Chapter 2

Introduction to probability

In this chapter, we provide a compact review of probability theory. There are very few ideas, and each is relatively simple when considered separately. However, they combine to form a powerful language for describing uncertainty.

2.1 Random variables

A random variable x denotes a quantity that is uncertain. The variable may denote the result of an experiment (e.g., flipping a coin) or a real-world measurement of a fluctuating property (e.g., measuring the temperature). If we observe several instances $\{x_i\}_{i=1}^{I}$, then it might take a different value on each occasion. However, some values may occur more often than others. This information is captured by the probability distribution $Pr(x)$ of the random variable.

A random variable may be *discrete* or *continuous*. A discrete variable takes values from a predefined set. This set may be ordered (the outcomes 1–6 of rolling a die) or unordered (the outcomes "sunny," "raining," "snowing," upon observing the weather). It may be finite (there are 52 possible outcomes of drawing a card randomly from a standard pack) or infinite (the number of people on the next train is theoretically unbounded). The probability distribution of a discrete variable can be visualized as a histogram or a Hinton diagram (figure 2.1). Each outcome has a positive probability associated with it and the sum of the probabilities for all outcomes is always one.

Continuous random variables take values that are real numbers. These may be finite (the time taken to finish a 2-hour exam is constrained to be greater than 0 hours and less than 2 hours) or infinite (the amount of time until the next bus arrives is unbounded above). Infinite continuous variables may be defined on the whole real range or may be bounded above or below (the 1D velocity of a vehicle may take any value, but the speed is bounded below by 0). The probability distribution of a continuous variable can be visualized by plotting the *probability density function* (pdf). The probability density for an outcome represents the relative propensity of the random variable to take that value (see figure 2.2). It may take any positive value. However, the integral of the pdf always sums to one.

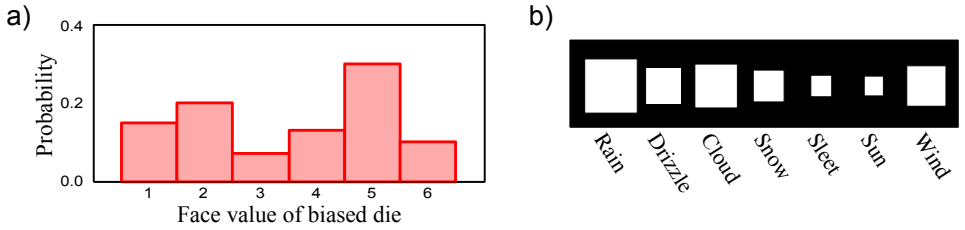
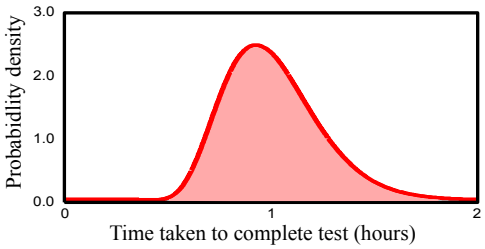


Figure 2.1 Two different representations for discrete probabilities a) A bar graph representing the probability that a biased six-sided die lands on each face. The height of the bar represents the probability: the sum of all heights is one. b) A Hinton diagram illustrating the probability of observing different weather types in England. The area of the square represents the probability, so the sum of all areas is one.

Figure 2.2 Continuous probability distribution (probability density function or pdf for short) for time taken to complete a test. Note that the probability density can exceed one, but the area under the curve must always have unit area.



2.2 Joint probability

Problem 2.1

Consider two random variables, x and y. If we observe multiple paired instances of x and y, then some combinations of the two outcomes occur more frequently than others. This information is encompassed in the *joint* probability distribution of x and y, which is written as $Pr(x, y)$. The comma in $Pr(x, y)$ can be read as the English word "and" so $Pr(x, y)$ is the probability of x *and* y. A joint probability distribution may relate variables that are all discrete or all continuous, or it may relate discrete variables to continuous ones (see figure 2.3). Regardless, the total probability of all outcomes (summing over discrete variables and integrating over continuous ones) is always one.

In general, we will be interested in the joint probability distribution of more than two variables. We will write $Pr(x, y, z)$ to represent the joint probability distribution of scalar variables x, y, and z. We may also write $Pr(\mathbf{x})$ to represent the joint probability of all of the elements of the multidimensional variable $\mathbf{x} = [x_1, x_2, \ldots, x_K]^T$. Finally, we will write $Pr(\mathbf{x}, \mathbf{y})$ to represent the joint distribution of all of the elements from multidimensional variables $\mathbf{x}$ and $\mathbf{y}$.

