Exercise Problems: Information Theory and Coding

Exercise 1

Prove that the information measure is additive: that the information gained from observing the combination of N independent events, whose probabilities are p_i for $i = 1....N$, is the *sum* of the information gained from observing each one of these events separately and in any order.

Solution: The information measure assigns $\log_2(p)$ bits to the observation of an event whose probability is p. The probability of the combination of N independent events whose probabilities are $p_1....p_N$ is: $\prod_{i=1}^{N} p_i$ Thus the information content of such a combination is:

$$\log_2(\prod_{i=1}^{N} p_i) = \log_2(p_1) + \log_2(p_2) + \cdots + \log_2(p_N)$$

which is the sum of the information content of all of the separate events.

Exercise 2

(a) What is the maximum possible entropy H of an alphabet consisting of N different letters? In such a maximum entropy alphabet, what is the probability of its most likely letter? What is the probability of its least likely letter? Why are fixed length codes inefficient for alphabets whose letters are not equiprobable? Discuss this in relation to Morse Code.

Solution: (a)
The maximum possible entropy of an alphabet consisting of N different letters is $H = \log_2 N$. This is only achieved if the probability of every letter is $1/N$. Thus $1/N$ is the probability of both the "most likely" and the "least likely" letter.

(b) Let X and Y represent random variables with associated probability distributions $p(x)$ and $p(y)$, respectively. They are not independent. Their conditional probability distributions are $p(x|y)$ and $p(y|x)$, and their joint probability distribution is $p(x, y)$.

1. What is the marginal entropy $H(X)$ of variable X, and what is the mutual information of X with itself?
2. In terms of the probability distributions, what are the conditional entropies $H(X|Y)$ and $H(Y|X)$?
3. What is the joint entropy $H(X, Y)$, and what would it be if the random variables X and Y were independent?
4. Give an alternative expression for $H(Y) - H(Y|X)$ in terms of the joint entropy and both marginal entropies.
5. What is the mutual information $I(X; Y)$?

Solution: (b)

1. $H(X) = -\sum_x p(x) \log_2 p(x)$ is both the marginal entropy of X, and its mutual information with itself.

2. $H(X|Y) = -\sum_y p(y) \sum_x p(x|y) \log_2 p(x|y) = -\sum_x \sum_y p(x,y) \log_2 p(x|y)$
   $H(Y|X) = -\sum_x p(x) \sum_y p(y|x) \log_2 p(y|x) = -\sum_x \sum_y p(x,y) \log_2 p(y|x)$

3. $H(X,Y) = -\sum_x \sum_y p(x,y) \log_2 p(x,y)$.
   If X and Y were independent random variables, then $H(X,Y) = H(X) + H(Y)$.

4. $H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y)$.

5. $I(X;Y) = \sum_x \sum_y p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)}$
   or: $\sum_x \sum_y p(x,y) \log_2 \frac{p(x|y)}{p(x)}$
   or: $I(X;Y) = H(X) - H(X|Y) = H(X) + H(Y) - H(X,Y)$

(c) Consider two independent integer-valued random variables, X and Y. Variable X takes on only the values of the eight integers {1, 2, ..., 8} and does so with uniform probability. Variable Y may take the value of *any* positive integer k, with probabilities $P\{Y = k\} = 2^{-k},\ k = 1, 2, 3, \ldots$.

1. Which random variable has greater uncertainty? Calculate both entropies $H(X)$ and $H(Y)$.
2. What is the joint entropy $H(X, Y)$ of these random variables, and what is their mutual information $I(X; Y)$?

Solution: (c)

1. Surprisingly, there is greater uncertainty about random variable X which is just any one of the first 8 integers, than about Y which can be *any* positive integer. The uniform probability distribution over the eight possibilities for X means that it has entropy $H(X) = 3$ bits. But the rapidly decaying probability distribution for Y has entropy

$$H(Y) = -\lim_{N\to\infty} \sum_{k=1}^{N} \frac{1}{2^k} \log_2(2^{-k})$$

   which is known to converge to just 2 bits.

2. Since random variables X and Y are independent, their joint entropy $H(X, Y)$ is $H(X) + H(Y) = 5$ bits, and their mutual information is $I(X; Y) = 0$ bits.

Exercise 3 (a)

(a) Calculate the entropy in bits for each of the following random variables:

(i) Pixel values in an image whose possible grey values are all the integers from 0 to 255 with uniform probability.

(ii) Humans classified according to whether they are, or are not, mammals.

