TEACHING YOUNG GROWNUPS HOW TO USE BAYESIAN NETWORKS

Stefan Krauss[1], Georg Bruckmaier[1] and Laura Martignon[2]
[1]Institute of Mathematics and Mathematics Education, University of Regensburg, Germany
[2]Ludwigsburg University of Education, Germany
Stefan1.krauss@mathematik.uni-regensburg.de

A Bayesian network, or directed acyclic graphical model is a probabilistic graphical model that represents conditional dependencies and conditional independencies of a set of random variables. Each node is associated with a probability function that takes as input a particular set of values for the node's parent variables and gives the probability of the variable represented by the node, conditioned on the values of its parent nodes. Links represent probabilistic dependencies, while the absence of a link between two nodes denotes a conditional independence between them. Bayesian networks can be updated by means of Bayes' Theorem. Because Bayesian networks are a powerful representational and computational tool for probabilistic inference, it makes sense to instruct young grownups on their use and even provide familiarity with software packages like Netica. We present introductory schemes with a variety of examples.

INTRODUCTION

Thomas Bayes (1702-1761) introduced a fundamental innovation to the science of probability by providing a mathematical translation of the phrase "the likelihood of an event A given that I know the likelihood of B". His famous formula for modelling this phrase is

$$P(A \mid B) = \frac{P(A,B)}{P(B)}$$

In his view empirical knowledge about events could be encoded by events and the conditional relationships between them. Thus, for instance, the conjoint probability of an event A and B becomes

$$P(A,B) = P(A \mid B)P(B)$$

Generalizing this formula to the case of any set of mutually exclusive events B_i, $i=1,2,\ldots,n$, covering the space of possible outcomes, one obtains

$$P(A) = \sum_i P(A \mid B_i)P(B_i)$$

Expressed in terms of beliefs this formula means that our belief in an event A is a weighted sum over the beliefs in all the distinct ways that A may be realized.

The core of Bayes' innovation in probability theory is that his modelling of conditional beliefs allows for an inversion:

If, on the one hand, $P(A \mid B) = \frac{P(A,B)}{P(B)}$ and, on the other, $P(B \mid A) = \frac{P(A,B)}{P(B)}$, then

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

This is the fundamental inversion that Bayes had originally expressed in terms of "evidence" and "hypothesis": if *H* is any hypothesis and *e* a piece of evidence on *H*, then

In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010), Ljubljana, Slovenia*. Voorburg, The Netherlands: International Statistical Institute. www.stat.auckland.ac.nz/~iase/publications.php [© 2010 ISI/IASE]

$$P(H \mid E) = \frac{P(e \mid H)P(H)}{P(e)}$$

Observe that the formula for conditional probabilities leads naturally to the formula for probabilistic independence: two events A and B are independent if A does not depend upon B, that is,

$$P(A \mid B) = \frac{P(A,B)}{P(B)} = P(A) \Rightarrow P(A,B) = P(A)P(B)$$

The conditional independence relation, extended and formalized by Lauritzen and Spiegelhalter (1988), goes one step further and looks at situations with more than two events. As an example, imagine a person who wants to find out whether it makes sense to hurry and run to the next bus stop now or to take one's time and wait at one's usual bus stop. What kind of knowledge is useful for making this quick decision? If the person knows at what time the bus stopped at the stop just before them, there is no point in enquiring at what time the bus stopped at any other stop before them. More generally, one often encounters situations in which the chances of A occurring given that both B and C occur, coincide with the chances that A occurs given that B occurs. Loosely speaking, knowing about C does not add knowledge about the outcome of A, once we know B. It is a feature of reasoning under uncertainty that knowledge of the outcomes of certain events may eliminate the necessity of knowing the outcomes of other events. This type of complexity reduction is especially convenient for good decision making. One typical example is the Markov Chain situation, in which only the near past matters for the probabilities involved at a given instance.

In probability theory, the notion of conditional independence captures and models the way dependencies change in response to new facts (Pearl, 1988). A proposition A is said to be independent of B, given the information K if

$$P(A \mid B,K) = P(A \mid K)$$

As an example consider a typical day in California. Assume you have a sprinkler in your outdoor area and assume the pavement is wet. What is the probability of falling given that we consider knowledge about rain last night, knowledge on whether the sprinkler was on, knowledge on whether the pavement in the outdoor area was wet, or knowledge about the chances of having ruined one of our shoes? The network in Figure 1 expresses dependencies.

