

SIMG-714 Information Theory for Imaging Science

Homework 1

1. A table of letter frequencies in English is given below. Compute the first order entropy of this model of English. The entropy is given by

$$H = - \sum_k p_k \log_2 p_k$$

Letter	Frequency	Letter	Frequency	Letter	Frequency
E	0.1310	D	0.0380	W	0.0130
T	0.1050	L	0.0330	B	0.0120
A	0.0860	F	0.0290	V	0.0092
O	0.0800	C	0.0280	K	0.0042
N	0.0710	M	0.0250	X	0.0017
R	0.0680	U	0.0250	J	0.0014
I	0.0630	G	0.0200	Q	0.0013
S	0.0610	Y	0.0200	Z	0.0012
H	0.0530	P	0.0200		

Table 1: Approximate letter frequencies.

2. Construct an instantaneous binary code for this alphabet. Compute the average length and efficiency for your code.
3. Suppose that we want to construct a language that consists only of words that contain at most one occurrence of any letter. How many different words can be constructed from an alphabet of size r ? Note that this language has a limit in size.
4. Consider another language where the words are constructed of symbols drawn from a binary alphabet $\mathcal{A}_2 = \{s_1, s_2\}$. In every word, s_1 occurs exactly n_1 times and s_2 occurs exactly n_2 times, with n_1 and n_2 fixed. How many words of length $n = n_1 + n_2$ can be formed in this manner?
5. In this problem we will investigate a formula that provides a good approximation to the number of words of length $n = n_1 + n_2$ that can be constructed with the alphabet $\mathcal{A}_2 = \{s_1, s_2\}$ in which the fraction of each letter is kept constant as the length is increased. Let p be a parameter in the interval $[0, 1]$. Let $n_1 = np$ and $n_2 = n(1 - p)$. Clearly, $n_1 + n_2 = n$. Let $f(n, p)$ be the exact function for the number of such words (constructed by substituting n_1 and n_2 into the result of the previous problem). Let $g(n, p) = -n[p \log_e p + (1 - p) \log_e(1 - p)]$. Let

$$r(n, p) = \frac{\log_e f(n, p)}{g(n, p)}$$

Plot $r(n, p)$ vs $\log_{10} n$ for $p = 0.5$. Use $n = 10, 100, 1000, \dots, 10^6$. What do you conclude about the quality of the approximation?

6. Derive the formula for $g(n, p)$ by starting with the formula for $f(n, p)$ and making use of the Stirling approximation

$$\log_e(N!) \approx \left(N + \frac{1}{2}\right) \log_e N - N + \frac{1}{2} \log_e 2\pi$$

Keep only the dominant terms as you let n increase.