CHAPTER 1

Probabilistic Reasoning

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, we might have $\mathrm{dom}(c) = \{\mathsf{heads}, \mathsf{tails}\}$ 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 (or performing some other operation) over a variable $\sum_x f(x)$, the interpretation is that all states of x are included, *i.e.* $\sum_x f(x) \equiv \sum_{\mathsf{s}\in\mathrm{dom}(x)} f(x = \mathsf{s})$.

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 simultaneously occur. 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 $p(x = \mathsf{tr})$ for the probability of event/variable x being in the state true and $p(x = \mathsf{fa})$ for the probability of event/variable x being in the state false.

The Rules of Probability

Definition 1 (Rules of Probability (Discrete Variables))**.**

The probability of an event x occurring is represented by a value between 0 and 1.

$p(x) = 1$ means that we are certain that the event does occur.

Conversely, $p(x) = 0$ means that we are certain that the event does not occur.

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

$$\sum_{\mathsf{x}} p(x = \mathsf{x}) = 1 \qquad (1.1.1)$$

Such probabilities are normalised. We will usually more conveniently write $\sum_x p(x) = 1$.

Two events x and y can interact through

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

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

$$p(x \text{ or } y) \equiv p(x \cup y), \qquad p(x, y) \equiv p(x \cap y) \tag{1.1.3}$$

Definition 3 (Marginals)**.** Given a *joint distribution* $p(x, y)$ the distribution of a single variable is given by

$$p(x) = \sum_y p(x, y) \tag{1.1.4}$$

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

$$p(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) = \sum_{x_i} p(x_1, \ldots, x_n) \tag{1.1.5}$$

An important definition that will play a central role in this book is conditional probability.

Definition 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

$$p(x|y) \equiv \frac{p(x, y)}{p(y)} \tag{1.1.6}$$

If $p(y) = 0$ then $p(x|y)$ is not defined.

Probability Density Functions

Definition 5 (Probability Density Functions)**.** For a single continuous variable x, the probability density $p(x)$ is defined such that

$$p(x) \geq 0 \tag{1.1.7}$$

$$\int_{-\infty}^{\infty} p(x) dx = 1 \tag{1.1.8}$$

$$p(a < x < b) = \int_a^b p(x)dx \tag{1.1.9}$$

As shorthand we will sometimes write $\int_{x=a}^{b} p(x)$, 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.

For continuous variables, formally speaking, events are defined for the variable occurring within a defined region, for example

$$p(x \in [-1, 1.7]) = \int_{-1}^{1.7} f(x)dx \tag{1.1.10}$$

where here $f(x)$ is the probability density function (pdf) of the continuous random variable x. 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 $p(X = x)$ 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.

Interpreting Conditional Probability

Imagine a circular dart board, split into 20 equal sections, labelled from 1 to 20 and Randy, a dart thrower who 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. In the above $p(\textit{region } 5, \textit{not region } 20) = p(\textit{region } \{5 \cap 1 \cap 2\cap, \ldots, \cap 19\}) = p(\textit{region } 5)$.

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 $p(A = \mathsf{a}|B = \mathsf{b})$.

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

Definition 6 (Independence)**.**

Events x and y are independent if knowing one event gives no extra information about the other event. 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.

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:

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

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, in English, suggests that there is no influence or 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 $p(x)$.

1.1.1 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 these three countries would live in England, Scotland or Wales, would be approximately 0.88, 0.08 and 0.04 respectively. We can write this as a vector (or probability table) :

$$\begin{pmatrix} p(Cnt = \mathsf{E}) \\ p(Cnt = \mathsf{S}) \\ p(Cnt = \mathsf{W}) \end{pmatrix} = \begin{pmatrix} 0.88 \\ 0.08 \\ 0.04 \end{pmatrix} \tag{1.1.14}$$

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, let's 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

$$\begin{array}{lll} p(MT=\mathsf{Eng}|Cnt=\mathsf{E})=0.95 & p(MT=\mathsf{Scot}|Cnt=\mathsf{E})=0.04 & p(MT=\mathsf{Wel}|Cnt=\mathsf{E})=0.01 \\ p(MT=\mathsf{Eng}|Cnt=\mathsf{S})=0.7 & p(MT=\mathsf{Scot}|Cnt=\mathsf{S})=0.3 & p(MT=\mathsf{Wel}|Cnt=\mathsf{S})=0.0 \\ p(MT=\mathsf{Eng}|Cnt=\mathsf{W})=0.6 & p(MT=\mathsf{Scot}|Cnt=\mathsf{W})=0.0 & p(MT=\mathsf{Wel}|Cnt=\mathsf{W})=0.4 \end{array} \tag{1.1.15}$$

