情報理論　#5

第5回講義　5月23日

情報科学研究科　　　井上純一

http://chaosweb.complex.eng.hokudai.ac.jp/~j_inoue/

情報源圧縮

(先週の復習)

ababbabaabbb という系列を ab|ab|ba|ba|aa|bb と2つずつに区切り、
次の表に従って01の並びからなる符号に変換する

アルファベット	符号語	符号長
aa	110	3
ab	0	1
ba	10	2
bb	111	3

この表による変換で

ababbabaabb 12ビット

→ 0010100111 10ビット

に圧縮される

対応表を逆にたどると元の系列
が復元できるので**可逆圧縮**である

復号

どの符号語も他の符号語の先頭部分になっていない
⇒ **語頭条件**がみたされている

クラフト不等式

(先週の復習)

$f: A \to B^{+}$は一意復号可能であるとし、$l(x)$をxに対する符号語$f(x)$の長さとすると

$$\sum_{x \in A} K^{-l(x)} \leq 1$$

が成り立つ。　KはBの要素数である。

適用例）

$$A = \{aa, ab, ba, bb\}, B = \{0,1\}$$

$$y_2(aa) = 00, y_2(ab) = 10, y_2(ba) = 11, y_2(bb) = 110$$

$$l(aa) = l(ab) = l(ba) = 2, l(bb) = 3, K = 2$$

のときは

$$\sum_{x \in A} K^{-l(x)} = 2^{-2} + 2^{-2} + 2^{-2} + 2^{-3} = \frac{7}{8} < 1$$

一意復号可能

クラフト不等式の証明 #1

(先週後回しにした部分)

J を任意の正の整数として、次が成り立つことに注意する

$$\left(\sum_{x\in A} K^{-l(x)}\right)^J = \boxed{\sum_{x_1\in A}\sum_{x_2\in A}\cdots\sum_{x_J\in A} K^{-(l(x_1)+l(x_2)+\cdots l(x_J))}}$$

$A_J(n)$ を全体の長さが n になるようなJ 個の符号語の並べ方の総数とすると

$$A_J(n) = \#\left\{x_1 x_2 \cdots x_J \in A^J : l(x_1)+l(x_2)+\cdots l(x_J)=n\right\}$$

であり、ある符号語に対し、次が成り立つ

J個の符号語を並べた符号長の最大値

$$\boxed{\sum_{n=1}^{Jl_{\max}} d\left(l(x_1)+\cdots+l(x_J),n\right)=1}$$

1をかける

クラフトの不等式の証明 #2

先週後回しにした部分）

$$\left(\sum_{x\in A} K^{-l(x)}\right)^J = \sum_{n=1}^{Jl_{\max}} \sum_{x_1\in A} \cdots \sum_{x_J\in A} d\left(l(x_1)+\cdots+l(x_J), n\right)K^{-n} = \sum_{n=1}^{Jl_{\max}} A_J(n)K^{-n}$$

$\forall_{a,a'} \in A$に対して$f(a) \neq f(a')$となるためには

$$A_J(n) \leq K^n$$

Bの記号をn個並べてできる語の総数

であることが必要。

よって

$$\left(\sum_{x\in A} K^{-l(x)}\right)^J = \sum_{n=1}^{Jl_{\max}} A_J(n)K^{-n} \leq \sum_{n=1}^{Jl_{\max}} 1 = Jl_{\max}$$

$$\longrightarrow \quad \sum_{x\in A} K^{-l(x)} \leq \left(Jl_{\max}\right)^{1/J} = 1 \ (J \to \infty)$$

情報源符号の平均符号長

$$\underline{A=\{a_1,a_2,\cdots,a_K\}},f:A\mapsto \underline{B^+=\{b_1,b_2,\cdots,b_M\}}$$

記号系列　　　　符号語

l_i ：**符号語**$f(a_i)$**の長さ**

p_i: **記号**a_i**の出現確率**$p(a_i)$

$$\boxed{L=\sum_{i=1}^{K} p_i l_i}$$

平均符号長

例）

出現確率

aa	00	1/4
ab	10	1/4
ba	01	1/4
bb	110	1/4

$$L=2\times(1/4)+2\times(1/4)+2\times(1/4)+3\times(1/4)=9/4$$

平均符号長

平均符号長の下限

$$L - H_M(X) = -\sum_{i=1}^{K} p_i \log_M M^{-l_i} + \sum_{i=1}^{K} p_i \log_M p_i$$

平均符号長

情報源のエントロピー

$$= -\sum_{i=1}^{K} p_i \log_M c r_i + \sum_{i=1}^{K} p_i \log_M p_i$$

$$= \boxed{\sum_{i=1}^{K} p_i \log\left\{\frac{p_i}{r_i}\right\}} - \log_M c \geq 0$$

KL情報量で正の値を持つ

確率の規格化条件

$$\sum_{i=1}^{K} r_i = c^{-1} \sum_{i=1}^{K} M^{-l_i} = 1$$

より

$$c = \boxed{\sum_{i=1}^{K} M^{-l_i} \leq 1}$$

クラフト不等式より

$$\boxed{L \geq H_M(X)}$$

一意復号可能な場合には、平均符号長をエントロピーよりも小さくすることができない

クラフト不等式からの考察

記号a_iを符号化したものの長さを次のようにおく

$$\boxed{l_i = \left\lceil \log_M \frac{1}{p_i} \right\rceil}$$

括弧内以上の最小な整数

$$\sum_{i=1}^{K} M^{-l_i} \le \sum_{i=1}^{K} M^{-\left\lceil \log_M \frac{1}{p_i} \right\rceil}$$

$$\le \sum_{i=1}^{K} M^{-\log_M \frac{1}{p_i}} = 1$$

クラフト不等式を満たす

$$\log_M \frac{1}{p_i} \le l_i < \log_M \frac{1}{p_i} + 1$$

$$\Rightarrow \sum_{i=1}^{K} p_i \log \frac{1}{p_i} \le \sum_{i=1}^{K} p_i l_i < \sum_{i=1}^{K} p_i \log \frac{1}{p_i} + 1$$

エントロピー　平均符号長　エントロピー

$$\boxed{H_M(X) \le L < H_M(X) + 1}$$