情報理論
第 5 回 情報量とエントロピー

堀田 政二
工学部 情報工学科

情報量とは

- 聞いて非常に驚く情報 $\cdots$ 情報量が大きい (人が犬を噛む)
- 聞いても驚かない情報 $\cdots$ 情報量が小さい (犬が人を噛む)

これを数学的に表現することを考える．例えば生起確率 (発生確率) が $p(a)$ の事象 a が実際に起きたとき，これを知ることによって得られる情報量を

$$I(a) \propto \frac{1}{p(a)}$$

と定義したとしよう．すなわち，情報量は生起確率に反比例する．

- $p(a)$ が小さい $\cdots$ $I(a)$ が大きい
- $p(a)$ が大きい $\cdots$ $I(a)$ が小さい

自己情報量 (self information)

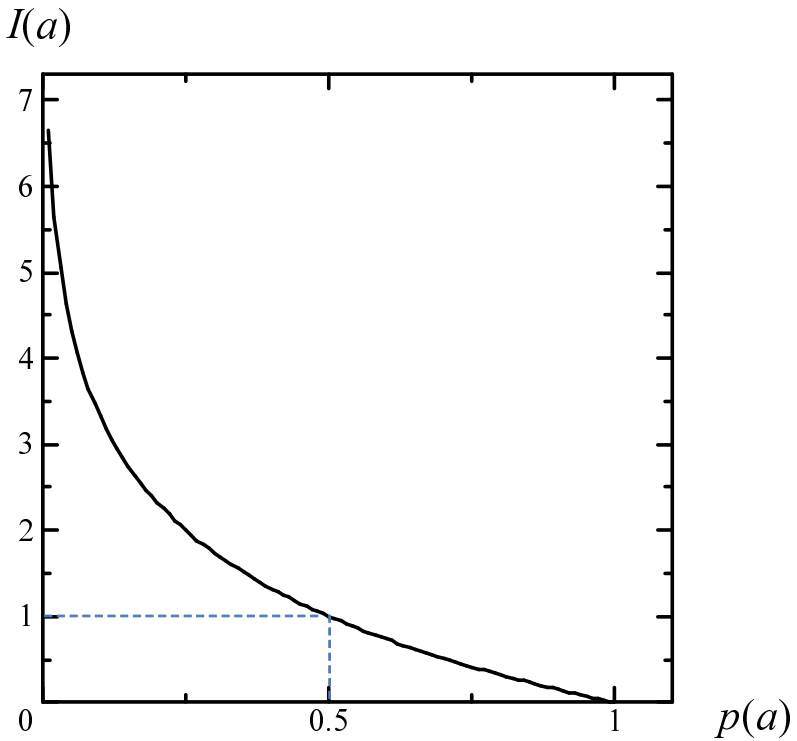
この定義では $p(a) = 1$ ならば $I(a) = 0$ とはならない．つまり必ず起きる事象が起きた時の驚きは $I(a) = 0$ であるにも関わらず，この定義では $I(a)$ が ∞ に近づいてしまう．そこで右辺の対数をとったものを事象 a の自己情報量と定義する．

自己情報量

$$I(a) = \log_2 \frac{1}{p(a)} = -\log_2 p(a)$$ 単位は bit

情報理論では，対数の底として通常 2 $(\log_2)$ を用いる

自己情報量のグラフ



- $p(a) = 1/2$ **の時，** $I(a) = 1$
- $p(a) = 1$ **の時，** $I(a) = 0$

自己情報量の例

【例 1】赤ん坊が生まれたとき，その男女比が $1:1$ とする．男が生まれる事象を boy**，女が生まれる事象を** girl **とすると，それぞれの自己情報量は下記の通りになる**:

- $I(\text{boy}) = -\log_2 \frac{1}{2} = 1$ bit
- $I(\text{girl}) = -\log_2 \frac{1}{2} = 1$ bit

