講義「情報理論」

第3回　情報量とエントロピー

情報理工学部門 情報知識ネットワーク研究室
喜田拓也

講義資料

2019/6/21

今日の内容

2.1 情報量

2.2 エントロピー

2.3 エントロピーの性質

2.4 結合エントロピー

2.5 条件付きエントロピー

2.6 相互情報量

情報には量がある！

確率が高いことを知らされても，
そのニュースは価値が低い

私には一人，妹がいます

フーン（´<_`　）

妹は女性です

で？

確率1の結果が知らされる → 得られる情報量は 0

情報には量がある！

確率が低いことを知らされたら，
そのニュースは価値が高い

私には12人，妹がいます

(;゜Д゜)(゜Д゜;(゜，゜;)
ナ、ナンダッテー!!

確率が 0 に近い事柄を知らされる → 情報量は大！

一つの結果を知ったときの情報量

確率pの事象の生起を知ったときに得られる情報量を$I(p)$とすると$I(p)$は次のような性質を満たすべき

① $I(p)$は$0 < p \leq 1$で単調減少な関数である

② 確率p_1, p_2で起こる二つの互いに独立な事象が同時に起こる確率$p_1 p_2$について$I(p_1 p_2) = I(p_1) + I(p_2)$

③ $I(p)$は$0 < p \leq 1$で連続な関数である

情報量の加法性

これらを満たす関数$I(p)$は

$$I(p) = \boxed{}$$

という形しかありえない（ただし$a > 1$）

証明は省略

情報量の定義

定義2.1

確率 p で生起する事象が起きたことを知ったときに得られる情報量 $I(p)$ を自己情報量と呼び，

$$I(p) = -\log_a p$$

と定義する．ただし，a は $a > 1$ の定数とする．

$a = 2$ の場合，単位はビット（bit）という

自然対数で計るときはナット（nat）　1 nat≒1.443 bit

10を底とする対数で計るときはハートレー（Hartley）

もしくはディット（dit）またはデシット（decit）　1 Hartley≒3.322 bit

確率1/2で生じる結果を知ったときの情報量 ＝1 [bit]

簡単な例題：　サイコロを1回振ったときの出目を知ったときに得られる情報量は何ビットか答えよ．ただし，サイコロの各出目が得られる確率はすべて等しく1/6とする．

平均情報量

定義2.2

M個の互いに排反な事象$a_1, a_2, \cdots, a_M$が起こる確率を$p_1, p_2, \cdots, p_M$とする（ただし，$p_1 + p_2 + \cdots + p_M = 1$）．
このうち1つの事象が起こったことを知ったときに得る情報量は$-\log_2 p_i$であるから，これを平均した期待値$\bar{I}$は，

$$\bar{I} = p_1(-\log_2 p_1) + p_2(-\log_2 p_2) + \cdots + p_M(-\log_2 p_M)$$

$$= \boxed{}$$

となる．これを平均情報量（単位はビット）という．

エントロピー

定義2.3

確率変数X がとりうる値が$x_1, x_2, \cdots, x_M$ とし，X がそれぞれの値をとる確率が$p_1, p_2, \ldots, p_M$（ただし，$p_1 + p_2 + \cdots + p_M = 1$）であるとき，確率変数 X のエントロピーを

$$H(X) = -\sum_{i=1}^{M} p_i \log_2 p_i$$

ビットと定義する.

例題2.1：　偏りのないコインを2回投げて表の出た枚数を確率変数 X とする. このとき，X のエントロピー$H(X)$ は何ビットか？

$$H(X) = -\frac{1}{4}\log_2\frac{1}{4} - \frac{1}{2}\times\log_2\frac{1}{2} - \frac{1}{4}\log_2\frac{1}{4}$$

$$= 2\times\frac{2}{4} + \frac{1}{2} = 1.5 \quad (\text{ビット})$$

X	0	1	2
確率	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

Try 練習問題2.1

エントロピーの性質

定理2.1

M個の値をとる確率変数Xのエントロピー$H(X)$は次の性質を満たす.

(1) [　　]

(2) $H(X)$が最小値0となるのは，ある値をとる確率が1で，他の$M-1$個の値をとる確率がすべて0のときに限る．すなわち，Xのとる値が初めから確定している場合のみである.

