21

Exact Inference by Complete Enumeration

We open our toolbox of methods for handling probabilities by discussing a brute-force inference method: complete enumeration of all hypotheses, and evaluation of their probabilities. This approach is an exact method, and the difficulty of carrying it out will motivate the smarter exact and approximate methods introduced in the following chapters.

21.1 The burglar alarm

Bayesian probability theory is sometimes called 'common sense, amplified'. When thinking about the following questions, please ask your common sense what it thinks the answers are; we will then see how Bayesian methods confirm your everyday intuition.

Example 21.1. Fred lives in Los Angeles and commutes 60 miles to work. Whilst at work, he receives a phone-call from his neighbour saying that Fred's burglar alarm is ringing. What is the probability that there was a burglar in his house today? While driving home to investigate, Fred hears on the radio that there was a small earthquake that day near his home. 'Oh', he says, feeling relieved, 'it was probably the earthquake that set off the alarm'. What is the probability that there was a burglar in his house? (After Pearl, 1988).

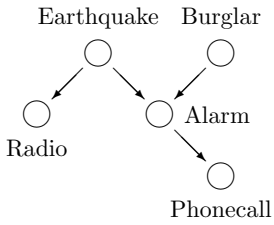


Figure 21.1. Belief network for the burglar alarm problem.

Let's introduce variables b (a burglar was present in Fred's house today), a (the alarm is ringing), p (Fred receives a phonecall from the neighbour reporting the alarm), e (a small earthquake took place today near Fred's house), and r (the radio report of earthquake is heard by Fred). The probability of all these variables might factorize as follows:

$$P(b, e, a, p, r) = P(b)P(e)P(a \mid b, e)P(p \mid a)P(r \mid e), \tag{21.1}$$

and plausible values for the probabilities are:

1. Burglar probability:

$$P(b=1) = \beta, \quad P(b=0) = 1 - \beta, \tag{21.2}$$

   e.g., $\beta = 0.001$ gives a mean burglary rate of once every three years.

2. Earthquake probability:

$$P(e=1) = \epsilon, \quad P(e=0) = 1 - \epsilon, \tag{21.3}$$

with, e.g., $\epsilon = 0.001$; our assertion that the earthquakes are independent of burglars, i.e., the prior probability of b and e is $P(b, e) = P(b)P(e)$, seems reasonable unless we take into account opportunistic burglars who strike immediately after earthquakes.

3. Alarm ringing probability: we assume the alarm will ring if *any* of the following three events happens: (a) a burglar enters the house, and triggers the alarm (let's assume the alarm has a reliability of $\alpha_b = 0.99$, i.e., 99% of burglars trigger the alarm); (b) an earthquake takes place, and triggers the alarm (perhaps $\alpha_e = 1\%$ of alarms are triggered by earthquakes?); or (c) some other event causes a false alarm; let's assume the false alarm rate f is 0.001, so Fred has false alarms from non-earthquake causes once every three years. [This type of dependence of a on b and e is known as a 'noisy-or'.] The probabilities of a given b and e are then:

$$\begin{array}{lcllcl}
P(a=0\,|\,b=0,\,e=0) &=& (1-f), & P(a=1\,|\,b=0,\,e=0) &=& f \\
P(a=0\,|\,b=1,\,e=0) &=& (1-f)(1-\alpha_b), & P(a=1\,|\,b=1,\,e=0) &=& 1-(1-f)(1-\alpha_b) \\
P(a=0\,|\,b=0,\,e=1) &=& (1-f)(1-\alpha_e), & P(a=1\,|\,b=0,\,e=1) &=& 1-(1-f)(1-\alpha_e) \\
P(a=0\,|\,b=1,\,e=1) &=& (1-f)(1-\alpha_b)(1-\alpha_e), & P(a=1\,|\,b=1,\,e=1) &=& 1-(1-f)(1-\alpha_b)(1-\alpha_e)
\end{array}$$

or, in numbers,

$$\begin{array}{lcllcl}
P(a=0\,|\,b=0,\,e=0) &=& 0.999, & P(a=1\,|\,b=0,\,e=0) &=& 0.001 \\
P(a=0\,|\,b=1,\,e=0) &=& 0.009\,99, & P(a=1\,|\,b=1,\,e=0) &=& 0.990\,01 \\
P(a=0\,|\,b=0,\,e=1) &=& 0.989\,01, & P(a=1\,|\,b=0,\,e=1) &=& 0.010\,99 \\
P(a=0\,|\,b=1,\,e=1) &=& 0.009\,890\,1, & P(a=1\,|\,b=1,\,e=1) &=& 0.990\,109\,9.
\end{array}$$

