

コンピュータによる自動認識 (1) : 確率論と情報理論

- 問： 情報とは何か . 情報を数量的にとらえることは可能なのか .
- C. E. Shannon が , 確率論を基礎に数量的に定式化した .

C. E. Shannon, "A mathematical theory of communication," Bell System Technical Journal, vol. 27, pp. 379-423 and 623-656, July and October, 1948.

- まずは確率の話

同時分布 , 周辺分布の例 :

	B_1 (風邪)	B_2 (風邪なし)	$p(A_i)$
A_1 (熱あり)	0.55	0.05	0.60
A_2 (熱なし)	0.10	0.30	0.40
$p(B_j)$	0.65	0.35	

熱があるかないかを知ってしまえば , 風邪であるかないかはほぼ検討がつく .
確率と情報量は何か関係がありそう .

確率と確率分布の基礎知識 :

- 同時確率分布: $p(A_i, B_j)$ A_i と B_j とが組になって同時に起こる確率
- 周辺確率分布: $p(A_i)$ 一方 (A_i) だけに着目した確率分布
- 条件付確率分布: $p(B_j|A_i)$ A に関する事象が A_i であることがわかっているときの B_j の確率

これらの間には次の関係が成立する .

$$\sum_{i,j} p(A_i, B_j) = 1, \quad \sum_i p(A_i) = \sum_j p(B_j) = 1 \tag{1}$$

$$\sum_j p(B_j|A_i) = \sum_i p(A_i|B_j) = 1 \tag{2}$$

$$p(A_i) = \sum_j p(A_i, B_j), \quad p(B_j) = \sum_i p(A_i, B_j) \tag{3}$$

$$p(A_i, B_j) = p(A_i)p(B_j|A_i) = p(B_j)p(A_i|B_j) \tag{4}$$

$$p(B_j|A_i) = \frac{p(B_j)p(A_i|B_j)}{p(A_i)} = \frac{p(A_i, B_j)}{p(A_i)} \tag{5}$$

$$p(A_i|B_j) = \frac{p(A_i)p(B_j|A_i)}{p(B_j)} = \frac{p(A_i, B_j)}{p(B_j)} \tag{6}$$

- 計算例 :

熱がある (A_1) という条件のもとでの風邪のあるなしに関する条件付確率分布

$$P(B_1|A_1) = \frac{P(A_1, B_1)}{P(A_1)} = \frac{0.55}{0.60} = \frac{11}{12} \approx 0.92$$

$$P(B_2|A_1) = \frac{P(A_1, B_2)}{P(A_1)} = \frac{0.05}{0.60} = \frac{1}{12} \approx 0.08$$

● 情報量とエントロピー

定義：情報量

確率 p の事象が実際に起ったことを知らせる情報に含まれている情報量を $-\log_2 p$ ビットと定義する。

- 例：サイコロ ($\log_2 6 \approx 2.58$) . コイン ($\log_2 2 = 1$) . 宝くじ ($-\log_2 2^{-100} = 100$) ...
- 不確実な状況を確定したいから情報がほしい .
- どういう情報がもらえるかは事前にはわからない .
もらえる情報量そのものはわからない .
- $\sum_{i=1}^n p_i = 1$ であるのでもらえる情報量の期待値はわかる .

定義：エントロピー

n 個の事象がそれぞれ確率 p_1, p_2, \dots, p_n で発生するとき、どれが発生したかの不確実度を

$$H(p_1, p_2, \dots, p_n) = - \sum_{i=1}^n p_i \log p_i$$

と定義し、これをエントロピーと呼ぶ。

- この値は、不確実な状況を確定するのに要する平均情報量。

定義：情報量

情報を受け取る以前の、事象 A についてのエントロピーを $H(A)$ 、情報を受け取った後のエントロピーを $H'(A)$ とするとき、事象 A に関するこの情報の情報量 I を

$$I = H(A) - H'(A)$$

と定義する。

● 計算例：

体温を計る前の風邪についてのエントロピー：

$$H(B) = - \sum_{j=1}^2 P(B_j) \log_2 P(B_j) = -0.65 \log 0.65 - 0.35 \log 0.35 = 0.93 \text{ ビット}$$

体温を計り、熱がある (A_1) と分かった後での風邪についてのエントロピー：

$$H(B|A_1) = - \sum_{j=1}^2 P(B_j|A_1) \log_2 P(B_j|A_1) = -0.92 \log 0.92 - 0.08 \log 0.08 = 0.41 \text{ ビット}$$

したがって、熱があることが分かったことにより、 $0.52 (= 0.93 - 0.41)$ ビットの情報量を得たことがわかる。