Learning from Data
Mathematics Primer

Gavin Brown

School of Computer Science
University of Manchester

This document reviews the background mathematics necessary for you to comfortably complete the Learning from Data theme. The first module will start with the basics, but ramp up quickly, especially as you start the second module.

You are **not expected** to understand all of this material before the first module begins. However, **if you believe, with a few weeks' hard work**, you could master most (not necessarily all) of it, then fine, if not, then maybe the course is not for you.

1 Algebra

1.1 Sums and products

We write a sum of N terms as,

$$\sum_{i=1}^{N} f(i) = f(1) + f(2) + \ldots + f(N) \ .$$

Sometimes we will write sums over a set of values which aren't in a natural sequence from 1 to N. For example, we might wish to sum up the number of each nucleotide (letter) in a DNA sequence. Call this number $n(x)$ where $x \in \{A, C, T, G\}$ is one of the letters in the DNA alphabet. We could then write the total number of nucleotides in the DNA sequence as,

$$\sum_{x \in \{A,C,T,G\}} n(x) \ = \ n(A) + n(C) + n(T) + n(G) \ .$$

In some cases it will be cumbersome to write the terms which the sum is over and we may just write,

$$\sum_{x} n(x) \ = \ n(A) + n(C) + n(T) + n(G)$$

where it is understood that the sum is over all possible values that x may take. This allows us to write general expressions where the range of the sum will depend on the context and is particularly useful when writing mathematical expressions which hold for many different cases.

In a similar way to sums of many terms we can write a product of many terms as,

$$\prod_{i=1}^{N} f(i) \quad = \quad f(1)f(2)f(3)\ldots f(N) \ .$$

1.2 Logarithms

A very important function is the logarithm. In general we define this function by the equation,

$$\log_a a^n = n$$

where $\log_a x$ is called the logarithm of x to base a. The most usual logarithms are base 2 ($\log_2 x$), base 10 ($\log_{10} x$) and base e ($\log_e x$, also written as $\ln x$).

All logarithms are proportional to one another, i.e. they only differ by a constant factor (e.g. $\log_2 x \propto \ln x$). In many formulas we will therefore not need to specify the base of logarithm used, since this will not usually affect the results. If one logged number is larger than another, this will remain the case when we change base.

Important formulas for manipulating powers and logarithms (to any base) are,

$$\begin{aligned}
(a^x)(a^y) &= a^{x+y} \ , \quad \frac{1}{a^x} = a^{-x} \ , \quad \frac{a^x}{a^y} = a^{x-y} \ , \\
\log(xy) &= \log x + \log y \ , \quad \log(x^n) = n \log x \ , \\
\log\left(\frac{x}{y}\right) &= \log x - \log y \ , \quad \log \prod_{i=1}^{N} f(i) = \sum_{i=1}^{N} \log f(i) \ .
\end{aligned}$$

Notice the last equation shows that a product of numbers can be converted into a sum of logarithms. Before the advent of calculators, log tables could be used to help with calculations involving the multiplication and division of large numbers. Logarithms are still useful now to help computers deal with arithmetic involving very large or very small numbers.

1.3 Vectors and matrices

This course deals extensively with vectors and matrices. Matrices are also the basic data-type used in the Matlab programming language used in the labs. We will mostly use bold fonts to denote vectors (lower case bold) and matrices (upper case bold).

For example, we could write a 3-dimensional vector $\boldsymbol{x}$ as,

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

where x_i is the value in the ith row of the vector. We write a 3×2 matrix $\boldsymbol{A}$ as,

$$\boldsymbol{A} = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \\ A_{31} & A_{32} \end{pmatrix}$$

where A_{ij} is the value in the ith row and jth column of the matrix. Notice that a column vector is just a matrix with a single column. This is important, because it means we can multiply vectors and matrices together and we can use matrix operations like transposition on vectors.

Addition and multiplication

We can add vectors and matrices if they are the same size. We just add all the corresponding elements,

$$\begin{pmatrix} 1 & 3 \\ 4 & 5 \\ 2 & 6 \end{pmatrix} + \begin{pmatrix} 5 & 4 \\ 10 & 2 \\ 1 & 6 \end{pmatrix} = \begin{pmatrix} 6 & 7 \\ 14 & 7 \\ 3 & 12 \end{pmatrix} \ .$$

Two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ can be multiplied together if the number of rows in $\boldsymbol{B}$ equals the number of columns in $\boldsymbol{A}$. In this case we write,

$$\boldsymbol{AB} = \boldsymbol{C} \quad \text{where} \quad C_{ik} = \sum_{j=1}^{N} A_{ij} B_{jk}$$

where N is the number of columns in $\boldsymbol{A}$ and rows in $\boldsymbol{B}$. E.g.

