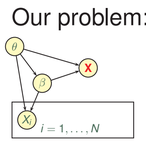
A Hierarchical Approach to Learning with Imprecise Probabilities

Serafín Moral
smc@decsai.ugr.es
Department of Computer Science and Artificial Intelligence. University of Granada, Spain

ISIPTA '19 - July 3-6, 2019, Ghent

1.- Introduction

Our problem:



- We have imprecise information about $\theta \in \Theta$ given by a coherent set of desirable gambles.
- Given $\theta \in \Theta$ we have a full Bayesian model about $\beta \in B$ and X_i.
- We have observations $\mathcal{O} : \mathbf{X}_1, \ldots, \mathbf{X}_N$ that are conditionally independent given θ and β.
- We want to estimate a model for $\mathbf{X}$ (with the same distribution than $\mathbf{X_i}$) under conditional independence.

1. Coherent Set of Desirable Gambles

A gamble: a mapping $f : \Theta \to \mathbb{R}$
$\mathcal{L}$ is the set of all gambles.
Coherent Set of Desirable Gambles $\mathcal{D}$

D1. $0 \notin \mathcal{D}$,
D2. if $f \in \mathcal{L}$ and $f > 0$ then $f \in \mathcal{D}$,
D3. if $f \in \mathcal{D}$ and $c \in \mathbb{R}$ with $c > 0$ then $cf \in \mathcal{D}$,
D4. if $f \in \mathcal{D}$ and $g \in \mathcal{D}$ then $f + g \in \mathcal{D}$.

$\mathcal{L}^+ = \{f \in \mathcal{L} : f > 0\}$.

Natural Extension

$$\text{posi}(\mathcal{K}) = \{\sum_{i=1}^{k} c_i f_i \ : \ c_i > 0, f_i \in \mathcal{K}, k \geq 1\},$$

If $\mathcal{A}$ is a set of gambles it avoids partial loss if there is a coherent set of gambles containing it.
If $\mathcal{A}$ avoid partial loss, its natural extension is equal to $\text{posi}(\mathcal{A} \cup \mathcal{L}^+)$: the intersection of all coherent sets of gambles containing $\mathcal{A}$.

Vacuous Coherent Credal Set

No information: $\mathcal{D}_0 = \mathcal{L}^+$.

Conditioning

Conditioning of a coherent set of desirable gambles $\mathcal{D}$ to a likelihood function $L : \Theta \to \mathbb{R}$,

$$\mathcal{D}_L = \{f \ : \ f.L \in \mathcal{D}\}$$

If A is a subset of Θ, then conditioning $\mathcal{D}$ to A, $\mathcal{D}_A$, means conditioning to the likelihood equal to the indicator function of A, I_A.

Credal Set

A set of desirable gambles on Θ defines a convex set of finitely additive probability measures (a credal set) on Θ:

$$\mathcal{M}_{\mathcal{D}} = \{P \mid P(f) \geq 0, \forall f \in \mathcal{D}\}, \quad (1)$$

Credal sets and coherent sets of desirable gambles are not equivalent.

2.- The Uniform Coherent Set

This uniform prior, $\mathcal{D}_u$, is the natural extension of the set of gambles:

$$\mathcal{K}_u = \{I_\theta - \alpha I_{\theta'} \ : \ \theta, \theta' \in \Theta, \alpha < 1\}$$

Finite Θ

Result
If Θ is finite, $\mathcal{D}_u$ is equal to the set of gambles f such that $\sum_{\theta \in \Theta} f(\theta) > 0$.

The set $\mathcal{M}_{\mathcal{D}_u}$ contains only one element: P_u, the uniform probability in Θ.

Infinite Θ

$\text{Support}^+(f) = \{\theta \in \Theta : f(\theta) > 0\}$
$\text{Support}^-(f) = \{\theta \in \Theta : f(\theta) < 0\}$.

Result
If Θ is infinite, D_u is the set of gambles f for which $\text{Support}^-(f)$ is finite and there is $H \subseteq \text{Support}^+(f)$ with H finite and $\sum_{\theta \in (H \cup \text{Support}^-(f))} f(\theta) > 0$.

