Chapter 2

Bayesian Decision Theory

2.11 *Bayesian Belief Networks

The methods we have described up to now are fairly general — all that we assumed, at base, was that we could parameterize the probability distributions by a vector $\boldsymbol{\theta}$. If we had prior information about $\boldsymbol{\theta}$, this too could be used. Sometimes our knowledge about a distribution is not directly expressed by a parameter vector, but instead about the statistical dependencies (or independencies) or the causal relationships among the component variables. (Recall that for some multidimensional distribution $p(\mathbf{x})$, if for two features we have $p(x_i, x_j) = p(x_i)p(x_j)$, we say those variables are statistically independent, as illustrated in Fig. 2.23.)

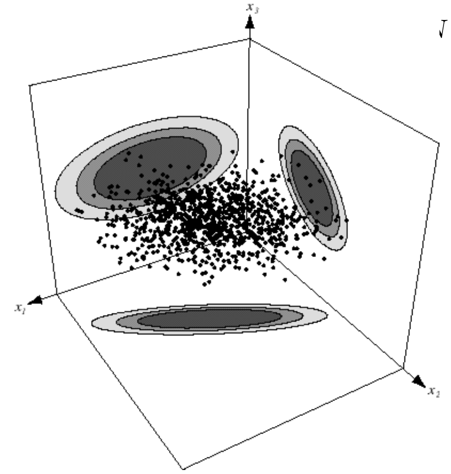


Figure 2.23: A three-dimensional distribution which obeys $p(x_1, x_3) = p(x_1)p(x_3)$; thus here x_1 and x_3 are statistically independent but the other feature pairs are not.

There are many such cases where we know — or can safely assume — which variables are or are not causally related, even if it may be more difficult to specify the precise probabilistic relationships among those variables. Suppose, for instance, we are describing the state of an automobile: temperature of the engine, pressure of the brake fluid, pressure of the air in the tires, voltages in the wires, and so on. Our basic knowledge of cars includes the fact that the oil pressure in the engine and the air pressure in a tire are *not* causally related while the engine temperature and oil temperature *are* causally related. Furthermore, we may know *several* variables that might influence another: The coolant temperature is affected by the engine temperature, the speed of the radiator fan (which blows air over the coolant-filled radiator), and so on. We shall now exploit this structural information when reasoning about the system and its variables.

We represent these causal dependencies graphically by means of *Bayesian belief nets*, also called *causal networks*, or simply *belief nets*. While these nets can represent continuous multidimensional distributions over their variables, they have enjoyed greatest application and success for discrete variables. For this reason, and because the calculations are simpler, we shall concentrate on the discrete case.

NODE

Each *node* (or unit) represents one of the system components, and here it takes on discrete values. We label nodes **A**, **B**, ... and their variables by the corresponding lowercase letter. Thus, while there are a discrete number of possible values of node **A** — for instance two, a_1 and a_2 — there may be continuous-valued *probabilities* on these discrete states. For example, if node **A** represents the automobile ignition switch —

$a_1 = on$, $a_2 = off$ — we might have $P(a_1) = 0.739, P(a_2) = 0.261$, or indeed any other probabilities. Each link in the net is directional and joins two nodes; the link represents the causal influence of one node upon another. Thus in the net in Fig. 2.24 **A** directly influences **D**. While **B** also influences **D**, such influence is indirect, through **C**. In considering a single node in a net, it is useful to distinguish the set of nodes immediately *before* that node — called its *parents* — and the set of those immediately *after* it — called its *children*. Thus in Fig. 2.24 the parents of **D** are **A** and **C** while the child of **D** is **E**.

PARENT

CHILD

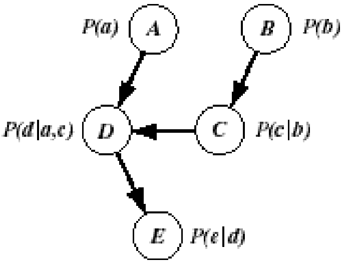


Figure 2.24: A belief network consists of nodes (labeled with uppercase bold letters) and their associated discrete states (in lowercase). Thus node **A** has states $\{a_1, a_2, ...\}$, which collectively are denoted simply **a**; node **B** has states $\{b_1, b_2, ...\}$, denoted **b**, and so forth. The links between nodes represent direct causal influence. For example the link from **A** to **D** represents the direct influence of **A** upon **D**. In this network, the variables at **B** may influence those at **D**, but only indirectly through their effect on **C**. Simple probabilities are denoted $P(\mathbf{a})$ and $P(\mathbf{b})$, and conditional probabilities $P(\mathbf{c}|\mathbf{b})$, $P(\mathbf{d}|\mathbf{a}, \mathbf{c})$ and $P(\mathbf{e}|\mathbf{d})$.