$$\begin{pmatrix} 1 & 3 & 2 \\ 4 & 5 & 1 \end{pmatrix} \begin{pmatrix} 5 & 4 \\ 10 & 2 \\ 1 & 6 \end{pmatrix} = \begin{pmatrix} (1\times 5 + 3\times 10 + 2\times 1) & (1\times 4 + 3\times 2 + 2\times 6) \\ (4\times 5 + 5\times 10 + 1\times 1) & (4\times 4 + 5\times 2 + 1\times 6) \end{pmatrix}$$
$$= \begin{pmatrix} 37 & 22 \\ 71 & 32 \end{pmatrix} .$$

Matrix multiplication is not commutative in general and $\boldsymbol{AB} \neq \boldsymbol{BA}$. One must therefore respect the order of multiplication, so that,

$$\begin{aligned} \boldsymbol{A}(\boldsymbol{B}+\boldsymbol{C}) &= \boldsymbol{AB} + \boldsymbol{AC} , \\ (\boldsymbol{B}+\boldsymbol{C})\boldsymbol{A} &= \boldsymbol{BA} + \boldsymbol{CA} . \end{aligned}$$

The top and bottom are not equal in general.

Identity and inverse

The identity matrix $\boldsymbol{I}$ is a square diagonal matrix (i.e. it has zeros off the diagonal) with ones on the diagonal,

$$\boldsymbol{I} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} .$$

The identity plays the role of 1 in scalar arithmetic,

$$\boldsymbol{AI} = \boldsymbol{IA} = \boldsymbol{A}$$

where $\mathbf{A}$ is any square matrix. If a square matrix $\boldsymbol{A}$ has an inverse then it is said to be invertible or *non-singular*,

$$\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I} ,$$

and if $\boldsymbol{A}$ and $\boldsymbol{B}$ have inverses then,

$$(\boldsymbol{AB})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1} .$$

A matrix is singular (non-invertible) if and only if it has a zero determinant (see below).

Transpose and dot-product

The transpose of a matrix is a matrix where the rows and columns have been switched. We put a capital T at the top right of a matrix to show that we are taking the transpose, e.g.

$$\begin{pmatrix} 1 & 3 & 2 \\ 4 & 5 & 1 \end{pmatrix}^{\mathrm{T}} = \begin{pmatrix} 1 & 4 \\ 3 & 5 \\ 2 & 1 \end{pmatrix} .$$

An important identity involves the transpose of a matrix product,

$$(\boldsymbol{AB})^{T} = \boldsymbol{B}^{T}\boldsymbol{A}^{T} .$$

The transpose of a column vector is a row vector and from the previous section we see that it is possible to multiply together column and row vectors of the same length to get a single number (a scalar), e.g.

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \quad \boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} \qquad \boldsymbol{x}^{\mathrm{T}}\boldsymbol{y} = x_1y_1 + x_2y_2 + x_3y_3 \ .$$

This particular form of multiplication is quite common and is given the name *dot-product* or *scalar product*. In general the dot-product of two length d column vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ with elements x_i and y_i is written,

$$\boldsymbol{x} \cdot \boldsymbol{y} \ = \ \boldsymbol{x}^{\mathrm{T}}\boldsymbol{y} \ = \ \sum_{i=1}^{d} x_i y_i \ .$$

The dot-product plays an important role in numerous aspects of machine learning.

Geometrical interpretation of vectors

It is sometimes helpful to think of vectors in geometrical terms. We can think of a vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)^T$ as representing a point in a d-dimensional space which is connected to the origin of some axes. The values of each element in the vector are the displacements along each of the corresponding axes. In figure 1 we show an example of two 2-dimensional vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ with angle θ between them. One can show that the dot-product is given in geometrical terms by,

$$\boldsymbol{x} \cdot \boldsymbol{y} = |\boldsymbol{x}||\boldsymbol{y}| \cos\theta$$

where $|\boldsymbol{x}|$ is the the length (also known as the vector norm or L^2-norm) of vector $\boldsymbol{x}$, which is defined,

$$|\boldsymbol{x}| \ = \ \sqrt{\boldsymbol{x} \cdot \boldsymbol{x}} \ = \ \sqrt{\sum_{i=1}^{d} x_i^2} \ .$$

This is often written using two bars, e.g. $||\boldsymbol{x}||$, to emphasize that it is the L^2-norm.

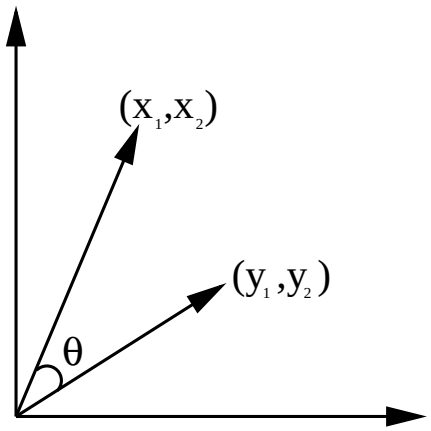

Figure 1: Vectors can be represented geometrically as arrows from the origin to a point with coordinates given by the vector elements. Here we show two 2-dimensional vectors $\boldsymbol{x} = (x_1, x_2)$ and $\boldsymbol{y} = (y_1, y_2)$. The *dot-product* of the two vectors is $\boldsymbol{x} \cdot \boldsymbol{y} = |\boldsymbol{x}||\boldsymbol{y}| \cos\theta$ where the length of vector $\boldsymbol{x}$ is defined $|\boldsymbol{x}| = \sqrt{\boldsymbol{x} \cdot \boldsymbol{x}}$.