(3) $H(X)$が最大値$\log_2 M$となるのは，M個の値がすべて$1/M$で等しい場合に限る.

エントロピー関数

定義2.4

エントロピー関数とは，$0 \leq x \leq 1$ で定義される関数

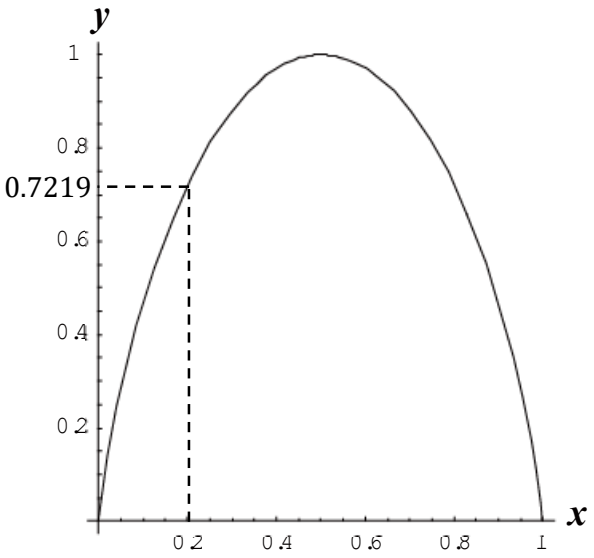
$$\mathcal{H}(x) = -x \log_2 x - (1-x) \log_2 (1-x)$$

のことをいう.

二つの値$1, 0$ をそれぞれ $0.2, 0.8$ の確率でとる確率変数Xのエントロピー$H(X)$は,

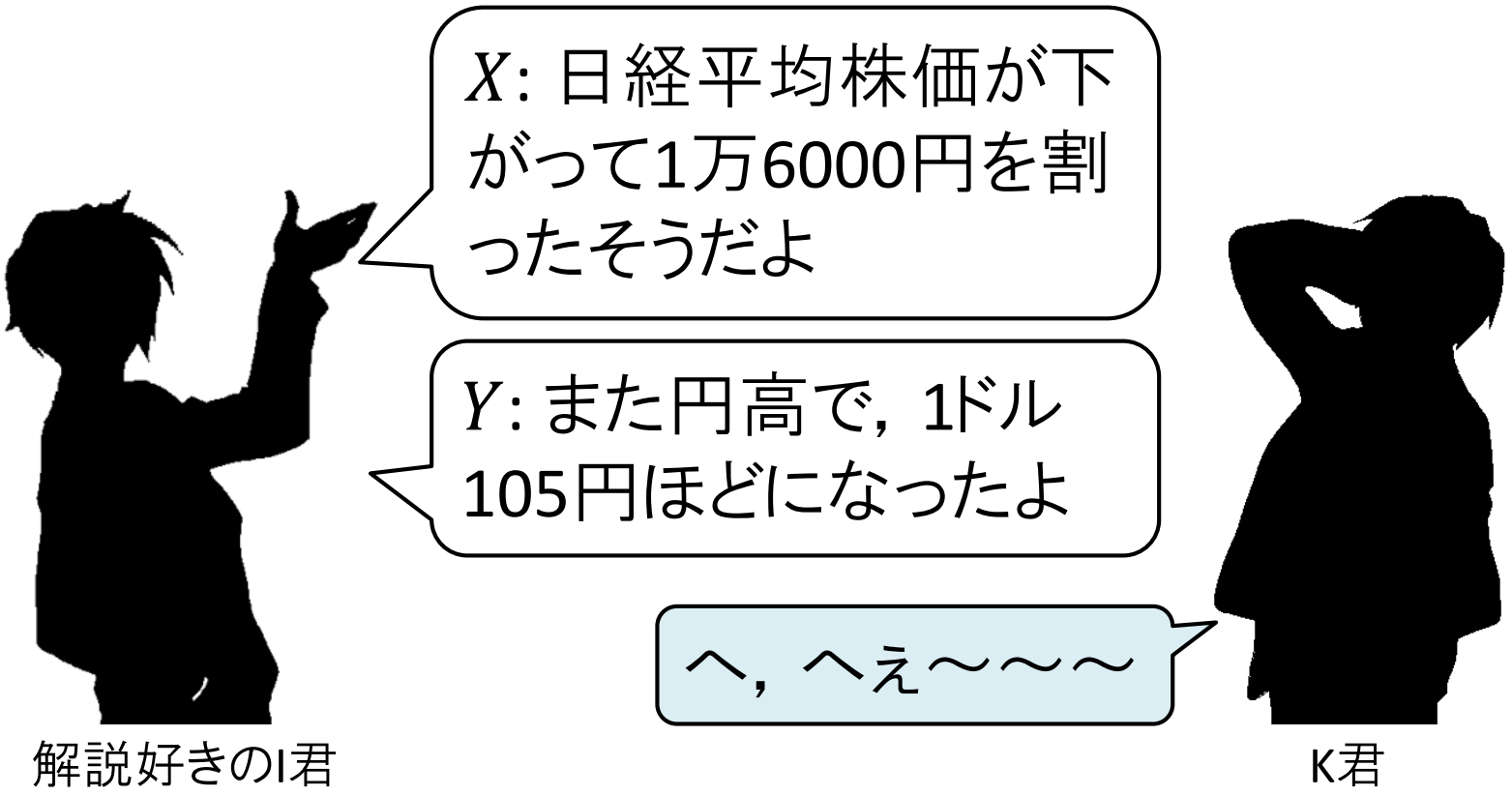
$$\begin{aligned} H(X) &= -\sum_{i=1}^{2} p_i \log p_i \\ &= -0.2 \log 0.2 - 0.8 \log 0.8 \\ &= \mathcal{H}(0.2) \\ &\cong 0.7219 \end{aligned}$$