We assume the neighbour would never phone if the alarm is not ringing [$P(p=1\,|\,a=0) = 0$]; and that the radio is a trustworthy reporter too [$P(r=1\,|\,e=0) = 0$]; we won't need to specify the probabilities $P(p=1\,|\,a=1)$ or $P(r=1\,|\,e=1)$ in order to answer the questions above, since the outcomes $p=1$ and $r=1$ give us certainty respectively that $a=1$ and $e=1$.

We can answer the two questions about the burglar by computing the posterior probabilities of all hypotheses given the available information. Let's start by reminding ourselves that the probability that there is a burglar, before either p or r is observed, is $P(b=1) = \beta = 0.001$, and the probability that an earthquake took place is $P(e=1) = \epsilon = 0.001$, and these two propositions are *independent*.

First, when $p=1$, we know that the alarm is ringing: $a=1$. The posterior probability of b and e becomes:

$$P(b, e\,|\,a=1) = \frac{P(a=1\,|\,b, e)P(b)P(e)}{P(a=1)}. \tag{21.4}$$

The numerator's four possible values are

$$\begin{array}{lcllcl}
P(a=1\,|\,b=0,\,e=0)\times P(b=0)\times P(e=0) &=& 0.001 & \times 0.999\times 0.999 &=& 0.000\,998 \\
P(a=1\,|\,b=1,\,e=0)\times P(b=1)\times P(e=0) &=& 0.990\,01 & \times 0.001\times 0.999 &=& 0.000\,989 \\
P(a=1\,|\,b=0,\,e=1)\times P(b=0)\times P(e=1) &=& 0.010\,99 & \times 0.999\times 0.001 &=& 0.000\,010\,979 \\
P(a=1\,|\,b=1,\,e=1)\times P(b=1)\times P(e=1) &=& 0.990\,109\,9 & \times 0.001\times 0.001 &=& 9.9\times 10^{-7}.
\end{array}$$

The normalizing constant is the sum of these four numbers, $P(a=1) = 0.002$, and the posterior probabilities are

$$\begin{array}{lcl}
P(b=0,\,e=0\,|\,a=1) &=& 0.4993 \\
P(b=1,\,e=0\,|\,a=1) &=& 0.4947 \\
P(b=0,\,e=1\,|\,a=1) &=& 0.0055 \\
P(b=1,\,e=1\,|\,a=1) &=& 0.0005.
\end{array} \tag{21.5}$$

To answer the question, 'what's the probability a burglar was there?' we *marginalize* over the earthquake variable e:

$$\begin{array}{rcl} P(b=0\,|\,a=1) & = & P(b=0,\, e=0\,|\,a=1) + P(b=0,\, e=1\,|\,a=1) \;=\; 0.505 \\ P(b=1\,|\,a=1) & = & P(b=1,\, e=0\,|\,a=1) + P(b=1,\, e=1\,|\,a=1) \;=\; 0.495. \end{array} \tag{21.6}$$

So there is nearly a 50% chance that there was a burglar present. It is important to note that the variables b and e, which were independent *a priori*, are now *dependent*. The posterior distribution (21.5) is not a separable function of b and e. This fact is illustrated most simply by studying the effect of learning that $e = 1$.