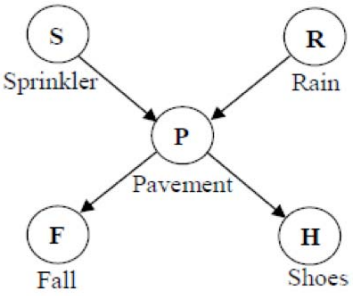

Figure 1. A network of dependencies for the outdoor area situation

Observe that the nodes "sprinkler" and "rain" are actually independent. Yet they become conditionally dependent upon pavement. If the pavement is wet and if it has not rained, the probability that the sprinkler is on rises. "Fall" and "shoes" are dependent upon each other, yet they become conditionally independent if we learn that the pavement is wet.

A Bayesian network is a directed acyclic graphical representation of a probabilistic model of the relationships among a set of variables. Each node in the network represents a variable, taking values in a set of mutually exclusive possibilities. From a node to another there may or not be an arrow. The presence of an arrow from one node to another indicates that the former has a probabilistic influence on the latter.

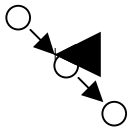

Figure 2. A possible network

In the situation depicted in Figure 2 the lowest node is dependent upon the middle node but conditionally independent of the first node at the top. The nodes with arrows ending at one specific node are called its *parents*. The node itself is called a *child* of its parents. The *Markov Blanket* of a node is the set of parents of a node, its children and the parents of its children.
Theorem (Pearl, 1988): In a Bayesian network the probability of a node variable conditioned on all the remaining variables coincides with its probability conditioned just on the Markov Blanket.

In order to illustrate the effectiveness of Bayesian networks we will now present an example (Gigerenzer, Todd & the ABC Group, 1999)─which stood at the core of the theory of human heuristics for decision-making compared to normative Bayesian models. Consider the question: "Which city has a larger population, Ulm or Nürnberg?" People, according to a large body of psychological research, answer such questions by thinking of cues of these two cities and making inferences on the population size based on these cues. Gigerenzer, Todd and the ABC Group (1999) put together the set of nine cues often used by people. They were: "Is the city a national capital (NC)?", "Is the city a state capital (SC)?", "Was the city once an exposition site (EX)?", "Is the city a member of the German industrial belt (IB)?", "Does the city have a soccer team in the National League (SO)?", "Is the city in West Germany (WG)?", "Does the city have a train station for ICE trains (IT)?", "Does the code in license plate of the city consist of one letter only (LP)?", "Does the city have a university (UN)?".

People rank these cues according to their validity, that is the probability of making a correct comparison when discriminating between two cities; for instance the cue "Is the city a national capital (NC)?" is highly valid in Germany, because Berlin, Germany's national capital, has, in fact, a larger population than any other city. The next cue in this ranking is the cue "ES", namely "Was the city once an exposition site?". What people do, according to the findings of Gigerenzer, Todd and the ABC Group (1999) is to look at each of these cues *lexicographically*. This means that the first cue that discriminates between two cities, declaring that one of the cities has a positive value on the cue while the other has a negative value on the cue, makes the decision that the city with the positive cue value has the larger population. To better explain how this lexicographic heuristic works, imagine the two German cities Ulm and Nürnberg once again. The ranking of cues according to their validities is as follows:

$$(NC) > (EX) > (SO) > (IT) > (SC) > (LP) > (UN) > (IB) > (WG).$$

Ulm's cue profile is (-1, -1, -1, 1, -1, -1, 1, -1, 1) while Nürnberg's cue profile is (-1, -1, 1, 1, -1, 1, 1, -1, 1). People judge that Nürnberg has a larger population than Ulm─which is true─because it is lexicographically larger that Ulm: the first discriminating cue, namely "Does the city have a soccer team in the national league (SO)?" is positive for Nürnberg (with a value of 1 in the profile) but negative for Ulm (with a value of -1 in the profile). What is surprising about the lexicographic heuristic people tend to use is that its predictive accuracy is quite high.

