Thus we can view the mutual information as the reduction in the uncertainty about \( x \) by virtue of being told the value of \( y \) (or vice versa). From a Bayesian perspective, we can view \( p(x) \) as the prior distribution for \( x \) and \( p(x|y) \) as the posterior distribution after we have observed new data \( y \). The mutual information therefore represents the reduction in uncertainty about \( x \) as a consequence of the new observation \( y \).

**Exercises**

1.1 (*) Consider the sum-of-squares error function given by (1.2) in which the function \( y(x, w) \) is given by the polynomial (1.1). Show that the coefficients \( w = \{w_i\} \) that minimize this error function are given by the solution to the following set of linear equations

\[
\sum_{j=0}^{M} A_{ij} w_j = T_i
\]

(1.122)

where

\[
A_{ij} = \sum_{n=1}^{N} (x_n)^{i+j}, \quad T_i = \sum_{n=1}^{N} (x_n)^i t_n.
\]

(1.123)

Here a suffix \( i \) or \( j \) denotes the index of a component, whereas \( (x)^i \) denotes \( x \) raised to the power of \( i \).

1.2 (*) Write down the set of coupled linear equations, analogous to (1.122), satisfied by the coefficients \( w_i \) which minimize the regularized sum-of-squares error function given by (1.4).

1.3 (**) Suppose that we have three coloured boxes \( r \) (red), \( b \) (blue), and \( g \) (green). Box \( r \) contains 3 apples, 4 oranges, and 3 limes, box \( b \) contains 1 apple, 1 orange, and 0 limes, and box \( g \) contains 3 apples, 3 oranges, and 4 limes. If a box is chosen at random with probabilities \( p(r) = 0.2 \), \( p(b) = 0.2 \), \( p(g) = 0.6 \), and a piece of fruit is removed from the box (with equal probability of selecting any of the items in the box), then what is the probability of selecting an apple? If we observe that the selected fruit is in fact an orange, what is the probability that it came from the green box?

1.4 (**) Consider a probability density \( p_x(x) \) defined over a continuous variable \( x \), and suppose that we make a nonlinear change of variable using \( x = g(y) \), so that the density transforms according to (1.27). By differentiating (1.27), show that the location \( \hat{y} \) of the maximum of the density in \( y \) is not in general related to the location \( \hat{x} \) of the maximum of the density over \( x \) by the simple functional relation \( \hat{x} = g(\hat{y}) \) as a consequence of the Jacobian factor. This shows that the maximum of a probability density (in contrast to a simple function) is dependent on the choice of variable. Verify that, in the case of a linear transformation, the location of the maximum transforms in the same way as the variable itself.

1.5 (*) Using the definition (1.38) show that \( \text{var}[f(x)] \) satisfies (1.39).
1.6  (⋆) Show that if two variables $x$ and $y$ are independent, then their covariance is zero.

1.7  (★★) In this exercise, we prove the normalization condition (1.48) for the univariate Gaussian. To do this consider, the integral

$$I = \int_{-\infty}^{\infty} \exp \left( -\frac{1}{2\sigma^2} x^2 \right) \, dx$$  \hspace{1cm} (1.124)

which we can evaluate by first writing its square in the form

$$I^2 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp \left( -\frac{1}{2\sigma^2} x^2 - \frac{1}{2\sigma^2} y^2 \right) \, dx \, dy.$$  \hspace{1cm} (1.125)

Now make the transformation from Cartesian coordinates $(x, y)$ to polar coordinates $(r, \theta)$ and then substitute $u = r^2$. Show that, by performing the integrals over $\theta$ and $u$, and then taking the square root of both sides, we obtain

$$I = \left(2\pi\sigma^2\right)^{1/2}.$$  \hspace{1cm} (1.126)

Finally, use this result to show that the Gaussian distribution $N(x|\mu, \sigma^2)$ is normalized.

1.8  (★★) By using a change of variables, verify that the univariate Gaussian distribution given by (1.46) satisfies (1.49). Next, by differentiating both sides of the normalization condition

$$\int_{-\infty}^{\infty} N(x|\mu, \sigma^2) \, dx = 1$$  \hspace{1cm} (1.127)

with respect to $\sigma^2$, verify that the Gaussian satisfies (1.50). Finally, show that (1.51) holds.

1.9  (⋆) Show that the mode (i.e. the maximum) of the Gaussian distribution (1.46) is given by $\mu$. Similarly, show that the mode of the multivariate Gaussian (1.52) is given by $\mu$.

