

情報理論09

藤田 一寿

津山工業高等専門学校情報工学科 講師
電気通信大学先進理工学科 協力研究員

情報源のエルゴード性

統計力学におけるエルゴード仮説

時間平均と集団(アンサンブル)平均は一致する

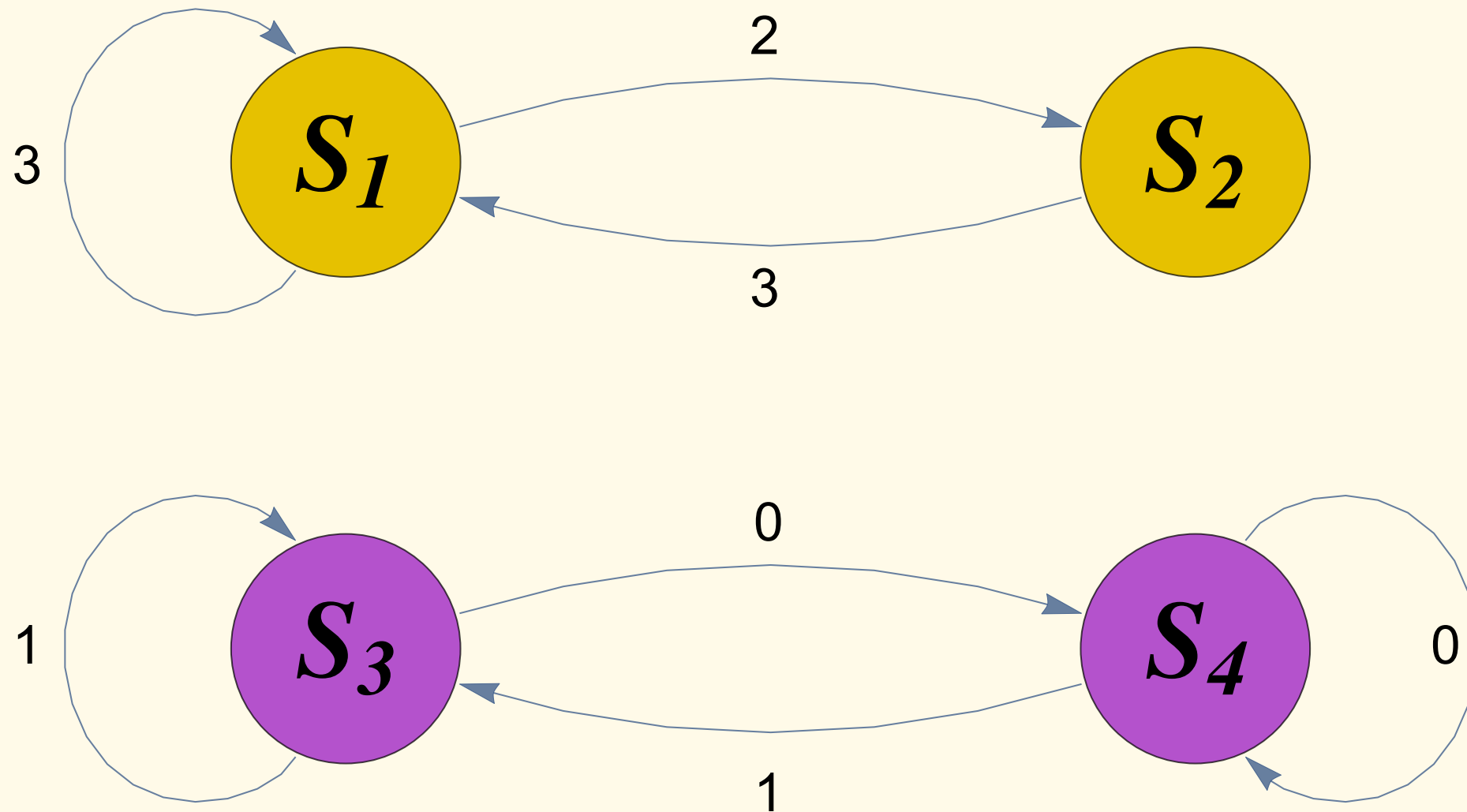
サイコロを100万回ふった時の平均と、
100万個のサイコロをふった時の平均は
一致する。

