1 情報源と情報量

1.1 情報源

情報の意味的な要素を捨てて客観的に扱おうとすれば, 頻度あるいは確率に注目するのが自然であろう.

定義 1. 情報源とは有限集合 Σ と σ 上の分布 $(p_\sigma)_{\sigma\in\Sigma}$ の組 $(\Sigma, (p_\sigma)_{\sigma\in\Sigma})$ のことを指す. また Σ の事を情報源アルファベットと呼ぶ.

Σ が有限集合のとき, Σ-値確率変数 X はその分布によって情報源とみなすことができる. あるいは, 頻度の抽象化として捉える場合は独立同分布確率変数列 $X_1, \ldots, X_N$ を考える方が自然である. これは無記憶情報源とよばれる.

1.2 情報量

数学的に「情報量」を定義する前に, 「情報量」が持つことを期待する性質を整理してみる:

1. 事象 A の「情報量」はその確率 $P(A)$ によって決める.
   [「情報量」はある関数 h を使って $h(P(A))$ と書ける.]
2. より確率が低い方が事象 A の「情報量」$h(P(A))$ は小さい様にする.
   [h は単調減少関数]
3. 「独立」な事象 A と B の「情報量」は各々の「情報量」の和に等しい.
   $$h(P(A)P(B)) = h(P(A\cap B)) = h(P(A)) + h(P(B))$$
   [h は積を和に移す.(逆方向の指数法則)]

この性質から h は 1 より小さい底を持つ対数関数を使うことが自然であろう. 特に数学的議論の煩雑さを避ける為に底として e^{-1} を使う. つまり

$$h(x) = \log_{e^{-1}} x = -\log x$$

として定めるのが次の定義である.

定義 2. $(\Sigma, (p_\sigma)_{\sigma\in\Sigma})$ を情報源とする. このとき

$$\sum_{\sigma\in\Sigma} p_\sigma \log \frac{1}{p_\sigma}$$

をその情報源の情報量 (エントロピー) といい, $H((\Sigma, (p_\sigma)_{\sigma\in\Sigma}))$ と表す,

例 1 (コイン投げ)**.** 情報源

$$\Sigma := \{0,1\}, \quad p_0 = p_1 = \frac{1}{2}$$

の情報量は

$$H((\Sigma, (p_\sigma)_{\sigma\in\Sigma})) = -p_0\log p_0 - p_1 \log p_1 = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = \log 2$$

である.

例 2 (一様分布)**.** 情報源

$$\Sigma := \{1, \ldots, m\}, \quad p_0 = p_1 = \cdots = p_m := \frac{1}{m}$$

の情報量は

$$H((\Sigma, (p_\sigma)_{\sigma\in\Sigma})) = \sum_{k=1}^{m} -\frac{1}{m}\log\frac{1}{m} = \log m$$

である.

1.3 情報量の単位

自然対数で定義した情報量

$$H((\Sigma, (p_\sigma)_{\sigma\in\Sigma}))$$

の単位は [nat](ナット) という.

導入で述べたように $h(x) = \log_{e^{-1}} x = -\log x$ の底の選び方は数学的議論の簡単さの為であって, 本質的には $\alpha > 1$ について $h_\alpha(x) = \log_{\alpha^{-1}} x = -\log_\alpha x$ を用いてもよかった. 特に実用上の利便さから $\alpha = 2$ で定めた情報量も頻繁に用いられる.

2 を底にして意義される値

$$H_2((\Sigma, (p_\sigma)_{\sigma\in\Sigma})) = \sum_{\sigma\in\Sigma} -p_\sigma \log_2 p_\sigma$$

もまた情報量と呼ばれ [bit](ビット) と言う単位が用いられる.

[bit] と [nat] の値は

$$H_2((\Sigma, (p_\sigma)_{\sigma\in\Sigma}))[\text{bit}] = \frac{1}{\log 2} H((\Sigma, (p_\sigma)_{\sigma\in\Sigma}))[\text{nat}]$$

によって変換できる.

例 3. 例 1 の情報量は

$$0.693[\text{nat}] \doteqdot \log 2[\text{nat}] = 1[\text{bit}]$$

である.

1.4 情報量の性質

例 2 を見よ (普通の) サイコロが与える情報は $\log_2 6[\text{bit}] \doteqdot 2.58[\text{bit}]$ である. それに対して以下の様に偏ったサイコロが与える情報量は小さい.

