Ch2-4: Bayesian Belief Networks

Also known as Bayesian Networks, Causal Probabilistic Networks, Probabilistic Influence Diagrams, etc.)

- We assumed, 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.

Conditional Independence

Variables *A* and *B* are conditionally independent given *C* if any of the following hold:

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

Knowing *C* tells me everything about *B*. I don't gain anything by knowing *A* (either because *A* doesn't influence *B* or because knowing *C* provides all the information knowing *A* would give)

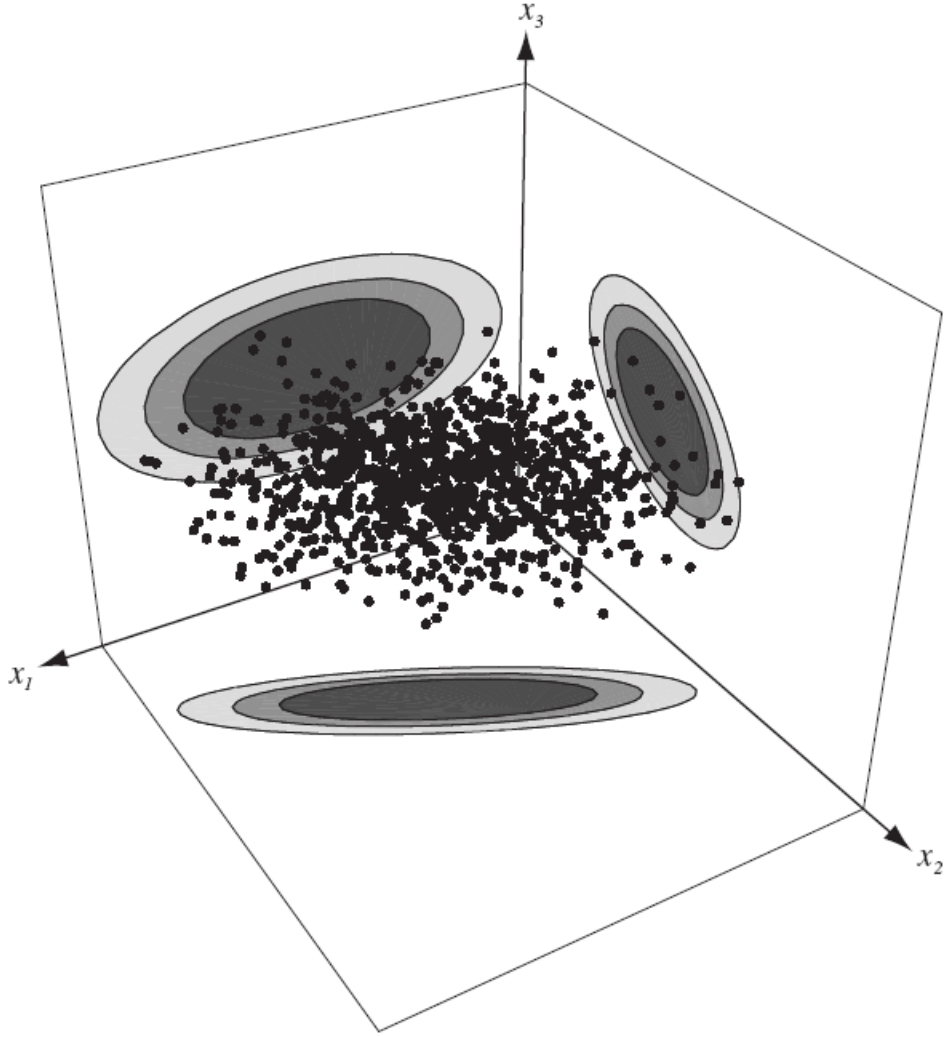
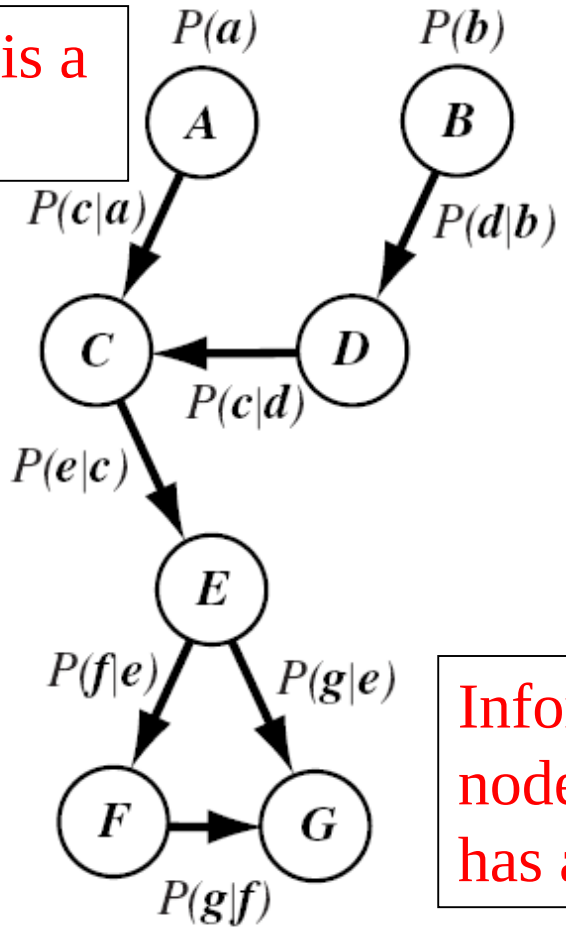


