Information Theoretic Modeling – Exercise 4

Haibo Jin
Student number: 014343698

1 Problem 1

(a)

The probabilities of symbols $\{a, b, c, !\}$ is given as $p(a)$=0.05, $p(b)$=0.5, $p(c)$=0.35, $p(!)$=0.1, we can create a table illustrating the first iteration of arithmetic coding, which is also known as Shannon-Fano-Elias coding. Please see Table 1 for more details.

x	$p(x)$	$F(x)$	$I(x)$
a	0.05	0.05	[0, 0.05)
b	0.5	0.55	[0.05, 0.55)
c	0.35	0.9	[0.55, 0.9)
!	0.1	1	[0.9, 1)

Table 1: First iteration of arithmetic coding, $p(x)$ is symbol distribution, $F(x)$ is cumulative function and $I(x)$ is the interval for symbol x

The target message is $cab!$, so we choose $I(c)$ from the first iteration and will be further split in the next iteration. So now we have $I(c) = [0.55, 0.9)$.

The second symbol is a, we can use similar method to allocate an interval for a. But the total range of the interval is no longer $[0, 1]$ but $I(c)$. Then the probability of a is $(0.9 - 0.55) \times 0.05 = 0.0175$. So $I(ca) = [0.55, 0.5675)$.

The third symbol is b. We know its probability is between $(0.5675 - 0.55) \times 0.05 = 0.000875$ and $(0.5675 - 0.55) \times 0.55 = 0.009625$. So $I(cab) = [0.550875, 0.559625)$.

The last symbol is ! and we can get its proportion of probability is $(0.559625 - 0.550875) \times 0.9 = 0.007875$. Finally, the interval $I(cab!)$ will be **[0.55875, 0.559625)**.

(b)

Based on the previous calculation, $I(cab!) = [0.55875, 0.559625)$

i.

The shortest codeword within the interval is **0.559**.

ii.

The codewords that satisfy the condition are **0.5588**,**0.5589**,**0.5590**,**0.5591**,**0.5592**,**0.5593**,**0.5594**,**0.5595**.

(c)

Since the symbol probabilities are slightly changed, we need to calculate the interval of *cab*! again and the method is basically the same as previous.

The new symbol probabilities are $p(a) = 2/32$, $p(b) = 16/32$, $p(c) = 11/32$, $p(!) = 3/32$.

We can get $I(c) = [18/32, 29/32)$, $I(ca) = [288/512, 299/512)$, $I(cab) = [4619/8192, 4707/8192)$, $I(cab!) = [18795/32768, 4707/8192)$.

By transforming the real numbers to binaries, $I(cab!)$ can be rephrased as $[0.100100101101011, 0.1001001100011)$.

Now we can answer that shortest codeword within the interval is **0.10010011**.

The shortest codeword that all its continuations are also within the interval is **0.10010010111**.

2 Problem 2

$D = 0011$

The two-part code-length can be represented as $L = l_1(\theta) + log_2 \frac{1}{p_\theta(D)}$.

For $\theta = 0.25$, $p_{0.25}(0011) = 0.25^2 \times 0.75^2 = 0.03515625$
$l(0.25) = 2$
$L = 2 + 4.83 = 6.83$

For $\theta = 0.5$, $p_{0.5}(0011) = 0.5^2 \times 0.5^2 = 0.0625$
$l(0.25) = 1$
$L = 1 + 4 = 5$

For $\theta = 0.75$, $p_{0.25}(0011) = 0.75^2 \times 0.25^2 = 0.03515625$
$l(0.25) = 2$
$L = 2 + 4.83 = 6.83$

So, the expected two-part code-length of sequence 0011 is $E(L) = (6.83 \times 2 + 5)/3 =$ **6.22**

3 Problem 3

Since the parameter prior is uniform, so $\omega(\theta) = 1$, then we can get

$p(D) = \int_\Theta p_\theta(D)\omega(\theta)d\theta = \int_\Theta p_\theta(D)d_\theta = \int_\Theta \theta^k(1-\theta)^{n-k}d_\theta$

Use Beta-Binomial distribution, $p(D)$ can be calculated easily since we can control the value of θ by controlling the parameter α and β in Beta-Binomial distribution. In our case, the target sequence is 0011, which is just a discrete point in Beta-Binomial distribution with $\alpha = 1$, $\beta = 1$, $n = 4$ and $k = 2$. So the probability of this point is

$pmf = \frac{B(k+\alpha,n-k+\beta)}{B(\alpha,\beta)} = \frac{B(2+1,4-2+1)}{B(1,1)} = \frac{B(3,3)}{B(1,1)} = 0.0333$

So the mixture code-length of 0011 is $-log_2 p(D) = \mathbf{4.9069}$.

4 Problem 4

(a)

$C = \binom{4}{0} \times 1 + \binom{4}{1} \times \frac{1}{4} \times \frac{3}{4}^3 + \binom{4}{2} \times \frac{2}{4} \times \frac{2}{4}^2 + \binom{4}{3} \times \frac{3}{4}^3 \times \frac{1}{4} + \binom{4}{4} \times 1$

$= 1 + \frac{27}{64} + \frac{3}{8} + \frac{27}{64} + 1$

$= \frac{103}{32} = \mathbf{3.21875}$

(b)

The maximum likelihood model is $p_{\hat{\theta}}(0011) = \frac{1}{2}^2 \times \frac{1}{2}^2 = \frac{1}{16}$

So the normalized maximum likelihood model is $p_{nml}(0011) = \frac{p_{\hat{\theta}}(0011)}{C} = \frac{1}{16} \times \frac{32}{103} = 0.0194$.

Then we can get the NML code-length as $L = -log_2 p_{nml}(0011) = \mathbf{5.6865}$