots & A_{Nk} \end{bmatrix} = \begin{bmatrix} C_1 & X_1 \\ \vdots & \vdots \\ C_i & X_i \\ \vdots & \vdots \\ C_N & X_N \end{bmatrix} = \begin{bmatrix} D_1 \\ \vdots \\ D_i \\ \vdots \\ D_N \end{bmatrix} = \boldsymbol{D}$$

  - C_i's are *class variables*, with values in $\mathcal{C}$
  - A_{ij}'s are *attribute variables*, with values in $\mathcal{A}_j$ for each (i,j)
  - Each row D_i in the matrix is called *unit*, with values in

$$\mathcal{D} = \mathcal{C}\times\mathcal{X} = \mathcal{C}\times\mathcal{A}_1\times\cdots\times\mathcal{A}_m$$

Learning – actual variables

- Consider one more random variable
- $\boldsymbol{O}$ is the *actual observation* of $\boldsymbol{D}$
  - $\boldsymbol{D}$ takes values $\boldsymbol{d}$ from $\mathcal{D}^N$
  - $\boldsymbol{O}$ takes values $\boldsymbol{o}$ from $\wp(\mathcal{D}^N)$
    - Rather than observing $\boldsymbol{d}$, you observe a set $\boldsymbol{o}$ that contains it
    - This is called *coarsening*, i.e., looking at $\boldsymbol{d}$ with different levels of detail
    - Note that $\boldsymbol{o}$ is simply a symbol, not a set, when regarded as the value of $\boldsymbol{O}$

Learning – assumptions

- Overall model = joint density $p(T, \boldsymbol{D}, \boldsymbol{O})$
  - T is a random parameter with values $\theta \in \Theta$
- Factorization
  - $p(T, \boldsymbol{D}, \boldsymbol{O}) = p(T)p(\boldsymbol{D}|T)p(\boldsymbol{O}|\boldsymbol{D})$
    - $p(T)$ is an imprecise prior for T, i.e., it belongs to a certain non-empty set
    - The IM depends only on $\boldsymbol{D}$
- Accuracy (of mechanism)
  - $p(\boldsymbol{o}|\boldsymbol{d}) = 0$ if $\boldsymbol{d}$ $\boldsymbol{o}$
  - Connection between complete and incomplete observations
- Positivity
  - Ideal data: $p(\boldsymbol{D}) > 0$
  - Actual observation: $p(\boldsymbol{o}) > 0$
    - There exists $\boldsymbol{d}$ s.t. $p(\boldsymbol{o}|\boldsymbol{d}) > 0$

Learning – problem

- Formulation of the learning problem
  - Using observed data $\boldsymbol{o}$ to update beliefs about a function $f : \Theta \to \mathbb{R}$
    - e.g., θ real number, $f(\theta) = \theta$
  - In the precise framework, one would compute $E(f|\boldsymbol{o})$
- $$E(f|\boldsymbol{o}) = \frac{\int f(\theta)p(\theta, \boldsymbol{o})d\theta}{p(\boldsymbol{o})} = \frac{\int f(\theta) \sum_{\boldsymbol{d}\in\boldsymbol{o}} p(\theta)p(\boldsymbol{d}|\theta)p(\boldsymbol{o}|\boldsymbol{d})d\theta}{\sum_{\boldsymbol{d}\in\boldsymbol{o}} p(\boldsymbol{o}|\boldsymbol{d})p(\boldsymbol{d})}$$
  - Using Factorization, Accuracy and Positivity
- What about $p(\boldsymbol{o}|\boldsymbol{d})$ and $p(\theta)$?
  - Imprecise knowledge

Learning – express ignorance about the IM

- Focus on $p(\boldsymbol{o}|\boldsymbol{d})$, $\boldsymbol{d}\in\boldsymbol{o}$
  - Call $p(\boldsymbol{o}|\boldsymbol{D})$ the vector with elements $p(\boldsymbol{o}|\boldsymbol{d})$, $\boldsymbol{d}\in\boldsymbol{o}$
- We can only constrain the set of possible vectors $p(\boldsymbol{o}|\boldsymbol{D})$

$$\sum_{\boldsymbol{d}\in\boldsymbol{o}} p(\boldsymbol{o}|\boldsymbol{d}) > 0$$
$$0 \le p(\boldsymbol{o}|\boldsymbol{d}) \le 1,\ \boldsymbol{d} \in \boldsymbol{o}$$

  - The restriction on the sum is due to Accuracy
  - The inequality following the sum is due to Positivity
- The inequalities define an open linear set called $\mathcal{P}(\boldsymbol{o}|\boldsymbol{D})$
- Call $\mathcal{P}_\varepsilon(\boldsymbol{o}|\boldsymbol{D})$ the approximating closed set

$$\sum_{\boldsymbol{d}\in\boldsymbol{o}} p(\boldsymbol{o}|\boldsymbol{d}) \ge \varepsilon$$
$$0 \le p(\boldsymbol{o}|\boldsymbol{d}) \le 1,\ \boldsymbol{d} \in \boldsymbol{o}$$

