Assignment 2 COMP 423 Prof. M. Langer

Instructions

- Posted Fri. Feb. 15, 2008. Hardcopy due **in class** on Wed. March 5, 2008.
- Assignment will be marked out of 20.
- Late penalty is 2 points per day (e-copy is due by midnight).
- Please submit a hardcopy of answers to the questions below.
- *You may use whatever programming language you like for this assignment.* You do not need to submit a hard copy of the source code this time. Instead, please use `handin` to submit any code that you write. Include a README file that explains the different programs you use.

  Complete instructions on how to use handin are given at
  `http://www.cs.mcgill.ca/socsinfo/handin/`
- Be neat and organized. Points will be taken off for excessive sloppiness.

Introduction

Write a program that generates a bit string, based on the transition matrix:

$$P(X_{j+1}|X_j) \ = \ \begin{bmatrix} \frac{1}{4} & \frac{3}{8} \\ \frac{3}{4} & \frac{5}{8} \end{bmatrix}$$

where, for example, $P(X_{j+1} = 1|X_j = 0) \ = \ \frac{3}{4}$. To start the sequence, assume $p(X_1)$ is uniform (a fair coin toss).

Use these generated bit strings to answer the following questions.

Question 1 (5 points)

Given a data sequence of length $n = 2^{16}$ generated by this program:

1. Estimate the conditional probabilities using k^{th} order models where $k = 0, 1, 2$. You will need to compute the frequencies of joint events and estimate probabilities based on these frequencies.
2. Compare these estimates to what you would expect from the transition matrix. (Note: the matrix is for a first order model only, so you will need to explain what you expect from the zeroth and second order model.)

Question 2 (10 points)

Write a program that parses a long data sequence, using the LZ3 algorithm. You do not need to encode the parsed sequence. You only need to compute how many phrases $\phi(n)$ there are.

1. Calculate $\phi(n)$ for $n = 2^{14}, 2^{15}, 2^{16}$ and give a 3×3 table showing $n, \phi(n), totbits(n)$ for the three values of n, where

$$totbits = \sum_{i=1}^{\phi(n)} (\lceil \log i \rceil + 1) \ .$$

2. Compare $totbits(n)$ from your experiments to the best and worst case bounds seen in class.

3. Compare the compression ratio, $totbits(n)/n$, to the conditional entropy $H(X_{j+1}|X_j)$, for the stationary case.

Question 3 (5 points)

Given a data sequence of length $n \approx 2^{16}$, partition this sequence into disjoint $m-$tuples where $m = 1, 2, 3, 4, 5$.

1. Estimate the joint probability functions $p(X_1, \ldots, X_m)$, based on frequencies of the $m-$tuples. Note that since $N = 2$, there are 2^m possible $m-$tuples and so you need this many joint probabilities.

2. Calculate the entropy $H(X_1, \ldots, X_m)$ and the entropy per symbol $H(X_1, \ldots, X_m)/m$.