Chapter 4

Information Theory

4.1 Introduction

This lecture covers entropy, joint entropy, mutual information and minimum description length. See the texts by Cover [12] and Mackay [36] for a more comprehensive treatment.

4.2 Measures of Information

Information on a computer is represented by binary bit strings. Decimal numbers can be represented using the following encoding. The position of the binary digit

Bit 1 ($2^3 = 8$)	Bit 2 ($2^2 = 4$)	Bit 3 ($2^1 = 2$)	Bit 4 ($2^0 = 1$)	Decimal
0	0	0	0	0
0	0	0	1	1
0	0	1	0	2
0	0	1	1	3
0	1	0	0	4
.	.	.	.	.
.	.	.	.	.
0	1	1	1	14
1	1	1	1	15

Table 4.1: *Binary encoding*

indicates its decimal equivalent such that if there are N bits the ith bit represents the decimal number 2^{N-i}. Bit 1 is referred to as the most significant bit and bit N as the least significant bit. To encode M different messages requires $\log_2 M$ bits.

4.3 Entropy

The table below shows the probability of occurrence $p(x_i)$ (to two decimal places) of selected letters x_i in the English alphabet. These statistics were taken from Mackay's book on Information Theory [36]. The table also shows the *information content* of a

x_i	$p(x_i)$	$h(x_i)$
a	0.06	4.1
e	0.09	3.5
j	0.00	10.7
q	0.01	10.3
t	0.07	3.8
z	0.00	10.4

Table 4.2: *Probability and Information content of letters*

letter

$$h(x_i) = \log \frac{1}{p(x_i)} \tag{4.1}$$

which is a measure of *surprise*; if we had to guess what a randomly chosen letter of the English alphabet was going to be, we'd say it was an A, E, T or other frequently occuring letter. If it turned out to be a Z we'd be surprised. The letter E is so common that it is unusual to find a sentence without one. An exception is the 267 page novel 'Gadsby' by Ernest Vincent Wright in which the author deliberately makes no use of the letter E (from Cover's book on Information Theory [12]). The *entropy* is the average information content

$$H(x) = \sum_{i=1}^{M} p(x_i) h(x_i) \tag{4.2}$$

where M is the number of discrete values that x_i can take. It is usually written as

$$H(x) = -\sum_{i=1}^{M} p(x_i) \log p(x_i) \tag{4.3}$$

with the convention that $0 \log 1/0 = 0$. Entropy measures uncertainty.

Entropy is maximised for a uniform distribution $p(x_i) = 1/M$. The resulting entropy is $H(x) = log_2 M$ which is the number of binary bits required to represent M different messages (first section). For $M = 2$, for example, the maximum entropy distribution is given by $p(x_1) = p(x_2) = 0.5$ (eg. an unbiased coin; biased coins have lower entropy).

The entropy of letters in the English language is 4.11 bits [12] (which is less than $log_2 26 = 4.7$ bits). This is however, the information content due to considering just the probability of occurence of letters. But, in language, our expectation of what the next letter will be is determined by what the previous letters have been. To measure this we need the concept of joint entropy. Because $H(x)$ is the entropy of a 1-dimensional variable it is sometimes called the scalar entropy, to differentiate it from the joint entropy.

4.4 Joint Entropy

Table 2 shows the probability of occurence (to three decimal places) of selected pairs of letters x_i and y_i where x_i is followed by y_i. This is called the joint probability $p(x_i, y_i)$. The table also shows the joint information content

x_i	y_j	$p(x_i, y_j)$	$h(x_i, y_j)$
t	h	0.037	4.76
t	s	0.000	13.29
t	r	0.012	6.38

Table 4.3: *Probability and Information content of pairs of letters*

$$h(x_i, y_j) = \log \frac{1}{p(x_i, y_j)} \tag{4.4}$$

The average joint information content is given by the *joint entropy*

$$H(x, y) = -\sum_{i=1}^{M} \sum_{j=1}^{M} p(x_i, y_j) \log p(x_i, y_j) \tag{4.5}$$

