Mathematical Preliminaries

Linear Algebra

Many problems in machine learning can be naturally expressed as systems of linear equations. In this section, we briefly describe how linear equations can be represented using the language of vectors and matrices.

Consider a set of N numbers $(x_1, x_2, \ldots, x_N)$. These numbers could represent, for example, the coordinates of a point in 3D space (when N=3); they could represent some measurement; they could even be the intensities of pixels in an image. We will require that they all have the same "unit" (distance, time, chemical concentraion, intensity etc). Such sets of numbers are extremely common in machine learning. Many machine learning methods will involve the construction of models that map one such set of numbers onto some other set of numbers, and the goal wil be to learn the model that performs this mapping. A common form of such a model is a *linear mapping* that constructs a second set of numbers $(y_1, y_2, \ldots, y_M)$ by adding up the x's in some weighted combination, i.e:

$$\begin{aligned}
y_1 &= A_{11}x_1 + A_{12}x_2 + \cdots + A_{1N}x_N \\
y_2 &= A_{21}x_1 + A_{22}x_2 + \cdots + A_{2N}x_N \\
&\ldots \\
y_M &= A_{M1}x_1 + A_{M2}x_2 + \cdots + A_{MN}x_N \qquad (1)
\end{aligned}$$

Note that the label on the A's has two components: one corresponding to the label on the associated y and one corresponding to the label on the associated x. This will be important.

It is clear that working with a model like this will quickly become very tiresome and tedious unless we can find a more efficient way to represent this problem. That language is linear algebra. In this problem, we have three mathematical "objects": the set of x's, the set of y's and the set of A's. The x and y each represent something – a position in space, an image etc. A defines the relationsip between them. That is, given the x's and A's, one can calculate the y's.

A very convenient way of working with this sort of problem is to adopt the following notation. We consider the x's and y's to be *components* of a quantity called a vector. We write

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix} \qquad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_M \end{pmatrix} \tag{2}$$

We will use lower-case, bold-face letters to represent vectors such as these. Note that we write the vector as a *column* rather than as a *row* for reasons of convenience.

We now need a similar representation for the *a*'s. The form of the equations above gives us a hint as to how we might right this down. Noting that the *a*'s have two labels, we will write them as a two-dimensional *matrix*

$$\mathbf{A} = \begin{pmatrix} A_{11} & A_{12} & \dots & A_{1N} \\ A_{21} & A_{22} & \dots & A_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ A_{M1} & A_{M2} & \dots & A_{MN} \end{pmatrix} \tag{3}$$

The labels (indices) on the *A*'s denote which row and column of the matrix they are in. We denote matrices using a bold-face upper-case letter.

This representation now allows us to write a compact shorthand for our linear mapping:

$$\mathbf{y} = \mathbf{A}\mathbf{x} \tag{4}$$

We will describe **A** and **x** as being *multiplied*, by analogy with the operation with which we are familiar, but clearly not quite the same. To see how this work, we write this expression out in full as

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_M \end{pmatrix} = \begin{pmatrix} A_{11} & A_{12} & \dots & A_{1N} \\ A_{21} & A_{22} & \dots & A_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ A_{M1} & A_{M2} & \dots & A_{MN} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix} \tag{5}$$

Comparing this with our original equations (1) we see how how the components of **A** and **x** are combined in matrix multiplication: each row of **y** is formed by taking the corresponding row of **A**, multiplying its components by the components of **x** and then summing the products, i.e.

$$y_i = \sum_{j=1}^{N} A_i j x_j. \tag{6}$$

Notice that this requires that **A** has the same number of rows as **y** has components, and the same number of columns as **x** has components.