FIGURE 2.23. A three-dimensional distribution which obeys $p(x_1,x_3) = p(x_1)p(x_3)$; thus here x_1 and x_3 are statistically independent but the other feature pairs are not.

- 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.
- For instance, we know 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.
- We represent these causal dependencies graphically by means of *Bayesian belief nets*, also called causal networks, or simply *belief nets*.

- 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.
- Each link in the net is directional and joins two nodes; the link represents the causal influence of one node upon another.
- Nodes immediately before that node — called its *parents* $\mathcal{P}$ — and the set of those immediately parent after it — called its *children* $\mathcal{C}$.

Each node in the graph is a random variable



Informally, an arrow from node *X* to node *Y* means *X* has a direct influence on *Y*

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, . . . , denoted simply **a**; node **B** has states b_1, b_2, . . . , denoted **b**, and so forth. The links between nodes represent conditional probabilities. For example, $P(\mathbf{c}|\mathbf{a})$ can be described by a matrix whose entries are $P(c_i|a_j)$. From: Richard O. Duda, Peter E. Hart, and David

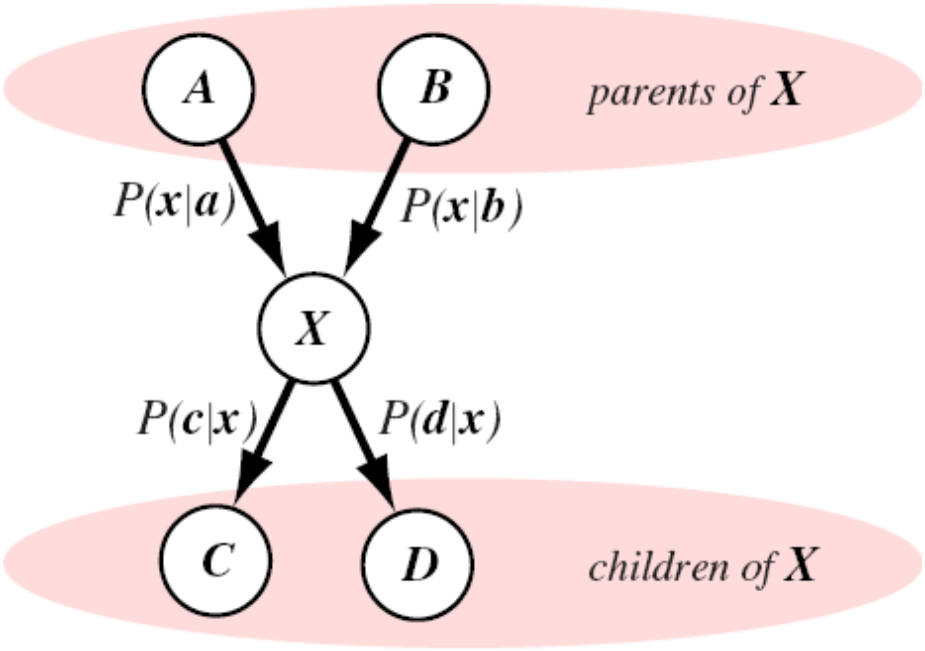


FIGURE 2.25. A portion of a belief network, consisting of a node **X**, having variable values *(*x_1, x_2, . . .*)*, its parents (**A** and **B**), and its children (**C** and **D**).

The belief

- The *belief* of a set of propositions $\mathbf{x} = (x_1, x_2, ...)$ on node ***X*** describes the relative belief probabilities of the variables given all the evidence **e** throughout the rest of the network, i.e., $P(\mathbf{x}|\mathbf{e})$ or $BEL(\mathbf{x})$.
- We can divide the dependency of the belief upon the parents and the children in the following way:

$$P(\mathbf{x}|\mathbf{e}) \propto P(\mathbf{e}^{\mathcal{C}}|\mathbf{x})P(\mathbf{x}|\mathbf{e}^{\mathcal{P}}),$$

