COMP6237 Logistic regression and model reduction

February 22, 2022

Abstract

Solutions for problem sheet 3 for COMP6237

1 Additivity of information

Prove that the information measure (slide 8 of the lecture slides) 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 $I = -log_2(p)$ bits to the observation of an event whose probability is p. The joint probability of a 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: $I_{\text{joint}} = -log_2(\prod_{i=1}^{N} p_i) = -\sum_{i=1}^{N} log_2 p_i = \sum_{i=1}^{N} I_i$. which is the sum of the information content of all of the separate events.

2 Entropy and information

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 $PY = k = 2^k$, $k = 1, 2, 3, ...$

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

solution:
The uniform probability distribution over the eight possibilities for X means that this random variable has entropy $H(X) = -\sum_{i=1}^{8} p_i log_2 p_i = 8 \times 1/8 \times 3 = 3$ bits. But the rapidly decaying probability distribution for

random variable Y has entropy $\sum_{i=1}^{\infty} 2^{-i} log_2 2^i = \sum_{i=1}^{\infty} i2^{-i} = 2$
The latter can be shown with some work using a standard trick by noting that $\sum_{k=1}^{\infty} = \frac{\partial}{\partial\alpha}_{|\alpha=1} \sum_{k=1}^{\infty} 2^{-\alpha k} = -\frac{1}{\ln 2} \frac{\partial}{\partial\alpha}_{|\alpha=1} \left(\sum_{k=0}^{\infty} 2^{-\alpha k} - 1\right)$. We can now use the well-known result for the geometric series $\sum_{k=0}^{\infty} q^k = 1/(1-q)$ for $-1 < q < 1$ and thus find $H(Y) = -\frac{1}{\ln 2} \frac{\partial}{\partial\alpha}_{|\alpha=1} \frac{1}{1-2^{-\alpha}} = -\frac{1}{\ln 2} \frac{-1}{(1-2^{-\alpha})^2} \times (-1) \times (-1) \times \ln 2 \times 2^{-\alpha}$ evaluated at $\alpha = 1$. Inserting $\alpha = 1$ gives the final result. The result may appear somewhat surprising, because $H(X) > H(Y)$ even though Y allows any positive integer whereas X allows for only eight possible events.
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.

3 Entropy

Assume that we have some random source that emits one of M symbols with equal likelihood. What is the entropy? Assume a source is restricted to emitting one of M symbols at a time. What is the distribution of probabilities over these symbols that maximises the average uncertainty of the receiver?

solution:
Part 1 of the question: The pdf of the given distribution is $p(i) = 1/M, i = 1, ..., M$. From the lecture, we remember that the entropies are given by (note that to be consistent with Boltzmann, here I use ln instead of $\log_2$ which only differs in a constant factor from the definition using $\log_2$) $S = -\sum_{i=1}^{M} p(i) \ln p(i) = -\sum_{i=1}^{M} 1/M \ln 1/M = \ln M$. This is Boltzmann's famous formula (which some of you might remember from school.) – the wiki pages have a bit more of a story around this if you are interested link to wiki page on Boltzmann.
Part 2: Suppose $p(x), x = 1, ..., M$ is an unknown pdf, i.e. normalized $\sum_{x=1}^{M} p(x) = 1$. We are interested in the pdf which maximizes uncertainty, i.e. we aim to maximize S subject to the constraint $\sum_{x=1}^{M} p(x) = 1$. Introducing the Lagrange multiplier λ, we thus aim to maximize $-\sum_{i=1}^{M} p(x) \ln p(x) + \lambda(\sum_x p(x) - 1)$. Taking partial derivatives with respect to the $p(x), x = 1, ..., M$, we obtain:

$$-\ln p(x) - 1 + \lambda = 1 \quad (1)$$

$$p(x) = \exp(\lambda - 1) \quad (2)$$

$$p(x) = \text{const.} \quad (3)$$

The Lagrange multiplier is to be determined from the constraint, i.e. $\sum_x p(x) = M \exp(\lambda - 1) = 1$, i.e. $\exp(\lambda - 1) = 1/M$ and hence $p(x) = 1/M, x = 1, ..., M$. Thus, (not surprisingly) the distribution that maximizes the entropy is the uniform distribution.

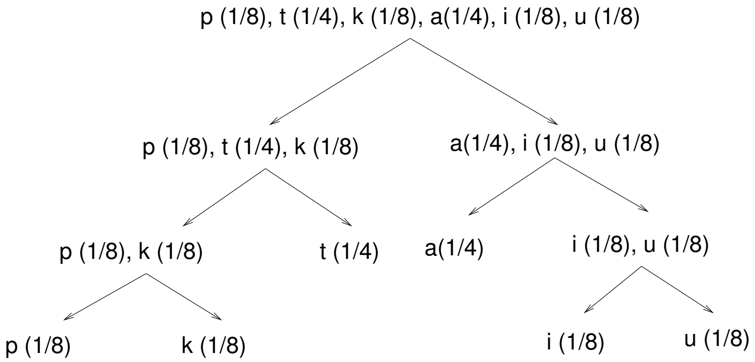

Figure 1: Figure illustrating the divide and conquer strategy applied to construct an optimal set of questions to encode the Polynesian alphabet.

4 Optimal codes for the Polynesian alphabet

Polynesian languages are famous for their small alphabets. Assume a language with the following letters and relative frequencies: $p(1/8)$, $t(1/4)$, $k(1/8)$, $a(1/4)$, $i(1/8)$, $u(1/8)$. What is the per-character entropy for this language? Design an (optimal, i.e. short) code to transmit a letter.

solution:
We start by developing a set of yes/no questions to identify which letter has been sampled. For this purpose, we use the divide and conquer strategy discussed in the lecture, always halfing the probability mass. This procedure is not unique, so you might make other choices than I have made here. For example, one could proceed as illustrated in Fig. 1. Thus, the first question is: Is the letter p, t, or k? If yes, we proceed on the left hand branch of the tree and next ask: Is it p or k? etc. The information content of each letter then corresponds to the number of questions we have to ask to identify this letter (or just $-\ln 1/p(x)$ for letter x). This yields the entropy: $S = 2 \times 1/4 \log 4 + 4 \times 1/8 \log 8 = 1/2(\log 4 + \log 8) = 5/2$. Optimal codes can also be read from the tree given in Fig. 1. For my choice of questions I obtain (and depending on your questions you might have obtained a different code, but with the same code length for each symbol): $p - 111$, $t - 10$, $k - 110$, $a - 01$, $i - 001$, $u - 000$.

5 Entropy

Find an example for three random variables X, Y, Z with negative interaction $I(X;Y|Z) < I(X;Y)$ and one for positive interaction $I(X;Y|Z) > I(X;Y)$.

solution:

My examples are taken from the wiki page for interaction information.

Positive interaction information is typical for common-cause structures. For example, clouds cause rain and also block the sun; therefore, the correlation between rain and darkness is partly accounted for by the presence of clouds, $I(rain; dark|cloud) < I(rain; dark)$.

A prototypical example of negative interaction information has X as the output of an XOR gate to which Y and Z are the independent random inputs. In this case $I(Y; Z)$ will be zero, but $I(Y; Z|X)$ will be positive (1 bit) since once output X is known, the value on input Y completely determines the value on input Z. Thus $I(Y; Z|X) > I(Y; Z)$.