

プログラミング演習 1 LaTeX 練習

自分の氏名

自分の学籍番号

1 はじめに

本レポートでは、第2章で情報量の定義を、第3章で符号化、特にハフマンの符号化について説明している。

2 情報量

事象 A の起こる確かさを示す量として、 A の生起確率 $P(A)$ をとる。当然、 $0 \leq P \leq 1$ である。もし、事象 A と B が独立であれば、 A と B がともに起こる確率は

$$P(A \cap B) = P(A)P(B)$$

で表される。このことを考慮して、事象 A が起こったことを知らされたとき、われわれが受けとる情報量を

$$I(A) = -\log_2 P(A)$$

で定義してみることにする。すると、 A が起こる確率が小さいほど、 $I(A)$ という情報量は大きくなる。また、独立事象 A, B について、それらが共に起こった場合に受けとる情報量は、

$$\begin{aligned} I(A \cap B) &= -\log_2 P(A \cap B) \\ &= -\log_2 [P(A)P(B)] \\ &= -\log_2 P(A) - \log_2 P(B) \\ &= I(A) + I(B) \end{aligned} \quad (1)$$

が成り立ち、情報量の持つ性質が満たされていることが分かる。また、事象 a_k ($k = 1, \dots, m$) の発生確率を $P(a_k)$ とすると、これらの事象における平均情報量 H は (2) 式で求められる。

$$H = \sum_{k=1}^m P(a_k) I(a_k) = -\sum_{k=1}^m P(a_k) \log_2 P(a_k) \quad (2)$$

3 情報符号化

情報の符号化には、大別して2つの方法がある。一つは、出現頻度の高いものに短い符号を割り当て、逆に頻度の低いものに長い符号を割り当てて、全体の情報量をできるだけ小さくする方法である。この場合、文字ごとに符号の長さが違うので計算機では取り扱いにくい。この方式を可変長符号化という。もう一つは、出現頻度に関係なしに同じ長さの符号を対応させる方式であり、固定長符号化という。

パターンの利用頻度の偏りを利用する符号化の代表例として、ハフマン (Huffman) の最適符号化法を挙げることができる。いま、例として A, B, C, D という4文字の符号化を考える。固定長ビットで符号化すれば、1文字あたり2ビット必要である。いま、各文字の出現頻度が $A:B:C:D=60:25:10:5$ であるとすると、この情報源の平均情報量は (3) 式で求められる。

$$\begin{aligned} H &= -0.60 \log_2 0.60 - 0.25 \log_2 0.25 - 0.10 \log_2 0.10 \\ &\quad - 0.05 \log_2 0.05 \\ &= 1.491 \text{ (ビット)} \end{aligned} \quad (3)$$

可変長符号を用いると、平均符号長を H または H に近い値にまで短くすることができるが、 H より短くすることは不可能であることが証明されている。この例では、 A の頻度が最大であるから、これに一桁の符号 '0' をあて、これと区別するために他の文字の符号は '1' から始めることにする。2番目に出現頻度の高い B に '10' という符号を与えると、残りの2文字は先頭の2桁を '11' にとることにより区別できるので、 C に '110'、 D に '111' という符号を与える。このようにハフマンの符号化を用いたとき、この符号の平均符号長 l は次式で与えられる。

$$l = 0.60 \times 1 + 0.25 \times 2 + 0.10 \times 3 + 0.05 \times 3 = 1.55$$

上式より、ハフマンの符号化を用いると固定長符号化時の2ビットよりも確かに短くなっていることが分かる。