エルゴード的情報源

- ・ 閉部分集合で、小さな閉部分集合を含まない。
 - ・ ある状態になる確率が0とならない。
- ・ この閉部分集合の測度(確率)だけ1であって他の集合の測度(確率)は0である。
- ・ 非周期的である。
- ・ この閉部分集合に含まれる文字列をエルゴード系列と呼ぶ。

非周期的で分離不可能なマルコフ情報源はエルゴード的である。エルゴード的情報源であるならば、時間平均(エントロピー)と集団平均(エントロピー)は等しくなる。

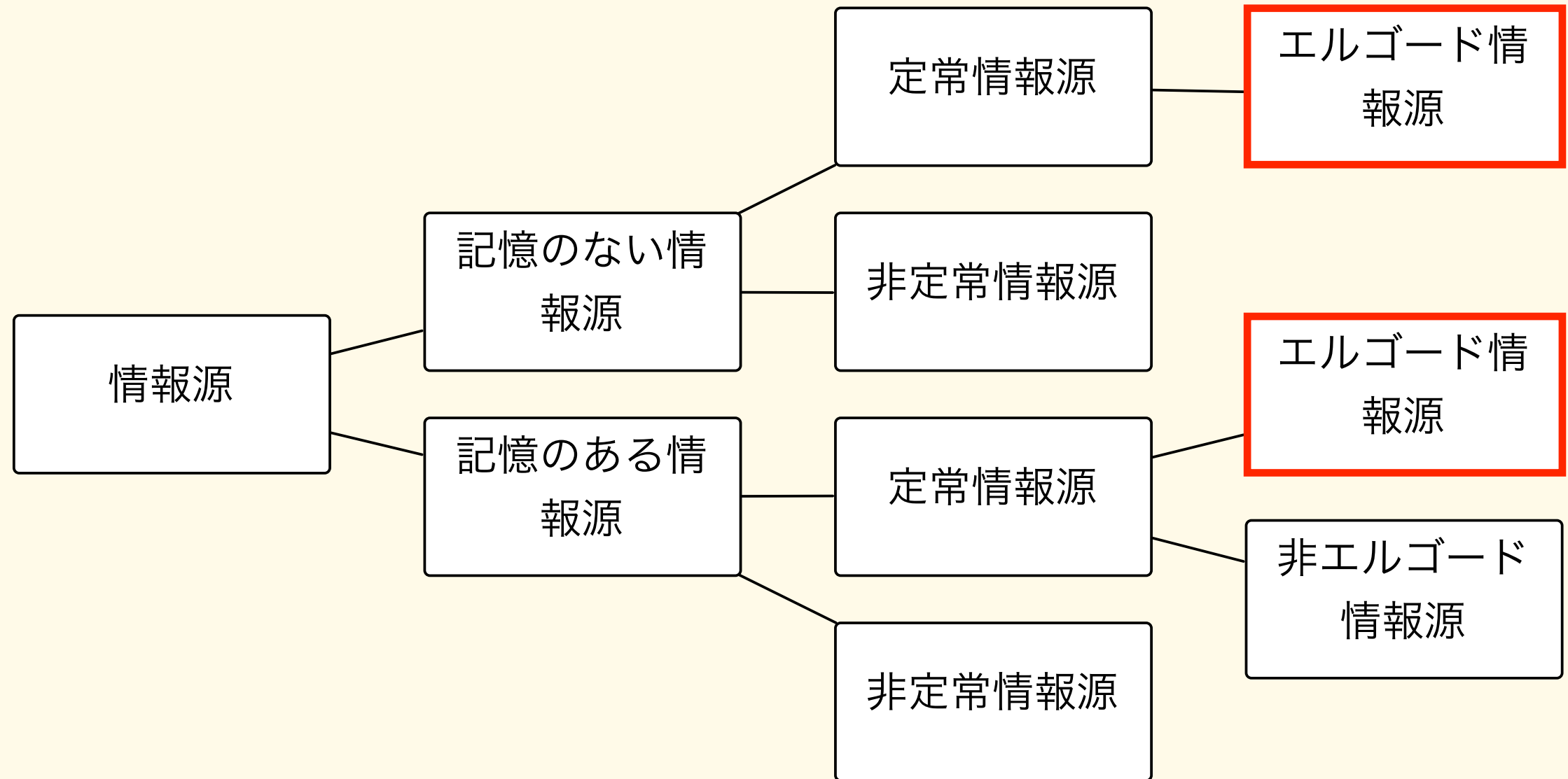
エルゴード的ではない情報源



4つの文字0, 1, 2, 3を発生させるマルコフ情報源

