From ProbLog to First Order Logic: A First Exploration

Joost Vennekens, Angelika Kimmig, Theofrastos Mantadelis, Bernd Gutmann, Maurice Bruynooghe, Luc De Raedt FIRSTNAME.LASTNAME@CS.KULEUVEN.BE
Katholieke Universiteit Leuven, CS Department, Celestijnenlaan 200A, 3001 Heverlee, Belgium

Abstract

ProbLog is a system that allows a user to compute (or approximate) the probability of a query in a theory consisting of Horn clauses guarded by probabilistic facts. This paper explores how to to generalize the approach towards theories consisting of first order logic formulas guarded by probabilistic facts.

1. Introduction

A problem for languages that combine probability theory with *expressive* logical formulas, is that probabilistic and logical knowledge might interact in complex ways, leading to a semantics that is too complicated to understand, and inference/learning that is too slow to be of practical use. Initial research into Probabilistic Logic Learning therefore mainly focused on adding probabilities to restricted logical languages, such as definite clause logic (Kersting & De Raedt, 2008; Sato & Kameya, 1997). Since then, the trend has been to extend the expressivity of the logical language, e.g., to normal clauses (Poole, 2000). More recently, (Richardson & Domingos, 2006) allows full first-order logic, while still offering practically feasible inference and learning. However, this is done at the expense of some probabilistic clarity: MLNs have no straightforward correspondence between the weight of a formula and its probability; instead, the probability of a formula depends non-linearly on all weights in the theory.

In this paper, we present an attempt at constructing a language in which (1) general first-order formulas are allowed, (2) efficient forms of inference are possible, and (3) weights are really probabilities. Our approach will be an extension of that of (Poole, 2000; Sato & Kameya, 1997; De Raedt et al., 2007). The basic idea is to restrict all "communication" between the probabilistic and logical part of the theory to a particular, and very limited, interface. The hope is that this will allow us to rule out the complex interactions that make reasoning tasks for logics such as (Nilson, 1986) hard.

Presented at *ILP-MLG-SRL*, Leuven, Belgium, 2009.

2. Language and its semantics

A theory T in our language consists of a probabilistic part PF and a logical part Φ. The predicates of T are likewise split into a set of probabilistic predicates Σ_P and a set of logical ones Σ_L. The probabilistic part PF is made up of a number of *probabilistic facts* of the form $pf(\overline{x}) : \alpha$, with α a non-zero probability and $pf \in \Sigma_P$. We require that PF contains precisely one such probabilistic fact for each $pf \in \Sigma_P$. Their meaning is that each ground instance $pf(\overline{x})\theta$ of such a fact has a probability α of being true. Different instances (of the same or of different probabilistic facts) are probabilistically independent. We call an assignment of truth values to all ground instances of all probabilistic facts an *atomic choice*. The probability $prob(s)$ of such an atomic choice s is the product of all α_i for true instances $pf_i(\overline{x})\theta : \alpha_i$ and all $(1 - \alpha_j)$ for false instances $pf_j(\overline{x})\theta : \alpha_j$. Next, we get to the logical part Φ of the theory. This consists of a set of implications $\forall \overline{x}\ pf(\overline{x}) \rightarrow F(\overline{x})$. Here, $F(\overline{x})$ is a first-order formula, containing no probabilistic predicates, and $pf(\overline{x})$ is an atom with a probabilistic predicate.

Fix a domain D and consider the set $\mathcal{W}_D$ of all possible worlds in D, i.e., each I in $\mathcal{W}_D$ is an interpretation for the vocabulary of the theory with D as its domain. We define the semantics of our logic as follows.

Def. 1 *Let $T = (PF, \Phi)$ be a theory and D a domain. A distribution μ over $\mathcal{W}_D$ is a* model *of T, denoted $\mu \models T$, if and only if, for each atomic choice s, $\mu(s) = \sum_{I \models s} prob(s)$ and $\mu(I) = 0$ for all $I \not\models \Phi$.*

This definition is straightforward: a distribution μ is a model if it assigns the right probabilities to the atomic choices and is consistent with the standard FO semantics for the logical part of the theory. We illustrate with an example of (Richardson & Domingos, 2006), stating that friends of friends tend to be friends, friendless people tend to smoke, smoking tends to cause cancer, and friends tend to share smoking habits.

$$\forall x, y, z \; pf_1(x, y, z) \rightarrow (fr(x, y) \wedge fr(y, z) \rightarrow fr(x, z)))$$
$$\forall x : pf_2(x) \rightarrow \neg \exists y \; fr(x, y) \rightarrow sm(x)$$
$$\forall x : \; pf_3(x) \rightarrow (sm(x) \rightarrow Ca(x)).$$
$$\forall xy : pf_4(x, y) \rightarrow (fr(x, y) \rightarrow (sm(x) \leftrightarrow sm(y)))$$
$$pf_1(x, y, z) : 0.3 \; pf_2(x) : 0.9 \; pf_3(x, y) : 0.7 \; pf_4(x, y) : 0.5$$

Consistency. A first question is when a theory is consistent, in the sense of having at least one model μ.

