２０１９年４月８日（川端）

データ圧縮基礎資料

1　エントロピーの概念

1.1　情報源符号化

集合 $\mathcal{A}$ を有限の文字からなるアルファベットとする．これを**符号アルファベット**とよぶ．例としては，二進アルファベット $\{0,1\}$ がある．符号アルファベット $\mathcal{A}$ からなる長さ l の文字列の集合を $\mathcal{A}^l$ と書く．すなわち，$\mathcal{A}_1 = \mathcal{A}$, $\mathcal{A}_2 = \mathcal{A}$,..., $\mathcal{A}_l = \mathcal{A}$ とするとき，$\mathcal{A}^l = \mathcal{A}_1 \times \mathcal{A}_2 \times \ldots \times \mathcal{A}_l$ は文字列 $c_1 c_2 \ldots c_l$ （ただし，$c_1 \in \mathcal{A}_1$, $c_2 \in \mathcal{A}_2$, ..., $c_l \in \mathcal{A}_l$）の集合であると考える．これを，集合 $\mathcal{A}$ の l **次拡大**という．さらに，空列を λ と表すものとすると，これは $\mathcal{A}^0$ の要素である．そうすると，集合 $\mathcal{A}^* := \mathcal{A}^0 \cup \mathcal{A}^1 \cup \mathcal{A}^2 \cup \ldots$ は空列もふくめて有限の長さの $\mathcal{A}$ の文字列の全体を表すことになる．

さて，情報源符号化の考えを，電報局での電報を抽象化したものとして導入する．分りやすくするために，祝電や弔電などは，可能な**メッセージ**の組として用意されており依頼者はそのうち一つを選んで配達を依頼する．メッセージの集合 $\mathcal{X}$ を**メッセージ情報源**と呼ぶ．電報局では $\mathcal{X}$ の中から発生したメッセージを，決められた規則に従いアルファベット $\mathcal{A}$ の文字列に変換し，その文字列を配達先の電報局に電送する．この変換を**情報源符号化**という．この変換は数学的にはメッセージ情報源 $\mathcal{X}$ の要素に対して $\mathcal{A}^*$ の要素を対応させる写像 φ である．そこでこの写像 φ を $\mathcal{A}$ を用いた $\mathcal{X}$ の情報源符号化ということにする．一般に符号化とは文字列の集合から文字列の集合への写像である．（この写像は固定長 (Fixed length) の文字列を可変長 (Variable length) 文字列への変換とみなせるため正確には, **FV(Fixed-to-Variable Block Length) 情報源符号化**という.）

定義 1 (情報源符号化 (source coding))

$$\varphi : \mathcal{X} \to \mathcal{A}^*$$

すなわち，情報源符号化は $\mathcal{X}$ を入力とし $\mathcal{A}^*$ を出力としている．$\mathcal{X}$ が離散集合の場合，その各要素の番号をつければ，対応する φ の像のリスト $\{\varphi(x)\}_{x\in\mathcal{X}}$ が得られる．これを**符号** (code) といい，符号の各要素を**符号語** (codeword) という．　一般に 1 対 1 写像であるような符号化は**非特異**であるという．もし，一度だけ符号化を行うときには，非特異な符号の出力からは，入力を一意に復元することができる．一般に，符号化された入力を出力から復元する操作を**復号化**という．また文字列をいくつかつなげることを**連接**するという．連接された符号語列から，情報源列を復号するためには，符号が非特異であることは必要条件であるが，十分条件ではない．必要十分条件は「2 通りの符号語列の連接結果が決して等しくなることがない」ことであり，この条件を満たす符号を**一意に復号が可能な符号**という．

1.2　Kraft の不等式

$\mathcal{X}$ を離散集合とする. $B = \{0,1\}$ を二進アルファベットとする. 前節でのべた一意復号可能な符号について以下の事実が知られている。一意復号可能な符号 $c : \mathcal{X} \to B^*$ の符号長関数，すなわち符号語の長さを表す関数 $l : \mathcal{X} \to \mathcal{N}$ は不等式：

$$\sum_{x\in\mathcal{X}} 2^{-l(x)} \leq 1$$

をみたす（必要性）．　この不等式はクラフト (Kraft) の不等式という．逆にクラフトの不等式を満足する符号長関数をもつような, 一意復号可能な符号は必ず存在すること（十分性）も証明できる.

