講義「情報理論」

第３回　情報量とエントロピー（2）

情報理工学専攻 情報知識ネットワーク研究室
喜田拓也

情報量の定義（おさらい）

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 $(a \neq 1)$は定数とする．

$a = 2$ 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]

エントロピー（おさらい）

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	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

今日の内容

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

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) $\boxed{}$

(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.

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. □

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

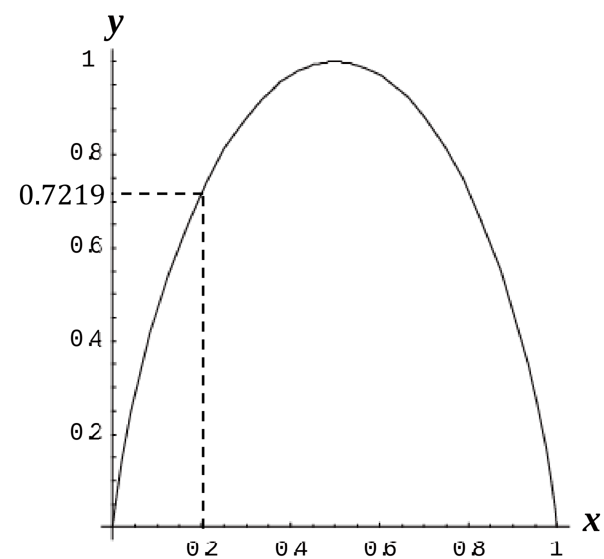
$$\mathcal{H}(x) = -x\log_2 x - (1-x)\log_2(1-x)$$

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，

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

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？

X: 日経平均株価が下がって1万6000円を割ったそうだよ

Y: また円高で，1ドル105円ほどになったよ

へ，へぇ～～～

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$x_1, x_2, \dots, x_{M_X}$の値をとり，Yは$y_1, y_2, \dots, y_{M_Y}$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,

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

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, $\{x_1, x_2, \dots, x_{M_X}\}$および$\{y_1, y_2, \dots, y_{M_Y}\}$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．

定理2.2の証明

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

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	Y: 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,

により定義される. ただし, $\{x_1, x_2, \dots, x_{M_X}\}$および$\{y_1, y_2, \dots, y_{M_Y}\}$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,

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

ベイズの定理

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

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

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，

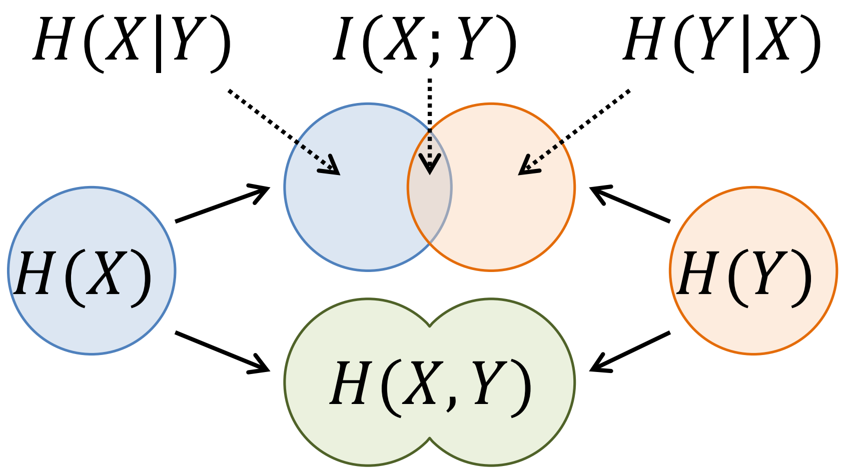
$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(X) + H(Y) - H(X,Y) \\ &= H(Y) - H(Y|X) \\ &= I(Y;X) \\ &= \sum_i \sum_j p(x_i, y_j) \log \frac{p(x_i,y_j)}{p(x_i)p(y_j)} \end{aligned}$$

が成り立つ．

XとYに関して対称

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

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.