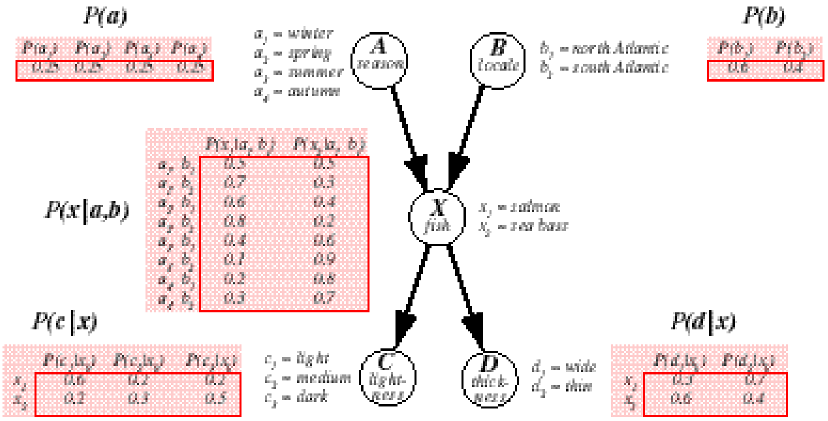
Suppose we have a belief net, complete with causal dependencies indicated by the topology of the links. Through a direct application of Bayes rule, we can determine the probability of any configuration of variables in the joint distribution. To proceed, though, we also need the *conditional probability tables*, which give the probability of any variable at a node for each conditioning event — that is, for the values of the variables in the parent nodes. Each row in a conditional probability table sums to 1, as its entries describe all possible cases for the variable. If a node has no parents, then the table just contains the prior probabilities of the variables. (There are sophisticated algorithms for learning the entries in such a table based on a data set of variable values. We shall not address such learning here as our main concern is how

CONDITIONAL PROBABILITY TABLE

to represent and reason about this probabilistic information.) Since the network and conditional probability tables contain all the information of the problem domain, we can use them to calculate any entry in the joint probability distribution, as illustrated in Example 4.

Example 4: Belief Network for fish

Consider again the problem of classifying fish, but now we want to incorporate more information than the measurements of the lightness and width. Imagine that a human expert has constructed the simple belief network in the figure, where node **A** represents the time of year and can have four values: a_1 = *winter*, a_2 = *spring*, a_3 = *summer* and a_4 = *autumn*. Node **B** represents the locale where the fish was caught: b_1 = *north Atlantic* and b_2 = *south Atlantic*. Node **X**, which represents the fish, has just two possible values: x_1 = *salmon* and x_2 = *sea bass*. **A** and **B** are the parents of the **X**. Similarly, our expert tells us that the children nodes of **X** represent lightness, **C**, with c_1 = *dark*, c_2 = *medium* and c_3 = *light*, as well as thickness, **D**, with d_1 = *thick* and d_2 = *thin*. Thus the season and the locale determine directly what kind of fish is likely to be caught; the season and locale also determine the fish's lightness and thickness, but only indirectly through their effect on **X**.



$P(a)$

$P(a_1)$	$P(a_2)$	$P(a_3)$	$P(a_4)$
0.25	0.25	0.25	0.25

$P(b)$

$P(b_1)$	$P(b_2)$
0.6	0.4

$P(x|a,b)$

	$P(x_1\|a_i,b_j)$	$P(x_2\|a_i,b_j)$
a_1, b_1	0.5	0.5
a_1, b_2	0.7	0.3
a_2, b_1	0.6	0.4
a_2, b_2	0.8	0.2
a_3, b_1	0.4	0.6
a_3, b_2	0.1	0.9
a_4, b_1	0.2	0.8
a_4, b_2	0.3	0.7

$P(c|x)$

	$P(c_1\|x_i)$	$P(c_2\|x_i)$	$P(c_3\|x_i)$
x_1	0.6	0.2	0.2
x_2	0.2	0.3	0.5

$P(d|x)$

	$P(d_1\|x_i)$	$P(d_2\|x_i)$
x_1	0.3	0.7
x_2	0.6	0.4

A simple belief net for the fish example. The season and the locale (and many other stochastic variables) determine directly the probability of the two different types of fish caught. The type of fish directly affects the lightness and thickness measured. The conditional probability tables quantifying these relationships are shown in pink.

