情報理論 配布資料 #5

担当：井上 純一 (情報エレクトロニクス系棟 8-13)

平成 16 年 5 月 24 日

演習問題 3 に関するコメント

(2) での「定常的単純マルコフ情報源」の場合のエントロピーレートの計算で, 教科書 p .26 の式 (2・34) を用いた人が多数いましたが, この式 :

$$\mathcal{H}(X) \quad = \quad -\sum_{x\in\mathcal{X}} P(x) \sum_{x'\in\mathcal{X}} \log P(x^{'}|x) \log P(x^{'}|x)$$

自体が間違っています (誤植です). 正しくは前回配布の解答にあるように

$$\mathcal{H}(X) \quad = \quad -\sum_{x\in\mathcal{X}} P(x) \sum_{x'\in\mathcal{X}} P(x^{'}|x) \log P(x^{'}|x)$$

であり, こうなる理由は直前の式である条件付きエントロピー $H(X_2|X_1)$ の定義からも明らかでしょう.

答案レポートの中には教科書の間違いに気づいてはいるのでしょうが,「教科書を信じればこうなる」と, 但し書きをしてから解答を書いて下さった方もいました. いくら教科書 (私の今書いているこの資料もそうですが •••) とは言え, 誤植や間違いは必ずありますから, 教科書と自分の答えが違ったときでも, ある程度は自分の答えを信じて書くことが重要ではないでしょうか. なお, 教科書ではエントロピーレートの導出に「エントロピーのチェイン則」を用いていますが, 解答例にも示したように, エントロピーに寄与する結合確率 $P(x_1, x_2, \cdots, x_n)$ に立ち返って考えてみればチェイン則を用いないでも計算できます (もちろん, $P(x_1, x_2, \cdots, x_n)$ の構造自体が複雑になり過ぎれば, どのような方法であれ, 解析的に計算するのは極めて困難となります).

演習問題 4 の解答例

1.

(1) クラフトの不等式が満たされているかどうかを調べればよい. $K = 3, l_1 = 1, l_2 = 2, l_3 = l_4 = 3$ であるから, クラフトの不等式は

$$3^{-1} + 3^{-2} + 3^{-3} + 3^{-3} \quad = \quad \frac{1}{3} + \frac{1}{9} + \frac{1}{27} + \frac{1}{27} = \frac{14}{27} < 1 \tag{1}$$

となり満たされている. よってこの符号は一意復号可能である.

(2) 例えば $\underline{x_1 = 0}$ と選ぶと, 語頭条件から $l_2 = 2$ である x_2 として $01, 00, 02$ の 3 つは使えない. 従って, $3^2 = 9$ 通りの可能性の中で, x_2 として用いることのできるのは $10, 11, 12, 20, 21, 22$ の 6 通りで

ある. ここでは $\underline{x_2 = 20}$ と選ぶことにしょう. 符号長が 3 である x_3, x_4 に対しては語頭条件のために $3^3 = 27$ 通りの可能性の中で 0 を先頭に持つ $000, 001, 002, 010, 011, 012, 020, 021, 022$, 及び, 20 を先頭に持つ $200, 201, 202$ の計 12 通りの符号は用いることができない. これ以外の符号長 3 の符号として例えば $\underline{x_3 = 100}$, $\underline{x_4 = 101}$ のように選ぶことができる. よって上記を表にまとめれば

x_1	0
x_2	20
x_3	110
x_4	101

となる.

(3) 符号の木を描くと図 1 のようになる.

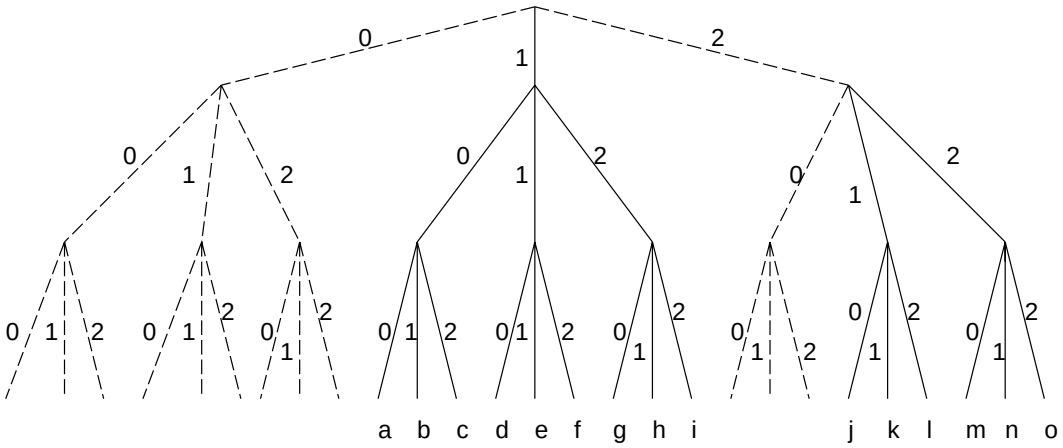

図 1: この問題での符号語に対する符号の木. $x_1 = 0, x_2 = 20$ のように符号を選んだ場合, x_3, x_4 として採用できる長さ 3 の符号としては図の a〜o の中のいずれかとなる. (破線でたどられる符号を用いることはできない)

2.

