Bayesian Networks and Probabilistic Inference

Drew Kristensen

December 24, 2017

Abstract

A Bayesian Network (BN) is a probabilistic model which can be used to model beliefs about a system in order to compute probabilities for outcomes given the environment[2]. These are represented via a directed acyclic graph (DAG) which related all the observed variables to the variables we wish to sample via a network of nodes and edges, where each node has certain conditional probabilities associated with it in order to model realistic situations. Dynamic Bayesian Networks (DBN) are a class of Bayesian Networks in which the observed variables at a given time can take into account the expected outcomes from the previous time slice or time slices. One example of a DBN is a Hidden Markov Model, an extension of the Markov chain with both emission and transition probabilities. This paper focuses on why inference methods work and provides an example of how to run through the processes.

1 Introduction

1.1 Independence and Dependence

There are two types of independence in probability: independence and conditional independence. A Bayesian network makes use of both forms in order to optimize and simplify the model as best it can. Two random variables X, Y are independent if and only if $P(X = x, Y = y) = P(X = x)P(Y = y)$. This is useful, since when we try to perform our probabilistic inference, we have to compute much fewer probabilities if we can factor our independent variables out from a larger probability equation. For example, if we ask a question about the probability of three independent variables taking on certain values, we can simply multiply the probabilities for each variables together instead of needing to marginalize over the variables. Conditional independence is similar to independence but requires an outside variable for the variables to be conditionally independent over. Two variables X, Y are conditionally independent given Z if $P(X, Y \mid Z) = P(X \mid Z)P(Y \mid Z)$. This is another identity that is used heavily in probabilistic inference using Bayesian networks since we can again simplify our equations.

Outside of independence, there is an idea of dependence, where if two variables are not independent, then they are dependent. That is to say that the outcome of one variable influences the probabilities of the outcome of the other. This is what is represented in our Bayesian network and makes sense from a logical perspective, since this gives us a way to represent cause and effect relationships.

1.2 Probabilistic Inference

Probabilistic inference refers to computing some probability of a random variable (or variables) given other known probabilities. We can split our random variables within our Bayesian network into three categories; query, evidence, and hidden variables. The query variables are the variables for which we are investigating, the evidence variables are the variables with values we have observed, and the hidden variables are the remaining variables over which we marginalize. Using these, we can estimate the underlying probabilities of random variables whose probability density functions are intractable. An important feature of probabilistic inference in Bayesian networks is that the conditional probability of the query variables given the evidence is proportional to the joint probability of the two sets, since by Bayes' Theorem,

$$P(Q \mid E) = \frac{P(Q, E)}{P(E)}$$

Thus, if we compute all of the joint probabilities for our query variables, then we can compute the conditional probabilities since we would only need to marginalize by the probability of the evidence variables (but this is non-negligible amount of work).

1.3 Bayesian Networks

Bayesian Networks (BN),also known as belief networks, are a probabilistic way to model events given hypothesized dependencies. The Bayesian part of the Bayesian network comes from the use of Bayesian random variables - where the variables are broad representations of unknown quantities, observable values, or unknown parameters. These could formally be described as events, since observing a Bayesian random variable is the same process as observing the outcome for an event. BNs can be represented as a directed acyclic graph (DAG) with each node being a random variable in our belief network. A graph is simply a set of nodes or vertices that are connected by a set of edges that link two vertices together. A directed graph has the property that edges between two nodes have a direction - that is, you can only travel one direction along the edge, and not back. An directed acyclic graph is a directed graph that has no cycles (ie loops) in the graph. This means that once you traverse an edge, you will be unable to reach the vertex you moved from.