となる.



エントロピー関数

二つの情報を一度に聞いたときの情報量は？



はたして，$H(\,(X,Y)\,) = H(X) + H(Y)$ だろうか？

例 2.2

二つの確率変数X, Yを考える. Xは$x_1, x_2, \ldots, x_{M_X}$の値をとり, Yは$y_1, y_2, \ldots, y_{M_Y}$の値をとるものとする. 確率変数の組$(X, Y)$の値が$(x, y)$となる結合確率分布を$P(x, y)$と書く

表2.1 ある日の天気Xとコンビニのアイスクリームの売上高Yの結合確率分布$P(x, y)$

$P(x, y)$		Y		$P(x)$
		1万円以上	1万円未満	
X	晴	0.5	0.1	0.6
	雨	0.2	0.2	0.4
$P(y)$		0.7	0.3	

組(X, Y)をまとめて考えると, 4つの値をとる確率変数 Z のエントロピー $H(Z)$ として考えることができる

結合エントロピー

定義2.5

確率変数XとYの結合エントロピー$H(X,Y)$は,

$$H(X,Y) = -\sum_{i=1}^{M_X}\sum_{j=1}^{M_Y} P(x_i, y_j)\log_2 P(x_i, y_j)$$

により定義される. これを結合エントロピーと呼ぶ. ただし,

$\{x_1, x_2, \ldots, x_{M_X}\}$および$\{y_1, y_2, \ldots, y_{M_Y}\}$は, それぞれ$X$と$Y$が取りうる値の集合とする.

表2.1から, (X,Y) の結合エントロピーは,

$$\begin{aligned} H(X,Y) &= -0.5 \times \log 0.5 - 0.1 \times \log 0.1 \\ &\quad -0.2 \times \log 0.2 - 0.2 \times \log 0.2 \\ &\fallingdotseq 1.76 \ \text{(ビット)}. \end{aligned}$$