This defines a neat and tidy notation through which we can represent linear mappings. Notice that sets of simultaneous linear equations can be represented in this notation. For example, the equations

$$\begin{aligned} 2a - b &= 3 \\ a + b &= 3 \end{aligned}$$

can be equivalently written as

$$\begin{pmatrix} 2 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 3 \\ 3 \end{pmatrix}$$

Operations on matrices and vectors

Matrices and vectors can be added together provided that they are of the same size (number of rows and columns). This is a simple operation: the result of an addition (or subtraction) is a matrix or vector of the same size, with the components added (subtracted). That is,

$$\mathbf{a} = \mathbf{b} + \mathbf{c} \quad \to \quad a_i = b_i + c_i \tag{7}$$

$$\mathbf{A} = \mathbf{B} + \mathbf{C} \quad \to \quad A_{ij} = B_{ij} + C_{ij} \tag{8}$$

Similarly, multiplication by a scalar number multiplies each of the components:

$$\mathbf{a} = \alpha\mathbf{b} \quad \to \quad a_i = \alpha b_i \tag{9}$$

$$\mathbf{A} = \alpha\mathbf{B} \quad \to \quad A_{ij} = \alpha B_{ij} \tag{10}$$

The *length* or *magnitude* of a vector can be calculated as the square-root of the sum of the squares of its components (courtesy of Pythagoras' Theorem for calculating the hypotenuse of a triangle):

$$|x| = \sqrt{x_1^2 + x_2^2 + \cdots + x_N^2} = \sqrt{\sum_{i=1}^{N} x_i} = \sqrt{\mathbf{x}^{\mathrm{T}}\mathbf{x}} \tag{11}$$

where $\mathbf{x}^{\mathrm{T}}$ is the *transpose* of $\mathbf{x}$, which is the vector written as a row rather than as a column. If one considers a vector to be a special case of a matrix with only one row/column, then it is easy to verify that this is entirely consistent with the rules of matrix/vector multiplication.

It is also possible to combine two matrices together. this is very similar to matrix-vector multiplication, except we have multiple columns:

$$\mathbf{A} = \mathbf{XY} \quad \to \quad A_{ik} = \sum_j X_{ij} Y_{jk} \tag{12}$$

For example:

$$\begin{pmatrix} 1 & 2 \\ 2 & 3 \\ 3 & 1 \end{pmatrix} \begin{pmatrix} 2 & 1 & 4 & 1 \\ 1 & 3 & 2 & 1 \end{pmatrix} = \begin{pmatrix} 1\times2+2\times1 & 1\times1+2\times3 & 1\times4+2\times2 & 1\times1+2\times1 \\ 2\times2+3\times1 & 2\times1+3\times3 & 2\times4+3\times2 & 2\times1+3\times1 \\ 3\times2+1\times1 & 3\times1+1\times3 & 3\times4+1\times2 & 3\times1+1\times1 \end{pmatrix} = \begin{pmatrix} 4 & 7 & 8 & 3 \\ 7 & 11 & 14 & 5 \\ 7 & 6 & 14 & 4 \end{pmatrix}$$

This can be used to construct *compositions* of linear mappings: the matrix $\mathbf{A}$ represents the linear mapping $\mathbf{Y}$ followed by the linear mapping $\mathbf{X}$. Note the order here: in the multiplication $\mathbf{XYx}$, $\mathbf{X}$ is applied to $\mathbf{Yx}$.

One special case of this is of particular interest. A special matrix that we will see quite frequently is the *identity* matrix

$$\mathbf{I} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix} \tag{13}$$

This matrix has the property that $\mathbf{Ix} = \mathbf{x}$ (and similarly $\mathbf{IA} = \mathbf{A}$). A special case of matrix compositions is the case where $\mathbf{XY} = \mathbf{I}$. That is, $\mathbf{XYx} = \mathbf{x}$. The linear transformation performed by $\mathbf{X}$ has reversed the transformation performed by $\mathbf{Y}$. We will refer to $\mathbf{X}$ as being the *inverse* of $\mathbf{Y}$ and will write this is $\mathbf{X} = \mathbf{Y}^{-1}$. As an example of where this can be useful, consider the linear equations we wrote down earlier:

$$\begin{pmatrix} 2 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 3 \\ 3 \end{pmatrix} \tag{14}$$

$$\mathbf{Ax} = \mathbf{y} \tag{15}$$

Using the tools we have developed so far, we do not know how to find a and b. However, the concept of a *matrix inverse* exactly what we need, and if we can find $\mathbf{A}^{-1}$, we can do the calculation

$$\mathbf{Ax} = \mathbf{y} \quad \rightarrow \quad \mathbf{A}^{-1}\mathbf{Ax} = \mathbf{A}^{-1}\mathbf{y} \quad \rightarrow \quad \mathbf{Ix} = \mathbf{A}^{-1}\mathbf{y} \quad \rightarrow \quad \mathbf{x} = \mathbf{A}^{-1}\mathbf{y} \tag{16}$$