If we fix x to, say x_i then the probability of y taking on a particular value, say y_j, is given by the *conditional probability*

$$p(y = y_j | x = x_i) = \frac{p(x = x_i, y = y_j)}{p(x = x_i)} \tag{4.6}$$

For example, if $x_i = t$ and $y_j = h$ then the joint probability $p(x_i, y_j)$ is just the probability of occurrence of the pair (which from table 2 is 0.037). The conditional probability $p(y_j|x_i)$, however, says that, given we've seen the letter t, what's the probability that the next letter will be h (which from tables 1 and 2 is $0.037/0.07 = 0.53$). Re-arranging the above relationship (and dropping the $y = y_j$ notation) gives

$$p(x, y) = p(y|x)p(x) \tag{4.7}$$

Now if y does *not* depend on x then $p(y|x) = p(y)$. Hence, for independent variables, we have

$$p(x, y) = p(y)p(x) \tag{4.8}$$

This means that, for independent variables, the joint entropy is the sum of the individual (or *scalar* entropies)

$$H(x, y) = H(x) + H(y) \tag{4.9}$$

Consecutive letters in the English language are not independent (except either after or during a bout of serious drinking). If we take into account the statistical dependence on the previous letter, the entropy of English reduces to 3.67 bits per letter (from 4.11). If we look at the statistics of not just pairs, but triplets and quadruplets of letters or at the statistics of words then it is possible to calculate the entropy more accurately; as more and more contextual structure is taken into account the estimates of entropy reduce. See Cover's book ([12] page 133) for more details.

4.5 Relative Entropy

The *relative entropy* or *Kullback-Liebler Divergence* between a distribution $q(x)$ and a distribution $p(x)$ is defined as

$$D[q||p] = \sum_x q(x) \log \frac{q(x)}{p(x)} \tag{4.10}$$

Jensen's inequality states that for any convex function [1] $f(x)$ and set of M positive coefficients $\{\lambda_j\}$ which sum to one

$$f(\sum_{j=1}^{M} \lambda_j x_j) \geq \sum_{j=1}^{M} \lambda_j f(x_j) \tag{4.11}$$

A sketch of a proof of this is given in Bishop ([3], page 75). Using this inequality we can show that

$$\begin{aligned} -D[q||p] &= \sum_x q(x) \log \frac{p(x)}{q(x)} \\ &\leq \log \sum_x p(x) \\ &\leq \log 1 \end{aligned} \tag{4.12}$$

Hence

$$D[q||p] \geq 0 \tag{4.13}$$

The KL-divergence will appear again in the discussion of the EM algorithm and Variational Bayesian learning (see later lectures).

4.6 Mutual Information

The *mutual information* is defined [12] as the relative entropy between the joint distribution and the product of individual distributions

$$I(x;y) = D[p(X,Y)||p(X)p(Y)] \tag{4.14}$$

Substuting these distributions into 4.10 allows us to express the mutual information as the difference between the sum of the individual entropies and the joint entropy

$$I(x;y) = H(x) + H(y) - H(x,y) \tag{4.15}$$

Therefore if x and y are independent the mutual information is zero. More generally, $I(x;y)$ is a measure of the *dependence* between variables and this dependence will be captured if the underlying relationship is linear *or* nonlinear. This is to be contrasted with Pearsons correlation coefficient, which measures only linear correlation (see first lecture).

[1] A convex function has a negative second derivative.

4.7 Minimum Description Length

Given that a variable has a determinisitic component and a random component the *complexity* of that variable can be defined as the length of a concise description of that variables regularities [19].