Try 練習問題2.2

結合エントロピーの性質

定理2.2

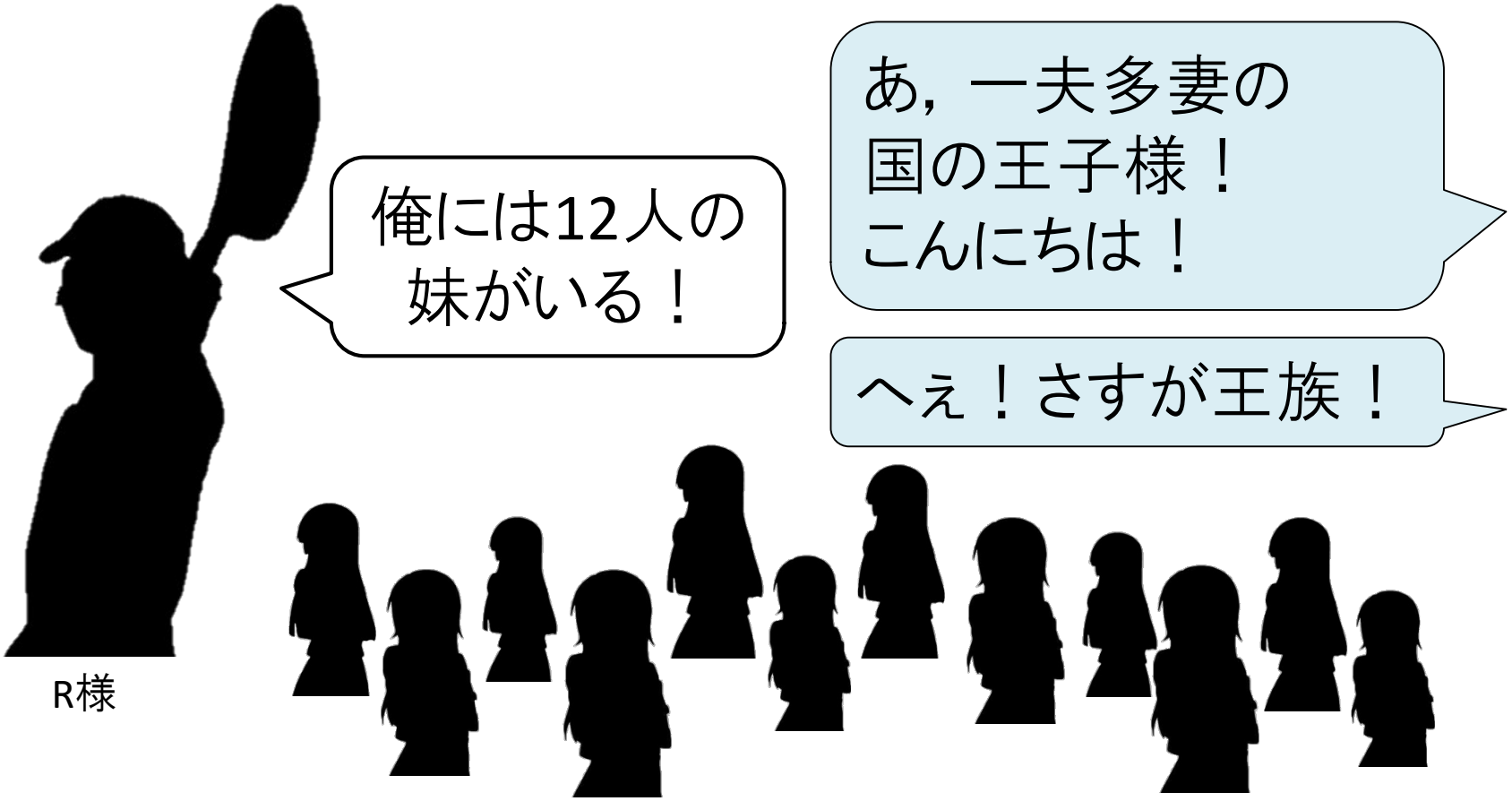
確率変数XとYの結合エントロピー$H(X,Y)$に対し,

$$0 \leq H(X,Y) \leq H(X) + H(Y)$$

が成り立つ. また $H(X,Y) = H(X) + H(Y)$ となるのは,
X と Y が独立のときのみである.

