Problem Set #1: A Bayesian Approach to Data Compression

Joe Suzuki

July 17, 2017

Problem (Corrected)

1. Let $A = \{\alpha, \beta\}$, $P(X = \alpha)\frac{1}{3}$ and $P(X = \beta) = \frac{2}{3}$. Calculate the entropy $H(X)$. Compute the codelength of each sequence and the average codelength for $n = 1, 2, 3$.

2. Let θ be unknown and $a = b = 0.5$. Calculate the codelength of each sequence of length $n = 4$. Assume that the length is $\lceil -\log Q(x^n) \rceil$ when the assigned probability of x^n is $Q(x^n)$.

3. Let $x \sim y$ denote that $x - y$ is bounded by above and below, and Use Stirling's formula
$$\log \Gamma(z) \sim -z + (z - \frac{1}{2}) \log z$$
to derive each $\sim$ below.
$$\log(n+1) \sim \log n$$
$$c \log\{c + \frac{1}{2}\} \sim c \log c$$
$$\log \Gamma(c + \frac{1}{2}) \sim -(c + \frac{1}{2}) + c \log(c + \frac{1}{2}) \sim -c + c \log c$$
$$\log \Gamma(n+1) \sim -(n+1) + (n + \frac{1}{2}) \log(n+1) \sim -n + (n + \frac{1}{2}) \log n$$
$$-\log Q(x^n) \sim -c \log \frac{c}{n} - (n-c) \log \frac{n-c}{n} + \frac{1}{2} \log n$$
for
$$Q(x^n) = \frac{1}{\Gamma(n+1)} \cdot \frac{\Gamma(c+1/2)}{\Gamma(1/2)} \cdot \frac{\Gamma(n-c+1/2)}{\Gamma(1/2)} \ .$$

Hints

1. The entropy can be computed as $-\frac{1}{3}\log_2 \frac{1}{3} - \frac{2}{3}\log_2 \frac{2}{3}$. The code lengths for $n = 1$ and $n = 2$ can be computed as

0	$\lceil -\log_2 \frac{1}{3} \rceil = 2$
1	$\lceil -\log_2 \frac{2}{3} \rceil = 1$

00	$\lceil -\log_2 \frac{1}{3^2} \rceil = 4$
01	$\lceil -\log_2 \frac{2}{3^2} \rceil = 3$
10	$\lceil -\log_2 \frac{2}{3^2} \rceil = 3$
11	$\lceil -\log_2 \frac{2^2}{3^2} \rceil = 2$

Generate the table for $n = 3$.

The average code lengths can be computed as follows. Calculate the concrete values.
$$\frac{1}{3}\lceil -\log_2 \frac{1}{3} \rceil + \frac{2}{3}\lceil -\log_2 \frac{2}{3} \rceil$$

$$\frac{1}{3^2}\lceil -\log_2 \frac{1}{3^2}\rceil + 2\cdot\frac{2}{3^2}\lceil -\log_2 \frac{2}{3^2}\rceil + \frac{2^2}{3^2}\lceil -\log_2 \frac{2^2}{3^2}\rceil$$

$$\frac{1}{3^3}\lceil -\log_2 \frac{1}{3^3}\rceil + 3\cdot\frac{2}{3^3}\lceil -\log_2 \frac{2}{3^2}\rceil + 3\cdot\frac{2^2}{3^3}\lceil -\log_2 \frac{2^2}{3^2}\rceil + \frac{2^3}{3^3}\lceil -\log_2 \frac{2^3}{3^3}\rceil$$

2. For $x^4 = (0,0,0,0)$, $\frac{0+1/2}{0+1}\cdot\frac{1+1/2}{1+1}\cdot\frac{2+1/2}{2+1}\cdot\frac{3+1/2}{3+1} = \frac{35}{2^7}$, so that the length is $\lceil -\log_2 \frac{35}{2^7}\rceil = 2$.
   For $x^4 = (0,0,0,1)$, $\frac{0+1/2}{0+1}\cdot\frac{1+1/2}{1+1}\cdot\frac{2+1/2}{2+1}\cdot\frac{0+1/2}{3+1} = \frac{5}{2^7}$, so that the length is $\lceil -\log_2 \frac{5}{2^7}\rceil = 5$.
   Compute the values for the other sequences. You do not have to compute J_n but compare the values like $\frac{Q(x^n, y^n)}{Q(x^n)Q(y^n)}$

3. From $\log(z+1) \le z$ for $z \ge 0$, we have

$$\log n \le \log(n+1) = \log n + \log(1+\frac{1}{n}) \le \log n + \frac{1}{n}$$

$$c\log(c+\frac{1}{2}) \le c\log c + c\log(1+\frac{1}{2c}) \le c\log c + c\cdot\frac{1}{2c} = c\log c + \frac{1}{2}$$

$$0 \le (n+\frac{1}{2})\log(1+\frac{1}{n}) \le (n+\frac{1}{2})\cdot\frac{1}{n} \le 1 + \frac{1}{2n}$$

   From $\log\Gamma(z) \sim -z + (z-1/2)\log z$ for $z > 0$, we have

$$\log\Gamma(c+\frac{1}{2}) \sim -(c+\frac{1}{2}) + c\log(c+\frac{1}{2})$$

$$\log\Gamma(n+1) \sim -(n+1) + (n+\frac{1}{2})\log(n+1) \sim -(n+1) + (n+\frac{1}{2})\log n + (n+\frac{1}{2})\log(1+\frac{1}{n})$$

Based on those equations, complete the proof.