Learning from Data 1
Mathematical Preliminaries*

Division of Informatics, University of Edinburgh

Dr Chris Williams

October 1999

1 Mathematical Preliminaries

1.1 Vectors

The course assumes that you are familiar with the basics of vectors and vector calculations. Let $\mathbf{x}$ denote the n-dimensional vector with components

$$(x_1, x_2, \cdots, x_n)$$

Then $|\mathbf{x}|$ denotes the length of this vector, using the usual Euclidian definition:

$$|\mathbf{x}| = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$$

The inner product $\mathbf{w}.\mathbf{x}$ is defined as:

$$\mathbf{w}.\mathbf{x} = \sum_{i=1}^{n} w_i x_i$$

and has a natural geometric interpretation as:

$$\mathbf{w}.\mathbf{x} = |\mathbf{w}|\,|\mathbf{x}|\cos(\theta)$$

where θ is the angle between the two vectors. Thus if the lengths of two vectors are fixed their inner product is largest when $\theta = 0$, whereupon one is just some constant multiple of the other.

Exercise: This 'natural geometric interpretation' is not black magic. Show that

$$\sum_{i=1}^{n} w_i x_i = |\mathbf{w}|\,|\mathbf{x}|\cos(\theta)$$

by using the Theorem of Pythagoras. Start with a triangle whose corners are the origin and the points $\mathbf{w}$ and $\mathbf{x}$. This triangle is not necessarily right-angled, of course; the angle at the origin is of size θ.

[0] The material in this handout derives from material written by Peter Ross for lecture notes in "Connectionist Computing".

1.2 Matrices

The course assumes some familiarity with matrices, which are shown as upper-case bold letters such as $\mathbf{A}$. If the element of the i-th row and j-th column is a_{ij}, then $\mathbf{A}^T$ denotes the matrix that has a_{ji} there instead - the *transpose* of $\mathbf{A}$. So, for example if $\mathbf{A}$ is a 3×3 matrix:

$$\mathbf{A} = \begin{pmatrix} 2 & 3 & 4 \\ 4 & 5 & 9 \\ 6 & 7 & 1 \end{pmatrix}$$

then the transpose (written $\mathbf{A}^T$) is:

$$\mathbf{A}^T = \begin{pmatrix} 2 & 4 & 6 \\ 3 & 5 & 7 \\ 4 & 9 & 1 \end{pmatrix}$$

The product of two matrices $\mathbf{A}$ and $\mathbf{B}$ has $\sum_k a_{ik}b_{kj}$ in the i-th row and j-th column.

The matrix $\mathbf{I}$ is the identity or unit matrix, necessarily square, with 1s on the diagonal and 0s everywhere else. If $\det(\mathbf{A})$ denotes the determinant of a square matrix $\mathbf{A}$ then the equation

$$\det(\mathbf{A} - \lambda\mathbf{I}) = 0$$

is called the *characteristic polynomial* of $\mathbf{A}$. Using the example above, the characteristic polynomial would be:

$$\begin{vmatrix} 2-\lambda & 3 & 4 \\ 4 & 5-\lambda & 9 \\ 6 & 7 & 1-\lambda \end{vmatrix} = 0$$

which is

$$(2-\lambda)((5-\lambda)(1-\lambda)-63) - 3(4(1-\lambda)-54) + 4(28-6(5-\lambda)) = 0$$

which simplifies to:

$$-\lambda^3 + 8\lambda^2 + 82\lambda + 26 = 0$$

Note that a square matrix must satisfy its own characteristic polynomial, by definition of the polynomial, so (pre- or post-multiplying through by $\mathbf{A}^{-1}$) it provides a way to calculate the inverse of a matrix using only matrix multiplication, if that inverse exists. Clearly the inverse exists if and only if the matrix is square and $\det(\mathbf{A}) \neq 0$ (note that $\det(\mathbf{A})$ is the constant term in the characteristic polynomial).

The roots of the characteristic polynomial are called the *eigenvalues* of the matrix. Note that if $\mathbf{A}$ is an $m \times n$ matrix and $\mathbf{x}$ is an n-dimensional (column) vector, then

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

represents a linear map into an m-dimensional space. If $\mathbf{A}$ happens to be a square matrix then any vector which is transformed by the linear map into a scalar multiple of itself is called an *eigenvector* of that matrix. Obviously, in that case $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ for some λ. The

eigenvectors can be found by finding the eigenvalues and then solving the linear equation set:

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = 0$$

An *orthogonal matrix* is a square matrix $\mathbf{A}$ such that $\mathbf{A}^T = \mathbf{A}^{-1}$. Such matrices represent a mapping from one rectangular co-ordinate system to another. For such a matrix,

$$\mathbf{A}\mathbf{A}^T = \mathbf{I}$$

- the inner product of any two different rows is 0 and the inner product of any row with itself is 1.

1.3 Basic combinatorics

The number of ways of selecting k items from a collection of n items is

$$\left(\begin{array}{c} n \\ k \end{array} \right) = \frac{n!}{k!(n-k)!}$$

if the ordering of the selection doesn't matter. This quantity is also the coefficient of x^k in the expansion of $(1+x)^n$. Stirling's formula provides a useful approximation for dealing with large factorials:

$$n! \approx n^n e^{-n}\sqrt{2\pi n}$$

There are a huge number of formulae involving combinations. For example, since $(1+x)^{n+1} = (1+x)^n(1+x)$ it is clear that

$$\left(\begin{array}{c} n \\ k \end{array} \right) + \left(\begin{array}{c} n \\ k+1 \end{array} \right) = \left(\begin{array}{c} n+1 \\ k+1 \end{array} \right)$$

and so on.

