Variational Bayes inference for logic-based probabilistic models on BDDs

Masakazu Ishihata, Yoshitaka Kameya, and Taisuke Sato

Tokyo institute of Technology
2-12-1 Ookayama, Meguro-ku Tokyo, Japan
{ishihata,kameya,sato}@mi.cs.titech.ac.jp

Abstract. Statistical abduction is an attempt to define a probability distribution over explanations derived by abduction and to evaluate them using their probabilities. In statistical abduction, deterministic knowledge like rules and facts are described as logic formulas. However non-deterministic knowledge like preference and frequency seems difficult to represent as logic. Bayesian inference can reflect such knowledge on a *prior* and *variational Bayes* (VB) is known as an approximation method for it. In this paper, we propose VB for logic-based probabilistic models and show that our proposed method is efficient in evaluating abductive explanations about failure in a logic circuit and a metabolic pathway.

1 Introduction

Abduction is one of logical inference to find explanations E from knowledge-base KB for an observation O such that $KB \wedge E \vdash O$ and $KB \wedge E$ is consistent. *Statistical abduction* defines a probability distribution over explanations and attempts to evaluate them by their probabilities. A couple of frameworks for statistical abduction have been proposed [1, 2] but they have restrictions on KB to realize efficient probability computation and learning. To relax these restrictions, the *BO-EM algorithm* which is an EM algorithm for *propositional logic-based probabilistic models* (PBPMs) was proposed [3]. A PBPM $p(b|\theta)$ is a joint distribution over boolean random variables b and defines probabilities for any boolean formulas. Probabilistic events (observations O) in a PBPM $p(b|\theta)$ are described as boolean formulas (explanations E) and a parameter θ can be learned by BO-EM in a dynamic programming manner on a BDD representing E. In statistical abduction, *deterministic knowledge* like rules and facts is described by logic as KB. However, *non-deterministic knowledge* like preference and frequency is difficult to represent by logic. To explicitly reflect such knowledge, *Bayesian inference* which is a method for statistical inference is useful. It assumes θ as a random variable and introduces a *prior* $p(\theta|\alpha)$ corresponding to preference or frequency of θ. Then, it computes a *posterior* $p(\theta|O, \alpha)$ which is a modified distribution by observations O. However the computation of the posterior requires complex summation and integration. *Variational Bayes* (VB) inference is an deterministic approximation of Bayesian inference and the VB-EM algorithm [4] is known as an EM like iterative computation for VB.

In this paper, we propose a VB-EM algorithm for PBPMs which is a generalization of the BO-EM algorithm for VB. We show that the complexity of our method is the same as BO-EM and also show that it runs efficiently in evaluating abductive explanations for a logic circuit and a metabolic pathway.

2 Preliminaries

2.1 Propositional logic based probabilistic models (PBPMs)

Let $\theta_j \equiv \{\theta_{jv}\}_{v=1}^{M_j}$ $(0 \leq \theta_{jv} \leq 1, \sum_{v=1}^{M_j} \theta_{jv} = 1)$ be a parameter of a *categorical distribution* $\mathrm{Cat}(\theta_j)$ corresponding to an M_j-sided dice. Also let $x_i \equiv \{x_{iv}\}_{v=1}^{N_i}$ $(x_{iv} \in \{0,1\}, \sum_{v=1}^{N_i} x_{iv} = 1)$ be a 1-of-N_i expression of a value drawn from $\mathrm{Cat}(\theta_{j_i})$ $(1 \leq j_i \leq M)$. We use v_i to denote v such that $x_{iv} = 1$. Then, the probability of x_i is computed by $p(x_i \mid \theta_{j_i}) = \prod_{v=1}^{N_i} \theta_{j_i v}^{x_{iv}} = \theta_{j_i v_i}$. Let x and θ be $\{x_i\}_{i=1}^{N}$ and $\{\theta_j\}_{j=1}^{M}$, respectively. Then, the probability of x is computed by

$$p(x \mid \theta) = \prod_{j=1}^{M} \prod_{v=1}^{M_j} \theta_{jv}^{\sigma_{jv}(x)}, \qquad \sigma_{jv}(x) \equiv \sum_{i:j_i=j} x_{iv}.$$

