CHAPTER 1

Probabilistic Reasoning

We have intuition about how uncertainty works in simple cases. To reach sensible conclusions in complicated situations, however – where there may be many (possibly) related events and many possible outcomes – we need a formal 'calculus' that extends our intuitive notions. The concepts, mathematical language and rules of probability give us the formal framework we need. In this chapter we review basic concepts in probability – in particular, conditional probability and Bayes' rule, the workhorses of machine learning. Another strength of the language of probability is that it structures problems in a form consistent for computer implementation. We also introduce basic features of the BRMLtoolbox *that support manipulating probability distributions.*

1.1 Probability Refresher

Variables, States and Notational Shortcuts

Variables will be denoted using either upper case X or lower case x and a set of variables will typically be denoted by a calligraphic symbol, for example $\mathcal{V} = \{a, B, c\}$.

The *domain* of a variable x is written $\mathrm{dom}(x)$, and denotes the states x can take. States will typically be represented using sans-serif font. For example, for a coin c, $\mathrm{dom}(c) = \{\mathsf{heads}, \mathsf{tails}\}$ and $p(c = \mathsf{heads})$ represents the probability that variable c is in state heads. The meaning of $p(\mathsf{state})$ will often be clear, without specific reference to a variable. For example, if we are discussing an experiment about a coin c, the meaning of $p(\mathsf{heads})$ is clear from the context, being shorthand for $p(c = \mathsf{heads})$. When summing over a variable $\sum_x f(x)$, the interpretation is that all states of x are included, *i.e.* $\sum_x f(x) \equiv \sum_{\mathsf{s} \in \mathrm{dom}(x)} f(x = \mathsf{s})$. Given a variable, x, its domain $\mathrm{dom}(x)$ and a full specification of the probability values for each of the variable states, $p(x)$, we have a *distribution* for x. Sometimes we will not fully specify the distribution, only certain properties, such as for variables x, y, $p(x, y) = p(x)p(y)$ for some unspecified $p(x)$ and $p(y)$. When clarity on this is required we will say distributions with structure $p(x)p(y)$, or a distribution class $p(x)p(y)$.

For our purposes, *events* are expressions about random variables, such as *Two heads in 6 coin tosses*. Two events are *mutually exclusive* if they cannot both be true. For example the events *The coin is heads* and *The coin is tails* are mutually exclusive. One can think of defining a new variable named by the event so, for example, $p(\textit{The coin is tails})$ can be interpreted as $p(\textit{The coin is tails} = \mathsf{true})$. We use the shorthand $p(x = \mathsf{tr})$ for the probability of event/variable x being in the state true and $p(x = \mathsf{fa})$ for the probability of variable x being in the state false.

Definition 1.1 (Rules of Probability for Discrete Variables)**.**

The probability $p(x = \mathsf{x})$ of variable x being in state x is represented by a value between 0 and 1. $p(x = \mathsf{x}) = 1$ means that we are certain x is in state x. Conversely, $p(x = \mathsf{x}) = 0$ means that we are certain x is not in state x. Values between 0 and 1 represent the degree of certainty of state occupancy.

The summation of the probability over all the states is 1:

$$\sum_{\mathsf{x} \in \mathrm{dom}(x)} p(x = \mathsf{x}) = 1 \tag{1.1.1}$$

This is called the normalisation condition. We will usually more conveniently write $\sum_x p(x) = 1$.

Two variables x and y can interact through

$$p(x = \mathsf{a} \text{ or } y = \mathsf{b}) = p(x = \mathsf{a}) + p(y = \mathsf{b}) - p(x = \mathsf{a} \text{ and } y = \mathsf{b}) \tag{1.1.2}$$

Or, more generally, we can write

$$p(x \text{ or } y) = p(x) + p(y) - p(x \text{ and } y) \tag{1.1.3}$$

We will use the shorthand $p(x, y)$ for $p(x \text{ and } y)$. Note that $p(y, x) = p(x, y)$ and $p(x \text{ or } y) = p(y \text{ or } x)$.

Definition 1.2 (Set notation)**.** An alternative notation in terms of set theory is to write

$$p(x \text{ or } y) \equiv p(x \cup y), \qquad p(x, y) \equiv p(x \cap y) \tag{1.1.4}$$

Definition 1.3 (Marginals)**.** Given a *joint distribution* $p(x, y)$ the distribution of a single variable is given by

$$p(x) = \sum_y p(x, y) \tag{1.1.5}$$

Here $p(x)$ is termed a *marginal* of the joint probability distribution $p(x, y)$. The process of computing a marginal from a joint distribution is called *marginalisation*. More generally, one has

$$p(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) = \sum_{x_i} p(x_1, \ldots, x_n) \tag{1.1.6}$$

Definition 1.4 (Conditional Probability / Bayes' Rule)**.** The probability of event x conditioned on knowing event y (or more shortly, the probability of x given y) is defined as

$$p(x|y) \equiv \frac{p(x, y)}{p(y)} \tag{1.1.7}$$

If $p(y) = 0$ then $p(x|y)$ is not defined. From this definition and $p(x, y) = p(y, x)$ we immediately arrive at Bayes' rule

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)} \tag{1.1.8}$$

Since Bayes' rule trivially follows from the definition of conditional probability, we will sometimes be loose in our language and use the terms Bayes' rule and conditional probability as synonymous.

As we shall see throughout this book, Bayes' rule plays a central role in probabilistic reasoning since it helps

us 'invert' probabilistic relationships, translating between $p(y|x)$ and $p(x|y)$.

Definition 1.5 (Probability Density Functions)**.** For a continuous variable x, the probability density $f(x)$ is defined such that

$$f(x) \geq 0, \qquad \int_{-\infty}^{\infty} f(x)dx = 1 \tag{1.1.9}$$

and the probability that x falls in an interval $[a, b]$ is given by

$$p(a \leq x \leq b) = \int_a^b f(x)dx \tag{1.1.10}$$

As shorthand we will sometimes write $\int_x f(x)$, particularly when we want an expression to be valid for either continuous or discrete variables. The multivariate case is analogous with integration over all real space, and the probability that x belongs to a region of the space defined accordingly. Unlike probabilities, probability densities can take positive values greater than 1.

Formally speaking, for a continuous variable, one should not speak of the probability that $x = 0.2$ since the probability of a single value is always zero. However, we shall often write $p(x)$ for continuous variables, thus not distinguishing between probabilities and probability density function values. Whilst this may appear strange, the nervous reader may simply replace our $p(x)$ notation for $\int_{x\in\Delta} f(x)dx$, where Δ is a small region centred on x. This is well defined in a probabilistic sense and, in the limit Δ being very small, this would give approximately $\Delta f(x)$. If we consistently use the same Δ for all occurrences of pdfs, then we will simply have a common prefactor Δ in all expressions. Our strategy is to simply ignore these values (since in the end only relative probabilities will be relevant) and write $p(x)$. In this way, all the standard rules of probability carry over, including Bayes' Rule.

Remark 1.1 (Subjective Probability)**.** Probability is a contentious topic and we do not wish to get bogged down by the debate here, apart from pointing out that it is not necessarily the rules of probability that are contentious, rather what interpretation we should place on them. In some cases potential repetitions of an experiment can be envisaged so that the 'long run' (or frequentist) definition of probability in which probabilities are defined with respect to a potentially infinite repetition of experiments makes sense. For example, in coin tossing, the probability of heads might be interpreted as 'If I were to repeat the experiment of flipping a coin (at 'random'), the limit of the number of heads that occurred over the number of tosses is defined as the probability of a head occurring.'

