Solutions to Written Exam in

Data compression

TSBK08

20th March 2023

1 a) See the course literature.

b) See the course literature.

c) See the course literature.

2 a) See the course literature.

b) See the course literature.

c) See the course literature.

3 See the course literature.

4 Probabilities $p(x_i, x_{i+1}) = p(x_i) \cdot p(x_{i+1})$ for pairs of symbols:

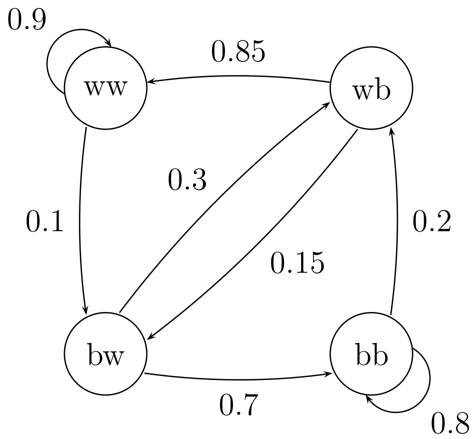
$$p(a,a) = 0.36, \quad p(a,b) = 0.21, \quad p(a,c) = 0.03$$

$$p(b,a) = 0.21, \quad p(b,b) = 0.1225, \quad p(b,c) = 0.0175$$

$$p(c,a) = 0.03, \quad p(c,b) = 0.0175, \quad p(c,c) = 0.0025$$

A Huffman code for this distribution gives the mean codeword length $\bar{l} = 2.435$ bits/codeword and average data rate $R = \frac{\bar{l}}{2} = 1.2175$ bits/symbol.

5 a) Given states (x_i, x_{i-1}), the state diagram looks like



The stationary probabilities for this model is

$$w_{ww} = \frac{17}{28},\ w_{wb} = \frac{2}{28},\ w_{bw} = \frac{2}{28},\ w_{bb} = \frac{7}{28}$$

These probabilities are also probabilities for pairs $p(x_i, x_{i-1})$.

b) The pair probabilities from above gives us the entropy $H(X_i, X_{i-1}) \approx 1.4810$. Probabilties for single symbols can be found as marginal probabilities

$$p(w) = p(w,w) + p(w,b) = \frac{19}{28},\quad p(b) = p(b,w) + p(b,b) = \frac{9}{28}$$

which gives us the entropy $H(X_i) \approx 0.9059$. Using the chain rule, we find $H(X_i|X_{i-1}) = H(X_i, X_{i-1}) - H(X_{i-1}) \approx 0.5751$. Finally, we need probabilities for three symbols

$$p(x_i, x_{i-1}, x_{i-2}) = p(x_{i-1}, x_{i-2}) \cdot p(x_i|x_{i-1}, x_{i-2})$$

$$\begin{aligned} p(w,w,w) &= \frac{153}{280}, \quad p(b,w,w) = \frac{17}{280} \\ p(w,w,b) &= \frac{17}{280}, \quad p(b,w,b) = \frac{3}{280} \\ p(w,b,w) &= \frac{6}{280}, \quad p(b,b,w) = \frac{14}{280} \\ p(w,b,b) &= \frac{14}{280}, \quad p(b,b,b) = \frac{56}{280} \end{aligned}$$

This gives us the entropy $H(X_i, X_{i-1}, X_{i-2}) \approx 2.0527$.
Using the chain rule we find
$H(X_i|X_{i-1}, X_{i-2}) = H(X_i, X_{i-1}, X_{i-2}) - H(X_{i-1}, X_{i-2}) \approx 0.5717$.

6 Assuming that we always place the b interval closest to 0, the sequence corresponds to the interval $[0.311824 \;\; 0.32203)$. The interval size is 0.010206 and thus we will need at least $\lceil -\log_2 0.010206 \rceil = 7$ bits in our codeword, maybe one more.

Write the two interval limits as binary numbers:

$$\begin{aligned} 0.311824 &= 0.010011111101001\ldots \\ 0.32203 &= 0.010100100111000\ldots \end{aligned}$$

The smallest seven bit number inside the interval is 0.0101000, and all numbers starting with these bits are also inside the interval (ie smaller than the upper interval limit). Thus, seven bits are enough.

The codeword is **0101000**.

7 The decoded sequence is

$$bedbadbedededebabebabbabb\ldots$$

and the dictionary looks like

index	word	index	word	index	word	index	word
0	a	5	be	10	dbe	15	$babb$
1	b	6	ed	11	ede	16	$babb*$
2	c	7	db	12	$edeb$	17	
3	d	8	ba	13	bab	18	
4	e	9	ad	14	beb	19	

where we don't know * until we have decoded the next index.

8 Inverse mtf gives the vector $L = [c\ c\ c\ a\ a\ b\ b\ b]$.

Sort the sequence to get the vector $F = [a\ a\ b\ b\ b\ c\ c\ c]$ which gives us the vector $T = [3\ 4\ 5\ 6\ 7\ 0\ 1\ 2]$ (the position in L where you find each symbol in F).

Inverse BWT gives the sequence *cabcbcab*.