符号化と通信の理論

新居俊作

2023 年 6 月 5 日

データ送受信には様々な様式がある。例えばインターネットではマンチェスター・コードと呼ばれる通信方式が使われている。いずれの方式でも必ずノイズによる誤りが入り込む。以下では誤りを発見して修正する方法を考える。

1 先ずは素朴に Try and Error

パリティ検査

一度に 10 ビットの 0, 1 の列を送るとする。

誤りの起こる確率が各ビットあたり 1 万分の 1 とすると、長さ 10 ビットのメッセージの 1 つの記号が誤る確率は千分の 1 である。二つの記号が誤る確率は非常に小さいので無視することにする。

10 ビットのメッセージの終わりにパリティ(偶奇) を表すパリティビットを 1 つ付け加える。*i.e.*

メッセージ内の 1 の数が偶数なら 0

メッセージ内の 1 の数が奇数なら 1

をメッセージの後に加える。

例

メッセージが 1101011001 なら最後に 0 を付け加えて 11010110010 を送る。

受け取ったメッセージのパリティとパリティビットが異なれば誤りがあることになるのでもう一度メッセージを送り直す。単にメッセージを送る場合と比べて 10% 効率が落ちるが誤りを見つけられるようになる。

この方法の欠点：

- パリティビットに誤りが生じると、メッセージに誤りがなくても誤りと判定される。
- どこに誤りが生じたか分からないので、誤りを修正出来ず、もう一度送り直すしか無い。

改良方法

メッセージを 10×10 の行列に並べ、各行と各列の最後にパリティビットを付け加え、全体の最後に全メッセージのパリティビットを加える：

1	1	0	$\cdots$	$\cdots$	0	0	1	1
0	1						$\vdots$	
1		$\ddots$						行
$\vdots$								の
$\vdots$		メ	ッ	セ	-	ジ		パ
$\vdots$								リ
$\vdots$							$\vdots$	テ
1	$\cdots$					$\cdots$	1	ィ
1		列	の	パ	リ	テ	ィ	1

全メッセージのパリティ↗

メッセージに誤りが生じると行と列の両方のパリティビットが合わなくなるので、どこに誤りが生じたかがわかる。$\longrightarrow$ 誤りが訂正できる。

注意点

メッセージの 1 つのビットとその行 (列) のパリティビットが同時に誤ると、列のパリティビットの誤りに見えてしまう。この場合は全メッセージのパリティと全メッセージのパリティビットを比較するとメッセージの誤りとわかる。この場合はどこに誤りがあるか分からないので、メッセージを送り直す。

定義 1. 符号化法の冗長度 R を

$$R := \frac{\text{全体で使われるビット数}}{\text{メッセージのビット数}}$$

で定義する。また、$R-1$ を超過冗長度とよぶ。

上の例では $R = \frac{11\times11}{10\times10} = 1.21$ だが、一度に送るメッセージの長さを長くすれば R は小さくなる。しかし、二つ以上の誤りが入り込む確率も大きくなって無視できなくなる。

ハミング符号

長さ 11 ビットのメッセージを送るのに、4 ビットのパリティ検査ビットを付けて 15 ビットの符号をとする。例えば

10111011011

を送るのに 4 桁の検査ビット a,b,c,d を 1, 2, 4,8 番目に置く *i.e.*

位置	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
符号	a	b	1	c	0	1	1	d	1	0	1	1	0	1	1

ここで、
a：奇数位置 1,3,5,... のパリティが偶数になる様に決める。例では 0
b：位置 2,3,6,7,10,11,14,15 のパリティが偶数になる様に決める。例では 0
c：位置 4,5,6,7,12,13,14,15 のパリティが偶数になる様決める。例では 1
d：位置 8,9,10,11,12,13,14,15 のパリティが偶数になる様決める。例では 1

従って最終的な符号は

101101111011011

となる。

1 〜 15 を二進数で表すと、次のような 4 桁になる。

1	2	3	4	5	6	7	8	9	10
0001	0010	0011	0100	0101	0110	0111	1000	1001	1010

11	12	13	14	15
1011	1100	1101	1110	1111

これを見ると、
奇数位置では下から 1 桁目が 1
位置 2,3,6,7,10,11,14,15 では下から 2 桁目が 1
位置 4,5,6,7,12,13,14,15 では下から 3 桁目が 1
位置 8,9,10,11,12,13,14,15 では下から 4 桁目が 1

