Learning the Probability Distribution for Probabilistic Expert Systems

Michelle Baker and Kenneth Roberts
Department of Computer Science
Columbia University

No. CUCS-490-89

Abstract

This paper describes a method for learning the joint probability distribution of a set of variables from a sample of instances from the domain. The method is based on a straightforward application of Bayes Law to the problem of estimating individual probabilities from a probability distribution. We use a maximum entropy distribution as an initial estimate and show how this estimate can be easily updated each time an additional example is observed. Although developed for the purpose of estimating the conditional probabilities required for Bayesian inference networks, this method can be adopted to simplify knowledge acquisition in any expert system that uses knowledge in the form of probabilities.

1. Introduction

The acquisition of knowledge for expert systems operating in probabilistic domains is especially difficult because it is necessary to specify probabilities for most of the information included in the domain model. In the worst case it is necessary to specify and store the full joint probability distribution for the predicates that are used in a domain model. For example if there are 15 predicates in the domain model there are $2^{15}-1$ individual probabilities to deal with. One of the advantages of Bayesian inference networks is that they reduce the number of probabilities required by excluding all those that represented independent relationships among the predicates. Nevertheless, the difficulty of specifying subsets of the joint probability distribution remains a serious problem. Even when an expert is willing to attempt this task they are unlikely be able to do so with a high degree of accuracy.

Research that has addressed the problems of knowledge acquisition and storage of probabilities has assumed that knowledge of the distribution is fixed. There are two basic approaches in this work. One approach is to use functional forms to specify large segments of the joint distribution. As long as the functions are easy to compute probabilities need not be stored and can be computed as needed. Furthermore knowledge acquisition is simplified because an expert need only specify a single functional form to describe a potentially large set of probabilities. Along these lines, Cheeseman (Cheeseman, 1983) has shown how functions to compute maximum entropy probabilities that take arbitrary constraints on the form of a distribution into account can be derived. Cooper (Cooper, 1988) has described a class of functions called *prototypical probability functions* which have been used in systems built on Bayesian inference networks (Pearl, 1986, Pearl, 1987, Geffner and Pearl, 1987, Cooper, 1984). The second approach is to use decision theory to limit the knowledge that is acquired (Heckerman and Jimison, 1987, Horvitz, 1987). In these methods the domain model is only specified up to the point to which it is determined to be worth the effort. Large segments of the probability distribution may then be ignored.

In this paper we describe a Bayesian method for learning the joint probability distribution from a set of instances. There are two features of our work that set it apart from other research into the acquisition of probabilistic knowledge. First, we use actual instances from the domain to estimate a probability distribution. Secondly, we do not assume that knowledge of the probabilities is fixed. This method has the advantage that estimates of the probabilities can be easily updated as new instances are observed. Thus, a system using this method will become more accurate as its experience solving problems grows.

2. Derivation of the Updating Formula

This work is a straightforward application of Bayes Law,

$$f(\theta|y) = \frac{h(\theta)g(y|\theta)}{\int_a^b h(\theta)g(y|\theta)d\theta}$$

where $f(\theta|y)$ is the posterior density function of θ given that y is known and $h(\theta)$ is the prior density function of θ.

In our application of Bayes formula we take θ_i to be the probability of the i^{th} vector, where each vector represents a point in the Cartesian product of a set of

nominal valued variables, V_i, i.e. $V_1 \times V_2 \times \cdots V_n$. In otherwords, if the cardinality of $V_1 \times V_2 \times \cdots V_n$ is m then there are m possible vector outcomes. Denote the i^{th} outcome, $(v_1, v_2, \ldots v_n)$, by A_i. From an alternative perspective, imagine that all the possible instantiations of a frame in a simple frame system[1] have been listed in some order. Then there are m frame instantiations in the list, A_i is the i^{th} frame, and θ_i is the probability of A_i.

Given that we have $n=n_1+n_2+\cdots n_m$ observations such that there have been n_i observations of A_i.

Then the problem is to find an estimate, $\overline{\theta}'$, of $\overline{\theta} = (\theta_1, \theta_2, \cdots \theta_m)$ such that $\sum_{i=1}^{m} \theta_i' = 1$.

