目次

★3　パターン情報の表現 (2) — データ圧縮と情報量（承前）

「承前」って？ ⇒ 辞書引きましょう

★ 3.6　情報量の概念（承前）

★ 3.6.3　平均情報量（エントロピー）

n 個の独立な事象 $E_1, E_2, \ldots, E_n$ がそれぞれ確率 $p_1, p_2, \ldots, p_n$ で生起する場合を考える．$\sum_{i=1}^{n} p_i = 1$ とする．このとき，事象 E_i が「起こった」という知らせの情報量 $I(E_i)$ は $I(E_i) = -\log p_i$ である．それでは，「$E_1, E_2, \ldots, E_n$ のどれが起こったかまだ知らない状況でどれが起こったかを知らせてもらう場合，その知らせは平均としてどれだけの情報量をもつと期待できるか」を考えてみよう．その値を H とおくと，H は得られる情報量の期待値であるから，

$$H = \sum_{i=1}^{n} p_i I(E_i) = -\sum_{i=1}^{n} p_i \log p_i \qquad (1)$$

ただし，$\lim_{p\to+0} p\log p = 0$ より，$p = 0$ の場合の $p\log p$ の値は 0 として扱う．

となる．これは，「どれが起こったか知らない不確定な状況を確定させることで得られる情報量の平均値」を表しており，**平均情報量**あるいは**エントロピー**と呼ばれる．「知ろうとしている状況の不確かさの度合い」を表していると考えてもよい．

エントロピー: entropy. 熱力学でででてくるエントロピーと同じようなものと考えられる．

エントロピーについてはいろいろ興味深い話があるが，この授業では割愛する．

例: 確率 p で表が，確率 $1-p$ で裏が出るコインがある．このコインを投げたときに得られるエントロピーは，(底を 2 とすると)

$$H = p \times I(\text{表}) + (1-p) \times I(\text{裏}) = -p\log_2 p - (1-p)\log_2(1-p) \quad [\text{bit}] \qquad (2)$$

となる．右図からわかるように，このエントロピーは $p = 0$ または $p = 1$ のときに 0（投げる前から結果がわかっている）となり，$p = \frac{1}{2}$ のときに最大 (1[bit]) となる（最も不確かな状況）．これは，「確率に偏りがある ⇔ エントロピーが小さい ⇔ 不確かさが小さい」ということを意味している．

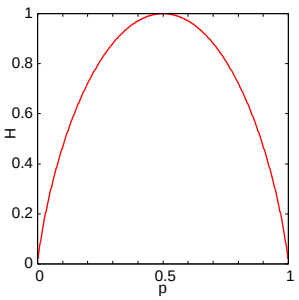

図: コイン投げのエントロピー

Q1. 「ほげおくんが 'H' という文字を書いた」という事象を E_1 と表すことにする．同様に，'O','G','E' を書いたという事象をそれぞれ E_2, E_3, E_4 と表記する．これら 4 つの事象 H,O,G,E がそれぞれ独立に確率 $\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}$ で生起する場合を考える (☆ 1)．次の問に答えなさい．

(1) 事象 E_i が起こったという知らせの情報量 $I(E_i)$ を求めなさい $(i = 1, 2, 3, 4)$．
(2) 平均情報量を求めなさい．
(3) これら 4 つの事象をビットパターンに対応させて区別したい．すべて同じ bit 数で表すなら，1 つの事象を何 bit のビットパターンに対応させればよいか．
(4) ほげおくんが 'H' ばかり書くようになった（(E_1の生起確率) $\to 1$）ら，平均情報量はどうなるか．

☆ 1) $\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{8} = 1$ ですね．ほげおくんが他の文字を書くことはないらしい...

