We can write the following general marginalization (summing out) rule for any sets of variables $Y$ and $Z$:

$$P(Y) = \sum_{z \in Z} P(Y, z)$$

Here $\sum_{z \in Z}$ means summing over all the possible combinations of values of the set of variables $Z$.

For example:

$$P(\text{Cavity}) = \sum_{z \in \{\text{Catch, Toothache}\}} P(\text{Cavity}, z)$$

was just used above.

A variant of this rule involves conditional probabilities instead of joint probabilities, using the product rule:

$$P(Y) = \sum_z P(Y \mid z) P(z)$$

This rule is called *conditioning*.

Marginalization and conditioning turn out to be useful rules for all kinds of derivations involving probability expressions.
Computing a conditional probability

\[
P(\text{cavity} | \text{toothache}) = \frac{P(\text{cavity} \land \text{toothache})}{P(\text{toothache})} = \frac{(0.108 + 0.012)}{(0.108 + 0.012 + 0.016 + 0.064)} = \frac{0.12}{0.2} = 0.6
\]

Respectively

\[
P(\neg \text{cavity} | \text{toothache}) = \frac{(0.016 + 0.064)}{0.2} = 0.4
\]

The two probabilities sum up to one, as they should

\[
\begin{array}{|c|c|c|c|}
\hline
& \text{toothache} & \neg \text{toothache} \\
\hline
\text{catch} & 0.108 & 0.012 & 0.072 & 0.008 \\
\text{\neg catch} & 0.016 & 0.064 & 0.144 & 0.576 \\
\hline
\end{array}
\]

- \(1/P(\text{toothache}) = 1/0.2 = 5\) is a normalization constant ensuring that the distribution \(P(\text{Cavity} | \text{toothache})\) adds up to 1

- Let \(\alpha\) denote the normalization constant

\[
P(\text{Cavity} | \text{toothache}) = \alpha P(\text{Cavity, toothache})
= \alpha [P(\text{Cavity, toothache, catch})
+ P(\text{Cavity, toothache, catch\neg})]
= \alpha [(0.108, 0.016) + (0.012, 0.064)]
= \alpha (0.12, 0.08)
= (0.6, 0.4)
\]

- In other words, we can calculate the conditional probability distribution without knowing \(P(\text{toothache})\) using normalization
• More generally:
  • we need to find out the distribution of the query variable \( X \) (Cavity),
  • evidence variables \( E \) (Toothache) have observed values \( e \), and
  • the remaining unobserved variables are \( Y \) (Catch)

• Evaluation of a query:
  \[
P(X \mid e) = \alpha P(X, e) = \alpha \sum_y P(X, e, y),
\]
  where the summation is over all possible \( ys \); i.e., all possible combinations of values of the unobserved variables \( Y \)

• \( P(X, e, y) \) is simply a subset of the joint probability distribution of variables \( X, E, \) and \( Y \)
• \( X, E, \) and \( Y \) together constitute the complete set of variables for the domain
• Given the full joint distribution to work with, the equation in the previous slide can answer probabilistic queries for discrete variables
• It does not scale well
• For a domain described by \( n \) Boolean variables, it requires an input table of size \( O(2^n) \) and takes \( O(2^n) \) time to process the table
• In realistic problems the approach is completely impractical
13.4 Independence

- If we expand the previous example with a fourth random variable *Weather*, which has four possible values, we have to copy the table of joint probabilities four times to have 32 entries together.
- Dental problems have no influence on the weather, hence:
  \[
P(\text{Weather} = \text{cloudy} \mid \text{toothache, catch, cavity}) = P(\text{Weather} = \text{cloudy})
  \]
- By this observation and product rule
  \[
P(\text{toothache, catch, cavity, Weather} = \text{cloudy}) = P(\text{Weather} = \text{cloudy}) P(\text{toothache, catch, cavity})
  \]
- A similar equation holds for the other values of the variable *Weather*, and hence
  \[
P(\text{toothache, catch, cavity, Weather}) = P(\text{toothache, catch, cavity}) P(\text{Weather})
  \]
- The required joint distribution tables have 8 and 4 elements.
- Propositions \(a\) and \(b\) are independent if
  \[
P(a \mid b) = P(a) \iff P(b \mid a) = P(b) \iff P(a \land b) = P(a)P(b)
  \]
- Respectively variables \(X\) and \(Y\) are independent of each other if
  \[
P(X \mid Y) = P(X) \iff P(Y \mid X) = P(Y) \iff P(X,Y) = P(X)P(Y)
  \]
- Independent coin flips:
  \[
P(C_1, ..., C_n) \text{ can be represented as the product of } n \text{ single-variable distributions } P(C_i)
  \]
13.5 Bayes’ Rule and Its Use