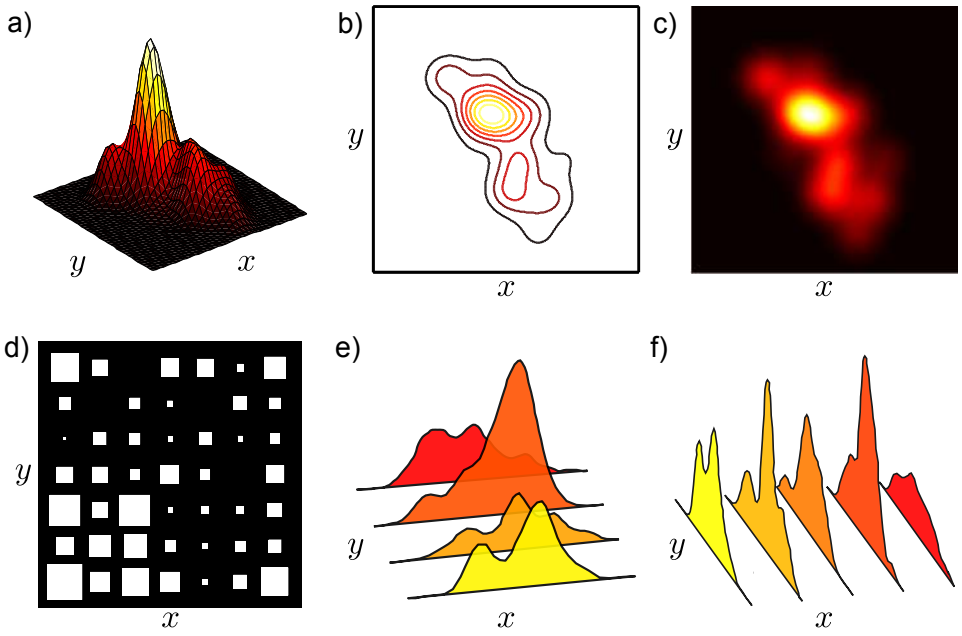


Figure 2.3 Joint probability distributions between variables x and y. a-c) The same joint pdf of two continuous variables represented as a surface, contour plot, and image, respectively. d) Joint distribution of two discrete variables represented as a 2D Hinton diagram. e) Joint distribution of a continuous variable x and discrete variable y. f) Joint distribution of a discrete variable x and continuous variable y.

2.3 Marginalization

We can recover the probability distribution of any single variable from a joint distribution by summing (discrete case) or integrating (continuous case) over all the other variables (figure 2.4). For example, if x and y are both continuous and we know $Pr(x,y)$, then we can recover the distributions $Pr(x)$ and $Pr(y)$ using the relations

$$\begin{aligned} Pr(x) &= \int Pr(x,y)\, dy, \\ Pr(y) &= \int Pr(x,y)\, dx. \end{aligned} \tag{2.1}$$

The recovered distributions $Pr(x)$ and $Pr(y)$ are referred to as *marginal* distributions, and the process of integrating/summing over the other variables is called *marginalization*. Calculating the marginal distribution $Pr(x)$ from the joint distribution $Pr(x,y)$ by marginalizing over the variable y has a simple interpretation:

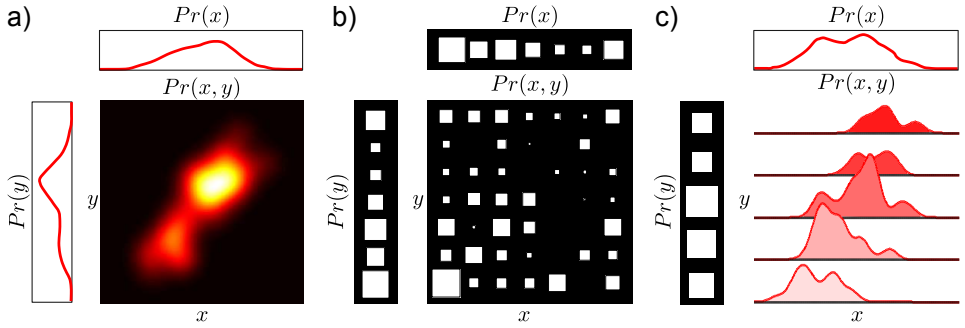


Figure 2.4 Joint and marginal probability distributions. The marginal probability $Pr(x)$is found by summing over all values of y (discrete case) or integrating over y (continuous case) in the joint distribution $Pr(x, y)$. Similarly, the marginal probability $Pr(y)$ is found by summing or integrating over x. Note that the plots for the marginal distributions have different scales from those for the joint distribution (on the same scale, the marginals would look larger as they sum all of the mass from one direction). a) Both x and y are continuous. b) Both x and y are discrete. c) The random variable x is continuous and the variable y is discrete.

we are finding the probability distribution of x regardless of (or in the absence of information about) the value of y.

