Modeling Deterministic Equations in Discrete Bayesian Networks by Credal Networks

Alessandro Antonucci

Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (IDSIA)
Manno-Lugano (Switzerland)

E-mail: `alessandro@idsia.ch`

Abstract. We focus on the problem of embedding deterministic equations in discrete Bayesian networks. This is typically achieved by a discretization of both input and output variables and a degenerate quantification of the corresponding conditional probability table. We note that, generally speaking, this approach based on the classical Bayesian theory of probability cannot properly take into account the information loss induced by the discretization. Consequently, we propose a conservative modeling of such *epistemic* uncertainty by means of *credal sets*, i.e., sets of probability mass functions. This approach transforms the original Bayesian network in a *credal network*, returning interval-valued inferences, that are robust with respect to the discretization. Procedures for optimal discretization in this setup are discussed. The proposed method can be used for both developing knowledge-based expert systems as well as machine learning based on probabilistic graphical models.

1. Introduction
Bayesian networks [1] are popular probabilistic graphical models often used in AI for both machine learning and the implementation of knowledge-based decision-support systems. Originally designed for discrete variable only [2], Bayesian networks have been extended to continuous variables [3]. Yet, when coping with knowledge-based models, continuous variables are often discretized in order to simplify the elicitation of the probabilities from the experts [4]. Even in machine learning tasks based on Bayesian networks, discretization can provide more accurate inferences [5]. This might be an issue if the Bayesian network embeds any kind of deterministic knowledge in the form of equations over some of its variables. The typical approach consists in discretizing the variables involved in the equation and providing a degenerate (i.e., one for a state, zero for all the other ones) quantification of the conditional probability mass functions. This reflects the fact that a specification of the input in the equation deterministically produces a single output with no uncertainty. Yet, this does not take into account the *epistemic* uncertainty induced by the discretization [6]. To do that, many authors advocate the need of non-classical approaches to uncertainty modeling such as fuzzy systems [7], evidence theory [6], and imprecise probabilities [8]. The goal of this paper is to discuss these ideas in the framework of Bayesian networks and adopt an imprecise-probability model in the form of a set of distributions, or *credal set*, to achieve that. We open the discussion with a motivating example.

2. A Motivating Example

The *body mass index* (BMI) I of a person whose weight (in Kg) is W and height (in m) is H is defined as $I := W/H^2$. The two tables here below depict, respectively, eight canonical categories based on BMI ranges (left), and the categories for different height/weight combinations assuming measurements on ranges of 5 Kg for W and .05 m for H (right). Such discretization cannot distinguish between a person with (W, H) equal to $(89, 1.71)$ and another with $(86, 1.74)$. Yet, the first is *overweight* (BMI $\simeq$ 28.4) and the second *moderately obese* (BMI $\simeq$ 30.4). This simple example shows that the epistemic uncertainty[1] induced by the discretization of the input variables W and H propagates through the deterministic relation defining BMI and, as in this case, might lead to a non-unique identification of the state of the output variable.

	Category	BMI	W \ H	1.55-1.60	1.60-1.65	1.65-1.70	1.70-1.75	1.75-1.80	1.80-1.85	1.85-1.90	1.90-1.95	1.95-2.00
			55-60	4	4	4	3-4	3-4	3-4	2-3	1-3	1-2
1	V. severely underweight	$\leq$15	60-65	4-5	4-5	4	4	4	3-4	3-4	2-3	2-3
2	Severely underweight	15-16	65-70	5	4-5	4-5	4	4	4	3-4	3-4	3
3	Underweight	16-18.5	70-75	5-6	5	4-5	4-5	4	4	4	3-4	3-4
4	Normal	18.5-25	75-80	5-6	5-6	5	4-5	4-5	4	4	4	4
5	Overweight	25-30	80-85	6-7	5-6	5-6	5	4-5	4-5	4	4	4
6	Moderately obese	30-35	85-90	6-7	6-7	5-6	5-6	5	4-5	4-5	4	4
7	Severely obese	35-40	90-95	7	6-7	6	5-6	5-6	5	4-5	4-5	4
8	Very severely obese	$\geq$40	95-100	7	6-7	6-7	6	5-6	5-6	5	4-5	4-5
			100-105	7	7	6-7	6-7	6	5-6	5-6	5	5
			105-110	7	7	7	6-7	6-7	6	5-6	5-6	5
			110-115	7	7	7	7	6-7	6-7	6	5-6	5-6

