0.1 Bayesian Networks

Suppose that W is a set of possible worlds characterized by n binary random variables $\mathcal{X} = \{X_1, \ldots, X_n\}$ (or, equivalently, n primitive propositions). That is, a world $w \in W$ is a tuple $(x_1, \ldots, x_n)$, where $x_i \in \{0,1\}$ is the value of X_i. That means that there are 2^n worlds in W, say $w_1, \ldots, w_{2^n}$. A naive description of a probability measure on W requires $2^n - 1$ numbers, $\alpha_1, \ldots, \alpha_{2^n-1}$, where α_i is the probability of world w_i. (Of course, the probability of w_{2^n} is determined by the other probabilities, since they must sum to 1.)

If n is relatively small, describing a probability measure in this naive way is not so unreasonable, but if n is, say, 1000 (certainly not unreasonable in many practical applications), then it is completely infeasible. Bayesian networks provide a tool for describing and working with probability measures that is computationally far more feasible.

A (qualitative) *Bayesian network* (sometimes called a *belief network*) is a *dag*, that is, a directed acyclic graph, whose nodes are labeled by random variables. (For readers not familiar with graph theory, a *directed graph* just consists of a collection of *nodes* or *vertices* joined by directed edges. Formally, an edge is just an ordered pair of nodes; the edge (u, v) can be drawn by joining u and v by a line with an arrow at the end point from u to v. A directed graph is *acyclic* if there is no *cycle*; that is, there does not exist a sequence of vertices $v_0, \ldots, v_k$ such that $v_0 = v_k$ and there is an edge from v_i to v_{i+1} for $i = 0, \ldots, k-1$.) Informally, the edges in a Bayesian network can be thought of as representing causal influence. For example, suppose that we were interested in reasoning about the relationship between smoking and cancer. Our model might include binary random variables such as C for "has cancer", SH for "exposed to second-hand smoke", PS for "at least one parent smokes", and S for "smokes". The Bayesian network G_s in Figure 0.1 might represent the situation.

G_s is best thought of in causal terms. Intuitively, it says that whether or not a patient has cancer is directly influenced by whether he is exposed to second-hand smoke and whether he smokes. Both of these random variables, in turn, are influenced by whether his parents smoke. Whether or not his parents smoke also clearly influences whether or not he has cancer, but this influence is mediated through the random variables SH and S. Once we know whether he smokes and was exposed to second-hand smoke, finding out whether his parents smoke gives no additional information. That is, C is independent of PS given SH and S.

More generally, given a Bayesian network G and a node X in G, think of the *ancestors* of X in the graph, where Y is an ancestor of X if there is a directed path from Y to X in G (i.e., a sequence $(Y_1, \ldots, Y_k)$ of nodes

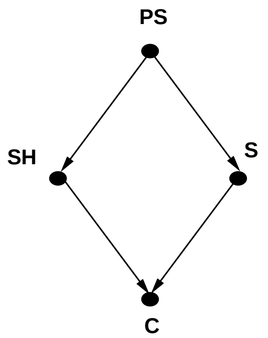

Figure 0.1: A Bayesian network that represents smoking.

such that $Y_1 = Y$, $Y_k = X$, and there is a directed edge from Y_i to Y_{i+1} for $i = 1, \ldots, k-1$), as those random variables that have a potential influence on X. This influence is mediated through the *parents* of X, those ancestors of X directly connected to X. More formally, that means that X should be conditionally independent of its ancestors, given its parents. The formal definition requires that X be independent not only of its ancestors, but of its *nondescendants*, given its parents, where the nondescendants of X are those nodes Y such that X is not the ancestor of Y.