Problem 2.2

In general, we can recover the joint probability of any subset of variables, by marginalizing over all of the others. For example, given variables, w, x, y, z, where w is discrete and z is continuous, we can recover $Pr(x, y)$ using

$$Pr(x, y) = \sum_{w} \int Pr(w, x, y, z)\, dz. \tag{2.2}$$

2.4 Conditional probability

The conditional probability of x given that y takes value y^* tells us the relative propensity of the random variable x to take different outcomes given that the random variable y is fixed to value y^*. This conditional probability is written as $Pr(x|y = y^*)$. The vertical line "|" can be read as the English word "given."

The conditional probability $Pr(x|y = y^*)$ can be recovered from the joint distribution $Pr(x, y)$. In particular, we examine the appropriate slice $Pr(x, y = y^*)$ of the joint distribution (figure 2.5). The values in the slice tell us about the relative probability that x takes various values having observed $y = y^*$, but they do not themselves form a valid probability distribution; they cannot sum to one as they constitute only a small part of the joint distribution which did itself sum to one. To calculate the conditional probability distribution, we hence normalize by the total probability in the slice:

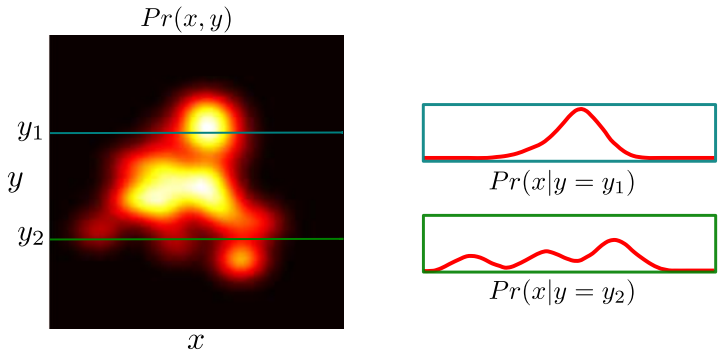


Figure 2.5 Conditional probability. Joint pdf of x and y and two conditional probability distributions $Pr(x|y=y_1)$ and $Pr(x|y=y_2)$. These are formed by extracting the appropriate slice from the joint pdf and normalizing so that the area is one. A similar operation can be performed for discrete distributions.

$$Pr(x|y=y^*) = \frac{Pr(x,y=y^*)}{\int Pr(x,y=y^*)dx} = \frac{Pr(x,y=y^*)}{Pr(y=y^*)}, \tag{2.3}$$

where we have used the marginal probability relation (Equation 2.1) to simplify the denominator. It is common to write the conditional probability relation without explicitly defining the value $y=y^*$ to give the more compact notation

$$Pr(x|y) = \frac{Pr(x,y)}{Pr(y)}. \tag{2.4}$$

This relationship can be re-arranged to give

$$Pr(x,y) = Pr(x|y)Pr(y), \tag{2.5}$$

and by symmetry we also have

$$Pr(x,y) = Pr(y|x)Pr(x). \tag{2.6}$$

When we have more than two variables, we may repeatedly take conditional probabilities to divide up the joint probability distribution into a product of terms

Problem 2.3

$$\begin{aligned} Pr(w,x,y,z) &= Pr(w,x,y|z)Pr(z) \\ &= Pr(w,x|y,z)Pr(y|z)Pr(z) \\ &= Pr(w|x,y,z)Pr(x|y,z)Pr(y|z)Pr(z). \end{aligned} \tag{2.7}$$

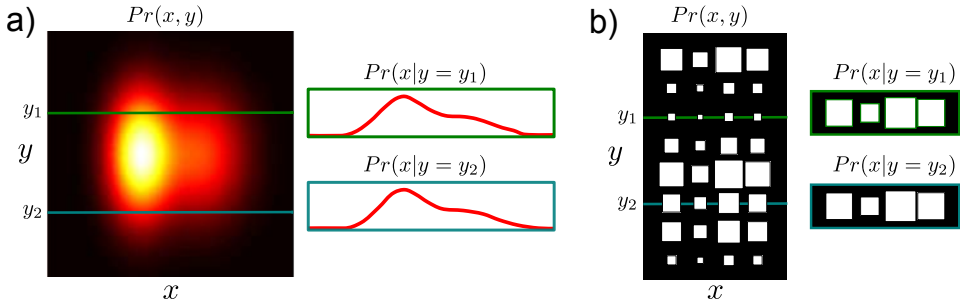


Figure 2.6 Independence. a) Joint pdf of continuous independent variables x and y. The independence of x and y means that every conditional distribution is the same: the value of y tells us nothing about x and vice-versa. Compare this to figure 2.5 which illustrated variables that were dependent. b) Joint distribution of discrete independent variables x and y. The conditional distributions of x given y are all the same.

