Part I

Inference in probabilistic models

Probabilistic models explicitly take into account uncertainty and deal with our imperfect knowledge of the world. Such models are of fundamental significance in Machine Learning since our understanding of the world will always be limited by our observations and understanding. We will focus initially on using probabilistic models as a kind of expert system.

In Part I, we assume that the model is fully specified. That is, given a model of the environment, how can we use it to answer questions of interest? We will relate the complexity of inferring quantities of interest to the structure of the graph describing the model. In addition, we will describe operations in terms of manipulations on the corresponding graphs. As we will see, provided the graphs are simple tree-like structures, most quantities of interest can be computed efficiently.

Part I deals with manipulating mainly discrete variable distributions and forms the background to all the later material in the book.

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 six 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 10 000 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 select a male at random, Bob. Before phoning Bob and asking him, what does knowing Alice's birthday 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

[1] We will discuss issues related to causality further in Section 3.4.

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. 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:

$$\{a \not\perp\!\!\!\perp b, b \not\perp\!\!\!\perp c\} \Rightarrow a \not\perp\!\!\!\perp c. \tag{1.1.20}$$

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 $x \perp\!\!\!\perp y$.

1.1.2 Probability tables

Based on the populations 60 776 238, 5 116 900 and 2 980 700 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

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

$$= \frac{\frac{9}{10} \times \frac{1}{100\,000}}{\frac{1}{2}} = 1.8 \times 10^{-5}. \tag{1.2.2}$$

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

$$\frac{\frac{9}{10} \times \frac{1}{100\,000}}{\frac{1}{1000}} \approx 1/100. \tag{1.2.3}$$

   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:

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

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

$$\begin{aligned} &p(\text{knife used}|B = \text{not murderer}, \quad M = \text{not murderer}) = 0.3 \\ &p(\text{knife used}|B = \text{not murderer}, \quad M = \text{murderer}) \quad = 0.2 \\ &p(\text{knife used}|B = \text{murderer}, \quad M = \text{not murderer}) = 0.6 \\ &p(\text{knife used}|B = \text{murderer}, \quad M = \text{murderer}) \quad = 0.1. \end{aligned} \tag{1.2.6}$$

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,

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

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`)

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

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,

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

is computed to be 0.456. 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*

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

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.