1.4 Basic probability and distributions

A random variable X is a variable which, in different experiments carried out under the same conditions, assumes different values x_i, each of which then represents a random event. A discrete random variable can take one of only a finite, or perhaps a countably infinite, set of values. A continuous random variable can take any value in a finite or infinite interval. Random variables are completely characterised by their probability density and distribution functions.

For a discrete random variable, if $p(X = x)$ is the probability that it takes the value x then

$$F(x) = p(X < x)$$

is the distribution function of X. For a continuous random variable, there is a probability density function $f(x)$ such that

$$\int_{-\infty}^{\infty} f(x)\, dx = 1$$

and the distribution function is then:

$$F(x) = \int_{-\infty}^{x} f(t)\,dt$$

For a discrete random variable, the mean value μ is

$$\mu = \sum x_i p(X = x_i)$$

and for a continuous variable it is

$$\mu = \int_{-\infty}^{\infty} t f(t)\,dt$$

The variance σ^2 is, for a discrete variable:

$$\sigma^2 = \sum (x_i - \mu)^2 p(X = x_i)$$

and for a continuous variable:

$$\sigma^2 = \int_{-\infty}^{\infty} (t - \mu)^2 f(t)\,dt$$

There are several widely-occurring distributions that are worth knowing about. Suppose that some event will happen with fixed probability p. Then the probability that it will happen exactly k times in n trials is

$$\left(\begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k}$$

and this is the binomial distribution. It has mean np and variance $np(1-p)$. If one lets $n \to \infty$ one gets the Gaussian or normal distribution, typically parameterised by two constants a and b; it has density function

$$\frac{1}{a\sqrt{2\pi}} e^{-(x-b)^2/(2a^2)}$$

with mean b and variance a^2. If one starts with the binomial distribution and lets $n \to \infty$ and $p \to 0$ with the extra assumption that $np = a$, where a is some constant, then one gets the Poisson distribution with density function

$$\frac{a^k e^{-a}}{k!}$$

with mean and variance both a.

1.5 Partial differentiation

If $z = f(x_1, x_2, \cdots, x_n)$ is a function of n independent variables then one can form the partial derivative of the function with respect to one variable (say x_i),

$$\frac{\partial f}{\partial x_i}$$

by treating all other variables as constant. For example, if

$$f = xy + y^3$$

then

$$\frac{\partial f}{\partial x} = y \qquad \frac{\partial f}{\partial y} = x + 3y^2$$

The geometric significance of a quantity such as $\frac{\partial f}{\partial x}$ is as follows. If the function f is plotted and represents some suitably well-behaved surface, then this partial derivative represents the slope of the surface in a direction parallel to the x-axis at any given point (x, y). The total derivative dz is given by

$$dz = \sum \frac{\partial z}{\partial x_i} dx_i$$

and clearly, if all the x_i are functions of one variable t then

$$\frac{dz}{dt} = \sum \frac{\partial z}{\partial x_i} \frac{dx_i}{dt}$$

There is a directly analogous version of this 'chain rule' for the case where the x_i are each functions of several variables and you wish to find the partial derivative of z with respect to one of those variables.

Exercise: Find the partial derivatives of the function

$$f(x, y, z) = (x + 2y)^2 \sin(xy)$$

1.6 Optimization: Lagrange multipliers[1]

Suppose that you wish to find the stationary points (maxima or minima) of some n-argument function $f(\mathbf{x}) = f(x_1, \cdots, x_n)$, subject to the m constraints $g_1(\mathbf{x}) = 0, \cdots, g_m(\mathbf{x}) = 0$. Lagrange showed that they could be found as the solution of the $(n+m)$ equations in the $(n+m)$ variables $x_1, \cdots, x_n, \lambda_1, \cdots, \lambda_m$:

$$\begin{aligned}
\frac{\partial f}{\partial x_1} - \sum_{j=1}^{m} \lambda_j \frac{\partial g_j}{\partial x_1} &= 0 \\
&\cdots \\
\frac{\partial f}{\partial x_n} - \sum_{j=1}^{m} \lambda_j \frac{\partial g_j}{\partial x_n} &= 0 \\
g_1(\mathbf{x}) &= 0 \\
&\cdots \\
g_m(\mathbf{x}) &= 0
\end{aligned}$$

where the λ_j are m specially-introduced variables called Lagrange multipliers. This theorem is certainly not obvious, but should at least be fairly natural-looking, and it provides

[1] The material in this section is not strictly necessary for the LfD1 course.

a handy way to tackle a range of optimization problems. Notice that the above equations are the $(n+m)$ partial derivatives of the function

$$f - \sum_{j=1}^{m} \lambda_j g_j$$

each set to zero.

For example, to find the maximum of $f(x,y) = x+y$ subject to the constraint $x^2+y^2 = 1$, solve:

$$\begin{aligned} 1 - 2\lambda x &= 0 \\ 1 - 2\lambda y &= 0 \\ x^2 + y^2 - 1 &= 0 \end{aligned}$$

to get $x = y = \lambda = \pm 1/\sqrt{2}$, after which you should then check to determine which of these two solutions is the true maximum.

Exercise: Find the maximum of $y - x$ subject to the constraint that $y + x^2 = 4$.

You can find the answer to the same problem experimentally as follows. Plot the graph of $y = 4 - x^2$ and the graph of $y = x + m$, and find the largest value of m such that the two graphs still intersect.