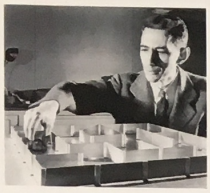
情報理論

Information Theory

柳田研究室

情報理論



情報理論とは「情報」を数学的に扱うことにより，情報を，効率よく，確実に，伝送したり，蓄積したりするための理論である．

1948年C.E.Shannonによる論文『通信の数学的理論』(A Mathematical Theory of Communication)によって情報理論の基礎が確立された．

彼は通信の概念をより明確にするため，通信システムを機能に応じた要素に分解したモデルを導入した．

そして，情報の量(エントロピー)を数学的に定式化することによってそれらモデルを特徴付け，それらを扱う上で重要ないくつもの法則を与えた．

Shannonの通信モデル

エントロピー関数の導出

取り得る値の種類が多い
それぞれの値になる確率が散らばっている

大 ⟷ 小

取り得る値の種類が少ない
それぞれの値になる確率が偏っている

不確かさの「量」 → エントロピー

得た情報の量(エントロピー)を減った不確かさの量と考え，満たすべき性質からエントロピー関数を導出する．

■ n 個の値のどれもが平等に現れる可能性があるとき，その不確かさを n の関数とみて，$u(n)$で表すことにする．

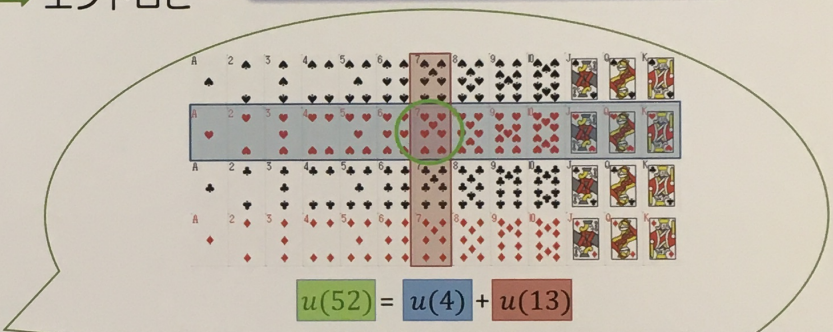
✓ nが大きくなれば不確かさも大きくなると考えられる．

$$0 = u(1) < u(2) < u(3) \cdots$$

✓ ある直積集合から平等に現れる可能性の不確かさについて考える．直積の一つの成分がどの要素をとったかが分かると，その成分についての不確かさが減り，もう片方の成分についての不確かさだけが残ると考えられる．

$$u(mn) = u(m) + u(n)$$

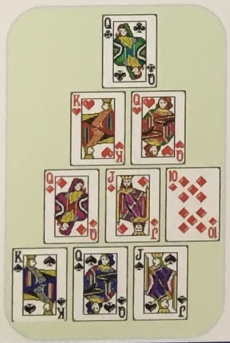
$$u(n) = \log_b n \quad (b > 1)$$



$u(52) = u(4) + u(13)$

52枚のカードから無作為に一枚選ぶ場合，どのマークが選ばれたかを知ると，52枚から選ばれる可能性の不確かさから，13枚から選ばれる可能性の不確かさに減る．

■ 確率分布 $p_1, p_2, \ldots, p_n$ の不確かさを $H(p_1, p_2, \ldots, p_n)$ と表すと，$H\left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right) = u(n)$ となる．



■ 任意の確率分布に対するエントロピー関数を求めるため，例として左図のカード群から無作為に一枚選ぶ場合の不確かさについて考える．

✓ 9枚のカードから1枚選ばれる可能性の不確かさは $u(9)$

✓ どのマークが選ばれたのか知ったのちに残る数字に関する不確かさの平均値は

$$\frac{1}{9}u(1) + \frac{2}{9}u(2) + \frac{3}{9}u(3) + \frac{3}{9}u(3)$$

よってマークを知ったことにより減った不確かさの量$H\left(\frac{1}{9}, \frac{2}{9}, \frac{3}{9}, \frac{3}{9}\right)$は以下を満たすと考えられる．

$$H\left(\frac{1}{9}, \frac{2}{9}, \frac{3}{9}, \frac{3}{9}\right) = u(9) - \left(\frac{1}{9}u(1) + \frac{2}{9}u(2) + \frac{3}{9}u(3) + \frac{3}{9}u(3)\right)$$

$$= -\frac{1}{9}\log\frac{1}{9} - \frac{2}{9}\log\frac{2}{9} - \frac{3}{9}\log\frac{3}{9} - \frac{3}{9}\log\frac{3}{9}$$

以上の議論に確率分布に関して連続であるという仮定を加えると，エントロピー関数を以下のように定義することができる．

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

エントロピー関数を用いた定理

■ 情報源を，メッセージをある確率分布に従って記号ごとに出力するものと考える．情報源から出力される記号を符号化するとは，何らかのルールを用いて記号を別の記号列(符号語)に置き換える操作をいう．

■ 符号化された符号語列を一意に元の記号列に戻せる(復号できる)という条件の下では，その符号の一記号あたりの平均の符号語の長さは，情報源が従う確率分布のエントロピーの値を下回ることはない．またそのエントロピーの値にいくらでも近い平均符号語長をもつ符号を求めることが出来る．

情報源符号化定理

以下を満たす一意復号可能な符号が存在する．$l_1, l_2, \ldots, l_n$を一意復号可能な符号の各符号語長とする．

$$H(p_1, \ldots, p_n) \leq \sum_{i=1}^{n} p_i l_i < 1 + H(p_1, \ldots, p_n)$$

■ ‘C.E.SHANNON’という記号列を，最適符号(平均符号語長が最も小さい符号)の1つであるハフマンコードを用いて符号化した例．(下図では各記号と符号語の対応を色ごとに表している．)

C.E.SHANNON → 符号化 → 110000111000100101001001010011101

$H(p_1, \ldots, p_n)$

$H\left(\frac{1}{11}, \frac{2}{11}, \frac{1}{11}, \frac{1}{11}, \frac{1}{11}, \frac{1}{11}, \frac{3}{11}, \frac{1}{11}\right)$

≈ 1.972

$\leq$

$\sum p_i l_i$

$\frac{1}{11}\cdot 3 + \frac{2}{11}\cdot 3 + \frac{1}{11}\cdot 3 + \frac{1}{11}\cdot 3 + \frac{1}{11}\cdot 3 + \frac{1}{11}\cdot 4 + \frac{3}{11}\cdot 2 + \frac{1}{11}\cdot 4 \approx 2.909$