1.10  (⋆) Suppose that the two variables $x$ and $z$ are statistically independent. Show that the mean and variance of their sum satisfies

$$\mathbb{E}[x + z] = \mathbb{E}[x] + \mathbb{E}[z]$$  \hspace{1cm} (1.128)

$$\text{var}[x + z] = \text{var}[x] + \text{var}[z].$$  \hspace{1cm} (1.129)

1.11  (⋆) By setting the derivatives of the log likelihood function (1.54) with respect to $\mu$ and $\sigma^2$ equal to zero, verify the results (1.55) and (1.56).
1.12 (★★) Using the results (1.49) and (1.50), show that
\[ \mathbb{E}[x_n x_m] = \mu^2 + I_{nm} \sigma^2 \] (1.130)
where \( x_n \) and \( x_m \) denote data points sampled from a Gaussian distribution with mean \( \mu \) and variance \( \sigma^2 \), and \( I_{nm} \) satisfies \( I_{nm} = 1 \) if \( n = m \) and \( I_{nm} = 0 \) otherwise. Hence prove the results (1.57) and (1.58).

1.13 (*) Suppose that the variance of a Gaussian is estimated using the result (1.56) but with the maximum likelihood estimate \( \mu_{ML} \) replaced with the true value \( \mu \) of the mean. Show that this estimator has the property that its expectation is given by the true variance \( \sigma^2 \).

1.14 (★★) Show that an arbitrary square matrix with elements \( w_{ij} \) can be written in the form \( w_{ij} = w_{ij}^S + w_{ij}^A \) where \( w_{ij}^S \) and \( w_{ij}^A \) are symmetric and anti-symmetric matrices, respectively, satisfying \( w_{ij}^S = w_{ji}^S \) and \( w_{ij}^A = -w_{ji}^A \) for all \( i \) and \( j \). Now consider the second order term in a higher order polynomial in \( D \) dimensions, given by
\[ \sum_{i=1}^{D} \sum_{j=1}^{D} w_{ij} x_i x_j. \] (1.131)
Show that
\[ \sum_{i=1}^{D} \sum_{j=1}^{D} w_{ij} x_i x_j = \sum_{i=1}^{D} \sum_{j=1}^{D} w_{ij}^S x_i x_j \] (1.132)
so that the contribution from the anti-symmetric matrix vanishes. We therefore see that, without loss of generality, the matrix of coefficients \( w_{ij} \) can be chosen to be symmetric, and so not all of the \( D^2 \) elements of this matrix can be chosen independently. Show that the number of independent parameters in the matrix \( w_{ij}^S \) is given by \( D(D + 1)/2 \).

1.15 (★★★) In this exercise and the next, we explore how the number of independent parameters in a polynomial grows with the order \( M \) of the polynomial and with the dimensionality \( D \) of the input space. We start by writing down the \( M^{th} \) order term for a polynomial in \( D \) dimensions in the form
\[ \sum_{i_1=1}^{D} \sum_{i_2=1}^{D} \cdots \sum_{i_M=1}^{D} w_{i_1 i_2 \cdots i_M} x_{i_1} x_{i_2} \cdots x_{i_M}. \] (1.133)
The coefficients \( w_{i_1 i_2 \cdots i_M} \) comprise \( D^M \) elements, but the number of independent parameters is significantly fewer due to the many interchange symmetries of the factor \( x_{i_1} x_{i_2} \cdots x_{i_M} \). Begin by showing that the redundancy in the coefficients can be removed by rewriting this \( M^{th} \) order term in the form
\[ \sum_{i_1=1}^{D} \sum_{i_2=1}^{D} \cdots \sum_{i_M=1}^{D} \tilde{w}_{i_1 i_2 \cdots i_M} x_{i_1} x_{i_2} \cdots x_{i_M}. \] (1.134)
Note that the precise relationship between the $\tilde{w}$ coefficients and $w$ coefficients need not be made explicit. Use this result to show that the number of independent parameters $n(D, M)$, which appear at order $M$, satisfies the following recursion relation

$$n(D, M) = \sum_{i=1}^{D} n(i, M - 1). \quad (1.135)$$

Next use proof by induction to show that the following result holds

$$\sum_{i=1}^{D} \frac{(i + M - 2)!}{(i - 1)!(M - 1)!} = \frac{(D + M - 1)!}{(D - 1)! M!} \quad (1.136)$$

which can be done by first proving the result for $D = 1$ and arbitrary $M$ by making use of the result $0! = 1$, then assuming it is correct for dimension $D$ and verifying that it is correct for dimension $D + 1$. Finally, use the two previous results, together with proof by induction, to show

$$n(D, M) = \frac{(D + M - 1)!}{(D - 1)! M!}. \quad (1.137)$$

To do this, first show that the result is true for $M = 2$, and any value of $D \geq 1$, by comparison with the result of Exercise 1.14. Then make use of (1.135), together with (1.136), to show that, if the result holds at order $M - 1$, then it will also hold at order $M$.

