Machine Learning and Artificial Intelligence - COS 402

Written Homework Assignment 1

Due Date: one week from announcement in class, due in class

(1) Consulting other students from this course is allowed. In this case - clearly state whom you consulted with for each problem separately.

(2) Searching the internet or literature for solutions is NOT allowed.

(3) Submit your homework in separate pages for the different questions, each including your name and email address (this is to help the graders).

I Compute the entropy of the following distributions:

- The distribution on integers from one to $n \geq 2$, where i has probability proportional to 2^{-i} (scaled such that all probabilities sum up to one). Stated equivalently, for this distribution it holds that

$$\frac{\Pr[i]}{\Pr[i+1]} = 2$$

- The uniform distribution on all binary strings of length n, with exactly k ones.

II In this exercise we show that entropy is a lower bound on lossless compression. Suppose files are sequences of m bits, of which $m \cdot p$ are 1 and $m \cdot (1-p)$ are 0. Here $p \in (0,1)$ is some fraction.

- Give an expression for the total number of distinct files.
- Let N be the number computed in the previous part. Show that

$$\lim_{m \to \infty} \frac{1}{m} \log N = H(X_p),$$

  where X_p is a Bernoulli random variable with parameter p.

  You may use Stirling's approximation:

$$n! \approx \sqrt{2\pi n} \left(\frac{n}{e}\right)^n.$$

- Imagine a file compression algorithm that, given any file of length m, compresses it to $\tilde{m}$ bits. Show that if $\tilde{m} < m \cdot (H(X_p) - \varepsilon)$ for some $\varepsilon > 0$, then it

must necessarily be a lossy compression; meaning that two different files must correspond to the same compressed file.