Here's a problem that is typical of the kind of scenario one might face in a machine learning situation. A film enthusiast joins a new online film service. Based on expressing a few films a user likes and dislikes, the online company tries to estimate the probability that the user will like each of the 10000 films in their database. If we were to define probability as a limiting case of infinite repetitions of the same experiment, this wouldn't make much sense in this case since we can't repeat the experiment. However, if we assume that the user behaves in a manner consistent with other users, we should be able to exploit the large amount of data from other users' ratings to make a reasonable 'guess' as to what this consumer likes. This *degree of belief* or *Bayesian* subjective interpretation of probability sidesteps non-repeatability issues – it's just a framework for manipulating real values consistent with our intuition about probability[158].

1.1.1 Interpreting Conditional Probability

Conditional probability matches our intuitive understanding of uncertainty. For example, imagine a circular dart board, split into 20 equal sections, labelled from 1 to 20. Randy, a dart thrower, hits any one of the 20 sections uniformly at random. Hence the probability that a dart thrown by Randy occurs in any one of the 20 regions is $p(\textit{region } i) = 1/20$. A friend of Randy tells him that he hasn't hit the 20 region. What is the probability that Randy has hit the 5 region? Conditioned on this information, only regions 1 to 19 remain possible and, since there is no preference for Randy to hit any of these regions, the probability is 1/19. The

conditioning means that certain states are now inaccessible, and the original probability is subsequently distributed over the remaining accessible states. From the rules of probability :

$$p(\textit{region } 5|\textit{not region } 20) = \frac{p(\textit{region } 5, \textit{not region } 20)}{p(\textit{not region } 20)} = \frac{p(\textit{region } 5)}{p(\textit{not region } 20)} = \frac{1/20}{19/20} = \frac{1}{19}$$

giving the intuitive result. An important point to clarify is that $p(A = \mathsf{a}|B = \mathsf{b})$ should not be interpreted as 'Given the event $B = \mathsf{b}$ has occurred, $p(A = \mathsf{a}|B = \mathsf{b})$ is the probability of the event $A = \mathsf{a}$ occurring'. In most contexts, no such explicit temporal causality is implied[1] and the correct interpretation should be ' $p(A = \mathsf{a}|B = \mathsf{b})$ is the probability of A being in state a under the constraint that B is in state b'.

The relation between the conditional $p(A = \mathsf{a}|B = \mathsf{b})$ and the joint $p(A = \mathsf{a}, B = \mathsf{b})$ is just a normalisation constant since $p(A = \mathsf{a}, B = \mathsf{b})$ is not a distribution in A – in other words, $\sum_{\mathsf{a}} p(A = \mathsf{a}, B = \mathsf{b}) \neq 1$. To make it a distribution we need to divide : $p(A = \mathsf{a}, B = \mathsf{b}) / \sum_{\mathsf{a}} p(A = \mathsf{a}, B = \mathsf{b})$ which, when summed over a does sum to 1. Indeed, this is just the definition of $p(A = \mathsf{a}|B = \mathsf{b})$.

Definition 1.6 (Independence)**.**

Variables x and y are independent if knowing the state (or value in the continuous case) of one variable gives no extra information about the other variable. Mathematically, this is expressed by

$$p(x, y) = p(x)p(y) \tag{1.1.11}$$

Provided that $p(x) \neq 0$ and $p(y) \neq 0$ independence of x and y is equivalent to

$$p(x|y) = p(x) \Leftrightarrow p(y|x) = p(y) \tag{1.1.12}$$

If $p(x|y) = p(x)$ for all states of x and y, then the variables x and y are said to be independent. If

$$p(x, y) = kf(x)g(y) \tag{1.1.13}$$

for some constant k, and positive functions $f(\cdot)$ and $g(\cdot)$ then x and y are independent and we write $x \perp\!\!\!\perp y$.

Example 1.1 (Independence)**.** Let x denote the day of the week in which females are born, and y denote the day in which males are born, with $\mathrm{dom}(x) = \mathrm{dom}(y) = \{1, \ldots, 7\}$. It is reasonable to expect that x is independent of y. We randomly select a woman from the phone book, Alice, and find out that she was born on a Tuesday. We also randomly select a male at random, Bob. Before phoning Bob and asking him, what does knowing Alice's birth day add to which day we think Bob is born on? Under the independence assumption, the answer is nothing. Note that this doesn't mean that the distribution of Bob's birthday is necessarily uniform – it just means that knowing when Alice was born doesn't provide any extra information than we already knew about Bob's birthday, $p(y|x) = p(y)$. Indeed, the distribution of birthdays $p(y)$ and $p(x)$ are non-uniform (statistically fewer babies are born on weekends), though there is nothing to suggest that x are y are dependent.

Deterministic Dependencies

Sometimes the concept of independence is perhaps a little strange. Consider the following : variables x and y are both binary (their domains consist of two states). We define the distribution such that x and y are always both in a certain joint state:

$$p(x = \mathsf{a}, y = 1) = 1, \quad p(x = \mathsf{a}, y = 2) = 0, \quad p(x = \mathsf{b}, y = 2) = 0, \quad p(x = \mathsf{b}, y = 1) = 0$$

Are x and y dependent? The reader may show that $p(x = \mathsf{a}) = 1$, $p(x = \mathsf{b}) = 0$ and $p(y = 1) = 1$, $p(y = 2) = 0$. Hence $p(x)p(y) = p(x, y)$ for all states of x and y, and x and y are therefore independent.

[1] We will discuss issues related to causality further in section(3.4).

This may seem strange – we know for sure the relation between x and y, namely that they are always in the same joint state, yet they are independent. Since the distribution is trivially concentrated in a single joint state, knowing the state of x tells you nothing that you didn't anyway know about the state of y, and vice versa. This potential confusion comes from using the term 'independent' which may suggest that there is no relation between objects discussed. The best way to think about statistical independence is to ask whether or not knowing the state of variable y tells you something more than you knew before about variable x, where 'knew before' means working with the joint distribution of $p(x, y)$ to figure out what we can know about x, namely $p(x)$.

Definition 1.7 (Conditional Independence)**.**

$$\mathcal{X} \perp\!\!\!\perp \mathcal{Y} | \mathcal{Z} \tag{1.1.14}$$

denotes that the two sets of variables $\mathcal{X}$ and $\mathcal{Y}$ are independent of each other provided we know the state of the set of variables $\mathcal{Z}$. For conditional independence, $\mathcal{X}$ and $\mathcal{Y}$ must be independent given *all* states of $\mathcal{Z}$. Formally, this means that

$$p(\mathcal{X}, \mathcal{Y}|\mathcal{Z}) = p(\mathcal{X}|\mathcal{Z})p(\mathcal{Y}|\mathcal{Z}) \tag{1.1.15}$$

for all states of $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$. In case the conditioning set is empty we may also write $\mathcal{X} \perp\!\!\!\perp \mathcal{Y}$ for $\mathcal{X} \perp\!\!\!\perp \mathcal{Y} | \emptyset$, in which case $\mathcal{X}$ is (unconditionally) independent of $\mathcal{Y}$.

If $\mathcal{X}$ and $\mathcal{Y}$ are not conditionally independent, they are conditionally dependent. This is written

$$\mathcal{X} \top\!\!\!\top \mathcal{Y} | \mathcal{Z} \tag{1.1.16}$$

Similarly $\mathcal{X} \top\!\!\!\top \mathcal{Y} | \emptyset$ can be written as $\mathcal{X} \top\!\!\!\top \mathcal{Y}$.

Intuitively, if x is conditionally independent of y given z, this means that, given z, y contains no additional information about x. Similarly, given z, knowing x does not tell me anything more about y. Note that $\mathcal{X} \perp\!\!\!\perp \mathcal{Y} | \mathcal{Z} \Rightarrow \mathcal{X}' \perp\!\!\!\perp \mathcal{Y}' | \mathcal{Z}$ for $\mathcal{X}' \subseteq \mathcal{X}$ and $\mathcal{Y}' \subseteq \mathcal{Y}$.