Learning – goal

- Goal: $\underline{E}(f|\boldsymbol{o}) = \inf_{p(T)\in\mathcal{P}(T)} \inf_{p(\boldsymbol{o}|\boldsymbol{D})\in\mathcal{P}(\boldsymbol{o}|\boldsymbol{D})} E(f|\boldsymbol{o})$
- Focus on $\min_{p(\boldsymbol{o}|\boldsymbol{D})\in\mathcal{P}_\varepsilon(\boldsymbol{o}|\boldsymbol{D})} E(f|\boldsymbol{o})$
  - i.e., $\min_{p(\boldsymbol{o}|\boldsymbol{D})\in\mathcal{P}_\varepsilon(\boldsymbol{o}|\boldsymbol{D})} \frac{\sum_{\boldsymbol{d}\in\boldsymbol{o}} p(\boldsymbol{o}|\boldsymbol{d}) \int f(\theta)p(\theta)p(\boldsymbol{d}|\theta)d\theta}{\sum_{\boldsymbol{d}\in\boldsymbol{o}} p(\boldsymbol{o}|\boldsymbol{d})p(\boldsymbol{d})}$
  - Objective function is ratio of linear functions
- Fractional programming theorem
  - Consider $\min_{x\in S} \frac{q(x)}{r(x)}$ where S = compact subset of $\mathbb{R}^v$
    - q, r continuous and r positive on S
  - Define $h : \mathbb{R} \to \mathbb{R}$, $h(\mu) = \min_{x\in S}[q(x) - \mu r(x)]$
  - Then $\mu^* = \operatorname{argmin}_{x\in S} \frac{q(x)}{r(x)} \iff h(\mu^*) = 0$
    - h has a single zero

Learning – solution: conservative learning rule

- $$h(\mu) = \min_{p(\boldsymbol{o}|\boldsymbol{D})\in\mathcal{P}_\varepsilon(\boldsymbol{o}|\boldsymbol{D})} \left[\sum_{\boldsymbol{d}\in\boldsymbol{o}} p(\boldsymbol{o}|\boldsymbol{d}) \int f(\theta)p(\theta)p(\boldsymbol{d}|\theta)d\theta - \mu \sum_{\boldsymbol{d}\in\boldsymbol{o}} p(\boldsymbol{o}|\boldsymbol{d})p(\boldsymbol{d})\right]$$
  $$= \min_{p(\boldsymbol{o}|\boldsymbol{D})\in\mathcal{P}_\varepsilon(\boldsymbol{o}|\boldsymbol{D})} \sum_{\boldsymbol{d}\in\boldsymbol{o}} p(\boldsymbol{o}|\boldsymbol{d})p(\boldsymbol{d}) \left[\int f(\theta)p(\theta|\boldsymbol{d})d\theta - \mu\right]$$
  $$= \min_{p(\boldsymbol{o}|\boldsymbol{D})\in\mathcal{P}_\varepsilon(\boldsymbol{o}|\boldsymbol{D})} \sum_{\boldsymbol{d}\in\boldsymbol{o}} p(\boldsymbol{o}|\boldsymbol{d})p(\boldsymbol{d}) \left[E(f|\boldsymbol{d}) - \mu\right]$$
- $\mu^* = \min_{\boldsymbol{d}\in\boldsymbol{o}} E(f|\boldsymbol{d}) \implies h(\mu^*) = 0$
- $\min_{p(\boldsymbol{o}|\boldsymbol{D})\in\mathcal{P}_\varepsilon(\boldsymbol{o}|\boldsymbol{D})} E(f|\boldsymbol{o}) = \min_{\boldsymbol{d}\in\boldsymbol{o}} E(f|\boldsymbol{d})$

⇩

$$\underline{E}(f|\boldsymbol{o}) = \inf_{p(T)\in\mathcal{P}(T)} \min_{\boldsymbol{d}\in\boldsymbol{o}} E(f|\boldsymbol{d})$$

Obtaining stronger conclusions: the mixed rules

- CAR might hold in some cases
  - This should not be neglected in order to strengthen conclusions
- Some attribute variables are CAR, the others non-ignorable
- Mixed rules
  - Learning: $\underline{E}(f|\boldsymbol{o}, \hat{\boldsymbol{o}}) = \inf_{p(T)\in\mathcal{P}(T)} \min_{\boldsymbol{d}\in\boldsymbol{o}} E(f|\boldsymbol{d}, \hat{\boldsymbol{d}} \in \hat{\boldsymbol{o}})$
  - Classification: $\underline{E}_C(g|\boldsymbol{o}^+, \hat{\boldsymbol{o}}^+) = \inf_{p(T)\in\mathcal{P}(T)} \min_{\boldsymbol{d}^-\in\boldsymbol{o}^-} E_C(g|\boldsymbol{d}^-, \hat{\boldsymbol{d}}^- \in \hat{\boldsymbol{o}}^-)$

Inference of the NCC from incomplete data

- Test of credal dominance with an incomplete learning sample

$$\min_{\boldsymbol{d}\in\boldsymbol{o}} \inf_{\boldsymbol{t}} \frac{p(c', x|\boldsymbol{d}, s, \boldsymbol{t})}{p(c'', x|\boldsymbol{d}, s, \boldsymbol{t})} = \inf \left[\left(\frac{n_{c''} + st_{c''}}{n_{c'} + st_{c'}}\right)^{m-1} \prod_{j=1}^{m} \frac{\underline{n}_{c',a_j}}{\overline{n}_{c'',a_j} + st_{c''}}\right]$$
$$\text{s.t. } t_{c'} + t_{c''} = 1$$
$$t_{c'}, t_{c''} > 0$$

- Same complexity as with complete data