where **e** represents all evidence (i.e., values of variables on nodes other than **X**), $\mathbf{e}^P$ the evidence on the parent nodes, and $\mathbf{e}^C$ the children nodes.

$$
\begin{aligned}
P(\mathbf{e}^{\mathcal{C}}|\mathbf{x}) &= P(\mathbf{e}_{\mathcal{C}_1}, \mathbf{e}_{\mathcal{C}_2}, ..., \mathbf{e}_{\mathcal{C}_{|\mathcal{C}|}}|\mathbf{x}) \\
&= P(\mathbf{e}_{\mathcal{C}_1}|\mathbf{x})P(\mathbf{e}_{\mathcal{C}_2}|\mathbf{x}) \cdots P(\mathbf{e}_{\mathcal{C}_{|\mathcal{C}|}}|\mathbf{x}) \\
&= \prod_{j=1}^{|\mathcal{C}|} P(\mathbf{e}_{\mathcal{C}_j}|\mathbf{x}),
\end{aligned}
$$

where C_j represents the *j*th child node and $\mathbf{e}_{Cj}$ the values of the probabilities of its states. |*C*| denotes the *cardinality* of set *C* — the number of elements in the set — a convenient notation for indicating the full range of summations or products.

This equation simply states that the probability of a given set of states throughout all the children nodes of **X** is the product of the (independent) probabilities in the individual children nodes. For instance

$$P(\mathbf{e}_{\mathbf{C}}, \mathbf{e}_{\mathbf{D}}|\mathbf{x}) = P(\mathbf{e}_{\mathbf{C}}|\mathbf{x})P(\mathbf{e}_{\mathbf{D}}|\mathbf{x}).$$

Incorporating evidence from parent nodes is a bit more subtle. We have:

$$
\begin{aligned}
P(\mathbf{x}|\mathbf{e}^{\mathcal{P}}) &= P(\mathbf{x}|\mathbf{e}_{\mathcal{P}_1}, \mathbf{e}_{\mathcal{P}_2}, ..., \mathbf{e}_{\mathcal{P}_{|\mathcal{P}|}}) \\
&= \sum_{all\ \ i,j,...,k} P(\mathbf{x}|\mathcal{P}_{1i}, \mathcal{P}_{2j}, ..., \mathcal{P}_{|\mathcal{P}|k}) P(\mathcal{P}_{1i}, \mathcal{P}_{2j}, ..., \mathcal{P}_{|\mathcal{P}|k}|\mathbf{e}_{\mathcal{P}_1}, ..., \mathbf{e}_{\mathcal{P}_{|\mathcal{P}|}}) \\
&= \sum_{all\ \ i,j,...,k} P(\mathbf{x}|\mathcal{P}_{1i}, \mathcal{P}_{2j}, ..., \mathcal{P}_{|\mathcal{P}|k}) P(\mathcal{P}_{1i}|\mathbf{e}_{\mathcal{P}_1}) \cdots P(\mathcal{P}_{|\mathcal{P}|k}|\mathbf{e}_{\mathcal{P}_{|\mathcal{P}|k}}), \quad (82)
\end{aligned}
$$