When we learn $e=1$, the posterior probability of b is given by $P(b\,|\,e=1,\, a=1) = P(b, e=1\,|\,a=1)/P(e=1\,|\,a=1)$, i.e., by dividing the bottom two rows of (21.5), by their sum $P(e=1\,|\,a=1) = 0.0060$. The posterior probability of b is:

$$\begin{array}{rcl} P(b=0\,|\,e=1,\, a=1) & = & 0.92 \\ P(b=1\,|\,e=1,\, a=1) & = & 0.08. \end{array} \tag{21.7}$$

There is thus now an 8% chance that a burglar was in Fred's house. It is in accordance with everyday intuition that the probability that $b=1$ (a possible cause of the alarm) reduces when Fred learns that an earthquake, an alternative explanation of the alarm, has happened.

Explaining away

This phenomenon, that one of the possible causes ($b=1$) of some data (the data in this case being $a=1$) becomes *less* probable when another of the causes ($e=1$) becomes more probable, even though those two causes were independent variables *a priori*, is known as *explaining away*. Explaining away is an important feature of correct inferences, and one that any artificial intelligence should replicate.

If we believe that the neighbour and the radio service are unreliable or capricious, so that we are not certain that the alarm really is ringing or that an earthquake really has happened, the calculations become more complex, but the explaining-away effect persists; the arrival of the earthquake report r simultaneously makes it *more* probable that the alarm truly is ringing, and *less* probable that the burglar was present.

In summary, we solved the inference questions about the burglar by enumerating all four hypotheses about the variables (b, e), finding their posterior probabilities, and marginalizing to obtain the required inferences about b.

▷ Exercise 21.2.[2] After Fred receives the phone-call about the burglar alarm, but before he hears the radio report, what, from his point of view, is the probability that there was a small earthquake today?

▶ 21.2 Exact inference for continuous hypothesis spaces

Many of the hypothesis spaces we will consider are naturally thought of as continuous. For example, the unknown decay length λ of section 3.1 (p.48) lives in a continuous one-dimensional space; and the unknown mean and standard deviation of a Gaussian μ, σ live in a continuous two-dimensional space. In any practical computer implementation, such continuous spaces will necessarily be discretized, however, and so can, in principle, be enumerated – at a grid of parameter values, for example. In figure 3.2 we plotted the likelihood

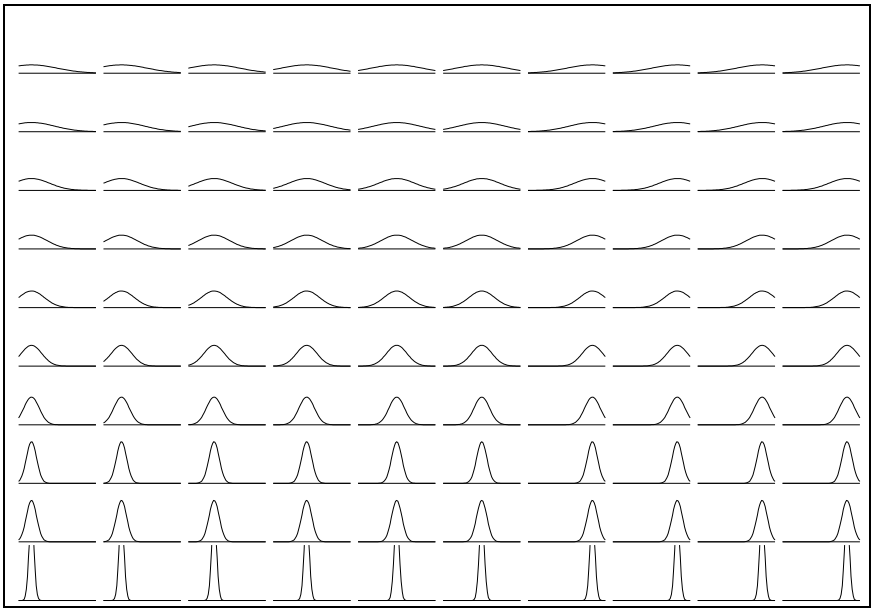


Figure 21.2. Enumeration of an entire (discretized) hypothesis space for one Gaussian with parameters μ (horizontal axis) and σ (vertical).

function for the decay length as a function of λ by evaluating the likelihood at a finely-spaced series of points.