Eigenvalues and eigenvectors

An important construction in linear algebra is the eigen-decomposition of a square matrix $\boldsymbol{A}$. In the equation,

$$\boldsymbol{A}\boldsymbol{u} = \lambda \boldsymbol{u}$$

we say that $\boldsymbol{u}$ is an eigenvector of the matrix $\boldsymbol{A}$ with an associated eigenvalue (or singular value) λ. We will limit our discussion to symmetric matrices with real-valued entries, in which case the eigenvalues are also real valued.

Notice that the eigenvector can be multiplied by any constant scalar and the above equation remains true. Therefore it is common to normalise the eigenvector so that it has unit length, e.g. $\boldsymbol{u}^T\boldsymbol{u} = 1$. One can find d different (linearly independent) eigenvectors for a d-dimensional matrix,

$$\boldsymbol{A}\boldsymbol{u}_i = \lambda_i \boldsymbol{u}_i \quad \text{for} \quad i = 1, 2, \ldots, d.$$

The number of non-zero eigenvalues is known as the rank of a matrix and a full-rank d-dimensional square matrix has d non-zero eigenvalues. It is useful to choose a set of orthogonal eigenvectors (a basis set),

$$\boldsymbol{u}_i \cdot \boldsymbol{u}_j = \delta_{ij} \quad \text{where} \quad \delta_{ij} = \begin{cases} 1 \text{ if } i = j \ , \\ 0 \text{ if } i \neq j \ . \end{cases}$$

Here, the function δ_{ij} is known as the Kronecker delta function or indicator function. If all the eigenvalues are distinct ($\lambda_i \neq \lambda_j \forall\, i \neq j$) then the eigenvectors are uniquely defined and are guaranteed to be orthogonal.

In matrix notation we can write the set of eigenvalue equations as,

$$\boldsymbol{A}\boldsymbol{U} = \boldsymbol{U}\boldsymbol{\Lambda} \tag{1}$$

where $\boldsymbol{U}$ is a matrix who's columns are the eigenvectors and the corresponding eigenvalues are in the diagonal matrix $\boldsymbol{\Lambda}$,

$$\boldsymbol{\Lambda} = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_d \end{pmatrix} .$$

The orthogonality of the eigenvectors can also be expressed in matrix form,

$$\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{U}\boldsymbol{U}^T = \boldsymbol{I}$$

and we say that $\boldsymbol{U}$ is an orthogonal matrix. For an orthogonal matrix $\boldsymbol{U}^T = \boldsymbol{U}^{-1}$ which should be obvious from above. By multiplying both sides of equation (1) by $\boldsymbol{U}^T$ on the right one obtains

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$$

which is known as the Singular Value Decomposition (SVD) of the matrix $\boldsymbol{A}$. This is the decomposition used in Principal Component Analysis (PCA) which we will meet during the course.

The covariance matrix $\boldsymbol{C}$ defined in equation (9) is an example of a symmetric matrix, because $C_{ij} = C_{ji}$. A full-rank covariance matrix is positive definite which means that it has all positive eigenvalues.

Determinant

The determinant of a square matrix can written in terms of the eigenvalues,

$$\det(\boldsymbol{A}) = |\boldsymbol{A}| = \prod_{i=1}^{d} \lambda_i \ .$$

This shows that a singular real symmetric matrix (i.e. with real-valued eigenvalues) must have at least one zero eigenvalue.

1.4 Differential calculus

Differentiation

The most fundamental quantity in differential calculus is the *derivative* of a function. The derivative of a function (of a single variable) at a point is the slope or *gradient* of the function at that point, i.e. how much y changes for a small constant increase in x. In fact it is the ratio of the change in y over the increase in x. Points which are in regions of the plot which slope up to the right are points with positive derivative (gradient). Points which are in regions of the plot which slope down to the right are points with negative derivative (gradient). Points which are in flat regions or lie between sloping regions have zero derivative (gradient).

In figure 2 we have plotted $y = f(x)$ with $f(x) = 2x - x^3$, a cubic polynomial. The gradient of the curve at $x = 1/\sqrt{3}$ is one, i.e. there is a slope of 45^o at this point. There are some equivalent notations for derivative,

$$f'(x) \quad \text{or} \quad \frac{d}{dx}f(x) \quad \text{or} \quad \frac{dy}{dx} \quad (\text{when } y = f(x)) \ .$$

The notation on the right reminds us that the gradient is the ratio of a small change in y, dy to a small change in x, dx. The notation on the left reminds us that the derivative is different at different parts of the function, so the derivative itself is a function of x.