Definition 0.1.1 Given a qualitative Bayesian network G, let $\mathrm{Par}_G(X)$ be the *parents* of the random variable X in G; let $\mathrm{Des}_G(X)$ be all the *descendants* of X, that is, X and all those nodes Y such that X is an ancestor of Y; let $\mathrm{NonDes}_G(X)$, the *nondescendants of* X consist of $\mathcal{X} - \mathrm{Des}_G(X)$. Note that all ancestors of X are nondescendants of X. The Bayesian network G *(qualitatively) represents*, or *is compatible with*, the probability measure μ if $I_\mu^{rv}(X, \mathrm{NonDes}_G(X)|\mathrm{Par}(X))$, that is, X is conditionally independent of its nondescendants given its parents, for all $X \in \mathcal{X}$. ∎

Definition 0.1.1 says that G represents μ if, in a certain sense, it captures the conditional independence relations in μ. But what does this notion of representation buy us? Why is it at all interesting? Suppose without loss of generality that the variables in $\mathcal{X}$ are *topologically sorted*, that is, if X_i is a parent of X_j, then $i < j$. That means that $\{X_1, \ldots, X_{i-1}\} \subseteq \mathrm{NonDes}_G(X_i)$, for $i = 1, \ldots, k$. It is immediate from the definition of conditional probability that

$$\begin{array}{ll} & \mu(X_1 = x_1 \cap \ldots \cap X_n = x_n) \\ = & \mu(X_n = x_n | X_1 = x_1 \cap \ldots \cap X_{n-1} = x_{n-1}) \times \mu(X_1 = x_1 \cap \ldots \cap X_{n-1} = x_{n-1}). \end{array}$$

(Assume for now all the relevant probabilities are positive, so that the conditional probabilities are well defined.) Applying this observation inductively, it follows that

$$\begin{array}{ll} & \mu(X_1 = x_1 \cap \ldots \cap X_n = x_n) \\ = & \mu(X_n = x_n | X_1 = x_1 \cap \ldots \cap X_{n-1} = x_{n-1}) \times \\ & \quad \mu(X_{n-1} = x_{n-1} | X_1 = x_1 \cap \ldots \cap X_{n-2} = x_{n-2}) \times \\ & \quad \ldots \times \mu(X_2 = x_2 | X_1 = x_1) \times \mu(X_1 = x_1). \end{array} \tag{0.1}$$

This is an instance of what is called the *chain rule* for conditional probability.

Now suppose that G represents μ. Since I assumed that the nodes in $\mathcal{X}$ were topologically sorted, that means that $\{X_1, \ldots, X_{k-1}\} \subseteq \mathrm{NonDes}_G(X_k)$, for $k = 1, \ldots, n$; all the descendants of X_k must have subscripts greater than k. Thus, all the nodes in $\{X_1, \ldots, X_{k-1}\}$ are independent of X_k given $\mathrm{Par}_G(X_k)$. It follows that

$$\mu(X_k = x_k | X_{k-1} = x_{k-1} \cap \ldots \cap X_1 = x_1) = \mu(X_k = x_k | \cap_{X_i \in \mathrm{Par}(X_k)} X_i = x_i).$$

Thus, if G represents μ, then (0.1) reduces to

$$\begin{array}{ll} & \mu(X_1 = x_1 \cap \ldots \cap X_n = x_n) \\ = & \mu(X_n = x_n | \cap_{X_i \in \mathrm{Par}(X_n)} X_i = x_i) \times \\ & \mu(X_{n-1} = x_{n-1} | X_1 = \cap_{X_i \in \mathrm{Par}(X_{n-1})} X_i = x_i) \times \ldots \times \mu(X_1). \end{array} \tag{0.2}$$

A *quantitative Bayesian network* is a pair (G, f) consisting of a qualitative Bayesian network G and a function f that associates with each node X in G a *conditional probability table (cpt)* that quantifies the effects of the parents of X on X. There is an entry in $[0,1]$ in the cpt for each possible setting of the parents of X. Intuitively, the entries in the cpt for X describe the probability that $X = 1$ conditional on all the possible values of X's parents. If X is a root of G, then the cpt for X can be thought of as giving the unconditional probability that $X = 1$.