ちょっと休憩

関連情報を事前に知っていた時の情報量は？



関連する情報が既知だと，驚きは少なくなる

→ エントロピーは小さくなっているはず！

例2.2 (p.17)

アイスクリームの売上高が「1万円以上」だったとき，実際の天気についての曖昧さ（エントロピー）は，晴と雨の確率がそれぞれ5/7と2/7であるから，

$$H(X|1\text{万円以上}) = \mathcal{H}(5/7) \fallingdotseq 0.8631 \text{ (bit)}.$$

同様に，売上高が「1万円未満」のときは，

$$H(X|1\text{万円未満}) = \mathcal{H}(1/3) \fallingdotseq 0.9183 \text{ (bit)}.$$

売上高が「1万円以上」「1万円未満」となる確率は，それぞれ0.7と0.3なので，この割合でエントロピーを平均すると，

$$H(X|Y) \fallingdotseq 0.7 \times 0.8631 + 0.3 \times 0.9183$$

$$\fallingdotseq 0.8797 \text{ (bit)}$$

となる．これは，Xのエントロピー

$$H(X) = \mathcal{H}(0.6) = 0.9710 \text{ (bit)}$$

と比べて確かに小さい．

表2.2 Y で条件付けた X の確率分布

$P(X \mid Y)$		Y	
		1万円以上	1万円未満
X	晴	5/7	1/3
	雨	2/7	2/3

条件付きエントロピー

定義2.6

確率変数Yで条件を付けたXの条件付きエントロピー$H(X|Y)$は，

$$\sum_{y_j \in \{y_1, y_2, \ldots, y_{M_Y}\}} P(y_j) H(X|y_j)$$

により定義される．ただし，$\{x_1, x_2, \ldots, x_{M_X}\}$および$\{y_1, y_2, \ldots, y_{M_Y}\}$は，それぞれ$X$と$Y$が取りうる値の集合とする．

Try 練習問題2.3

結合エントロピーと条件付きエントロピーの関係

定理2.3

$\{x_1, x_2, \ldots, x_{M_X}\}$および$\{y_1, y_2, \ldots, y_{M_Y}\}$をとりうる値の集合とする確率変数$X$および$Y$に関し，以下が成り立つ．

(1) $H(X|Y) = -\sum_{i=1}^{M_X} \sum_{j=1}^{M_Y} P(x_i, y_j) \log_2 P(x_i|y_j)$

(2)

(3) $0 \leq H(X|Y) \leq H(X)$
（$H(X|Y) = H(X)$はXとYが独立の時のみ成立）

(4) $0 \leq H(Y|X) \leq H(Y)$
（$H(Y|X) = H(Y)$はXとYが独立の時のみ成立）

別の情報を得ると，エントロピーは変化しないか減少する

定理2.3(2)の証明

［証明］ 結合エントロピーと条件付き確率の定義から,

$$H(X,Y) = -\sum_{i=1}^{M_X}\sum_{j=1}^{M_Y} P(x_i, y_j)\log_2 P(x_i, y_j)$$

$$= -\sum_{i=1}^{M_X}\sum_{j=1}^{M_Y} P(x_i, y_j)\log_2 \frac{P(x_i, y_j)P(x_i)}{P(x_i)}$$

ベイズの定理

$$= -\sum_{i=1}^{M_X}\sum_{j=1}^{M_Y} P(x_i, y_j)\{\log_2 P(x_i) + \log_2 P(y_j|x_i)\}$$

$$= H(X) + H(Y|X)$$

が成立する.

$H(X,Y) = H(Y) + H(X|Y)$ も同様にして証明できる. □

相互情報量の定義 [定義2.7]

例2.2において，天気Xについての曖昧さは，

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i) \fallingdotseq 0.9710 \ \text{(bit)}.$$

アイスクリームの売上高Yを聞いたとき，残っている曖昧さは，

$$H(X|Y) \fallingdotseq 0.8797 \ \text{(bit)}.$$

したがって，売上高Yを聞くことで，天気Xについて

$$I(X;Y) = H(X) - H(X|Y)$$
$$\fallingdotseq 0.9710 - 0.8797 = 0.0913 \ \text{(bit)}$$

だけ，曖昧さが減少する.

言い換えると，売上高Yを聞くことで天気Xに関する情報量が，

（平均として）$I(X;Y) \fallingdotseq 0.0913$ (bit) 得られることを意味する.

この$I(X;Y)$をXとYの相互情報量（mutual information）と呼ぶ.