となっている。従って、a,b,c,d を含んだ符号でパリティ検査を行うと、**パリティビットが誤っている場合を含めて**誤りの位置がわかる。この例では、もし受け取った符号が

$$101101\underline{0}11011011$$

ならば、
奇数位置のパリティが奇数で間違っており、誤り位置の下 1 桁目は 1
位置 2,3,6,7,10,11,14,15 のパリティも奇数で、誤り位置の下 2 桁目も 1
位置 4,5,6,7,12,13,14,15 のパリティも奇数で、誤り位置の下 3 桁目も 1
位置 8,9,10,11,12,13,14,15 のパリティは偶数で、誤り位置の下 4 桁目は 0
以上より誤りの位置の 2 進表示は 0111(= 7) と分かる。

この様な符号化をハミング符号 (Hamming code) とよぶ。

二重誤りを発見するには、最後に全体のパリティを表す符号を付け加えて 16 桁の符号にする。誤りが一つだけなら全体のパリティが合わなくなるので、メッセージ中の一つの誤りを上記の方法で見つけて訂正すればよい。(誤りが見つけられなければ全体のパリティビットの誤り。) 二重誤りの場合は 5 桁のパリティ検査は誤りを示すのに全体のパリティは合う。この場合はどこに誤りがあるか分からないので符号全体を送り直す。

ハミング符号の冗長度を考える。m 桁のパリティ検査ビットで $2^m - 1$ 桁の符号の誤りを判定できる。よってメッセージの長さを n とすると、$2^m - 1 \geq n + m$ の範囲 i.e. $n = 2^m - m - 1$ 桁までのメッセージが送れる。

二重誤りまで含めて $n = 2^{m-1} - (m-1) - 1$、 従って

$$R = \frac{n+m}{n} = \frac{2^{m-1}}{2^{m-1} - m}$$

$m = 5$ で $R = 1.46$、$m = 8$ で $R = 1.067$。m を大きくすると R は小さくなるが、二重三重の誤りが起こる確率が大きくなり無視できなくなる。

2 理論的な考察

<u>ハミング距離</u>

定義 2. 長さ m の二つの $0, 1$ の列

$$\boldsymbol{a} = a_1 a_2 \cdots a_m, \qquad \boldsymbol{b} = b_1 b_2 \cdots b_m$$

に対し、ハミング距離 $d(\boldsymbol{a}, \boldsymbol{b})$ を以下の様に定める：

$$d(\boldsymbol{a}, \boldsymbol{b}) := \sum_{i=1}^{m} |a_i - b_i| \tag{2.1}$$

メッセージ $\boldsymbol{a}$ の 1 個のビットが誤って $\boldsymbol{a}'$ になると $d(\boldsymbol{a}, \boldsymbol{a}') = 1$。同様に e 個の誤りが起こって $\boldsymbol{a}''$ になると $d(\boldsymbol{a}, \boldsymbol{a}'') = e$ になる。

長さ M のメッセージ空間 $\{0,1\}^M$ を長さ M_c の符号空間 $\{0,1\}^{M_c}$ に埋め込むことを考える：

$$\varphi\colon \{0,1\}^M \to \{0,1\}^{M_c} \tag{2.2}$$

このとき $\boldsymbol{a}, \boldsymbol{b} \in \{0,1\}^M$ に対し、

$$\boldsymbol{a} \neq \boldsymbol{b} \Longrightarrow d(\varphi(\boldsymbol{a}), \varphi(\boldsymbol{b})) \geq 2e + 1 \tag{2.3}$$

となる様に出来れば、メッセージの e 個の誤りまで訂正できる。

問題 1. (2.3) が成り立つならば、メッセージの e 個の誤りまで訂正できることを示せ。

効率的な φ を作るには数学の一分野である代数幾何学が使われている。(代数幾何符号。)