【例 2】ある試験では合格する可能性が $1/8$ である．この，試験に合格した場合の自己情報量は

- $I(\textbf{合格}) = -\log_2 \frac{1}{2^3} = 3$ bit

となる．一方，不合格になった時の自己情報量は

- $I(\textbf{不合格}) = -\log_2 \frac{7}{8} = -\log_2 7 + \log_2 8 = -2.807 + 3 = 0.193$ bit

となる．

情報量の加法性

ある事象 E は二つの事象 E_1 と E_2 の積だとする．この時，事象 E の自己情報量は

$$I(E) = I(E_1) + I(E_2)$$

となる．例えば，ジョーカーを除いた 52 枚のトランプを相手に引いて貰い，その内容を教えてもらうことを考える．

- 引いたカードが ♠ の A であることを知ったときの情報量は
  $I(♠ \cap A) = -\log_2 \frac{1}{52} \sim 5.7$ bit
- 引いたカードが ♠ であることのみを知ったときの情報量は
  $I(♠) = -\log_2 \frac{1}{4} = 2$ bit
- 引いたカードが A であることのみを知ったときの情報量は
  $I(A) = -\log_2 \frac{1}{13} \sim 3.7$ bit
- したがって $I(♠ \cap A) = I(♠) + I(A)$

対数の計算 (復習)

1. $\log_a b = c \Leftrightarrow b = a^c$ (**対数の定義**)
2. $\log_a b = \dfrac{\log_{10} b}{\log_{10} a}$
3. $\log_a(xy) = \log_a(x) + \log_a(y)$
4. $\log_a \dfrac{x}{y} = \log_a(x) - \log_a(y)$
5. $\log_a x^y = y \log_a x$
6. $-\log_a \dfrac{1}{x} = \log_a x$ (**式 5 で** $y = -1$ **のとき**)

なお， $\log_2 x$ **を計算するには，式 2 を利用すれば良い．すなわち**

$$\log_2 x = \log_{10} x / \log_{10} 2 = \log_{10} x / 0.3010 \sim 3.3223 \times \log_{10} x$$

平均情報量 (average information)

情報量の平均 (期待値) について考えよう．いま，ある事象系 A を $A = \{a_1, a_2, ..., a_n\}$ とする．これら n 個の事象は互いに排反で，その生起確率 $p(a_i)$ の総和は 1 とする (完全事象系)．情報量 $I(a_i)$ の期待値を $H(A)$ とすると

$$H(A) = \sum_{i=1}^{n} p(a_i) I(a_i) = -\sum_{i=1}^{n} p(a_i) \log_2 p(a_i)$$

簡単のために $p(a_i)$ を p_i と略記すると

平均情報量

$$H(A) = -\sum_{i=1}^{n} p_i \log_2 p_i \quad \text{bit}$$

平均情報量の性質

- **平均情報量の取りうる値は** $0 \leq H(A) \leq \log_2 n$ bit
- **事象系** A **のうち，一つの事象** a_i **の生起確率が** $p(a_i) = 1$ **で，その他の事象の生起確率がすべて** 0 **の時，** $H(A) = 0$**．これは結果を聞く前から結果が既知なので驚き** 0**．**
- **事象系** A **のすべての事象の生起確率が** $p(a_i) = 1/n$ **と一様の場合は平均情報量は最大の** $H(A) = \log_2 n$ **となる．これはどれが起きるか全く予想できない状態．**

平均情報量の例

小金井の 8 月 1 日の天気の生起確率が以下の通りだとする:

$$
\begin{aligned}
p(\textbf{晴}) &= \frac{1}{4}, \quad p(\textbf{雨}) = \frac{1}{2} \\
p(\textbf{曇}) &= \frac{1}{4}, \quad p(\textbf{雪}) = 0
\end{aligned}
$$

この時の平均情報量を求めると

$$
\begin{aligned}
H(A) &= -\sum_{i=1}^{4} p_i \log_2 p_i \\
&= -\frac{1}{4}\log_2\frac{1}{4} - \frac{1}{2}\log_2\frac{1}{2} - \frac{1}{4}\log_2\frac{1}{4} - 0\log_2 0 \\
&= \frac{2}{4} + \frac{1}{2} + \frac{2}{4} - 0 = 1.5 \text{ bit}
\end{aligned}
$$

