Probability theory

1 Definitions and terminology

The rudiments of probability theory used in the development of decision theory are now presented. Firstly, we introduce the idea of an experiment or observation. The set of all possible outcomes is called the sample space, $\Omega$, of the model, with each possible outcome being a sample point. An event, $A$, is a set of experimental outcomes, and corresponds to a subset of points in $\Omega$.

A probability measure is a function, $P(A)$, with a set, $A$, as argument. It can be regarded as the expected proportion of times that $A$ actually occurs and has the following properties

1. $0 \leq P(A) \leq 1$
2. $P(\Omega) = 1$
3. If $A$ and $B$ are mutually exclusive events (disjoint sets) then
   $$P(A \cup B) = P(A) + P(B)$$

More generally, when $A$ and $B$ are not necessarily exclusive,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

A random variable is a function that associates a number with each possible outcome $\omega \in \Omega$. We denote random variables by upper-case letters here, but in the main body of work we generally use the same symbol for a random variable and a measurement, the meaning being clear from context. Although the argument of $P$ is a set, it is usual to use a description of the event as the argument, regarding this as equivalent to the corresponding set. Thus, we write $P(X > Y)$ for the probability that $X$ is greater than $Y$, rather than $P(\{\omega : X(\omega) > Y(\omega)\})$.

The cumulative distribution function, sometimes simply called the distribution function, of a random variable is the probability that the random variable is less than or equal to some specified value $x$; that is,

$$P_X(x) = \text{probability that } X \leq x$$

(1)

Usually, when there is no ambiguity, we drop the subscript $X$. The cumulative distribution function is a monotonic function of its argument with the property that $P(-\infty) = 0$; $P(\infty) = 1$. The derivative of the distribution function,

$$p(x) = \frac{dP}{dx}$$
is the probability density function of the random variable $X$. For sample spaces such as the entire real line, the probability density function, $p(x)$, has the following properties

\[
\int_{-\infty}^{\infty} p(x)dx = 1
\]
\[
\int_{a}^{b} p(x)dx = \text{probability that } X \text{ lies between } a \text{ and } b
\]
\[
= P(a \leq X \leq b)
\]
\[
p(x) \geq 0
\]

Much of the discussion in this book will relate to vector quantities, since the inputs to many pattern classification systems may be expressed in vector form. Random vectors are defined similarly to random variables, associating each point in the sample space, $\Omega$, to a point in $\mathbb{R}^p$

\[X : \Omega \to \mathbb{R}^p\]

The joint distribution of $X$ is the $p$-dimensional generalisation of (1) above

\[P_X(x) = P_X(x_1, \ldots, x_p) = \text{probability that } X_1 \leq x_1, \ldots, X_p \leq x_p\]

and the joint density function is similarly given by

\[p(x) = \frac{\partial^p P(x)}{\partial x_1 \ldots \partial x_p}\]

Given the joint density of a set of random variables $X_1, \ldots, X_p$, then a smaller set $X_1, \ldots, X_m$ ($m < p$) also possesses a probability density function determined by

\[p(x_1, \ldots, x_m) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} p(x_1, \ldots, x_p)dx_{m+1} \ldots dx_p\]

This is sometimes known as the marginal density of $X_1, \ldots, X_m$, although the expression is more usually applied to the single-variable densities, $p(x_1), p(x_2), \ldots, p(x_p)$ given by

\[p(x_i) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} p(x_1, \ldots, x_p)dx_1 \ldots dx_{i-1}dx_{i+1} \ldots dx_p\]

### 1.1 Expectation

The expected vector, or mean vector, of a random variable $x$, is defined by

\[m = E[X] = \int x p(x)dx\]
where \( dx \) denotes \( dx_1 \ldots dx_p \) and the integral is over the entire space (and unless otherwise stated \( \int = \int_{-\infty}^{\infty} \)) and \( E[. \] denotes the expectation operator. The \( i \)th component of the mean can be calculated by

\[
m_i = E[X_i] = \int \cdots \int x_i p(x_1, \ldots, x_p) dx_1 \ldots dx_p = \int x_i p(x) dx = \int_{-\infty}^{\infty} x_i p(x_i) dx_i
\]

where \( p(x_i) \) is the marginal density of the single variable \( X_i \) given above.