It is straightforward to verify (by establishing that $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$) that

$$\mathbf{A} = \begin{pmatrix} 2 & -1 \\ 1 & 1 \end{pmatrix} \longrightarrow \mathbf{A}^{-1} = \frac{1}{3} \begin{pmatrix} 1 & 1 \\ -1 & 2 \end{pmatrix}, \tag{17}$$

and this allows us to solve the equations to find that $a = 2$ and $b = 1$, but we will not discuss the methods used to compute a matrix inverse (one such method is *Cramer's rule*) because it is not a calculation that is usually necessary in practice. The inverse of a matrix can only be computed exactly for square matrices with certain properties, and cannot be computed at all for non-square matrices ($N \neq M$). Even in the cases when the inverse can be computed, it is a computationally expensive operation and there are alternative methods such as Gaussian elimination that can be to solve the problem (finding a and b) directly without inverting $\mathbf{A}$. These methods are typically much more efficient, especially for very large systems of equations.

Students who would like to develop their knowledge of linear algebra are advised to consider Gibert Strang's excellent online course which has been made available vi MIT Open Courseware and can be found at `https://ocw.mit.edu/courses/mathematics/18-06-linear-algebra-spring-2010/`.

A Very Brief Introduction to Differentiation

A recurring theme in machine learning is the idea that we can describe how well a model is performing by defining a cost or *loss* function measures how closely the model matches the data and therefore takes high values for poor models, and low values for good models. The process of learning then becomes one of finding the model that minimises the cost function which is normally defined such that the cost of a model that perfectly matches the data is zero, and this is an absolute lower bound. Our model will not normally be able to reach this lower bound, but there will be some optimal value of its parameters that gets us as close as possible. If the optimal parameters are changed, even by a small amount, the cost of the model will increase, as shown in a very simple case in Figure 1, where the value $x = 5$ minimises the value of y.

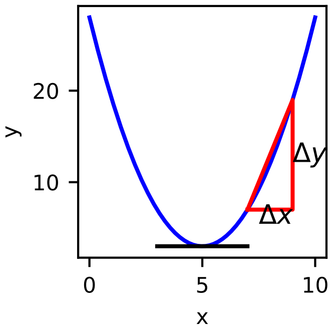

Figure 1: A schematic illustration of the process of computing the derivate (gradient) of a function.

The minimum point of a function that behaves in this way has a special property: at the minimum, it has a gradient, or slope, of zero (it is perfectly "flat") as indicated by the black line. The process of finding the minimum of the function then becomes one of finding the point where the gradient is zero.

The process of computing the gradient (also called the derivative) of an function is known as *differentiation*. This can be an intricate process and we will only sketch out the principles here. The basic idea is that we approximate the function as a straight line segment and compute the gradient of that straight line. We then make the straight line shorter and figure out how the derivative behaves as the length of the segment approaches zero to compute the gradient at a single point.

Consider the red triangle of width Δx and height Δy shown in Figure 1. The gradient of the hypotenuse of this triangle is $m = \Delta y / \Delta x$. At a general point x, the gradient of a function $f(x)$ will be

$$m \approx \frac{\Delta y}{\Delta x} = \frac{f(x + \Delta x) - f(x)}{\Delta x} \tag{18}$$

We now reduce Δx to zero to find the gradient at a single point. Formally, we will write this as

$$m_x = \frac{\mathrm{d}y}{\mathrm{d}x} = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x} \tag{19}$$

Where $\lim_{\Delta x \to 0}$ means "take Δx to 0", m_x is the gradient of $f(x)$as a function of x, and the notation $\frac{\mathrm{d}y}{\mathrm{d}x}$ is a formal way of writing the derivative which implies the process of taking the limit. Let us illlustrate with a couple of examples. First, one where we know the answer: $y = 3x + 2$, which is of the well-know form for a straight line ($y = mx + c$) and is known to have gradient $m = 3$.

$$\begin{aligned} m_x = \frac{\mathrm{d}y}{\mathrm{d}x} = \lim_{\Delta x \to 0} & \frac{f(x + \Delta x) - f(x)}{\Delta x} \\ & = \frac{3(x + \Delta x) + 2 - 3x + 2)}{\Delta x} = \frac{3\Delta x}{\Delta x} = 3. \end{aligned} \tag{20}$$