ただし， $x \to 0$ **のとき** $x\log_2 x \to 0$

エントロピー (entropy)

熱力学における分子の無秩序さを表す尺度

熱力学におけるエントロピー

$$H = -K \sum_k n_k \ln n_k$$

ここで，K はボルツマン定数，n_k は気体分子の k 番目のエネルギー状態にある確率．

情報理論におけるエントロピー

$$H = -\sum_i p_i \log_2 p_i$$

熱力学におけるエントロピーと平均情報量は定数倍，対数の底を除いて一致する．そのため，平均情報量を (情報) エントロピーと呼ぶことにする．

(11)

エントロピーの例

- ある日の K 市の天気予報が
- 晴 40%，曇 30%，雨 30% の時，エントロピーは

$$H = -0.4 \log_2 0.4 - 0.3 \log_2 0.3 - 0.3 \log_2 0.3 = 1.57 \text{ bit}$$

- 晴 100% のとき

$$H = -1.0 \log_2 1.0 - 0 \log_2 0 - 0 \log_2 0 = 0 \text{ bit}$$

晴れが 100% のときは結果が一つに決まっているのでエントロピーは 0，すなわち曖昧さがない

最大エントロピー (maximum entropy)

エントロピーが最大になるのはどのような場合かを考える．二つの事象からなる事象系 (2 元事象系) を次のように表す:

$$\mathbf{A} = \begin{pmatrix} a_1 & a_2 \\ p_1 & p_2 \end{pmatrix}$$

- **一行目は二つの互いに排反な事象を表し，どちらか一方の事象のみが起きる**
- **二行目は各事象の生起確率** $(p_1 + p_2 = 1)$

この場合のエントロピーは

$$H = -p_1 \log_2 p_1 - p_2 \log_2 p_2$$

となり， $p_1 + p_2 = 1$ **という制約条件のもとで** H **の最大値を求めるにはラグランジュの未定乗数法を使って解けばよい:**

$$L = -p_1 \log_2 p_1 - p_2 \log_2 p_2 + \lambda(1 - p_1 - p_2)$$

$\partial L/\partial p_i = -\log_2 p_i + 1 - \lambda = 0$，$\partial L/\partial \lambda = 1 - p_1 - p_2 = 0$ **の連立方程式を解けば，** $\log_2 p_1 = \log_2 p_2$ **のときエントロピーが** $H_{\max} = -\log_2 \frac{1}{2} = 1$ bit **と最大になることが分かる．** (13)

ラグランジュ未定乗数法

制約条件のもとで関数の極値を求める方法の一つ．

問題設定

制約条件 $g(\boldsymbol{x}) = 0$ のもとで関数 $f(\boldsymbol{x})$ の極値を求めよ

- ラグランジュ乗数 λ を用いてラグランジュ関数を導入

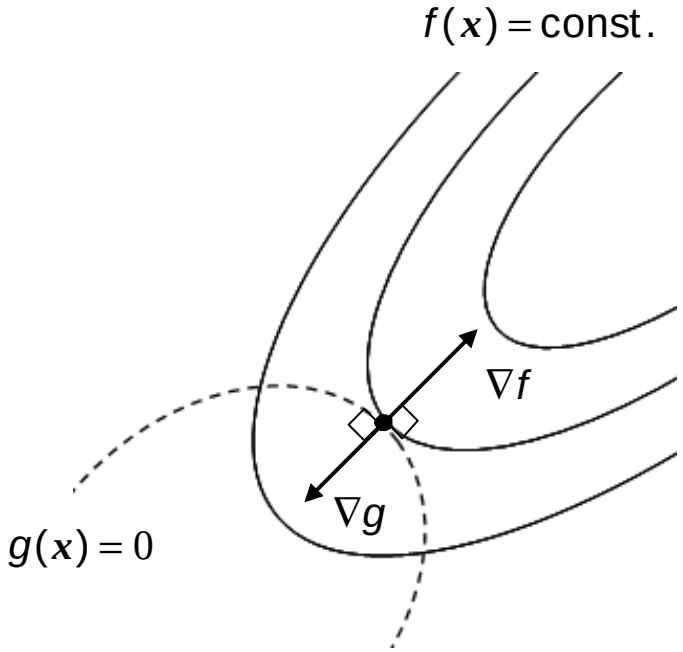
$$L = f(\boldsymbol{x}) - \lambda g(\boldsymbol{x})$$