Result
If Θ is infinite, then $\mathcal{M}_{\mathcal{D}_u}$ is equal to the set of all the finitely additive probability measures in Θ such that $P(H) = 0$ for any $H \subseteq \Theta$ finite.

This is a very large set of probabilities. In particular it contains all the probability measures associated to continuous densities (it is almost vacuous).

Result
If H is finite, the conditioning of $\mathcal{D}_u$ to H, D_{uH}, will contain all the gambles f such that $\sum_{\theta \in H} f(\theta) > 0$, and the associated credal set is given by the uniform probability in H (equivalent to the uniform in H).

3. The Discounted Uniform Credal Set

Discounting a Coherent Credal Set

If $\mathcal{D}$ is a set of desirable gambles and $\epsilon \in [0, 1]$, then the discounting of $\mathcal{D}$ by ϵ is

$$\mathcal{D}^\epsilon = \{f - \epsilon \inf(f) I_{\text{Support}(f)} \ : \ f \in \mathcal{D}\} \setminus \{0\}.$$

- If $\mathcal{D}$ is coherent, then $\mathcal{D}^\epsilon$ is coherent
- $\mathcal{D}^\epsilon \subseteq \mathcal{D}$
- $\mathcal{D}^0 = \mathcal{D}$
- $\mathcal{D}^1$ is the vacuous set of gambles $\mathcal{D}_0$.
- $\mathcal{M}_{\mathcal{D}^\epsilon} = (1 - \epsilon)\mathcal{M}_{\mathcal{D}} + \epsilon\mathcal{M}_0 = \{(1 - \epsilon)P + \epsilon Q \ : \ P \in M_{\mathcal{D}}, Q \in \mathcal{M}_0\}$
  $\mathcal{M}_0$ is the vacuous credal set (all the probabilities).

The Discounted Imprecise Model $\mathcal{D}_u^\epsilon$

Result
If f is a gamble with finite support, then $f \in \mathcal{D}_u^\epsilon$ if and only if $(1 - \epsilon) \sum_{\theta \in \text{Support}(f)} f(\theta) + \epsilon \inf(f) |\text{Support}(f)| > 0$.

Gambles with two points of support

Assume f in $\mathcal{D}_u^\epsilon$ θ_1, θ_2 and such that $f(\theta_1) > 0 > f(\theta_2)$.
f is desirable in $\mathcal{D}_u^\epsilon$ if and only if

$$(1 - \epsilon)f(\theta_1) + (1 + \epsilon)f(\theta_2) > 0.$$

Or, equivalently, if

$$\frac{-f(\theta_2)}{f(\theta_1)} < \frac{1 - \epsilon}{1 + \epsilon}.$$

4. Information of B, $\mathbf{X}$ given θ

- We have precise probabilities $P(\beta|\theta)$ and $P(\mathbf{X}|\beta, \theta)$.
- We are interested in learning a model for variables $\mathbf{X}$ given a dataset $\mathcal{O}$ of observations for variables $\mathbf{X}$.
- Learning with a fixed $\theta \in \Theta$ is equivalent to conditioning the parameter β to the observations $\mathcal{O}$, i.e. computing $P(\beta|\theta, \mathcal{O})$ (Bayesian model averaging):

$$P(\mathbf{X}|\mathcal{O}, \theta) = \int_B P(\beta|\theta, \mathcal{O}) P(\mathbf{X}|\beta, \theta) d\beta$$

- The set of observations defines a likelihood in Θ:

$$L(\theta) = \int_B P(\beta|\theta).P(\mathcal{O}|\beta, \theta) d\beta.$$

5. Full Model Selection Procedure

- Solve the decision problem in Θ with $\mathcal{D}_u^\epsilon$: there is a decision d_θ for each $\theta \in \Theta$, identified with a gamble $d_\theta(\theta')$ (the utility of selecting $\theta \in \Theta$ when the true value of the parameter is θ'). A 0-1 utility is considered: $d_\theta(\theta') = 1$ if $\theta = \theta'$ and 0,
  1. Compute the conditional set of desirable gambles $\mathcal{D}_{uL}^\epsilon$ for the likelihood associated to the observations.
  2. Select the set of maximal (non-dominated) decisions on the parameter space Θ, i.e. the set

$$H_L = \{d_\theta \ : \ d_{\theta'} - d_\theta \notin \mathcal{D}_{uL}^\epsilon, \forall \theta' \in \Theta\}.$$

- Finally, consider the set of models about $\mathbf{X}$ associated with these decisions,

$$\mathcal{M}_{\mathcal{O}} = \{P(\mathbf{X}|\mathcal{O}, \theta) \ : \ d_\theta \in H_L\},$$

where $P(\mathbf{X}|\mathcal{O}, \theta)$ is the Bayesian model averaging expression.