Th. 1 *Let $T = (PF, \Phi)$ be a theory and $L(T)$ the set of formulas $\forall \overline{x} \; F(\overline{x})$ for which Φ contains $\forall \overline{x} \; pf(\overline{x}) \rightarrow F(\overline{x})$. There exist a distribution μ over $\mathcal{W}_D$ such that $\mu \models T$ if and only if $L(T)$ is consistent in D*[1].

For instance, $\Phi = \{pf_1 \rightarrow a; pf_2 \rightarrow \neg a\}$ violates this condition (in any D), and indeed, no theory (PF, Φ) has a model. Such an inconsistency can arise if a user makes a mistake, but for instance also if we are trying to merge theories elicted from different domain experts. As in *paraconsistent reasoning*, it might then be desirable to look at distributions that respect the theory *as best as possible*. To this end, we can normalize the original probability distribution *prob* to rule out inconstistencies. For an atomic choice s, let T_s be the logical theory obtained by adding to Φ all literals $\neg pf(\overline{c})$ for which $s(pf(\overline{c})) = \mathit{false}$ and all atoms $pf(\overline{c})$ for which $s(pf(\overline{c})) = \mathit{true}$. Let $Cons$ be the set of all s for which T_s is consistent in D. We define the normalized probability $\widehat{prob}(s)$ as $prob(s) / \sum_{s \in Cons} prob(s)$ for all $s \in Cons$ and as zero for all $s \notin Cons$. Replacing *prob* by $\widehat{prob}$ in Def. 1 provides a generalized notion of a model, which we denote by $\hat{\models}$. If $L(T)$ is consistent, then $\models$ and $\hat{\models}$ coincide. However, as long as the set of formulas that are asserted with probability 1 is consistent, generalized models always exist.

Meaning of probabilistic facts. As is evident from e.g. the smokers example, a probabilistic predicate might have no meaning of its own, and simply be a device for speaking about the probability of the formula that it implies. It might be more meaningful, then, to leave out this contrivance and refer directly to this probability. That is, instead of: $\{pf(\overline{x}) : \alpha, \forall \overline{x} \; pf(\overline{x}) \rightarrow \phi(\overline{x})\}$, we might want to write more simply: $\forall \overline{x} \; \mathbf{P}(\phi(\overline{x})) \geq \alpha$, with the obvious meaning that the probability of $\phi(\overline{x})$ has to be at least α for each $\overline{x}$. A theory of such statements, together with the assumption that all probabilities are independent, is *almost* the same as what we originally had. The only difference occurs when the formulas $\phi(\overline{x})$ are not logically independent in the following sense.

[1] This means that there exists a model with domain D.

Def. 2 *Let F be a set of formulas $\phi_i(\overline{x_i})$ and let F' be the set of all sentences that can be produced by filling in the free variables $\overline{x_i}$ of some $\phi_i(\overline{x_i}) \in F$. F is* logically independent *if there does not exist a subtheory $G \subseteq F'$ such that G implies either a formula $\psi \in F' \setminus G$ or the negation $\neg\psi$ of some such $\psi \in F' \setminus G$.*

Let us suppose, for instance, that the theory contains two formulas ϕ and ψ, such that $\phi \models \psi$. In this case, saying that $P(\phi)$ and $P(\psi)$ are independent implies the equality $P(\phi) = P(\phi \wedge \psi) = P(\phi)P(\psi)$, which means that $P(\psi) = 1$ or $P(\phi) = 0$. On the other hand, if we just have that $pf_1 \rightarrow \phi$ and $pf_2 \rightarrow \psi$, then saying that pf_1 and pf_2 are independent implies no such thing.

Th. 2 *Let T be a theory. If $L(T)$ is logically independent, then the models of T are precisely those probability distributions μ for which $\mu(\phi_i(\overline{x})) \geq \alpha_i$, for each i and $\overline{x}$, and $\mu(\phi_i(\overline{x})\theta)$ and $\mu(\phi_j(\overline{x})\theta')$ are independent, for all i, j and θ, θ' such that $i \neq j$ or $\theta \neq \theta'$.*

So, under this condition, theories in our language have a straightforward interpretation, in which the probabilistic predicates no longer play a role. If the condition is not satisfied, e.g. if the theory contains some $pf_i \rightarrow \phi_i$ and $pf_j \rightarrow \phi_j$ such that ϕ_i implies ϕ_j, then the pf_i do play a relevant role. One way of explaning their meaning in this case might be to say that, even though the formulas ϕ_i and ϕ_j themselves are not independent, the probabilistic facts pf_i and pf_j might represent *independent reasons* for believing them.

For instance, suppose that a sociologist explains to us how society currenltly imposes expectation patterns which lead to a predominantly male student population in engineering. We might then write: $\forall x \; Sociologist(x) \rightarrow (Engineering(x) \rightarrow Male(x))$. Here, the probabilistic fact $Sociologist(x)$ represents that x is a student to whom the sociologist's line of reasoning actually applies. If we believe that we have been talking to a good sociologist, then we should assign it a high probability. Now, suppose that we also run into a statistician who has been keeping track of enrolment into the engineering departement over the years and uses this to predict that a proportion of engineering students are male. We could add: $\forall x \; Statistician(x) \rightarrow (Engineering(x) \rightarrow Male(x))$. Now, $Statistician(x)$ means that whatever pattern was present in the statistician's data is applicable to x, and its probability should reflect how widespread we expect this pattern to be. Adding these formulas to our theory now provides an additional reason to believe that engineering students are male, and it seems not unreasonable to assume that these two reasons are independent, so that this additional reasons increases our belief in the statement.

