Expressive Probability Models in Science

Stuart Russell

Computer Science Division, University of California, Berkeley, CA 94720, USA,
russell@cs.berkeley.edu,
WWW home page: http://www.cs.berkeley.edu/~russell

Abstract. The paper is a brief summary of an invited talk given at the Discovery Science conference. The principal points are as follows: first, that probability theory forms the basis for connecting hypotheses and data; second, that the expressive power of the probability models used in scientific theory formation has expanded significantly; and finally, that still further expansion is required to tackle many problems of interest. This further expansion should combine probability theory with the expressive power of first-order logical languages. The paper sketches an approximate inference method for representation systems of this kind.

1 Data, hypotheses, and probability

Classical philosophers of science have proposed a "deductive-nomological" approach in which observations are the logical consequence of hypotheses that explain them. In practice, of course, observations are subject to noise. In the simplest case, one attributes the noise to random perturbations within the measuring process. For example, the standard least-squares procedure for fitting linear models to data effectively assumes constant-variance Gaussian noise applied to each datum independently.

In more complex situations, uncertainty enters the hypotheses themselves. In Mendelian genetics, for instance, characters are inherited through a random process; experimental data quantify, but do not eliminate, the uncertainty in the process. Probability theory provides the mathematical basis relating data to hypotheses when uncertainty is present. Given a set of data D, a set of possible hypotheses $\mathcal{H}$, and a question X, the predicted answer according to the full Bayesian approach is

$$P(X|D) = \alpha \sum_{H \in \mathcal{H}} P(X|H)P(D|H)P(H)$$

where α is a normalizing constant, $P(X|H)$ is the prediction of each hypothesis H, $P(D|H)$ is the likelihood of the data given the hypothesis H and therefore incorporates the measurement process, and $P(H)$ is the prior probability of H. Because $\mathcal{H}$ may be large, it is not always possible to calculate exact Bayesian predictions. Recently, Markov chain Monte Carlo (MCMC) methods have shown great promise for approximate Bayesian calculations [3], and will be discussed further below. In other cases, a single *maximum a posteriori* (MAP) hypothesis can be selected by maximizing $P(D|H)P(H)$.

2 Expressive power

Probability models (hypotheses) typically specify probability distributions over "events." Roughly speaking, a model family $\mathcal{L}_2$ is at least as expressive as $\mathcal{L}_1$ iff, for every e_1 in $\mathcal{L}_1$, there is some e_2 in $\mathcal{L}_2$ such that e_1 and e_2 denote "the same" distribution. Crude expressive power in this sense is obviously important, but more important is *efficient* representation: we prefer desirable distributions to be represented by subfamilies with "fewer parameters," enabling faster learning.

The simplest probability model is the "atomic" distribution that specifies a probability explicitly for each event. Bayesian networks, generalized linear models, and, to some extent, neural networks provide compact descriptions of multivariate distributions. These models are all analogues of *propositional* logic representations.

Additional expressive power is needed to represent temporal processes: hidden Markov models (HMMs) are atomic temporal models, whereas Kalman filters and dynamic Bayesian networks (DBNs) are multivariate temporal models. The efficiency advantage of multivariate over atomic models is clearly seen with DBNs and HMMs: while the two families are equally expressive, a DBN representing a sparse process with n bits of state information requires $O(n)$ parameters whereas the equivalent HMM requires $O(2^n)$. This appears to result in greater statistical efficiency and hence better performance for DBNs in speech recognition [10]. DBNs also seem promising for biological modelling in areas such as oncogenesis [2] and genetic networks.

In areas where still more complex models are required, there has been less progress in creating general-purpose representation languages. Estimating the amount of dark matter in the galaxy by counting rare MACHO observations [9] required a detailed statistical model of matter distribution, observation regime, and instrument efficiency; no general tools are available that can handle this kind of aggregate modelling. Research on phylogenetic trees [1] and genetic linkage analysis [8] uses models with repeating parameter patterns and different tree structures for each data set; again, only special-purpose algorithms are used. What is missing from standard tools is the ability to handle distributions over attributes and relations of multiple objects—the province of first-order languages.