- By the product rule $P(a \land b) = P(a \mid b) \ P(b)$ and the commutativity of conjunction $P(a \land b) = P(b \mid a) \ P(a)$
- Equating the two right-hand sides and dividing by $P(a)$, we get the Bayes’ rule (or law or theorem)
  $$P(b \mid a) = \frac{P(a \mid b) \ P(b)}{P(a)}$$
- The more general case of multivalued variables $X$ and $Y$ conditionalized on some background evidence $e$
  $$P(Y \mid X, e) = \frac{P(X \mid Y, e) \ P(Y \mid e)}{P(X \mid e)}$$
- Using normalization Bayes’ rule can be written as
  $$P(Y \mid X) = \alpha P(X \mid Y) \ P(Y)$$

Meningitis patients have a stiff neck 70% of the time
$$P(s \mid m) = 0.7$$

The prior probability of meningitis is $1/50\,000$:
$$P(m) = 1/50\,000$$

Prior probability of that any patient has stiff neck is $1\%$
$$P(s) = 0.01$$

What is the probability that a patient complaining about a stiff neck has meningitis?
$$P(m \mid s) = \frac{P(s \mid m) \ P(m)}{P(s)} = \frac{0.7 \times 1/50000}{0.01} = 0.0014$$
Perhaps the doctor knows that a stiff neck implies meningitis in 1 out of 5,000 cases. The doctor, hence, has quantitative information in the diagnostic direction from symptoms to causes, and no need to use Bayes’ rule. Unfortunately, diagnostic knowledge is often more fragile than causal knowledge. If there is a sudden epidemic of meningitis, the unconditional probability of meningitis $P(m)$ will go up. The causal information $P(s \mid m)$, however, is unaffected by the epidemic. The doctor who derived diagnostic probability $P(m \mid s)$ directly from statistical observation of patients before the epidemic will have no idea how to update the value. The doctor who computes $P(m \mid s)$ from the other three values will see $P(m \mid s)$ go up proportionally with $P(m)$.

All modern probabilistic inference systems are based on the use of Bayes’ rule. On the surface the relatively simple rule does not seem very useful. However, as the previous example illustrates, Bayes’ rule gives a chance to apply existing knowledge. We can avoid assessing the probability of the evidence – $P(s)$ – by instead computing a posterior probability for each value of the query variable – $m$ and $\neg m$ – and then normalizing the result

$$P(M \mid s) = \alpha (P(s \mid m) P(m), P(s \mid \neg m) P(\neg m))$$

Thus, we need to estimate $P(s \mid \neg m)$ instead of $P(s)$.

Sometimes easier, sometimes harder.
When a probabilistic query has more than one piece of evidence the approach based on full joint probability will not scale up

\[ P(\text{Cavity} \mid \text{toothache} \land \text{catch}) = a(0.108, 0.016) \approx (0.871, 0.129) \]

Neither will applying Bayes’ rule scale up in general

\[ a P(\text{toothache} \land \text{catch} \mid \text{Cavity}) P(\text{Cavity}) \]

We would need variables to be independent, but variable Toothache and Catch obviously are not:

- if the probe catches in the tooth, it probably has a cavity and that probably causes a toothache
- Each is directly caused by the cavity, but neither has a direct effect on the other
- Catch and Toothache are conditionally independent given Cavity

Conditional independence:

\[ P(\text{toothache} \land \text{catch} \mid \text{Cavity}) = P(\text{toothache} \mid \text{Cavity}) P(\text{catch} \mid \text{Cavity}) \]

Plugging this into Bayes’ rule yields

\[ P(\text{Cavity} \mid \text{toothache} \land \text{catch}) = a P(\text{Cavity}) P(\text{toothache} \mid \text{Cavity}) P(\text{catch} \mid \text{Cavity}) \]

Now we only need three separate distributions

The general definition of conditional independence of variables \( X \) and \( Y \), given a third variable \( Z \) is

\[ P(X,Y \mid Z) = P(X \mid Z) P(Y \mid Z) \]

Equivalently, \( P(X \mid Y,Z) = P(X \mid Z) \) and \( P(Y \mid X,Z) = P(Y \mid Z) \)
If all effects are conditionally independent given a single cause, the exponential size of knowledge representation is cut to linear.

A probability distribution is called a naïve Bayes (NB) model if all effects $E_1, ..., E_n$ are conditionally independent, given a single cause $C$.