- 制約条件のもとで関数が極値をとる点は次式を満たす：

$$\frac{\partial}{\partial \boldsymbol{x}} L = \nabla f - \lambda \nabla g = \mathbf{0}, \ \frac{\partial L}{\partial \lambda} = 0$$

- 上記から $d+1$ 個の方程式が得られる．一方，未知数は $x_1,\ldots,x_d, \lambda$ の $d+1$ 個なので，方程式の解を求めることができる

ラグランジュ未定乗数法の直感的な理解



二変数の場合の例

- **制約条件 $g(x) = 0$ と $f(x)$ の等高線の法線ベクトルが極値で平行**

$$\nabla f = \lambda \nabla g \tag{15}$$

n 次元事象系の場合の最大エントロピー

2 元事象系を一般化した n 次元事象系における最大エントロピーを考える．

$$\mathbf{A} = \begin{pmatrix} a_1 & a_2 & \cdots & a_n \\ p_1 & p_2 & \cdots & p_n \end{pmatrix}$$

この場合のエントロピーは

$$H = -\sum_{i=1}^{n} p_i \log_2 p_i$$

2 元事象系の場合と同様にしてラグランジュの未定乗数法を使って解けばよい:

$$L = -\sum_{i=1}^{n} p_i \log_2 p_i + \lambda \left(1 - \sum_{i=1}^{n} p_i\right)$$

$\partial L/\partial p_i = -\log_2 p_i + 1 - \lambda = 0$，$\partial L/\partial \lambda = 1 - \sum_{i=1}^{n} p_i = 0$ の連立方程式を解けば，$p_1 = p_2 = \cdots = p_n$ のときエントロピーが $H_{\max} = -\log_2 \frac{1}{n}$ bit と最大になることが分かる．

最大エントロピーの例

以下の例はいずれも各事象の生起確率が等確率と仮定する．

- **サイコロを一回振る時の最大エントロピー**

$$H_{\max} = -\sum_{i=1}^{6} \frac{1}{6} \log_2 \frac{1}{6} = \log_2 6 = 2.585 \text{ bit}$$

- **英数字** (A∼Z **と空白，計** 27 **文字**) **の最大エントロピー**

$$H_{\max} = -\sum_{i=1}^{27} \frac{1}{27} \log_2 \frac{1}{27} = \log_2 27 = 4.755 \text{ bit}$$

- **常用漢字** (1945 **文字**) **の最大エントロピー**

$$H_{\max} = -\sum_{i=1}^{1945} \frac{1}{1945} \log_2 \frac{1}{1945} = \log_2 1945 = 10.925 \text{ bit}$$

(17)

エントロピー関数 (entropy function)