証明：（必要性）$c : \mathcal{X} \to B$ が一意復号可能な符号とする．任意の $k \geq 0$ に対してその k 次拡大符号 $c^k : \mathcal{X}^k \to B^*$ $(x^k \mapsto c(x_1)c(x_2)\dots c(x_k))$ は 1 対 1 写像である．$l^k : \mathcal{X}^k \to \mathcal{N}$ を c^k の符号長関数とする．

$$\begin{aligned}\Big(\sum_{x\in\mathcal{X}} 2^{-l(x)}\Big)^k &= \sum_{x^k\in\mathcal{X}^k} 2^{-l^k(x^k)} \\ &= \sum_{0\leq L\leq k\max_x l(x)} 2^{-L} \sum_{x^k : l^k(x^k)=L} 1 \\ &\leq \sum_{0\leq L\leq k\max_x l(x)} 2^{-L} \sum_{b^L\in B^L} 1 \\ &= \sum_{0\leq L\leq k\max_x l(x)} 1 \\ &= 1 + k\max_x l(x)\end{aligned}$$

が成立つことに注意する．ただし不等式では c^k はその像を B^L に制限してもやはり 1 対 1 性を保持することを用いている．もし c についてクラフトの和が 1 より大であるなら左辺は k について指数的に増大するが，右辺は線形的にしか増大しない．これは矛盾である．

（十分性）符号長関数 $l : \mathcal{X} \to \mathcal{N}$ がクラフトの不等式を満足するときこの符号長を有する語頭符号を構成することを示す.（語頭符号は一意復号符号であることに注意.）集合 $\{l(x)\}_{x\in\mathcal{X}}$ の要素を昇順に並べた並びを $l_1 \leq l_2 \leq \ldots \leq l_{\max}$, とする．ただし $\max = \sharp\mathcal{X}$ とする．($\sharp$ 集合は集合の要素数を表す．)　次に長さ $l_{\max}$ の $2^{l_{\max}}$ 個のブロックの集合 $B^{l_{\max}}$ を辞書順に並べたブロック列を考え，その先頭から上の並びに従って，ブロックの個数

$$2^{l_{\max}-l_1}, 2^{l_{\max}-l_2}, \ldots, 2^{l_{\max}-l_{\max}}$$

のグループを順に取り分けることを行う．符号長がクラフトの不等式を満足することから最後まで不足なく取り分けることができる．各グループは共通の長さ

$$l_1, l_2, \ldots, l_{\max}$$

の語頭をもつことに注意する．これら共通の語頭の集まりが求める語頭符号をなす.

1.3　エントロピー（初等確率論による）

情報理論の基本概念であるエントロピーについて述べる．離散集合 $\mathcal{X}$ はメッセージの集合であると解釈する．たとえばジャンケンの手がメッセージならばその集合は $\mathcal{X} = \{a, b, c\}$ と表せる．メッセージは確率的に選ばれるものとする．メッセージが選ばれる頻度が確率（例えば a が 0.3, b が 0.3, c が 0.4 等のように足して 1 になる非負の数のこと）として定まっているものとする．$\mathcal{X}$ 上ランダムに選ばれるメッセージ X を記述するには $x \in \mathcal{X}$ が選ばれる確率の組 $\{p(x), x \in \mathcal{X}\}$, 即ち $\sum_{x\in\mathcal{X}} p(x) = 1$ を満たす $p(x) \geq 0, x \in \mathcal{X}$ なる関数, を与えればよい．この確率の組を確率分布とよび, ランダムなメッセージ X を確率変数と呼ぶ．確率変数を $X : \Omega \to \mathcal{X}$ のような関数であると定義すると便利である．ここで定義域 Ω はランダムネスの全体をあらわす集合でその要素 $\omega \in \Omega$ が決まると関連する物事をすべてきめてしまうものと考える．値域 $\mathcal{X}$ はメッセージ全体の集合とするとメッセージ $X(\omega) \in \mathcal{X}$ が一意に決まってしまうというわけである．

さて，符号 $c : \mathcal{X} \to B^*$ の符号長関数 $l : \mathcal{X} \to \mathcal{N}$ はクラフトの不等式：

$$\sum_{x\in\mathcal{X}} 2^{-l(x)} \leq 1$$

を満たすものとしよう.

$$H(X) = \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{p(x)}$$

を X のエントロピーという. ただし，対数 log の底は 2 であるものとし，以後特に断りがなければそれを仮定する．また $0 \log 0$，$0 \log 1/0$ は 0 であると仮定する．エントロピーが確率分布の関数であることを明示するために, $H(X) = H(p(\cdot))$ と書くときもある. この量は確率変数 $\log(1/p(X))$ の期待値でもある. あらゆる一意復号可能な符号の期待符号長 $El(X)$ はエントロピー $H(X)$ を下限にもつ．証明は以下のとおりである.