1.16 (***) In Exercise 1.15, we proved the result (1.135) for the number of independent parameters in the $M$th order term of a $D$-dimensional polynomial. We now find an expression for the total number $N(D, M)$ of independent parameters in all of the terms up to and including the $M$th order. First show that $N(D, M)$ satisfies

$$N(D, M) = \sum_{m=0}^{M} n(D, m) \quad (1.138)$$

where $n(D, m)$ is the number of independent parameters in the term of order $m$. Now make use of the result (1.137), together with proof by induction, to show that

$$N(d, M) = \frac{(D + M)!}{D! M!}. \quad (1.139)$$

This can be done by first proving that the result holds for $M = 0$ and arbitrary $D \geq 1$, then assuming it holds at order $M$, and hence showing that it holds at order $M + 1$. Finally, make use of Stirling’s approximation in the form

$$n! \simeq n^n e^{-n} \quad (1.140)$$

for large $n$ to show that, for $D \gg M$, the quantity $N(D, M)$ grows like $D^M$, and for $M \gg D$ it grows like $M^D$. Consider a cubic ($M = 3$) polynomial in $D$ dimensions, and evaluate numerically the total number of independent parameters for (i) $D = 10$ and (ii) $D = 100$, which correspond to typical small-scale and medium-scale machine learning applications.
1.17 **WWW** The gamma function is defined by

\[ \Gamma(x) \equiv \int_0^\infty u^{x-1} e^{-u} \, du. \]  \hfill (1.141)

Using integration by parts, prove the relation \( \Gamma(x + 1) = x\Gamma(x) \). Show also that \( \Gamma(1) = 1 \) and hence that \( \Gamma(x + 1) = x! \) when \( x \) is an integer.

1.18 **WWW** We can use the result (1.126) to derive an expression for the surface area \( S_D \), and the volume \( V_D \), of a sphere of unit radius in \( D \) dimensions. To do this, consider the following result, which is obtained by transforming from Cartesian to polar coordinates

\[ \prod_{i=1}^D \int_{-\infty}^\infty e^{-x_i^2} \, dx_i = S_D \int_0^\infty e^{-r^2} r^{D-1} \, dr. \]  \hfill (1.142)

Using the definition (1.141) of the Gamma function, together with (1.126), evaluate both sides of this equation, and hence show that

\[ S_D = \frac{2\pi^{D/2}}{\Gamma(D/2)}. \]  \hfill (1.143)

Next, by integrating with respect to radius from 0 to 1, show that the volume of the unit sphere in \( D \) dimensions is given by

\[ V_D = \frac{S_D}{D}. \]  \hfill (1.144)

Finally, use the results \( \Gamma(1) = 1 \) and \( \Gamma(3/2) = \sqrt{\pi}/2 \) to show that (1.143) and (1.144) reduce to the usual expressions for \( D = 2 \) and \( D = 3 \).

1.19 **WWW** Consider a sphere of radius \( a \) in \( D \)-dimensions together with the concentric hypercube of side \( 2a \), so that the sphere touches the hypercube at the centres of each of its sides. By using the results of Exercise 1.18, show that the ratio of the volume of the sphere to the volume of the cube is given by

\[ \frac{\text{volume of sphere}}{\text{volume of cube}} = \frac{\pi^{D/2}}{D 2^{D-1} \Gamma(D/2)}. \]  \hfill (1.145)

Now make use of Stirling’s formula in the form

\[ \Gamma(x + 1) \simeq (2\pi)^{1/2} e^{-x} x^{x+1/2} \]  \hfill (1.146)

which is valid for \( x \gg 1 \), to show that, as \( D \to \infty \), the ratio (1.145) goes to zero. Show also that the ratio of the distance from the centre of the hypercube to one of the corners, divided by the perpendicular distance to one of the sides, is \( \sqrt{D} \), which therefore goes to \( \infty \) as \( D \to \infty \). From these results we see that, in a space of high dimensionality, most of the volume of a cube is concentrated in the large number of corners, which themselves become very long ‘spikes’!
1.20  (⋆⋆) In this exercise, we explore the behaviour of the Gaussian distribution in high-dimensional spaces. Consider a Gaussian distribution in $D$ dimensions given by

$$p(x) = \frac{1}{(2\pi\sigma^2)^{D/2}} \exp \left( -\frac{\|x\|^2}{2\sigma^2} \right). \quad (1.147)$$