For example, we may have the following cpt for the random variable C in Figure 0.1, where the first line describes the conditional probability that $C = 1$ given $S = 1 \cap SH = 1$, the second line describes the conditional probability that $C = 1$ given $S = 1 \cap SH = 0$, and so on.

S	SH	C
1	1	.6
1	0	.4
0	1	.1
0	0	.01

Since the probability that $C = 0$ given some setting of S and SH is 1 minus the probability that $C = 1$ given that setting of S and SH, there is no need to describe explicitly the conditional probability that $C = 0$ given some setting of S and SH in the cpt.

Similarly, we could have the following cpts for S, C, and PS:

PS	S
1	.4
0	.2

PS	SH
1	.8
0	.3

PS
.3

For future reference, let f_S be the function associating the random variables in G_s with the cpts described above.

Definition 0.1.2 A quantitative Bayesian network (G, f) *(quantitatively) represents*, or *is compatible with*, the probability measure μ if G qualitatively represents μ and the cpts agree with μ, in the sense that, for each random variable X, the entry in the cpt for X given some setting $Y_1 = y_1, \ldots, Y_k = y_k$ of its parents is $\mu(X = 1 | Y_1 = y_1 \cap \ldots \cap Y_k = y_k)$ if $\mu(Y_1 = y_1 \cap \ldots \cap Y_k = y_k) \neq 0$. (It does not matter what the cpt entry for $Y_1 = y_1, \ldots, Y_k = y_k$ is if $\mu(Y_1 = y_1 \cap \ldots \cap Y_k = y_k) = 0$.) ∎

It follows immediately from (0.2) that if (G, f) quantitatively represents μ, then we can completely reconstruct μ from (G, f). More precisely, (0.2) shows that we can compute the 2^n values $\mu(X_1 = x_1 \cap \ldots \cap X_n = x_n)$ from (G, f); from these values, we can easily compute $\mu(A)$ for all $U \subseteq W$.

Proposition 0.1.3 *A quantitative belief network (G, f) always quantitatively represents a unique probability measure, the one determined by using (0.2).*

Proof See Exercise 0.2. ∎

It is easy to calculate that for the unique probability measure μ represented by the quantitative Bayesian network (G_s, f_s),

$$\begin{array}{rl} & \mu(PS = 0 \cap S = 0 \cap SH = 1 \cap C = 1) \\ = & \mu(C = 1 | S = 0 \cap SH = 1) \times \mu(S = 0 | PS = 0) \times \mu(SH = 1 | PS = 0) \times \mu(PS = 0) \\ = & .1 \times .8 \times .3 \times .7 \\ = & .0168 \end{array}$$

(These numbers have been made up purely for example, and bear no necessary relationship to reality!)

Proposition 0.1.3, while straightforward, is important because it shows that there is no choice of numbers in the cpts that can be inconsistent with probability. Whatever the numbers are in the cpt, (as long as they are in the interval $[0,1]$) there is always a probability measure that is compatible with (G, f).

What about the converse? Can every probability measure on W be represented by a quantitative belief network? It can, and in general there are many ways of doing so.

Let $Y_1, \ldots, Y_n$ be any permutation of the random variables in $\mathcal{X}$. (Think of $Y_1, \ldots, Y_n$ as describing an ordering of the variables in $\mathcal{X}$.) Construct a qualitative Bayesian network as follows: Start with the nodes $Y_1, \ldots, Y_n$. For each k, find a minimal subset of $Y_1, \ldots, Y_k$, call it $\mathbf{P}_k$, such that $I_\mu^{rv}(\{Y_1, \ldots, Y_{k-1}\}, Y_k | \mathbf{P}_k)$. (Clearly there is a subset with this property, namely $\{Y_1, \ldots, Y_k\}$ itself. It follows that there must be a minimal subset with this property.) Then add edges from each of the nodes in $\mathbf{P}_k$ to Y_k. Call the resulting graph G.