$$\begin{aligned}
& El(X) - H(X) && (1)\\
&= \sum_{x \in \mathcal{X}} p(x)\{l(x) + \log p(x)\} && (2)\\
&= \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{2^{-l(x)}} && (3)\\
&\geq \sum_{x \in \mathcal{X}: p(x) \neq 0} p(x) \frac{1}{\ln 2} \{1 - \frac{2^{-l(x)}}{p(x)}\} && (4)\\
&= \frac{1}{\ln 2} \{ \sum_{x \in \mathcal{X}: p(x) \neq 0} p(x) - \sum_{x \in \mathcal{X}: p(x) \neq 0} 2^{-l(x)} \} && (5)\\
&\geq 0 && (6)
\end{aligned}$$

ただし, 最初の不等号を得るには基本的な関係式 $\ln z \leq (z-1), z > 0$ の変形

$$\log \frac{1}{z} \geq \frac{1}{\ln 2}(1 - z)$$

を用いた. 全ての x について $p(x) = 2^{-l(x)}$ のときだけ等号が成り立つ.

2 期待符号長の上界

さて一般の確率分布 $p(x), x \in \mathcal{X}$ について,

$$l(x) = \lceil \log \frac{1}{p(x)} \rceil$$

と定義すると,

$$\sum_{x \in \mathcal{X}} 2^{-l(x)} \leq \sum_{x \in \mathcal{X}} p(x) = 1$$

である. 即ち, $l(x)$ はある一意復号可能な符号の符号長関数である. 期待符号長を計算すると,

$$\sum_{x \in \mathcal{X}} p(x) l(x) \leq \sum_{x \in \mathcal{X}} p(x) \{\log \frac{1}{p(x)} + 1\} \leq H(X) + 1$$

がなりたつ. 即ち, エントロピープラス 1 の期待符号長をもつ一意復号可能な符号が存在する.

3 確率論

3.1 確率の公理

標本空間を Ω その要素である標本を小文字 ω で表す. (例：無限回の公平なコイン投げ. 各標本 ω はコインの表裏の無限列とする. Ω はその全体.) 標本の集合 A は, ランダム標本が A に含まれれば A が起こり, 含まれなければ A が起こらなかったという. このような A を事象（出来事）という.

以下では, 以下では事象 A が起こる確率を非負の実数値 $P(A)$ で表すことを考える. そのために確率の定義域を明確にする必要がある. Ω の部分集合族 $\mathcal{B}$ であって, Ω 自身を含み, 高々可算回の集合演算（和, 差, 交わり）に関して閉じたものを考える. より正確には

$\Omega \in \mathcal{B}$

$A \in \mathcal{B} \Longrightarrow A^c \in \mathcal{B}$

$A_1, A_2, A_3, \ldots \in \mathcal{B} \Longrightarrow \cup_{l=1,2,3\ldots} A_l \in \mathcal{B}$

を全て満たすものを $\sigma-$ **加法族**という.

確率はある $\sigma-$ 加法族 $\mathcal{B}$ を定義域として定義される実数値関数である. ある標本空間についてその事象の全体を**事象系**という. 即ち, **事象系**とは Ω の $\sigma-$ **加法族**のことである. 確率が定義されるならば標本空間ならびにある $\sigma-$ 加法族 $\mathcal{B}$ が事象系として定義されていること [1] と考えるとよい.

全体集合 Ω に関する事象 A の補集合 A^c を A の**余事象**という. 事象 A, B について, それらが共通の標本をもたないとき互いに素であるという. 事象系はそれに属するどの 2 事象も互いに素であるとき, 互いに素であるという.

さて確率は以下の 3 つの性質を満たす写像 $P : \mathcal{B} \to \mathcal{R}$ であると考えられる.

非負性 $\forall A \in \mathcal{B}, 0 \leq P(A) \leq 1$

全確率性 $P(\Omega) = 1$

可算加法性（完全加法性） $A_1, A_2, \ldots$ が互いに素な出来事の可算無限列であるとする. このとき

$$P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$$

これを確率の公理とする. 確率の公理から導かれる基本的性質を列挙しよう.

空事象の確率 $P(\phi) = 0$

有限加法性 加法性は無限集合列を有限集合列に置き換えても成立する

相補性 $P(A) + P(A^c) = 1$

事象には事象の特性関数が付随する. 逆に標本に関する命題にはその真値を与える事象が対応する. この事象の確率, 即ち命題が真である確率を $P($ 命題 $)$ と書く.