Here $\mathcal{P}_{mn}$ denotes a particular value for state n on parent node $\mathcal{P}_m$.

For the purposes of clarity and for computing **x**, each term at the extreme right, $P(\mathcal{P}_{1i}|\mathbf{e}_{P_1})$ can be considered to be $P(\mathcal{P}_{1i})$ — the probability of state i on the first parent node.

For the sake of computing the probabilities at **X** we temporarily ignore the dependencies beyond the parents and children of **X**.

Thus we rewrite Eq. 82 as

$$P(\mathbf{x}|\mathbf{e}^{\mathcal{P}}) = \sum_{all\ \mathcal{P}_{mn}} P(\mathbf{x}|\mathcal{P}_{mn}) \prod_{i=1}^{|\mathcal{P}|} P(\mathcal{P}_i|\mathbf{e}_{\mathcal{P}_i})$$

We put these results together for the general case with |*P*| parent nodes and |*C*| children nodes and find

$$P(\mathbf{x}|\mathbf{e}) \propto \underbrace{\prod_{j=1}^{|\mathcal{C}|} P(\mathbf{e}_{\mathcal{C}_j}|\mathbf{x})}_{P(\mathbf{e}^{\mathcal{C}}|\mathbf{x})} \underbrace{\left[\sum_{all\ \mathcal{P}_{mn}} P(\mathbf{x}|\mathcal{P}_{mn}) \prod_{i=1}^{|\mathcal{P}|} P(\mathcal{P}_i|\mathbf{e}_{\mathcal{P}_i}) \right]}_{P(\mathbf{x}|\mathbf{e}^{\mathcal{P}})} .$$

The first factor is due the children (the product of children's independent likelihoods). The second is the sum over all possible configurations of states on the parent nodes of the prior probabilities of their values and the conditional probabilities of the **x** variables given those parent values. The final values must be normalized to represent probabilities.

The conditional probability tables

- 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.

- 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.

Belief network for fish

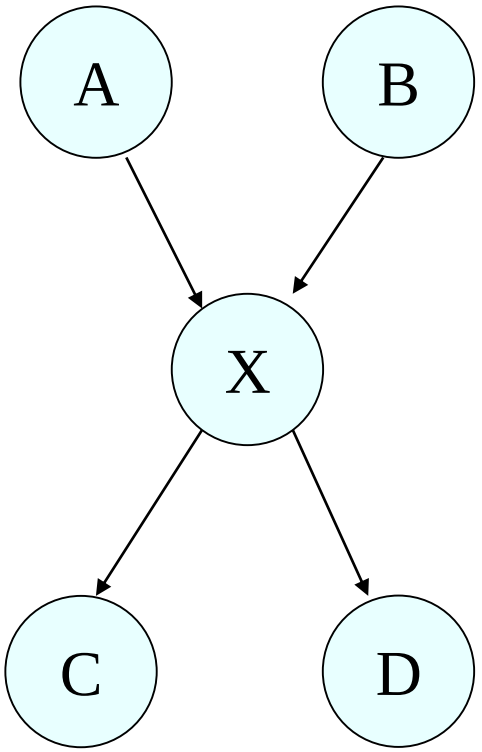
A season: a_1 = *winter*, a_2 = *spring*, a_3 = *summer* and a_4 = *autumn*.