### 1.2 Covariance

The **covariance** of two random variables provides a measure of the extent to which the deviations of the random variables from their respective mean values tend to vary together. The covariance of random variables \( X_i \) and \( X_j \), denoted by \( C_{ij} \), is given by

\[
C_{ij} = E[(X_j - E[X_j])(X_i - E[X_i])]
\]

which may be expressed as

\[
C_{ij} = E[X_i X_j] - E[X_i]E[X_j]
\]

where \( E(X_i X_j) \) is the **autocorrelation**. The matrix with \( (i,j) \)th component \( C_{ij} \) is the covariance matrix, \( C \)

\[
C = E[(X - m)(X - m)^T]
\]

Two random variables \( X_i \) and \( X_j \) are **uncorrelated** if the covariance between the two variables is zero; that is, \( C_{ij} = 0 \), which implies

\[
E[X_i X_j] = E[X_i]E[X_j]
\]

Two vectors, \( X \) and \( Y \), are uncorrelated if

\[
E[X^T Y] = E[X]^T E[Y]
\]

In the special case where the means of the vectors are zero, so that the relation above becomes \( E[X^T Y] = 0 \), then the random variables are said to be **mutually orthogonal**.

#### 1.2.1 Independence

Two events, \( A \) and \( B \), are **statistically independent** if

\[
P(A \cap B) = P(A)P(B)
\]

and two random variables \( X_i \) and \( X_j \) are independent if

\[
p(x_i, x_j) = p(x_i)p(x_j)
\]
If the random variables $X_1, X_2, \ldots, X_p$ are independent then the joint density function may be written as a product of the individual densities:

$$p(x_1, \ldots, x_p) = p(x_1) \cdots p(x_p)$$

If two variables are independent then the expectation of $(X_1X_2)$ is given by

$$E[X_1X_2] = \int \int x_1x_2p(x_1, x_2)dx_1dx_2$$

and using the independence property

$$E[X_1X_2] = \int x_1p(x_1)dx_1 \int x_2p(x_2)dx_2 = E[X_1]E[X_2]$$

This shows that $X_1$ and $X_2$ are uncorrelated. However, this does not imply that two variables that are uncorrelated are statistically independent.

### 1.3 Transformations

Often we shall want to consider a functional transformation from a given set of random variables \( \{X_1, X_2, \ldots, X_p\} \) represented by the vector $X$ to a set \( \{Y_1, Y_2, \ldots, Y_p\} \) represented by the vector $Y$. How do probability density functions change under such a transformation? Let the transformation be given by $Y = g(X)$, where $g = (g_1, g_2, \ldots, g_p)^T$. Then the density functions of $X$ and $Y$ are related by

$$p_Y(y) = \frac{p_X(x)}{|J|}$$

where $|J|$ is the absolute value of the Jacobian determinant

$$J(x_1, \ldots, x_p) = \begin{vmatrix} \frac{\partial g_1}{\partial x_1} & \ldots & \frac{\partial g_1}{\partial x_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_p}{\partial x_1} & \ldots & \frac{\partial g_p}{\partial x_p} \end{vmatrix}$$

A simple transformation is the linear one

$$Y = AX + B$$

Then if $X$ has the probability density $p_X(x)$, the probability density of $Y$ is

$$p_Y(y) = \frac{p_X(A^{-1}(y - B))}{|A|}$$

where $|A|$ is the absolute value of the determinant of the matrix $A$. 