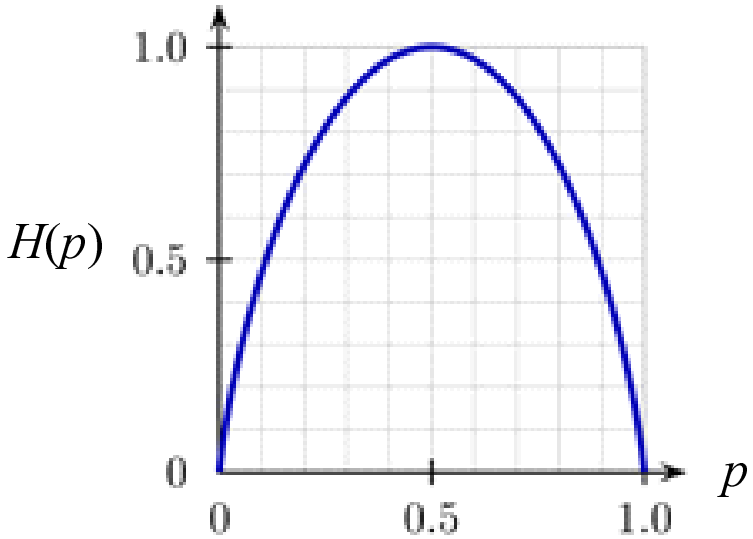
2 元事象系のエントロピー

$$H = -p_1 \log_2 p_1 - p_2 \log_2 p_2$$

において，$p_1 = p$，$p_2 = 1 - p$ と置くと

$$\mathcal{H}(p) = -p \log_2 p - (1-p) \log_2 (1-p)$$

となる．この関数 $\mathcal{H}(p)$ をエントロピー関数と呼ぶ．



エントロピー関数の利用法

ある試験を受けて，合格する確率が A 君は 0.6 **(不合格の確率** 0.4**)，B 君は** 0.9 **(不合格の確率** 0.1**) の場合，それぞれのエントロピーは**

- $\mathcal{H}(\text{A 君}) = \mathcal{H}(0.6) = \mathcal{H}(0.4) \sim 0.971$ bit
- $\mathcal{H}(\text{B 君}) = \mathcal{H}(0.9) = \mathcal{H}(0.1) \sim 0.496$ bit

B 君の方が合格する可能性が高いので，曖昧さ，不確実さは A 君より少なくなる

結合エントロピー (joint entropy)

二つの事象系を考える:

$$\boldsymbol{A} = \begin{pmatrix} a_1 & a_2 \\ p(a_1) & p(a_2) \end{pmatrix} \quad \boldsymbol{B} = \begin{pmatrix} b_1 & b_2 \\ p(b_1) & p(b_2) \end{pmatrix}$$

二つの事象系 A **と** B **が同時に起きる事象を結合事象系と呼び，** $A \otimes B$**，または単に** AB **と表す**:

$$\boldsymbol{AB} = \begin{pmatrix} (a_1, b_1) & (a_1, b_2) & (a_2, b_1) & (a_2, b_2) \\ p_{11} & p_{12} & p_{21} & p_{22} \end{pmatrix}$$

ただし， $(a_i, b_j) = a_i \cap b_j$**，** $p_{ij} = p(a_i \cap b_j)$ **とする．このとき** AB **の平均情報量**

$$H(AB) = -\sum_i \sum_j p_{ij} \log_2 p_{ij}$$

を結合エントロピーと呼ぶ．

(20)

条件付きエントロピー (conditional entropy)

結合エントロピー $H(AB)$ を変形すると

$$
\begin{aligned}
H(AB) = & - \sum_i \sum_j p(a_i \cap b_j) \log_2 p(a_i \cap b_j) \\
= & - \sum_i \sum_j p(a_i) p(b_j|a_i) \log_2 p(a_i) p(b_j|a_i) \\
= & - \sum_i \sum_j p(a_i) p(b_j|a_i) \{\log_2 p(a_i) + \log_2 p(b_j|a_i)\} \\
= & - \sum_i \sum_j p(a_i) p(b_j|a_i) \log_2 p(a_i) \\
& - \sum_i \sum_j p(a_i) p(b_j|a_i) \log_2 p(b_j|a_i) \\
= & - \sum_i p(a_i) \log_2 p(a_i) \sum_j p(b_j|a_i) \\
& - \sum_i p(a_i) \sum_j p(b_j|a_i) \log_2 p(b_j|a_i)
\end{aligned}
$$

$\sum_j p(b_j|a_i) = 1$ **であるため，第 1 項は** $H(A)$ **である．**

(21)

条件付きエントロピーの続き

一方，第 2 項の $\sum_j p(b_j|a_i)\log_2 p(b_j|a_i)$ は条件 a_i のもとでの b_j のエントロピーである．したがって，第 2 項全体はその a_i に関する平均値であることから $H(B|A)$ を意味している．すなわち，

$$H(B|A) = -\sum_i p(a_i) \sum_j p(b_j|a_i)\log_2 p(b_j|a_i)$$

この $H(B|A)$ を条件付きエントロピーと呼ぶ．
以上から，$H(AB)$ は $H(AB) = H(A) + H(B|A)$ と書けることが分かる．同様にして $H(AB) = H(B) + H(A|B)$ も示すことができることから $H(AB) = H(BA)$ である．