Try 練習問題2.4

相互情報量の性質(1) [定理2.4(1)]

相互情報量の定義

$$I(X;Y) = H(X) - H(X|Y)$$

と，先ほどの結合エントロピーと条件付きエントロピーの関係

$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$$

から，

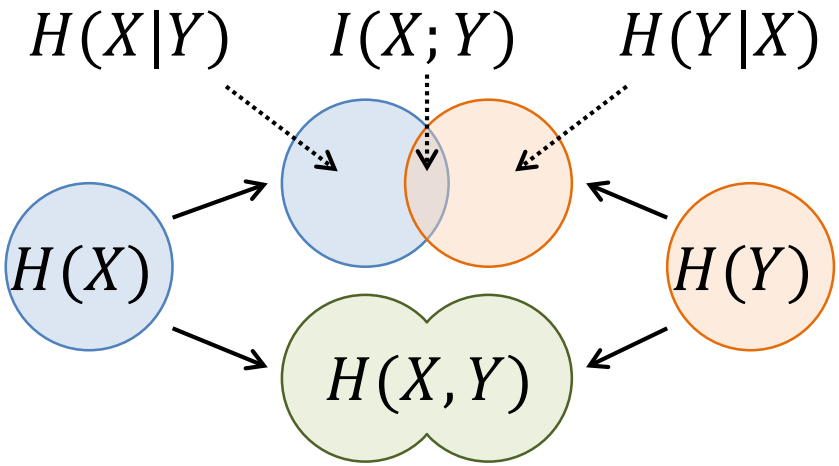
$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(X) + H(Y) - H(X,Y) \\ &= H(Y) - H(Y|X) \\ &= I(Y;X) \\ &= \sum_i \sum_j p(x_i, y_j) \log \frac{p(x_i,y_j)}{p(x_i)p(y_j)} \end{aligned}$$

XとYに関して対称

が成り立つ.

※ $H(X|Y) = -\sum_{j=1}^{m} p(y_i) \sum_{i=1}^{n} p(x_i|y_j) \log p(x_i|y_j) = -\sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i, y_j) \log p(x_i|y_j)$

相互情報量の性質(2) [定理2.4(2)]



相互情報量$I(X;Y)$は，XとYに共通して含まれる情報の量を表すと解釈できる．$I(X;Y)$の範囲は，次式のとおりである．

$$0 \leq I(X;Y) \leq \min\{H(X), H(Y)\}$$

[証明]

$I(X;Y) = H(X) - H(X|Y)$と，$H(X|Y) \leq H(X)$の関係から，左側は明らか．右側の不等式についても，$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$の関係と，$H(X|Y) \geq 0$，$H(Y|X) \geq 0$であることから導ける．

相互情報量の計算例

前回のガンの検査の例について，ガンである確率変数をX，検査の結果の確率変数をYとして相互情報量を計算してみよう.

$$P_{Y|X}(A|C) = P_{Y|X}(A^c|C^c) = 0.95, P_X(C) = 0.01 \quad \text{なので,}$$

$$\begin{aligned} P_Y(A) &= P_{Y|X}(A|C)P_X(C) + P_{Y|X}(A|C^c)P_X(C^c) \\ &= 0.95 \times 0.01 + 0.05 \times 0.99 \\ &= 0.0095 + 0.0495 = 0.059 \ . \end{aligned}$$

$$\therefore H(Y) = \mathcal{H}(0.059) \fallingdotseq 0.323 \ .$$

次に，$H(Y|X = C) = \mathcal{H}(0.95) \fallingdotseq 0.286$ ，

$H(Y|X = C^c) = \mathcal{H}(0.05) \fallingdotseq 0.286$ なので,

$$\begin{aligned} H(Y|X) &\fallingdotseq 0.01 \times 0.286 + 0.99 \times 0.286 \\ &= 0.286 \ . \end{aligned}$$

したがって，相互情報量$I(X;Y)$は,

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &\fallingdotseq 0.323 - 0.286 \\ &= 0.037 \ \text{(bit)}. \end{aligned}$$