Imagine that fishing boats go out throughout the year; then the probability distribution on the variables at **A** is uniform. Imagine, too, that boats generally spend more time in the north than the south Atlantic areas, specifically the probabilities that any fish came from those areas are 0.6 and 0.4, respectively. The other conditional probabilities are similarly given in the tables.

Now we can determine the value of any entry in the joint probability, for instance the probability that the fish was caught in the summer in the north Atlantic and is a sea bass that is dark and thin:

$$\begin{aligned} P(a_3, b_1, x_2, c_3, d_2) &= P(a_3)P(b_1)P(x_2|a_3, b_1)P(c_3|x_2)P(d_2|x_2) \\ &= 0.25 \times 0.6 \times 0.4 \times 0.5 \times 0.4 \\ &= 0.012. \end{aligned}$$

Note how the topology of the net is captured by the probabilities in the expression. Specifically, since **X** is the only node to have two parents, only the $P(x_2|\cdot,\cdot)$ term has two conditioning variables; the other conditional probabilities have just one each. The product of these probabilities corresponds to the assumption of statistical independence.

We now illustrate more fully how to exploit the causal structure in a Bayes belief net when determining the probability of its variables. Suppose we wish to determine the probability distribution over the variables $d_1, d_2, ...$ at **D** in the left network of Fig. 2.25 using the conditional probability tables and the network topology.

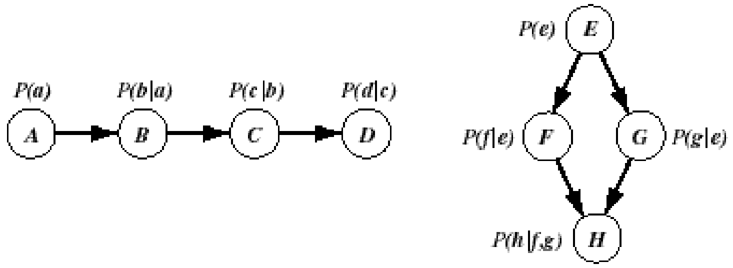


Figure 2.25: Two simple belief networks. The one on the left is a simple linear chain, the one on the right a simple loop. The conditional probability tables are indicated, for instance, as $P(\mathbf{h}|\mathbf{f}, \mathbf{g})$.

We evaluate this by summing the full joint distribution, $P(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d})$, over all the variables other than $\mathbf{d}$:

$$\begin{aligned} P(\mathbf{d}) &= \sum_{\mathbf{a},\mathbf{b},\mathbf{c}} P(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}) \\ &= \sum_{\mathbf{a},\mathbf{b},\mathbf{c}} P(\mathbf{a})P(\mathbf{b}|\mathbf{a})P(\mathbf{c}|\mathbf{b})P(\mathbf{d}|\mathbf{c}) \\ &= \underbrace{\sum_{\mathbf{c}} P(\mathbf{d}|\mathbf{c}) \underbrace{\sum_{\mathbf{b}} P(\mathbf{c}|\mathbf{b}) \underbrace{\sum_{\mathbf{a}} P(\mathbf{b}|\mathbf{a})P(\mathbf{a})}_{P(\mathbf{b})}}_{P(\mathbf{c})}}_{P(\mathbf{d})}. \end{aligned} \tag{96}$$

In Eq. 96 the summation variables can be split simply, and the intermediate terms have simple interpretations, as indicated. If we wanted the probability of a *particular* value of $\mathbf{D}$, for instance d_2, we would compute

$$P(d_2) = \sum_{\mathbf{a},\mathbf{b},\mathbf{c}} P(\mathbf{a}, \mathbf{b}, \mathbf{c}, d_2), \tag{97}$$

and proceed as above. In either case, the conditional probabilities are simple because of the simple linear topology of the network.

Now consider computing the probabilities of the variables at $\mathbf{H}$ in the network with the loop on the right of Fig. 2.25. Here we find

$$\begin{aligned} P(\mathbf{h}) &= \sum_{\mathbf{e},\mathbf{f},\mathbf{g}} P(\mathbf{e}, \mathbf{f}, \mathbf{g}, \mathbf{h}) \\ &= \sum_{\mathbf{e},\mathbf{f},\mathbf{g}} P(\mathbf{e})P(\mathbf{f}|\mathbf{e})P(\mathbf{g}|\mathbf{e})P(\mathbf{h}|\mathbf{f}, \mathbf{g}) \\ &= \sum_{\mathbf{e}} P(\mathbf{e})P(\mathbf{f}|\mathbf{e})P(\mathbf{g}|\mathbf{e}) \sum_{\mathbf{f},\mathbf{g}} P(\mathbf{h}|\mathbf{f}, \mathbf{g}). \end{aligned} \tag{98}$$

Note particularly that the expansion of the full sum differs somewhat from that in Eq. 96 because of the $P(\mathbf{h}|\mathbf{f}, \mathbf{g})$ term, which itself arises from the loop topology of the network.