Result
$d_{\theta'} - d_\theta \in \mathcal{D}_{uL}^\epsilon$ if and only if

$$\frac{L(\theta)}{L(\theta')} < \frac{1 - \epsilon}{1 + \epsilon}.$$

θ is dominated if and only if

$$\frac{L(\theta)}{L(\hat{\theta})} < \frac{1 - \epsilon}{1 + \epsilon}.$$

where $\hat{\theta}$ is the maximum likelihood value.

6. Estimating Multinomial Probabilities

A single variable X with K possible values $\{x_1, x_2, \ldots, x_K\}$. We want to estimate $P(X = x_i) = \theta_i, (i = 1, \ldots, K)$

Maximum Likelihood

- $\Theta = \{\theta = (\theta_1, \ldots, \theta_K) \mid \sum_{i=1}^K \theta_i = 1, \theta_i \geq 0\}$
- $B = \{\beta\}$ (equivalent to no parameter)
- $\epsilon = 0$ (no discounting)

θ is dominated if and only if there is θ' with $\frac{L(\theta)}{L(\theta')} < 1$, where

$$L(\theta) = \theta_1^{n_1} \ldots . \theta_K^{n_K}.$$

Maximum likelihood estimation: $\hat{\theta}_i = n_i/N$ (n_i observations of $X = x_i$).

Maximum Likelihood Confidence Regions

Same setting, but $\epsilon > 0$ (real discounting).
The non dominated θ

$$\frac{L(\theta)}{L(\hat{\theta})} \geq \frac{1 - \epsilon}{1 + \epsilon},$$

This is a pure likelihood-based confidence region:

$$\{\theta \ : \ \frac{L(\theta)}{L(\hat{\theta})} \geq c\}$$

. where the parameter in this case is: $c = \frac{1-\epsilon}{1+\epsilon}$.

Bayesian Point Estimation

- $\Theta = \{s\}$ is a single hyperparameter (the equivalent sample size)
- $B = \{\theta = (\theta_1, \ldots, \theta_K) \mid \sum_{i=1}^K \theta_i = 1, \theta_i \geq 0\}$.
- The prior probability in B conditioned to s is Dirichlet:

$$P(\theta_1, \ldots, \theta_K|s) = \frac{(\Gamma(s/K))^K}{\Gamma(s)} \theta_1^{s/K} \ldots . \theta_1^{s/K},$$

  where $\Gamma()$ is the Gamma function.
- The probability of the observations is $P(X = x_i|\theta, s) = \theta_i$.

We only have Bayesian model averaging, estimating $P(X = x_i|\mathcal{O}, s)$ by $\frac{n_i + s/K}{N+s}$.

Example
Binary variable X with values in $\{0, 1\}$.
100 independent values of X ($X_1, \ldots, X_{100}$) which have been partially observed: We only know the difference between the number of 0s and the number of 1
Assume that the obsertation is = 100,
If $P(X = 0) = \theta$, then the likelihood function is:

$$L(\theta) = P(X_1 = X_2 = \cdots = X_{100}|\theta) = \theta^{100} + (1 - \theta)^{100}.$$

The estimated value is $\theta^* = 0.5$, however, given the observations we have a strong belief that the probabilities will be extreme.
The likelihood confidence region with $c = 0.95$ is $[0, 0.02951305] \cup [0.970487, 1.0]$, showing that the parameter value should be close to 0 or 1.