B Locale: b_1 = *north Atlantic* and b_2 = *south Atlantic*.

X fish**:** x_1 = *salmon* and x_2 = *sea bass*

C lightness: c_1 = *dark*, c_2 =*medium* and c_3 = *light*.

D thickness: d_1 = *thick* and d_2 = *thin*.



The probability matrixes

$$P(x_i|a_j):\quad \begin{array}{l} \\ \text{winter} \\ \text{spring} \\ \text{summer} \\ \text{autumn} \end{array} \begin{array}{c} \begin{array}{cc} \text{salmon} & \text{sea bass} \end{array} \\ \begin{pmatrix} .9 & .1 \\ .3 & .7 \\ .4 & .6 \\ .8 & .2 \end{pmatrix} \end{array}$$

$$P(x_i|b_j):\quad \begin{array}{l} \\ \text{north} \\ \text{south} \end{array} \begin{array}{c} \begin{array}{cc} \text{salmon} & \text{sea bass} \end{array} \\ \begin{pmatrix} .65 & .35 \\ .25 & .75 \end{pmatrix} \end{array}$$

each row is normalized, like $P(x_1|a_1) + P(x_2|a_1) = 1.$

The conditional probabilities for the variables in the children nodes are as follows:

$$P(c_i|x_j): \quad \begin{array}{c} \\ \text{salmon} \\ \text{sea bass} \end{array} \begin{array}{c} \begin{array}{ccc} \text{light} & \text{medium} & \text{dark} \end{array} \\ \begin{pmatrix} .33 & .33 & .34 \\ .8 & .1 & .1 \end{pmatrix} \end{array}$$

$$P(d_i|x_j): \quad \begin{array}{c} \\ \text{salmon} \\ \text{sea bass} \end{array} \begin{array}{c} \begin{array}{cc} \text{wide} & \text{thin} \end{array} \\ \begin{pmatrix} .4 & .6 \\ .95 & .05 \end{pmatrix} \end{array}$$

Now we turn to the problem of using such a belief net to infer the identity of a fish. We have no direct information about the identity of the fish, and thus $P(x_1) = P(x_2) = 0.5$. This might be a reasonable starting point, expressing our lack of knowledge of the identity of the fish. Our goal now is to estimate the probabilities $P(x_1|\mathbf{e})$ and $P(x_2|\mathbf{e})$.

$$
\begin{aligned}
P(x_1) &= \sum_{i,j,k,l} P(x_1, a_i, b_j, c_k, d_l) \\
&= \sum_{i,j,k,l} P(a_i)P(b_j)P(x_1|a_i, b_j)P(c_k|x_1)P(d_l|x_1) \\
&= \sum_{i,j} P(a_i)P(b_j)P(x_1|a_i, b_j) \\
&= (0.25)(0.5) \sum_{i,j} P(x_1|a_i, b_j) \\
&= (0.25)(0.5)(0.9 + 0.3 + 0.4 + 0.7 + 0.8 + 0.2 + 0.1 + 0.6) \\
&= 0.5,
\end{aligned}
$$

and thus $P(x_1) = P(x_2)$, as we would expect.

Now we collect evidence for each node, $\{e_{\mathbf{A}}, e_{\mathbf{B}}, e_{\mathbf{C}}, e_{\mathbf{D}}\}$, assuming they are independent of each other.

If it is winter, thus $P(a_1|e_{\mathbf{A}}) = 1$ and $P(a_i|e_{\mathbf{A}}) = 0$ for $i = 2, 3, 4$.

we assume that $P(b_1|e_{\mathbf{B}}) = 0.2$ and $P(b_2|e_{\mathbf{B}}) = 0.8$.

We measure the fish and find that it is fairly light, and set by hand to be $P(e_{\mathbf{C}}|c_1) = 1$, $P(e_{\mathbf{C}}|c_2) = 0.5$, and $P(e_{\mathbf{C}}|c_3) = 0$.

