

情報理論 (情報源符号化)

- 伝えるべき情報をより効率良く伝えるには
- 「効率の良さ」を計る
 - ★ 伝えるべき「情報の量」を計る
 - ★ 伝える為の「手間」を計る

→ **Shannon**

「情報の量は伝えるのに必要な手間と一致」

例: モールス符号 (Morse code)

1 文字のための符号長が区々

← 頻度の高い文字は短く、低い文字は長く

→ 頻度まで考慮して符号長の期待値を短く

… 頻度 (出現確率) を考慮して
符号効率の定式化を考える

各符号語の長さが異なると問題も生ずる

→ 一意復号可能か？

→ 一意復号可能としても瞬時復号可能か？

情報源符号化の定式化

情報源 alphabet S : 有限集合

$S^+ := \bigsqcup_{n \geq 1} S^n$: S の元の 1 個以上の列

$S^0 := \{\varepsilon\}$: 空語

$S^* := \bigsqcup_{n \geq 0} S^n$: S の元の 0 個以上の列

$$= S^+ \sqcup \{\varepsilon\}$$

$w \in S^n$ に対し、 $|w| := n$ (文字列の長さ)

情報源符号化の定式化

$$P : S \longrightarrow [0, 1] \subset \mathbf{R} : \underline{\text{生起確率}} \left(\sum_{s \in S} P(s) = 1 \right)$$

情報源 $S := (S, P)$

: 文字 $s \in S$ を確率 $P(s)$ で次々と発生
→ $w \in S^+$ を発生

(ここでの) 仮定:

各 $s \in S$ の生起確率は、 s のみで決まり、
先立って発生した文字に依らない。

情報源符号化の定式化

符号 (伝送) alphabet T : 有限集合
(しばしば $T = \{0, 1\}$)

$C : S \longrightarrow T^+$: 符号 (code)

→ 文字列を並べて $C^* : S^* \longrightarrow T^*$ に延長

$L(C) := \sum_{s \in S} P(s) |C(s)|$: C の 平均符号長

符号への要請

- 一意符号: $C^* : S^* \longrightarrow T^* : \text{単射}$
- 瞬時符号: $C(x) = C(s)w \implies x = sy$
(最初に届いた符号語で最初の文字が復元できる)
(以上は生起確率 P には依らない)
- 効率が良い... 平均符号長 $L(C)$ が小さい
(これは生起確率 P に依る)

瞬時符号の性質

- C : 瞬時符号 $\implies C$: 一意符号
- C : 瞬時符号 $\iff C$: 語頭符号
($C(s') = C(s)x \implies s' = s, x = \varepsilon$)

瞬時符号の作り方

「符号語木」を考えよう

瞬時符号の性質

- C : 瞬時符号 $\implies C$: 一意符号
- C : 瞬時符号 $\iff C$: 語頭符号
($C(s') = C(s)x \implies s' = s, x = \varepsilon$)

瞬時符号の作り方

「符号語木」を考えよう

Kraft の不等式

$$S = \{s_1, \dots, s_k\}, \quad \#T = r$$

自然数列 (ℓ_1, \dots, ℓ_k) に対し、

各符号語長 $|C(s_i)| = \ell_i$ なる
 r 元 瞬時符号 が存在

$$\iff \sum_{i=1}^k \frac{1}{r^{\ell_i}} \leq 1$$

McMillan の不等式

$$S = \{s_1, \dots, s_k\}, \quad \#T = r$$

自然数列 (ℓ_1, \dots, ℓ_k) に対し、

各符号語長 $|C(s_i)| = \ell_i$ なる
 r 元 一意符号 が存在

$$\iff \sum_{i=1}^k \frac{1}{r^{\ell_i}} \leq 1$$

母関数

数列 (a_n) から関数を作る

→ 解析的手法の利用

- $\sum_{n \geq 0} a_n X^n$: (通常の) 母関数
- $\sum_{n \geq 0} \frac{a_n}{n!} X^n$: 指数型母関数
- $\sum_{n \geq 1} \frac{a_n}{n} X^n$: 対数型母関数
- $\sum_{n \geq 1} \frac{a_n}{n^s}$: **Dirichlet** 級数

例: $k = 2$ の時

情報源の長さ 1:

$$\begin{array}{r|l} w_1 & a_1 X^{\ell_1} \\ w_2 & a_2 X^{\ell_2} \\ \hline & a_1 X^{\ell_1} + a_2 X^{\ell_2} \end{array}$$

情報源の長さ 2:

$$\begin{array}{r|l} w_1 w_1 & a_1^2 X^{2\ell_1} \\ w_1 w_2 & a_1 a_2 X^{\ell_1 + \ell_2} \\ w_2 w_1 & a_1 a_2 X^{\ell_1 + \ell_2} \\ w_2 w_2 & a_2^2 X^{2\ell_2} \\ \hline & (a_1 X^{\ell_1} + a_2 X^{\ell_2})^2 \end{array}$$

瞬時符号・一意符号の基本性質を見た。

符号の効率に移ろう。

(平均符号長を小さくする)

平均符号長の小さい符号の構成

→ Huffman 符号

瞬時符号・一意符号の基本性質を見た。

符号の効率に移ろう。
(平均符号長を小さくする)

平均符号長の小さい符号の構成

→ Huffman 符号