<u>シャノンの定理</u>

長さ M のメッセージを符号化して長さ M_c になったとする。各ビットが誤って伝わる確率を q とするとき、誤りをいくらでも確実に訂正できるためには M_c を M に対してどれくらい長くする必要があるか。

定理 1 (シャノンの定理)**.**

$$\frac{M}{M_c} \leq 1 - \left(\log_2 \frac{1}{q} + (1-q) \log_2 \frac{1}{1-q} \right) \tag{2.4}$$

<u>例</u>

q	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{4}$	$\frac{1}{10}$	$\frac{1}{100}$	$\frac{1}{1000}$
$\frac{M}{M_c}$	0	0.082	0.19	0.53	0.919	0.998

注

- シャノンの定理は具体的に M_c を作る方法は与えない抽象的な定理。
- 具体的な方法でこの上限に近づけようとすると M_c を非現実的に大きくしなければならない。

  例

  惑星探査機との通信は各ビットの誤り確率が $\frac{1}{3}$ 程度なので、$\frac{M}{M_c}$ は 8% 強で、メッセージ 1 ビット当り検査ビットを 12 ビット送ればよい計算になるが、そのためには M と M_c を非現実的に長くしなければならないので、実際には小さな M と M_c で、メッセージ 1 ビット当り 150 ビットの検査ビットを送っている。

シャノンの定理の説明

長さ M のメッセージを長さ M_c の符号に符号化するとする。平均 k 個の誤りが発生するとすると、誤りの起こる確率が q ならば $k = M_c q$ である。これを訂正するには、各 $a \in \{0,1\}^M$ について $\varphi(a) \in \{0,1\}^{M_c}$ 中心半径 k の球が $\{0,1\}^{M_c}$ に重なり合わないように埋め込まれている必要がある。

$\{0,1\}^{M_c}$ 内で半径 k の球に含まれる点の数は

$$\frac{M_c!}{k!(M_c-k)!}$$

以上である。よって、長さ M のメッセージ各々に半径 k の球を対応させると、点の個数は合計で

$$\frac{M_c!}{k!(M_c-k)!}2^M$$

以上ある。これが長さ M_c の符号全体の点の数より少なくなければいけないので

$$2^{M_c} \geq \frac{M_c!}{k!(M_c-k)!}2^M$$

となる。

Stirling の公式より $\log n! \approx n\log n - n$ なので (対数の底は e)、両辺底を e とする対数をとって、Stirling の公式を用いた後に底の変換公式で底を 2 とすると、

$$M_c \geq M_c \log_2 M_c - M_c \log_2(M_c-k) + k\log_2(M_c-k) - k\log_2 k + M \quad (2.5)$$

ここで $k = M_c q$ なので、

$$
\begin{aligned}
&= M_c \log_2 M_c - M_c \log_2 M_c(1-q) + M_c q \log_2 M_c(1-q) - M_c q \log_2 M_c q + M \\
&= M_c \left\{ q \log_2 \frac{1}{q} + (1-q) \log_2 \frac{1}{1-q} \right\} + M
\end{aligned}
\tag{2.6}
$$

全体を M_c で割って

$$
1 \geq q \log_2 \frac{1}{q} + (1-q) \log_2 \frac{1}{1-q} + \frac{M}{M_c} \tag{2.7}
$$

i.e.

$$
\frac{M}{M_c} \leq 1 - \left\{ q \log_2 \frac{1}{q} + (1-q) \log_2 \frac{1}{1-q} \right\} \tag{2.8}
$$

問題 2. シャノンの定理は $M, M_c \to \infty$ の極限では任意の精度で成り立つことを以下の方法で示せ。

誤りの分布は二項分布に従い平均値が $k = M_c q$ で標準偏差が $\sigma = \sqrt{M_c q(1-q)}$ であると考えられる。従って、誤りが訂正できない確率を与えられた許容範囲以下に抑えたければ、それに応じて $k' = k + g\sigma$ 個 (g は定数) までの誤りを訂正できるようにすればよい。例えば $g = 2$ とすれば、96% の確率で誤りを訂正でき、$g = 20$ ならば誤りが訂正できない確率は 10^{-100} 以下になる。