The question becomes then, which are even better decision models, at least from the strictly *normative* point of view? The Bayesian networks, as shown by Martignon & Laskey

(1999), constitute the normative benchmark for the comparison problem. Figure 3 illustrates one possible Bayesian network for the task of comparing two German cities:

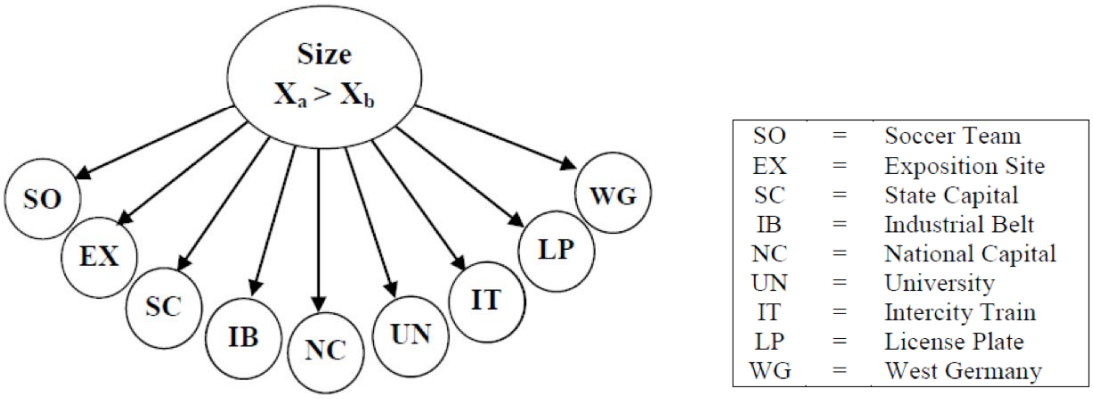

Figure 3. The naive Bayesian network for comparing the sizes of German cities

In the network of Figure 3 all cues are assumed to be conditionally independent upon the criterion. In other words, the network assumes that once we know that one city is larger than the other, then knowing that one has a station for Intercity trains does not additionally influence our belief on whether the other city has a university. The Bayesian network in Figure 3 has 10 variables: one for the criterion and one for each of the nine cues. The criterion variable can take on two possible values: the first city, say *a*, is larger than the second city, say *b*, or vice versa. Recall that in the network of Figure 1, "Fall" and "Shoes" become conditionally independent once we learn that the pavement is wet; analogously, the network of Figure 3 assumes that having or not having an Intercity train station becomes conditionally independent from having or not having a university, once we learn which one of the two cities is larger.

This Bayesian network is called "naive", because it assumes conditional independence of cues given criterion, without trying to find out whether other conditional dependencies between cues do exist. Nevertheless, one may be interested in finding out which is "the true", or "the truest" network for the given set of data? The situation being much more complex than that of the sprinkler and the wet pavement described above, we do have trouble imagining which arrows have to be placed in and which arrows have to be omitted in the network. One way of producing a Bayesian network that describes the situation of the German cities adequately is finding those links between nodes (here the nodes are the criterion and the nine cues) that are *robust and informative for prediction* (Friedman and Goldszmit, 1996).

Today there is a plethora of programmes that produce a Bayesian network based on data about the node variables. Some of these programmes are relatively simple, some very sophisticated. The problem of finding the adequate Bayesian network for a set of data can be solved by performing a search in the space of all possible networks and attributing a certain measure of goodness to each network based, for instance, in the amount of information the network adds to all networks having one link less than itself. The most popular software package for producing Bayesian networks is NETICA which is used for most straightforward applications and has excellent tutorials for use (produced by Norsys Software Corporation). It is not the most sophisticated among the software packages but definitely both the easiest to use and the most popular.

For the problem of comparing German cities Martignon and Laskey (1999) used a programme constructed by Neil Friedman and Matt Goldszmith (1996), which produced a savvy Bayesian network for the comparison problem with the German cities. This programme is based on a smart search algorithm across the space of all possible Bayesian networks for a given data set which makes use of the BIC (Bayesian Information Criterion) for evaluating the contribution of a specific network with respect to all its sub-networks, i.e., the networks formed by proper subsets of that specific network.

