FMCS Assessment Four

Miles Osborne
Informatics
miles@inf.ed.ac.uk

March 12, 2007

A probability distribution can be thought of as a *model* of some process. The *entropy* of that distribution in turn measures how good that model is: all things being equal, we prefer a model with a lower entropy over one with a higher entropy.

For this assessment, you are going to compute entropies of two models: the first one is $P(C)$, where C is a random variable which can take on characters (for example, *a b d ! ?* etc; note that a space is also a character. The second model will be $P(C_1 \mid C_2)$, which computes the probability of a character, given some preceding character. For example, we might work out $P(n \mid e)$ –the probability that the letter n follows the letter e.

Now, given a set of letters, we can compute probabilities as follows:

$$P(C) = \frac{\text{freq}(C)}{N}$$

Here, N is the total of all letter frequencies and freq(C) is the frequency of some letter. Likewise, we can compute a conditional probability as follows:

$$P(C_1 \mid C_2) = \frac{\text{freq}(C_1, C_2)}{\text{freq}(C_2)}$$

For this assessment, you are to:

1. Download the course web page: http://www.inf.ed.ac.uk/teaching/courses/fmcs1/ and save it as a text file.
2. Work-out the set of all distinct characters which occur in that file. Remember to include space characters. What are they?
3. Work out $P(C)$ and $P(C_1 \mid C_2)$. You can assume that the very first line starts with a dummy symbol and that new lines are ignored. Compute $H(P(C))$ and $H(P(C_1 \mid C_2))$. What are these values? What can you conclude about how predictable Web pages are?
4. Add one to the count of every distinct letter you found in the Web page and recompute the two entropy values. What happens to the two entropy values? Can you explain this?