2.5 Bayes' rule

In equations 2.5 and 2.6 we expressed the joint probability in two ways. We can combine these formulations to find a relationship between $Pr(x|y)$ and $Pr(y|x)$,

$$Pr(y|x)Pr(x) = Pr(x|y)Pr(y), \tag{2.8}$$

or, rearranging, we have

$$\begin{aligned} Pr(y|x) &= \frac{Pr(x|y)Pr(y)}{Pr(x)} \\ &= \frac{Pr(x|y)Pr(y)}{\int Pr(x,y)\, dy} \\ &= \frac{Pr(x|y)Pr(y)}{\int Pr(x|y)Pr(y)\, dy}, \end{aligned} \tag{2.9}$$

Problem 2.4

where in the second and third lines we have expanded the denominator using the definitions of marginal and conditional probability, respectively. These three equations are all commonly referred to as *Bayes' rule*.

Each term in Bayes' rule has a name. The term $Pr(y|x)$ on the left-hand side is the *posterior*. It represents what we know about y given x. Conversely, the term $Pr(y)$ is the *prior* as it represents what is known about y before we consider x. The term $Pr(x|y)$ is the *likelihood* , and the denominator $Pr(x)$ is the *evidence*.

In computer vision, we often describe the relationship between variables x and y in terms of the conditional probability $Pr(x|y)$. However, we may be primarily interested in the variable y, and in this situation Bayes' rule is exploited to compute the probability $Pr(y|x)$.

2.6 Independence

If knowing the value of variable x tells us nothing about variable y (and vice-versa) then we say x and y are independent (figure 2.6). Here, we can write

Problem 2.5

$$\begin{aligned} Pr(x|y) &= Pr(x) \\ Pr(y|x) &= Pr(y). \end{aligned} \tag{2.10}$$

Substituting into equation 2.5, we see that for independent variables the joint probability $Pr(x, y)$ is the product of the marginal probabilities $Pr(x)$ and $Pr(y)$,

Problem 2.6
Problem 2.7

$$Pr(x, y) = Pr(x|y)Pr(y) = Pr(x)Pr(y). \tag{2.11}$$

2.7 Expectation

Given a function $f[\bullet]$ that returns a value for each possible value x^* of the variable x and a probability $Pr(x = x^*)$ that each value of x occurs, we sometimes wish to calculate the *expected* output of the function. If we drew a very large number of samples from the probability distribution, calculated the function for each sample, and took the average of these values, the result would be the *expectation*. More precisely, the expected value of a function $f[\bullet]$ of a random variable x is defined as

Problem 2.8

$$\begin{aligned} \mathrm{E}[f[x]] &= \sum_x f[x]Pr(x), \\ \mathrm{E}[f[x]] &= \int f[x]Pr(x)\, dx, \end{aligned} \tag{2.12}$$

for the discrete and continuous cases, respectively. This idea generalizes to functions $f[\bullet]$ of more than one random variable so that, for example,

$$\mathrm{E}[f[x, y]] = \iint f[x, y]Pr(x, y)\, dx\, dy. \tag{2.13}$$

For some choices of the function $f[\bullet]$, the expectation is given a special name (table 2.1). Such quantities are commonly used to summarize the properties of complex probability distributions.

There are four rules for manipulating expectations, which can be easily proved from the original definition (equation 2.12):

Problem 2.9
Problem 2.10

1. The expected value of a constant κ with respect to the random variable x is just the constant itself:
$$\mathrm{E}[\kappa] = \kappa. \tag{2.14}$$

Function $f[\bullet]$	Expectation
x	mean, μ_x
x^k	k^{th} moment about zero
$(x-\mu_x)^k$	k^{th} moment about the mean
$(x-\mu_x)^2$	variance
$(x-\mu_x)^3$	skew
$(x-\mu_x)^4$	kurtosis
$(x-\mu_x)(y-\mu_y)$	covariance of x and y

Table 2.1 Special cases of expectation. For some functions $f(x)$, the expectation $\mathrm{E}[f(x)]$ is given a special name. Here we use the notation μ_x to represent the mean with respect to random variable x and μ_y the mean with respect to random variable y.