Theorem 0.1.4 *G qualitatively represents μ.*

Proof Note that $Y_1, \ldots, Y_k$ represents a topological sort of G; edges always go from nodes in $\{Y_1, \ldots, Y_{k-1}\}$ to Y_k. It follows that G is acyclic; i.e., it is a dag. The construction guarantees that $\mathbf{P}_k = \mathrm{Par}_G(Y_k)$ and that $I_\mu^{rv}(\{Y_1, \ldots, Y_{k-1}\}, Y_k | \mathrm{Par}_G(Y_k))$. Using CIRV1-5, it can be shown that $I_\mu^{rv}(\mathrm{NonDes}_G(Y_k), Y_k | \mathrm{Par}_G(Y_k))$ (Exercise 0.3). Thus, G qualitatively represents μ. ∎

How much does this buy us? That depends on how sparse the graph is, that is, how many parents each node has. If a node has k parents, then its conditional probability table has 2^k entries. For example, the cpt for C above has four entries, and the cpts for SH and S each have two entries, while the one for PS has only one entry. That means if each node has at most k parents in the graph, then there are at most $n2^k$ entries in all the cpts. If k is small, then $n2^k$ can be much smaller than $2^n - 1$. In the quantitative Bayesian network (G_s, f_s) for the smoking example, there are nine entries altogether in the cpts. A naive description of the probability distribution would involve fifteen numbers. For large n, the discrepancy can be much greater. (The numbers 2^k and $2^n - 1$ arise because I have considered only binary random variables. If the random variables can have m values, say $0, 1, \ldots, m-1$, the conditional probability table for a random variable X with m parents would have to describe the probability that $X = j$, for $j = 1, \ldots, m-1$, for each of the m^k possible settings of its parents, so would involve $(m-1)m^k$ entries.) Not only does a well-designed Bayesian network (I discuss what "well-designed" means below) typically require far fewer numbers to represent a probability measure, the numbers are typically easier to obtain. For example, for (G_s, f_s), it is typically easier to obtain entries in the cpt like $\mu(C = 1 | S = 1 \cap SH = 0)$—the probability that someone gets cancer given that they smoke and are not exposed to second-hand smoke—than it is to obtain $\mu(C = 1 \cap S = 1 \cap SH = 0 \cap PS = 0)$.

Note that the Bayesian network constructed in Theorem 0.1.4 depends on the ordering of the random variables. For example, the first element in the ordering is guaranteed to be a root of the Bayesian network. Thus, there are many Bayesian networks that represent a given probability measure. But not all orderings lead to equally useful Bayesian networks. In a well-designed Bayesian network, the nodes are ordered so that if X has a causal influence on Y, then X precedes Y in the ordering. This typically leads both to simpler Bayesian networks (in the sense of having fewer edges) and to conditional probabilities that are easier to obtain in practice. For example, it is possible to construct a Bayesian network that represents the same probability measure as (G_s, f_s) but has S as the root, by applying Theorem 0.1.4, with the ordering S, C, PS, SH (Exercise 0.4). However, not only does this network have more edges, the conditional probability tables require entries that are harder to elicit in practice. It is easier to elicit from medical experts the probability that someone will smoke given that at least one of his parents smoke ($\mu(S = 1|PS = 1)$) than the probability that at least one of a smoker's parents also smokes ($\mu(PS = 1|S = 1)$.

By definition, a node in a Bayesian network that represents a probability measure μ if it is independent of its nonancestors, given its parents. We are in general interested in knowing about other conditional independencies. A Bayesian network lets us read them off in an elegant way. There is a notion of *d-separation*, which I am about to define, with the property that $\mathbf{X}$ is conditionally independent of $\mathbf{Y}$ given $\mathbf{Z}$ if the nodes in $\mathbf{Z}$ d-separate every node in $\mathbf{X}$ from every node in $\mathbf{Y}$.