Remark 1.2 (Independence implications)**.** It's tempting to think that if a is independent of b and b is independent of c then a must be independent of c:

$$\{a \perp\!\!\!\perp b, b \perp\!\!\!\perp c\} \Rightarrow a \perp\!\!\!\perp c \tag{1.1.17}$$

However, this does not follow. Consider for example a distribution of the form

$$p(a, b, c) = p(b)p(a, c) \tag{1.1.18}$$

From this

$$p(a, b) = \sum_c p(a, b, c) = p(b) \sum_c p(a, c) \tag{1.1.19}$$

Hence $p(a, b)$ is a function of b multiplied by a function of a so that a and b are independent. Similarly, one can show that b and c are independent. However, a is not necessarily independent of c since the distribution $p(a, c)$ can be set arbitrarily.

Similarly, it's tempting to think that if a and b are dependent, and b and c are dependent, then a and c must be dependent:

$$\{a \top\!\!\!\top b, b \top\!\!\!\top c\} \Rightarrow a \top\!\!\!\top c \tag{1.1.20}$$

However, this also does not follow. We give an explicit numerical example in exercise(3.17).

Finally, note that conditional independence $x \perp\!\!\!\perp y | z$ does not imply marginal independence $x \perp\!\!\!\perp y$. See also exercise(3.20).

1.1.2 Probability Tables

Based on the populations 60776238, 5116900 and 2980700 of England (E), Scotland (S) and Wales (W), the a priori probability that a randomly selected person from the combined three countries would live in England, Scotland or Wales, is approximately 0.88, 0.08 and 0.04 respectively. We can write this as a vector (or probability table) :

$$\begin{pmatrix} p(Cnt = \mathsf{E}) \\ p(Cnt = \mathsf{S}) \\ p(Cnt = \mathsf{W}) \end{pmatrix} = \begin{pmatrix} 0.88 \\ 0.08 \\ 0.04 \end{pmatrix} \tag{1.1.21}$$

whose component values sum to 1. The ordering of the components in this vector is arbitrary, as long as it is consistently applied.

For the sake of simplicity, we assume that only three Mother Tongue languages exist : English (Eng), Scottish (Scot) and Welsh (Wel), with conditional probabilities given the country of residence, England (E), Scotland (S) and Wales (W). We write a (fictitious) conditional probability table

$$\begin{array}{lll} p(MT = \mathsf{Eng}|Cnt = \mathsf{E}) = 0.95 & p(MT = \mathsf{Eng}|Cnt = \mathsf{S}) = 0.7 & p(MT = \mathsf{Eng}|Cnt = \mathsf{W}) = 0.6 \\ p(MT = \mathsf{Scot}|Cnt = \mathsf{E}) = 0.04 & p(MT = \mathsf{Scot}|Cnt = \mathsf{S}) = 0.3 & p(MT = \mathsf{Scot}|Cnt = \mathsf{W}) = 0.0 \\ p(MT = \mathsf{Wel}|Cnt = \mathsf{E}) = 0.01 & p(MT = \mathsf{Wel}|Cnt = \mathsf{S}) = 0.0 & p(MT = \mathsf{Wel}|Cnt = \mathsf{W}) = 0.4 \end{array} \tag{1.1.22}$$

From this we can form a joint distribution $p(Cnt, MT) = p(MT|Cnt)p(Cnt)$. This could be written as a 3×3 matrix with columns indexed by country and rows indexed by Mother Tongue:

$$\begin{pmatrix} 0.95 \times 0.88 & 0.7 \times 0.08 & 0.6 \times 0.04 \\ 0.04 \times 0.88 & 0.3 \times 0.08 & 0.0 \times 0.04 \\ 0.01 \times 0.88 & 0.0 \times 0.08 & 0.4 \times 0.04 \end{pmatrix} = \begin{pmatrix} 0.836 & 0.056 & 0.024 \\ 0.0352 & 0.024 & 0 \\ 0.0088 & 0 & 0.016 \end{pmatrix} \tag{1.1.23}$$

The joint distribution contains all the information about the model of this environment. By summing the columns of this table, we have the marginal $p(Cnt)$. Summing the rows gives the marginal $p(MT)$. Similarly, one could easily infer $p(Cnt|MT) \propto p(MT|Cnt)p(Cnt)$ from this joint distribution by dividing an entry of equation (1.1.23) by its row sum.

For joint distributions over a larger number of variables, $x_i, i = 1, \ldots, D$, with each variable x_i taking K_i states, the table describing the joint distribution is an array with $\prod_{i=1}^{D} K_i$ entries. Explicitly storing tables therefore requires space exponential in the number of variables, which rapidly becomes impractical for a large number of variables. We discuss how to deal with this issue in chapter(3) and chapter(4).

A probability distribution assigns a value to each of the joint states of the variables. For this reason, $p(T, J, R, S)$ is considered equivalent to $p(J, S, R, T)$ (or any such reordering of the variables), since in each case the joint setting of the variables is simply a different index to the same probability. This situation is more clear in the set theoretic notation $p(J \cap S \cap T \cap R)$. We abbreviate this set theoretic notation by using the commas – however, one should be careful not to confuse the use of this indexing type notation with functions $f(x, y)$ which are in general dependent on the variable order. Whilst the variables to the left of the conditioning bar may be written in any order, and equally those to the right of the conditioning bar may be written in any order, moving variables across the bar is not generally equivalent, so that $p(x_1|x_2) \neq p(x_2|x_1)$.

1.2 Probabilistic Reasoning

The central paradigm of probabilistic reasoning is to identify all relevant variables $x_1, \ldots, x_N$ in the environment, and make a probabilistic model $p(x_1, \ldots, x_N)$ of their interaction. Reasoning (inference) is then performed by introducing *evidence* that sets variables in known states, and subsequently computing probabilities of interest, conditioned on this evidence. The rules of probability, combined with Bayes' rule make for a complete reasoning system, one which includes traditional deductive logic as a special case[158]. In the examples below, the number of variables in the environment is very small. In chapter(3) we will discuss

reasoning in networks containing many variables, for which the graphical notations of chapter(2) will play a central role.

Example 1.2 (Hamburgers)**.** Consider the following fictitious scientific information: Doctors find that people with Kreuzfeld-Jacob disease (KJ) almost invariably ate hamburgers, thus $p(\textit{Hamburger Eater}|KJ) = 0.9$. The probability of an individual having KJ is currently rather low, about one in 100,000.

1. Assuming eating lots of hamburgers is rather widespread, say $p(\textit{Hamburger Eater}) = 0.5$, what is the probability that a hamburger eater will have Kreuzfeld-Jacob disease?

   This may be computed as

$$p(KJ\,|\textit{Hamburger Eater}) = \frac{p(\textit{Hamburger Eater}, KJ)}{p(\textit{Hamburger Eater})} = \frac{p(\textit{Hamburger Eater}|KJ)p(KJ)}{p(\textit{Hamburger Eater})} \tag{1.2.1}$$

$$= \frac{\frac{9}{10} \times \frac{1}{100000}}{\frac{1}{2}} = 1.8 \times 10^{-5} \tag{1.2.2}$$

2. If the fraction of people eating hamburgers was rather small, $p(\textit{Hamburger Eater}) = 0.001$, what is the probability that a regular hamburger eater will have Kreuzfeld-Jacob disease? Repeating the above calculation, this is given by

$$\frac{\frac{9}{10} \times \frac{1}{100000}}{\frac{1}{1000}} \approx 1/100 \tag{1.2.3}$$

   This is much higher than in scenario (1) since here we can be more sure that eating hamburgers is related to the illness.