例 4 (6 に偏ったサイコロ)**.** 情報源

$$\Sigma := \{1, 2, 3, 4, 5, 6\}, \qquad p_0 = p_1 = \cdots = p_5 := \frac{1}{7}, \quad p_6 := \frac{2}{7}$$

の情報量は

$$H_2((\Sigma, (p_\sigma)_{\sigma\in\Sigma})) = \log_2 7 - 1 \doteqdot 1.81$$

偏ったサイコロの目は普通のサイコロの目より当てやすい.[1] この様な意味で, 予想がつきにくいことについての情報元ほど情報量も大きくなっている.

命題 1. 情報源アルファベット Σ の元の数が m 個であるとき

$$0 \le H((\Sigma, (p_\sigma)_{\sigma\in\Sigma})) \le \log m$$

で特に均等分布の時最大値 $\log m$ をとる.

Proof. log は上に凸であるので

$$H((\Sigma, (p_\sigma)_{\sigma\in\Sigma})) = \sum_{\sigma\in\Sigma} p_\sigma \log \frac{1}{p_\sigma} \le \log \sum_{\sigma\in\Sigma} p_\sigma \frac{1}{p_\sigma} = \log m$$

□

[1] この場合 6 と予想すれば 2/7 の確率で当たるのに対し, 普通のサイコロの場合はどの目を予想しても当たる確率は 1/6 で, 2/7 より低い.

逆に確定した情報しか与えない情報源の情報量は 0 である.

命題 2. 次の条件は同値.

1. ある $\hat{\sigma} \in \Sigma$ で $p_{\hat{\sigma}} = 1$ なるものが存在する.
2. ある $\hat{\sigma} \in \Sigma$ で $\sigma \neq \hat{\sigma} \Rightarrow p_\sigma = 0$ をみたすものが存在する.
3. 情報量が 0.

2 複数の確率変数と結合分布の情報量

2.1 定義

Σ が有限集合のとき, Σ-値確率変数 X はその分布によって情報源とみなすことができた. この情報量は

$$\sum_{\sigma\in\Sigma} -p_\sigma \log p_\sigma = \sum_{\sigma\in\Sigma} -P(X=\sigma)\log P(X=\sigma) = E[-\log p_X]$$

と等しい. 確率変数 X を情報源と見なす時, その情報量を $H(X)$ で表す.

更に有限集合 Σ_1,Σ_2 と各々Σ_1-値,Σ_2-値の確率変数 X_1 と X_2 があるとき,

$$p_{\sigma_1,\sigma_2} := P(X_1=\sigma_1, X_2=\sigma_2)$$

は $\Sigma := \Sigma_1 \times \Sigma_2$ 上の分布を定める (X_1 と X_2 の結合分布と呼ぶ). この分布の情報量を $H(X_1, X_2)$ で表す. 即ち,

$$H(X_1,X_2) = \sum_{(\sigma_1,\sigma_2)\in\Sigma} -p_{(\sigma_1,\sigma_2)}\log p_{(\sigma_1,\sigma_2)} = \sum_{\sigma_1\in\Sigma_1}\sum_{\sigma_2\in\Sigma_2} -P(X_1=\sigma_1,\, X_2=\sigma_2)\log P(X_1=\sigma_1,\, X_2=\sigma_2).$$

確率変数が 3 個以上のときも同様に定める.

2.2 結合分布の情報量の性質

補題 1. $h(x) = -\log x$ は凸関数. 即ち $p_1 + \cdots + p_n = 1$ なる $p_i > 0$ と任意の x_i について

$$\sum_{i=1}^n p_i h(x_i) \geq h(\sum_{i=1}^n p_i x_i)$$

特に $x_1 = \cdots = x_n = \sum_{i=1}^n p_i x_i$ のとき, またその時のみ等式が成り立つ.

命題 3. Z を Σ-値確率変数, $h : \Sigma \to \Sigma'$,$X := h(Z)$ とする. このとき

$$H(X) \leq H(Z)$$

$P(g(X) = Z) = 1$ なる $g : \Sigma' \to \Sigma$ が存在するとき, またその時のみ等号が成り立つ.