[1] $\Omega = [-\infty, \infty] =: \mathcal{R}$ において, $\mathcal{B}$ としてあらゆる半無限区間 $[-\infty, x], x \in \mathcal{R})$ を含む最小の $\sigma-$ **加法族**を考えることができる. これを（半無限区間から生成される）Borel 集合族という. 連続確率分布を論じるときに用いられる. （半無限区間から生成される）Borel 集合族の部分集合族として全ての要素が区間 $[0, 1]$ の部分集合からなるものと考えるとこれも区間 $[0, 1]$ を全体集合とする $\sigma-$ **加法族**である. これはあらゆる半区間の全体 $\{[0, x], x \in [0, 1]\}$ を含む最小の $\sigma-$ **加法族**でもある.

3.2 確率変数と確率分布

標本空間があるとそれとは別の集合上にランダムに分布する値を考えると便利である．(たとえば Ω が無限コイン投げの場合始めて表が出るまでに投げたコインの回数.) ランダムな値 X を確率変数とよびこの確率変数がとる値の集合を分布空間という．正確には, **確率変数**とは標本空間から集合 $\mathcal{X}$ への関数

$$X : \Omega \to \mathcal{X}$$

であって, 値をとる確率が定義される（すねわち' 逆関数の値' が事象である）[2] ものをいう.

$\mathcal{X}$ は $X(\omega)$ が分布する集合なので, X の分布空間であるという．分布空間の例は, 離散集合（有限集合, 可算無限集合）, 連続集合やそれらの直積集合である.

分布空間の部分集合には集合の大きさが定められていることが多い．集合の大きさとは集合に非負の数値を対応させる関数で, 確率と同様加法性をもつ集合の関数である. 例えば離散集合の場合には集合の大きさとは集合の要素数である．また, n 次元連続集合の場合その n 次元体積である. これは $n = 1$ では長さ, $n = 2$ では面積, $n = 3$ では体積である．集合 A の大きさを $\mu(A)$ と記述する.

確率変数 X に付随して関数

$$p_X : \mathcal{X} \to [0, 1]$$

（ただし右辺は単位実数区間を表す）を

$$p_X(x) = \lim_{A \downarrow \{x\}} \frac{P(X \in A)}{\mu(A)}$$

のように定める. ただし $\mu(A)$ は集合 A の体積である.

この関数を X の**確率分布**（あるいは $\mathcal{X}$ が離散の場合には**確率関数**, 連続の場合には**確率密度密度**）という．引数として確率変数の小文字を用いる場合添え字を省略し単に $p(x)$ と略記する．上の定義において, 特に離散集合の場合には $p_X(x) = P(X = x)$ となる.

4 条件付確率

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

事象 A を全事象とみたときの事象 $A \cap B$ の相対確率を表す. これを事象 B の事象 A に関する条件付き確率と定義する. また $P(A) = 0$ の場合 $P(B|A) = 0$ であり比は未定義であるが, $[0, 1]$ に属する任意数とする.

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

が成り立つ. 特に

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

のとき, 事象 A と B は独立であるという.

確率変数 $X : \Omega \to \mathcal{X}$ と $Y : \Omega \to \mathcal{Y}$ の同時確率分布を

$$p_{XY}(x, y) = \lim_{A \downarrow \{(x,y)\}} \frac{P((X, Y) \in A)}{\mu(A)}$$

[2]分布空間 $\mathcal{X}$ が離散集合の場合には, $X^{-1}(x), x \in \mathcal{X}$, がすべて事象であればよい. $\mathcal{X}$ が連続集合の場合には, ボレル集合族 $\mathcal{B}_{\mathcal{X}}$ が付随するが, $X^{-1}(B), B \in \mathcal{B}_{\mathcal{X}}$, がすべて事象であることが要請される.

と定義する. これは確率変数 $(X,Y):\Omega\to\mathcal{X}\times\mathcal{Y}$ の確率分布 $p_{(X,Y)}((x,y))$ の定義において, 余計な括弧とカンマを省いたものになっている. 引数から関数形が理解できるならば添え字を省き $p(x,y)$ と記述する. 特に, 離散確率変数の場合には, $p(x,y)=P(X=x,Y=y)$ である.

X の Y に関する条件付確率分布を

$$p_{X|Y}(x|y)=\frac{p_{XY}(x,y)}{p_Y(y)}$$

と定義する. これは離散の場合,

$$\frac{P(X=x,Y=y)}{P(Y=y)}$$

に他ならない. $p_{X|Y}(x|y)$ も $p(x|y)$ と略記する. $\mathcal{X}$ が離散集合の場合, 同時確率分布 $p(x,y)$ について, $p(x)=\sum_y p(x,y)$ を X に関する周辺分布という. Y に関する周辺分布 $p(y)$ も同様に定義される. $p(x,y)=p(x)p(y)$ のとき, 確率変数 X と Y は独立であるという.