★ 3.7　エントロピー符号化

前節の Q の条件で，ほげおくんが書いた文字の並び（‘H’,‘O’,‘G’,‘E’ の 4 種のみでできている）を単純に 1 文字あたり 2bit で符号化するなら，この文字の並びのデータ量は当然 1 文字あたり 2bit となる．しかし，各文字の出現頻度に偏りがあるせいで，彼の書いた文字の並びの平均情報量（エントロピー）は 2bit よりも小さくなっていた．何か工夫することで，1 文字あたり 2bit より少ないデータ量で符号化できる方法があるんではないだろうか？

実は，データ圧縮の手法として，エントロピーの概念に基づく**エントロピー符号化**という符号化法が知られている．

★ 3.7.1　エントロピー符号化の考え方

表に示す 6 つの記号が，**それらを並べた記号列中の並び方や位置に無関係に**(☆ 2)，表に示す確率で出現するとする．この記号列を 0,1 で符号化したい．

☆ 2) H のあとに O が出てきやすいとか，最初の方は H が出てきやすい，というようなことがない，という意味．

記号	D	E	F	G	H	O
確率	$\frac{10}{100}$	$\frac{2}{100}$	$\frac{15}{100}$	$\frac{13}{100}$	$\frac{20}{100}$	$\frac{40}{100}$

最も単純な方法は，各記号を 3bit の符号で表現する（記号 D に 000，E に 001，等），というものである．この場合，たとえば `HOGEHOHOOO` という 10 文字の並びは，$10 \times 3 = 30$bit のビット列となる．この例では，どの記号にも 3bit を割り当てることになるから，「1 つの記号を平均何 bit で表せるか」を表す**平均符号語長**を考えると，その値は 3bit となる．

しかし，これらの記号の出現頻度には偏りがあるので，平均情報量はもっと小さいと考えられる (☆ 3)．この例の場合に実際に計算してみると，

$$H = -\frac{10}{100}\log_2\frac{10}{100} - \frac{2}{100}\log_2\frac{2}{100} - \cdots - \frac{40}{100}\log_2\frac{40}{100} = \text{約 } 2.23[\text{bit}] \quad (3)$$

となる．このことは，符号の割り当て方を工夫すれば，平均符号語長をここまで小さくできる可能性があることを意味している．たとえば，右表のように出現確率の高いものに短い符号を割り当てるようにしてやると，平均符号語長を

$$\frac{10}{100}\times 4 + \frac{2}{100}\times 4 + \frac{15}{100}\times 3 + \frac{13}{100}\times 3 + \frac{20}{100}\times 3 + \frac{40}{100}\times 1 = 2.32[\text{bit}] \quad (4)$$

にすることができる．実際に上記の 10 文字の並びの例でやってみると，

$$11101011000011101111000 \quad (5)$$

となり，30bit だったものを 21bit に減らせている．

このように，エントロピー符号化とは，記号の出現頻度（確率）に偏りがある，すなわち平均情報量が小さい場合に，その偏りを利用して効率的な（平均符号語長の短い）符号化を実現しようとするものである．上記の例では平均情報量が約 2.23bit であるから，理想的なエントロピー符号化ができれば平均符号長をそこまで減らせることを意味している．

☆ 3) ちなみに，6 通りの記号が等確率で出現する場合でも平均情報量は $-6\times\frac{1}{6}\log_2\frac{1}{6} = \log_2 6 = $ 約 2.58[bit] である．

記号	符号
D	1001
E	1000
F	110
G	101
H	111
O	0

Q2. 式 (5) の符号を表に従って先頭から順に復号していくとちゃんと元通りになることを確認してみよう．

★ 3.7.2 ハフマン符号化

エントロピー符号化の一種である**ハフマン符号化**は，記号毎に算出した出現頻度を用いて，次のような手順で木（ハフマン木）を作ることで，それぞれの記号に割り当てる符号を求める（授業中にもう少し詳しく説明します）．

ハフマン符号化: Huffman coding. 1950 年代前半に D. A. Huffman が提案した符号化法．画像圧縮方式 JPEG でも，離散コサイン変換（そのうち紹介するかも）したデータの符号化に用いられている．