2. The expected value of a constant κ times a function $f[x]$ of the random variable x is κ times the expected value of the function:

$$\mathrm{E}[\kappa f[x]] = \kappa \mathrm{E}[f[x]]. \tag{2.15}$$

3. The expected value of the sum of two functions of a random variable x is the sum of the individual expected values of the functions:

$$\mathrm{E}[f[x] + g[x]] = \mathrm{E}[f[x]] + \mathrm{E}[g[x]]. \tag{2.16}$$

4. The expected value of the product of two functions $f[x]$ and $g[y]$ of random variables x and y is equal to the product of the individual expected values if the variables x and y are independent:

$$\mathrm{E}[f[x]g[y]] = \mathrm{E}[f[x]]\mathrm{E}[g[y]], \qquad \text{if } x, y \text{ independent.} \tag{2.17}$$

Discussion

The rules of probability are remarkably compact and simple. The concepts of marginalization, joint and conditional probability, independence, and Bayes' rule will underpin all of the machine vision algorithms in this book. There is one remaining important concept related to probability, which is *conditional independence*. We discuss this at length in chapter 10.

Notes

For a more formal discussion of probability, the reader is encouraged to investigate one of the many books on this topic (e.g., Papoulis 1991). For a view of probability from a machine learning perspective, consult the first chapter of Bishop (2006).

Problems

Problem 2.1 Give a real-world example of a joint distribution $Pr(x, y)$ where x is discrete and y is continuous.

Problem 2.2 What remains if I marginalize a joint distribution $Pr(v, w, x, y, z)$ over five variables with respect to variables w and y? What remains if I marginalize the resulting distribution with respect to v?

Problem 2.3 Show that the following relation is true:

$$Pr(w, x, y, z) = Pr(x, y)Pr(z|w, x, y)Pr(w|x, y).$$

Problem 2.4 In my pocket there are two coins. Coin 1 is unbiased, so the likelihood $Pr(h = 1|c = 1)$ of getting heads is 0.5 and the likelihood $Pr(h = 0|c = 1)$ of getting tails is also 0.5. Coin 2 is biased, so the likelihood $Pr(h = 1|c = 2)$ of getting heads is 0.8 and the likelihood $Pr(h = 0|c = 2)$ of getting tails is 0.2. I reach into my pocket and draw one of the coins at random. There is an equal prior probability I might have picked either coin. I flip the coin and observe a head. Use Bayes' rule to compute the posterior probability that I chose coin 2.

Problem 2.5 If variables x and y are independent and variables x and z are independent, does it follow that variables y and z are independent?

Problem 2.6 Use equation 2.3 to show that when x and y are independent, the marginal distribution $Pr(x)$ is the same as the conditional distribution $Pr(x|y = y^*)$ for any y^*.

Problem 2.7 The joint probability $Pr(w, x, y, z)$ over four variables factorizes as

$$Pr(w, x, y, z) = Pr(w)Pr(z|y)Pr(y|x, w)Pr(x).$$

Demonstrate that x is independent of w by showing that $Pr(x, w) = Pr(x)Pr(w)$.

Problem 2.8 Consider a biased die where the probabilities of rolling sides $\{1, 2, 3, 4, 5, 6\}$ are $\{1/12, 1/12, 1/12, 1/12, 1/6, 1/2\}$, respectively. What is the expected value of the die? If I roll the die twice, what is the expected value of the sum of the two rolls?

Problem 2.9 Prove the four relations for manipulating expectations.

$$\begin{aligned}
\mathrm{E}[\kappa] &= \kappa, \\
\mathrm{E}[\kappa f[x]] &= \kappa \mathrm{E}[f[x]], \\
\mathrm{E}[f[x] + g[x]] &= \mathrm{E}[f[x]] + \mathrm{E}[g[x]], \\
\mathrm{E}[f[x]g[y]] &= \mathrm{E}[f[x]]\mathrm{E}[g[y]], \qquad \text{if } x, y \text{ independent.}
\end{aligned}$$

For the last case, you will need to use the definition of independence (see section 2.6).

Problem 2.10 Use the relations from problem 2.9 to prove the following relationship between the second moment around zero and the second moment about the mean (variance):

$$E\left[(x-\mu)^2\right] = E\left[x^2\right] - \mathrm{E}[x]\mathrm{E}[x].$$