2018 Fall Data Compression Homework #1

EE 248583 **106368002 張昌祺 Justin, Chang-Qi Zhang**
Advisor: 電子所 高立人 副教授 justin840727@gmail.com
Due Date: November 12 2018

Problem 1 Entropy

Let X be a random variable with an alphabet $H = \{1, 2, 3, 4, 5\}$. Please determine $H(X)$ for the following three cases of probability mass function $p(i) = prob[X = i]$. (15%)

(a) $P(1) = P(2) = 1/2$:

Ans

$$\begin{aligned} H(X) &= -(P(1)\log_2 P(1) + P(2)\log_2 P(2)) \\ &= -(0.5\log_2(0.5) + 0.5\log_2(0.5)) \\ &= -(-0.5 - 0.5) \\ &= 1\ bits/symbol \end{aligned}$$

(b) $P(i) = 1/4, for\ i = 1, 2, 3, and\ p(4) = p(5) = 1/8$:

Ans

$$\begin{aligned} H(X) &= -(3 \times P(1)\log_2 P(1) + P(4)\log_2 P(4) + P(5)\log_2 P(5)) \\ &= -(3 \times 0.25\log_2(0.25) + 2 \times 0.125\log_2(0.125)) \\ &= -(-1.5 - 0.75) \\ &= 2.25\ bits/symbol \end{aligned}$$

(c) $P(i) = 2^{-i}, for\ i = 1, 2, 3, 4, and\ p(5) = 1/16$:

Ans

$$\begin{aligned} H(X) &= -(\sum_{i=1}^{4} 2^{-i}\log_2 2^{-i} + \frac{1}{16}\log_2\frac{1}{16}) \\ &= -(0.5 \times (-1) + 0.25 \times (-2) + 0.125 \times (-3) + 0.0625 \times (-4) + 0.0625 \times (-4)) \\ &= 1.875\ bits/symbol \end{aligned}$$

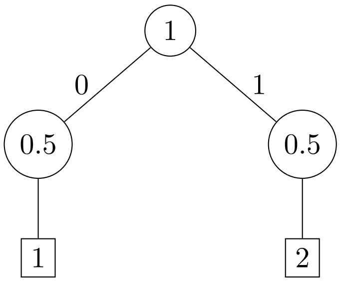
Problem 2 Huffman Code

Design a Huffman code C for the source in Problem 1. (15%)

(a) Specify your codewords for individual pmf model in Problem 1.

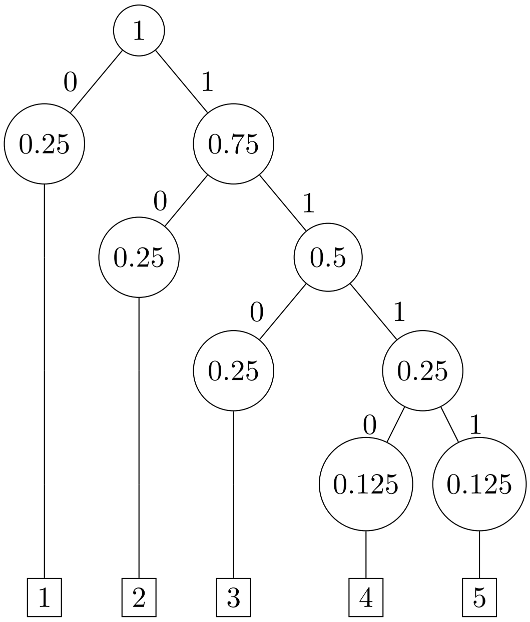
Ans

1.(a)



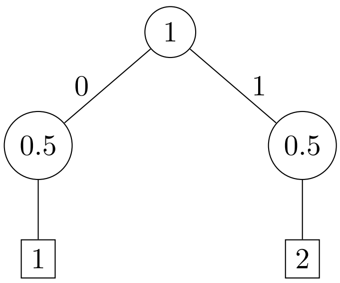
Alphabet	P	Codeword
1	0.5	0
2	0.5	1

1.(b)



Alphabet	P	Codeword
1	0.25	0
2	0.25	10
3	0.25	1110
4	0.125	11110
5	0.125	11111

1.(c)



Alphabet	P	Codeword
1	0.5	0
2	0.25	1
3	0.125	1
4	0.0625	1
5	0.0625	1

(b) Compute the expected codeword length and compare with the entropy for your codes in (a).

Ans

(c) Design a code with minimum codeword length variance for the pmf model in Problem 1.(b)

Ans

Problem 3 Empirical Distribution C++

Empirical distribution. In the case a probability model is not known, it can be estimated from empirical data. Let's say the alphabet is $H = \{1, 2, 3, \dots, m\}$. Given a set of observations of length N , the empirical distribution is given by $p = total\ number\ of\ symbol\ 1/N,\ for\ i = 1, 2, 3, ..., m$. Please determine the empirical distribution for **santaclaus.txt**, which is an ASCII file with only lower-cased English letters (i.e., $a \sim z$), space and CR (carriage return), totally 28 symbols. The file can be found on the class web site. Compute the entropy. (14%)

Ans

Problem 4 Huffman Code Encode C++

Write a program that designs a Huffman code for the given distribution in Problem 3. (14

Ans

Problem 5 Adaptive Huffman Tree

Let X be a random variable with an alphabet H , i.e., the 26 lower-case letters. Use adaptive Huffman tree to find the binary code for the sequence **a a b b a**. (24%)
You are asked to use the following 5bits fixed-length binary code as the initial codewords for the 26 letters. That is
a: 00000
b: 00001
⋮
z: 11001
Note: Show the Huffman tree during your coding process.

Ans

Problem 6 Golomb Encoding and Decoding.

(a) Find the Golomb code of n=21 when m=4.

Ans

(b) Find the Golomb code of n=14 when m=4.

Ans

(c) Find the Golomb code of n=21 when m=5.

Ans

(d) Find the Golomb code of n=14 when m=5.

Ans

(e) A two-integer sequence is encoded by Golomb code with m=4 to get the bitstream 11101111000. What's the decoded two-integer sequence?

Ans

(f) A two-integer sequence is encoded by Golomb code with m=5 to get the bitstream 11101111000 (the same bitstream as that in (e)). What's the decoded two-integer sequence?

Hint: The unary code for a positive integer q is simply q 1s followed by a 0.

Ans