ちなみに
$H(X) \fallingdotseq 0.0808$

$P(Y\|X)$		X	
		C	C^c
Y	A	0.95	0.05
	A^c	0.05	0.95

$P(X,Y)$		X	
		C	C^c
Y	A	0.0095	0.0495
	A^c	0.0005	0.9405

今日のまとめ

2.1 情報量

確率pで起こる事象の自己情報量 $I(p) = -\log_a p$

2.2 エントロピー

確率変数Xの平均情報量$H(X) = -\sum_{i=1}^{M} p_i \log_2 p_i$

2.3 エントロピーの性質

$$0 \leq H(X) \leq \log_2 M$$

二つの確率変数に対するエントロピー

2.4 結合エントロピー $H(X,Y)$

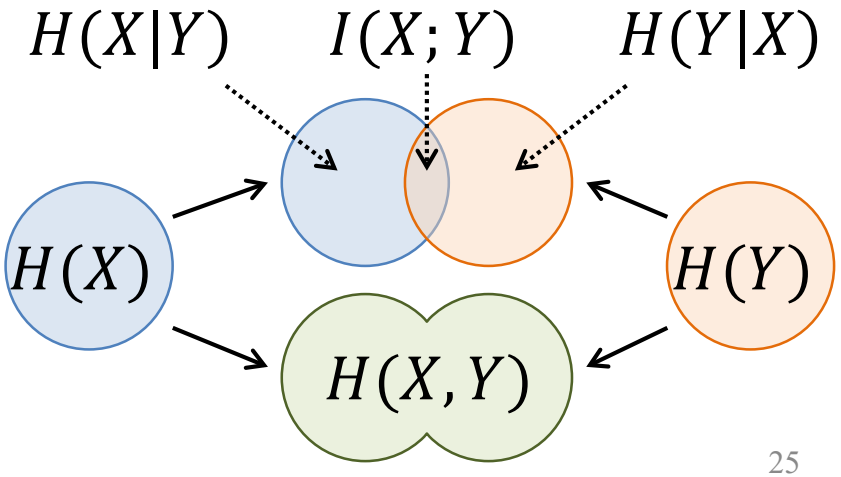
2.5 条件付きエントロピー

$H(X|Y), H(Y|X)$

2.6 相互情報量 $I(X;Y)$

次回

情報源のモデルについて



補助定理A.1[シャノンの補助定理]

補助定理A.1

$p_1, p_2, \cdots, p_M$ および $q_1, q_2, \cdots, q_M$ を

$$p_1 + p_2 + \cdots + p_M = 1,$$

$$q_1 + q_2 + \cdots + q_M \leq 1$$

を満たす任意の非負の数とする（ただし，$p_i \neq 0$ のときは $q_i \neq 0$ とする）．このとき，

$$-\sum_{i=1}^{M} p_i \log_2 q_i \geq -\sum_{i=1}^{M} p_i \log_2 p_i \qquad \text{(A.3)}$$

が成立する．等号は $q_i = p_i$ $(i = 1, 2, \cdots, M)$のとき，またそのときに限って成立する．

証明は教科書を参照

つまり，確率分布$P = \{p_i\}_{i=1}^{M}$ とちょっと違う分布 q_i（ただし総和が1以下）を持ってきて，$\log_2$ の内側の p_i と置き換えると，元よりも少し大きくなる．

定理2.1の証明

Xのエントロピー$H(X)$は

$$H(X) = -\sum_{i=1}^{M} p_i \log_2 p_i \ .$$

$-\log_2 p_i \geq 0$ だから

$0 \leq p_i \leq 1$ なので，明らかに $0 \leq H(S)$ であり，$H(S) = 0$ が成立するのは，$p_1, p_2, \cdots, p_k$ のうち一つが 1 で他が 0 の場合である．

$\sum_{i=1}^{M} p_i = 1$ だから

補助定理A.1（シャノンの補助定理）を$q_i = 1/M$として適用すると，

$$\begin{aligned} H(X) &= -\sum_{i=1}^{M} p_i \log_2 p_i \\ &\leq -\sum_{i=1}^{M} p_i \log_2 \frac{1}{M} \\ &= \log_2 M \ . \end{aligned}$$

補助定理A.1より

等号が成立するのは $p_i = q_i = 1/M$ のときのみである．□

定理2.2の証明

［証明］ 結合エントロピーの定義より$0 \le H(X,Y)$は明らかである.
よって, $H(X,Y) \le H(X)+H(Y)$を証明する.