Example 1.3 (Inspector Clouseau)**.** Inspector Clouseau arrives at the scene of a crime. The victim lies dead in the room alongside the possible murder weapon, a knife. The Butler (B) and Maid (M) are the inspector's main suspects and the inspector has a prior belief of 0.6 that the Butler is the murderer, and a prior belief of 0.2 that the Maid is the murderer. These beliefs are independent in the sense that $p(B, M) = p(B)p(M)$. (It is possible that both the Butler and the Maid murdered the victim or neither). The inspector's *prior* criminal knowledge can be formulated mathematically as follows:

$$\mathrm{dom}(B) = \mathrm{dom}(M) = \{\mathsf{murderer}, \mathsf{not\ murderer}\}, \mathrm{dom}(K) = \{\mathsf{knife\ used}, \mathsf{knife\ not\ used}\} \tag{1.2.4}$$

$$p(B = \mathsf{murderer}) = 0.6, \qquad p(M = \mathsf{murderer}) = 0.2 \tag{1.2.5}$$

$$\begin{array}{lll} p(\mathsf{knife\ used}|B = \mathsf{not\ murderer}, & M = \mathsf{not\ murderer}) & = 0.3 \\ p(\mathsf{knife\ used}|B = \mathsf{not\ murderer}, & M = \mathsf{murderer}) & = 0.2 \\ p(\mathsf{knife\ used}|B = \mathsf{murderer}, & M = \mathsf{not\ murderer}) & = 0.6 \\ p(\mathsf{knife\ used}|B = \mathsf{murderer}, & M = \mathsf{murderer}) & = 0.1 \end{array} \tag{1.2.6}$$

In addition $p(K, B, M) = p(K|B, M)p(B)p(M)$. Assuming that the knife is the murder weapon, what is the probability that the Butler is the murderer? (Remember that it might be that neither is the murderer). Using b for the two states of B and m for the two states of M,

$$p(B|K) = \sum_m p(B, m|K) = \sum_m \frac{p(B, m, K)}{p(K)} = \frac{\sum_m p(K|B, m)p(B, m)}{\sum_{m,b} p(K|b, m)p(b, m)} = \frac{p(B)\sum_m p(K|B, m)p(m)}{\sum_b p(b)\sum_m p(K|b, m)p(m)} \tag{1.2.7}$$

where we used the fact that in our model $p(B, M) = p(B)p(M)$. Plugging in the values we have (see also `demoClouseau.m`)

$$p(B = \text{murderer}|\text{knife used}) = \frac{\frac{6}{10}\left(\frac{2}{10}\times\frac{1}{10} + \frac{8}{10}\times\frac{6}{10}\right)}{\frac{6}{10}\left(\frac{2}{10}\times\frac{1}{10} + \frac{8}{10}\times\frac{6}{10}\right) + \frac{4}{10}\left(\frac{2}{10}\times\frac{2}{10} + \frac{8}{10}\times\frac{3}{10}\right)} = \frac{300}{412} \approx 0.73 \quad (1.2.8)$$

Hence knowing that the knife was the murder weapon strengthens our belief that the butler did it.

Remark 1.3. The role of $p(\text{knife used})$ in the Inspector Clouseau example can cause some confusion. In the above,

$$p(\text{knife used}) = \sum_b p(b) \sum_m p(\text{knife used}|b, m)p(m) \quad (1.2.9)$$

is computed to be 0.412. But surely, $p(\text{knife used}) = 1$, since this is given in the question! Note that the @@ quantity $p(\text{knife used})$ relates to the *prior* probability the model assigns to the knife being used (in the absence of any other information). If we know that the knife is used, then the *posterior*

$$p(\text{knife used}|\text{knife used}) = \frac{p(\text{knife used}, \text{knife used})}{p(\text{knife used})} = \frac{p(\text{knife used})}{p(\text{knife used})} = 1 \quad (1.2.10)$$

which, naturally, must be the case.

Example 1.4 (Who's in the bathroom?)**.** Consider a household of three people, Alice, Bob and Cecil. Cecil wants to go to the bathroom but finds it occupied. He then goes to Alice's room and sees she is there. Since Cecil knows that only either Alice or Bob can be in the bathroom, from this he infers that Bob must be in the bathroom.

To arrive at the same conclusion in a mathematical framework, we define the following events

$$A = \text{Alice is in her bedroom}, \quad B = \text{Bob is in his bedroom}, \quad O = \text{Bathroom occupied} \quad (1.2.11)$$

We can encode the information that if either Alice or Bob are not in their bedrooms, then they must be in the bathroom (they might both be in the bathroom) as

$$p(O = \text{tr}|A = \text{fa}, B) = 1, \qquad p(O = \text{tr}|A, B = \text{fa}) = 1 \quad (1.2.12)$$

The first term expresses that the bathroom is occupied if Alice is not in her bedroom, wherever Bob is. Similarly, the second term expresses bathroom occupancy as long as Bob is not in his bedroom. Then

$$p(B = \text{fa}|O = \text{tr}, A = \text{tr}) = \frac{p(B = \text{fa}, O = \text{tr}, A = \text{tr})}{p(O = \text{tr}, A = \text{tr})} = \frac{p(O = \text{tr}|A = \text{tr}, B = \text{fa})p(A = \text{tr}, B = \text{fa})}{p(O = \text{tr}, A = \text{tr})} \quad (1.2.13)$$

where

$$p(O = \text{tr}, A = \text{tr}) = p(O = \text{tr}|A = \text{tr}, B = \text{fa})p(A = \text{tr}, B = \text{fa}) + p(O = \text{tr}|A = \text{tr}, B = \text{tr})p(A = \text{tr}, B = \text{tr}) \quad (1.2.14)$$

Using the fact $p(O = \text{tr}|A = \text{tr}, B = \text{fa}) = 1$ and $p(O = \text{tr}|A = \text{tr}, B = \text{tr}) = 0$, which encodes that if Alice is in her room and Bob is not, the bathroom must be occupied, and similarly, if both Alice and Bob are in their rooms, the bathroom cannot be occupied,

$$p(B = \text{fa}|O = \text{tr}, A = \text{tr}) = \frac{p(A = \text{tr}, B = \text{fa})}{p(A = \text{tr}, B = \text{fa})} = 1 \quad (1.2.15)$$

This example is interesting since we are not required to make a full probabilistic model in this case thanks to the limiting nature of the probabilities (we don't need to specify $p(A, B)$). The situation is common in limiting situations of probabilities being either 0 or 1, corresponding to traditional logic systems.

@@

Example 1.5 (Aristotle : Modus Ponens)**.** According to logic, the statements 'All apples are fruit' and 'All fruits grow on trees' lead to the conclusion that 'All apples grow on trees'. This kind of reasoning is a form of transitivity : from the statements $A \Rightarrow F$ and $F \Rightarrow T$ we can infer $A \Rightarrow T$.

To see how this might be deduced using Bayesian, we assume that 'All apples are fruit' corresponds to $p(F = \mathsf{tr}|A = \mathsf{tr}) = 1$ and 'All fruit grows on trees' corresponds to $p(T = \mathsf{tr}|F = \mathsf{tr}) = 1$. We then want to show that this implies $p(T = \mathsf{tr}|A = \mathsf{tr}) = 1$. Showing this is equivalent to showing $p(T = \mathsf{fa}|A = \mathsf{tr}) = 0$ which (assuming $p(A = \mathsf{tr}) > 0$) is in turn equivalent to showing that $p(T = \mathsf{fa}, A = \mathsf{tr}) = 0$. Consider

$$p(T = \mathsf{fa}, A = \mathsf{tr}) = p(T = \mathsf{fa}, A = \mathsf{tr}, F = \mathsf{tr}) + p(T = \mathsf{fa}, A = \mathsf{tr}, F = \mathsf{fa}) \tag{1.2.16}$$

We can show that both terms on the right are zero. First, consider

$$p(T = \mathsf{fa}, A = \mathsf{tr}, F = \mathsf{tr}) \leq p(T = \mathsf{fa}, F = \mathsf{tr}) = p(T = \mathsf{fa}|F = \mathsf{tr})p(F = \mathsf{tr}) \tag{1.2.17}$$

This is zero since, by assumption, $p(T = \mathsf{fa}|F = \mathsf{tr}) = 1 - p(T = \mathsf{tr}|F = \mathsf{tr}) = 1 - 1 = 0$. Similarly,

$$p(T = \mathsf{fa}, A = \mathsf{tr}, F = \mathsf{fa}) \leq p(A = \mathsf{tr}, F = \mathsf{fa}) = p(F = \mathsf{fa}|A = \mathsf{tr})p(A = \mathsf{tr}) \tag{1.2.18}$$

where again, by assumption, $p(F = \mathsf{fa}|A = \mathsf{tr}) = 0$.