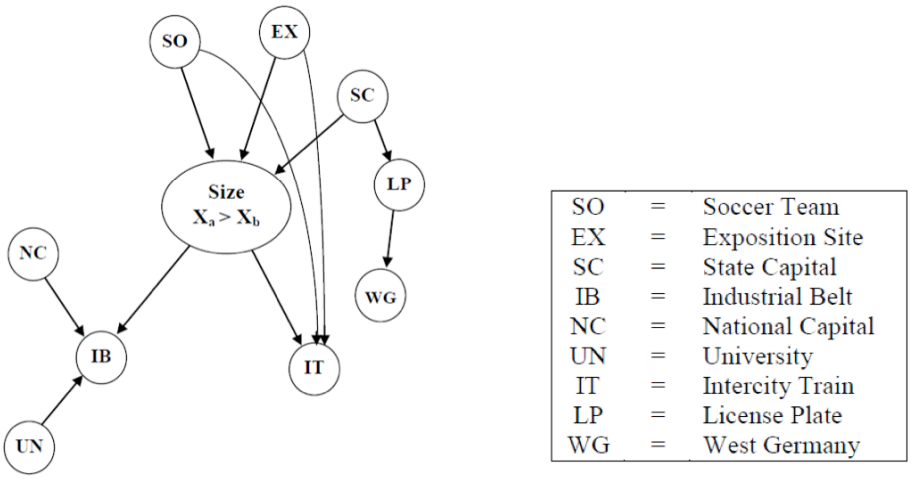

Figure 4. Full Bayesian Network for the German city data

Figure 4 represents the Bayesian network constructed by the method designed by Friedman and Goldszmith (1996). The Markov blanket of the node "Size" in this network includes all other nodes with the exception of LP ("Does the code in license plate of the city consist of one letter only (LP)?") and WG ("Is the city in West Germany (WG)?"), which becomes irrelevant to computing city size, given knowledge on the other cues.

What is important about this Bayesian network is that when trained on 50% of the German cities it is able to generalize to the remaining 50% cities remarkably well. Compared with several other models, including Multiple Regression, CART and neural networks, the Bayesian network performes with higher predictive accuracy (in the sense that it makes the highest number of correct inferences) across a very large collection of comparison problems in a variety of different environments (Martignon & Laskey, 1999). Bayesian networks have proven to be the best performing model on a variety of other tasks and across a variety of applications, particularly when generalizing from training set to test set. They represent the normative benchmark for inference (Jensen, 2001; Neapolitan, 2003).

BAYESIAN NETWORKS FOR YOUNG GROWNUPS

Bayes' theorem has long been accepted as a topic of probabilistic education in schools, not just in the Anglo Saxon countries but in most countries of the world. It is usually taught at the end of a session on conditional probabilities and motivated by typical tasks regarding the validity of a cue for classifying an item as belonging or not to a certain category. A typical task is: What is the probability that a patient has a disease if a given test result is positive? The importance of Bayesian reasoning for decision-making and for scientific discovery is seldom made clear by school texts, partly because the time dedicated to probabilistic reasoning is, in itself, short. Here we propose instructing young adults not just in probabilistic conditioning and probabilistic independence but also on the concept inherent to Bayesian networks, namely conditional independence. This concept leads to an enhancement of probabilistic reasoning techniques and their representation. Bayesian networks constitute a normative benchmark for categorization and decision-making. It is true, that human decision-making tends to be fast and frugal (Gigerenzer, Todd & the ABC Group, 1999), especially when time is limited and information costly, yet being familiar with the normative benchmarks allows an evaluation of the fast and frugal decision heuristics used by humans. Given heuristic and benchmark algorithms for decision making it is possible to assess the trade-off between computational cost and accuracy. This broad vision on decision making should be conveyed to grown-up students, especially if they have the chance of learning to use software for algorithm design. First explorative studies on the success of this type of instruction have been implemented at the Ludwigsburg University of Education. The results are extremely encouraging

and motivate the conception of specific units on Bayesian networks as part of standard courses on probability theory.

REFERENCES

Jensen, F. (2001). *Bayesian Networks and Decision Graphs*. New York: Springer-Verlag.

Friedman, N., & Goldszmit, M. (1996). Learning Bayesian Networks with local structure. In *Proceedings of the 12th Conference on Uncertainty in Artificial Intelligence* (UAI) (pp. 252-262). San Mateo, CA: Morgan Kaufmann.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, *102*(4), 684-704.

Gigerenzer, G., Todd, P., & the ABC Group (1999). *Simple heuristics that make us smart*. New York: Oxford University Press.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under Uncertainty: Heuristics and Biases*. New York: Cambridge University Press.

Lauritzen, S. L., & Spiegelhalter, D. J. (1988). Local computations with probabilities on graphical structures and their applications to expert systems. *Journal of the Royal Statistical Society, Series B*, *50*(2), 154-227.

Martignon, L., & Laskey, K. (1999). Bayesian benchmarks for fast and frugal heuristics. In G. Gigerenzer, P. Todd & the ABC Group. *Simple heuristics that makes us smart*. New York: Oxford University Press.

Martignon, L., & Krauss, S. (2009). Hands-On Activities for Fourth Graders: A Tool Box for Decision-Making and Reckoning with Risk. *International Electronic Journal of Mathematics Education*, *4*(3), 227-258.

Neapolitan, R. (2003). *Learning Bayesian Networks*. New York: Prentice Hall.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of Plausible Inference.* San Francisco: Morgan Kauffman.