Now for the formal definition: A node X is *d-separated* (the *d* is for *directed*) from a node Y by a set of nodes $\mathbf{Z}$ if, for every *undirected path* from X to Y (an undirected path is a path that ignores the arrows; for example, $(SH,\ PS,\ S)$ is an undirected path from SH to S in G_s), there is a node Z' on the path such that either:

(a) $Z' \in \mathbf{Z}$ and there is an arrow on the path leading in to Z' and an arrow leading out;

(b) $Z' \in \mathbf{Z}$ and has both path arrows leading out; or

(c) Z' has both path arrows leading in, and neither Z' nor any of its descendants are in $\mathbf{Z}$.

Consider the graph G_s. The set $\{SH, S\}$ d-separates PS from C. One path from PS to C is blocked by SH and the other by S, according to clause (a), since both S and SH have an arrow leading in and one leading out. Similarly $\{PS\}$ d-separates SH from S. The (undirected) path

(SH, PS, S) is blocked by PS according to clause (b), and the undirected path (SH, C, S) is blocked by $C \notin \{PS\}$ according to clause (c). On the other hand, $\{PS, C\}$ does *not* d-separate SH from S, since there is no node on the path (SH, C, S) that satisfies any of (a), (b), or (c).

These examples may also help explain the intuition behind each of the clauses of the definition. Clause (a) is quite straightforward. Clearly if there is a directed path from PS to C, then PS can influence C, so PS and C are not independent. However, once we condition on $\{SH, S\}$, then all paths from PS to C are blocked, so C is conditionally independent of C given $\{SH, S\}$. The situation in clause (b) is exemplified by the edges leading out from PS to SH and S. Intuitively, smoking (S) and being exposed to second-hand smoke (SH) are not independent because they have a common cause, a parent smoking (PS). Finding out that $S = 1$ increases the likelihood that $PS = 1$ which in turn increases the likelihood that $SH = 1$. However, S and SH are conditionally independent given PS.

Clause (c) in the definition of d-separation seems the most puzzling. Why should the *absence* of a node in $\mathbf{Z}$ cause X and Y to be d-separated? Again this can be understood in terms of the graph G_s. Knowing $C = 1$ makes S and SH dependent, because, for example, finding out that $S = 0$ increases the likelihood that $SH = 1$; finding out that $S = 1$ decreases the likelihood that $S = 0$. Put another way, finding out that $C = 1$ makes S and SH become negatively correlated. Since they are both potential causes of $C = 1$, finding out that one holds decreases the likelihood that the other holds. To understand the role of descendants in clause (c), suppose that we add a node D (for "early death") to G_s with an edge from C to D. Finding out that $D = 1$ makes it more likely that $C = 1$, and thus also makes S and SH negatively correlated.

The following theorem says that d-separation completely characterizes conditional independence in Bayesian networks.

Theorem 0.1.5 *If $\mathbf{X}$ is d-separated from $\mathbf{Y}$ by $\mathbf{Z}$ in the Bayesian network G, then $I_\mu^{rv}(\mathbf{X}, \mathbf{Y}|\mathbf{Z})$ holds for all probability measures μ compatible with G. On the other hand, if $\mathbf{X}$ is* not *d-separated from $\mathbf{Y}$ by $\mathbf{Z}$, then there is a probability measure μ compatible with G such that $I_\mu^{rv}(\mathbf{X}, \mathbf{Y}|\mathbf{Z})$ does not hold.*

The first half says that d-separation really does imply conditional independence in Bayesian networks. Its proof again uses only properties CIRV1-5 and the fact that, by definition, $I_\mu^{rv}(\mathrm{NonDes}_G(X), X|\mathrm{Par}_G(X))$ holds for every μ compatible with G. The second half says that there is no more that we can say about conditional independence in qualitative Bayesian networks than can be said by d-separation.