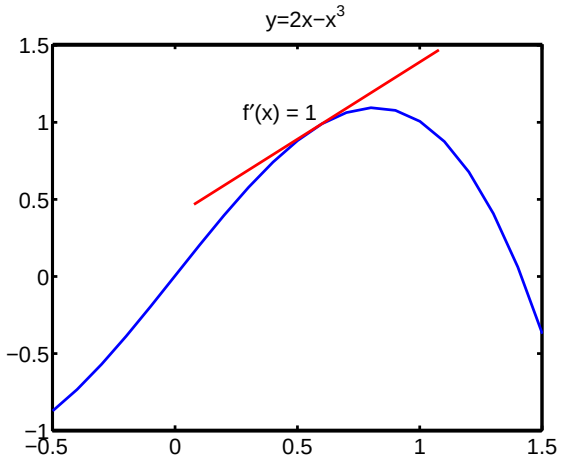

Figure 2: We plot a cubic polynomial. The gradient at the point $x = 1/\sqrt{3}$ is one ($f'(x) = 1$), corresponding to a 45^o slope to the right.

One important use for derivatives is in determining the *stationary points* of a function. Examples of stationary points are the maxima and minima of the function. At these points the derivative vanishes. In figure 3 we show a maximum and minimum of $f(x) = 2x - x^3$ which are at $x = \pm\sqrt{2/3}$. These are actually *local* or *relative* minima and maxima because the function does take higher and lower values for other values of x. This should be contrasted with the minimum of the function $y = x^2$ at $x = 0$ which is a *global* minimum, i.e. at $x = 0$ the function reaches its lowest point over the entire range of x. We can find stationary points by setting the derivative to zero and solving the corresponding equation for x.

We can calculate the derivative of any function using a small number of rules. For example,

$$\begin{aligned} \frac{d}{dx}\left(af(x) + b\,g(x)\right) &= a\,\frac{d}{dx}f(x) + b\,\frac{d}{dx}g(x) \ , \\ \frac{d}{dx}\left(x^n\right) &= nx^{n-1} \ , \end{aligned}$$

$$\begin{aligned}\frac{d}{dx}\exp(ax) &= a\exp(ax)\ ,\\ \frac{d}{dx}\ln(x) &= \frac{1}{x}\ ,\end{aligned}$$

where a and b are numbers or possibly other functions which are independent of x. We can use these rules to calculate the derivative of many functions. For example, consider the polynomial plotted in figures 2 and 3, $f(x) = 2x - x^3$,

$$\frac{d}{dx}\left(2x - x^3\right) = 2\frac{d}{dx}x - \frac{d}{dx}x^3 = 2 - 3x^2$$

where we have used the top two rules above. We can now work out why the points in the figures have the derivatives shown. Setting the derivative to one we get,

$$\begin{aligned}f'(x) &= 2 - 3x^2 = 1\\ \rightarrow \quad x &= \pm\tfrac{1}{\sqrt{3}}\ .\end{aligned}$$

The positive solution $x = 1/\sqrt{3}$ is the point highlighted in figure 2. A similar equation can be used to determine positions for the maximum and minimum shown in figure 3. We just set the derivative to zero in this case,

$$\begin{aligned}f'(x) &= 2 - 3x^2 = 0\\ \rightarrow \quad x &= \pm\sqrt{\tfrac{2}{3}}\ .\end{aligned}$$

These are the points in figure 3. We can use the second derivative at these points to determine whether they are a minimum or a maximum,

$$f''(x) = \frac{d^2}{dx^2}\left(2x - x^3\right) = -6x$$

For $x = \sqrt{2/3}$ the second derivative is negative, indicating that this is a maximum. For $x = -\sqrt{2/3}$ we have a positive second derivative indicating that this is a minimum. These conditions should be familiar, but later we will generalise them to conditions to functions of more then one variable.

Often we will want to take the derivative of a function containing another function. In this case it is useful to use the primed notation $f'(x)$ for the derivative,

$$\frac{d}{dx}g(f(x)) = f'(x)g'(f(x))\ .$$

For example,

$$\frac{d}{dx}\exp(f(x)) = f'(x)\exp(f(x)) \quad \text{and} \quad \frac{d}{dx}\ln(f(x)) = \frac{f'(x)}{f(x)}\ .$$

Multivariate functions

We will often deal with functions of more than one variable – multivariate functions. In this case we will often use a *partial derivative* which is a derivative with respect to one variable assuming the others are constant, e.g.

$$\begin{aligned}\frac{\partial}{\partial x}(x^2 - yx\mathrm{e}^x) &= 2x - y(\mathrm{e}^x + x\mathrm{e}^x)\\ \frac{\partial}{\partial y}(x^2 - yx\mathrm{e}^x) &= -x\mathrm{e}^x\\ \frac{\partial^2}{\partial x\partial y}(x^2 - yx\mathrm{e}^x) &= -(\mathrm{e}^x + x\mathrm{e}^x)\end{aligned}$$