We wish to find the density with respect to radius in polar coordinates in which the direction variables have been integrated out. To do this, show that the integral of the probability density over a thin shell of radius $r$ and thickness $\epsilon$, where $\epsilon \ll 1$, is given by $p(r)\epsilon$ where

$$p(r) = \frac{S_D r^{D-1}}{(2\pi\sigma^2)^{D/2}} \exp \left( -\frac{r^2}{2\sigma^2} \right) \quad (1.148)$$

where $S_D$ is the surface area of a unit sphere in $D$ dimensions. Show that the function $p(r)$ has a single stationary point located, for large $D$, at $\hat{r} \simeq \sqrt{D}\sigma$. By considering $p(\hat{r} + \epsilon)$ where $\epsilon \ll \hat{r}$, show that for large $D$,

$$p(\hat{r} + \epsilon) = p(\hat{r}) \exp \left( -\frac{3\epsilon^2}{2\sigma^2} \right) \quad (1.149)$$

which shows that $\hat{r}$ is a maximum of the radial probability density and also that $p(r)$ decays exponentially away from its maximum at $\hat{r}$ with length scale $\sigma$. We have already seen that $\sigma \ll \hat{r}$ for large $D$, and so we see that most of the probability mass is concentrated in a thin shell at large radius. Finally, show that the probability density $p(x)$ is larger at the origin than at the radius $\hat{r}$ by a factor of $\exp(D/2)$. We therefore see that most of the probability mass in a high-dimensional Gaussian distribution is located at a different radius from the region of high probability density. This property of distributions in spaces of high dimensionality will have important consequences when we consider Bayesian inference of model parameters in later chapters.

1.21  (⋆⋆) Consider two nonnegative numbers $a$ and $b$, and show that, if $a \leq b$, then $a \leq (ab)^{1/2}$. Use this result to show that, if the decision regions of a two-class classification problem are chosen to minimize the probability of misclassification, this probability will satisfy

$$p(\text{mistake}) \leq \int \{p(x, C_1)p(x, C_2)\}^{1/2} \, dx. \quad (1.150)$$

1.22  (⋆) Given a loss matrix with elements $L_{kj}$, the expected risk is minimized if, for each $x$, we choose the class that minimizes (1.81). Verify that, when the loss matrix is given by $L_{kj} = 1 - I_{kj}$, where $I_{kj}$ are the elements of the identity matrix, this reduces to the criterion of choosing the class having the largest posterior probability. What is the interpretation of this form of loss matrix?

1.23  (⋆) Derive the criterion for minimizing the expected loss when there is a general loss matrix and general prior probabilities for the classes.
1.24 (** www) Consider a classification problem in which the loss incurred when an input vector from class \( C_k \) is classified as belonging to class \( C_j \) is given by the loss matrix \( L_{kj} \), and for which the loss incurred in selecting the reject option is \( \lambda \). Find the decision criterion that will give the minimum expected loss. Verify that this reduces to the reject criterion discussed in Section 1.5.3 when the loss matrix is given by \( L_{kj} = 1 - I_{kj} \). What is the relationship between \( \lambda \) and the rejection threshold \( \theta \)?

1.25 (*) www Consider the generalization of the squared loss function (1.87) for a single target variable \( t \) to the case of multiple target variables described by the vector \( t \) given by
\[
E[L(t, y(x))] = \int \int \|y(x) - t\|^2 p(x, t) \, dx \, dt. \tag{1.151}
\]

Using the calculus of variations, show that the function \( y(x) \) for which this expected loss is minimized is given by \( y(x) = E_t[t|x] \). Show that this result reduces to (1.89) for the case of a single target variable \( t \).

1.26 (*) By expansion of the square in (1.151), derive a result analogous to (1.90) and hence show that the function \( y(x) \) that minimizes the expected squared loss for the case of a vector \( t \) of target variables is again given by the conditional expectation of \( t \).

1.27 (** www) Consider the expected loss for regression problems under the \( L_q \) loss function given by (1.91). Write down the condition that \( y(x) \) must satisfy in order to minimize \( E[L_q] \). Show that, for \( q = 1 \), this solution represents the conditional median, i.e., the function \( y(x) \) such that the probability mass for \( t < y(x) \) is the same as for \( t \geq y(x) \). Also show that the minimum expected \( L_q \) loss for \( q \to 0 \) is given by the conditional mode, i.e., by the function \( y(x) \) equal to the value of \( t \) that maximizes \( p(t|x) \) for each \( x \).