シャノンの基本不等式 (Shannon's fundamental inequality)

条件付きエントロピーに関して次の関係が成り立つ:

シャノンの基本不等式

$$H(A|B) \leq H(A), \quad H(B|A) \leq H(B)$$

上式は，情報を得る前よりも，情報を得た後の方がエントロピーは小さい (曖昧さが減少する) ことを意味している．例えば

- A: **雨が降るという事象**
- B: **台風が接近しているという事象**

とすると B **を知れば** A **が起きるであろうことは，より確実に予想可能になる．この不等式と** $H(AB) = H(A) + H(B|A)$ **から**

$$H(AB) = H(A) + H(B|A) \leq H(A) + H(B)$$

なる関係が導かれる．等号が成り立つ場合は A **と** B **が独立の時．**

(23)

シャノンの基本不等式の続き

不等式

$$H(AB) = H(A) + H(B|A) \leq H(A) + H(B)$$

において，等号が成り立つ場合は A **と** B **が独立の時．例えば**

- A: **犬が子供を産むという事象**
- B: **台風が接近しているという事象**

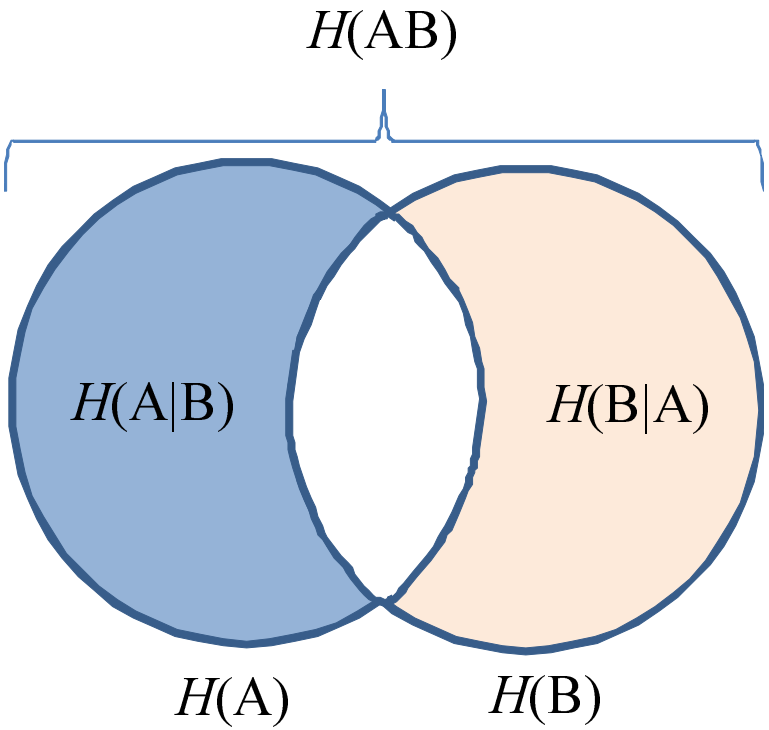
の場合には，A **と** B **の事象は互いに独立なので等号が成り立つ．**
これまでの議論をまとめると

$$0 \leq H(A|B) \leq H(A) \leq H(AB)$$

なる関係が成り立つ．

(24)

各種エントロピーの関係



$$H(AB) = H(A) + H(B|A) = H(B) + H(A|B)$$

(25)

問題

- 【5.1】ある都市のある日の天気予報が，晴 45%，曇 35%，雨 12%，雪 8% のとき，エントロピー H を小数第 2 位まで求めよ．
- 【5.2】平仮名 48 文字の生起確率がすべて等しいと仮定した場合の平均情報量を小数第 3 位まで求めよ．
- 【5.3】A 君が 3 年後に大学を卒業できる確率は 75%，A 君の父が 3 年後に会社で重役になれる確率を 30% とする．このとき，二つの事象の結合エントロピーを求めよ．