Bayesian High-Density Regions

Selecting a region H with high value. With classical uniform density: the same than maximum likelihood confidence regions (same shape), but parameter is now γ and we have to have $P(H|\mathcal{O}) \geq \gamma$

The Imprecise Dirichlet Model (IDM)

- There is an hyperparameter s (the equivalent sample size) and $\Theta = \{\alpha = (\alpha_1, \ldots, \alpha_k) \ : \ \sum_{i=1}^K \alpha_i = s, \alpha_i > 0\}$
- $\epsilon = 1$, i.e. vacuous information on Θ and there is no dominance (all the parameters in Θ are considered)
- $B = \{\theta = (\theta_1, \ldots, \theta_K) \mid \sum_{i=1}^K \theta_i = 1, \theta_i \geq 0\}$, the conditional density is Dirichlet $(\alpha_1, \ldots, \alpha_K)$.
- $P(X = x_i|\alpha, \theta) = \theta_i$.

The estimation of $P(X = x_i|\alpha, \mathcal{O})$ is $\frac{n_i + \alpha_i}{N+s}$ for an α, then $P(X = x_i|\mathcal{O}) \in [\frac{n_i}{n_i+s}, \frac{n_i+s}{n_i+s}]$.

6. Estimating Multinomial Probabilities (Cont.)

Empirical Bayes

- $\Theta = [s_1, s_2]$. No discounting of the uniform model.
- $B = \{\theta = (\theta_1, \ldots, \theta_K) \mid \sum_{i=1}^K \theta_i = 1, \theta_i \geq 0\}$, with $P(\theta_1, \ldots, \theta_K|s)$ a Dirichlet$(s/K, \ldots, s/K)$.
- $P(X = x_i|\theta, s) = \theta_i$

The likelihood in $[s_1, s_2]$ given the observations is:

$$L(s) = \frac{\Gamma(s)}{\Gamma(N + s)} \prod_{i=1}^K \frac{\Gamma(n_i + s/k)}{\Gamma(s/K)}.$$

The only parameter in Θ that is not dominated is parameter $\hat{s}$, maximizing $L(s)$ and the final estimation $\frac{n_i + \hat{s}/K}{N+\hat{s}}$.

Imprecise Sample Size Dirichlet Model

- The same setting, but discounting the uniform distribution by $\epsilon = 1$ (no dominance):

$$[\min\{\frac{n_i + s_1/K}{N + s_2}, \frac{n_i + s_2/K}{N + s_1}\}, \max\{\frac{n_i + s_1/K}{N + s_1}, \frac{n_i + s_2/K}{N + s_2}\}].$$

- An intermediate solution: the ϵ-discounted uniform. If $\hat{s}$ is maximum likelihood, then $H_L = \{s \in \Theta \mid L(s) \geq \frac{1-\epsilon}{1+\epsilon}\}$ and $P(X_i = x_i|\mathcal{O})$ is

$$[\min_{s \in H_L} \frac{n_i + s/K}{N + s}, \max_{s \in \Theta'} \frac{n_i + s/K}{N + s}]$$

.

The Imprecise Dirichlet Model with α-cut Conditioning

Problems of IDM with non-direct observations: too wide intervals (Piatti, Zaffalon, 2007): f X is not directly observed, but we observe Y which is X with a very small error: estimated intervals are vacuous.
α-cut conditioning Cattaneo (2004) : same setting than IDM with discounting $\epsilon < 1$. His parameter is $c = \frac{1-\epsilon}{1+\epsilon}$.
We must compute: $\hat{\alpha} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_K)$ maximizing the marginal likelihood,

$$L(\alpha) = L(\alpha_1, \ldots, \alpha_K) = \frac{\Gamma(s)}{\Gamma(N + s)} \prod_{i=1}^K \frac{\Gamma(n_i + \alpha_i)}{\Gamma(\alpha_i)}.$$

Then, $\Theta = H_L = \{\alpha \ : \ L(\alpha) \geq L(\hat{\alpha})^{\frac{1-\epsilon}{1+\epsilon}}\}$.
Estimated probabilities:

$$\left[\min_{\alpha \in H_L} \frac{n_i + \alpha_i}{N + s}, \max_{\alpha \in H_L} \frac{n_i + \alpha_i}{N + s}\right].$$

Partitioned Sets of Parameters

Example
Two observed variables (X, Y) urns 10 balls in two different colours, red (R) and white (W).
- X and Y are selected from different urns of unknown composition.
- X and Y are selected from the same urn (with replacement)

Θ two parts:
- $\Theta_1 = \{(D, r_{i1}, r_{j2}) \mid r_{i1}, r_{i2} \in \{0, 1, \ldots, 10\}\}$, representing the first situation, in which r_{i1}, r_{i2} are the number of red balls in urns 1 and 2 respectively.
- $\Theta_2 = \{(E, r_i) \mid r_i \in \{0, 1, \ldots, 10\}\}$, representing the case of one urn for the two extractions, being r_i the number of red balls.