In fact, we didn't need to take a limit here because all of the non-constant terms cancelled out. Let us try a different example: $y = x^2$:

$$\begin{aligned} m_x = \frac{\mathrm{d}y}{\mathrm{d}x} &= \lim_{\Delta x \to 0} \frac{f(x+\Delta x) - f(x)}{\Delta x} \\ &= \lim_{\Delta x \to 0} \frac{(x+\Delta x)^2 - x^2)}{\Delta x} \\ &= \lim_{\Delta x \to 0} \frac{x^2 + 2x\Delta x + (\Delta x)^2 - x^2}{\Delta x} \\ &= \lim_{\Delta x \to 0} \frac{2x\Delta x + (\Delta x)^2}{\Delta x} \end{aligned}$$

Now, in the limit $\Delta x \to 0$, the term $(\Delta x)^2$ becomes vanishingly small and we are left with the result that $\frac{\mathrm{d}y}{\mathrm{d}x} = 2x$ for $y = x^2$.

A general and frequently used result is that for polynomials is that for $y = x^n$, $\frac{\mathrm{d}y}{\mathrm{d}x} = nx^{n-1}$. Note also that differentiation is a linear operation: if $y = f(x) + g(x)$

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{\mathrm{d}f(x)}{\mathrm{d}x} + \frac{\mathrm{d}g(x)}{\mathrm{d}x}. \tag{21}$$

Random Variables

A variable that is said to be random is one that is subject to some degree of nondeterminism. For example, the draw of a card from a deck, the role of a die, or the arrival time of a bus/train all have some element of randomness. In this section, we will briefly review some important properties of random variables and present some important results that we will need to use.

Let us consider first a simple case of rolling a die. There are six distinct outcomes, and in the case of an unbiased die, these all have an equal chance of occurring. We can write down the probability of each outcome very easily, provided that we follow some simple rules.

1. A probability of 1 for an outcome means that outcome is certain. For examples, if all faces of the die had three dots, $P(3) = 1$.

2. One of the possible outcomes *must* occur, and no other outcome can occur, so the probabilities must add to 1.

3. No outcome can have a probability of greater than one – this would imply that an outcome occurs more frequently than every time!

4. No outcome can have a negative probability.

The "probability matrix" for the role of a single fair die, following these rules, is shown in Table 1. This is an example of a *discrete* probability distribution.

X	1	2	3	4	5	6
P(X)	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Table 1: The probability matrix of the outcomes of a single role of a die.

Now consider how this matrix would change if we had a biased die, for which the probability of rolling a 6 is twice that of rolling any of the other numbers. We may be tempted to change $P(X = 6)$

to $\frac{2}{6}$ in Table 1, but this would not be correct: it would mean that the sum of the probabilities of the possible outcomes would be $\frac{7}{6}$, which violates rule 2 above. We need to *normalise* the outcomes such that this rule is obeyed. We do this by rescaling all of the probabilities by the sum of the outcomes, that is

$$P(X = X_i) \rightarrow \frac{P(X = X_i)}{\sum_i P(X = X_i)} \tag{22}$$

In the case of our biased die, the relative probabilities of the size outcomes are (in order): $\{1, 1, 1, 1, 1, 2\}$ and so $\sum_i P(X = X_i) = 7$, and the probabilities becomes $\left\{\frac{1}{7}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7}, \frac{2}{7}\right\}$.

Now consider the case of rolling *two* die. Each die has six possible outcomes, and so there are 36 possible outcomes of rolling two dice. Again, following our rules, these are shown in Table 2.

This is known as the *joint* probability of X_1 and X_2. Because X_1 and X_2 are *independent* random variables, the joint distribution of the outcomes is formed from the product of the probabilities of the independent variables, that is, $P(X_1, X_2) = P(X_1)P(X_2)$ or in general:

$$P(X_1, X_2, \ldots, X_N) = \prod_{i=1}^{N} P(X_i) \tag{23}$$

X_1 X_2	1	2	3	4	5	6
1	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
2	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
3	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
4	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
5	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
6	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$

Table 2: The joint distribution of the outcomes of rolling two fair dice.