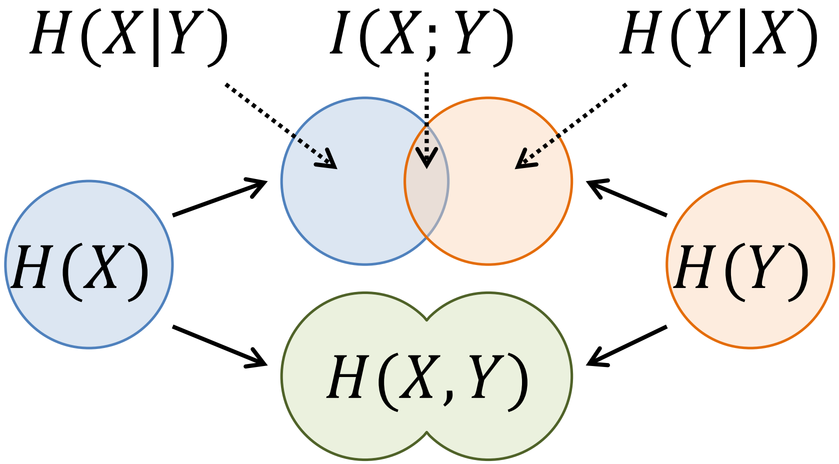
今日のまとめ

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

補足： 確率の統計的な定義

標本集合のどの要素（根源事象）も同程度に確からしく起こる場合は古典的定義で問題ない．

根源事象の
生起にかたより
がある場合

この仮定が適用できない場合は困る．

試行をn回繰り返し行ったときに事象Aが起こった回数を$n(A)$とする．このとき，

$$P(A) = \lim_{n \to \infty} \frac{n(A)}{n}$$

を事象Aの確率とする．

しかし，現実に無限回の試行は不可能なので困る．

補足： 条件付き確率と独立性

ある事象Aが起こったという条件のもとで別の事象Bが起こる確率を考える.

n回の試行を行って事象A，$A \cap B$がそれぞれ$n(A)$回，$n(A \cap B)$回起こったとする. これらの事象の起こる頻度の比（統計的定義による確率）はそれぞれ

$$\frac{n(A)}{n}, \qquad \frac{n(A \cap B)}{n}.$$

である. また，事象Aが起こったという条件のもとで事象Bが起こる頻度の比率を考えると,

$$\frac{n(A \cap B)}{n(A)} = \frac{\frac{n(A \cap B)}{n}}{\frac{n(A)}{n}}.$$

補足：条件付き確率と独立性（つづき）

$P(A) > 0$のとき，

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

を，事象Aが起こったもとでの事象Bの条件付き確率と定義する．このとき，明らかに次式が成り立つ．

$$P(A \cap B) = P(A)P(B|A).$$

事象Bの起こる確率が事象Aの生起に無関係な場合，すなわち

$$P(B|A) = P(B)$$

が成り立つとき，事象Aと事象Bは独立であるという．このとき，明らかに次式が成り立つ．

$$P(A \cap B) = P(A)P(B).$$

補足： 条件付き確率の簡単な例

1個のサイコロを振る試行を考える.

サイコロの出目が偶数である事象をA,
サイコロの出目が4以上である事象をBとする

事象Aの確率$P(A)$は $P(A)=\frac{3}{6}=\frac{1}{2}$

偶数である確率

事象Bの確率$P(B)$は $P(B)=\frac{3}{6}=\frac{1}{2}$

4,5,6である確率

事象$A\cap B$の確率は$P(A\cap B)=\frac{2}{6}=\frac{1}{3}$

4か6である確率

このとき，事象Aで条件付けた事象Bの確率$P(B|A)$は

$$P(B|A)=\frac{P(A\cap B)}{P(A)}=\frac{\frac{1}{3}}{\frac{1}{2}}=\frac{2}{3}.$$

出目が偶数であること
は分かっている場合で
4以上である確率

補足： ベイズの定理

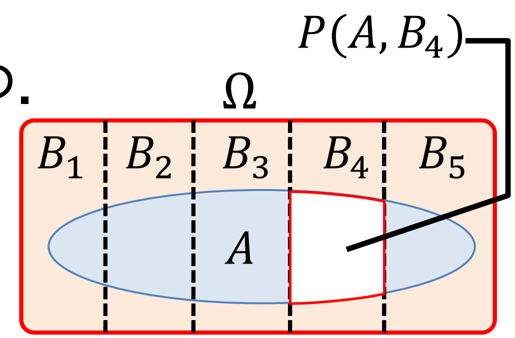
ベイズの定理

全事象 Ω を互いに排反な n個の事象 $B_1, B_2, \ldots, B_n$ に分割する. すなわち,

$$B_i \cap B_j = \emptyset \qquad (i \neq j),$$
$$B_1 \cup B_2 \cup \cdots \cup B_n = \Omega\,.$$