Let f be a function of x. Then, a boolean function $f_y(x) \equiv$ "$f(x) = y$" corresponds to a probabilistic event and its probability is computed as $p(f_y \mid \theta) = \sum_x f_y(x) p(x \mid \theta)$. We use "$x_i = v$" to denote $x_{iv} = 1$ and $x_{iv'} = 0$ $(v' \neq v)$. Then, f_y can be represented as a boolean formula as follows:

$$f_y = \bigvee_{x:f(x)=y} \bigwedge_{x_i \in x} \text{“}x_i = v_i\text{”}.$$

Probabilistic events "$x_i = v$" and "$x_i = v'$" $(v \neq v')$ depend on each other. However, they can be described as a boolean formula of independent boolean random variables $b \equiv \{b_{iv} \mid b_{iv} \equiv \text{“}x_i \leq v \mid x_i \geq v\text{”}, 1 \leq i \leq N, 1 \leq v < N_i\}$ as follows [3]:

$$\text{“}x_i = v\text{”} \equiv \begin{cases} b_{iv} \wedge \bigwedge_{v'=1}^{v-1} \neg b_{iv'} & 1 \leq v < N_i \\ \bigwedge_{v'=1}^{v-1} \neg b_{iv'} & v = N_i \end{cases},$$

where the probability of b_{iv} is defined as follows:

$$p(b_{iv}|\theta) \equiv \frac{\theta_{j_i v}}{\phi_{j_i v}}, \qquad p(\neg b_{iv}|\theta) \equiv \frac{\phi_{j_i, v+1}}{\phi_{j_i v}}, \qquad \phi_{jv} \equiv \sum_{v'=v}^{M_j} \theta_{jv}.$$

Then, the probability of "$x_i = v$" $(1 \leq v < N_i)$ is computed by

$$\begin{aligned} p(\text{“}x_i = v\text{”} \mid \theta) &= p(b_{iv} \mid \theta) \prod_{v'=1}^{j-1} p(\neg b_{iv'} \mid \theta) \\ &= \frac{\theta_{j_i v}}{\phi_{j_i v}} \prod_{v'=1}^{v-1} \frac{\phi_{j_i, v'+1}}{\phi_{j_i v'}} \\ &= \theta_{j_i v} \quad (= p(x_i = v \mid \theta)). \end{aligned}$$

Consequently, $p(x \mid \theta)$ can be computed by $p(b \mid \theta)$. The propositionalized distribution $p(b \mid \theta)$ is called a *propositional logic-based probabilistic model* (PBPM) for $p(x \mid \theta)$ and defines probabilities for any boolean formulas in b.

2.2 Maximum likelihood estimation and the EM algorithm

We assume that a probabilistic event f_y is sampled from $p(f_y, x \mid \theta)$ and that x and θ are unobservable. Maximum likelihood estimation (MLE) estimates the parameter θ as $\hat{\theta} \equiv \mathrm{argmax}_\theta\, p(f_y \mid \theta)$. When there are hidden variables like x, it is popular to use the EM algorithm for MLE. It iterates the following update:

$$\theta_{jv}^{(t+1)} = \frac{\mathrm{E}[\sigma_{jv}(x)]_{p(x|f_y,\theta^{(t)})}}{\sum_{v'=1}^{M_j} \mathrm{E}[\sigma_{jv'}(x)]_{p(x|f_y,\theta^{(t)})}}, \tag{1}$$

but it only converges to a local maximum of $p(f_y \mid \theta)$. The above expectations $\mathrm{E}[\sigma_{iv}(x)]_{p(x|f_y,\theta^{(t)})}$ can be computed using the PBPM $p(f_y, b \mid \theta)$. The BO-EM algorithm which is an EM algorithm for PBPMs computes them on a BDD representing f_y in time linear in the BDD size [3].