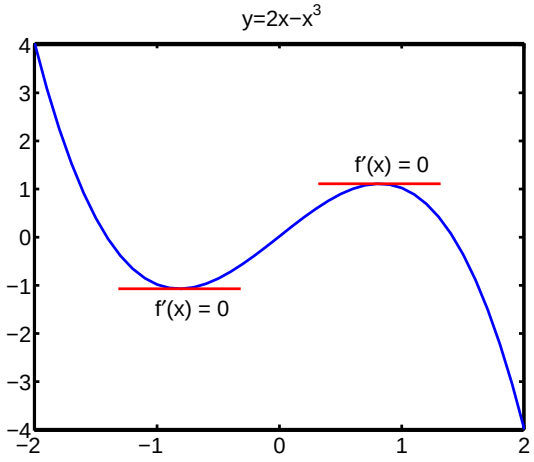

Figure 3: We plot a cubic polynomial which has a local maximum at $x = \sqrt{2/3}$ and a local minimum at $x = -\sqrt{2/3}$. The derivative is zero at these stationary points and the second derivative can be used to determine whether they are a relative maximum or a relative minimum.

The last equation is obtained by taking the derivative of the top equation with respect to y or the middle equation with respect to x.

It is sometimes useful to think of a function of more than one variable as a function of a vector of variables $\boldsymbol{x}$ which we can write $f(\boldsymbol{x})$. The argument is a vector while the output is a scalar. Recall the geometrical interpretation of a vector in figure 1. In this two-dimensional example, one can think of the function $f(\boldsymbol{x})$ as a surface in a three-dimensional plot with the height of the third axis giving the value of the function. If this surface is smooth then it is useful to generalise the ideas from differential calculus to the multivariate setting.

The generalisation of a derivative in this context is the gradient function (a vector function) which is a vector of partial derivatives,

$$\nabla_{\boldsymbol{x}} f(\boldsymbol{x}) = \begin{pmatrix} \frac{\partial f(\boldsymbol{x})}{x_1} \\ \frac{\partial f(\boldsymbol{x})}{x_2} \\ \vdots \\ \frac{\partial f(\boldsymbol{x})}{x_d} \end{pmatrix} .$$

The stationary points of a multivariate function are those for which the gradient is equal to a vector of zeros,

$$\nabla_{\boldsymbol{x}} f(\boldsymbol{x}) = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} .$$

This could correspond to a maximum, minimum or a *saddle point*. Numerical optimisation methods exist to find such points – examples are gradient descent, conjugate gradient descent or quasi-newton methods; we will use these methods during the neural networks lab.

We can find out what type of stationary point we are dealing with by examining the matrix of second

derivatives, which is known as the *Hessian matrix*,

$$\boldsymbol{H} = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_d} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_d} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_d \partial x_1} & \frac{\partial^2 f}{\partial x_d \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_d^2} \end{pmatrix} .$$

The Hessian matrix plays an analogous role to the second derivative in the univariate case. The Hessian is a real symmetric matrix and therefore it has real-valued eigenvalues. The eigenvalues tell us what kind of stationary point we are at, generalising the univariate case described previously. For a relative maximum of $f(\boldsymbol{x})$ all the eigenvalues will be negative at the stationary point and the Hessian matrix is negative definite. For a relative minimum all the eigenvalues will be positive and the Hessian is positive definite. For a saddle point there will be both positive and negative eigenvalues. If any of the eigenvalues are zero then there are directions in which the function is flat (has zero curvature) and we may have to consider higher order derivatives to decide on the type of stationary point.

Lagrange multipliers

In some cases a function is constrained in some way, and we would like to find the maximum or minimum of the function $f(\boldsymbol{x})$ given a constraint $g(\boldsymbol{x}) = 0$. One can solve this problem by finding the stationary points of a function called a *Lagrangian*,

$$\mathcal{L}(\boldsymbol{x}, \lambda) = f(\boldsymbol{x}) - \lambda g(\boldsymbol{x})$$

where λ is known as a Lagrange multiplier[1]. We then solve,

$$\nabla_{\boldsymbol{x}} \mathcal{L} = 0 \ , \quad \frac{\partial \mathcal{L}}{\partial \lambda} = 0 \ .$$

The second equation just gives the constraint equation. The first equation becomes,

$$\nabla_{\boldsymbol{x}} f(\boldsymbol{x}) - \lambda \nabla_{\boldsymbol{x}} g(\boldsymbol{x}) \ = \ \mathbf{0} \ .$$

This equation must be solved for $\boldsymbol{x}$.