From the joint probability, we can compute the probability of any possible outcome. For example, $P(X_1 + X_2)$ is obtained by summing the relevant cells in Table 2, as shown in Table 3 :

$X_1 + X_2$	2	3	4	5	6	7	6	9	10	11	12
$P(X_1 + X_2)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Table 3: The probability of the sum of two rolls of a fair die.

For our biased die, we can form a similar table (Table 4). This is a little more interesting because it has some heterogeneity in it, but the two variables are still independent. Notice that this table, and indeed all of the tables obey the rules of probability: no outcome has a probability greater than one, all have probabilities greater than zero, and the sum of all the probabilities is one.

X_1 X_2	1	2	3	4	5	6
1	$\frac{1}{49}$	$\frac{1}{49}$	$\frac{1}{49}$	$\frac{1}{49}$	$\frac{1}{49}$	$\frac{2}{49}$
2	$\frac{1}{49}$	$\frac{1}{49}$	$\frac{1}{49}$	$\frac{1}{49}$	$\frac{1}{49}$	$\frac{2}{49}$
3	$\frac{1}{49}$	$\frac{1}{49}$	$\frac{1}{49}$	$\frac{1}{49}$	$\frac{1}{49}$	$\frac{2}{49}$
4	$\frac{1}{49}$	$\frac{1}{49}$	$\frac{1}{49}$	$\frac{1}{49}$	$\frac{1}{49}$	$\frac{2}{49}$
5	$\frac{1}{49}$	$\frac{1}{49}$	$\frac{1}{49}$	$\frac{1}{49}$	$\frac{1}{49}$	$\frac{2}{49}$
6	$\frac{2}{49}$	$\frac{2}{49}$	$\frac{2}{49}$	$\frac{2}{49}$	$\frac{2}{49}$	$\frac{4}{49}$

Table 4: The joint distribution of the outcomes of two rolls of a biased die.

Let us consider a different example with more illustrative power. The variables are again independent, but we can ask richer questions. Our chosen example is a diagnostic medical test, in which the outcome of the test depends on the whether a patient has a disease. The test has the following properties:

- If the patient has the disease, the probability of a positive test is 0.9 (a *true positive*).
- If the patient does not have the disease, the probability of a positive test is 0.05 (*false positive*).

From these, we can also determine that the probability of the patient with a disease giving a negative test is 0.1, and that the probability of the patient without a disease giving a negative test is

0.95, by observing that the probabilitie of positive and negative tests must add up to one.

These are *conditional* probabilities, which we denote $P(X|Y)$: the value of X (the test) depends on the value of Y (the disease state). We use X to denote the test, and Y to denote the didease state, and noting that both can be either true (T) or false (F) we have

$$\begin{aligned} P(X=T|Y=T) &= 0.9 \quad (24)\\ P(X=T|Y=F) &= 0.05 \quad (25)\\ P(X=F|Y=T) &= 0.1 \quad (26)\\ P(X=F|Y=F) &= 0.95 \quad (27) \end{aligned}$$

We may also have access to some additional information. We know that the disease is quite rare, with only 0.1 of the population suffering from it, so $P(Y = T) = 0.1$ and $P(Y = F) = 0.9$. This information allows us to form the joint distribution by noting that (for example) $P(X = T, Y = T)$ is formed by weighting the conditional probability $P(X = T|Y = T)$ by the probability that $Y = T$. That is, $P(X = T, Y = T) = P(X = T|Y = T)P(Y = T)$. The full joint probability distribution is shown in Table 5

X Y	T	F
T	0.09	0.01
F	0.045	0.855

Table 5: The joint distribution of the outcome of a medical test X and the prevalence of the disease Y.

In building this table we have made explicit use of two fundamental rules of probability: the **sum rule**

$$P(X) = \sum_{\{Y\}} P(X,Y) \quad (28)$$

(where $\{Y\}$ is the set of all possible values Y can take); and the **product rule**

$$P(X,Y) = P(X|Y)P(Y). \quad (29)$$

We can ask some interesting questions of this data. First, we verify what we already know. Using the sum rule, we add across the rows to compute the probabilities of disease/not disease: $P(Y = T) = 0.09 + 0.01 = 0.1$ and $P(Y = F) = 0.045 + 0.855 = 0.9$.

