1 Probability Theory


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

Probabilities of alternative outcomes add: \( \Pr(A \in \{a, a'\}) = \Pr(A = a) + \Pr(A = a') \)

The probabilities of all outcomes must sum to one: \( \sum_{a} \Pr(A = a) = 1 \)

\( \Pr(A = a, B = b) \) is the joint probability that both \( A = a \) and \( B = b \) occur.

Variables can be “summed out” of joint distributions:

\[
\Pr(A = a) = \sum_{b} \Pr(A = a, B = b)
\]

\( \Pr(A = a | B = b) \) is the probability \( A = a \) occurs given the knowledge \( B = b \).

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

Bayes rule can be derived from the above:

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

Marginalizing over all possible \( a \) gives the evidence or normalizing constant:

\[
\sum_{a} \Pr(A = a, B = b | \mathcal{H}) = \Pr(B = b | \mathcal{H})
\]

The following hold, for all \( a \) and \( b \), **if and only if** \( A \) and \( B \) are independent:

\[
\begin{align*}
\Pr(A = a | B = b) &= \Pr(A = a) \\
\Pr(B = b | A = a) &= \Pr(B = b) \\
\Pr(A = a, B = b) &= \Pr(A = a) \Pr(B = b).
\end{align*}
\]

All the above theory basically still applies to continuous variables if the sums are converted into integrals\(^4\). The probability that \( X \) lies between \( x \) and \( x + dx \) is \( p(x) \, dx \), where \( p(x) \) is a *probability density function* with range \([0, \infty]\).

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

The expectation or mean under a probability distribution is:

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

\(^4\)Integrals are the equivalent of sums for continuous variables, e.g. \( \sum_{i=1}^{n} f(x_i) \Delta x \) 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 \( x_i = a + i \Delta x \).
2 Linear Algebra

This complements Sam Roweis’s “Matrix Identities”: 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, \( ^\top \) (’ in Matlab), swaps the rows and columns:

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

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

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

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

<table>
<thead>
<tr>
<th>( x^\top x )</th>
<th>( xx^\top )</th>
<th>( Ax )</th>
<th>( AA^\top )</th>
<th>( A^\top A )</th>
<th>( x^\top Ax )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( 1 \times 1 )</td>
<td>( n \times n )</td>
<td>( m \times 1 )</td>
<td>( m \times m )</td>
<td>( n \times n )</td>
<td>( 1 \times 1 )</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:

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

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

2.1 Square Matrices

A square matrix has equal number of rows and columns, e.g. \( B \) with dimensions \( n \times n \).

A diagonal matrix is a square matrix with all off-diagonal elements being zero: \( B_{ij} = 0 \) if \( i \neq j \).

The identity matrix is a diagonal matrix with all diagonal elements equal to one.

\[ \text{“I is the identity matrix” } \iff I_{ij} = 0 \text{ if } i \neq j \text{ and } I_{ii} = 1 \text{ } \forall i \]

The identity matrix leaves vectors and matrices unchanged upon multiplication.

\[ Ix = x \quad IB = B = BI \quad x^\top I = x^\top \]
Some square matrices have inverses:

$$B^{-1}B = BB^{-1} = \mathbb{I} \quad (B^{-1})^{-1} = B,$$

which have the additional properties:

$$ (BC)^{-1} = C^{-1}B^{-1} \quad (B^{-1})^\top = (B^\top)^{-1} $$

Linear simultaneous equations could be solved this way:

if $Bx = y$ then $x = B^{-1}y$

Some other commonly used matrix definitions include:

- **Symmetry**: $B_{ij} = B_{ji} \iff \text{"B is symmetric"}$

- **Trace**: $\text{Trace}(B) = \text{Tr}(B) = \sum_{i=1}^{n} B_{ii} = \text{"sum of diagonal elements"}$

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

$$\text{Tr}(BCD) = \text{Tr}(DBC) = \text{Tr}(CDB) \quad x^\top Bx = \text{Tr}(x^\top Bx) = \text{Tr}(xx^\top B)$$

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: $|B| = 0 \Rightarrow B^{-1}$ 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 normalizing 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.

$$B\mathbf{v}^{(i)} = \lambda^{(i)}\mathbf{v}^{(i)} \iff \lambda^{(i)} \text{ is an eigenvalue of } B \text{ with eigenvector } \mathbf{v}^{(i)}$$

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 flavor so you understand other references and Sam’s crib sheet. More detailed history and overview is here: [http://www.wikipedia.org/wiki/Determinant](http://www.wikipedia.org/wiki/Determinant)
3 Differentiation

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

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

\[
y'(x) = \lim_{\Delta x \to 0} \frac{y(x+\Delta x) - y(x)}{\Delta x}, \quad \text{which could be differentiated again: } \quad y'' = \frac{d^2y}{dx^2} = \frac{dy'}{dx}
\]

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

\[
\begin{align*}
f(x) & : & c & cx & cx^n & \log_e(x) & e^x \\
f'(x) & : & 0 & c & cnx^{n-1} & 1/x & e^x \\
\end{align*}
\]

At a maximum or minimum the function is rising on one side and falling on the other. In between the gradient must reach zero somewhere. Therefore

\[
\text{maxima and minima satisfy: } \frac{df(x)}{dx} = 0 \quad \text{or} \quad \frac{df(x)}{dx} = 0 \iff \frac{df(x)}{dx_i} = 0 \ \forall i
\]

If we can’t solve this analytically, 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. But there’s no guarantee that we would find all maxima and minima.

A function may be approximated by a straight line\(^3\) about any point \( a \).

\[
f(a+x) \approx f(a) + xf'(a), \quad \text{e.g.: } \log(1+x) \approx \log(1+0) + x \frac{1}{1+0} = x
\]

The derivative operator is linear:

\[
\frac{d(f(x) + g(x))}{dx} = \frac{df(x)}{dx} + \frac{dg(x)}{dx}, \quad \text{e.g.: } \frac{d(x + \exp(x))}{dx} = 1 + \exp(x).
\]

Dealing with products is slightly more involved:

\[
\frac{d(u(x)u(x))}{dx} = \frac{du}{dx}u + u \frac{dv}{dx}, \quad \text{e.g.: } \frac{d(x \cdot \exp(x))}{dx} = \exp(x) + x \exp(x).
\]

The “chain rule” \( \frac{df(u)}{dx} = \frac{du}{dx} \frac{df(u)}{du} \), allows results to be combined.

\[
\text{For example: } \frac{d\exp(ay^n)}{dy} = \frac{d(ay^n)}{dy} \cdot \frac{d\exp(ay^n)}{d(ay^n)} \quad \text{“with } u = ay^n
\]

\[
= amy^{n-1} \cdot \exp(ay^n)
\]

Convince yourself of the following:

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

Note that \( a, b, c \) and \( e \) are constants and \( 1/u = u^{-1} \). This might be hard if you haven’t done differentiation (for a long time). You may want to grab a calculus textbook for a review.

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