Background material crib-sheet

Iain Murray <i.murray+ta@gatsby.ucl.ac.uk>, October 2003

Here are a summary of results with which you should be familiar. If anything here is unclear you should do some further reading and exercises.

1 Probability Theory

Chapter 2, sections 2.1–2.3 of David MacKay’s book covers this material:

The probability a discrete variable $A$ takes value $a$ is: $0 \leq P(A=a) \leq 1$  

Probabilities of alternatives add: $P(A=a \text{ or } a') = P(A=a) + P(A=a')$  

Variables can be “summed out” of joint distributions:  

\[ P(A=a) = \sum_{\text{all possible } b} P(A=a, B=b) \]

$P(A=a|B=b)$ is the probability $A=a$ occurs given the knowledge $B=b$.  

$P(A=a, B=b) = P(A=a) P(B=b|A=a) = P(B=b) P(A=a|B=b)$  

The following hold, for all $a$ and $b$, if and only if $A$ and $B$ are independent:  

\[ P(A=a|B=b) = P(A=a) \]
\[ P(B=b|A=a) = P(B=b) \]
\[ P(A=a, B=b) = P(A=a) P(B=b). \]

Otherwise the product rule above must be used.

Bayes rule can be derived from the above:  

\[ P(A=a|B=b, \mathcal{H}) = \frac{P(B=b|A=a, \mathcal{H}) P(A=a|\mathcal{H})}{P(B=b|\mathcal{H})} \propto P(A=a, B=b|\mathcal{H}) \]

Note that here, as with any expression, we are free to condition the whole thing on any set of assumptions, $\mathcal{H}$, we like. Note $\sum_a P(A=a, B=b|\mathcal{H}) = P(B=b|\mathcal{H})$ gives the normalising constant of proportionality.
All the above theory basically still applies to continuous variables if sums are converted into integrals\(^1\). The probability that \(X\) lies between \(x\) and \(x + dx\) is \(p(x)\,dx\), where \(p(x)\) is a probability density function with range \([0, \infty)\).

\[
P(x_1 < X < x_2) = \int_{x_1}^{x_2} p(x)\,dx, \quad \int_{-\infty}^{\infty} p(x)\,dx = 1 \quad \text{and} \quad p(x) = \int_{-\infty}^{x} p(x, y)\,dy.
\]

The expectation or mean under a probability distribution is:

\[
\langle f(a) \rangle = \sum_a P(A = a) f(a) \quad \text{or} \quad \langle f(x) \rangle = \int_{-\infty}^{\infty} p(x) f(x)\,dx
\]

## 2 Linear Algebra

*This is designed as a prequel to Sam Roweis’s “matrix identities” sheet:  
http://www.cs.toronto.edu/~roweis/notes/matrixid.pdf*

Scalars are individual numbers, vectors are columns of numbers, matrices are rectangular grids of numbers, eg:

\[
x = 3.4, \quad x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \quad A = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m1} & A_{m2} & \cdots & A_{mn} \end{pmatrix}
\]

In the above example \(x\) is \(1 \times 1\), \(x\) is \(n \times 1\) and \(A\) is \(m \times n\).

The transpose operator, \(^T\) (‘ in Matlab), swaps the rows and columns:

\[
x^T = x, \quad x^T = \begin{pmatrix} x_1 & x_2 & \cdots & x_n \end{pmatrix}, \quad (A^T)_{ij} = A_{ji}
\]

Quantities whose inner dimensions match may be “multiplied” by summing over this index. The outer dimensions give the dimensions of the answer.

\[
A\!x \text{ has elements } (A\!x)_i = \sum_{j=1}^{n} A_{ij}x_j \quad \text{and} \quad (AA^T)_{ij} = \sum_{k=1}^{n} A_{ik}(A^T)_{kj} = \sum_{k=1}^{n} A_{ik}A_{jk}
\]

All the following are allowed (the dimensions of the answer are also shown):

<table>
<thead>
<tr>
<th>(x^T!x)</th>
<th>(xx^T)</th>
<th>(A!x)</th>
<th>(AA^T)</th>
<th>(A^TA)</th>
<th>(x^T!A)</th>
</tr>
</thead>
<tbody>
<tr>
<td>scalar</td>
<td>matrix</td>
<td>vector</td>
<td>matrix</td>
<td>matrix</td>
<td>scalar</td>
</tr>
</tbody>
</table>

while \(xx\), \(AA\) and \(xA\) do not make sense for \(m \neq n \neq 1\). Can you see why?

An exception to the above rule is that we may write: \(xA\). Every element of the matrix \(A\) is multiplied by the scalar \(x\).

Simple and valid manipulations:

\[
(AB)C = A(BC) \quad A(B+C) = AB+AC \quad (A+B)^T = A^T+B^T \quad (AB)^T = B^T\!A^T
\]

Note that \(AB \neq BA\) in general.

---

\(^1\)Integrals are the equivalent of sums for continuous variables. Eg: \(\sum_{i=1}^{n} f(x_i)\Delta x\) becomes the integral \(\int_{a}^{b} f(x)\,dx\) in the limit \(\Delta x \to 0, n \to \infty\), where \(\Delta x = \frac{b-a}{n}\) and \(x_i = a + i\Delta x\). Find an A-level text book with some diagrams if you have not seen this before.
2.1 Square Matrices

Now consider the square \( n \times n \) matrix \( B \).