One of the main criticisms of the use of probability in AI applications such as expert systems used to be that probability measures were too hard to work with. Bayesian networks allow us to deal with part of the criticism by providing a (potentially) compact representation of probability measures, one that experience has shown can be effectively constructed in realistic domains. For example, they have been used by PATHFINDER, a diagnostic expert system for lymph-node diseases. The first step in the use of Bayesian networks for PATHFINDER was for experts to decide what the relevant random variables were. At one stage in the design, they used 60 binary random variables to represent diseases (did the agent have the disease or not) and over 100 random variables for findings (symptoms and test results). Deciding on the appropriate vocabulary took 8 hours, constructing the appropriate qualitative Bayesian network took 35 hours, and making the assessments to fill in the cpts took another 40 hours. This is considered a perfectly acceptable length of time to spend in constructing a significant expert system.

But, of course, constructing the Bayesian network is only part of the problem. Once we have represented the probability measure, we want to reason about it. For example, a doctor using the PATHFINDER system will typically want to know the probability of a given disease given certain findings. Even if the disease is a parent of the symptom, computing the probability of the disease given the symptom requires some effort. For example, suppose that we want to compute $\mu_s(C = 1|SH = 1)$ for the unique probability measure μ_s compatible with (G_s, f_s). The cpt tells us that $\mu_s(C = 1|SH = 1 \cap S = 1) = .6$ and that $\mu(C = 1|SH = 1 \cap S = 0) = .1$. To compute $\mu_s(C = 1|SH = 1)$, we need to use the identity

$$\begin{aligned} \mu_s(C = 1|SH = 1) = \quad & \mu_s(C = 1|SH = 1 \cap S = 1) \times \mu_s(S = 1) + \\ & \mu_s(C = 1|SH = 1 \cap S = 0) \times \mu_s(S = 0). \end{aligned}$$

That means that we still have to compute $\mu_S(S = 1)$ and $\mu_S(S = 0)$. Efficient algorithms for such computations have been developed (and continue to be improved), which take advantage of the dag structure of a Bayesian network. It would take us too far afield here to go into the details; see the notes at the end of this chapter for references.

Exercises

0.1 Show using CIRV1-5 that $I_\mu^{rv}(\mathbf{X}, \mathbf{Y}|\mathbf{Z})$ iff $I_\mu^{rv}(\mathbf{X} - \mathbf{Z}, \mathbf{Y} - \mathbf{Z}|\mathbf{Z})$. This shows that without loss of generality, we can restrict attention to conditional independence statements $I_\mu^{rv}(\mathbf{X},\mathbf{Y}|\mathbf{Z})$ where $\mathbf{Z}$ is disjoint from both $\mathbf{X}$ and $\mathbf{Y}$.

0.2 Prove Proposition 0.1.3. Note that what requires proof here is that the required independence relations hold for (G, f) to represent the probability measure μ that it determines.

0.3 Complete the proof of Theorem 0.1.4 by showing that, for all nodes Y_k,

$$I_\mu^{rv}(\mathrm{NonDes}_G(Y_k), Y_k|\mathrm{Par}_G(Y_k)),$$

using CIRV1-5. (Hint: let $\mathbf{Z}_m = \mathrm{NonDes}_G(Y_k) \cap \{Y_1, \ldots, Y_m\}$. Prove by induction on m that $I_\mu^{rv}(\mathrm{NonDes}_G(Y_k), Y_k|\mathrm{Par}_G(Y_k))$, using CIRV1-5.)

0.4 Consider the quantitative Bayesian network (G_s, f_s) described in Section 0.1.

(a) Notice that $\{S, SH\}$ blocks both paths from PS to C. What does this say about the relationship between PS and C in probabilistic terms?

(b) Calculate $\mu_s(C = 1|PS = 1)$ for the unique probability measure μ_s represented by (G_s, f_s).

(c) Use the construction of Theorem 0.1.4 to construct two qualitative Bayesian networks representing μ_s, but with S as the root.

(d) Suppose that you believe that there is a gene (that can be inherited) that results both in a predisposition to smoke and to have cancer, but otherwise smoking and cancer are unrelated. Draw a Bayesian network describing these beliefs, using the variables PG (at least one parent has this gene), G (has this gene), PS (at least one parent smokes), S (smokes), and C (has cancer). Explain why you put in each edge that you did.