3 First-order languages

First-order probabilistic logics (FOPL) knowledge bases specify probability distributions over events that are models of the first-order language [5]. The ability to handle objects and relations gives such languages a huge advantage in representational efficiency in certain situations: for example, the rules of chess can be written in about one page of Prolog but require perhaps millions of pages in a propositional language. Recent work by Koller and Pfeffer [6] has succeeded in devising restricted sublanguages that make it relatively easy to specify complete distributions over events (see also [7]). Current inference methods, however, are somewhat restricted in their ability to handle uncertainty in the relational structure of the domain or in the number of objects it contains. It is possible that algorithms based on MCMC will yield more flexibility. The basic idea is to construct a Markov chain on the set of logical models of the FOPL theory. In each model, any query can be answered simply; the MCMC method samples models so that estimates of the query probability converge to the true posterior. For example, suppose we have the following simple knowledge base:

$$\begin{aligned}
&\forall x, y\ P(CoveredBy(x, y)) = 0.3 \\
&\forall x\ P(Safe(x) \mid \exists y\ CoveredBy(x, y)) = 0.8 \\
&\forall x\ P(Safe(x) \mid \neg\exists y\ CoveredBy(x, y)) = 0.5
\end{aligned}$$

and we want to know the probability that everyone who is safe is covered only by safe people, given certain data on objects a, b, c, d:

$$P(\forall x\, Safe(x) \Rightarrow (\forall y\, CoveredBy(x, y) \Rightarrow Safe(y)) \mid Safe(a), Safe(b), \neg Safe(d))$$

Figure 1 shows the results of running MCMC for this query.

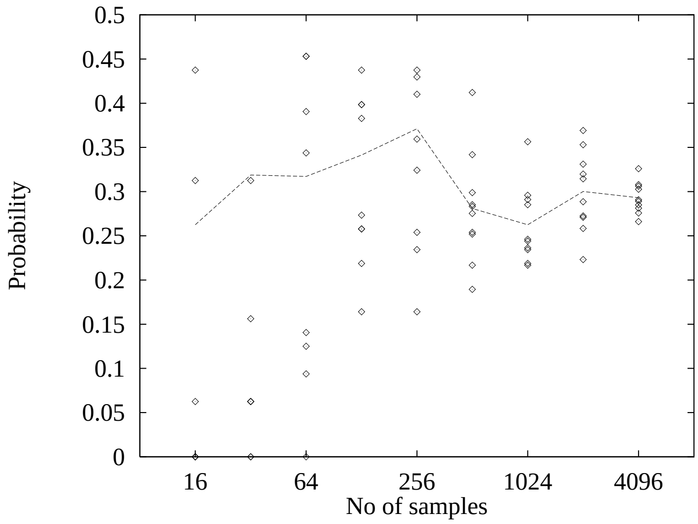

Fig. 1. Solution of a first-order query by MCMC, showing the average probability over 10 runs as a function of sample size (*solid line*)

Much work remains to be done, including convergence proofs for MCMC on infinite model spaces, analyses of the complexity of approximate inference for various sublanguages, and incremental query answering between models. The approach has much in common with the *general pattern theory* of Grenander [4], but should benefit from the vast research base dealing with representation, inference, and learning in first-order logic.

References

1. J. Felsenstein. Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. *Methods in Enzymology*, 266:418–427, 1996.
2. Nir Friedman, Kevin Murphy, and Stuart Russell. Learning the structure of dynamic probabilistic networks. In *Uncertainty in Artificial Intelligence: Proceedings of the Fourteenth Conference*, Madison, Wisconsin, July 1998. Morgan Kaufmann.
3. W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, editors. *Markov chain Monte Carlo in practice*. Chapman and Hall, London, 1996.
4. Ulf Grenander. *General pattern theory*. Oxford University Press, Oxford, 1993.
5. J. Y. Halpern. An analysis of first-order logics of probability. *Artificial Intelligence*, 46(3):311–350, 1990.
6. D. Koller and A. Pfeffer. Probabilistic frame-based systems. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI-98)*, Madison, Wisconsin, July 1998. AAAI Press.
7. S. Muggleton. Stochastic logic programs. Journal of Logic Programming, to appear, 1999.
8. M. Polymeropoulos, J. Higgins, L. Golbe, et al. Mapping of a gene for Parkinson's disease to chromosome 4q21-q23. *Science*, 274(5290):1197–1199, 1996.
9. W. Sutherland. Gravitational microlensing: A report on the MACHO project. *Reviews of Modern Physics*, 71:421–434, January 1999.
10. Geoff Zweig and Stuart J. Russell. Speech recognition with dynamic Bayesian networks. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI-98)*, Madison, Wisconsin, July 1998. AAAI Press.