4
1.4 Conditional distributions

Given a random system and any two events $A$ and $B$ that can occur together, we can form a new system by taking only those trials in which $B$ occurs. The probability of $A$ in this new system is called the conditional probability of $A$ given $B$ and is denoted by $P(A|B)$ and if $P(B) > 0$ it is given by

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

or

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

This is the total probability theorem.

Now, since $P(A \cap B) = P(B \cap A)$, we have from (4)

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

or

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

This is Bayes’ theorem.

Now, if $A_1, A_2, \ldots, A_N$ are events which partition the sample space (that is, they are mutually exclusive and their union is $\Omega$) then

$$P(B) = \sum_{i=1}^{N} P(B \cap A_i)$$

$$= \sum_{i=1}^{N} P(B|A_i)P(A_i)$$

and we obtain a more practical form of Bayes’ theorem

$$P(A_j|B) = \frac{P(B|A_j)P(A_j)}{\sum_{i=1}^{N} P(B|A_i)P(A_i)}$$

(6)

In pattern classification problems, $B$ is often an observation event and the $A_j$ are pattern classes. The term a priori probability is often used for the quantity $P(A_i)$ and the objective is to find $P(A_i|B)$, which is termed the a posteriori probability of $A_i$.

The conditional distribution, $P_X(x|A)$, of a random variable $X$ given the event $A$ is defined as the conditional probability of the event $\{X \leq x\}$

$$P(x|A) = \frac{P(\{X \leq x\}, A)}{P(A)}$$

and $P(\infty|A) = 1$; $P(-\infty|A) = 0$. The conditional density $p(x|A)$ is the derivative of $P(x|A)$

$$p(x|A) = \frac{dP}{dx} = \lim_{\Delta x \to 0} \frac{P(x \leq X \leq x + \Delta x|A)}{\Delta x}$$
The extension of the result (5) to the continuous case gives

$$p(x) = \sum_{i=1}^{N} p(x|A_i)p(A_i)$$

where $p(x)$ is the mixture density and we have taken $B = \{X \leq x\}$, and the continuous version of Bayes’ theorem may be written

$$p(x|A) = \frac{p(A|x)p(x)}{p(A)} = \frac{p(A|X = x)p(x)}{\int_{-\infty}^{\infty} p(A|X = x)p(x)dx}$$

The conditional density of $x$ given that the random vector $Y$ has some specified value, $y$, is obtained by letting $A = \{y \leq Y \leq y + \Delta y\}$ and taking the limit

$$\lim_{\Delta y \to 0} p(x|\{y \leq Y \leq y + \Delta y\}) = \lim_{\Delta y \to 0} \frac{p(x, \{y \leq Y \leq y + \Delta y\})}{p(\{y \leq Y \leq y + \Delta y\})}$$

giving

$$p(x|y) = \frac{p(x,y)}{p(y)} \quad (7)$$

where $p(x,y)$ is the joint density of $X$ and $Y$ and $p(y)$ is the marginal density

$$p(y) = \int p(x,y)dx \quad (8)$$

Equations (7) and (8) lead to the density form of Bayes’ theorem

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)} = \frac{p(y|x)p(x)}{\int p(y|x)p(x)dx}$$

A generalisation of (7) is the conditional density of the random variables $X_{k+1} \ldots X_p$ given $X_1 \ldots X_k$ which is given by

$$p(x_{k+1}, \ldots, x_p|x_1, \ldots, x_k) = \frac{p(x_1 \ldots x_p)}{p(x_1 \ldots x_k)} \quad (9)$$

This leads to the chain rule

$$p(x_1 \ldots x_p) = p(x_p|x_1 \ldots x_{p-1})p(x_{p-1}|x_1 \ldots x_{p-2}) \ldots p(x_2|x_1)p(x_1)$$

The results of (8) and (9) allow unwanted variables in a conditional density to be removed. If they occur to the left of the vertical line, then integrate with respect to them. If they occur to the right, then multiply by the conditional density of the variables given the remaining variables on the right and integrate. For example

$$p(a|l, m, n) = \int p(a, b, c|l, m, n)dbdc$$

$$p(a, b, c|m) = \int p(a, b, c|l, m, n)p(l, n|m)dldn$$
2 Normal distribution

We shall now illustrate some of the definitions and results of this section using
the Gaussian or normal distribution (we use the two terms interchangeably in the
book). It is a distribution to which we often referred in our discussion of pattern
recognition algorithms in this book.