$$H(X) = -\sum_{i=1}^{M_X} P(x_i)\log_2 P(x_i) = -\sum_{i=1}^{M_X}\sum_{j=1}^{M_Y} P(x_i,y_j)\log_2 P(x_i),$$

$$H(Y) = -\sum_{j=1}^{M_Y} P(y_j)\log_2 P(y_j) = -\sum_{j=1}^{M_Y}\sum_{i=1}^{M_X} P(x_i,y_j)\log_2 P(y_j).$$

したがって,

$$H(X)+H(Y) = -\sum_{j=1}^{M_Y}\sum_{i=1}^{M_X} P(x_i,y_j)\log_2 P(x_i)P(y_j)$$

定理2.2の証明(つづき)

A.1節の補助定理A.1(シャノンの補助定理)を適用すると,

$$-\sum_{j=1}^{M_Y}\sum_{i=1}^{M_X} P(x_i, y_j)\log_2 P(x_i)P(y_j)$$

$$\geq -\sum_{j=1}^{M_Y}\sum_{i=1}^{M_X} P(x_i, y_j)\log_2 P(x_i, y_j)$$

が成り立つ. すなわち, $H(X)+H(Y)\geq H(X,Y)$ となる.

等号が成り立つのは, シャノンの補助定理の統合条件より, すべてのi, jに対して$P(x_i, y_j) = P(x_i)P(y_j)$が成立する場合である. これは, X と Y が独立であるときに他ならない. □

自己情報量が対数関数である理由(1/3)

まず，コーシー（Cauchy）の関数方程式

$$f(x+y)=f(x)+f(y)$$

を満たす連続関数が $f(x)=ax$（aは定数）であることを示す．

$x=y=0$ を代入すると，$f(0)=f(0)+f(0)$ より，$f(0)=0$.
次に，$y=-x$ を代入すると，

$$f(x-x)=f(x)+f(-x),$$
$$0=f(x)+f(-x).$$

より，$f(-x)=-f(x)$ が成り立つ（つまり，$f(x)$ は奇関数）．
n が自然数のとき，

$$f(n)=f\big(1+(n-1)\big)=f(1)+f(n-1)$$
$$=f(1)+f(1)+f(n-2)=\cdots=nf(1).$$

$f(x)$が奇関数であることから，$f(-n)=-nf(1)$も成り立つ．すなわち，任意の整数について$f(n)=nf(1)$が成り立つ．

自己情報量が対数関数である理由(2/3)

同様の考えにより，任意の実数 x と自然数 m に対して，
$f(mx) = mf(x)$ が成り立つ．m が自然数，n が整数のとき，

$$f(n) = f\left(\frac{n}{m} \times m\right) = mf\left(\frac{n}{m}\right)$$

より，

$$f\left(\frac{n}{m}\right) = \frac{f(n)}{m} = \frac{n}{m} f(1).$$

したがって，任意の有理数 x に対して，次が成り立つ．

$$f(x) = xf(1).$$

$f(1)$は定数なので，これを a と置くと，$f(x) = ax$ と書ける．
有理数の稠密性から，連続関数に限定するとコーシーの関数方程式を満たす解は $f(x) = ax$ のみであることが言える．

どんなに微小な区間をとっても，その間に有理数が存在する

自己情報量が対数関数である理由(3/3)

コーシーの関数方程式の解を応用して，自己情報量の三つの性質を満たす関数 $I(p)$ が対数関数で表されることを示す．

ある実数 $a > 1$ をとり，$g(x) = I(a^x)$ とおく．このとき，

$$g(x+y) = I(a^{x+y}) = I(a^x \cdot a^y) = I(a^x) + I(a^y)$$
$$= g(x) + g(y)$$

$I(p)$の加法性から

が成り立つ．コーシーの関数方程式の解から，

$$g(x) = bx$$

と書ける（bは定数）．すなわち，$p = a^x$ とおくと，

$$I(p) = g(\log_a p) = b \log_a p\,.$$

$I(p)$ は $0 \leq p < 1$ で単調減少関数なので，$a > 1$ のとき $b < 0$ でなければならない．$b = -1$ とおけば，$I(p) = -\log_a p$ となる．