第 8 回
学部 3 年後期ゼミナール発表資料

青山和樹

2024 年 12 月 9 日

発表の目的

テキスト [1] の

- 5.1 Examples of codes
- 5.2 Kraft inequality

について発表する.

目次

エントロピーの定義に対して, 情報の圧縮に関する基本的な限界という意味を与える. 圧縮は, データソースに対して, 頻繁に出現するものに短い説明を与え, 出現頻度の低いものに長い説明を割り当てることによって達成される. ここでは, 「瞬時符号（instantaneous code)」という概念を定義し, その後クラフトの不等式（Kraft inequality）を証明する.

1 Example of codes

定義 1.1（情報源符号（codeword））

確率変数 X に対する情報源符号 C とは $\mathcal{X}$（X の値域）から $\mathcal{D}^*$（D 元アルファベットからなる有限長の文字列の集合）への写像である. $C(x)$ を x に対応する符号語とし, $l(x)$ を $C(x)$ の長さとする.

$$C : \mathcal{X} \to \mathcal{D}^* \tag{1}$$

例 1.2　$\mathcal{X} = \{\text{red}, \text{blue}\}$, アルファベット $\mathcal{D} = \{0, 1\}$ に対して

- $C(\text{red}) = 00$
- $C(\text{blue}) = 11$

となるような情報源符号が考えられる.

定義 1.3（平均符号語長（expected length））

平均符号語長 $L(C)$ は確率関数 $p(x)$ に対する確率変数 X の情報源符号 C によって

$$\begin{aligned} L(C) &= \sum_{x \in \mathcal{X}} p(x) l(x) && (2) \\ &= \mathbb{E}[l(X)] && (3) \end{aligned}$$

で表される.

注意 1.4　一般性と失うことなく, D 元アルファベットを $\mathcal{D} = \{0, 1, \ldots, D-1\}$ と仮定できる

例 1.5　確率変数 X を次の分布と符号語の割り当てを持つとする.

$$\begin{aligned} &\mathbb{P}[X = 1] = \frac{1}{2}, \quad \text{符号語 } C(1) = 0, \\ &\mathbb{P}[X = 2] = \frac{1}{4}, \quad \text{符号語 } C(2) = 10, \\ &\mathbb{P}[X = 3] = \frac{1}{8}, \quad \text{符号語 } C(3) = 110, \\ &\mathbb{P}[X = 4] = \frac{1}{8}, \quad \text{符号語 } C(4) = 111, \end{aligned} \tag{4}$$

$$H(X) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{4}\log\frac{1}{4} - \frac{1}{8}\log\frac{1}{8} - \frac{1}{8}\log\frac{1}{8} \tag{5}$$
$$= 1.75 \quad \text{bits} \tag{6}$$

$$L(C) = \mathbb{E}[l(X)] \tag{7}$$
$$= \frac{1}{2}1 + \frac{1}{4}2 + \frac{1}{8}3 + \frac{1}{8}3 \tag{8}$$
$$= 1.75 \quad \text{bits} \tag{9}$$

X のエントロピー $H(X)$ は 1.75 ビットであり, この情報源符号の平均符号語長 $L(C)$ も 1.75 ビットである. よって今回の場合では, エントロピーと平均符号語長が一致する情報源符号が得られた. またこの符号では, 任意のビット列を X のシンボル列に一意に復号できる. 例えば, ビット列 0110111100110 はシンボル列 134213 に復号される.

例 1.6　別の簡単な例として, 確率変数 X を次の分布と符号語の割り当てを持つとする.

$$\begin{aligned} \mathbb{P}[X=1] &= \frac{1}{3}, \quad \text{符号語 } C(1) = 0, \\ \mathbb{P}[X=2] &= \frac{1}{3}, \quad \text{符号語 } C(2) = 10, \\ \mathbb{P}[X=3] &= \frac{1}{3}, \quad \text{符号語 } C(3) = 11, \end{aligned} \tag{10}$$

例 1.5 と同様に, この符号も一意に復号可能である. しかしこの場合, エントロピーは $l\log 3 \approx 1.58$ ビットであり, 平均符号語長は 1.66 ビットである. よって $\mathbb{E}[l(X)] > H(X)$ となる.

例 1.7　モールス符号

モールス符号は, 英字アルファベットを表現するために 4 種類の記号トン, ツー, 文字間スペース, 単語間スペースからなる, 比較的効率的な情報源符号である. 頻繁に使用される文字（例：E）は短い符号列（トン）で表され, 頻度の低い文字（例：Q）は長い符号列（ツーツートンツー）で表される.