Suppose that due to occlusion, we cannot measure the width of the fish; we thus set $P(e_{\mathbf{D}}|d_1) = P(e_{\mathbf{D}}|d_2)$.

$$
\begin{aligned}
P_{\mathcal{P}}(x_1) \quad &\propto \quad P(x_1|a_1,b_1)P(a_1)P(b_1) \\
&\quad\quad +P(x_1|a_1,b_2)P(a_1)P(b_2) \\
&\quad\quad +P(x_1|a_2,b_1)P(a_2)P(b_1) \\
&\quad\quad +P(x_1|a_2,b_2)P(a_2)P(b_2) \\
&\quad\quad +P(x_1|a_3,b_1)P(a_3)P(b_1) \\
&\quad\quad +P(x_1|a_3,b_2)P(a_3)P(b_2) \\
&\quad\quad +P(x_1|a_4,b_1)P(a_4)P(b_1) \\
&\quad\quad +P(x_1|a_4,b_2)P(a_4)P(b_2) \\
&= \quad 0.82.
\end{aligned}
$$

A similar calculation gives $P_P(x_2) \propto 0.18$.

We now turn to the children nodes we find

$$
\begin{aligned}
P_C(x_1) &\propto P(e_{\mathbf{C}}|x_1)P(e_{\mathbf{D}}|x_1) \\
&= [P(e_{\mathbf{C}}|c_1)P(c_1|x_1) + P(e_C|c_2)P(c_2|x_1) + P(e_{\mathbf{C}}|c_3)P(c_3|x_1)] \\
&\quad \times [P(e_{\mathbf{D}}|d_1)P(d_1|x_1) + P(e_{\mathbf{D}}|d_2)P(d_2|x_1)] \\
&= [(1.0)(0.33) + (0.5)(0.33) + (0)(0.34)] \times [(1.0)(0.4) + (1.0)(0.6)] \\
&= 0.495.
\end{aligned}
$$