1. 個々の記号に対応した葉節点をつくる．出現確率をそれぞれの値とする．
2. 親のいない節点の中で最も小さい値をもつもの 2 つを選び，それらの親となる節点をつくる．親節点の値は，2 つの子節点の値の和とする．
3. 親を持たない節点が一つになるまで 2. を繰り返す．
4. 「左」を 0,「右」を 1 として，できた木の各節点から左右に接続した枝のそれぞれに 0,1 を割り当てる（「左」を 1 としてもよい）．
5. 根から葉までたどる際に現れる 0,1 のならびをそれぞれの葉の記号に対応した符号とする．

ハフマン符号化は純粋なエントロピー符号化法であり，記号の出現頻度のみを考慮して符号を決めるアルゴリズムである．したがって，出現頻度に偏りがないと有効な圧縮ができない (☆ 4)．また，特定の記号が何度も繰り返すとか，特定の並び（HOGE みたいな）が何度も出てくるとか，ある記号のあとに特定の記号が続くことが多い（G のあとには E がよく出てくるみたいな）というような記号の並び方は，データ圧縮の性能に影響しない（それを活かした圧縮はできない）．

☆ 4) 復号の際にハフマン木を使うので，データにハフマン木の情報も加える必要がある．そのため，極端な場合には元よりデータ量が増えてしまうこともありえる．

Q3. 上の（授業中に説明した）Huffman 木を用いて，次のデータを復号しなさい．1011000111010101110

Q4. HOGEHOGEHOGEHOGE... とひたすら繰り返すだけのテキストデータがあったとする．このデータにハフマン符号化を適用しても，有効なデータ圧縮はできない．その理由を述べなさい．

★ 3.7.3 よだんだよん

- エントロピー符号化の方法としては，Huffman 符号化よりも高効率な算術符号化という方法もある．
- 符号化は何もデータ圧縮のためだけに行なうものではない．雑音等のせいで送信データに誤りが生じてしまう状況で誤りを見つける（誤り検出）／誤りを訂正する（誤り訂正）ためにも用いられる (☆ 5)．最も原始的な誤り検出の符号化法は，ビット列の最後に，ビット列中の 1 の数が常に偶数になるように 0 か 1 を付け加える，というものである．興味のある人はいろいろ調べてみるとおもしろいかも．
- 符号化のさらに別の目的として，暗号化がある．

☆ 5) 情報通信機器では当たり前に使われている．図書の ISBN，様々な商品の JAN コード（バーコードの元になってるコード）などにも使われている．

情報理論の参考書: 高橋の手元には「情報理論」甘利俊一，ダイヤモンド社，ISBN4-478-82000-7，といういい本があるのですが，残念ながら現在は手に入らないようです (☆ 6)．他にいい本ないか探索中．
データ圧縮の参考書: 「圧縮アルゴリズム 符号化の原理と C 言語による実装」昌達 K'z，ソフトバンクパブリッシング，ISBN4-7973-2552-6

☆ 6) 最近，ちくま学芸文庫で復活しました．

宿題

ある記号列を調べたところ，6 通りの記号が次表のような頻度で出現することがわかった．このような記号列をハフマン符号化法でビット列に符号化するとして，次の問に答えなさい（答えのみでよい）．

記号	あ	か	0	2	3	4
頻度	0.45	0.16	0.25	0.02	0.04	0.08

(1)–(6) 上記の表にもとづいてハフマン木を求めると，右図のようになった．(1) から (6) までの各葉節点に割り当てられる記号を答えなさい．ただし，このハフマン木は，授業中に説明したものと同様に，子をもつ全ての節点について，(左の子の値) < (右の子の値) となるように描かれているものとする．

(7) 右図のハフマン木の子をもつ全ての節点について，**左の子にむかってのびた枝に符号 1** を，右の子にむかってのびた枝に符号 0 を割り当てたとする．このハフマン木を用いて，次のビット列を記号列に復号しなさい．

1101001110000010010

(8) ハフマン符号化法はデータ圧縮のためによく用いられるが，データの性質によっては有効でない（圧縮率が高くならない）ことがある．それはどのような場合か，理由も含めて述べなさい．