The standard normal density of a random variable \( X \) has zero mean and unit
variance and has the form

\[
p(x) = \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{x^2}{2} \right\} - \infty < x < \infty
\]

The distribution function is given by

\[
P(X) = \int_{-\infty}^{X} p(x) \, dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{X} \exp \left\{ -\frac{1}{2} x^2 \right\} \, dx = \frac{1}{2} + \frac{1}{2} \text{erf} \left( \frac{X}{\sqrt{2}} \right)
\]

where \( \text{erf}(x) \) is the error function \( \frac{2}{\sqrt{\pi}} \int_{0}^{x} \exp(-t^2) \, dt \).

For the function \( Y = \mu + \sigma X \) of the random variable \( X \), the density function
of \( Y \) is

\[
p(y) = \frac{1}{\sqrt{2\sigma^2\pi}} \exp \left\{ -\frac{1}{2} \left( \frac{y - \mu}{\sigma} \right)^2 \right\}
\]

which has mean \( \mu \) and variance \( \sigma^2 \), and we write \( Y \sim N(\mu, \sigma^2) \).

If \( X_1, X_2, \ldots, X_p \) are independently and identically distributed, each following
the standard normal distribution, then the joint density is given by

\[
p(x_1, x_2, \ldots, x_p) = \prod_{i=1}^{p} p(x_i) = \frac{1}{(2\pi)^{p/2}} \exp \left\{ -\frac{1}{2} \sum_{i=1}^{p} x_i^2 \right\}
\]

The transformation \( Y = AX + \mu \) leads to the density function for \( Y \) (using
Equation (3))

\[
p(y) = \frac{1}{(2\pi)^{p/2}|A|} \exp \left\{ -\frac{1}{2} (y - \mu)^T (A^{-1})^T (A^{-1}) (y - \mu) \right\}
\]

and since the covariance matrix, \( \Sigma \), of \( Y \) is

\[
\Sigma = \text{E}[(Y - \mu)(Y - \mu)^T] = AA^T
\]

Equation (10) is usually written

\[
p(y) = N(y|\mu, \Sigma) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2} (y - \mu)^T \Sigma^{-1} (y - \mu) \right\}
\]

This is the multivariate normal distribution.

Recall from the previous section that if two variables are independent then they
are uncorrelated, but that the converse is not necessarily true. However, a special
property of the normal distribution is that if two variables are joint normally
distributed and uncorrelated, then they are independent.

The marginal densities and conditional densities of a joint normal distribution
are all normal.
3 Probability distributions

We introduce some of the more commonly-used distributions (further probability distributions are listed by Bernardo and Smith, 1994). If $x$ has a specific probability density function, $R(x|\alpha)$, where $\alpha$ is the set of parameters of the specific functional form $R$, then for shorthand notation we may write $x \sim R(\alpha)$; similarly, we use $x|\beta \sim R(f(\beta))$, for some function $f$ to mean $p(x|\beta) = R(x|f(\beta))$.

$N(x|\mu, \Sigma)$, \hspace{1cm} Normal 1

\[ p(x) = \frac{1}{|\Sigma|^{1/2}(2\pi)^{p/2}} \exp\left\{ -\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu) \right\} \]

$\Sigma$, symmetric positive definite; $E[x] = \mu$; $V[x] = \Sigma$.

Sometimes it is convenient to express the normal with the inverse of the covariance matrix as a parameter.