Proof.

$$\begin{aligned}
H(X) - H(Z) &= \sum_{\sigma'\in\Sigma'} h(P(X=\sigma'))P(X=\sigma') - \sum_{\sigma\in\Sigma} h(P(Z=\sigma))P(Z=\sigma) \\
&= \sum_{\sigma'\in\Sigma'} h(P(X=\sigma')) \sum_{\sigma:h(\sigma)=\sigma'} P(Z=\sigma) - \sum_{\sigma\in\Sigma} h(P(Z=\sigma))P(Z=\sigma) \\
&= \sum_{\sigma'\in\Sigma'}\sum_{\sigma:h(\sigma)=\sigma'} h(P(X=h(\sigma)))P(Z=\sigma) - \sum_{\sigma\in\Sigma} h(P(Z=\sigma))P(Z=\sigma) \\
&= \sum_{\sigma\in\Sigma} h(\frac{P(X=h(\sigma))}{P(Z=\sigma)})P(Z=\sigma)
\end{aligned}$$

h の凸性によって

$$H(X) - H(Z) = \sum_{\sigma\in\Sigma} h(\frac{P(X=h(\sigma))}{P(Z=\sigma)})P(Z=\sigma) \geq h(\sum_{\sigma\in\Sigma} P(Z=\sigma)\frac{P(X=h(\sigma))}{P(Z=\sigma)}) = h(1) = 0$$

g が存在すれば同様に逆の不等号が導ける.

逆に $H(X) - H(Z) = 0$ を仮定する.

$$\bar{\Sigma} := \{\sigma \in \Sigma \mid P(Z=\sigma) > 0\}, \quad \bar{\Sigma}' := \{h(\sigma) \mid \sigma \in \bar{\Sigma}\}$$

とおけば $\sigma \notin \bar{\Sigma}$ ならば $P(Z=\sigma) = 0$ であるので

$$P(X \in \bar{\Sigma}') = P(Z \in \bar{\Sigma}) = \sum_{\sigma\in\bar{\Sigma}} P(Z=\sigma) = \sum_{\sigma\in\Sigma} P(Z=\sigma) = P(\Omega) = 1.$$

一方で, $H(X) - H(Z) = 0$ をの仮定と h の狭義凸性から, $\sigma \in \bar{\Sigma}$ について $\frac{P(X=h(\sigma))}{P(Z=\sigma)} = 1$ であるから

$$\sum_{\sigma\in\bar{\Sigma}} P(X=h(\sigma)) = \sum_{\sigma\in\bar{\Sigma}} P(Z=\sigma) = 1.$$

よって

$$P(X \in \bar{\Sigma}') = P(\bigcup_{\sigma\in\bar{\Sigma}} \{X = h(\sigma)\}) = 1 = \sum_{\sigma\in\bar{\Sigma}} P(X=h(\sigma))$$

つまり加法性が成り立っているがこれは $\{X = h(\sigma)\}$ が互いに素であることを導く. よって h は $\bar{\Sigma}$ 上で単射であり, $\bar{\Sigma}'$ からの逆写像が存在する. □

各成分への射影を考えれば, Proposition 3 の特別な場合として

$$H(X) \leq H(X,Y), \quad H(Y) \leq H(X,Y)$$

が成り立つ.

結合分布は上からの評価として和で押さえられる:

命題 4. X,Y が確率変数とする

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

Proof. 情報量の定義から

$$H(X) + H(Y) - H(X,Y) = \sum_{\sigma,\sigma'} P(X=\sigma, Y=\sigma')h(\frac{P(X=\sigma)P(Y=\sigma')}{P(X=\sigma, Y=\sigma')})$$

h の凸性を使えばよい. □

不等号は, 「情報の重なり」を表していると考えられる. 実際, 等号は次のときに成り立つ

命題 5. X,Y を確率変数とする.

1. Y が X によって決まる (ある $h : \Sigma_1 \to \Sigma_2$ で $P(Y = h(X)) = 1$ なるものが存在する) とき, またその時のみ
$$H(X,Y) = H(X)$$
2. X,Y が独立であるとき, またその時のみ
$$H(X,Y) = H(X) + H(Y)$$

が成り立つ.

$H(X,Y)$ と $H(X)$ の差, つまり「X を知っている状態で Y を知ったときに, 新たに得られる情報量」を

$$H(Y \mid X) := H(X,Y) - H(X)$$

で書く. 上の命題の主張は,

1. Y が X によって決まるとき $H(Y|X) = 0$
2. X,Y が独立であるならば $H(Y \mid X) = H(Y)$

と書ける.