この図で S_1 か S_2 の状態にある確率と S_3 と S_4 の状態にある確率は等しいとする。

情報源の分類



情報源の冗長度

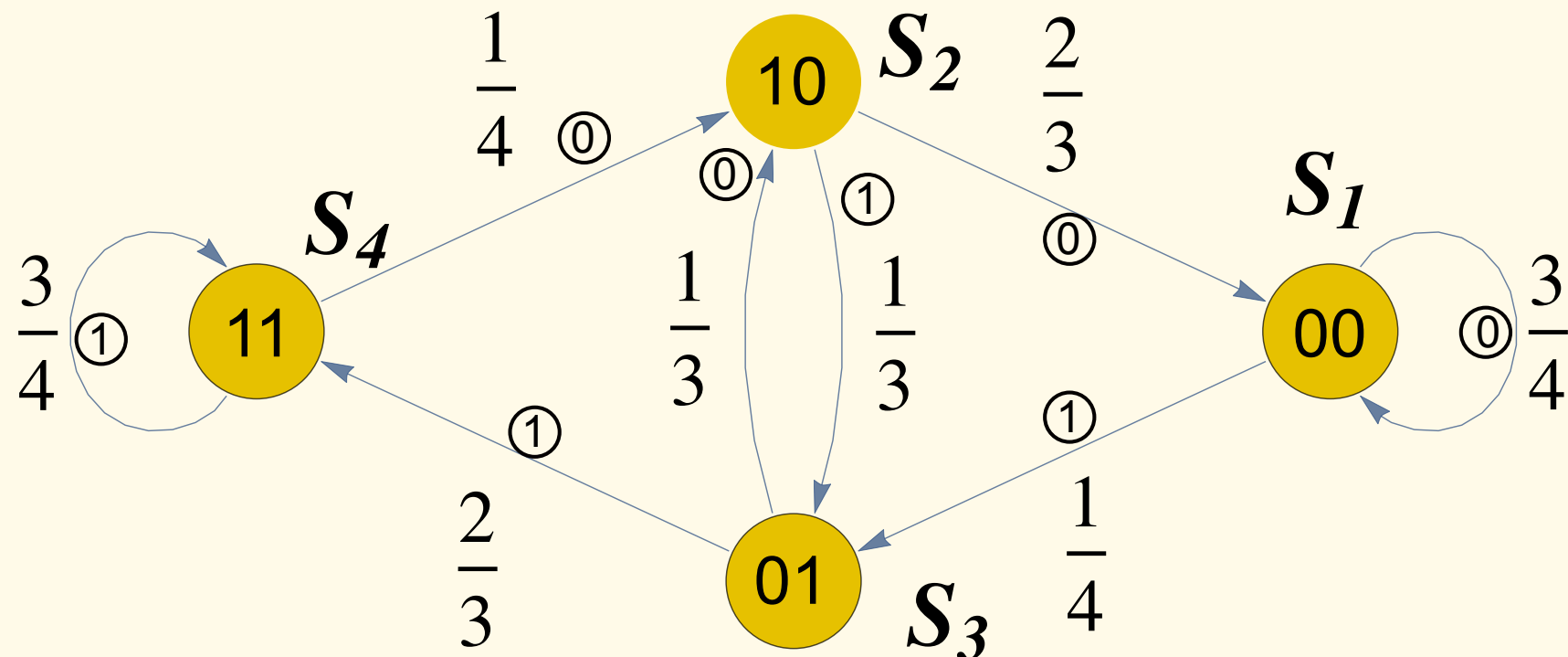
冗長

- ・ 必要以上に物事が多く無駄なこと、長いこと、またはその様子。
- ・ wikipediaより

情報源のエントロピー

情報源のエントロピーは、次に出てくる文字のエントロピーの期待値であるので、

$$\begin{aligned} H &= \frac{3}{22} \times 0.92 \times 2 + \frac{8}{22} \times 0.81 \times 2 \\ &= 0.84 \end{aligned}$$