We also compute the probability of positive/negative tests: $P(X = T) = 0.09 + 0.045 = 0.135$ and $P(X = F) = 0.01 + 0.855 = 0.865$. This is interesting: tests are much more likely to be negative, because the test is quite accurate and the disease prevalence is low.

A much more interesting and powerful question is to ask: if the test is positive, what is the probability that the patient has the disease (i.e. $P(Y = T|X = T)$)? Our test appears to be very good: it has a high false positive and low false positive rate, but how do we really interpret its results?

To answer this question, we make an observation about conditional probabilities. Consider the product rule $P(X,Y) = P(X|Y)P(Y)$. The joint distribution is symmetric by definition: $P(X,Y) = P(Y,X)$ so we also have $P(X,Y) = P(Y|X)P(X)$. It follows that $P(X|Y)P(Y) = P(Y|X)P(X)$, and that leads to

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \quad (30)$$

This is known as *Bayes' Rule*, and it is a very powerful tool with which to reason about probabilities. A loosely intuitive explanation of Bayes' rule is that is our knowledge of Y following (*posterior* to) a measurement of X, $P(Y|X)$, depends on the *likelihood* $P(X|Y)$ of measuring X given Y, given *prior* knowledge of Y, $P(Y)$. $P(X)$ is a normalising factor.

Let us formulate our question in these terms. We want to compute $P(Y = T|X = T)$ which, from Bayes' rule, is $P(Y = T|X = T) = P(X = T|Y = T)P(Y = T)/P(X = T)$. We know all of these numbers: $P(X = T|Y = T) = 0.9$, $P(Y = T) = 0.1$, and $P(X = T) = 0.135$. We therefore have that $P(Y = T|X = T) = 0.9 \times 0.1/0.135 = 0.67$: only 2 out of 3 positive tests is from a patient who has the disease. Notice that we could also have derived this directly from the joint distribution: $P(Y = T|X = T) = P(X = T, Y = T)/P(X = T)$.

Despite the fact that the test is very good, the outcomes are surprisingly unreliable. This is a consequence of the rarity of the disease: although the false positive rate is very low (5%), the proportion of patients who could give rise to this (90%) is very large. This demonstrates the importance of taking into account prior information when modelling an inference or learning problem. Inference via Bayesian means will be an important tool in our studies of machine learning.

Our discussions to the point have been restricted to variables that take discrete values, but everything was have discussed can be transferred quite straightforwardly to the case of continuous variables, subject to a few technical modifications. Whilst for discrete variables we have worked with tables of probabilities, for continuous variables we will work with **probability density functions** (PDF) which describes how the possible values of the variable are distributed. These are entirely analogous to the tables of probabilities that we have considered in the discrete case with one important difference: in the discrete case, we model the probability that the random variable takes a certain value; in the continuous case we cannot do this, because there are infinitely many values and by definition the probability that the variable will take a precise value is zero. However, the probability that the variable will lie within some range of values is finite, and is given by the area under the PDF in the range of interest.

All PDFs obey some important properties that are almost the same as discrete probability distributions, but with some important differences. A PDF $P(x)$ must have the following properties:

1. $P(x) \geq 0$ for all values of x
2. The area under $P(x)$ must be equal to one.
3. The probablility that x lies between value a and b $P(a \leq x \leq b)$ is given by the *area* under the PDF in that range of x values

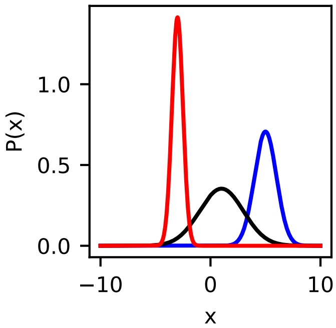

Figure 2: Three examples of normal distributions. Red: $\mu = -3, \sigma = 0.5$; Blue: $\mu = 5, \sigma = 1$; Black: $\mu = 1, \sigma = 2$.

The main differences are that because continuous variable can take infinitely many values, by definition the probability that the variable will take a precise value is zero and it is only meaningful to talk about ranges of values (point 3 above). A second difference is that the PDF can take values greater than one provided that the total area under the PDF is exactly equal to one.