Example 1.6 (Aristotle : Inverse Modus Ponens)**.** According to Logic, from the statement : 'If A is true then B is true', one may deduce that 'if B is false then A is false'. To see how this fits in with a probabilistic reasoning system we can first express the statement : 'If A is true then B is true' as $p(B = \mathsf{tr}|A = \mathsf{tr}) = 1$. Then we may infer

$$\begin{aligned} p(A = \mathsf{fa}|B = \mathsf{fa}) &= 1 - p(A = \mathsf{tr}|B = \mathsf{fa}) \\ &= 1 - \frac{p(B = \mathsf{fa}|A = \mathsf{tr})p(A = \mathsf{tr})}{p(B = \mathsf{fa}|A = \mathsf{tr})p(A = \mathsf{tr}) + p(B = \mathsf{fa}|A = \mathsf{fa})p(A = \mathsf{fa})} = 1 \end{aligned} \tag{1.2.19}$$

This follows since $p(B = \mathsf{fa}|A = \mathsf{tr}) = 1 - p(B = \mathsf{tr}|A = \mathsf{tr}) = 1 - 1 = 0$, annihilating the second term.

Both the above examples are intuitive expressions of deductive logic. The standard rules of Aristotelian logic are therefore seen to be limiting cases of probabilistic reasoning.

Example 1.7 (Soft XOR Gate)**.**
A standard XOR logic gate is given by the table on the right. If we observe that the output of the XOR gate is 0, what can we say about A and B? In this case, either A and B were both 0, or A and B were both 1. This means we don't know which state A was in – it could equally likely have been 1 or 0.

A	B	A xor B
0	0	0
0	1	1
1	0	1
1	1	0

Consider a 'soft' version of the XOR gate given on the right, so that the gate stochastically outputs $C = 1$ depending on its inputs, with additionally $A \perp\!\!\!\perp B$ and $p(A = 1) = 0.65$, $p(B = 1) = 0.77$. What is $p(A = 1|C = 0)$? ++

A	B	$p(C = 1\|A, B)$
0	0	0.1
0	1	0.99
1	0	0.8
1	1	0.25

$$\begin{aligned} p(A = 1, C = 0) &= \sum_B p(A = 1, B, C = 0) = \sum_B p(C = 0|A = 1, B)p(A = 1)p(B) \\ &= p(A = 1)\left(p(C = 0|A = 1, B = 0)p(B = 0) + p(C = 0|A = 1, B = 1)p(B = 1)\right) \\ &= 0.65 \times (0.2 \times 0.23 + 0.75 \times 0.77) = 0.405275 \end{aligned} \tag{1.2.20}$$

$$\begin{aligned} p(A = 0, C = 0) &= \sum_B p(A = 0, B, C = 0) = \sum_B p(C = 0|A = 0, B)p(A = 0)p(B) \\ &= p(A = 0)\left(p(C = 0|A = 0, B = 0)p(B = 0) + p(C = 0|A = 0, B = 1)p(B = 1)\right) \\ &= 0.35 \times (0.9 \times 0.23 + 0.01 \times 0.77) = 0.075145 \end{aligned}$$

Then

$$p(A = 1|C = 0) = \frac{p(A = 1, C = 0)}{p(A = 1, C = 0) + p(A = 0, C = 0)} = \frac{0.405275}{0.405275 + 0.075145} = 0.8436 \tag{1.2.21}$$

Example 1.8 (Larry)**.** Larry is typically late for school. If Larry is late, we denote this with $L = \text{late}$, otherwise, $L = \text{not late}$. When his mother asks whether or not he was late for school he never admits to being late. The response Larry gives R_L is represented as follows

$$p(R_L = \text{not late}|L = \text{not late}) = 1, \qquad p(R_L = \text{late}|L = \text{late}) = 0 \tag{1.2.22}$$

The remaining two values are determined by normalisation and are

$$p(R_L = \text{late}|L = \text{not late}) = 0, \qquad p(R_L = \text{not late}|L = \text{late}) = 1 \tag{1.2.23}$$

Given that $R_L = \text{not late}$, what is the probability that Larry was late, *i.e.* $p(L = \text{late}|R_L = \text{not late})$?

Using Bayes' we have

$$\begin{aligned} p(L = \text{late}|R_L = \text{not late}) &= \frac{p(L = \text{late}, R_L = \text{not late})}{p(R_L = \text{not late})} \\ &= \frac{p(L = \text{late}, R_L = \text{not late})}{p(L = \text{late}, R_L = \text{not late}) + p(L = \text{not late}, R_L = \text{not late})} \end{aligned} \tag{1.2.24}$$

In the above

$$p(L = \text{late}, R_L = \text{not late}) = \underbrace{p(R_L = \text{not late}|L = \text{late})}_{=1} p(L = \text{late}) \tag{1.2.25}$$

and

$$p(L = \text{not late}, R_L = \text{not late}) = \underbrace{p(R_L = \text{not late}|L = \text{not late})}_{=1} p(L = \text{not late}) \tag{1.2.26}$$

Hence

$$p(L = \text{late}|R_L = \text{not late}) = \frac{p(L = \text{late})}{p(L = \text{late}) + p(L = \text{not late})} = p(L = \text{late}) \tag{1.2.27}$$

Where we used normalisation in the last step, $p(L = \text{late}) + p(L = \text{not late}) = 1$. This result is intuitive – Larry's mother knows that he never admits to being late, so her belief about whether or not he really was late is unchanged, regardless of what Larry actually says.

Example 1.9 (Larry and Sue)**.** Continuing the example above, Larry's sister Sue always tells the truth to her mother as to whether or not Larry was late for School.

$$p(R_S = \text{not late}|L = \text{not late}) = 1, \quad p(R_S = \text{late}|L = \text{late}) = 1 \tag{1.2.28}$$

The remaining two values are determined by normalisation and are

$$p(R_S = \text{late}|L = \text{not late}) = 0, \quad p(R_S = \text{not late}|L = \text{late}) = 0 \tag{1.2.29}$$

We also assume $p(R_S, R_L|L) = p(R_S|L)p(R_L|L)$. We can then write

$$p(R_L, R_S, L) = p(R_L|L)p(R_S|L)p(L) \tag{1.2.30}$$

Given that $R_S = \text{late}$ and $R_L = \text{not late}$, what is the probability that Larry was late?