A two-parameter model

Let's look at the Gaussian distribution as an example of a model with a two-dimensional hypothesis space. The one-dimensional Gaussian distribution is parameterized by a mean μ and a standard deviation σ:

$$P(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \equiv \text{Normal}(x; \mu, \sigma^2). \tag{21.8}$$

Figure 21.2 shows an enumeration of one hundred hypotheses about the mean and standard deviation of a one-dimensional Gaussian distribution. These hypotheses are evenly spaced in a ten by ten square grid covering ten values of μ and ten values of σ. Each hypothesis is represented by a picture showing the probability density that it puts on x. We now examine the inference of μ and σ given data points x_n, $n = 1, \ldots, N$, assumed to be drawn independently from this density.

Figure 21.3. Five datapoints $\{x_n\}_{n=1}^5$. The horizontal coordinate is the value of the datum, x_n; the vertical coordinate has no meaning.

Imagine that we acquire data, for example the five points shown in figure 21.3. We can now evaluate the posterior probability of each of the one hundred subhypotheses by evaluating the likelihood of each, that is, the value of $P(\{x_n\}_{n=1}^5 \mid \mu, \sigma)$. The likelihood values are shown diagrammatically in figure 21.4 using the line thickness to encode the value of the likelihood. Subhypotheses with likelihood smaller than e^{-8} times the maximum likelihood have been deleted.

Using a finer grid, we can represent the same information by plotting the likelihood as a surface plot or contour plot as a function of μ and σ (figure 21.5).

A five-parameter mixture model

Eyeballing the data (figure 21.3), you might agree that it seems more plausible that they come not from a single Gaussian but from a mixture of two Gaussians, defined by two means, two standard deviations, and two mixing

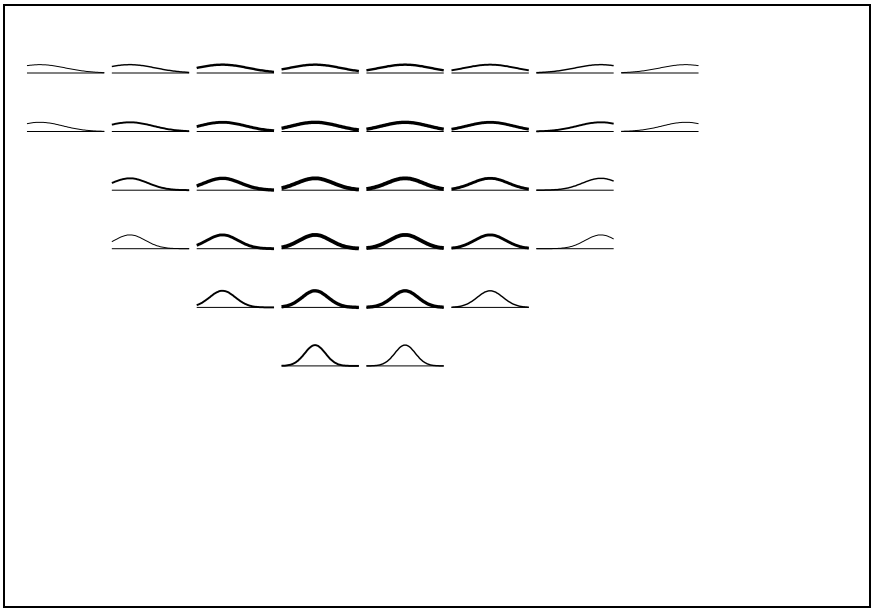


Figure 21.4. Likelihood function, given the data of figure 21.3, represented by line thickness. Subhypotheses having likelihood smaller than e^{-8} times the maximum likelihood are not shown.