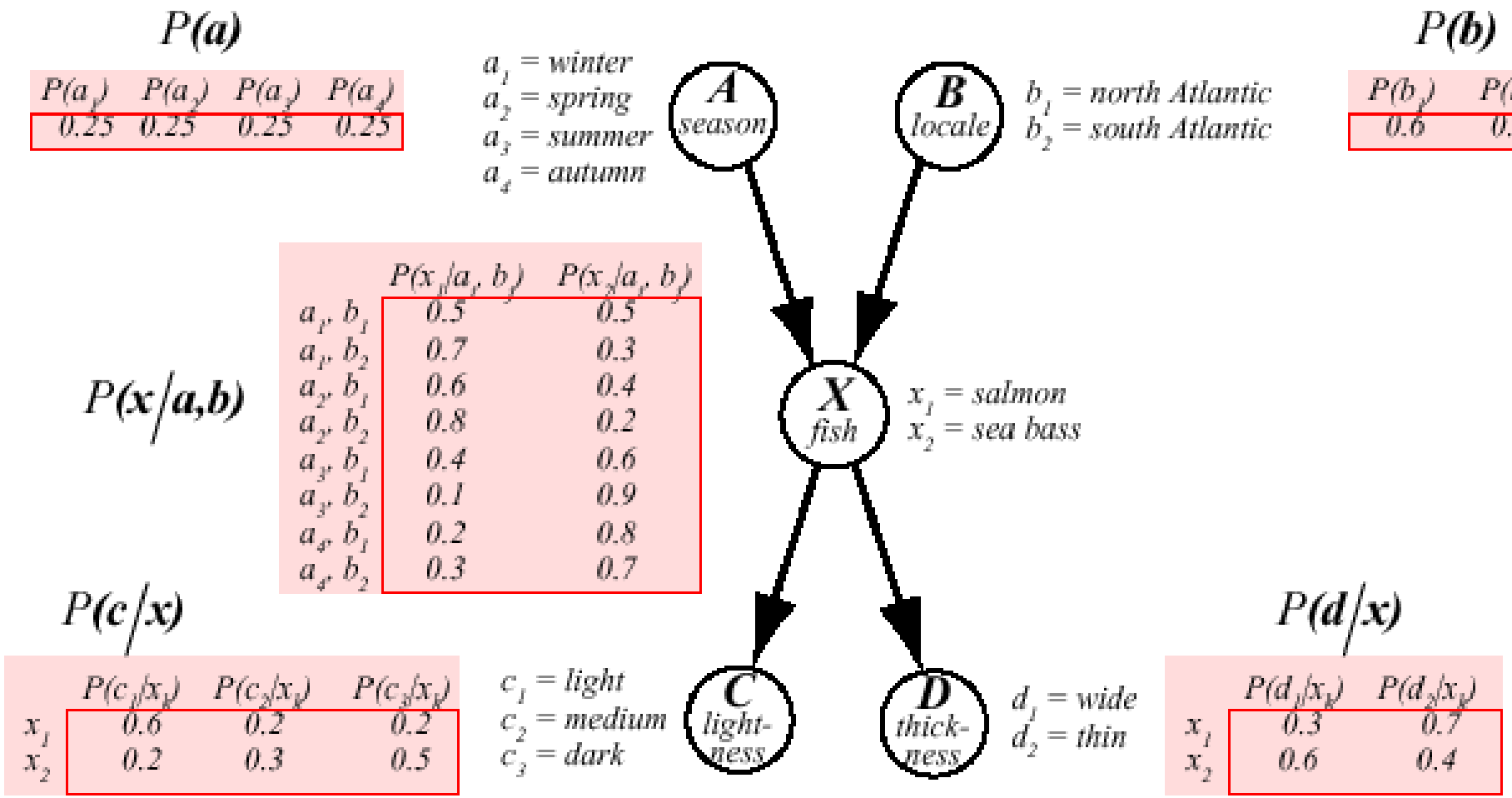
A similar calculation gives $P_C(x_2) \propto 0.85$. We put these estimates together as products $P(x_i) \propto P_C(x_i)P_P(x_i)$ and renormalize (i.e., divide by their sum). Thus our final estimates for node **X** are:

$$
P(x_1|\mathbf{e}) = \frac{(0.82)(0.495)}{(0.82)(0.495) + (0.18)(0.85)} = 0.726
$$

$$
P(x_2|\mathbf{e}) = \frac{(0.18)(0.85)}{(0.82)(0.495) + (0.18)(0.85)} = 0.274.
$$

Thus given all the evidence throughout the belief net, the most probable outcome is x_1 = *salmon*.

Example 4: Belief Network for fish



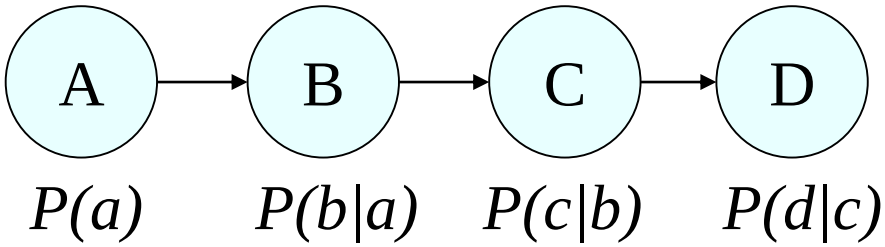
Conditional Probability Tables

- There exist algorithms for learning these probabilities from data…
- We can compute the probability of any configuration of variables in the joint density distribution:

- 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:

  $P(a_3, b_1, x_2, c_3, d_2) = P(a_3) \times P(b_1) \times P(x_2|a_3, b_1) \times P(c_3|x_2) \times P(d_2|x_2) = 0.25 \times 0.6 \times 0.6 \times 0.5 \times 0.4 = 0.018.$

- 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 next Figure using the conditional probability tables and the network topology.