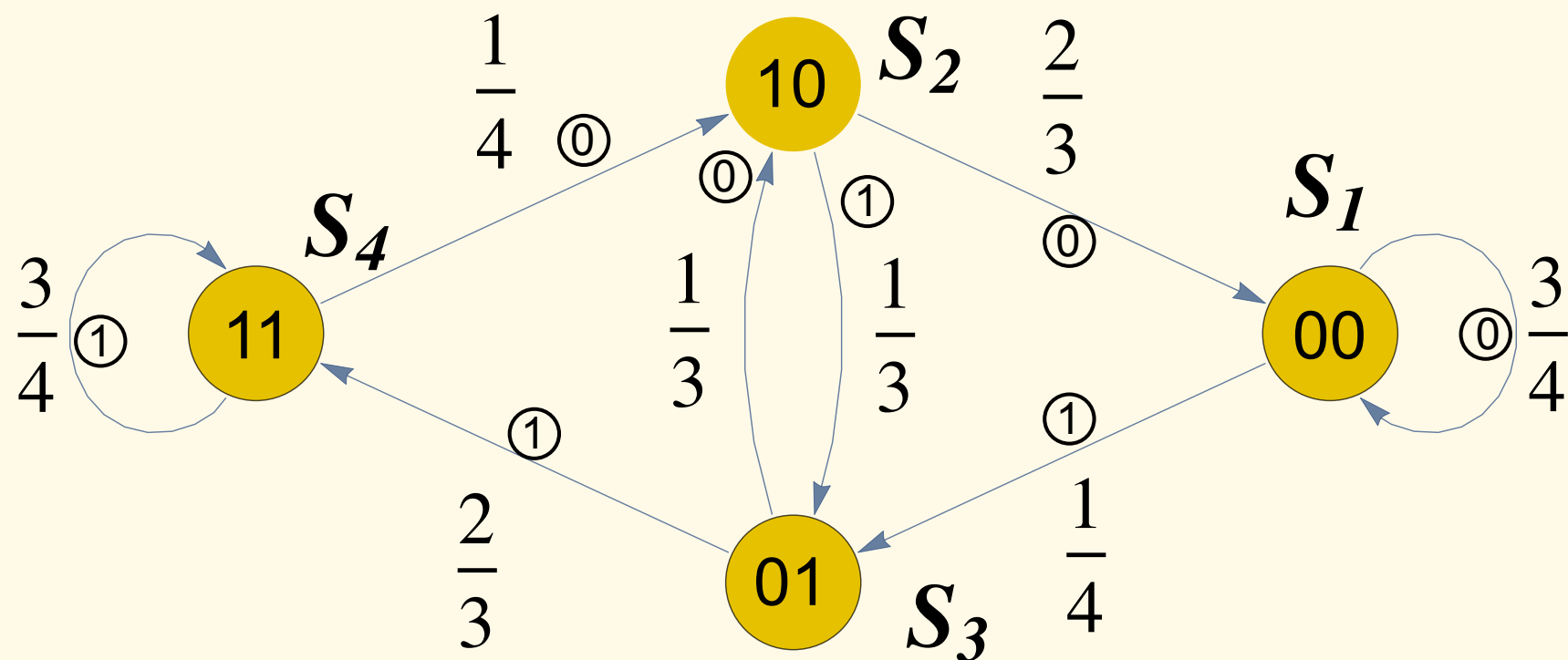


文字の発生確率が独立な場合との比較

この情報源は、実は0と1が出る確率は等しい。

しかし、エントロピーは1ビットではない。

前後の文字の発生確率が独立かつ確率が等しい場合の
2元情報源のエントロピーは1ビットになる



この情報源は前後の文字の発生確率は独立ではないため、1ビットよりエントロピーが小さくなる。

0が出た場合、次も0になりやすい。

1が出た場合、次も1になりやすい。

言い方を変えると、00となりやすいので、0が出たあとに0が発生しても得られる情報は少ないと言える。

その分、エントロピーが文字の前後関係が独立な場合より少ない。

情報源の最大エントロピー

A_1, \dots, A_k の k 個の文字を持っている情報源は、以前に発生した文字が何であろうと関係なく次の文字が出る場合最大のエントロピーを持つ。この時のエントロピーは

$$\begin{aligned} H_0 &= -\frac{1}{k} \log \frac{1}{k} \times k \\ &= \log k \end{aligned}$$