2 Probabilistic Inference

2.1 Conditional probability factoring into joint probabilities

By definition of independence, if events A and B are independent, then $P(A \mid B) = P(A)$ and $P(B \mid A) = P(B)$. Furthermore, if event C is not independent of A or B, then $P(A \mid B, C) = P(A \mid C)$ and $P(B \mid A, C) = P(B \mid C)$. When we combine these with Bayes' Theorem, we develop the Chain Rule for probability, that is for any random variables $\{A, B, C, \cdots, N\}$, $P(A, B, C, \cdots, N) = P(A \mid B, C, \cdots, N)P(B, C, \cdots, N)$, since we can simply multiply both sides by the denominator used to normalize the joint. This rule can be applied repeatedly to factor the joint into a series of conditionals, which can be simplified to only include its parent variables. These key facts are used to simplify our calculations and understand difference dependencies when we come to probabilistic inference. For example, in the model in figure 3[1] on page 9 can be factored into

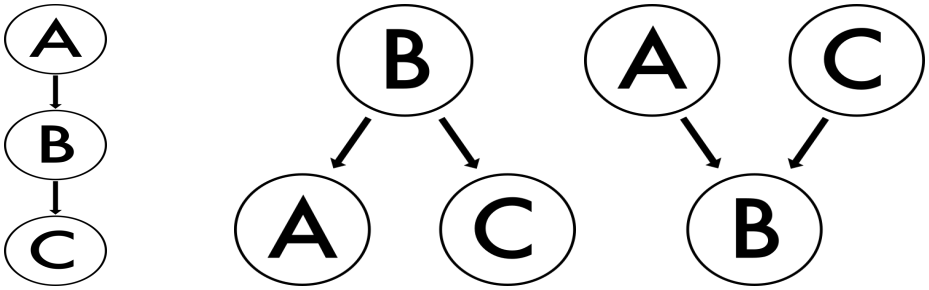
$$P(A, B, E, J, M) = P(A \mid B, E)P(B)P(E)P(J \mid A)P(M \mid A).$$

This allows us to preform probabilistic inference much easier, since we have to preform less computations marginalizing over our hidden variables. For our example, when we ask questions about variables lower down in our belief network, we can simply pull out our summations over B and E to the front since neither depend on the outcome of another

2.2 D-separation and Markov Blankets

D-separation is an algorithm that factors a Bayesian network from a massive joint probability equation to a product of conditional probabilities by detecting independences between nodes. The idea behind how it works is that there are only 3 distinct ways for a three nodes to be connected with two edges: either the middle node has two outgoing connections (diverging), the middle node has two incoming connections (converging, or the middle node has one incoming and one outgoing connection (linear). For each of these, we can prove conditional independence based on the probabilities of the situation.

For a linear connection, we have the case where $A \to B \to C$, where B is our middle node. From this, we know that we can factor our entire joint probability into $P(A, B, C) = P(A)P(B \mid A)P(C \mid B)$, since we have defined B to be dependent on A and C to be dependent on B. In this case, our end nodes A and C are dependent when B is not observed. We can see this by looking



(a) Linear **(b)** Diverging **(c)** Converging

Figure 1: Examples of all three relationships

at the probability of the two by marginalizing over the hidden variable B.

$$\begin{aligned} P(A,C) =& P(A)\sum_B P(B \mid A)P(C \mid B) \\ \frac{P(A,C)}{P(A)} =& \sum_B P(B \mid A)P(C \mid B) \\ P(C \mid A) =& \sum_B P(B \mid A)P(C \mid B) \end{aligned}$$

Since $P(C \mid A) \neq P(C)$, A and C are not independent. However, when we 'observe' B, that is, we bring B into our set of evidence variables, A and C are rendered independent. We know

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

we can expand our joint as described above to get

$$\begin{aligned} P(A,C \mid B) =& \frac{P(A)P(B \mid A)P(C \mid B)}{P(B)} \\ =& \frac{P(A)\frac{P(A,B)}{P(A)}\frac{P(B,C)}{P(B)}}{P(B)} \\ =& \frac{P(A,B)P(B,C)}{P(B)^2} \\ =& \frac{P(A,B)}{P(B)}\frac{P(B,C)}{P(B)} \\ P(A,C \mid B) =& P(A \mid B)P(C \mid B). \end{aligned}$$

So we have that A and C are conditionally independent given B. So, for the linear case, the end nodes are dependent until the middle node is observed.

This is also the case of the divergent connections. The divergent connection for a node B is given by $A \leftarrow B \rightarrow C$. For this, we factor the joint probability into $P(A,B,C) = P(A \mid B)P(B)P(C \mid B)$. For events A, C, we can break their conditional into

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

Since we have that we can get the probability of both A and C by marginalizing over B (from the joint distribution equation), then we can expand these probabilities into

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

Since we cannot factor the numerator to pull out the factor in the denominator, we have that $P(A \mid C) \neq P(A)$, meaning A and C are not independent.