The full joint probability distribution can be written as

$$P(C, E_1, ..., E_n) = P(C) \prod_i P(E_i | C)$$

It is often used as a simplifying assumption even in cases where the effect variables are not conditionally independent given the cause variable.

In practice, NB systems can work surprisingly well, even when the independence assumption is not true.

---

14 PROBABILISTIC REASONING

A Bayesian network is a directed graph in which each node is annotated with quantitative probability information:

1. A set of random variables makes up the nodes of the network. Variables may be discrete or continuous.
2. A set of directed links (arrows) connects pairs of nodes. If there is an arrow from node $X$ to node $Y$, then $X$ is said to be a parent of $Y$.
3. Each node $X_i$ has a conditional probability distribution $P(X_i | Parents(X_i))$.
4. The graph has no directed cycles and hence is a directed, acyclic graph DAG.
The intuitive meaning of arrow $X \rightarrow Y$ is that $X$ has a direct influence on $Y$

The topology of the network — the set of nodes and links — specifies the conditional independence relations that hold in the domain

Once the topology of the Bayesian network has been laid out, we need only specify a conditional probability distribution for each variable, given its parents.

The combination of the topology and the conditional distributions specify implicitly the full joint distribution for all the variables.

---

**Example (from LA)**

- A new burglar alarm ($A$) has been installed at home.
- It is fairly reliable at detecting a burglary ($B$), but also responds on occasion to minor earthquakes ($E$).
- Neighbors John ($J$) and Mary ($M$) have promised to call you at work when they hear the alarm.
- John always calls when he hears the alarm, but sometimes confuses the telephone ringing with the alarm and calls then, too.
- Mary, on the other hand, likes loud music and sometimes misses the alarm altogether.
- Given the evidence of who has or has not called, we would like to estimate the probability of a burglary.
The topology of the network indicates that:
- Burglary and earthquakes affect the probability of the alarm’s going off.
- Whether John and Mary call depends only on the alarm.
- They do not perceive any burglaries directly, they do not notice minor earthquakes, and they do not confer before calling.

Mary listening to loud music and John confusing phone ringing to the sound of the alarm can be read from the network only implicitly as uncertainty associated to calling at work.

The probabilities actually summarize a potentially infinite set of circumstances:
- The alarm might fail to go off due to high humidity, power failure, dead battery, cut wires, a dead mouse stuck inside the bell, etc.
- John and Mary might fail to call and report an alarm because they are out to lunch, on vacation, temporarily deaf, passing helicopter, etc.
The *conditional probability tables* in the network give the probabilities for the values of the random variable depending on the combination of values for the parent nodes.

Each row must sum to 1, because the entries represent exhaustive set of cases for the variable.

Above all variables are Boolean, and therefore it is enough to know that the probability of a true value is $p$, the probability of false must be $1 - p$.

In general, a table for a Boolean variable with $k$ parents contains $2^k$ independently specifiable probabilities.

A variable with no parents has only one row, representing the prior probabilities of each possible value of the variable.

### 14.2 Semantics of Bayesian Networks

Every entry in the full joint probability distribution can be calculated from the information in a Bayesian network.

A generic entry in the joint distribution is the probability of a conjunction of particular assignments to each variable:

$$ P(X_1 = x_1 \land \ldots \land X_n = x_n), \text{ abbreviated as } P(x_1, \ldots, x_n) $$

The value of this entry is

$$ P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i \mid \text{parents}(X_i)) , $$

where $\text{parents}(X_i)$ denotes the specific values of the variables $\text{Parents}(X_i)$.

- $P(j \land m \land a \land \neg b \land \neg e)$

  $$ = P(j \mid a) \cdot P(m \mid a) \cdot P(a \mid \neg b \land 
  \neg e) \cdot P(\neg b) \cdot P(\neg e) $$

  $$ = 0.90 \times 0.70 \times 0.001 \times 0.999 \times 0.998 $$

  $$ = 0.000628 $$
Constructing Bayesian networks

- We can rewrite an entry in the joint distribution $P(x_1, \ldots, x_n)$, using the product rule, as
  
  $$P(x_n | x_{n-1}, \ldots, x_1) P(x_{n-1}, \ldots, x_1)$$

- Repeat the process, reducing each conjunctive probability to a conditional probability and a smaller conjunction.

- We end up with one big product:
  
  $$P(x_n | x_{n-1}, \ldots, x_1) P(x_{n-1} | x_{n-2}, \ldots, x_1) \cdots P(x_2 | x_1) P(x_1)$$

  $$= \prod_{i=1}^{n} P(x_i | x_{i-1}, \ldots, x_1)$$

- This identity holds true for any set of random variables and is called the chain rule.

