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

★ 3.4　データ圧縮の例—ランレングス圧縮

ある地域の天気を 1 日毎に「晴」「曇」「雨」「雪」「霙」(☆ 1) などの 8 種類に分類したデータがあったとしよう．例えば，

晴 晴 晴 晴 晴 晴 晴 曇 曇 雨

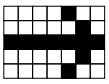
といったものである．このように同じ値が連続して現れることの多いデータの場合，値とその連続する数をペアにして

晴 7 曇 2 雨 1

のように表すと，データ量を減らせそうである．このようなアイデアに基づくデータ圧縮の手法を，**ランレングス圧縮**（ランレングス符号化）という．

「減らせそう」でごまかさないで具体的に考えてみよう．この例では，1 日の天気の値を 3bit で表現することができる．その場合，元のデータを表すためのデータ量は $3 \times 10 = 30$[bit] となる．一方，ランレングス圧縮したデータの方は，値だけでなくその連長も符号化する必要がある．例えば長さも 3bit で表現するとしたなら (☆ 2)，データ量は $(3+3) \times 3 = 18$[bit] ということになる．したがって，この例ではランレングス圧縮を行なうことでデータ量を元の 6 割にまで削減できている．

ランレングス圧縮は，**二値画像**（白と黒の二通りの画素で構成された画像）を扱うのに特に適しているので，ファクシミリ等に応用されている（例えば文書などを読み取って二値画像にすると，白画素ばかりになるから）．例えば右図の画像 の画素値を左上から右に向かって順に符号化していくと「白 4 黒 1 白 7 黒 1 白 1 黒 7 白 5 黒 1 白 5 黒 1 白 2」のように表せるが，二値かつ白から数えると仮定すれば，さらに省略して「4 1 7 1 1 7 5 1 5 1 2」としてもちゃんと復号できる．



上述のことからわかるように，ランレングス圧縮のアルゴリズムは非常に単純である (☆ 3) しかし，単純なランレングス圧縮の方法には次のような問題点がある

- 同じ値があまり続かない場合には逆にデータ量を増やしてしまう
- 同じ値が長く続くからといってその長さをそのまま符号化すると，長さを表すために大きな bit 数を割り当てる必要が生じて圧縮率が落ちてしまう

そのため，実用的にはもう少し凝ったアルゴリズムを考える必要がある (☆ 4)．

Q1. 以下は，7×7 の格子状に並んだ画素の値（白か黒）をランレングス圧縮して得られるデータである．ただし，左上の画素から右に向かって符号化してある（最初は白の数）とする．これを伸長するとどんなパターンが得られるか．(^_-)

815111137151158

☆ 1) 霙：みぞれと読むそうです.

ランレングス: run-length. 連長ともいう.

☆ 2) 3bit なら 1 から 8 までの長さを表せる.

☆ 3) プログラミングの好きな人は，例えば，整数のならびをキーボードから入力するとそれをランレングス圧縮したものを表示するプログラムを考えてみると楽しいでしょう．手頃な練習問題になります.

☆ 4) 興味のある人は，データ圧縮関連の文献などを調べてみるとよい．下記の参考文献には，これらの問題点に対応するための改良版ランレングス圧縮を含め各種データ圧縮手法の C プログラムが載っている.

★ 3.5　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

★ 3.6　情報量の概念

★ 3.6.1　情報量の定義

ある事象が起こったことを知らせる際に伝達される「情報の量」を考えてみよう.「1 億枚中 1 枚しか当たりのないくじではずれをひいた」という知らせよりも「そのくじで当たりをひいた」という知らせの方が情報が多い気がしないだろうか (☆ 5).「犬が人を噛んだ」というニュースよりも「人が犬を噛んだ」という方が情報が多い気がしないだろうか．実は，このような素朴な議論から出発して，**情報量**というものを次のように定義するとよいことが知られている．

☆ 5) 「びっくり度」が高い，と言ってもよいかも．

ある事象 E の生起確率が $P(E)$ であるとき，E が起こったことを知らせる際に伝達される情報量 $I(E)$ を次式で定義する：

$$I(E) = \log\left(\frac{1}{P(E)}\right) = -\log P(E) \tag{1}$$

対数の底は任意に定めればよい (☆ 6) が，bit との対応づけができるため，2 を選ぶことが多い．その場合，情報量の単位は [bit] となる．

☆ 6) ある二つのことの情報量に注目すると，底の選び方を変えてもそれらの大小関係は変化しない．ただし底の選び方によって情報量の値そのものは変化する．

例： 確率 $\frac{1}{2}$ で赤旗か白旗のいずれかが上がるのを観測する場合,「赤（白）の旗が揚がった」という知らせの情報量は $-\log_2 \frac{1}{2} = 1[\text{bit}]$ となる (☆ 7).