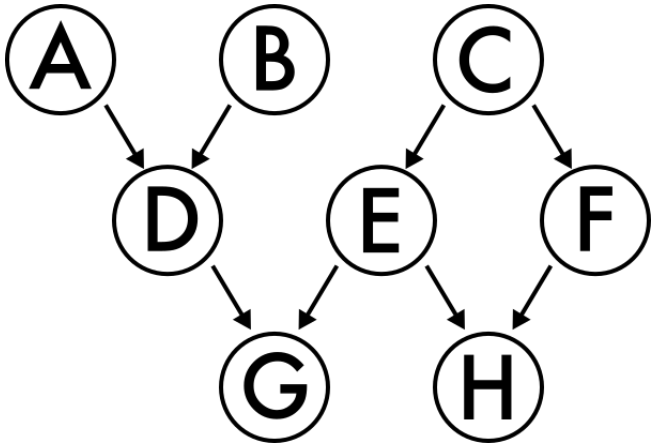


Figure 2: Example belief network

From the joint probability, it is trivial to see that A, C are conditionally independent given B since $P(A, C \mid B) = \frac{P(A,B,C)}{P(B)}$ since the $P(B)$ in the numerator and denominator cancel each other out to leave $P(A, C \mid B) = P(A \mid B)P(C \mid B)$.

The last case is the convergent connections where $A \rightarrow B \leftarrow C$. Here, we can factor our joint probability here to have $P(A, B, C) = P(A)P(B \mid A, C)P(C)$. Interestingly, we have that the end nodes of A, C are independent until B is observed. To see this, we can first see quickly that our ends are independent when B is not part of our evidence. By definition of the connections,

$$P(A, C) = P(A)P(C)$$

So A, C are independent without B. The interesting case comes when B is incorporated into the evidence. This case gives us

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

we can expand our joint as described above to get

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

Since $P(B)$ is the marginal probability of B over A and C, we can expand the denominator to be

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

Since we now have an expression in which $P(A, C \mid B) \neq P(A \mid C)P(C \mid B)$, then A and C are not independent given B.

Now that we have all the connections explained, we can formalize how we define independence between two nodes. For two nodes A, B in our Bayesian network, if every path from A to B has two end nodes M, N where M, N are independent, then A, B are independent of each other. If there exists any path between A and B where all end nodes are not independent of each other, then A and B are not independent of each other.

We will run through an example. In figure 2, we can ask a question as to whether or not D and F are independent. Running through d-separation, since our path from D to F starts with $D \rightarrow G \leftarrow E$, and we aren't observing G, then D and E are independent. Since the only path from D to F is 'blocked' by an independence, then D and F are independent. However, we can also ask if D and F are independent if we observe G. Here, in our path from $D \rightarrow G \leftarrow E$, observing G renders D and E dependent. When we check if E and F are independent, we check the path $E \rightarrow H \leftarrow F$. Since we aren't observing H, this path is blocked. However, there is also the path $E \leftarrow C \rightarrow G$. Since we aren't observing C, E and F are dependent. Since there exists a path from D to F, namely $D \rightarrow G \leftarrow E \leftarrow C \rightarrow F$, that is not blocked, D and F are not independent when we observe G.

This algorithm leads to the idea known as the Markov Blanket - where some node X in our network is conditionally independent of all other nodes given its parents, children, and children's

parents. This makes computing probabilities based on evidence easier and quicker, since the amount of probabilities you need to compute drops significantly. For example, for a boolean Bayesian network with 10 nodes, there are 2^{10} different settings of the boolean values in the network, so reducing the number of calculations proves to be a very useful practice. If each node can instead be dependent on only 3 other nodes, then the number of computations goes from 1024 all the way down to $10(2^4) = 160$.