Using Bayes' rule, we have

$$\begin{aligned} &p(L = \text{late}|R_L = \text{not late}, R_S = \text{late}) \\ &\qquad = \frac{1}{Z} p(R_S = \text{late}|L = \text{late}) p(R_L = \text{not late}|L = \text{late}) p(L = \text{late}) \end{aligned} \tag{1.2.31}$$

where the normalisation Z is given by

$$\begin{aligned} &p(R_S = \text{late}|L = \text{late}) p(R_L = \text{not late}|L = \text{late}) p(L = \text{late}) \\ &\quad + p(R_S = \text{late}|L = \text{not late}) p(R_L = \text{not late}|L = \text{not late}) p(L = \text{not late}) \end{aligned} \tag{1.2.32}$$

Hence

$$p(L = \text{late}|R_L = \text{not late}, R_S = \text{late}) = \frac{1 \times 1 \times p(L = \text{late})}{1 \times 1 \times p(L = \text{late}) + 0 \times 1 \times p(L = \text{not late})} = 1 \tag{1.2.33}$$

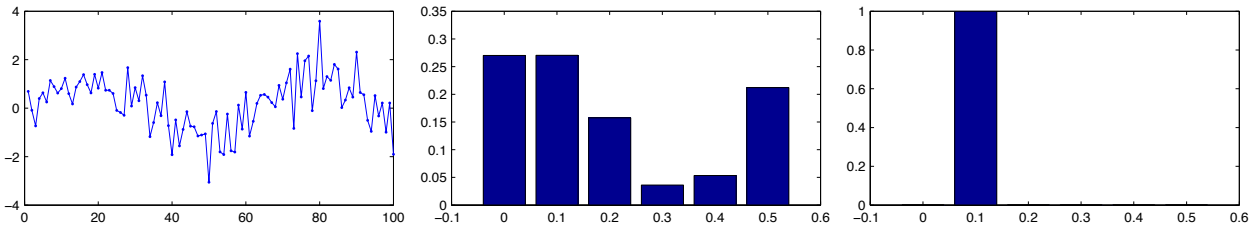
This result is also intuitive – Since Larry's mother knows that Sue always tells the truth, no matter what Larry says, she knows he was late.

Example 1.10 (Luke)**.** Luke has been told he's lucky and has won a prize in the lottery. There are 5 prizes available of value £10, £100, £1000, £10000, £1000000. The prior probabilities of winning these 5 prizes are p_1, p_2, p_3, p_4, p_5, with p_0 being the prior probability of winning no prize. Luke asks eagerly 'Did I win £1000000?!'. 'I'm afraid not sir', is the response of the lottery phone operator. 'Did I win £10000?!' asks Luke. 'Again, I'm afraid not sir'. What is the probability that Luke has won £1000?

Note first that $p_0 + p_1 + p_2 + p_3 + p_4 + p_5 = 1$. We denote $W = 1$ for the first prize of £10, and $W = 2, \ldots, 5$ for the remaining prizes and $W = 0$ for no prize. We need to compute

$$\begin{aligned} p(W = 3|W \neq 5, W \neq 4, W \neq 0) &= \frac{p(W = 3, W \neq 5, W \neq 4, W \neq 0)}{p(W \neq 5, W \neq 4, W \neq 0)} \\ &= \frac{p(W = 3)}{p(W = 1 \text{ or } W = 2 \text{ or } W = 3)} = \frac{p_3}{p_1 + p_2 + p_3} \end{aligned} \tag{1.2.34}$$

where the term in the denominator is computed using the fact that the events W are mutually exclusive (one can only win one prize). This result makes intuitive sense : once we have removed the impossible states of W, the probability that Luke wins the prize is proportional to the prior probability of that prize, with the normalisation being simply the total set of possible probability remaining.



(a) (b) (c)

Figure 1.1: **(a)**: Noisy observations of displacements $x_1, \ldots, x_{100}$ for a pendulum. **(b)**: The prior belief on 5 possible values of θ. **(c)**: The posterior belief on θ.

1.3 Prior, Likelihood and Posterior

Much of science deals with problems of the form : tell me something about the variable θ given that I have observed data $\mathcal{D}$ and have some knowledge of the underlying data generating mechanism. Our interest is then the quantity

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} = \frac{p(\mathcal{D}|\theta)p(\theta)}{\int_\theta p(\mathcal{D}|\theta)p(\theta)} \tag{1.3.1}$$

This shows how from a forward or *generative model* $p(\mathcal{D}|\theta)$ of the dataset, and coupled with a *prior* belief $p(\theta)$ about which variable values are appropriate, we can infer the *posterior* distribution $p(\theta|\mathcal{D})$ of the variable in light of the observed data. The *most probable a posteriori* (*MAP*) setting is that which maximises the posterior, $\theta_* = \arg\max_\theta p(\theta|\mathcal{D})$. For a 'flat prior', $p(\theta)$ being a constant, not changing with θ, the MAP solution is equivalent to the *maximum likelihood*, namely that θ that maximises the likelihood $p(\mathcal{D}|\theta)$ of the model generating the observed data. We will return to a discussion of such summaries of the posterior and parameter learning in chapter(9).

This use of a generative model sits well with physical models of the world which typically postulate how to generate observed phenomena, assuming we know the model. For example, one might postulate how to generate a time-series of displacements for a swinging pendulum but with unknown mass, length and damping constant. Using this generative model, and given only the displacements, we could infer the unknown physical properties of the pendulum.

Example 1.11 (Pendulum)**.** As a prelude to scientific inference and the use of continuous variables, we consider an idealised pendulum for which x_t is the angular displacement of the pendulum at time t. Assuming that the measurements are independent, given the knowledge of the parameter of the problem, θ, we have that the likelihood of a sequence of observations $x_1, \ldots, x_T$ is given by

$$p(x_1, \ldots, x_T|\theta) = \prod_{t=1}^{T} p(x_t|\theta) \tag{1.3.2}$$

If the model is correct and our measurement of the displacements x is perfect, then the physical model is

$$x_t = \sin(\theta t) \tag{1.3.3}$$

where θ represents the unknown physical constants of the pendulum ($\sqrt{g/L}$, where g is the gravitational attraction and L the length of the pendulum). If, however, we assume that we have a rather poor instrument to measure the displacements, with a known variance of σ^2 (see chapter(8)), then

$$x_t = \sin(\theta t) + \epsilon_t \tag{1.3.4}$$

where ϵ_t is zero mean Gaussian noise with variance σ^2. We can also consider a set of possible parameters θ and place a prior $p(\theta)$ over them, expressing our prior belief (before seeing the measurements) in the appropriateness of the different values of θ. The posterior distribution is then given by

$$p(\theta|x_1,\ldots,x_T) \propto p(\theta)\prod_{t=1}^{T}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}(x_t-\sin(\theta t))^2} \tag{1.3.5}$$

Despite noisy measurements, the posterior over the assumed possible values for θ becomes strongly peaked for a large number of measurements, see fig(1.1).

1.3.1 Two dice : what were the individual scores?

Two fair dice are rolled. Someone tells you that the sum of the two scores is 9. What is the posterior distribution of the dice scores[2]?
The score of die a is denoted s_a with $\mathrm{dom}(s_a) = \{1,2,3,4,5,6\}$ and similarly for s_b. The three variables involved are then s_a, s_b and the total score, $t = s_a + s_b$. A model of these three variables naturally takes the form

$$p(t,s_a,s_b) = \underbrace{p(t|s_a,s_b)}_{\text{likelihood}}\underbrace{p(s_a,s_b)}_{\text{prior}} \tag{1.3.6}$$

The prior $p(s_a,s_b)$ is the joint probability of score s_a and score s_b without knowing anything else. Assuming no dependency in the rolling mechanism,

$$p(s_a,s_b) = p(s_a)p(s_b) \tag{1.3.7}$$

Since the dice are fair both $p(s_a)$ and $p(s_b)$ are uniform distributions, $p(s_a) = p(s_b) = 1/6$.

Here the likelihood term is

$$p(t|s_a,s_b) = \mathbb{I}\left[t = s_a + s_b\right] \tag{1.3.8}$$

which states that the total score is given by $s_a + s_b$. Here $\mathbb{I}\left[A\right]$ is the *indicator function* defined as $\mathbb{I}\left[A\right] = 1$ if the statement A is true and 0 otherwise.