Θ_1 and Θ_2 do not have the same size (121 for Θ_1 and 11 for Θ_2).
$\Theta_1 \preceq \Theta_2$

Finite Θ: Generalized uniform, a gamble $aI_{\theta_i} + bI_{\theta_j}$ is desirable with $\theta_i \in \Theta_i, \theta_j \in \Theta_j$, when $a/|\Theta_i| + b/|\Theta_j| > 0$
When Θ_1 and Θ_2 represents different numbers of parameters (M_1 and M_2) from a set of cardinal m (Similar to BIC or Akaike):

$$\log(L(\theta_1)) - M_1 \log(m) > \log(L(\theta_2)) - M_2 \log(m).$$

Infinite Θ: A virtual size (BIC, Akaike)

Learning Credal Networks

Several Possibilities:

- Θ the set of all directed acyclic graphs G for variables $\mathbf{X} = (X_1, \ldots, X_m)$, $\mathcal{G}$
  B being the set of parameters $\{\theta_G \in \Theta_G \mid G \in \mathcal{G}\}$.
  Given G, θ_G, follows independent Dirichlet distributions for the parameters of the network.
  Without discounting: selecting the graph with maximum BDEu Score:

$$BDEu(G) = P(\mathcal{O}|G) = \prod_{i=1}^m \prod_{j=1}^{R_i} \frac{\Gamma(s/R_i)}{\Gamma(n_{ij} + s/R_i)} \prod_{k=1}^{K_i} \frac{\Gamma(n_{ijk} + s/(R_i K_i))}{\Gamma(s/(R_i K_i))}.$$

- $\Theta = \{(G, \theta_G) \ : \ G \in \mathcal{G}, \theta_G \in \Theta_G\}$, which is partitioned as $\Theta = \bigcup_{G \in \mathcal{G}} \Theta'_G$, where $\Theta'_G = \{(G, \theta_G) \mid \theta_G \in \Theta_G\} = \{G\} \times \Theta_G$. We can obtain Akaike score:

$$AIC(G, \theta_G) = \log(L(G, \theta)) = \sum_{i=1}^m \sum_{j=1}^{R_i} \sum_{k=1}^{K_i} n_{ijk} \log(\theta_{ijk}) - \sum_{i=1}^m R_i(K_i - 1)$$

If the uniform model is discounted:

- Compute the pairs $(G, \hat{\theta}_G)$ where $BDEu(G) \geq \frac{1-\epsilon}{1+\epsilon} \max_{G' \in \mathcal{G}} BDEu(G')$.
- In the Akaike information criterion we must compute the pair $(\hat{G}, \hat{\theta}_G)$ by maximizing it, and then all the pairs (G, θ_G) such that $AIC(G, \theta_G) \geq \log\left(\frac{1-\epsilon}{1+\epsilon}\right) + AIC(\hat{G}, \hat{\theta}_G)$.

Main References

[1] Marco EGV Cattaneo. A continuous updating rule for imprecise probabilities. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 426–435. Springer, 2014.

[2] Peter Gärdenfors and Nils-Eric Sahlin. Unreliable probabilities, risk taking, and decision making. *Synthese*, 53(3): 361–386, 1982.

[3] **Serafín Moral. Learning with imprecise probabilities as model selection and averaging. *International Journal of Approximate Reasoning*, 109:111–124, 2019.**

[4] Peter Walley. Towards a unified theory of imprecise probability. *International Journal of Approximate Reasoning*, 24: 125–148, 2000.

Acknowledgements

Research supported by the Spanish Ministry of Economy and Competitiveness (project TIN2016-77902-C3-2-P), and the European Regional Development Fund (FEDER).