

情報理論07

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

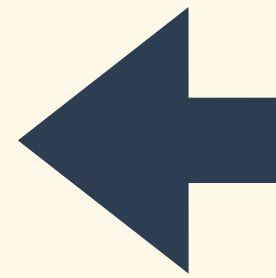
情報は湧いてくる

- ▶ 天気は、毎日変わっていく。
 - ▶ 晴れ、曇、晴れ、晴れ、曇、雨、晴れ、…
- ▶ 文章も一文字一文字が続いて出てくる。
- ▶ 事象が次々に起こるとき、その事象の発生を規則を確率的に考えることができる。事象と呼ぶのは面倒なので文字と呼ぶことにする。

情報源

- ▶ つまり、どこかから情報が次々に提供されると考えることができる。
- ▶ そして、提供される情報は、過去の情報に依存して確率的に決まる。
- ▶ そのような、情報が出てくる源を情報源と呼ぶ。

$A_3 A_2 A_3 A_1 A_6 A_5 \dots$



情報源

$A_1, A_2, A_3, \dots, A_k$

A_1 から A_k までの文字を持っている情報源から、文字が発生する。この文字の集合をアルファベットとよぶ。

時刻 t ($t = 1, 2, 3, \dots$)に発生する文字を x_t と表すと、その時刻に出た文字が A_i なら

$$x_t = A_i$$

と表される。

M元情報源

情報源のアルファベットが

$$\mathbf{A} = (a_1, a_2, \dots, a_M)$$

で表されるとき、M元情報源という。

2進信号：2元情報源

サイコロ：6元情報源

Tさんが情報源の場合

今日もLinux日和だ。

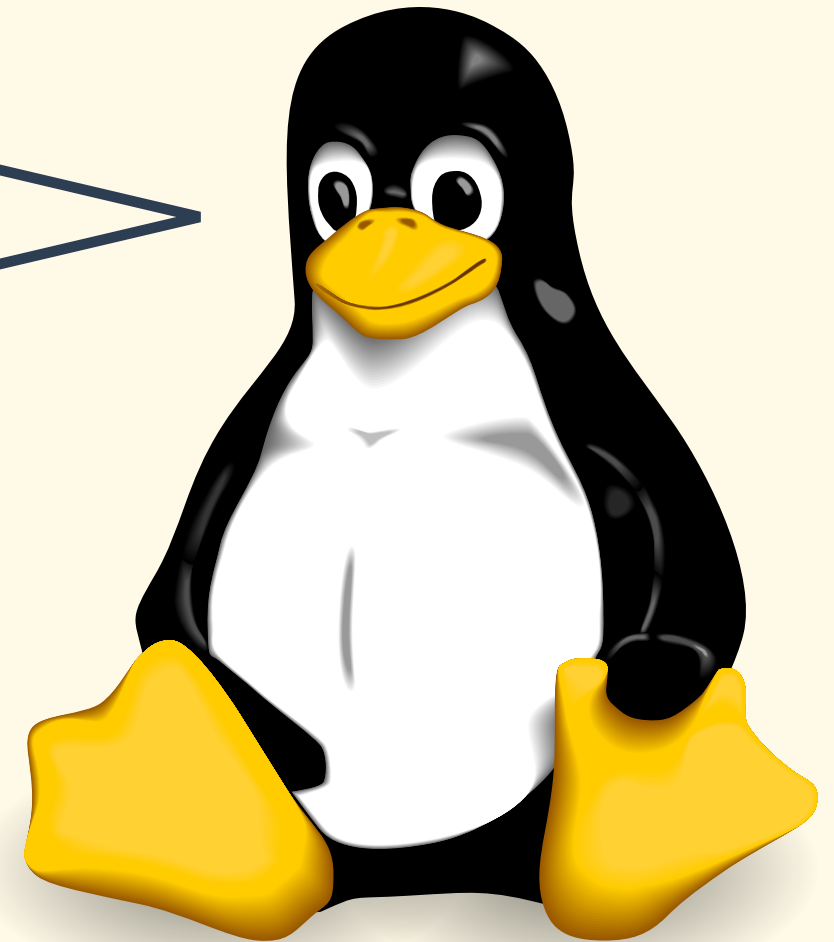
x1=今

x2=日

・

・

・



Tさん

文字で区切るか単語で区切るかは問題設定によって変わる。

発生する文字列は独立に出てくるのではなく、以前出た文字列に依存すると考える。

- ・ 曇の日の次の日は雨がふるかもしれない。
- ・ エントロピーという文字列がよく出てくる講義に、
「クーロンの法則」という文字列は現れにくい。

時刻 t に発生する文字 x_t が、以前出た文字列に $\cdots x_{t-2}x_{t-1}$ 依存するとする。以前出た s 個の文字の系列が

$$x_{t-s} \cdots x_{t-2}x_{t-1}$$

であるとき、 x_t の出る確率は

$$p(x_t | x_{t-s} \cdots x_{t-1})$$

と書ける。ここで文字列 $x_{t-s}\cdots x_{t-2}x_{t-1}$ が出る確率を

$$p(x_{t-s} \cdots x_{t-1})$$

と書くとする、文字列 $x_t\cdots x_{t-1}$ が出る確率は

$$\begin{aligned}
p(x_{t-s} \dots x_{t-1} x_t) &= p(x_t | x_{t-s} \dots x_{t-1}) p(x_{t-s} \dots x_{t-1}) \\
&= p(x_t | x_{t-s} \dots x_{t-1}) p(x_{t-1} | x_{t-s} \dots x_{t-2}) p(x_{t-s} \dots x_{t-2}) \\