Hence, our complete model is

$$p(t,s_a,s_b) = p(t|s_a,s_b)p(s_a)p(s_b) \tag{1.3.9}$$

where the terms on the right are explicitly defined.

The posterior is then given by,

$$p(s_a,s_b|t=9) = \frac{p(t=9|s_a,s_b)p(s_a)p(s_b)}{p(t=9)} \tag{1.3.10}$$

where

$$p(t=9) = \sum_{s_a,s_b} p(t=9|s_a,s_b)p(s_a)p(s_b) \tag{1.3.11}$$

$p(s_a)p(s_b)$:

	$s_a=1$	$s_a=2$	$s_a=3$	$s_a=4$	$s_a=5$	$s_a=6$
$s_b=1$	1/36	1/36	1/36	1/36	1/36	1/36
$s_b=2$	1/36	1/36	1/36	1/36	1/36	1/36
$s_b=3$	1/36	1/36	1/36	1/36	1/36	1/36
$s_b=4$	1/36	1/36	1/36	1/36	1/36	1/36
$s_b=5$	1/36	1/36	1/36	1/36	1/36	1/36
$s_b=6$	1/36	1/36	1/36	1/36	1/36	1/36

$p(t=9|s_a,s_b)$:

	$s_a=1$	$s_a=2$	$s_a=3$	$s_a=4$	$s_a=5$	$s_a=6$
$s_b=1$	0	0	0	0	0	0
$s_b=2$	0	0	0	0	0	0
$s_b=3$	0	0	0	0	0	1
$s_b=4$	0	0	0	0	1	0
$s_b=5$	0	0	0	1	0	0
$s_b=6$	0	0	1	0	0	0

$p(t=9|s_a,s_b)p(s_a)p(s_b)$:

	$s_a=1$	$s_a=2$	$s_a=3$	$s_a=4$	$s_a=5$	$s_a=6$
$s_b=1$	0	0	0	0	0	0
$s_b=2$	0	0	0	0	0	0
$s_b=3$	0	0	0	0	0	1/36
$s_b=4$	0	0	0	0	1/36	0
$s_b=5$	0	0	0	1/36	0	0
$s_b=6$	0	0	1/36	0	0	0

$p(s_a,s_b|t=9)$:

	$s_a=1$	$s_a=2$	$s_a=3$	$s_a=4$	$s_a=5$	$s_a=6$
$s_b=1$	0	0	0	0	0	0
$s_b=2$	0	0	0	0	0	0
$s_b=3$	0	0	0	0	0	1/4
$s_b=4$	0	0	0	0	1/4	0
$s_b=5$	0	0	0	1/4	0	0
$s_b=6$	0	0	1/4	0	0	0

The term $p(t=9) = \sum_{s_a,s_b} p(t=9|s_a,s_b)p(s_a)p(s_b) = 4\times 1/36 = 1/9$. Hence the posterior is given by equal mass in only 4 non-zero elements, as shown.

[2]This example is due to Taylan Cemgil.

1.4 Summary

- The standard rules of probability are a consistent, logical way to reason with uncertainty.
- Bayes' rule mathematically encodes the process of inference.

A useful introduction to probability is given in [292]. The interpretation of probability is contentious and we refer the reader to [158, 197, 193] for detailed discussions. The website `understandinguncertainty.org` contains entertaining discussions on reasoning with uncertainty.

1.5 Code

The BRMLtoolbox code accompanying this book is intended to give the reader some insight into representing discrete probability tables and performing simple inference. We provide here only the briefest of descriptions of the code and the reader is encouraged to experiment with the demos to understand better the routines and their purposes.

1.5.1 Basic Probability code

At the simplest level, we only need two basic routines. One for multiplying probability tables together (called potentials in the code), and one for summing a probability table. Potentials are represented using a structure. For example, in the code corresponding to the Inspector Clouseau example `demoClouseau.m`, we define a probability table as

```
>> pot(1)
ans =
    variables: [1 3 2]
        table: [2x2x2 double]
```

This says that the potential depends on the variables $1, 3, 2$ and the entries are stored in the array given by the table field. The size of the array informs how many states each variable takes in the order given by `variables`. The order in which the variables are defined in a potential is irrelevant provided that one indexes the array consistently. A routine that can help with setting table entries is `setstate.m`. For example,

```
>> pot(1) = setstate(pot(1),[2 1 3],[2 1 1],0.3)
```

means that for potential 1, the table entry for variable 2 being in state 2, variable 1 being in state 1 and variable 3 being in state 1 should be set to value 0.3.

The philosophy of the code is to keep the information required to perform computations to a minimum. Additional information about the labels of variables and their domains can be useful to interpret results, but is not actually required to carry out computations. One may also specify the name and domain of each variable, for example

```
>>variable(3)
ans =
    domain: {'murderer'  'not murderer'}
      name: 'butler'
```

The variable name and domain information in the Clouseau example is stored in the structure `variable`, which can be helpful to display the potential table:

```
>> disptable(pot(1),variable);
knife   =   used        maid    = murderer        butler  =  murderer      0.100000
knife   =   not used    maid    = murderer        butler  =  murderer      0.900000
knife   =   used        maid    = not murderer    butler  =  murderer      0.600000
```

```
knife   =   not used    maid    = not murderer    butler  =  murderer      0.400000
knife   =   used        maid    = murderer        butler  =  not murderer  0.200000
knife   =   not used    maid    = murderer        butler  =  not murderer  0.800000
knife   =   used        maid    = not murderer    butler  =  not murderer  0.300000
knife   =   not used    maid    = not murderer    butler  =  not murderer  0.700000
```

Multiplying Potentials

In order to multiply potentials, (as for arrays) the tables of each potential must be dimensionally consistent – that is the number of states of variable i must be the same for all potentials. This can be checked using `potvariables.m`. This consistency is also required for other basic operations such as summing potentials.

`multpots.m`: Multiplying two or more potentials
`divpots.m`: Dividing a potential by another

Summing a Potential

`sumpot.m`: Sum (marginalise) a potential over a set of variables
`sumpots.m`: Sum a set of potentials together

Making a conditional Potential

`condpot.m`: Make a potential conditioned on variables

Setting a Potential

`setpot.m`: Set variables in a potential to given states
`setevpot.m`: Set variables in a potential to given states and return also an identity potential on the given states

The philosophy of BRMLtoolbox is that all information about variables is local and is read off from a potential. Using `setevpot.m` enables one to set variables in a state whilst maintaining information about the number of states of a variable.

Maximising a Potential

`maxpot.m`: Maximise a potential over a set of variables
See also `maxNarray.m` and `maxNpot.m` which return the N-highest values and associated states.

Other potential utilities

`setstate.m`: Set a potential state to a given value
`table.m`: Return a table from a potential
`whichpot.m`: Return potentials which contain a set of variables
`potvariables.m`: Variables and their number of states in a set of potentials
`orderpotfields.m`: Order the fields of a potential structure
`uniquepots.m`: Merge redundant potentials by multiplication and return only unique ones
`numstates.m`: Number of states of a variable in a domain
`squeezepots.m`: Find unique potentials and rename the variables 1,2,...
`normpot.m`: Normalise a potential to form a distribution

1.5.2 General utilities

`condp.m`: Return a table $p(x|y)$ from $p(x, y)$
`condexp.m`: Form a conditional distribution from a log value
`logsumexp.m`: Compute the log of a sum of exponentials in a numerically precise way
`normp.m`: Return a normalised table from an unnormalised table
`assign.m`: Assign values to multiple variables
`maxarray.m`: Maximize a multi-dimensional array over a subset

1.5.3 An example

The following code highlights the use of the above routines in solving the Inspector Clouseau, example(1.3), and the reader is invited to examine the code to become familiar with how to numerically represent probability tables.
`demoClouseau.m`: Solving the Inspector Clouseau example