Let's consider a simple example to demonstrate how a Lagrange multiplier could be used. A packaging firm want to produce a box using the minimum amount of material (surface area) for a given volume. This type of problem is important in manufacturing, where we want to maximise utility while minimising cost. Let the sides of the box be x_1, x_2 and x_3. The volume V is fixed which provides us with the constraint $V = x_1 x_2 x_3$. We can write this constraint by defining the function,

$$g(\boldsymbol{x}) = x_1 x_2 x_3 - V \ ,$$

and writing the constraint as $g(\boldsymbol{x}) = 0$. We want to minimise the surface area which is the sum of the areas of each side,

$$f(\boldsymbol{x}) = 2(x_1 x_2 + x_1 x_3 + x_2 x_3) \ .$$

The condition for the minimum surface area is then given by,

$$\nabla_{\boldsymbol{x}} f(\boldsymbol{x}) - \lambda \nabla_{\boldsymbol{x}} g(\boldsymbol{x}) = \begin{pmatrix} \frac{\partial f}{x_1} - \lambda \frac{\partial g}{x_1} \\ \frac{\partial f}{x_2} - \lambda \frac{\partial g}{x_2} \\ \frac{\partial f}{x_3} - \lambda \frac{\partial g}{x_3} \end{pmatrix} = \begin{pmatrix} 2(x_2 + x_3) - \lambda x_2 x_3 \\ 2(x_1 + x_3) - \lambda x_1 x_3 \\ 2(x_1 + x_2) - \lambda x_1 x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \ .$$

[1] not to be confused with an eigenvalue – we typically use λ for both

The solution to this system of equations is $x_1 = x_2 = x_3 = 4/\lambda$ (you should check this). We therefore get the intuitively obvious result that the shape of box that minimises surface area for a given volume is a cube. We can also find the value of λ from the constraint equation $g(\boldsymbol{x}) = 0$ but that is not necessary for us to determine the shape. It is often the case that the actual value of the Lagrange multipliers is not important.

Notice that there was an alternative way to cast the box problem. We could equally have maximised the volume for a fixed surface area, and that would have given us the same conclusion. There is often more than one way to solve an optimisation problem.

This idea can easily be extended to any number of linear constraints by adding a new Lagrange multiplier λ_i for each constraint $g_i(\boldsymbol{x}) = 0$. Lagrange multipliers will be an important tool for explaining Principal Component Analysis (PCA). Lagrange multipliers can also be used for inequality constraints and in this form they play an important role in the Support Vector Machine (SVM). We will also use them in maximum likelihood applied to simple sequence models.

Integration

Integration is the inverse of differentiation. If $f(x)$ is a function of x with derivative $f'(x)$ then,

$$f(x) = \int f'(x)\, dx \ .$$

This is known as an *indefinite integral*. A *definite integral* is one where the limits of integration are specified. If we have a graph of a function $f(x)$ then we can integrate in the range a to b in order to calculate the area under the function in this range for a one-dimensional (univariate) function,

$$\text{Area under } f(x) \text{ between } a \text{ and } b \ = \ \int_a^b f(x) \ .$$

If the integral is over the whole range of integration then we often don't bother specifying the limits. For a multivariate function defined over the whole space we will often write,

$$\int f(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} = \int_{-\infty}^{\infty} \mathrm{d}x_1 \int_{-\infty}^{\infty} \mathrm{d}x_2 \cdots \int_{-\infty}^{\infty} \mathrm{d}x_d f(\boldsymbol{x}) \ .$$

where the range of the integrals is determined by the existence of the function $f(\boldsymbol{x})$ depending on the context. Working out integrals in a high-dimensional space is often very difficult and we usually resort to numerical or other approximate methods.

2 Probability theory

Much of this course will be concerned with making inferences from uncertain, or noisy, data sets. In this case the most principled way to proceed (arguably the only principled way to proceed) is by using the calculus of probability. Below we outline some of the main elements of probability theory, without too much detail or rigor. There are many textbooks available on statistics and probability theory which give a much more thorough introduction.

2.1 Laws of probability

Let A and B denote two events which can occur. Define the probability of seeing event A to be $P(A)$ and the probability of seeing both A and B by $P(A, B)$. Further let $P(A|B)$ denote the probability of seeing A once B is known to have occurred. The axioms of probability theory are then,

$$\begin{aligned} P(\text{not } A) + P(A) &= 1 \ , \\ P(A, B) &= P(A|B)P(B) \ . \end{aligned} \tag{2}$$

The first of these states that the probability of all possibilities will add to one (either A occurs or it doesn't). The second axiom shows how probabilities of different, non-exclusive events should be combined. Note that the second expression could equally well be written,

$$P(B, A) = P(B|A)P(A) \ ,$$

where we have simply exchanged the events A and B. Since $P(A, B) = P(B, A)$ (both are the probability of seeing A and B) we can equate the right hand sides of these equivalent expressions to get,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \ . \tag{3}$$