This definition has the merit that both random data and highly regular data will have a low complexity and so we have a correspondence with our everyday notion of complexity [2]

The length of a description can be measured by the number of binary bits required to encode it. If the probability of a set of measurements D is given by $p(D|\boldsymbol{\theta})$ where $\boldsymbol{\theta}$ are the parameters of a probabilistic model then the minimum length of a code for representing D is, from Shannon's coding theorem [12], the same as the information content of that data under the model (see eg. equation 4.1)

$$L = -\log p(D|\boldsymbol{\theta}) \tag{4.16}$$

However, for the recevier to decode the message they will need to know the parameters $\boldsymbol{\theta}$ which, being real numbers are encoded by truncating each to a finite precision $\Delta\theta$. We need a total of $-k \log \Delta\theta$ bits to encode the This gives

$$L_{tx} = -\log p(D|\boldsymbol{\theta}) - k \log \Delta\theta \tag{4.17}$$

The optimal precision can be found as follows. First, we expand the negative log-likelihood (ie. the error) using a Taylor series about the Maximum Likelihood (ML) solution $\hat{\boldsymbol{\theta}}$. This gives

$$L_{tx} = -\log p(D|\hat{\boldsymbol{\theta}}) + \frac{1}{2}\Delta\boldsymbol{\theta}^T H \Delta\boldsymbol{\theta} - k \log \Delta\theta \tag{4.18}$$

where $\Delta\boldsymbol{\theta} = \boldsymbol{\theta} - \hat{\boldsymbol{\theta}}$ and H is the Hessian of the error which is identical to the inverse covariance matrix (the first order term in the Taylor series disappears as the error gradient is zero at the ML solution). The derivative is

$$\frac{\partial L_{tx}}{\partial \Delta\theta} = H\Delta\boldsymbol{\theta} - \frac{k}{\Delta\theta} \tag{4.19}$$

If the covariance matrix is diagonal (and therefore the Hessian is diagonal) then, for the case of linear regression (see equation 1.47) the diagonal elements are

$$h_i = \frac{N\sigma_{x_i}^2}{\sigma_e^2} \tag{4.20}$$

where σ_e^2 is the variance of the errors and $\sigma_{x_i}^2$ is the variance of the ith input. More generally, eg. nonlinear regression, this last variance will be replaced with the variance

[2] This is not the case, however, with measures such as the Algorithm Information Content (AIC) or Entropy as these will be high even for purely random data.

of the derivative of the output wrt. the ith parameter. But the dependence on N remains. Setting the above derivative to zero therefore gives us

$$(\Delta\theta)^2 = \frac{1}{N} \times constant \tag{4.21}$$

where the constant depends on the variance terms (when we come to take logs of $\Delta\theta$ this constant becomes an additive term that does'nt scale with either the number of data points or the number of parameters in the model; we can therefore ignore it). The *Minimum Description Length (MDL)* is therefore given by

$$MDL(k) = -\log p(D|\boldsymbol{\theta}) + \frac{k}{2}\log N \tag{4.22}$$

This may be minimised over the number of parameters k to get the optimal model complexity.

For a linear regression model

$$-\log p(D|\boldsymbol{\theta}) = \frac{N}{2}\log \sigma_e^2 \tag{4.23}$$

Therefore

$$MDL_{Linear}(k) = \frac{N}{2}\log \sigma_e^2 + \frac{k}{2}\log N \tag{4.24}$$

which is seen to consist of an accuracy term and a complexity term. This criterion can be used to select the optimal number of input variables and therefore offers a solution to the bias-variance dilemma (see lecture 1). In later lectures the MDL criterion will be used in autoregressive and wavelet models.

The MDL complexity measure can be further refined by integrating out the dependence on $\boldsymbol{\theta}$ altogether. The resulting measure is known as the *stochastic complexity* [54]

$$I(k) = -\log p(D|k) \tag{4.25}$$

where

$$p(D|k) = \int p(D|\boldsymbol{\theta}, k)p(\boldsymbol{\theta})d\boldsymbol{\theta} \tag{4.26}$$

In Bayesian statistics this quantity is known as the 'marginal likelihood' or 'evidence'. The stochastic complexity measure is thus equivalent (after taking negative logs) to the Bayesian model order selection criterion (see later). See Bishop ([3], page 429) for a further discussion of this relationship.