1.28 (*) In Section 1.6, we introduced the idea of entropy \( h(x) \) as the information gained on observing the value of a random variable \( x \) having distribution \( p(x) \). We saw that, for independent variables \( x \) and \( y \) for which \( p(x, y) = p(x)p(y) \), the entropy functions are additive, so that \( h(x, y) = h(x) + h(y) \). In this exercise, we derive the relation between \( h \) and \( p \) in the form of a function \( h(p) \). First show that \( h(p^2) = 2h(p) \), and hence by induction that \( h(p^n) = nh(p) \) where \( n \) is a positive integer. Hence show that \( h(p^{n/m}) = (n/m)h(p) \) where \( m \) is also a positive integer. This implies that \( h(p^x) = xh(p) \) where \( x \) is a positive rational number, and hence by continuity when it is a positive real number. Finally, show that this implies \( h(p) \) must take the form \( h(p) \propto \ln p \).

1.29 (*) www Consider an \( M \)-state discrete random variable \( x \), and use Jensen’s inequality in the form (1.115) to show that the entropy of its distribution \( p(x) \) satisfies \( H[x] \leq \ln M \).

1.30 (** Evaluate the Kullback-Leibler divergence (1.113) between two Gaussians \( p(x) = \mathcal{N}(x|\mu, \sigma^2) \) and \( q(x) = \mathcal{N}(x|m, s^2) \).
Table 1.3  The joint distribution $p(x, y)$ for two binary variables $x$ and $y$ used in Exercise 1.39.

<table>
<thead>
<tr>
<th>$x$</th>
<th>0</th>
<th>1/3</th>
<th>1/3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1/3</td>
<td>1/3</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1/3</td>
<td></td>
</tr>
</tbody>
</table>

### Exercises

1.31  (**WWW**) Consider two variables $x$ and $y$ having joint distribution $p(x, y)$. Show that the differential entropy of this pair of variables satisfies

$$H[x, y] \leq H[x] + H[y]$$  \hspace{1cm} (1.152)

with equality if, and only if, $x$ and $y$ are statistically independent.

1.32  (*) Consider a vector $x$ of continuous variables with distribution $p(x)$ and corresponding entropy $H[x]$. Suppose that we make a nonsingular linear transformation of $x$ to obtain a new variable $y = Ax$. Show that the corresponding entropy is given by $H[y] = H[x] + \ln |A|$ where $|A|$ denotes the determinant of $A$.

1.33  (***) Suppose that the conditional entropy $H[y|x]$ between two discrete random variables $x$ and $y$ is zero. Show that, for all values of $x$ such that $p(x) > 0$, the variable $y$ must be a function of $x$, in other words for each $x$ there is only one value of $y$ such that $p(y|x) \neq 0$.

1.34  (**WWW**) Use the calculus of variations to show that the stationary point of the functional (1.108) is given by (1.108). Then use the constraints (1.105), (1.106), and (1.107) to eliminate the Lagrange multipliers and hence show that the maximum entropy solution is given by the Gaussian (1.109).

1.35  (*) WWW Use the results (1.106) and (1.107) to show that the entropy of the univariate Gaussian (1.109) is given by (1.110).

1.36  (*) A strictly convex function is defined as one for which every chord lies above the function. Show that this is equivalent to the condition that the second derivative of the function be positive.

1.37  (*) Using the definition (1.111) together with the product rule of probability, prove the result (1.112).

1.38  (**WWW**) Using proof by induction, show that the inequality (1.114) for convex functions implies the result (1.115).

1.39  (**++) Consider two binary variables $x$ and $y$ having the joint distribution given in Table 1.3.

Evaluate the following quantities

(a) $H[x]$  \hspace{1cm} (c) $H[y|x]$  \hspace{1cm} (e) $H[x, y]$

(b) $H[y]$  \hspace{1cm} (d) $H[x|y]$  \hspace{1cm} (f) $I[x, y]$.

Draw a diagram to show the relationship between these various quantities.
1.40 (∗) By applying Jensen’s inequality (1.115) with \( f(x) = \ln x \), show that the arithmetic mean of a set of real numbers is never less than their geometrical mean.

1.41 (∗) Using the sum and product rules of probability, show that the mutual information \( I(x, y) \) satisfies the relation (1.121).