This is Bayes' theorem which plays a fundamental role in probabilistic inference problems. Because probabilities must add to one we know that,

$$P(A|B) + P(\text{not } A|B) = 1 \ ,$$

which gives us an expression for the denominator in Bayes' theorem,

$$P(B) = P(B|A)P(A) + P(B|\text{not } A)P(\text{not } A) \ .$$

2.2 Discrete random variables

Above we considered the probability of two events. In general we may be interested in a *random variable* X which maps a number to each possible outcome of an experiment involving many events. There may be a finite number of possible outcomes in which case X takes values from some finite set $\{x_1, x_2, \ldots, x_n\}$. In this case the *probability mass function* is defined $p(x) \equiv P(X = x)$ and gives the probability of each possible outcome. A simple example is the Bernoulli distribution (where $f \in [0, 1]$),

$$p(x) = \begin{cases} f & \text{if } x = 1 \\ 1 - f & \text{if } x = 0 \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

For example, if $f = 1/2$ this is the appropriate distribution for describing the probability of the outcome of a fair coin toss.

The equivalent of the first axiom in equation (2) can now be written in terms of random variables as,

$$\sum_{i=1}^{n} p(x_i) \ = \ 1 \ .$$

This is called the *normalisation* condition, which ensures that probabilities add up to one. The mean (or expectation value) and variance of a discrete random variable are defined by,

$$\begin{aligned} \mu &= \sum_{i=1}^{n} p(x_i) x_i \ , \\ \sigma^2 &= \sum_{i=1}^{n} p(x_i)(x_i - \mu)^2 \ . \end{aligned}$$

Calculate the mean and variance for the Bernoulli distribution defined in equation (4).

A very useful probability distribution for discrete random variables is the binomial distribution, which is the probability of getting $X = 1$ r times out of n when sampling from the Bernoulli distribution defined in equation (4). In this case the probability mass function for r is,

$$p(r) = \frac{n!}{(n-r)!\, r!} f^r (1-f)^{n-r} \ . \tag{5}$$

The mean and variance of the binomial distribution are,

$$\begin{aligned} \mu &= nf \ , \\ \sigma^2 &= nf(1-f) \ . \end{aligned}$$

2.3 Continuous random variables

In some cases the random variable X maps onto a continuum and the probability of any particular value $P(X = x)$ will be zero for most probability distributions. In this case it makes more sense to define a probability density function rather than a probability mass function. A probability density function $p(x)$ is defined on a vanishingly small interval between x and $x + \mathrm{d}x$,

$$p(x) = \lim_{\mathrm{d}x \to 0} \frac{P(x < X < x + \mathrm{d}x)}{\mathrm{d}x} \ .$$

The probability density is similar to the probability mass function except that now we replace weighted sums by integrals (we use the same notation for both, since it should be clear which we mean by the context). For example, the normalisation condition above becomes,

$$\int_{-\infty}^{\infty} p(x) \mathrm{d}x \ = \ 1$$

and the definition of the mean and variance are now,

$$\begin{aligned} \mu &= \int_{-\infty}^{\infty} p(x) x \, \mathrm{d}x \ , \\ \sigma^2 &= \int_{-\infty}^{\infty} p(x)(x-\mu)^2 \, \mathrm{d}x \ . \end{aligned}$$

It should be noted that the integral will extend over the range of integration (not necessarily over the whole real line as in the above examples). A very useful probability density function is the normal distribution, which is defined in equation (6) (the multivariate generalisation is given in equation (7)).

Central limit theorem and the normal distribution

A frequency histogram will sometimes be well approximated by some standard distribution. A very useful example is the normal distribution (also known as a Gaussian distribution) which has the following density function,

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2} \right) \ , \tag{6}$$

where μ and σ^2 are the mean and variance respectively. The central limit theorem states that there are a class of random processes for which the associated frequency histograms will approach a normal distribution as the number of terms increases. Roughly speaking, these are problem in which terms contributing to the frequencies are randomly distributed according to some distribution with finite variance.

2.4 Multivariate data

In the previous examples each data instance is a scalar quantity, i.e. the number of sixes in a sequence of throws. In this case the data are said to be one-dimensional, since a scalar can be considered a one-dimensional vector. In contrast to this, all of the data we will deal with during the course will have dimension greater than one. In this case we will have to generalize some of the concepts introduced previously to higher dimensions. Data instances become data vectors and vectors are collected together to form the columns or rows of a matrix (see section 1.3). A particularly important multivariate probability distribution is the multivariate normal (or Gaussian) distribution. The Gaussian density function for column data vector $\boldsymbol{x}$ is given by,

$$p(\boldsymbol{x}) = \frac{\mathrm{e}^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{C}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}}{\sqrt{2\pi\det(\boldsymbol{C})}} \tag{7}$$

where the distribution has mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{C}$ defined by,

$$\boldsymbol{\mu} = \int p(\boldsymbol{x})\,\boldsymbol{x}\,\mathrm{d}\boldsymbol{x}\ , \tag{8}$$

$$\boldsymbol{C} = \int p(\boldsymbol{x})\,(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\,\mathrm{d}\boldsymbol{x}\ , \tag{9}$$

where the integral over the vector $\boldsymbol{x}$ is interpreted as,

$$\int f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \int_{-\infty}^{\infty}\mathrm{d}x_1\int_{-\infty}^{\infty}\mathrm{d}x_2\cdots\int_{-\infty}^{\infty}\mathrm{d}x_N f(\boldsymbol{x})\ .$$

From the above expressions we see that $\boldsymbol{C}$ should be symmetric and invertible (non-singular). We will sometimes use the shorthand notation $\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{C})$.