1.6 Exercises

Exercise 1.1. *Prove*

$$p(x, y|z) = p(x|z)p(y|x, z) \tag{1.6.1}$$

and also

$$p(x|y, z) = \frac{p(y|x, z)p(x|z)}{p(y|z)} \tag{1.6.2}$$

Exercise 1.2. *Prove the* Bonferroni inequality

$$p(a, b) \geq p(a) + p(b) - 1 \tag{1.6.3}$$

Exercise 1.3 (Adapted from [181])**.** *There are two boxes. Box 1 contains three red and five white balls and box 2 contains two red and five white balls. A box is chosen at random* $p(box = 1) = p(box = 2) = 0.5$ *and a ball chosen at random from this box turns out to be red. What is the posterior probability that the red ball came from box 1?*

Exercise 1.4 (Adapted from [181])**.** *Two balls are placed in a box as follows: A fair coin is tossed and a white ball is placed in the box if a head occurs, otherwise a red ball is placed in the box. The coin is tossed again and a red ball is placed in the box if a tail occurs, otherwise a white ball is placed in the box. Balls are drawn from the box three times in succession (always with replacing the drawn ball back in the box). It is found that on all three occasions a red ball is drawn. What is the probability that both balls in the box are red?*

A secret government agency has developed a scanner which determines whether a person is a terrorist. The scanner is fairly reliable; 95% of all scanned terrorists are identified as terrorists, and 95% of all upstanding citizens are identified as such. An informant tells the agency that exactly one passenger of 100 aboard an aeroplane in which you are seated is a terrorist. The police haul off the plane the first person for which the scanner tests positive. What is the probability that this person is a terrorist? @@

Exercise 1.6. *Consider three variable distributions which admit the factorisation*

$$p(a, b, c) = p(a|b)p(b|c)p(c) \tag{1.6.4}$$

where all variables are binary. How many parameters are needed to specify distributions of this form?

Exercise 1.7. *Repeat the Inspector Clouseau scenario, example(1.3), but with the restriction that either the maid or the butler is the murderer, but not both. Explicitly, the probability of the maid being the murderer and not the butler is 0.04, the probability of the butler being the murderer and not the maid is 0.64. Modify* `demoClouseau.m` *to implement this.*

Exercise 1.8. *Prove*

$$p(a, (b \text{ or } c)) = p(a, b) + p(a, c) - p(a, b, c) \tag{1.6.5}$$

Exercise 1.9. *Prove*

$$p(x|z) = \sum_y p(x|y, z)p(y|z) = \sum_{y,w} p(x|w, y, z)p(w|y, z)p(y|z) \tag{1.6.6}$$

Exercise 1.10. *As a young man Mr Gott visits Berlin in 1969. He's surprised that he cannot cross into East Berlin since there is a wall separating the two halves of the city. He's told that the wall was erected 8 years previously. He reasons that : The wall will have a finite lifespan; his ignorance means that he arrives uniformly at random at some time in the lifespan of the wall. Since only 5% of the time one would arrive in the first or last 2.5% of the lifespan of the wall he asserts that with 95% confidence the wall will survive between* $8/0.975 \approx 8.2$ *and* $8/0.025 = 320$ *years. In 1989 the now Professor Gott is pleased to find that his prediction was correct and promotes his prediction method in prestigious journals. This 'delta-t' method is widely adopted and used to form predictions in a range of scenarios about which researchers are 'totally ignorant'. Would you 'buy' a prediction from Prof. Gott? Explain carefully your reasoning.*

Exercise 1.11. *Implement the soft XOR gate, example(1.7) using* BRMLtoolbox. *You may find* `condpot.m` *of use.*

Exercise 1.12. *Implement the hamburgers, example(1.2) (both scenarios) using* BRMLtoolbox. *To do so you will need to define the joint distribution* $p(hamburgers, KJ)$ *in which* $dom(hamburgers) = dom(KJ) = \{tr, fa\}$.

Exercise 1.13. *Implement the two-dice example, section(1.3.1) using* BRMLtoolbox.

Exercise 1.14. *A redistribution lottery involves picking the correct four numbers from 1 to 9 (without replacement, so 3,4,4,1 for example is not possible). The order of the picked numbers is irrelevant. Every week a million people play this game, each paying £1 to enter, with the numbers 3,5,7,9 being the most popular (1 in every 100 people chooses these numbers). Given that the million pounds prize money is split equally between winners, and that any four (different) numbers come up at random, what is the expected amount of money each of the players choosing 3,5,7,9 will win each week? The least popular set of numbers is 1,2,3,4 with only 1 in 10,000 people choosing this. How much do they profit each week, on average? Do you think there is any 'skill' involved in playing this lottery?*

Exercise 1.15. *In a test of 'psychometry' the car keys and wrist watches of 5 people are given to a medium. The medium then attempts to match the wrist watch with the car key of each person. What is the expected number of correct matches that the medium will make (by chance)? What is the probability that the medium will obtain at least 1 correct match?*

Exercise 1.16. 1. *Show that for any function* f

$$\sum_x p(x|y) f(y) = f(y) \tag{1.6.7}$$

2. *Explain why, in general,*

$$\sum_x p(x|y) f(x, y) \neq \sum_x f(x, y) \tag{1.6.8}$$

Exercise 1.17 (Inspired by `singingbanana.com`)**.** *Seven friends decide to order pizzas by telephone from Pizza4U based on a flyer pushed through their letterbox. Pizza4U has only 4 kinds of pizza, and each person chooses a pizza independently. Bob phones Pizza4U and places the combined pizza order, simply stating how many pizzas of each kind are required. Unfortunately, the precise order is lost, so the chef makes seven randomly chosen pizzas and then passes them to the delivery boy.*

1. *How many different combined orders are possible?*

2. *What is the probability that the delivery boy has the right order?*

Exercise 1.18. *Sally is new to the area and listens to some friends discussing about another female friend. Sally knows that they are talking about either Alice or Bella but doesn't know which. From previous conversations Sally knows some independent pieces of information: She's 90% sure that Alice has a white car, but doesn't know if Bella's car is white or black. Similarly, she's 90% sure that Bella likes sushi, but doesn't know if Alice likes sushi. Sally hears from the conversation that the person being discussed hates sushi and drives a white car. What is the probability that the friends are talking about Alice? Assume maximal uncertainty in the absence of any knowledge of the probabilities.*

Exercise 1.19. *The weather in London can be summarised as: if it rains one day there's a 70% chance it will rain the following day; if it's sunny one day there's a 40% chance it will be sunny the following day.*

1. *Assuming that the prior probability it rained yesterday is 0.5, what is the probability that it was raining yesterday given that it's sunny today?*
2. *If the weather follows the same pattern as above, day after day, what is the probability that it will rain on any day (based on an effectively infinite number of days of observing the weather)?*
3. *Use the result from part 2 above as a new prior probability of rain yesterday and recompute the probability that it was raining yesterday given that it's sunny today.*

Exercise 1.20. *A game of Battleships is played on a* 10×10 *pixel grid. There are two 5-pixel length ships placed uniformly at random on the grid, subject to the constraints that (i) the ships cannot overlap and (ii) one ship is vertical and the other horizontal. After 10 unsuccessful 'misses' in locations* $(1,10), (2,2), (3,8), (4,4), (5,6), (6,5), (7,4), (7,7), (9,2), (9,9)$ *calculate which pixel has the highest probability of containing a ship. State this pixel and the value of the highest probability.* ++

Exercise 1.21. *A game of Battleships is played on a* 8×8 *grid. There are two 5-pixel length ships placed horizontally and two 5-pixel length ships placed vertically, subject to the same constraints as in the previous question. Given 'misses' in locations* $(1,1), (2,2)$ *and a 'hit' in location* $(5,5)$*, which pixel most likely contains a ship and what is that probability?* ++