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 (say) rows indexed by country and columns indexed by Mother Tongue:

$$\begin{pmatrix} 0.95 \times 0.88 & 0.7 \times 0.08 & 0.6 \times 0.04 \\ 0.04 \times 0.88 & 0.3 \times 0.08 & 0.0 \times 0.04 \\ 0.01 \times 0.88 & 0.0 \times 0.08 & 0.4 \times 0.04 \end{pmatrix} = \begin{pmatrix} 0.836 & 0.056 & 0.024 \\ 0.0352 & 0.024 & 0 \\ 0.0088 & 0 & 0.016 \end{pmatrix} \tag{1.1.16}$$

The joint distribution contains all the information about the model of this environment. By summing a column of this table, we have the marginal $p(Cnt)$. Summing the row gives the marginal $p(MT)$. Similarly, one could easily infer $p(Cnt|MT) \propto p(Cnt|MT)p(MT)$ from this joint distribution.

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.

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 $p(x_1|x_2) \neq p(x_2|x_1)$.

1.1.2 Interpreting Conditional Probability

Together with the rules of probability, conditional probability enables one to reason in a rational, logical and consistent way. One could argue that much of science deals with problems of the form : tell me something about the parameters θ given that I have observed data $\mathcal{D}$ and have some knowledge of the underlying data generating mechanism. From a modelling perspective, this requires

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} = \frac{p(D|\theta)p(\theta)}{\int_\theta p(\mathcal{D}|\theta)p(\theta)} \tag{1.1.17}$$

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 parameter values are appropriate, we can infer the *posterior* distribution $p(\theta|\mathcal{D})$ of parameters in light of the observed data.

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 correct parameters of 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, such as its mass, length and friction damping constant.

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 axioms 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 another 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 consistent framework for manipulating real values consistent with our intuition about probability[145].

1.2 Probabilistic Reasoning

The axioms of probability, combined with Bayes' rule make for a complete reasoning system, one which includes traditional deductive logic as a special case[145].

Remark 1. 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*[2] that sets variables in known states, and subsequently computing probabilities of interest, conditioned on this evidence.

Example 1 (Hamburgers)**.** Consider the following fictitious scientific information: Doctors find that people with Kreuzfeld-Jacob disease (KJ) almost invariably ate hamburgers, thus $p(Hamburger\ Eater|KJ\ ) = 0.9$. 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 $p(Hamburger\ Eater) = 0.5$, what is the probability that a hamburger eater will have Kreuzfeld-Jacob disease?

   This may be computed as

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

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

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

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

Intuitively, this is much higher than in scenario (1) since here we can be more sure that eating hamburgers is related to the illness. In this case only a small number of people in the population eat hamburgers, and most of them get ill.

Example 2 (Inspector Clouseau)**.** Inspector Clouseau arrives at the scene of a crime. The victim lies dead in the room and the inspector quickly finds the murder weapon, a Knife (K). The Butler (B) and Maid (M) are his main suspects. The inspector has a prior belief of 0.8 that the Butler is the murderer, and a prior belief of 0.2 that the Maid is the murderer. These probabilities 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:

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

$$p(B = \textsf{murderer}) = 0.8, \qquad p(M = \textsf{murderer}) = 0.2 \tag{1.2.5}$$

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

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,

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

Plugging in the values we have

$$p(B = \textsf{murderer}|\textsf{knife used}) = \frac{\frac{8}{10}\left(\frac{2}{10}\times\frac{1}{10} + \frac{8}{10}\times\frac{6}{10}\right)}{\frac{8}{10}\left(\frac{2}{10}\times\frac{1}{10} + \frac{8}{10}\times\frac{6}{10}\right) + \frac{2}{10}\left(\frac{2}{10}\times\frac{2}{10} + \frac{8}{10}\times\frac{3}{10}\right)} = \frac{200}{228} \approx 0.877 \tag{1.2.8}$$

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

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

is computed to be 0.456. But surely, $p(\textsf{knife used}) = 1$, since this is given in the question! Note that the quantity $p(\textsf{knife used})$ 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*

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

which, naturally, must be the case.

Another potential confusion is the choice

$$p(B = \textsf{murderer}) = 0.8, \qquad p(M = \textsf{murderer}) = 0.2 \tag{1.2.11}$$

which means that $p(B = \textsf{not murderer}) = 0.2$, $p(M = \textsf{not murderer}) = 0.8$. These events are not exclusive and it's just 'coincidence' that the numerical values are chosen this way. For example, we could have also chosen

$$p(B = \textsf{murderer}) = 0.6, \qquad p(M = \textsf{murderer}) = 0.9 \tag{1.2.12}$$

which means that $p(B = \textsf{not murderer}) = 0.4$, $p(M = \textsf{not murderer}) = 0.1$