そこで、上述の説明中の k を k' で置き換えて $M, M_c \to \infty$ の極限をとることで、この場合も極限でシャノンの定理が成り立つことを示せ。

<u>情報量</u>

アルファベットと "+ , . ⋯" で英語は 30 文字を使う。N 文字で送れる英文の数はどれだけあるかを考える。当然 30^N よりははるかに少ない。なぜなら殆どのアルファベットの列は英語として意味を成さない。日本語でも俳句の総数は 50^{17} よりはるかに少ない。英語では、ある文字に続く可能性のある文字の平均数は約 5 ($\ll$26)。

N 文字の英文で送ることの出来る英文の数を n とすると、これらの英文に番号を付けておいて、その番号を送るのが最も効率が良い情報伝達の方法。この時 $2^I = n$ となる $I = I(N)$ *i.e.* $I(N) = \log_2 n$ を N 文字の英文で送れる情報量と呼ぶ。N が十分大きくて、単語が入りきらなかったりしない場合を考えて

$$
r := \frac{I(N)}{N}
$$

を文字あたりの情報と呼ぶ。

一般に M 個の文字 $a_1, a_2, \ldots, a_M$ から構成されるメッセージにおいて、各 a_i の出現頻度が確率 p_i であるとする。このとき長さ N のメッセージの数は N が大きければ

$$\frac{N!}{(Np_1)!(Np_2)!\cdots(Np_M)!}$$

となる。

よって、長さ N のメッセージの情報量の期待値は

$$\begin{aligned} < I > &= \log_2 \left(\frac{N!}{(Np_1)!(Np_2)!\cdots(Np_M)!} \right) \\ &\approx N \log_2 N - N - \sum_{i=1}^{M} \{(Np_i)\log_2(Np_i) - (Np_i)\} \\ &= N \sum_{i=1}^{M} (-p_i \log_2 p_i) \end{aligned} \tag{2.9}$$

これより 1 文字当の平均情報量は

$$\frac{< I >}{N} = \sum_{i=1}^{M} (-p_i \log_2 p_i) \tag{2.10}$$

これを情報エントロピーとよぶ。

定理 2 (情報源符号化定理).

(平均符号長) $\geq$ (エントロピー)

うまく符号化することでエントロピーにいくらでも近付けることが出来る。例えば、出現頻度の高い単語には桁数の小さい番号を付け、頻度の低い単語には桁数の大きい番号を付ける。

例

ハフマン符号、モールス符号 (例えば、頻度の高い "E" には "・"、低い "Q" は "– – ・ –")。

3　取りこぼした話題

標本化定理

アナログデータである音声データをデジタル化するには一定の時間間隔で飛び飛びのデータを記録する。その場合当然元のデータとは誤差が生じるが、それが許容範囲におさまるための条件を与えるのが**標本化定理**である。

人間の耳は高い音は 20kHz まで聞き取ることができるとされている。(低い方は 20Hz 以上。) CD では、20kHz までの音を録音するために標本化定理によって必要とされる 40kHz より高い 44kHz の周波数でサンプリングを行っている。(ハイレゾでは 96 kHz や 192kHz。)

問題 3. 標本化定理について調べよ。

予測符号化

画像データは、画面中の広い領域が同じ色である、という著しい特徴がある。この特徴を利用して通信データ量を削減する方法として**予測符号化**という方法がある。

量子通信

計算理論でも量子計算によって特定の問題の計算時間が劇的に短くなることが知られていて、現在実用化に向けた研究が活発に行われているが、通信においても量子力学の原理を用いることによる新しい通信方式の研究が活発に行われている。それにより、例えば盗聴に対し非常に安全な (『今のところ』原理的に盗聴できないとされる) 通信が実用化されるとされている。

参考文献

[ファインマン] R. ファインマン 述, A. ヘイ, R. アレン 記, 原 康夫, 中山 健, 松田 和典 訳
「ファインマン計算機科学」
岩波書店 (1999 年)

[シャノン] クロード・E. シャノン, ワレン ウィーバー 著, 植松 友彦 訳
「通信の数学的理論」
ちくま学芸文庫 (2009 年)