(iii) Gender in a tri-sexual insect population whose three genders occur with probabilities 1/4, 1/4, and 1/2.

(iv) A population of persons classified by whether they are older, or not older, than the population's median age.

Solution: (a)

By definition, $H = -\sum_i p_i \log_2 p_i$ is the entropy in bits for a discrete random variable distributed over states whose probabilities are p_i. So:

(i) In this case each $p_i = 1/256$ and the ensemble entropy summation extends over 256 such equiprobable grey values, so $H = -(256)(1/256)(-8) = 8$ bits.

(ii) Since all belong to the single state (humans $\subset$ mammals), there is no uncertainty about this state and hence the entropy is 0 bits.

(iii) The entropy of this tri-state gender distribution is $-(1/4)(-2)-(1/4)(-2)-(1/2)(-1) = 1.5$ bits.

(iv) In this case both classes have probability 0.5, so the entropy is 1 bit.

Exercise 3 (b)

(*b*) Let $p(x)$ and $q(x)$ be two probability distributions specified over integers x.

(*i*) What is the *Kullback-Leibler distance* between these distributions?

(*ii*) If we have devised an optimally compact code for the random variable described by $q(x)$, what does the *Kullback-Leibler distance* tell us about the effectiveness of our code if the probability distribution is $p(x)$ instead of $q(x)$?

(*iii*) Which axiom of distance metrics is violated by this distance?

(*iv*) What happens to this metric if there are some forbidden values of x for which $p(x) = 0$, and other values of x for which $q(x) = 0$?

Solution: (b)

For $p(x)$ and $q(x)$ as probability distributions over the integers:

(*i*) The Kullback-Leibler distance between random variables is defined as

$$D_{KL}(p\|q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

(*ii*) $D_{KL}(p\|q)$ reveals the inefficiency that arises from basing the code on the wrong distribution. It specifies the number of additional bits that would be needed per codeword, on average, if the actual distribution is $p(x)$ instead of $q(x)$.

(*iii*) The symmetry axiom for distance metrics is violated in this case.

(*iv*) The Kullback-Leibler distance becomes infinite, or undefined, when some values of x have zero probability for either $p(x)$, or $q(x)$, but not both.

Exercise 4

Consider an alphabet of 8 symbols whose probabilities are as follows:

A	B	C	D	E	F	G	H
$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{64}$	$\frac{1}{128}$	$\frac{1}{128}$

1. If someone has selected one of these symbols and you need to discover which symbol it is by asking 'yes/no' questions that will be truthfully answered, what would be the most efficient sequence of such questions that you could ask in order to discover the selected symbol?
2. By what principle can you claim that each of your proposed questions is maximally informative?
3. On average, how many such questions will need to be asked before the selected symbol is discovered?
4. What is the entropy of the above symbol set?
5. Construct a uniquely decodable prefix code for the symbol set, and explain why it is uniquely decodable and why it has the prefix property.
6. Relate the bits in your prefix code to the 'yes/no' questions that you proposed in 1.

Solution:

1. For this symbol distribution, the most efficient sequence of questions to ask (until a 'yes' is obtained) would be just: (1) Is it A? (2) Is it B? (3) Is it C? (Etc.)
2. Each such 1-bit question is maximally informative because the remaining uncertainty is reduced by half (1 bit).
3. The probability of terminating successfully after exactly N questions is 2^{-N}. At most 7 questions might need to be asked. The weighted average of the interrogation durations is:

$$\frac{1}{2} + (2)(\frac{1}{4}) + (3)(\frac{1}{8}) + (4)(\frac{1}{16}) + (5)(\frac{1}{32}) + (6)(\frac{1}{64}) + (7)(\frac{2}{128}) = 1\frac{126}{128}$$

   In other words, on average just slightly less than <u>two</u> questions need to be asked in order to learn which of the 8 symbols it is.
4. The entropy of the above symbol set is calculated by the same formula, but over all 8 states (whereas at most 7 questions needed to be asked):

$$H = -\sum_{i=1}^{8} p_i \log_2 p_i = 1\frac{126}{128}$$

5. A natural code book to use would be the following:

A	B	C	D	E	F	G	H
1	01	001	0001	00001	000001	0000001	0000000

   It is uniquely decodable because each code corresponds to a unique letter rather than any possible combination of letters; and it has the prefix property because the code for no letter could be confused as the prefix for another letter.
6. The bit strings in the above prefix code for each letter can be interpreted as the history of answers to the 'yes/no' questions.