2.3 Using Probabilistic Inference

As mentioned previously, probabilistic inference is best suited for tasks where the underlying probability density function is unavailable or impossible to compute given the resources, but you still want some idea for the kinds of probabilities that make up your model. Probabilistic inference gives us an accurate representation of the conditional probabilities within our model, but are seldom used in computer science due to the time taken to compute. To compute the probability of some query node given some evidence nodes, we represent it as the joint probability, since we know the conditional and joint probabilities will be proportional to each other (the only difference is the factor of the probability of our evidence on the denominator of the conditional). Then, we have

$$P(Q|E) \propto \sum_{h \in H} p(h_1, h_2, \cdots h_k, q, e_1, e_2, \cdots e_l).$$

From this, we can break each random variable up into its conditional dependencies which we can compute from the D-separation approach discussed in subsection 2.2. Using the factoring properties from subsection 2.1, we can marginalize over hidden variables that we broke up from the factoring, and pull out variables from our summations that are independent of the summed variables. The process goes as follows:

1. Set the evidence variables to be consistent with the inference task
2. Sum over all the factored hidden variables' probabilities
3. Compute the probability for each query variable taking on each possible value in it's domain
4. Normalize the computed probabilities to sum to 1 to be a proper probability distribution

From this, we can compute conditional probabilities for any set of query and evidence variables within our belief network.

Example of Probabilistic Inference in a Bayesian network

Take our example Bayesian network in figure 3 on page 9, we can perform probabilistic inference to answer non-obvious questions, such as the probability that John would call given that there was a burglary and no earthquake? We need to compute the probabilities of all possible events given John calling. Since we are performing our inference over two variables, we will compute four probabilities, $P(b, e \mid j), P(b, \neg e \mid j), P(\neg b, e \mid j), P(\neg b, \neg e \mid j)$. We start by computing the a probability proportional to the probability of John calling given there was an earthquake and no burglary.

$$\begin{aligned}
P(b, \neg e \mid j) &\propto P(b, \neg e, j) \\
&= \sum_A \sum_M P(A, M, j, b, \neg e) \\
&= P(\neg e)P(b) \sum_A P(A \mid \neg e, b)P(j \mid A) \sum_M P(M \mid A) \\
&= (0.998)(0.001)\left[(0.94)(0.9) + (0.06)(0.05)\right] \\
P(b, \neg e \mid j) &\propto 0.000847302.
\end{aligned}$$

Next we compute the probabilities of all the events besides our $P(b, \neg e \mid j)$.

First we compute $P(b, e \mid j)$

$$\begin{aligned} P(b, e \mid j) &\propto P(b, e, j) \\ &= \sum_A \sum_M P(A, M, j, b, e) \\ &= P(e)P(b) \sum_A P(A \mid e, b)P(j \mid A) \sum_M P(M \mid A) \\ &= (0.002)(0.001)\left[(0.95)(0.9) + (0.04)(0.05)\right] \\ P(b, e \mid j) &\propto 0.000001714. \end{aligned}$$

Second, we compute $P(\neg b, e \mid j)$

$$\begin{aligned} P(\neg b, e \mid j) &\propto P(\neg b, e, j) \\ &= \sum_A \sum_M P(A, M, j, \neg b, e) \\ &= P(e)P(\neg b) \sum_A P(A \mid e, \neg b)P(j \mid A) \sum_M P(M \mid A) \\ &= (0.002)(0.999)\left[(0.29)(0.9) + (0.71)(0.05)\right] \\ P(\neg b, e \mid j) &\propto 0.000592407. \end{aligned}$$

Lastly, we compute $P(\neg b, \neg e \mid j)$

$$\begin{aligned} P(\neg b, \neg e \mid j) &\propto P(\neg b, \neg e, j) \\ &= \sum_A \sum_M P(A, M, j, \neg b, \neg e) \\ &= P(\neg e)P(\neg b) \sum_A P(A \mid \neg e, \neg b)P(j \mid A) \sum_M P(M \mid A) \\ &= (0.998)(0.999)\left[(0.05)(0.9) + (0.95)(0.05)\right] \\ P(\neg b, \neg e \mid j) &\propto 0.092222685. \end{aligned}$$

Now that we have all the probabilities, we can compute the probability of our specific event in question. To do this, we normalize the proportional probability of $P(b, \neg e \mid j)$ by the sum of all probabilities that we computed. So,

$$P(b, \neg e \mid j) = \frac{0.000847302}{0.000847302 + 0.000001714 + 0.000592407 + 0.092222685} \approx 0.00904617.$$

Thus, the probability of John calling after a burglary and no earthquake is about 0.00904617. Conversely, there is a probability of approximately 0.9846107 that John is calling given neither an earthquake nor a burglary occurred. This shows a classic probability property - when there are large and small probabilities, like the probabilities of earthquakes and burglaries occurring, then it is so much more likely for them not to occur that the probabilities associated with them occurring are drowned out.