このとき，任意の事象Aに対して次が成り立つ.

$$P(B_k|A) = \frac{P(B_k)P(A|B_k)}{\sum_{i=1}^{n} P(B_i)P(A|B_i)}$$



次の関係は$\Omega = B \cup B^c$の場合の簡略版

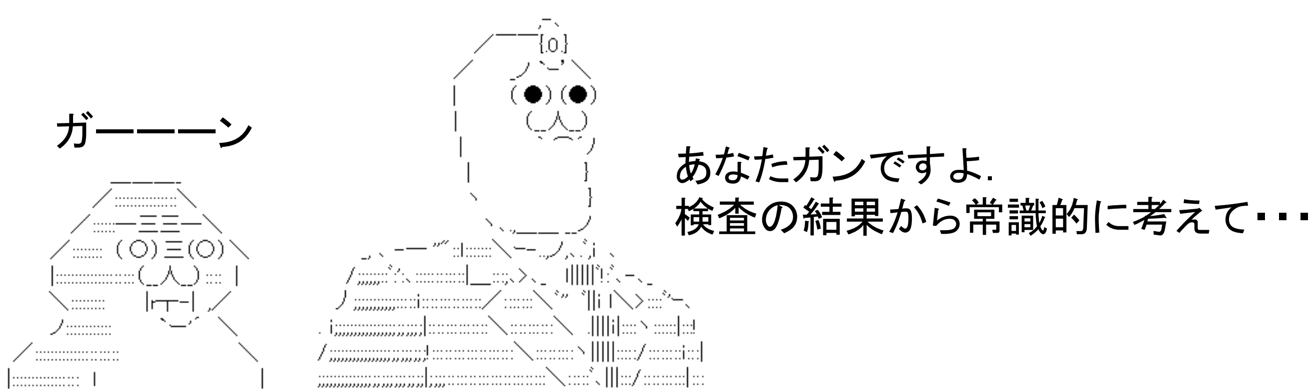
$$P(B|A) = \frac{P(B)P(A|B)}{P(B)P(A|B) + P(B^c)P(A|B^c)} = \frac{P(A,B)}{P(A)}\,.$$

$P(A,B)$ は $P(A \cap B)$ と同じ意味

補足：ベイズの定理を使う例題

[問題]

ガンを診断するための検査法があり，あなたがそれを受けると仮定しよう．
C を，被検査者がガンであるという事象，
A を，検査の結果が被検査者はガンであると示す事象とする．
（つまり検査結果が陽性となる事象）
検査法の認識力が95%（$P(A|C) = 0.95$ かつ $P(A^c|C^c) = 0.95$）であるとし，
検査を受ける人の中で実際にガンである割合が $P(C) = 0.01$であるとする．
このとき･･･



検査の結果が陽性だったとき，あなたが本当にガンである確率はいくらか？
（つまり，$P(C|A)$を求めなさい）

補足： ベイズの定理を使う例題の解答

［答え］ベイズの定理より,

$$P(C|A) = \frac{P(C)P(A|C)}{\sum_{i=1}^{2} P(C_i)P(A|C_i)} = \frac{P(C)P(A|C)}{P(C)P(A|C) + P(C^c)P(A|C^c)}$$

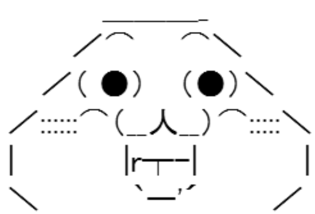
である. これに,

$$P(C) = 0.01, \qquad P(C^c) = 1 - P(C) = 0.99,$$

$$P(A|C) = 0.95, \qquad P(A|C^c) = 1 - P(A^c|C^c) = 0.05$$

を代入すると,

$$P(C|A) = \frac{0.01 \times 0.95}{0.01 \times 0.95 + 0.99 \times 0.05} = \frac{19}{118} \fallingdotseq 0.161.$$



たったの16%！
よかったお

補足: 相互情報量の計算例

最初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.

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

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

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

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，

$$H(Y|X) \fallingdotseq 0.01 \times 0.286 + 0.99 \times 0.286$$
$$= 0.286 .$$

したがって，相互情報量$I(X;Y)$は，

$$I(X;Y) = H(Y) - H(Y|X)$$
$$\fallingdotseq 0.323 - 0.286$$
$$= 0.037 \ \ (\mathrm{bit}).$$

ちなみに
$H(X) \fallingdotseq 0.0808$

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

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