☆ 7) 「赤旗が揚がった」と「白旗が〜」の 2 つに 2 進数を対応づけるためには 1 桁 (bit) の 2 進数が必要であるということに対応している．たとえば，0 が「赤旗が〜」で 1 が「白旗が〜」.

例： 天気が「晴れ」か「雨」のどちらかにしかならない地域があったとする．晴れの確率は 0.9 だという．この場合，この地域の天気が晴れだということ，および雨だということを知らせる情報の情報量は次のようになる．

$$I(\text{晴れ}) = -\log_2 0.9 = -\frac{\log 0.9}{\log 2} = \text{約 } 0.152[\text{bit}]$$

$$I(\text{雨}) = -\log_2 0.1 = -\frac{\log 0.1}{\log 2} = \text{約 } 3.32[\text{bit}]$$

例： 4 つの事象 A, B, C, D のどれかが起こるとして，それらの生起確率が等しい（つまり $\frac{1}{4}$ である）場合，どの事象が起こったかを知らせる情報の情報量は 2[bit] となる (☆ 8).

☆ 8) A, B, C, D に対応させるには 2 桁 (bit) の 2 進数が必要.

Q2. 20 通りの事象のいずれかが起こるとする．それらの生起確率が等しい場合，どの事象が起こったかを知らせる情報の情報量は何 bit か．必要ならば $\log_2 5 = 2.32$ を用いたらよい．

Q3. 「晴れ」,「曇り」,「雨」,「雪」の 4 通りの天気のいずれかが起こる地域があったとする．この地域の雪の確率が $\frac{1}{512}$ だとすると，この地域の天気が雪だということを知らせる情報の情報量は何 bit になるか．

★ 3.6.2 情報量の性質

例： $4 \times 13 = 52$ 枚のカードから成るトランプから無作為に 1 枚引くなら

- 「それが ♡ だった」という知らせの情報量は $-\log \frac{1}{4} = \log 4$
- 「エース (1) だった」という知らせの情報量は $-\log \frac{1}{13} = \log 13$
- 「♡ のエースだった」という知らせの情報量は $-\log \frac{1}{52} = \log 52$

となる．つまり，$I(♡) + I(\text{エース}) = I(♡ \text{ かつエース})$ である (☆ 9)．

ここから予想できるように，情報量には次のように**加法性**が成り立つ (☆ 10)：

事象 A の情報量が $I(A)$, 事象 B の情報量が $I(B)$ であるとき，A, B が独立ならば，事象 $A \cap B$ の情報量（A も B も起こったことを知らせる）は

$$I(A \cap B) = I(A) + I(B) \tag{2}$$

となる (☆ 11)．

Q4. さいころを振って出た目が「3 の倍数だった」,「偶数だった」,「6 だった」ということを知らされることで得られる情報量をそれぞれ求めなさい．ただし，単位は bit とすること．これらの間にはどのような関係があるだろう（ヒント：「3 の倍数」かつ「偶数だった」というのはどんな場合？）．

☆ 9) ここでは対数の底を e としているが，他の値にしても結論は変わらない．

☆ 10) 実際の理論の成り立ちは逆で,「素朴に考えると情報量というのは『生起確率の単調減少関数』でありかつ『加法性が成り立つ』ものでなければならない」というのを出発点にして，そのような性質を満たすのは式 (1) のような対数の形でなければならないことが導出されている．

☆ 11) 「ハートだった」を A,「赤いのだった」を B としてみると，どうなるだろう．

★ 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，

$$H = \sum_{i=1}^{n} p_i I(E_i) = -\sum_{i=1}^{n} p_i \log p_i \tag{3}$$

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）

$$H = p \times I(\text{表}) + (1-p) \times I(\text{裏}) = -p \log_2 p - (1-p) \log_2 (1-p) \quad [\text{bit}] \tag{4}$$

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．

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．

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．

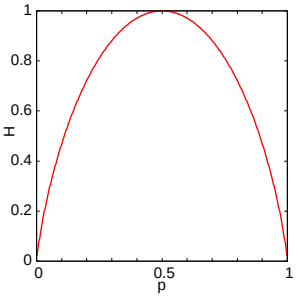


図: コイン投げのエントロピー

Q5. 「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 (☆ 12)．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．

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