2.4 Dynamic Bayesian Networks

A dynamic Bayesian network (DBN) is a Bayesian network that operates over successive periods of time. At each time step in the DBN, there is X_t, the 'state' of the model or the set of hidden variables at time t, and E_t, the set of evidence at time t that we observe. We will introduce some notation; let $A_{a:b}$ represent the set of the states of some A from step a through step b. For example, $A_{a:b} = \{A_a, A_{a+1}, A_{a+2}, \cdots, A_{b-1}, A_b\}$. These help us model how the state we are interested in can change over time, especially given specific parameters and starting environments. A dynamic Bayesian network has two parts - a transmission model, which specifies $P(X_t \mid X_{0:t-1})$, and an emission model, which specifies $P(E_t \mid X_{0:t-1}, E_{1:t-1})$. Put simply, the emission model specifies the probabilities of taking certain actions in response to the environment while the changes from state to state in the environment are specified by the transmission model, which can be represented as a single matrix, since we make the assumption that at all t, $P(E_t \mid X_{0:t}, E_{1:t-1}) = P(E_t \mid X_t)$.

The transmission model can be represented by a Markov chain, since we make the assumption that $P(X_t \mid X_{0:t-1}) = P(X_t \mid X_{t-1})$ for all t [this is known as the Markov assumption for Bayesian networks]. The simplest form of a dynamic Bayesian network is composed of single discrete state random variable - this is known as a hidden Markov model.

Since we cannot directly observe the state in some cases, we need to use inference methods in order to compute the probabilities of certain events happening. The inference method we will focus on is known as filtering - the method when we want to compute $P(X_t \mid E_{0:t})$

2.5 Inference in DBNs

Depending on the question asked, DBNs make use of different inference techniques, but it is common to ask questions about the current state given the past evidence since the state is the unobservable set of hidden variables. Since this is an algorithm, there is a sequence of steps we take to compute our probabilities.

1. Update the current state
2. Incorporate the new evidence
3. Repeat until at the desired time t

Filtering makes extensive use of Bayes' rule, since we know that we can factor conditional probabilities in the following way.

$$\begin{aligned}
P(X_t \mid E_{1:t}) =& \frac{P(X_t, E_{1:t-1}, E_t)}{P(E_{1:t})} \\
=& \frac{P(E_t \mid X_t, E_{1:t-1})P(X_t, E_{1:t-1})}{P(E_{1:t})} \\
=& \frac{P(E_t \mid X_t, E_{1:t-1})\frac{P(X_t|E_{1:t-1})}{P(E_{1:t-1})}}{P(E_{1:t})} \\
=& \frac{P(E_t \mid X_t, E_{1:t-1})P(X_t \mid E_{1:t-1})}{P(E_{1:t})^2}
\end{aligned}$$

Since we have made the Sensor Markov assumption, we have that E_t is only dependent on X_t, so $P(E_t \mid X_t, E_{1:t-1}) = P(E_t \mid X_t)$. Furthermore, since we know what our evidence is already, this factor in the denominator will be a constant for all values of X_t, so we can simply say that $P(X_t \mid E_{1:t}) \propto P(E_t \mid X_t)P(X_t \mid E_{1:t-1})$. We can again break up one of conditional probabilities by summing over all the previous probabilities in the state

$$\begin{aligned}
P(X_t \mid E_{1:t-1}) &= \sum_{X_{t-1}} P(X_t, X_{t-1} \mid E_{1:t-1}) \\
&= \sum_{X_{t-1}} P(X_t, X_{t-1}, E_{1:t-1})P(X_{t-1} \mid E_{1:t-1})
\end{aligned}$$

Since we made a first order Markov assumption for our transmission model along with the stationary state assumption, $P(X_t, X_{t-1}, E_{1:t-1}) = P(X_t \mid X_{t-1})$ for all t. This leaves us with

$$P(X_t \mid E_{1:t-1}) = \sum_{X_{t-1}} P(X_t \mid X_{t-1})P(X_{t-1} \mid E_{1:t-1})$$

which is helpful, since we can run the algorithm on the second term in the sum to compute the probability. By computing the probabilities from $t = 0$ up through whatever time we want, we skip the recursion and are left with a constructive algorithm. Using this, we can compute the probability of a state given all the past evidence up to the time we want. This is the simplest way of computing the probability since it linearizes the recursive calculation. This gives us an approach to answering questions such as "is it raining today if I saw people wearing coats the past three days?".