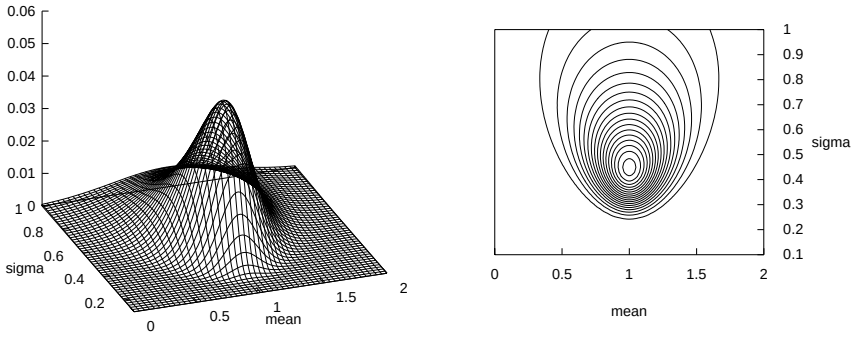


Figure 21.5. The likelihood function for the parameters of a Gaussian distribution. Surface plot and contour plot of the log likelihood as a function of μ and σ. The data set of $N = 5$ points had mean $\bar{x} = 1.0$ and $S = \sum(x - \bar{x})^2 = 1.0$.

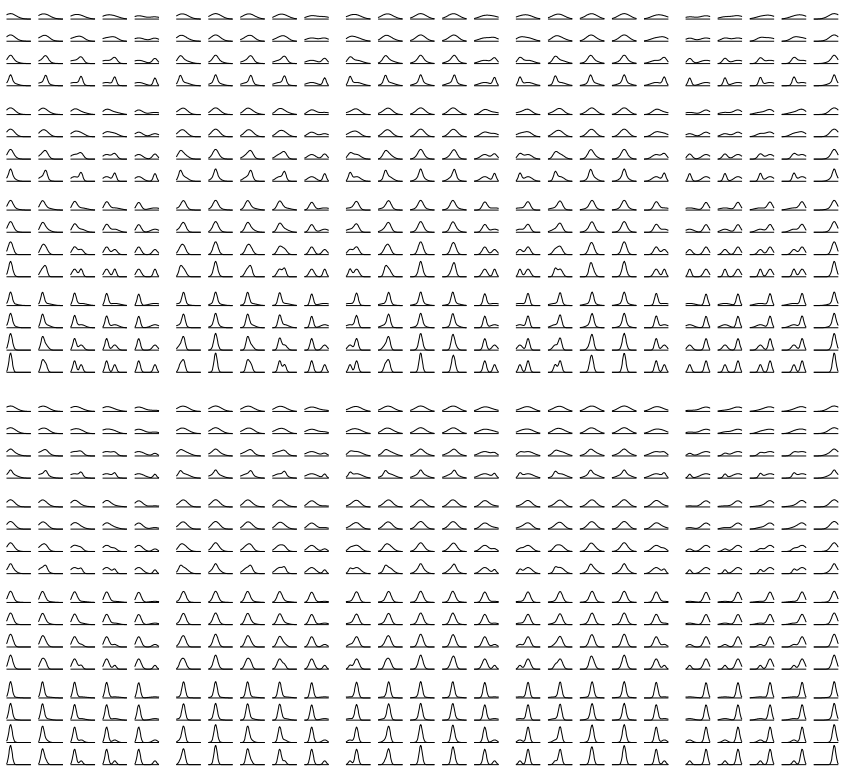


Figure 21.6. Enumeration of the entire (discretized) hypothesis space for a mixture of two Gaussians. Weight of the mixture components is $\pi_1, \pi_2 = 0.6, 0.4$ in the top half and $0.8, 0.2$ in the bottom half. Means μ_1 and μ_2 vary horizontally, and standard deviations σ_1 and σ_2 vary vertically.

coefficients π_1 and π_2, satisfying $\pi_1 + \pi_2 = 1$, $\pi_i \geq 0$.

$$P(x \,|\, \mu_1, \sigma_1, \pi_1, \mu_2, \sigma_2, \pi_2) = \frac{\pi_1}{\sqrt{2\pi}\sigma_1} \exp\left(-\tfrac{(x-\mu_1)^2}{2\sigma_1^2}\right) + \frac{\pi_2}{\sqrt{2\pi}\sigma_2} \exp\left(-\tfrac{(x-\mu_2)^2}{2\sigma_2^2}\right)$$