The specification of the joint distribution is thus equivalent to the general assertion that, for every variable $X_i$ in the network

$$P(X_i | X_{i-1}, \ldots, X_1) = P(X_i | \text{Parents}(X_i))$$

provided that $\text{Parents}(X_i) \subseteq \{X_{i-1}, \ldots, X_1\}$.

- The last condition is satisfied by labeling the nodes in any order that is consistent with the partial order implicit in the graph structure.

- Each node is required to be conditionally independent of its predecessors in the node ordering, given its parents.

- We need to choose as parents of a node all those nodes that directly influence the value of the variable.
For example, *MaryCalls* is certainly influenced by whether there is a *Burglary* or an *Earthquake*, but not directly influenced.

These events influence Mary’s calling behavior only through their effect on the alarm.

We believe the following conditional independence statement to hold:

\[ P(MaryCalls \mid JohnCalls, Alarm, Earthquake, Burglary) = P(MaryCalls \mid Alarm) \]

A Bayesian network is a complete and nonredundant representation of the domain.

It is also a more compact representation than the full joint distribution, due to its locally structured properties.

Each subcomponent interacts directly with only a bounded number of other components, regardless of the total number of components.

Local structure is usually associated with linear rather than exponential growth in complexity.

If each of the \( n \) random variables is influenced by at most \( k \) others, then specifying each CPT will require at most \( 2^k \) numbers.

Altogether \( n2^k \) numbers; the full joint distribution has \( 2^n \) cells.

E.g., \( n = 30 \) and \( k = 5 \): \( n2^k = 960 \) and \( 2^n > 10^9 \).
In a Bayesian network we can exchange accuracy with complexity.
It may be wiser to leave out very weak dependencies from the network in order to restrict the complexity, but this yields a lower accuracy.
Choosing the right topology for the network is a hard problem.
The correct order in which to add nodes is to add the “root causes” first, then the variables they influence, and so on, until we reach the “leaves,” which have no direct causal influence on other variables.
Adding nodes in the false order makes the network unnecessarily complex and unintuitive.
14.3 Representation of Conditional Distributions

- Filling in the CPT for a node requires up to $O(2^k)$ numbers
- Usually the relationship between the parents and the child is described by an easier-to-calculate canonical distribution
- The simplest example is a deterministic node, which has its value specified exactly by the values of its parents, with no uncertainty
- An example of a logical relationship is for instance a node Scandinavian whose value is a disjunction of its parents Swedish, Norwegian, Danish and Icelandic
- Uncertain relationships can often be characterized by so-called “noisy” logical relationships
- In the noisy-OR relation the causal relationship between parent and child may be inhibited
- It is assumed that the inhibition of each parent is independent of the inhibition of any other parents

\[
P(\neg \text{fever} \mid \text{cold}, \neg \text{flu}, \neg \text{malaria}) = 0.6
\]
\[
P(\neg \text{fever} \mid \neg \text{cold}, \text{flu}, \neg \text{malaria}) = 0.2
\]
\[
P(\neg \text{fever} \mid \neg \text{cold}, \neg \text{flu}, \text{malaria}) = 0.1
\]

<table>
<thead>
<tr>
<th>Cold</th>
<th>Flu</th>
<th>Malaria</th>
<th>$P(\text{Fever})$</th>
<th>$P(\neg \text{Fever})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F$</td>
<td>$F$</td>
<td>$F$</td>
<td>0.9</td>
<td>1.0</td>
</tr>
<tr>
<td>$F$</td>
<td>$F$</td>
<td>$T$</td>
<td>0.9</td>
<td>0.1</td>
</tr>
<tr>
<td>$F$</td>
<td>$T$</td>
<td>$F$</td>
<td>0.8</td>
<td>0.2</td>
</tr>
<tr>
<td>$F$</td>
<td>$T$</td>
<td>$T$</td>
<td>0.98</td>
<td>0.02 = 0.2 × 0.1</td>
</tr>
<tr>
<td>$T$</td>
<td>$F$</td>
<td>$F$</td>
<td>0.4</td>
<td>0.6</td>
</tr>
<tr>
<td>$T$</td>
<td>$F$</td>
<td>$T$</td>
<td>0.94</td>
<td>0.06 = 0.6 × 0.1</td>
</tr>
<tr>
<td>$T$</td>
<td>$T$</td>
<td>$F$</td>
<td>0.88</td>
<td>0.12 = 0.6 × 0.2</td>
</tr>
<tr>
<td>$T$</td>
<td>$T$</td>
<td>$T$</td>
<td>0.988</td>
<td>0.012 = 0.6 × 0.2 × 0.1</td>
</tr>
</tbody>
</table>
Bayesian nets with continuous variables