Example of Probabilistic Inference in a dynamic Bayesian network

We look at figure 4 on page 10 and take this to be our dynamic Bayesian network. In this example, we have three states; No Class, Lecture, and Exam. We assume that the transmission probabilities are the ones represented in the diagram. For each of the states, we have different probabilities. For

example, if there is an exam, the probability of a student crying is 0.40, whereas if there is a day without class, the probability of a student crying is 0. Now that we have a model, we can perform our filtering algorithm to compute some probability. For our example, say we want to know the probability that on the third day, the students were in lecture, given that the first day, we saw a happy student and the second and third days we saw sad students leaving the classroom. Starting at the bottom, we have

$$P(X_1 \mid E_1 = H) \propto P(E_1 = H \mid X_1) \sum_{X_0} P(X_1 \mid X_0) P(X_0 \mid \{\emptyset\})$$

For this, we let the probability X_0 taking on some value be equal for all values of X_0. Then

$$P(X_1 \mid E_1 = H) \propto P(E_1 = H \mid X_1) \left[P(X_1 \mid X_0 = N) \left(\frac{1}{3}\right) + P(X_1 \mid X_0 = L) \left(\frac{1}{3}\right) + P(X_1 \mid X_0 = E) \left(\frac{1}{3}\right) \right]$$

Using this, we can compute the probability for $P(X_2 \mid E_2 = S, E_1 = H)$. Since $P(X_2 \mid E_2 = S, E_1 = H) \propto P(E_2 = S \mid X_2) \sum_{X_1} P(X_2 \mid X_1) P(X_1 \mid E_1 = H)$, we will need to compute probabilities for all values of X_1 given $E_1 = H$. We compute these values using the formula we found.

a) $P(X_1 = N)$

$$\begin{aligned} P(X_1 = N \mid E_1 = H) &\propto (0.99) \left[(0.05) \left(\frac{1}{3}\right) + (0.05) \left(\frac{1}{3}\right) + (0.8) \left(\frac{1}{3}\right) \right] \\ &= 0.297. \end{aligned}$$

b) $P(X_1 = L)$

$$\begin{aligned} P(X_1 = L \mid E_1 = H) &\propto (0.5) \left[(0.85) \left(\frac{1}{3}\right) + (0.65) \left(\frac{1}{3}\right) + (0.2) \left(\frac{1}{3}\right) \right] \\ &= \frac{17}{60}. \end{aligned}$$

c) $P(X_1 = E)$

$$\begin{aligned} P(X_1 = E \mid E_1 = H) &\propto (0.1) \left[(0.1) \left(\frac{1}{3}\right) + (0.3) \left(\frac{1}{3}\right) + (0.0) \left(\frac{1}{3}\right) \right] \\ &= \frac{1}{75}. \end{aligned}$$

Now that we have values for X_1 taking on values, we can substitute in when we solve for the probabilities of X_2. Since $P(X_3 \mid E_3 = S, E_2 = S, E_1 = H) \propto P(E_3 = S \mid X_3) \sum_{X_2} P(X_3 \mid X_2) P(X_2 \mid E_2 = S)$, we need to do the same process as above to compute all probabilities for X_2.