(1) 問題文中に与えた表に従って平均符号長 L, 及び, エントロピー H を計算すると

$$\begin{aligned} L &= \sum_{x=\mathsf{aa,ab,ba,bb}} p(x)l(x) \\ &= p(\mathsf{aa})l(\mathsf{aa}) + p(\mathsf{ab})l(\mathsf{ab}) + p(\mathsf{ba})l(\mathsf{ba}) + p(\mathsf{bb})l(\mathsf{bb}) = \frac{9}{4} \qquad (2) \\ H &= -\sum_{x=\mathsf{aa,ab,ba,bb}} p(x)\log p(x) = \log 4 = 2 \qquad (3) \end{aligned}$$

となる.

(2) 情報源アルファベットの生成確率を $p(\mathsf{aa}) = p(\mathsf{ab}) = p(\mathsf{ba}) = p, p(\mathsf{bb}) = 1 - 3p$ のように p を用いて表現する場合, (1) と同様にして平均符号長, エントロピーとしてぞれぞれ

$$\begin{aligned} L &= 3 - 3p \qquad (4) \\ H &= -3p\log p - (1-3p)\log(1-3p) \qquad (5) \end{aligned}$$

が得られる. 従って, 両者の差である p の関数 Φ は

$$\begin{aligned}\Phi &= L - H \\ &= 3 - 3p + 3p\log p + (1-3p)\log(1-3p) \end{aligned} \tag{6}$$

であり, Φ を最小化する p の値は $(\partial\Phi/\partial p) = 0$ より

$$\log\left\{\frac{p}{1-3p}\right\} = 1 \tag{7}$$

すなわち, $(p/1-3p) = 2$, つまり, $p_* = 2/7$ となり, このときの Φ の値として

$$\Phi(p_*) = 3 - \log 7 \simeq \underline{0.19} \tag{8}$$

が得られる. この値は, (1) で調べた全ての情報源アルファベットが等確率で現れる場合の $\Phi(1/4) = 0.25$ と比べて小さくなっていることがわかる.

(3) $p(\mathsf{bb}) = 1/7$ であるから, 生成確率 (出現確率) の低い情報源アルファベットには長い符号を割り振り, 逆に, 生成確率の高いアルファベットには短い符号を与えることにより, 平均符号長をエントロピーに近づけることができる. Φ は非負であるから (つまり, どんなに頑張っても平均符号長はエントロピーより小さくできない), この手続きにより平均符号長を短くすることができる[1]

この例では 4 つの情報源アルファベットの生成確率を p で表し, それを Φ を最小化するという意味で最適化したが, 実際に個々の情報源アルファベットに生成確率が割りあてられた場合 (何らかの方法によりそれらの確率が事前に計測できた場合), いかにして符号を構成するかに関しては, 今回の講義で説明する**ハフマン符号**が有効な方法の一つとして知られている (このハフマン符号は「平均符号長を最小にする」という意味で最適な符号化法 (圧縮法) である).

演習問題 5

1. N_k を長さが k である符号化系列の総数としよう.

情報源アルファベット	符号語
x_1	0
x_2	10
x_3	11

表 1： この問題で考える符号.

[1] **演習問題 2** の 2. でみたように, n 値エントロピー関数が最大となるのは n 個の事象が等確率で現れる場合であったから, 全ての情報源アルファベットが等確率で生じるという仮定の下でエントロピーは最大となっており, 平均符号長はこの最大エントロピーを下回らないわけだから, 非常に長い値を持つことになる. 情報源アルファベットの出現確率が偏り始め, ある文字が出現しやすい状況になるとエントロピーが減少し始める. そしてその分だけ平均符号長の下限も低くなる. この状況下で適切な戦略の下に符号化すればその「下限」に一致させる最適符号を構成することができる.

表 1 の場合には

$$\begin{aligned} N_1 &= 1 \quad (x_1) \\ N_2 &= 3 \quad (x_1x_1, x_2, x_3) \\ N_3 &= 5 \quad (x_1x_1x_1, x_1x_2, x_1x_3, x_2x_1, x_3x_1) \end{aligned}$$

となる (括弧内は実際にその長さを与える情報源アルファベットの組み合わせ). このとき以下の問いに答えよ. ただし, この問題で考えるのは全て表 1 で与えられる符号であるとする.

(1) w_n を表 1 における長さ n の符号語の個数とする. つまり, $w_1 = 1, w_2 = 2, w_n = 0\,(n \geq 3)$ である. このとき, N_k を N_{k-1}, N_{k-2}, 及び, w_1, w_2 の中から必要なものを用いて表せ. ただし $k \geq 3$ とする.

(2) (1) で得られた N_k に関する漸化式の解として

$$N_k = \lambda^k$$

を仮定する. このとき (1) で得られた漸化式を λ に関する方程式に書き直せ.

(3) 初期条件 : $N_0 = N_1 = 1$ のもとで (2) で得られた方程式を解くことにより, N_k を求めよ. また, 得られた N_k の正当性をチェックするために, 長さ $k = 4$ の符号化系列を与える情報源アルファベットの組み合わせを全て列挙せよ.

2. 次の表 2 に与えた情報源アルファベット, 及び, その生成確率に対してハフマン符号を構成せよ. また, 得られるハフマン符号の平均符号長を求めよ.

情報源アルファベット	生成確率
x_1	0.2
x_2	0.18
x_3	0.10
x_4	0.10
x_5	0.10
x_6	0.061
x_7	0.059
x_8	0.04
x_9	0.04
x_{10}	0.04
x_{11}	0.04
x_{12}	0.03
x_{13}	0.01

表 2 : ハフマン符号を考えるアルファベットとその生成確率の表.