2.3 Variational Bayes inference and the VB-EM algorithm

Whereas MLE considers a parameter θ as an unknown constant and estimates it as what maximizes $p(f_y \mid \theta)$, Bayesian inference considers θ as a random variable and computes a distribution of θ given f_y. A prior distribution $p(\theta \mid \alpha)$ is given beforehand and a posterior distribution $p(\theta \mid f_y, \alpha)$ is computed by

$$p(\theta \mid f_y, \alpha) = p(f_y \mid \theta) p(\theta \mid \alpha) / p(f_y \mid \alpha), \quad p(f_y \mid \alpha) = \sum_x \int p(f_y, x \mid \theta) p(\theta \mid \alpha) d\theta.$$

While the number of possible x is too huge, it is difficult to compute the above posterior. The variational Bayes (VB) inference is known as a deterministic approximation of Bayesian inference. It approximates a joint posterior $p(x, \theta \mid f_y, \alpha)$ by a *variational distribution* $q(x, \theta) = q(x)q(\theta)$. Using $q(x)$ and $q(\theta)$, the log marginal likelihood $\ln p(f_y \mid \alpha)$ can be obtained as a sum of a *variational free energy* (VFE) $\mathrm{F}[q]$ and the *Kullback-Leibler* (KL) divergence $\mathrm{KL}(q||p)$ as follows:

$$\begin{aligned}
\ln p(f_y \mid \alpha) &= \sum_x \int q(x)q(\theta) \ln p(f_y \mid \alpha) d\theta \\
&= \sum_x \int q(x)q(\theta) \ln \frac{p(x, f_y, \theta \mid \alpha) q(x) q(\theta)}{p(x, \theta \mid f_y, \alpha) q(x) q(\theta)} d\theta \\
&= \mathrm{F}[q] + \mathrm{KL}(q||p)\,,
\end{aligned}$$

where

$$\mathrm{F}[q] = \sum_x \int q(x)q(\theta) \ln \frac{p(x, f_y, \theta|\alpha)}{q(x)q(\theta)} d\theta, \ \mathrm{KL}(q||p) = \sum_x \int q(x)q(\theta) \ln \frac{q(x)q(\theta)}{p(x, \theta|f_y, \alpha)} d\theta.$$

Since KL($q||p$) is non-negative, F[q] gives a lower bound of $\ln p(f_y \mid \alpha)$. Maximizing F[q] with respect to q results in minimizing KL($q||p$). The variational Bayes EM (VB-EM) algorithm is an EM like iterative algorithm for VB inference [4]. It attempts to maximize F[q] by iterating the following updates:

$$q^{(t+1)}(x) \propto \exp\bigl(\mathrm{E}[\ln p(x, f_y \mid \theta)]_{q^{(t)}(\theta)}\bigr)\,, \tag{2}$$

$$q^{(t+1)}(\theta) \propto p(\theta \mid \alpha)\exp\bigl(\mathrm{E}[\ln p(x, f_y \mid \theta)]_{q^{(t+1)}(x)}\bigr)\,. \tag{3}$$

The above updates are repeated until F[q] converges. When converged, the VB-EM algorithm outputs $q(\theta)$ as an approximation of the posterior $p(\theta \mid f_y, \alpha)$.

3 Proposed method

3.1 Bayesian inference for PBPMs

Let $\alpha_k \equiv \{\alpha_{kv}\}_{v=1}^{L_k}$ ($\alpha_{kv} > 0$) be a parameter of a *Dirichlet distribution* $\mathrm{Dir}\,(\alpha_k)$. We assume θ_j is sampled from $\mathrm{Dir}\,\bigl(\alpha_{k_j}\bigr)$. Then, the prior distribution $p(\theta \mid \alpha)$ ($\alpha \equiv \{\alpha_k\}_{k=1}^{L}$) is represented as a product of Dirichlet distributions:

$$p(\theta|\alpha) = \prod_{j=1}^{M} p(\theta_j|\alpha_{k_j}), \quad p(\theta_j|\alpha_k) = \frac{1}{\mathrm{Z}\,(\alpha_k)} \prod_{v=1}^{M_j} \theta_{jv}^{\alpha_{kv}-1}, \quad \mathrm{Z}\,(\alpha_k) \equiv \frac{\prod_{v=1}^{L_k} \Gamma(\alpha_{kv})}{\Gamma\Bigl(\sum_{v=1}^{L_k} \alpha_{kv}\Bigr)}.$$