となる。

冗長度

実際のエントロピーを H とすると $H_0 - H$ 分情報源が持つことができる最大のエントロピーより少ないということになる。逆に考えると、情報源は実際に吐き出す文字のエントロピーに比べ、余分にエントロピーを持つことが可能であると解釈できる(冗長であると考えられる)。

そこで $H_0 - H$ と H_0 の比を冗長度 r と定義する。

$$r = \frac{H_0 - H}{H_0}$$

冗長度を考察する

例で用いた情報源の持つことのできる最大のエントロピーは1ビット

実際は0.84ビット

冗長度は0.16

これは、最も効率のよい情報源を用いれば文字数を16%減らすことができる(圧縮できる)ことを示す。

情報源の大数の法則

文字列の発生確率

情報源が k 種類の文字を発生させるとすると、文字列の長さが N 個であった場合、その文字列は全部で k^N 個ある。

文字列 $x_1x_2\dots x_N$ が発生する確率を

$$p(x_1x_2\dots x_N)$$

と書く。

文字が発生する確率

もし、過去に $\dots x_1 x_2 \dots x_{i-1}$ という文字列が発生したとすると、次に発生する文字 x_i の発生確率は

$$p(x_i | \dots x_1 x_2 \dots x_{i-1})$$

と書ける。よって x_i が発生した時に得られる情報量は

$$I_i = -\log p(x_i | \dots x_{i-2} x_{i-1})$$

となる。

では I_i の期待値は

$$\bar{I}_i = - \sum_{x_i} p(x_i | \dots x_{i-2} x_{i-1}) \log p(x_i | \dots x_{i-2} x_{i-1})$$

ここまではある文字列が出てきたことが前提

さらに、これまで出た文字列について平均すると、それが情報源の1文字あたりのエントロピーになる。

$$H = - \sum_{\dots x_{i-1}} \sum_{x_i} p(\dots x_{i-1}) p(x_i | \dots x_{i-1}) \log p(x_i | \dots x_{i-1})$$

過去のすべての文字列がわかっているとき

過去の文字列を x^∞ と表すと、文字列 $x_1 x_2 \dots x_N$ が生成される確率は

$$p(x_1 x_2 \dots x_N | x^\infty) = p(x_1 | x^\infty) p(x_2 | x^\infty) \dots p(x_N | x^\infty)$$

と書ける。よって、情報量 I は

$$\begin{aligned} I &= -\log(p(x_1 | x^\infty) p(x_2 | x^\infty) \dots p(x_N | x^\infty)) \\ &= \sum_{i=1}^N I_i \end{aligned}$$

I_i はだいたい I_i 期待値である H と仮定すると

$$I \simeq NH$$

となることが期待できる。

$$I = -\log p = NH$$

エントロピーHは情報量の期待値

つまり、文字列の長さNを非常に大きくすれば情報量の平均がエントロピーに近づくと考えられる。

$$\frac{I}{N} \rightarrow H$$

$$\begin{aligned} I &= -\log(p(x_1|x^\infty)p(x_2|x^\infty)\dots p(x_N|x^\infty)) \\ &= \sum_{i=1}^N I_i \end{aligned}$$

$$I \simeq (H \pm \varepsilon)N$$

文字列の出現確率をpとすると

$$I = -\log p \simeq (H \pm \varepsilon)N$$

ある小さな正数 ε を考え,長さ N の文字列のうちで出現確率 p が

$$(H - \varepsilon)N < -\log p < (H + \varepsilon)N$$

$$\left| -\frac{\log p}{N} - H \right| < \varepsilon$$

を満たす文字列を $L1$ とし,それ以外のものを $L2$ とする.

$L1$ に属する文字列は出現確率が

$$-\log p \simeq NH$$

$$p \simeq 2^{-NH}$$

となる文字列がほとんどとなる。

任意の $\varepsilon > 0$ 、 $\delta > 0$ に対して、ある N_0 が存在し、 $N \geq N_0$ なら、 $\left| -\frac{\log p}{N} - H \right| < \varepsilon$ に属する文字列の発生する確率を $1-\delta$ より大きくすることができる。

$$\text{Prob} \left\{ \left| \frac{-\log p}{N} - H \right| \geq \varepsilon \right\} < \delta$$

L2に属する文字列
(数が多いがあまり出てこない
文字列)

$$p \neq 2^{-NH}$$

L1に属する文字列
(数は少ないがよく出
る文字列)

$$p \simeq 2^{-NH}$$

文字列全体