3. Modeling Epistemic Uncertainty by Credal Sets

Let us discuss the effects of the discretization more formally. Given a continuous, real-valued, variable X' defined on an interval $[\underline{x}', \overline{x}']$, call *discretization* an array of real numbers $\{x'_{(i)}\}_{i=0}^{d}$ such that $x'_{(0)} = \underline{x}'$, $x'_{(d)} = \overline{x}'$ and $x'_{(i)} < x'_{(j)}$ for each $i < j$, with $i, j = 0, \ldots, d$. Given a discretization, we obtain from X' a discrete variable X with d states, say $(x_1, \ldots, x_d)$. Knowing the actual state of X', say $X' = x'$, induces complete knowledge about X, i.e., $X = x_i$ with i such that $x' \in [x'_{(i-1)}, x'_{(i)}]$. Vice versa, knowing that $X = x_i$ only allows to conclude that x' certainly belongs to $[x'_{(i-1)}, x'_{(i)}]$. By construction, we have no reason to assume that a particular value in this interval should give the actual value of X. Yet, modeling such a condition by a probability density over the interval, would be equally unmotivated. The common approach of modeling such a condition of *ignorance* by means of of a uniform distribution, which is instead describing a condition of indifference among the various options, has been often criticized in the literature [9]. If this is the case, it basically means that classical probability theory is not able to describe such a condition of (strong) epistemic uncertainty, which requires instead a set of distributions or, in the imprecise-probability jargon, a *credal set* [10].

---

[1] The term *epistemic* uncertainty is used in the literature to contrast that of *aleatory* uncertainty. The first refers to the uncertainty affecting the model of the process, while the latter refers to the intrinsic randomness of a process. E.g., we decide to measure weight and height with a tolerance, which in principle can be reduced.

4. From Bayesian to Credal Networks

Let us first review some basic concept about Bayesian networks. Consider a set of discrete variables $\boldsymbol{X} := (X_1, \ldots, X_n)$. In a Bayesian network the variables in $\boldsymbol{X}$ are assumed in one-to-one correspondence with the nodes of a directed acyclic graph $\mathcal{G}$. The graph depicts conditional independence relations among variables in $\boldsymbol{X}$ by means of the *Markov condition*, i.e., every variable is conditionally independent of its non-descendants non-parents given its parents. This induces a factorization in the joint mass function:

$$P(\boldsymbol{x}) = \prod_{i=1}^{n} P(x_i|\mathrm{pa}_i),\qquad(1)$$

for each value $\boldsymbol{x}$ of $\boldsymbol{X}$, where the values of x_i and pa_i are those consistent with x_i and Pa_i is the joint variable denoting the predecessors of X_i according to $\mathcal{G}$ for each $i = 1, \ldots, n$. In other words, a Bayesian network allows for a compact specification of a joint mass function by means of a collection of conditional probability tables $\{P(X_i|\mathrm{Pa}_i)\}_{i=1}^{n}$. As an example, the graph here below depicts a conditional independence between the BMI and the gender given weight and height, which induces the decomposition of the joint mass function indicated under the graph.
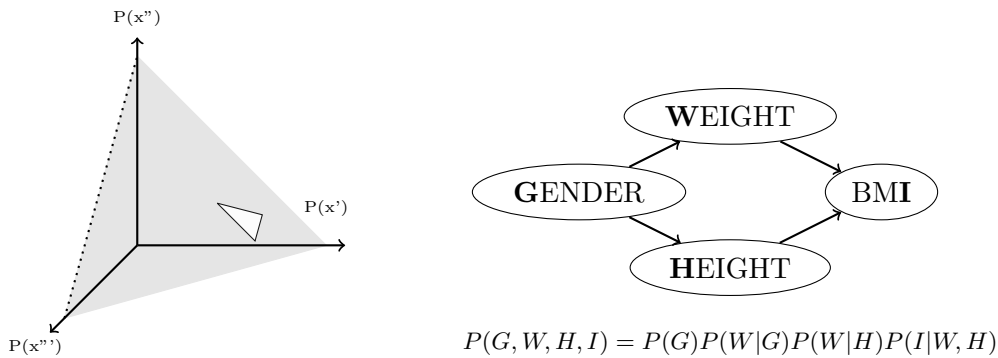

$$P(G, W, H, I) = P(G)P(W|G)P(W|H)P(I|W, H)$$

Figure 1. Credal sets for a ternary variable (left) and a directed acyclic graph (right).

A *credal set* $K(X)$ over X is defined as a convex set of probability mass functions over X. This represents a generalization of classical uncertainty models corresponding to a (single) probability mass function $P(X)$. The vacuous credal set is the largest credal set we can define, and it includes all the probability mass functions over X. This is commonly regarded as a model of *ignorance* about the state of X. Most informative credal sets are obtained by adding constraints to the vacuous credal set . A special class of credal sets is associated to *events*. Let Ω_X denote the set of possible values of X. Given event $A \subseteq \Omega_X$, the set of all the probability mass functions assigning probability one to A and zero to its complement is denoted as $K_A(X)$. The above figure (left) depicts a geometrical view of the vacuous credal set for a ternary variable (gray triangle), together with a more informative model (white triangle), and a credal set associated to event $A = \{x'', x'''\}$ (dotted segment).