Restating Bayes Law as it applies to our problem, we have the following formula for the posterior density function of $\overline{\theta}$ given a sample of instances represented by $(n_1, n_2, \ldots n_n)$.

$$f(\overline{\theta} | n_1, n_2, \ldots n_m) = \frac{h(\overline{\theta}) g(n_1, n_2, \ldots n_m | \overline{\theta})}{\int_0^1 \int_0^{1-\theta_1} \cdots \int_0^{1-(\theta_1+\theta_2+\cdots\theta_{m-2})} h(\overline{\theta}) g(n_1, n_2, \ldots n_m | \overline{\theta}) d\theta_{m-1} d\theta_{m-2} \ldots d\theta_1}$$

and, because we have the constraint, $\sum_{i=1}^{m} \theta_i' = 1$, we treat the $\theta_1, \theta_2, \cdots \theta_{m-1}$ as independent variables with $\theta_m = 1 - \sum_{i=1}^{m-1} \theta_i$.

The equation for $f()$ above gives us a density function for each θ_i but for most applications we need a point estimate of θ_i for each i. Therefore, we will use the expected value of θ_i which is given by,

$$\int_0^1 \int_0^{1-\theta_1} \cdots \int_0^{1-(\theta_1+\theta_2+\cdots\theta_{m-2})} \theta_i f(\overline{\theta} | n_1, n_2, \ldots n_m) d\theta_{m-1} d\theta_{m-2} \ldots d\theta_1$$

In order to solve our instantiation of the equation for Bayes Law, we have two subproblems, (1) we need to obtain the prior density function, $h(\overline{\theta})$, and (2) we need the conditional density function, $g(n_1, n_2, \ldots n_m | \overline{\theta})$.

For the prior distribution, $h(\overline{\theta})$, of the probabilities we are estimating we would like to assume as little as possible. Therefore, we select the maximum entropy distribution and assume the prior density of $(\theta_1, \theta_2, \cdots \theta_m)$ is uniform on the convex polyhedron bounded by the constraints, $\theta_i \geq 0$, which lies on the hyperplane, $\sum_{i=1}^{m} \theta_i = 1$. Thus,

---

[1] By "simple frame system" we mean one that does not include nesting or an object hierarchy.

$$h(\overline{\theta}) = h(\theta_1,\theta_2,\cdots\theta_m) = \frac{(m-1)!}{\sqrt{m}}$$

since the area of the polyhedron is

$$\frac{\sqrt{m}}{(m-1)!}.$$

To compute the conditional distribution, $g(n_1,n_2,\ldots n_m|\overline{\theta})$, we make the standard assumption that instances in our sample are independent. Therefore the probability of the sample is the product of the probabilities of the instances. The probability of an instance A_i is θ_i and we have n_i examples of A_i in our sample. Taking the product over the sample we get,

$$g(n_1,n_2,\ldots n_m|\overline{\theta}) = \theta_1{}^{n_1},\theta_2{}^{n_2},\ldots\theta_m{}^{n_m} = \prod_{i=1}^{m}\theta_i{}^{n_i}$$

Substituting $h()$ and $g()$ into the formula for $f()$ we get,

$$f() = \frac{\frac{(m-1)!}{\sqrt{m}}\theta_1{}^{n_1},\theta_2{}^{n_2},\ldots\theta_m{}^{n_m}}{\int_0^1\int_0^{1-\theta_1}\cdots\int_0^{1-(\theta_1+\theta_2+\cdots\theta_{m-2})}\frac{(m-1)!}{\sqrt{m}}\theta_1{}^{n_1},\theta_2{}^{n_2},\ldots(1-\sum_{i=1}^{m-1}\theta_i)^{n_m}d\theta_{m-1}d\theta_{m-2}\ldots d\theta_1}$$

Integrating and multiplying by θ_i to get the expected value we have an equation much like that above except that θ_i in the numerator has an exponent of n^i+1 rather than n^i:

$$\theta_i' = \frac{\int_0^1\int_0^{1-\theta_1}\cdots\int_0^{1-(\theta_1+\theta_2+\cdots\theta_{m-2})}\frac{(m-1)!}{\sqrt{m}}\theta_1{}^{n_1},\theta_2{}^{n_2},\ldots\theta_i{}^{n_i+1}\ldots\theta_m{}^{n_m}d\theta_{m-1}d\theta_{m-2}\ldots d\theta_1}{\int_0^1\int_0^{1-\theta_1}\cdots\int_0^{1-(\theta_1+\theta_2+\cdots\theta_{m-2})}\frac{(m-1)!}{\sqrt{m}}\theta_1{}^{n_1},\theta_2{}^{n_2},\ldots(\sum_{i=1}^{m-1}\theta_i)^{n_m}d\theta_{m-1}d\theta_{m-2}\ldots d\theta_1}$$

