情報量とエントロピー

機械学習ではデータから有益な情報を獲得することが目的とされる。これは情報に関する技術と言えるため、情報に関する理論が必要である。理論を構築する上で、対象を測る「ものさし」を定義することが重要である。ここでは、事象（または出来事）の発生が持つ情報の量を定義し、そのエントロピーについて考える。

1. 情報量

事象 x が発生することによって得られる情報量 $\mathcal{I}$ を以下の 2 つの条件を満たす量として定める ：

(I)　起こりにくい事象ほど多くの情報を持っている。
(II)　確率的に独立な 2 つの事象が持つ情報量は、それぞれの事象が持つ情報量の和で表される。

条件 (I) より、事象 x の持つ情報量は x が発生する確率 $P(x)$ の関数 $\mathcal{I} = \mathcal{I}(P(x))$ となる。また、2 つの事象 x, y がともに生じる確率は、結合確率を用い $P(x, y)$ で表せる。したがって、x と y の両方を知ることの情報量は $\mathcal{I}\left(P(x, y)\right)$ と表せる。2 つの確率変数が独立である場合、$P(x, y) = P(x)P(y)$ が成り立つので、条件 (II) は以下のように書ける。

$$\mathcal{I}\left(P(x)P(y)\right) = \mathcal{I}\left(P(x)\right) + \mathcal{I}\left(P(y)\right) \tag{1}$$

これを満たす関数として、対数関数が考えられる。そこで、条件 (I) を考慮して

$$\mathcal{I}\left(P(x)\right) = -\log P(x) \tag{2}$$

と定義する。$-\log P(x)$ を事象 x の情報量 (information) と呼ぶ。log の底の選び方は任意である。

（例 1） 2 つの事象の情報量

- 事象 x_1 ： 「8 月に雪が降った」
- 事象 x_2 ： 「2 月に雪が降った」

どちらの事象がニュースになりやすいか考えると、事象 x_1 の「8 月に雪が降った」の方であろう。一般に、我々が予想しなかったような（事前確率の小さい）事象が生起したとき、大きな情報量が与えられると考えるのはきわめて自然である。実際、それぞれの事象が起きる確率を考えると

$$P(x_1) \leq P(x_2)$$

であるので、式 (2) より

$$\mathcal{I}(P(x_1)) \geq \mathcal{I}(P(x_2))$$

となり、情報量を $-\log P(x)$ で定義することが妥当であることを示している。■

（例 2） 確実に起こる事象の情報量

確実に起こる事象 x の生じる確率は $P(x) = 1$ である。ゆえに情報量は

$$\mathcal{I}(P(x)) = -\log P(x) = 0$$

になる。実際、確実に起きる事象を知っても何の情報を得たことにもならないので、これが 0 となるのは情報量を $-\log P(x)$ で定義することが妥当であることを示している。■

2. エントロピー

ここでは事象は確率的に生じるとしているので、情報量 $-\log P(x)$ も確率的に決まる。そのため複数の分布の間での情報量の比較を行うために期待値を使うのが便利である。分布 $P(x)$ による情報量の

期待値をエントロピー (entropy) と呼び、$H(P)$ で表わす。

離散的確率変数に対しては、

$$H(P) = E_{P(x)}\left[-\log P(x)\right] = -\sum_{i=1}^{n} P(x_i)\log P(x_i) \tag{3}$$

となる。$P = 0$ のときは、ロピタルの定理により

$$\lim_{P\to+0} P\log P = \lim_{P\to+0} \frac{\log P}{\frac{1}{P}} = \lim_{P\to+0} \frac{\frac{1}{P}}{-\frac{1}{P^2}} = -\lim_{P\to+0} P = 0$$

となるので、$P(x)\log P(x) = 0$ と定義する。離散的確率変数に対する確率は $0 \le P \le 1$ であるので、式 (3) より、エントロピーの非負性：$H(P) \ge 0$ は明らかである。

一方、連続連続的確率変数に対しては、$p(x)$ を確率密度として、これを離散化した極限として考える。すなわち、実数をいくつかの区間に分割し、ある区間内の値はすべて 1 つの代表値で置き換える。一様分割の場合は、確率変数 x が区間

$$x_i - \frac{\Delta}{2} \le x < x_i + \frac{\Delta}{2} \tag{4}$$

にあれば、x を x_i に置き換える。x がこの区間に入る確率は

$$P(x_i) = \int_{x_i-\Delta/2}^{x_i+\Delta/2} p(x)\,dx \tag{5}$$

で与えられ、区間幅 Δ が十分小さければ、近似的に

$$P(x_i) \approx p(x_i)\Delta \tag{6}$$

したがって、連続的確率変数に対するエントロピーは、近似的に

$$\begin{aligned} H(p) &= -\sum_{i=1}^{n} P(x_i)\log P(x_i) \\ &\approx -\sum_{i=1}^{n} p(x_i)\Delta \log\left(p(x_i)\Delta\right) \\ &= -\sum_{i=1}^{n} p(x_i)\Delta \log\left(p(x_i)\right) - \sum_{i=1}^{n} p(x_i)\Delta\log\left(\Delta\right) \\ &= -\sum_{i=1}^{n} p(x_i)\Delta \log\left(p(x_i)\right) - \log\left(\Delta\right) \end{aligned} \tag{7}$$