しかし, この符号は 4 種類の記号でのアルファベットの最適な表現ではない. 実際, 多くの可能な符号語が使用されていない. その理由は, 文字の符号語に空白が含まれず, 各符号語の末尾にのみ文字間の空白が存在するためである. また, 空白の後に別の空白が続くことがない. この制約の下で構築できる符号列の数を計算することは興味深い問題である. この問題はシャノンによって 1948 年の元の論文で解かれた.

これから, 符号に対するより厳密な条件を段階的に定義していく. そこで x^n を $(x_1, x_2, \ldots, x_n)$ と表す.

定義 1.8（非特異（nonsingular））

情報源符号 C が非特異とは, X の値域のすべての要素が $\mathcal{D}^*$ の異なる文字列に対応するときである.

$$x \neq x' \Rightarrow C(x) \neq C(x') \tag{11}$$

解説 1.9　つまり情報源符号 C が単射であるとき

例 1.10　非特異符号
$\mathcal{X} = \{A, B, C\}$ として, $C(A) = 0, \quad C(B) = 10, \quad C(C) = 11$

例 1.11　特異符号
$\mathcal{X} = \{A, B, C\}$ として, $C(A) = 0, \quad C(B) = 0, \quad C(C) = 10$

解説 1.12　非特異は X の単一の値を曖昧なく記述するには十分である. しかし通常は, X の値の列を送信する場合が多い. このような場合は, 任意の 2 つの符号語の間に特別な記号（カンマなど）を追加することで, 復号可能性を保証できる. ただし, これは特別な記号の非効率的な使用となる. より効率的な方法として, 自己終端（self-punctuaring）または瞬時符号（instantaneous）のアイデアを発展させることができる. シンボル X の列を扱うことから, 符号を拡張する.

定義 1.13（符号 C の拡張（extension））

情報源符号 C の拡張 C^* は $\mathcal{X}$ の有限系列から $\mathcal{D}^*$ への写像であり,

$$C(x_1 x_2 \cdots x_n) = C(x_1)C(x_2)\cdots C(x_n) \tag{12}$$

で表される.
ここで, $C(x_1)C(x_2)\cdots C(x_n)$ は対応する符号語の連接（concatenation）を示す.

$$C^* : \mathcal{X}^* \to \mathcal{D}^* \tag{13}$$

例 1.14　$C(x_1) = 00, \quad C(x_2) = 11$ であれば, $C(x_1 x_2) = 0011$

定義 1.15（一意復号可能（uniquely decodable））

一意復号可能であるとは, その情報源符号の拡張 C^* が非特異であるときである.

解説 1.16　一意復号可能な符号では, 符号化された任意の文字列に対して, それを生成する情報源系列は 1 つしか存在しない. しかし, 元の文字列の最初のシンボルを特定するためにも, 文字列全体を確認しなければならない場合がある.

例 1.17　$\mathcal{X} = \{A, B, C\}$ として, $C(A) = 0, \quad C(B) = 01, \quad C(C) = 011$ とする. このとき 001011 は ABC に復号される.

定義 1.18（語頭符号（prefix code）・瞬時符号（instantaneous code））

符号が語頭符号または瞬時符号であるとは, どの符号語も他の符号語の語頭（先頭部分）にならないことである.

解説 1.19　瞬時符号は符号語の終わりが即座に認識できるため, 先の符号語を読まずに復号できる. 瞬時符号は自己終端符号でもある. 例えば例 1.5 の符号によって生成された 2 進数文字列 01011111010 は 0, 10, 111, 110, 10 として解釈される.

これらの定義の入れ子構造は図 1 に示される. 様々な種類の符号の違いを説明するために, $x \in \mathcal{X}$ に対する符号語 $C(x)$ の割り当て例を表 1 に示す.

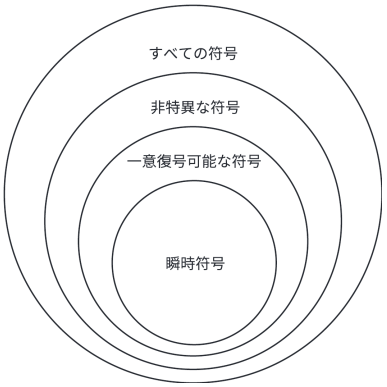

図 1　符号のクラス

表 1　符号のクラス

X	非特異	非特異であるが 一意復号可能ではない	一意復号可能であるが 瞬時符号でない	瞬時符号
1	0	0	10	0
2	0	010	00	10
3	0	01	11	110
4	0	10	110	111

非特異符号の場合, 符号列 010 は 2, 14, 31 の 3 つの可能な元の列が存在する. そのため, この符号は一意復号可能ではない. 一意復号可能な符号は語頭符号でないので, 瞬時符号でもない.