$N_p(x|\mu, \lambda)$, \hspace{1cm} Normal 2

\[ p(x) = \frac{|\lambda|^{1/2}}{(2\pi)^{p/2}} \exp\left\{ -\frac{1}{2}(x - \mu)^T \lambda(x - \mu) \right\} \]

$\lambda$, symmetric positive definite; $E[x] = \mu$; $V[x] = \lambda^{-1}$.

$W_{ip}(x|\alpha, \beta)$, \hspace{1cm} Wishart

\[ p(x) = \left[ \pi^{p(p-1)/4} \prod_{i=1}^{p} \left( \frac{1}{2}(2\alpha + 1 - i) \right) \right]^{-1} |\beta|^\alpha |x|^{\alpha-(p+1)/2} \exp(-Tr(\beta x)) \]

$x$, symmetric positive definite; $\beta$, symmetric non-singular; $2\alpha > p - 1$; $E[x] = \alpha \beta^{-1}$; $E[x^{-1}] = (\alpha - (p + 1)/2)^{-1} \beta$.

$St_{ip}(x|\mu, \lambda, \alpha)$, \hspace{1cm} Multivariate Student

\[ p(x) = \frac{\Gamma\left(\frac{1}{2}(\alpha + p)\right)}{\Gamma\left(\frac{1}{2}(\alpha p)\right)|\lambda|^{\frac{1}{2}}} \left[ 1 + \frac{1}{\alpha}(x - \mu)^T \lambda(x - \mu) \right]^{-(\alpha+p)/2} \]

$\alpha > 0$, $\lambda$ symmetric positive definite, $E[x] = \mu$, $V[x] = \lambda^{-1}(\alpha - 2)^{-1} \alpha$. 

8
\[ p(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x) \]

\( \alpha > 0, \beta > 0, \) \( \mathbb{E}[x] = \alpha \beta^{-1}, \mathbb{V}[x] = \alpha \beta^{-2} \).

\[ p(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-(\alpha+1)} \exp(-\beta/x) \]

\( \alpha > 0, \beta > 0, \) \( \mathbb{E}[x] = \beta / (\alpha - 1) \).

If \( 1/y \sim \text{Ga}(\alpha, \beta) \), then \( y \sim \text{Ig}(\alpha, \beta) \).

\[ p(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) \Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1} \]

\( \alpha > 0, \beta > 0, \) \( 0 < x < 1, \) \( \mathbb{E}[x] = \alpha / (\alpha + \beta) \).

\[ p(x) = \left( \frac{\Gamma \left( \sum_{i=1}^{C} a_i \right)}{\prod_{i=1}^{C} a_i} \right) \prod_{i=1}^{C} x_i^{a_i-1} \]

\( 0 < x_i < 1, \) \( a_i > 0, \) \( a = (a_1, \ldots, a_C), \) \( x = (x_1, \ldots, x_C), \) \( \sum_{i=1}^{C} x_i = 1, \) \( \mathbb{E}[x_i] = a_i / \sum a_i \).

\[ p(x) = \binom{n}{x} \theta^x (1-\theta)^{n-x} \]

\( 0 < \theta < 1, n = 1, 2, \ldots, x = 0, 1, \ldots, n. \) \( \mathbb{E}[x] = n\theta, \mathbb{V}[x] = n\theta(1-\theta). \)
Multinomial

\[ p(x) = \frac{n!}{\prod_{i=1}^{k} x_i!} \prod_{i=1}^{k} \theta_i^{x_i} \]

\[ 0 < \theta_i < 1, \quad \sum_{i=1}^{k} \theta_i = 1, \quad \sum_{i=1}^{k} x_i = n, \quad x_i = 0, 1, \ldots, n. \quad E[x_i] = n\theta_i, \quad V[x_i] = n\theta_i(1 - \theta_i), \quad \text{Cov}[x_i, x_j] = -n\theta_i\theta_j \]

References