For formulas involving vectors and matrices we should think of N-dimensional column vectors (like $\boldsymbol{x}$) as $N\times 1$ matrices. Recall that if matrix $\boldsymbol{A}$ is $N\times K$ while matrix $\boldsymbol{B}$ is $K\times M$ then $\boldsymbol{AB}$ is $N\times M$. The transpose $\boldsymbol{A}^{\mathrm{T}}$ has columns of $\boldsymbol{A}$ as rows and is $K\times N$. The same rules hold for the N-dimensional column vectors $\boldsymbol{x}$ and $\boldsymbol{y}$. For example,

$$\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y} = \sum_{i=1}^{N} x_i y_i = \boldsymbol{x}\cdot\boldsymbol{y} \quad \text{the dot or inner product which is } 1\times 1 \text{ (scalar)}\ , \tag{10}$$

$$\boldsymbol{x}\boldsymbol{y}^{\mathrm{T}} = \boldsymbol{Z} \quad \text{with} \quad Z_{ij} = x_i y_j \quad \text{the outer product which is } N\times N \text{ (matrix)}\ . \tag{11}$$

Similarly, the term in the exponent of equation (7) is a scalar,

$$(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{C}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = \sum_{i=1}^{N}\sum_{j=1}^{N}(x_i-\mu_i)C_{ij}^{-1}(x_j-\mu_j)\ ,$$

while the components of the covariance matrix defined in equation (9) are written,

$$C_{ij} = \int p(\boldsymbol{x})\,(x_i-\mu_i)(x_j-\mu_j)\,\mathrm{d}\boldsymbol{x}\ .$$

The covariance matrix gives the correlation of the various dimensions in the data. If two components x_i and x_j are positively correlated then $C_{ij}>0$. If they are anti-correlated then $C_{ij}<0$.

2.5 Information Theory

Consider the probability distribution $P(X)$ for a discrete random variable $X \in \{x_1, x_2, \ldots, x_n\}$. Let $p_i = P(X = x_i)$ be the probability mass vector. Shannon's entropy $H(P)$ is a function of the probability distribution,

$$H(P) = -\sum_{i=1}^{n} p_i \log p_i \ .$$

The units used to measure entropy depend on the base used for the logarithm. If we are using base 2 then entropy is measured in bits. If we are using natural logarithms then entropy is measured in nats. The entropy measures the uncertainty in a random experiment with n mutually exclusive outcomes having probabilities $p_1, p_2, \ldots, p_n$. Alternatively, it is the information gained when the outcome of such an experiment is observed.

The entropy is a positive quantity, because $-\log p_i \geq 0 \, \forall \, i$ since $p_i \leq 1$. It is maximised when all the probabilities are equal. It is zero when one outcome has probability one and the rest have zero probability, because in that case the outcome of the experiment is already certain before carrying it out.

The relative entropy, or Kullback-Leibler (KL) divergence, between two probability distributions P and Q is defined by,

$$D(P, Q) = \sum_{i=1}^{n} p_i \log \left(\frac{p_i}{q_i} \right) \ .$$

This can be considered to be a sort of "distance" between two probability distributions. $D(P, Q) > 0$ unless $P = Q$. However, the measure is not symmetric, $D(P, Q) \neq D(Q, P)$, and therefore it cannot be considered a true 'distance' — the word distance is a mathematical term reserved for measures having certain properties — look it up if you're interested. Hence, we call this the KL-divergence.

The mutual information $I(X, Y)$ is a measure of the dependence between two random variables $X \in \{x_1, x_2, \ldots, x_n\}$ and $Y \in \{y_1, y_2, \ldots, y_m\}$ (we consider discrete variables – the definitions carry over to the continuous case in a straightforward way). The variables have joint probability mass function $p(X, Y)$. We define the marginal probability distributions to be,

$$p(X) = \sum_{i=1}^{m} p(X, y_i) \ , \quad p(Y) = \sum_{i=1}^{n} p(x_i, Y) \ .$$

If X and Y are independent then $p(X, Y) = p(X)p(Y)$. The mutual information is defined,

$$I(X, Y) = D(p(X, Y), p(X)p(Y))$$

where $D(P, Q)$ is the KL-divergence between two distributions. If the variables X and Y are statistically independent then their mutual information is zero, otherwise it is positive.