EVIDENCE

Bayes belief nets are most useful in the case where are given the values of some of the variables — the *evidence* — and we seek to determine some particular configuration of other variables. Thus in our fish example we might seek to determine the probability that a fish came from the north Atlantic, given that it is springtime, and that the fish is a light salmon. (Notice that even here we may not be given the values of some variables such as the width of the fish.) In that case, the probability we seek is $P(b_1|a_2, x_1, c_1)$. In practice, we determine the values of several query variables (denoted collectively $\mathbf{X}$) given the evidence of all other variables (denoted $\mathbf{e}$) by

$$P(\mathbf{X}|\mathbf{e}) = \frac{P(\mathbf{x}, \mathbf{e})}{P(\mathbf{e})} = \alpha P(\mathbf{X}, \mathbf{e}), \tag{99}$$

where α is a constant of proportionality.

As an example, suppose we are given the Bayes belief net and conditional probability tables in Example 4. Suppose we know that a fish is light (c_1) and caught in the south Atlantic (b_2), but we do not know what time of year the fish was caught nor it thickness. How shall we classify the fish for minimum expected classification error? Of course we must compute the probability it is a salmon, and also the probability it is sea bass. We focus first on the relative probability the fish is a salmon given this evidence:

$$\begin{aligned} P(x_1|c_1, b_2) &= \frac{P(x_1, c_1, b_2)}{P(c_1, b_2)} \qquad (100) \\ &= \alpha \sum_{\mathbf{a},\mathbf{d}} P(x_1, \mathbf{a}, b_2, c_1, \mathbf{d}) \\ &= \alpha \sum_{\mathbf{a},\mathbf{d}} P(\mathbf{a})P(b_2)P(x_1|\mathbf{a}, b_2)P(c_1|x_1)P(\mathbf{d}|x_1) \\ &= \alpha P(b_2)P(c_1|x_1) \\ &\quad \times \left[\sum_{\mathbf{a}} P(\mathbf{a})P(x_1|\mathbf{a}, b_2)\right] \left[\sum_{\mathbf{d}} P(\mathbf{d}|x_1)\right] \\ &= \alpha P(b_2)P(c_1|x_1) \\ &\quad \times [P(a_1)P(x_1|a_1, b_2) + P(a_2)P(x_1|a_2, b_2) + P(a_3)P(x_1|a_3, b_2) + P(a_4)P(x_1|a_4, b_2)] \\ &\quad \times \underbrace{[P(d_1|x_1) + P(d_2|x_1)]}_{=1} \\ &= \alpha(0.4)(0.6)\,[(0.25)(0.7) + (0.25)(0.8) + (0.25)(0.1) + (0.25)(0.3)]\,1.0 \\ &= \alpha\ 0.114. \end{aligned}$$

Note that in this case,

$$\sum_{\mathbf{d}} P(\mathbf{d}|x_1) = 1, \tag{101}$$

that is, if we do not measure information corresponding to node **D**, the conditional probability table at **D** does not affect our results. A computation similar to that in Eq. 100 shows $P(x_2|c_1, b_2) = \alpha\ 0.066$. We normalize these probabilities (and hence eliminate α) and find $P(x_1|c_1, b_2) = 0.63$ and $P(x_2|c_1, b_2) = 0.27$. Thus given this evidence, we should classify this fish as a salmon.

When the dependency relationships among the features used by a classifier are unknown, we generally proceed by taking the simplest assumption, namely, that the features are conditionally independent given the category, that is,

$$P(\mathbf{x}|\mathbf{a}, \mathbf{b}) = P(\mathbf{x}|\mathbf{a})P(\mathbf{x}|\mathbf{b}). \tag{102}$$

NAIVE BAYES' RULE

In practice, this so-called *naive Bayes' rule* or *idiot Bayes' rule* often works quite well in practice, despite its manifest simplicity. Other approaches are to assume some functional form of conditional probability tables.

Belief nets have found increasing use in complicated problems such as medical diagnosis. Here the uppermost nodes (ones without their own parents) represent a fundamental biological agent such as the presence of a virus or bacteria. Intermediate nodes then describe diseases, such as flu or emphysema, and the lowermost nodes describe the symptoms, such as high temperature or coughing. A physician enters measured values into the net and finds the most likely disease or cause.