3. Inference and implementation

We are now interested in deciding what the theory T allows us to conclude over the probability of some additional query formula Q. For each given domain, T has a non-empty set $\hat{\mathcal{M}}$ of generalized models μ. Each $\mu \in \hat{\mathcal{M}}$ assigns a particular probability $\mu(Q) = \sum_{I \models Q} \mu(I)$ to Q, yielding, in general, a non-empty probability interval $[\min_{\mu \in \hat{\mathcal{M}}} \mu(Q), \max_{\mu \in \hat{\mathcal{M}}} \mu(Q)]$. We are now interested in the inference task of determining this interval. Because $\max_{\mu \in \hat{\mathcal{M}}} \mu(Q)$ must be equal to $1 - \min_{\mu \in \hat{\mathcal{M}}} \mu(\neg Q)$, we can restrict attention to the task of computing only the lower bound. As the following theorem shows, we can compute this lower bound without having to consider any specific model μ of T.

Th. 3 *Let $\hat{\mathcal{M}}$ be the non-empty set of models of a consistent theory T. Then $\min_{\mu \in \hat{\mathcal{M}}} \mu(Q) = \sum_{T_s \models Q} \widehat{prob}(s)$, where $T_s \models Q$ means, as usual, that Q holds in all $I \in \mathcal{W}_D$ such that $I \models T_s$.*

To illustrate, let us consider again the example of the smokers. Suppose that we obtain additional information about some specific people, say Ann, $Mary$ and Bob, such that $sm(Ann)$, $fr(Ann, Bob)$ and $fr(Bob, Mary)$, and want to figure out the probability that $Mary$ also smokes. We do this by adding this new information to our theory (with probability 1) and computing the lowerbound for the probability that $sm(Mary)$ according to this new theory. There are two sets of atomic choices that imply $sm(Mary)$. The first consists of all those in which Mary is friends with Ann through their mutual friend Bob ($pf_1(Ann, Mary)$) and she smokes because of this friendship ($pf_4(Ann, Mary)$); the probability of this is 0.15. The second consists of those in which Bob smokes because of his friendship with Ann ($pf_4(Ann, Bob)$) and Mary smokes because of her friendship with Bob ($pf_4(Bob, Mary)$); this probability is 0.25. The probability of the union of these two sets is $0.15 + 0.25 - 0.15 \cdot 0.25 = 0.3625$, and this is therefore our lower bound on $\mu(sm(Mary))$.

We believe that an implementation strategy similar to that of ProbLog (De Raedt et al., 2007) is feasible. The idea is to enumerate proofs of the query, but whereas ProbLog can use an SLD theorem prover, we need to use a full first order theorem prover (e.g. PTTP (Stickel, 1988)). Proofs return the atomic choice made while proving and this atomic choice is passed, similar to ProbLog, to a BDD module that calculates the probability of the query. However, there is a caveat in case $L(T)$ is not consistent. In that case we are using generalized models. One need to check that the atomic choice made in the proof results in a consistent possible world. This can be done by a model checker. Also, the probability of the query need to be normalised. This can be done by enumeration the proofs of inconsistency in the theories (again using the PTTP prover).

4. Conclusions and Future Work

We presented a first step towards a language that uses full first-order logic, allows probabilities to be directly specified, and is still useful for inference. However, a lot of work still remains. On the level of semantics, the relation to e.g. (Nilson, 1986) needs to be investigated. On the implementation level, a first prototype need to be built. Finally, more work on applications is needed to validate the approach.

Acknowledgments. This work is supported by GOA/08/008 Probabilistic Logic Learning. AK and BG are supported by FWO-Vlaanderen. JV is a postdoctoral researcher of the FWO-Vlaanderen.

References

De Raedt, L., Kimmig, A., & Toivonen, H. (2007). ProbLog: A probabilistic Prolog and its application in link discovery. *IJCAI* (pp. 2462–2467).

Kersting, K., & De Raedt, L. (2008). Basic principles of learning bayesian logic programs. *Probabilistic Inductive Logic Programming* (pp. 189–221).

Nilson, N. J. (1986). Probabilistic logic. *AI*, *28*, 71–87.

Poole, D. (2000). Abducing through negation as failure: stable models within the independent choice logic. *Journal of logic programming*, *44*, 5–35.

Richardson, M., & Domingos, P. (2006). Markov logic networks. *Machine Learning*, *62*, 107–136.

Sato, T., & Kameya, Y. (1997). PRISM: A language for symbolic-statistical modeling. *IJCAI*.

Stickel, M. E. (1988). A prolog technology theorem prover: implementation by an extended prolog compiler. *Journal of Automated Reasoning*, *4*, 353–380.