一意復号可能であることを確認するには, 任意の符号列を最初から復号すればよい. 最初の 2 ビットが 00 または 10 であれば, それらは瞬時に復号できる. また最初の 2 ビットが 11 である場合は次のビットを見る必要があり, 次が 1 であれば元のシンボルは 3 で, 続くビット列の長さが奇数で 0 であれば, 元のシンボルは 4 となる. この議論を繰り返すことで, 符号が一意復号可能かどうか確認できる.

表 1 の最後の符号はどの符号語も他の符号語の語頭ではないため, 瞬時符号である.

解説 1.20　サルティナス（Sarddinas）とパターソン（Patterson）は一意復号可能性を確認するための有限のテストを考案した. このテストでは, 符号語の可能な接尾辞の集合を形成し, それらを体系的に排除する方法を用いる.

2　Kraft inequality

与えられた情報源を記述する瞬時符号でのかで, 平均符号語長が最小のものを構築したい. しかし, 全ての情報源シンボルに語頭が被らず短い符号語を割り当てるのは不可能である. 瞬時符号において可能な符号語長の集合は次の不等式によって制限される.

定理 2.1（クラフトの不等式（Kraft inequality））

要素数 D のアルファベット上の全ての瞬時符号（語頭符号）について, 符号語の長さ $l_1, l_2, \ldots, l_m$ は 1

$$\sum_{i=1}^{m} D^{-l_i} \leq 1 \tag{14}$$

を満たさなければならない.

解説 2.2　逆にこの不等式を満たす符号語長の集合が与えられると, それらの長さを待つ瞬時符号が存在する.

証明 2.3　各ノードが D 個の子ノードを持つ D 分木を考える. 各枝がシンボルに対応する. 例えば, 根から生じる D 個の枝は符号語の最初のシンボルの D 個の可能な値を表す. この場合, 各符号語は木の葉に対応する. 根から葉までの道すじが符号語のシンボルを辿る. このような木の 2 分木を図 2 に示す.

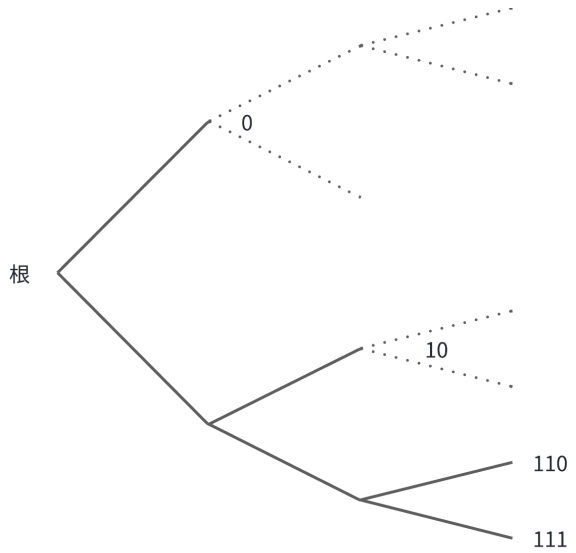

図 2　クラフトの不等式のための符号木

符号語に対する語頭条件は木上でどの符号語も他の符号語の祖先にならないことを意味する．したがって, 各符号語は, その子孫を符号語としての候補から排除する．

符号語集合における最長の符号語の長さを $l_{\max}$ とする．深さ $l_{\max}$ にあるすべてのノードを考える．それらは符号語であるもの, 符号語の子孫になるもの, そのどちらでもないものである．深さ l_i にある符号語は, 深さ $l_{\max}$ に $D^{l_{\max}-l_i}$ この子孫を持つ．これらの子孫の集合は互いに排反でなければならない．また, これらの集合内のノードの総数は $D^{l_{\max}}$ 以下でなければならない．したがって, 符号語について総和を取ると

$$\sum_{i=1}^{m} D^{l_{\max}-l_i} \leq D^{l_{\max}} \tag{15}$$

となり, 両辺を $D^{l_{\max}}$ で割ると,

$$\sum_{i=1}^{m} D^{-l_i} \leq 1 \tag{16}$$

となるため, クラフトの不等式を得る．

参考文献

[1] T.M.Cover and J.A.Thomas, Elements of Information Theory, Second Edition, John Wiley & Sons, 2006.