&= p(x_t | x_{t-s} \dots x_{t-1}) p(x_{t-1} | x_{t-s} \dots x_{t-2}) \dots p(x_{t-s+1} | x_{t-s}) p(x_{t-s})
\end{aligned}$$

と書ける。

情報源からは、逐次確率的に文字がされており、その生成する確率は過去に生成された文字が何であったかに依存する。このような情報源を記憶情報源という。

定常的な情報源

情報源から出る文字の発生確率分布が時間によらず変化しない情報源を定常的な情報源という。

情報源の性質が時間がたっても変わらないということは $p(x_t | x_{t-s} \dots x_{t-1})$ は t が何であってても同じだということである。

$$p(x_{t+r}) = p(x_t)$$

$$p(x_{t+r} | x_{t+r-s} \dots x_{t+r-1}) = p(x_t | x_{t-s} \dots x_{t-1})$$

Dさんの場合

朝



デーモンが好きです。ペンギンは嫌いです。

夜



デーモンが好きです。ペンギンも大好きです。

デーモン君が定常情報源ならば、時間によらずデーモン君はペンギンのあとに嫌いと言っていると考えられる。

情報源のエントロピー

時刻 t における情報源から生成される文字のエントロピーは

$$H(X_t) = - \sum p(x_t) \log p(x_t)$$

と表せる。 X_t は時刻 t の文字 x_t をあらわす確率変数である。
情報源は定常的であると考えると1文字のエントロピーは

$$H(X) = - \sum_i p(x = A_i) \log p(x = A_i)$$

となる。情報源から過去に出た文字を考えなければこれで良いが...

情報源から出てきた2文字のエントロピー

$$H(X_1 X_2) = - \sum_{x_1 x_2} p(x_1 x_2) \log p(x_1 x_2)$$

情報源から出てきたn文字のエントロピー

$$H(X^n) = - \sum_{x^n} p(x^n) \log p(x^n)$$

$$X^n = X_1 X_2 \dots X_n$$
$$x^n = x_1 x_2 \dots x_n$$

$H(X_n)$ の平均は

$$H_n = \frac{H(X^n)}{n}$$

1文字のエントロピーは

$$H = \lim_{n \rightarrow \infty} \frac{H(X^n)}{n}$$

$n \rightarrow$ 無限大の極限を取らないと期待値にならない

無記憶情報源

- ▶ 過去に発生した文字が、それ以降に発生する文字に影響しないような情報源を無記憶情報源と言う。
- ▶ 各文字の発生確率は互いに独立(i.i.d.)である。
- ▶ 文字列 $x_1 \cdots x_t$ が発生する確率は次のように書き表せる。

$$p(x_1 x_2 x_3 \cdots x_t) = p(x_1) p(x_2) p(x_3) \cdots p(x_t)$$

- ▶ 1文字のエントロピーは

$$H(X) = - \sum_x p(x) \log p(x)$$