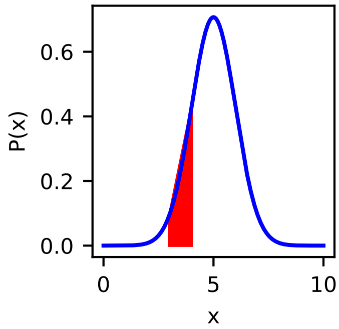

Figure 3: A example of a probability density function, in this case a normal distribution (Gaussian) $\mathcal{N}(x|5,1)$, ie with mean $\mu = 5$ and standard deviation $\sigma = 1$. The area of the red shaded region is the probability that x has a value between 3 and 4.

Let us consider an example to illustrate these ideas. The *Gaussian*, or *normal* distribution describes many common physical and statistical processes, especially those involving large numbers of independent random variables. In particular, the average values of random variables drawn from independent distributions can be shown to be normally distributed. Physical measurements, which are frequently the result of a large number of independent random processes are very often normally distributed. The normal distribution is also commonly used as a surrogate distribution for variables where the true distribution is unknown (although this should be done with care). For a single random variable, the normal distribution is given by

$$p(x|\mu,\sigma) = \mathcal{N}(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{31}$$

Notice that we have written this as a conditional probability: it is conditional on the parameters μ and σ which control the centre and width of the distribution respectively. Three examples of Gaussian PDFs for different combinations of μ and σ are shown in Figure 2. Notice that the PDF can take values of greater than one provided the total area underneath is exactly one.

As with all PDFs, the normal distribution describes the probability that a value drawn from it will lie in a certain range. Concretely, the area under any portion of this curve bounded by two values of x is the probability that a values drawn from that distribution will like between those two values. This is illustrated in Figure 3, in which the red shaded area represents $P(3 \leq x \leq 4)$.

The normal distribution can also be generalised to the multivariate case to represent a joint distribution. For two independent variables x and y, it has the form

$$\mathcal{N}(x|\mu_x,\mu_y,\sigma_x,\sigma_y) = \frac{1}{2\pi(\sigma_x\sigma_y)} \exp\left(-\left(\frac{(x-\mu_x)^2+}{2\sigma_x^2} + \frac{(y-\mu_y)^2+}{2\sigma_y^2}\right)\right) \tag{32}$$

and this is shown in Figure 4. In higher dimensions, the general form for dependent variables is

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^N\,|\boldsymbol{\Sigma}|}} \exp\left(-\frac{(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\,\boldsymbol{\Sigma}^{-1}\,(\mathbf{x}-\boldsymbol{\mu})}{2}\right) \tag{33}$$

where $\boldsymbol{\mu}$ is a vector of the mean values of each variable, $\boldsymbol{\Sigma}$ is the *covariance* matrix and $|\boldsymbol{\Sigma}|$ is its *determinant*. We will study these in more detail when ee need them.

Performing detailed calculations with PDFs generally requires an extensive knowledge of integral calculus to compute the desired areas under the PDF. This is beyond the scope of what we can

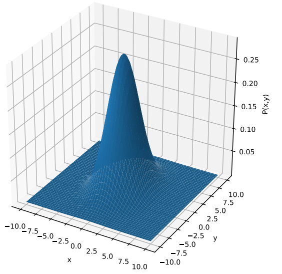

Figure 4: A example of a 2d normal PDF, with $\mu_x = \mu + y = 0$, $\sigma_x = 2$ and $\sigma_y = 3$.

cover in-depth during this module. For the moment, the key is to recognise the following points:

- Integration can be very loosely defined as a technical process for computing the area under a curve.
- Integration acheives this by summing a series of small strips under the curve, in the limit that those strips become vanishingly small.
- The process of integration is represented by the notation

  $$I(a, b) = \int_a^b f(x)dx \tag{34}$$

  which is a statement that we are computing the area under the curve $f(x)$ between $x = a$ and $x = b$.

You need to be able to recognise integration when you see it, and understand what is being achieved by integration. You will not be expected to calculate any integrals. It is worth noting that integration and differentiation are intimately related: they are essentially inverse processes. However there are some caveats to this. For example, differentiation irreversibly loses information about constant terms because they do not affect the gradient of a curve, whereas constant terms do affect the area under the curve, so some care is needed. Issues such as this will be explained when we meet them.