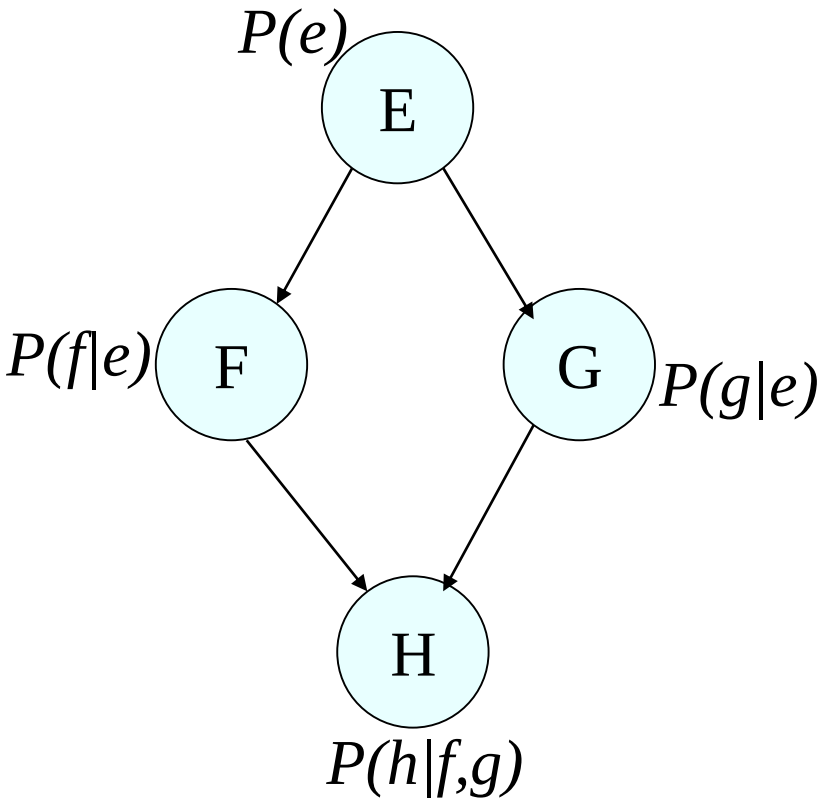


Figure 2.25: Two simple belief networks. The one on the left is a simple linear chain, the one on the right a simple loop. The conditional probability tables are indicated, for instance, as *P(h|f,g)*.

- We evaluate this by summing the full joint distribution, ***P(a, b, c, d)***, over all the variables other than ***d***:

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

If we wanted the probability of a particular value of **D**, for instance d_2, we would compute

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

Now consider computing the probabilities of the variables at **H** in the network with the loop on the right of Fig. 2.25. Here we find

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

▲

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.

In practice, we determine the values of several query variables (denoted collectively **X**) given the evidence of all other variables (denoted **e**) by

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

where α is a constant of proportionality.

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 its 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 a sea bass.

$$
\begin{aligned}
P(x_1|c_1,b_2) &= \frac{P(x_1,c_1,b_2)}{P(c_1,b_2)} \\
&= \alpha \sum_{\mathbf{a},\mathbf{d}} P(x_1,\mathbf{a},b_2,c_1,\mathbf{d}) \\
&= \alpha \sum_{\mathbf{a},\mathbf{d}} P(\mathbf{a})P(b_2)P(x_1|\mathbf{a},b_2)P(c_1|x_1)P(\mathbf{d}|x_1) \\
&= \alpha P(b_2)P(c_1|x_1) \\
&\quad \times \left[\sum_{\mathbf{a}} P(\mathbf{a})P(x_1|\mathbf{a},b_2)\right]\left[\sum_{\mathbf{d}} P(\mathbf{d}|x_1)\right]
\end{aligned}
$$

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

Note that in this case,

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

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 shows $P(x_2|c_1, b_2) = \alpha$ 0.042. We normalize these probabilities (and hence eliminate α) and find $P(x_1|c_1, b_2) =$ 0.73 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,

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

In practice, this so-called *naive Bayes' rule* or *idiot Bayes' rule* often works quite well in practice, despite its manifest simplicity.