Because a Dirichlet distribution is a conjugate prior of a categorical distribution, the posterior $p(\theta \mid f_y, \alpha)$ becomes a sum of products of Dirichlet distributions:

$$p(\theta \mid f_y, \alpha) \propto \sum_x f_y(x) \prod_{j=1}^{M} \frac{1}{\mathrm{Z}\,\bigl(\alpha_{k_j}\bigr)} \prod_{v=1}^{M_j} \theta_{jv}^{\alpha_{k_j v}+\sigma_{jv}(x)-1}.$$

3.2 VB-EM algorithm for PBPMs

We propose a VB-EM algorithm for PBPMs. By substituting the definitions of $p(x \mid \theta)$ and $p(\theta \mid \alpha)$ into (2) and (3), we get the following updates:

$$q^{(t+1)}(x) \propto \prod_{i=1}^{N} \prod_{v=1}^{N_i} \Bigl(\tilde{\theta}_{iv}^{(t+1)}\Bigr)^{\sigma_{iv}(x)}, \quad q^{(t+1)}(\theta) \propto \prod_{j=1}^{M} \frac{1}{\mathrm{Z}\,\Bigl(\tilde{\alpha}_j^{(t+1)}\Bigr)} \prod_{v=1}^{M_j} \theta_{jv}^{\tilde{\alpha}_{jv}^{(t+1)}-1},$$

where $\tilde{\theta}_{iv}^{(t+1)}$ and $\tilde{\alpha}_{kv}^{(t+1)}$ are defined as follows:

$$\tilde{\theta}_{iv}^{(t+1)} \equiv \exp\left(\Psi\Bigl(\tilde{\alpha}_{k_j v}^{(t)}\Bigr) - \Psi\left(\sum_{v'=1}^{M_i} \tilde{\alpha}_{k_j v'}^{(t)}\right)\right), \quad \tilde{\alpha}_{jv}^{(t+1)} \equiv \alpha_{k_j v} + \mathrm{E}[\sigma_{jv}(x)]_{q^{(t+1)}(x)},$$

where $\Psi(x)$ is the *digamma function* defined by $\Psi(x) \equiv \frac{d}{dx} \ln \Gamma(x)$. The point here is that the above expectation $\mathrm{E}[\sigma_{ij}(x)]_{q^{(k+1)}(x)}$ can be computed by substituting

$\tilde{\theta}_{iv}^{(t+1)}$ into $\theta_{iv}^{(t)}$ in the computation of $\mathrm{E}[\sigma_{jv}(x)]_{p(x|f_y,\theta^{(t)})}$ in (1). As mentioned in Section 2, the BO-EM algorithm [3] computes $\mathrm{E}[\sigma_{jv}(x)]_{p(x|f_y,\theta^{(t)})}$ by using $p(f_y, b \mid \theta)$ which is a PBPM for $p(f_y, x \mid \theta)$ and a BDD which represents f_y in linear time to the BDD size. Consequently, the VB-EM algorithm for PBPMs can also be executed on the BDD in the same time and space complexity as the BO-EM algorithm. Due to space limitations, the detail of our proposed algorithm is omitted.

4 Experiments

4.1 Artificial problem: diagnosis for failure in a logic circuit

We apply our method to a diagnosis for a 3-bit adder circuit involving error gates [1, 3]. An error gate is *stochastically* stuck at 0 or 1. The task is to find error gates in the circuit from observations that pairs of input and output values. The previous approach [3] learns probabilities of gates being stuck by the BO-EM algorithm and predicts where error gates are by using their probabilities.

In this experiment, we assume the average rate of a gate being normal, stuck at 0 and stuck at 1 is given as non-deterministic knowledge. To reflect the knowledge in prediction, we use our proposed method and obtain gate probabilities. We compare the prediction accuracy of our approach with that of the previous approach. The left side of Fig. 1 respectively shows precision, recall and F-measure of the previous approach and the right shows ours. These quantities are computed by predicting error gates in 10,000 randomly-generated 3-bit adder circuits while changing the number of observations $N = 10, 50, 100$. The result shows that our approach achieves higher F-measure value than the previous one and also shows that introducing non-deterministic knowledge is efficient in prediction.

