Recitation Oct. 5th, 2022

Information Theory

Problems from Cover and Thomas

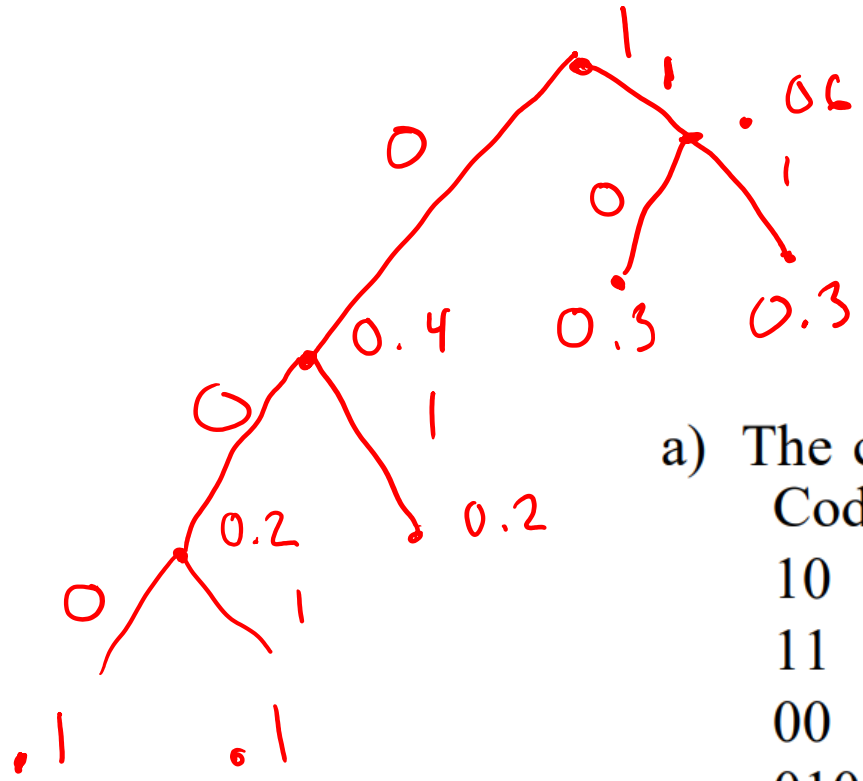
Problem 1: Huffman Codes

15) **Huffman codes.**

a) Construct a binary Huffman code for the following distribution on 5 symbols $\mathbf{p} = (0.3, 0.3, 0.2, 0.1, 0.1)$. What is the average length of this code?

b) Construct a probability distribution $\mathbf{p}'$ on 5 symbols for which the code that you constructed in part (a) has an average length (under $\mathbf{p}'$) equal to its entropy $H(\mathbf{p}')$.

Solution 1a



a) The code constructed by the standard Huffman procedure

Codeword	X	Probability				
10	1	0.3	0.3	0.4	0.6	1
11	2	0.3	0.3	0.3	0.4	
00	3	0.2	0.2	0.3		
010	4	0.1	0.2			
011	5	0.1				

The average length $= 2 * 0.8 + 3 * 0.2 = 2.2$ bits/symbol.

Solution 1b

b) The code would have a rate equal to the entropy if each of the codewords was of length $1/p(X)$. In this case, the code constructed above would be efficient for the distrubution (0.25.0.25,0.25,0.125,0.125).

$$\bar{\ell} = \sum p(x)\ell_x = \sum p(x)\log\frac{1}{p(x)} = H(X)$$

$$2^{\ell_x} = \frac{1}{p(x)} \qquad p(x) = \frac{1}{2^{\ell_x}}$$

Problem 2: Huffman Codes with costs

20) **Huffman codes with costs.** Words like Run! Help! and Fire! are short, not because they are frequently used, but perhaps because time is precious in the situations in which these words are required. Suppose that $X = i$ with probability $p_i, i = 1, 2, \ldots, m$. Let l_i be the number of binary symbols in the codeword associated with $X = i$, and let c_i denote the cost per letter of the codeword when $X = i$. Thus the average cost C of the description of X is $C = \sum_{i=1}^{m} p_i c_i l_i$.

a) Minimize C over all $l_1, l_2, \ldots, l_m$ such that $\sum 2^{-l_i} \leq 1$. Ignore any implied integer constraints on l_i. Exhibit the minimizing $l_1^*, l_2^*, \ldots, l_m^*$ and the associated minimum value C^*.

b) How would you use the Huffman code procedure to minimize C over all uniquely decodable codes? Let $C_{Huffman}$ denote this minimum.

c) Can you show that

$$C^* \leq C_{Huffman} \leq C^* + \sum_{i=1}^{m} p_i c_i ?$$

Solutions 2a

a) We wish to minimize $C = \sum p_i c_i n_i$ subject to $\sum 2^{-n_i} \leq 1$. We will assume equality in the constraint and let $r_i = 2^{-n_i}$ and let $Q = \sum_i p_i c_i$. Let $q_i = (p_i c_i)/Q$. Then $\mathbf{q}$ also forms a probability distribution and we can write C as

$$\begin{aligned} C &= \sum p_i c_i n_i && (403)\\ &= Q \sum q_i \log \frac{1}{r_i} && (404)\\ &= Q \left(\sum q_i \log \frac{q_i}{r_i} - \sum q_i \log q_i \right) && (405)\\ &= Q(D(\mathbf{q}||\mathbf{r}) + H(\mathbf{q})). && (406) \end{aligned}$$

Since the only freedom is in the choice of r_i, we can minimize C by choosing $\mathbf{r} = \mathbf{q}$ or

$$n_i^* = -\log \frac{p_i c_i}{\sum p_j c_j}, \quad (407)$$

where we have ignored any integer constraints on n_i. The minimum cost C^* for this assignment of codewords is

$$C^* = QH(\mathbf{q}) \quad (408)$$

Solutions 2b

b) If we use $\mathbf{q}$ instead of $\mathbf{p}$ for the Huffman procedure, we obtain a code minimizing expected cost.

Solutions 2c

c) Now we can account for the integer constraints.
Let

$$n_i = \lceil -\log q_i \rceil \tag{409}$$

Then

$$-\log q_i \leq n_i < -\log q_i + 1 \tag{410}$$

Multiplying by $p_i c_i$ and summing over i, we get the relationship

$$C^* \leq C_{Huffman} < C^* + Q. \tag{411}$$