Solving for the integral in the denominator we get,

$$\int_0^1\int_0^{1-\theta_1}\cdots\int_0^{1-(\theta_1+\theta_2+\cdots\theta_{m-2})}\theta_1{}^{n_1},\theta_2{}^{n_2},\ldots\theta_m{}^{n_m}d\theta_{m-1}d\theta_{m-2}\ldots d\theta_1 = \frac{\sqrt{m}\prod_{i=1}^{m}n_i!}{(\sum_{i=1}^{m}n_i+m-1)!}$$

Solve for the integral in the numerator in roughly the same way and you get a similar result but with n_i replaced by n_i+1. Then dividing the numerator by the denominator we get an updating formula for estimating the probabilities:

$$\theta_i' = \frac{n_i+1}{\sum_{i=1}^{m}(n_i+1)}$$

In addition to providing an initial estimate for the joint probability distribution, if we store the n_i, this formula enables us to easily update that distribution as we come across additional problem instances. Notice that we are not required to store an n_i for every point in the probability space. Instead we only store these numbers for the instances which we have seen. If we are attempting to estimate the joint probability distribution in the worst case we are required to store one number for every unique instance that we come across. If there is a great deal of regularity in the domain we can expect that the number of actual instances is small relative to the number of points in the space. Furthermore, if we are estimating the distribution for a Bayesian inference net we need only store one number for each conditional probability in the network for which we have seen at least one relevant instance.

3. Summary

This paper describes a Bayesian method for estimating the joint probability distribution for a domain from a set of instances. The method has the advantage that it can be used to easily update the estimated distribution each time an additional example is observed. Thus it not only simplifies initial knowledge acquisition for a probabilistic expert system but, unlike other approaches designed for this problem, it enables systems in which it is adopted to learn from experience solving problems.

We are currently working on determining how fast estimates based on this method can be expected to converge to the true distribution. This should enable us to develop a measure of confidence for how closely the estimated distribution can be expected to approximate the true distribution given the number of instances in a sample. Another important area of future research is to show how arbitrary constraints on the prior probability distribution can be incorporated into the method. This would make it possible for an expert to specify ranges for those of the probabilities about which they have uncertain prior knowledge.

References

Cheeseman, P. A Method of Computing Generalized Bayesian Probability Values for Expert Systems. In *Proceedings of the Eighth National Conference on Artificial Intelligence.* Los Altos, CA: Morgan Kaufmann, 1983.

Cooper, Gregory F. *NESTOR: A Computer-Based Medical Diagnostic Aid that Integrates Causal and Probabilistic Knowledge* (Tech. Rep. Memo HPP-84-48). Medical Computer Science Group, Stanford University, 1984.

Cooper, Gregory F. *Expert Systems Based on Belief Networks - Current Research Directions* (Tech. Rep. Memo KSL-87-51). Medical Computer Science Group, Stanford University, 1988.

Geffner, H. and Pearl, J. Distributed Diagnosis of Systems with Multiple Faults. In *Proceedings of the Third IEEE International Conference on AI Applications.* Los Altos, CA: Morgan Kaufmann, 1987.

Heckerman, D.E., and Jimison, H. A perspective on confidence and its use in focusing attention during knowledge acquisition. In *Proceedings of the AAAI Workshop on Uncertainty in Artificial Intelligence.* , 1987.

Horvitz, E.J. Reasoning about beliefs and actions under computational constraints. In *Proceedings of the AAAI Workshop on Uncertainty in Artificial Intelligence.* , 1987.

Pearl, Judea. Fusion, Propagation, and Structuring in Belief Networks. *Artificial Intelligence*, 1986, *28*, 241 - 288.

Pearl, Judea. Distributed Revision of Composite Beliefs. *Artificial Intelligence*, 1987, *33*, 173 - 215.