To compute $P(X_2 = N)$, we find

$$\begin{aligned} P(X_2 = N \mid E_2 = S) &\propto (0.01) \left[(0.05)\,(0.297) + (0.05) \left(\frac{17}{60}\right) + (0.8) \left(\frac{1}{75}\right) \right] \\ &= \frac{2381}{6000000}. \end{aligned}$$

By the same process, we get

$$P(X_2 = L \mid E_2 = S) = \frac{79071}{400000} \qquad \text{and} \qquad P(X_2 = E \mid E_2 = S) = \frac{1147}{20000}$$

Lastly, we compute all the probabilities of X_3 to find the distribution so we can answer our original question. Since we solved for the probabilities associated with X_2 taking on values given our evidence, an since we have a formula for X_3, we can perform the same process over the values of X_3 to get

a) $P(X_3 = N \mid E_3 = S) \approx 0.0005578371$
b) $P(X_3 = L \mid E_3 = S) = 0.0631339575$
c) $P(X_3 = E \mid E_3 = S) \approx 0.0296714666$

When we normalize these probabilities by dividing all of them by their sum, we get

$P(X_3 = N \mid E_{1:3}) \approx 0.0059749$
$P(X_3 = L \mid E_{1:3}) \approx 0.6762184$
$P(X_3 = E \mid E_{1:3}) \approx 0.3178066$.

Therefore, the answer to our question is that there is approximately a 0.67621 probability that given the observed evidence, the students were in lecture on the third day.

This was a long process, but it shows the expressive power of dynamic Bayesian networks. It allows you to compute probabilities that would be otherwise extremely difficult to compute and gives a process for a computation that seems intractable at first glance.

3 Conclusion

Bayesian networks are useful tools for computing probabilities for variables or events where the probability density function itself is unobtainable or intractable. The extension of Bayesian networks, dynamic Bayesian networks, have proven useful in modeling time sequence data, and we described one method of inference in the dynamic case. Using our inference methods, we can answer questions about our model by running through the steps discussed in this paper.

References

[1] America Chambers. "Lecture Slides from March 1 Class". Given at the University of Puget Sound. Mar. 2017.

[2] Drew Kristensen. "Lecture notes in Intro to Artificial Intelligence". Given at the University of Puget Sound. Mar. 2017.

Appendix

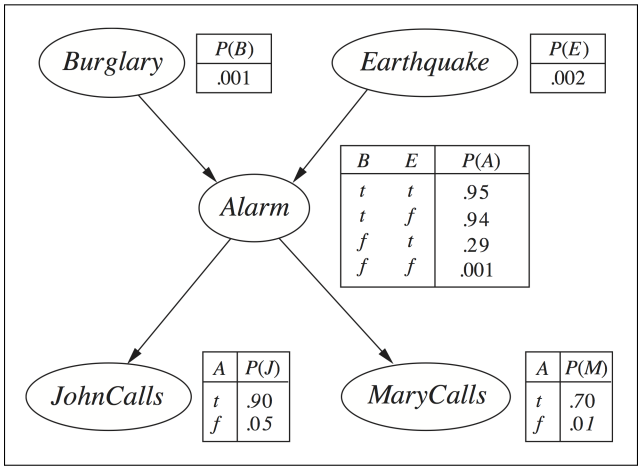


Figure 3: Simple Bayesian network for a toy problem[1]

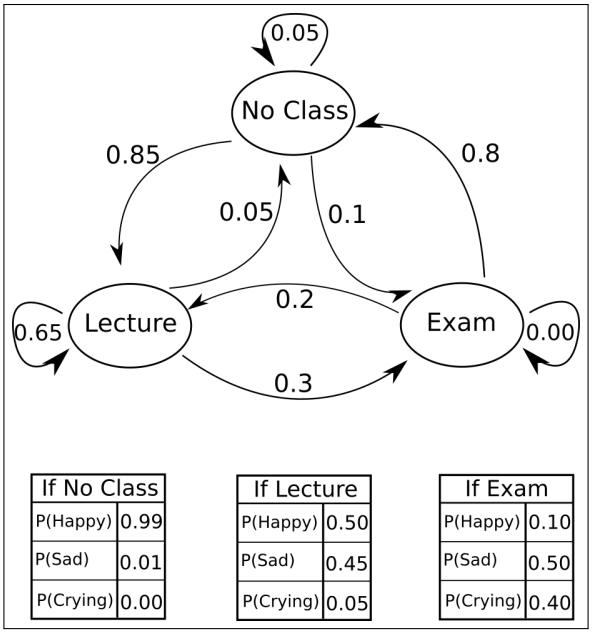


Figure 4: Example of a dynamic Bayesian network