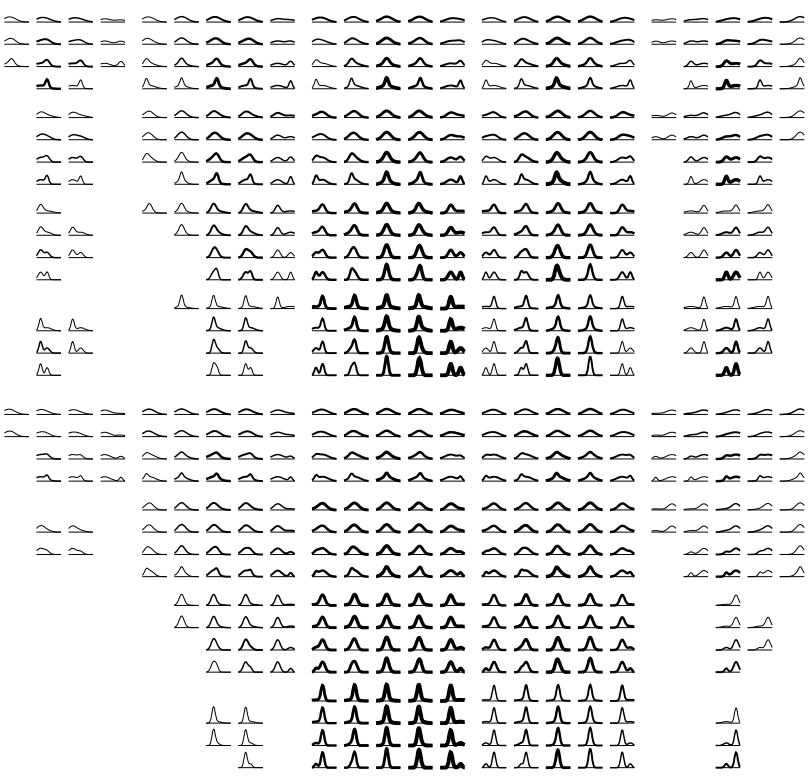
Let's enumerate the subhypotheses for this alternative model. The parameter space is five-dimensional, so it becomes challenging to represent it on a single page. Figure 21.6 enumerates 800 subhypotheses with different values of the five parameters $\mu_1, \mu_2, \sigma_1, \sigma_2, \pi_1$. The means are varied between five values each in the horizontal directions. The standard deviations take on four values each vertically. And π_1 takes on two values vertically. We can represent the inference about these five parameters in the light of the five datapoints as shown in figure 21.7.

If we wish to compare the one-Gaussian model with the mixture-of-two model, we can find the models' posterior probabilities by evaluating the marginal likelihood or evidence for each model $\mathcal{H}$, $P(\{x\} \,|\, \mathcal{H})$. The evidence is given by integrating over the parameters, $\boldsymbol{\theta}$; the integration can be implemented numerically by summing over the alternative enumerated values of $\boldsymbol{\theta}$,

$$P(\{x\} \,|\, \mathcal{H}) = \sum_{\boldsymbol{\theta}} P(\boldsymbol{\theta}) P(\{x\} \,|\, \boldsymbol{\theta}, \mathcal{H}), \tag{21.9}$$

where $P(\boldsymbol{\theta})$ is the prior distribution over the grid of parameter values, which I take to be uniform.

For the mixture of two Gaussians this integral is a five-dimensional integral; if it is to be performed at all accurately, the grid of points will need to be much finer than the grids shown in the figures. If the uncertainty about each of K parameters has been reduced by, say, a factor of ten by observing the data, then brute-force integration requires a grid of at least 10^K points. This



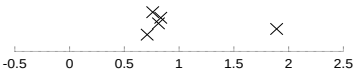


Figure 21.7. Inferring a mixture of two Gaussians. Likelihood function, given the data of figure 21.3, represented by line thickness. The hypothesis space is identical to that shown in figure 21.6. Subhypotheses having likelihood smaller than e^{-8} times the maximum likelihood are not shown, hence the blank regions, which correspond to hypotheses that the data have ruled out.

exponential growth of computation with model size is the reason why complete enumeration is rarely a feasible computational strategy.

Exercise 21.3.[1] Imagine fitting a mixture of ten Gaussians to data in a twenty-dimensional space. Estimate the computational cost of implementing inferences for this model by enumeration of a grid of parameter values.