4.2 Real problem: hypotheses finding in a metabolic pathway

Inoue et al. applied statistical abduction to find and rank explanations for a metabolic pathway data [3]. According to them there are two important reactions which tend to be inhibited and the knowledge is used to evaluate their ranking result. To explicitly reflect such knowledge in ranking, we apply our proposed method to ranking 66 explanations derived by them. Fig. 2 shows that explanations which have top 20 high probabilities. 22 out of 66 explanations are considered *good* because they include the above mentioned two important reactions as inhibited. The result shows that 16 out of top 20 explanations are good and also that our method can explicitly reflect non-deterministic knowledge in ranking explanations.

5 Conclusions and Related work

We propose a variational Bayes inference for PBPMs. In the context of Bayesian inference for statistical abduction, deterministic knowledge is described

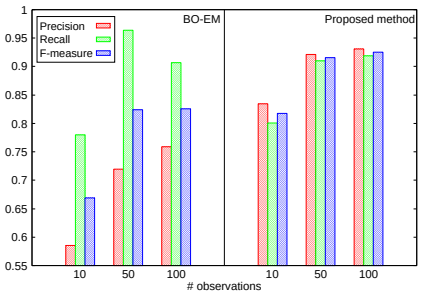

Fig. 1. Predicting result

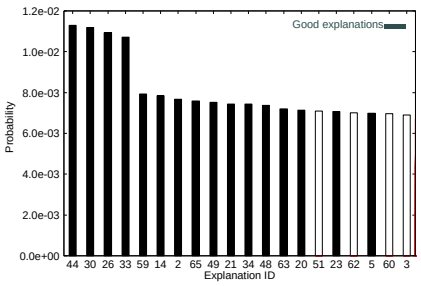

Fig. 2. Ranking result

as logic formulas whereas non-deterministic knowledge is represented as a prior distribution. We apply our proposed method to a diagnosis for failure in a logic circuit and to evaluating explanations for a metabolic pathway data. The experimental results show that introducing non-deterministic knowledge as a prior makes both of prediction and ranking results better.

PRISM [2] is a probabilistic extension of Prolog and variational Bayes inference for it has already proposed [5]. However, PRISM has the *exclusiveness condition* for explanations to realize efficient probability computation. Our proposed method can eliminate the exclusiveness condition because it can deal with any boolean formulas.

ProbLog [6] which is another probabilistic extension of Prolog also employs a BDD-based parameter learning algorithm [7]. However, variational Bayesian inference for ProbLog has not yet been proposed to our knowledge.

References

1. Poole, D.: Probabilistic Horn abduction and Bayesian networks. Artificial Intelligence **64**(1) (1993) 81–129
2. Sato, T., Kameya, Y.: Parameter Learning of Logic Programs for Symbolic-statistical Modeling. Journal of Artificial Intelligence Research **15** (2001) 391–454
3. Ishihata, M., Kameya, Y., Sato, T., Minato, S.: An EM algorithm on BDDs with order encoding for logic-based probabilistic models. In: Proc. of ACML'10. (2010)
4. Beal, M., Ghahramani, Z.: The Variational Bayesian EM Algorithm for Incomplete Data: with Application to Scoring Graphical Model Structures. Bayesian Statistics **7** (2003)
5. Sato, T., Kameya, Y., Kurihara, K.: Variational Bayes via propositionalized probability computation in PRISM. Annals of Mathematics and Artificial Inteligence **54**(1-3) (2009) 135–158
6. De Raedt, L., Kimming, A., Toivonen, H.: ProbLog: A probabilistic Prolog and its application in link discovery. In: Proc. of IJCAI'07. (2007) 2468–2473
7. Gutmann, B., Thon, I., De Raedt, L.: Learning the parameters of probabilistic logic programs from interpretations, Department of Computer Science, K.U.Leuven (2010)