Credal networks [11] are an extension of Bayesian networks based on credal sets. In practice, each conditional probability mass function $P(X_i|\mathrm{pa}_i)$ is replaced by a credal set $K(X_i|\mathrm{pa}_i)$ A credal network defines a joint mass function $K(\boldsymbol{X})$ whose elements are joint associated to Bayesian networks as in Eq. (1), with each conditional probability mass function $P(X_i|\mathrm{pa}_i)$ taking values in $K(X_i|\mathrm{pa}_i)$. The typical inference task in Bayesian networks is the computation of the conditional probability for a variable of interest given some observation of the other variables, e.g., $P(I|G)$ in the example. For credal networks, the analogous problem consists in the computation of the lower and upper bounds, e.g., $[\underline{P}(I|G), \overline{P}(I|G)]$, with respect to any Bayesian network specification consistent with the credal network.

5. Credal Modeling of Structural Equations

Let Y and $\boldsymbol{X}$ be some of the variables of a Bayesian network. Assume that $\boldsymbol{X}$ corresponds to the predecessors (or *parents*) of Y according to the graph of the network. Moreover, assume that both Y and $\boldsymbol{X}$ are obtained by discretization of the corresponding continuous variables, and these variables obey a deterministic equation $y' = f(\boldsymbol{x}')$. Following the discussion in the previous sections, for the quantification of $P(Y|\boldsymbol{x})$, we cannot take a representative point $\boldsymbol{x}'$ consistent with $\boldsymbol{x}$ and specify a degenerate mass function which gives probability one to the state y corresponding to $y' = f(\boldsymbol{x}')$. Our approach requires instead the evaluation of the extreme values returned by the function when the input variables are varying in the intervals associated to the discrete state, i.e.,

$$\underline{y}' := \min_{\boldsymbol{x}' \in \times_{i=1}^{n} [\underline{x}_i', \overline{x}_i']} f(\boldsymbol{x}') \, , \tag{2}$$

and analogously for the upper bound $\overline{y}$. Such optimization is trivial for the BMI example as the function can be easily differentiated and its restrictions are monotone. Generally speaking such a (linearly) constrained optimization can be achieved by gradient-descent techniques if information about the derivatives of f is available or by derivative-free technques such as quadratic approximation if this is not the case. Once the interval of $[\underline{y}', \overline{y}']$ has been computed, its overlap with the discretization intervals should be detected. In our procedure we therefore replace the Bayesian quantification of $P(Y|\boldsymbol{x})$ with a credal set $K(Y|\boldsymbol{x})$ assigning zero probability to the states of Y with empty overlap with $[\underline{y}', \overline{y}']$ being vacuous otherwise. If no constraints are imposed for the discretization of Y it is therefore possible to adopt a fine-grained discretization which does not need vacuous specification. If this requires too many states for Y, the task of keeping the number of credal specifications as low as possible can be easily reduced to an integer linear programming task.

6. Conclusions

A novel approach to the embedding of deterministic equation in discrete Bayesian networks has been proposed. This allows for a robust modeling of the epistemic uncertainty induced by the discretization. As a future work we intend to develop *ad hoc* algorithms for optimal discretization in this setup.

References

[1] D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques.* MIT Press, 2009.

[2] J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference.* Morgan Kaufmann, San Mateo, California, 1988.

[3] S. L. Lauritzen and F. Jensen. Stable local computation with conditional Gaussian distributions. *Statistics and Computing*, 11(2):191–203, 2001.

[4] H. Wang and M. J. Druzdzel. User interface tools for navigation in conditional probability tables and elicitation of probabilities in Bayesian networks. In *Proceedings of the Sixteenth conference on Uncertainty in Artificial Intelligence*, pages 617–625. Morgan Kaufmann Publishers Inc., 2000.

[5] N. Friedman and M. Goldszmidt. Discretizing continuous attributes while learning Bayesian networks. In *ICML*, pages 157–165, 1996.

[6] F. Tonon. Using random set theory to propagate epistemic uncertainty through a mechanical system. *Reliability Engineering & System Safety*, 85(1-3):169–181, 2004.

[7] D. Dubois and H. Prade. Gradualness, uncertainty and bipolarity: making sense of fuzzy sets. *Fuzzy Sets and Systems*, 192:3–24, 2012.

[8] M. Beer, S. Ferson, and V. Kreinovich. Imprecise probabilities in engineering analyses. *Mechanical systems and signal processing*, 37(1-2):4–29, 2013.

[9] P. Walley. Inferences from multinomial data: learning about a bag of marbles. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 3–57, 1996.

[10] P. Walley. *Statistical Reasoning with Imprecise Probabilities.* Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1991.

[11] F. G. Cozman. Credal networks. *Artificial intelligence*, 120(2):199–233, 2000.