All off-diagonal elements of diagonal matrices are zero. The “Identity matrix”, which leaves vectors and matrices unchanged on multiplication, is diagonal with each non-zero element equal to one.

\[
B_{ij} = 0 \text{ if } i \neq j \quad \text{“} B \text{ is diagonal”}
\]
\[
I_{ij} = 0 \text{ if } i \neq j \text{ and } I_{ii} = 1 \quad \forall i \quad \text{“} I \text{ is the identity matrix”}
\]
\[
Ix = x \quad I = B = B^T \quad x^T I = x^T
\]

Some square matrices have inverses:

\[
B^{-1} B = BB^{-1} = I \quad (B^{-1})^{-1} = B,
\]

which have these properties:

\[
(BC)^{-1} = C^{-1} B^{-1} \quad (B^{-1})^T = (B^T)^{-1}
\]

Linear simultaneous equations could be solved (inefficiently) this way:

if \( Bx = y \) then \( x = B^{-1} y \)

Some other commonly used matrix definitions include:

\[
B_{ij} = B_{ji} \quad \text{“} B \text{ is symmetric”}
\]
\[
\text{Trace}(B) = \text{Tr}(B) = \sum_{i=1}^{n} B_{ii} \quad \text{“sum of diagonal elements”}
\]

Cyclic permutations are allowed inside trace. Trace of a scalar is a scalar:

\[
\text{Tr}(BCD) = \text{Tr}(DBC) = \text{Tr}(CDB) \quad x^T B x = \text{Tr}(x^T B x) = \text{Tr}(xx^T B)
\]

The determinant\(^2\) is written \( \text{Det}(B) \) or \( |B| \). It is a scalar regardless of \( n \).

\[
|BC| = |B||C|, \quad |x| = x, \quad |xB| = x^n|B|, \quad |B^{-1}| = \frac{1}{|B|}.
\]

It determines if \( B \) can be inverted: \( |B| = 0 \Rightarrow B^{-1} \text{ undefined} \). If the vector to every point of a shape is pre-multiplied by \( B \) then the shape’s area or volume increases by a factor of \( |B| \). It also appears in the normalising constant of a Gaussian. For a diagonal matrix the volume scaling factor is simply the product of the diagonal elements. In general the determinant is the product of the eigenvalues.

\[
B e^{(i)} = \lambda^{(i)} e^{(i)} \quad \text{“} \lambda^{(i)} \text{ is an eigenvalue of } B \text{ with eigenvector } e^{(i)} \text{”}
\]

\[
|B| = \prod \text{eigenvalues} \quad \text{Trace}(B) = \sum \text{eigenvalues}
\]

If \( B \) is real and symmetric (eg a covariance matrix) the eigenvectors are orthogonal (perpendicular) and so form a basis (can be used as axes).

\(^2\)This section is only intended to give you a flavour so you understand other references and Sam’s crib sheet. More detailed history and overview is here: http://www.wikipedia.org/wiki/Determinant
3 Differentiation

Any good A-level maths text book should cover this material and have plenty of exercises. Undergraduate text books might cover it quickly in less than a chapter.

The gradient of a straight line \( y = mx + c \) is a constant \( y' = \lim_{\Delta x \to 0} \frac{y(x + \Delta x) - y(x)}{\Delta x} = m \).

Many functions look like straight lines over a small enough range. The gradient, the derivative, is not constant, but a new function:

\[
\frac{dy}{dx} = \lim_{\Delta x \to 0} \frac{y(x + \Delta x) - y(x)}{\Delta x},
\]

which could be differentiated again:

\[
\frac{d^2y}{dx^2} = \frac{dy'}{dx}.
\]

The following results are well known (\( c \) is a constant):

<table>
<thead>
<tr>
<th>Standard derivatives</th>
</tr>
</thead>
<tbody>
<tr>
<td>( f(x) ) : ( c ) ( cx ) ( cx^n )</td>
</tr>
<tr>
<td>( f'(x) ) : ( 0 ) ( c ) ( cnx^{n-1} )</td>
</tr>
</tbody>
</table>

At a maximum or minimum the function is rising on one side and falling on the other. In between the gradient must be zero. Therefore \( \maxima \) and \( \minima \) satisfy:

\[
\frac{df(x)}{dx} = 0 \quad \text{or} \quad \frac{df(x)}{dx} = 0 \quad \forall i
\]

If we can’t solve this we can evolve our variable \( x \), or variables \( x \), on a computer using gradient information until we find a place where the gradient is zero.

A function may be approximated by a straight line about any point \( a \).

\[
f(a + x) \approx f(a) + xf'(a),
\]

eg: \( \log(1 + x) \approx \log(1) + x \frac{1}{1 + 0} = x \)

The derivative operator is linear:

\[
\frac{d}{dx}(f(x) + g(x)) = \frac{df(x)}{dx} + \frac{dg(x)}{dx},
\]

eg: \( \frac{d}{dx}(x + \exp(x)) = 1 + \exp(x) \).

Dealing with products is slightly more involved:

\[
\frac{d}{dx}(u(x)v(x)) = v \frac{du}{dx} + u \frac{dv}{dx},
\]

eg: \( \frac{d}{dx}(x \cdot \exp(x)) = \exp(x) + x \exp(x) \).

For example:

\[
\frac{d}{dy}(ay^m) = \frac{d(ay^m)}{du} \cdot \frac{du}{dy} = amy^{m-1} \cdot \exp(ay^m)
\]

If you can’t show the following you could do with some practice:

\[
\frac{d}{dz} \left[ \frac{1}{b + cz} \exp(az) + e \right] = \exp(az) \left( \frac{a}{b + cz} - \frac{c}{(b + cz)^2} \right)
\]

Note that \( a, b, c \) and \( e \) are constants, that \( \frac{1}{a} = u^{-1} \) and this is hard if you haven’t done differentiation (for a long time). Again, get a text book.

---

\(^3\)More accurate approximations can be made. Look up Taylor series.