- We can avoid handling continuous variables by discretization, where the continuous domain is divided up into a fixed set of intervals.
- Using too few intervals may result in considerable loss of accuracy and using too many may lead to very large CPTs.
- Discretization can sometimes even be a provably correct approach for handling continuous domains.
- The other solution is to define standard families of probability density functions that are specified by a finite number of parameters.
- For example, a Gaussian (or normal) distribution \( N(\mu, \sigma^2)(x) \) has the mean \( \mu \) and variance \( \sigma^2 \) as parameters.

For continuous variables, there are an infinite number of values, and unless there are point spikes, the probability of any one value is 0.

- The probability density function \( P(x) \) for a random variable \( X \) is intuitively defined as the ratio of the probability that \( X \) falls into an interval around \( x \), divided by the width of the interval, as the interval width goes to zero:

\[
P(x) = \lim_{dx \to 0} P(x \leq x \leq x + dx)/dx
\]

- A probability density function must be nonnegative for all \( x \) and must have

\[
\int_{-\infty}^{+\infty} P(x)dx = 1
\]

- A cumulative probability density function \( F_X(x) \) is the probability of a random variable \( X \) being less than \( x \):

\[
F_X(x) = P(X \leq x) = \int_{-\infty}^{x} P(u)du
\]
The Gaussian distribution with mean $\mu$ and standard deviation $\sigma$ for a continuous random variable $x$ ranging from $-\infty$ to $+\infty$ is

$$P(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}$$

• A hybrid Bayesian network has both discrete and continuous variables
• Then we need to specify
  - the conditional distribution of a continuous variable, and
  - The conditional distribution of a discrete variable given continuous parents
• E.g., continuous Cost of fruits depends on continuous Harvest and binary Subsidy
• A customer’s discrete decision Buys depends only on the cost
• For the Cost variable, we need to specify
  \[ P(\text{Cost} \mid \text{Harvest, Subsidy}) \]
• The discrete parent is handled by explicit enumeration: \[ P(\text{Cost} \mid \text{Harvest, subsidy}) \] and \[ P(\text{Cost} \mid \text{Harvest, ~subsidy}) \]
• To handle Harvest we specify how the distribution over the cost $c$ depends on the continuous value $h$ of Harvest
I.e., we specify the parameters of the cost distribution as a function of $h$

The most common choice is the linear Gaussian distribution whose mean $\mu$ varies linearly with the value of the parent and whose standard deviation $\sigma$ is fixed

\[
P(c \mid h, \text{subsidy}) = N(a_t h + b_t, \sigma_t^2)(c) = \frac{1}{\sigma_t \sqrt{2\pi}} \exp \left( -\frac{1}{2} \left( (c - (a_t h + b_t))/\sigma_t \right)^2 \right)
\]

\[
P(c \mid h, \neg \text{subsidy}) = N(a_f h + b_f, \sigma_f^2)(c)
\]

Averaging over the two single-bump distributions eventually yields a two-bump distribution $P(c \mid h)$

A Bayesian network employing the linear Gaussian distribution “behaves well” (has an intuitive overall distribution)
Discrete variable $Buys$ depends on the cost of the product
- The customer will buy if the cost is low and will not buy if it is high
- The probability of buying varies smoothly in some intermediate region

In other words, the conditional distribution is like a “soft” threshold function

One way to make soft thresholds is to use the integral of the standard ($\mu = 0, \sigma^2 = 1$) normal distribution; a.k.a. the *probit* distribution

$$
\Phi(x) = \int_{-\infty}^{x} N(0,1)(x)dx
$$
• Then the probability of \textit{Buys} given \textit{Cost} might be
  \[ P(\text{buys} | \text{Cost} = c) = \Phi((-c + \mu) / \sigma) \]
  This means that the cost threshold occurs around \( \mu \), the width of the threshold region is proportional to \( \sigma \), and the probability of buying decreases as cost increases.
• An alternative to probit model is the \textit{logit} distribution, used widely in neural networks, which uses the sigmoid function to produce a soft threshold:
  \[ P(\text{buys} | \text{Cost} = c) = 1 / (1 + \exp(-2(-c + \mu) / \sigma)) \]
  The logit has much longer tails than the probit.
• The probit is often a better fit to real situations, but the logit is sometimes easier to deal with mathematically and, thus, common.