Lecture 5: Entropy rate for stochastic processes

Biology 429
Carl Bergstrom

January 23, 2008

This lecture loosely follows Cover and Thomas chapter 4. As usual, some of the text and equations are taken directly from that source.

Suppose we observe the following sequence:

$$5, 6, 6, 6, 1, 1, 1, 1, 1, 2, 2, 3, 3, 3, 4, 4, 4, 4, 5, 5, 6, 1$$

What is the entropy rate of this sequence?

Last week we saw how to compute the entropy rate of a set of i.i.d. random variables. In practice, most of the things that we observe or measure in biology — particularly those for which we want to use information theory — are not i.i.d. The aim to today's lecture is to extend previous results so that we can talk about the entropy rate of any stochastic process:

$$\Pr\Big[(X_1, X_2, \ldots, X_n) = (x_1, x_2, \ldots, x_n)\Big] = p(x_1, x_2, \ldots, x_n)$$

That said, we'll approach the problem by looking at *stationary processes*: stochastic processes for which the joint probability is unchanged for any time shift k:

$$\Pr\Big[(X_1, X_2, \ldots, X_n) = (x_1, x_2, \ldots, x_n)\Big] \quad =$$
$$\Pr\Big[(X_{1+k}, X_{2+k}, \ldots, X_{n+k}) = (x_1, x_2, \ldots, x_n)\Big]$$

In today's lecture, we'll pay particular attention to to a class of stochastic processes called *Markov processes.* A Markov process is a stochastic process for which the conditional probability distribution for the i-th observation X_i depends only on the value the previous observation $X_{i-1} = x_{i-1}$. More formally, X_i depends on X_{i-1} but is conditionally independent of all previous X_j.

Markov processes can be described as chains of conditional probabilities (and thus are sometimes called Markov chains):

$$p(x_1, x_2, \ldots, x_n) = p(x_1)p(x_2|x_1)p(x_3|x_2)\ldots p(x_n|x_{n-1})$$

We'll assume throughout that Markov chains are time-invariant, i.e., that these conditional probabilities $p(x_i|x_{i-1})$ do not change over time.

Returning to our sample sequence, we can now represent it as a graph. Alternatively, we can represent it as a matrix:

$$M = \begin{pmatrix} 1/2 & 1/2 & 0 & 0 & 0 & 0 \\ 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 1/2 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 1/2 & 0 & 0 & 0 & 0 & 1/2 \end{pmatrix}$$

For those of you who know a bit about the mathematics of Markov processes, if $\mathbf{p}(t)$ is the probability distribution at time t, then the probability distribution at time $t+1$ is simply

$$\mathbf{p}(t+1) = M\,\mathbf{p}(t).$$

More importantly, we can now see that at each draw X_i, there there are not six possible outcomes, but only two. In fact, I generated this data by flipping a coin at each stage and using the coin flip to choose which way to move.

Now, the key intuition is that if this is the process at work, we could code the entire sequence by simply recording the starting position, and the heads and tails outcome of each round.

Now these heads and tails are i.i.d. random variables, and we know from last time that a coin flip has an entropy rate of one bit (log[2]). So the

entropy rate of this stochastic process should be one bit. We'll spend some time now trying to prove this more formally and generally.

For any Markov process, we can write a probability transition matrix akin to the matrix M above, where $M_{ij} = \Pr[X_{n+1} = i | X_n = j]$. The probability distribution at time $t+1$ is then

$$\mathbf{p}(t+1) = M\,\mathbf{p}(t).$$

as before. If we have a probability distribution that doesn't change over time, i.e.,

$$\mathbf{p}(t+1) = M\,\mathbf{p}(t) = \mathbf{p}(t)$$

we call this a *stationary distribution*. If the initial state of a Markov process is a stationary distribution, then the Markov process is a stationary process. (There are various theorems about how Markov processes have stationary distributions and how "well-behaved" ones have unique stationary distributions.)

Definition 1 *The* entropy rate $H(\mathcal{X})$ *of a stochastic process* $\mathcal{X}$ *is defined as the limit where it exists*

$$H(\mathcal{X}) = \lim_{n\to\infty} \frac{1}{n} H(X_1, X_2, \ldots, X_n).$$

Example: If the X_i are i.i.d., this limit is simply the entropy rate of the i.i.d. random variable: $H(\mathcal{X}) = \lim_{n\to\infty} n\,H(X_1)/n = H(X)$.

Definition 2 *The* limiting conditional entropy $H'(\mathcal{X})$ *of a stochastic process* $\mathcal{X}$ *is defined as the limit where it exists*

$$H'(\mathcal{X}) = \lim_{n\to\infty} H(X_n | X_{n-1}, X_{n-2}, \ldots, X_1).$$

Theorem 1 *For a stationary stochastic process, both limits exist and they are equal:* $H'(\mathcal{X}) = H(\mathcal{X})$

The proof is straightforward but requires a bit of analysis; it is given in full on pages 64–65 of Cover and Thomas.

Using this proof, we can finally compute the entropy rate for a stationary Markov process. By theorem 1, the entrory rate of a Markov process M is equal to its limiting conditional entropy.

$$H(\mathcal{X}) = H'(\mathcal{X}) = \lim_{n\to\infty} H(X_n|X_{n-1}, X_{n-2}, \ldots, X_1).$$

But by the definition of a Markov process, the value of X_i depends only on X_{i-1}. Thus

$$\lim_{n\to\infty} H(X_n|X_{n-1}, X_{n-2}, \ldots, X_1) = \lim_{n\to\infty} H(X_n|X_{n-1}) = H(X_2|X_1),$$

where the last equality follows by stationarity. Thus we have, for Markov process M with stationary distribution μ,

$$H(\mathcal{X}) = H(X_2|X_1) = -\sum_{ij} \mu_j M_{ij} \log M_{ij}.$$

In other words, this is the average entropy of the next move $j \to i$ out of each state j, with the average weighted by the stationary probability of being in state j.