ここで、$p(x_i)\Delta$ は確率であるので、$\sum_i p(x_i)\Delta = 1$ であることを用いた。式 (7) において、Δ が十分に小さければ、式 (7) の第 1 項の総和は積分で近似できるから

$$H(p) = -\int_{-\infty}^{\infty} p(x)\log p(x)\,dx - \log\left(\Delta\right) \tag{8}$$

のように書ける。第 1 項は確率分布によって決まる項、第 2 項は分割幅によって決まる項である。

分割幅 Δ を無限に小さくして行けば、$\Delta \to 0$ の極限において、式 (8) の第 1 項の積分は近似的ではなく厳密に成立する。ところが、第 2 項は $\Delta \to 0$ の極限で無限大に発散してしまう。すなわち、連続的確率変数に対するエントロピーは無限大になる。

これは、少し考えれば至極当然なことであることが分かる。実数は、いかなる小さな区間をとっても、その中に無限個の実数が存在する。ある実数を完全に定めるには、小数点以下無限個の桁を必要とし、

これを定めるには無限個の情報が必要である。

しかし、ここで確率分布に依存するのは第 1 項だけであることに注目すると、第 1 項はそれぞれの相対的な情報を表わしている。したがって連続的確率変数のときは、この第 1 項だけをとり

$$H(p) = -\int_{-\infty}^{\infty} p(x) \log p(x)\, dx \tag{9}$$

をこの連続的確率変数のエントロピーと定義する。上の説明でもわかる通り、この場合のエントロピーは相対的な意味しか持っていない。

$p(x) = 1$ のときに、式 (9) は 0 になる。これは、ある単位区間、たとえば $[0, 1]$ の間に x が一様に分布した場合に相当する。したがって、式 (9) のエントロピーは、この状態を基準にして、これを 0 と定めたときの相対的な値を表わしていると考えることができる。

式 (7) のエントロピーの非負性については

$$\begin{aligned} H(p) &= -\sum_{i=1}^{n} p(x_i)\Delta \log\left(p(x_i)\right) - \log\left(\Delta\right) \\ &= -\sum_{i=1}^{n} p(x_i)\Delta \log\left(\left(p(x_i)\Delta\right)/\Delta\right) - \log\left(\Delta\right) \\ &= -\sum_{i=1}^{n} p(x_i)\Delta \log\left(p(x_i)\Delta\right) + \left(\sum_{i=1}^{n} p(x_i)\Delta\right) \log\left(\Delta\right) - \log\left(\Delta\right) \\ &= -\sum_{i=1}^{n} p(x_i)\Delta \log\left(p(x_i)\Delta\right) \geq 0 \end{aligned} \tag{10}$$

となる。ここで、$p(x_i)\Delta$ は確率であるので、$0 \leq p(x_i)\Delta \leq 1$ および $\sum_{i=1}^{n} p(x_i)\Delta = 1$ であることを用いた。

文献 1) では、式 (9) のエントロピーについても非負性が成立すると述べられているが、その理由は明確ではない（式 (9) は相対的な量であることに注意が必要）。

以下では、離散的な場合は積分を和で、$p(x)$ を $P(x_i)$ で置き換えるものとして、エントロピーを

$$H(p) = -\int p(x) \log p(x)\, dx \tag{11}$$

と書くことにする。

3. クロスエントロピー

確率分布 $q(x)$ で事象が生起しているときに、それとは異なる確率分布 $p(x)$ によって情報量を求めたときの情報量の期待値をクロスエントロピー（cross entropy）と呼び、$H(q, p)$ で表す。

$$H(q, p) = E_{q(x)}\left[-\log p(x)\right] = -\int q(x) \log p(x)\, dx \tag{12}$$

クロスエントロピーという名称は 2 つの分布 q と p をクロスさせた（交じり合わせた）エントロピーというところから来ている。

4. KL ダイバージェンス

クロスエントロピー $H(q, p)$ とエントロピー $H(p)$ との差を **KL ダイバージェンス**（カルバック・ライブラー情報量： Kullback-Leibler divergence）と呼び、$D\left(q||p\right)$ で表す。q と p の間に縦棒を 2 つ入

れるのが慣例である。これは縦棒ひとつでは条件付確率と紛らわしいためである。

$$
\begin{aligned}
D\left(q||p\right) &= H(q,\, p) - H(q) \qquad (13)\\
&= -\int q(x)\log p(x)\,dx + \int q(x)\log q(x)\,dx\\
&= \int q(x)\left(\log q(x) - \log p(x)\right)dx\\
&= \int q(x)\log\frac{q(x)}{p(x)}\,dx \qquad (14)
\end{aligned}
$$

KL ダイバージェンスの非負性を示すために以下の不等式を示す ： $\alpha \geq 0$ のとき

$$\alpha - 1 \geq \log\left(\alpha\right) \qquad (15)$$

（証明）

$y = \alpha - 1 - \log\left(\alpha\right)$ とおいて、α で微分すれば

$$\frac{dy}{d\alpha} = 1 - \frac{1}{\alpha}, \qquad \frac{d^2y}{d\alpha^2} = \frac{1}{\alpha^2} > 0$$

となり、y は下に凸で $\alpha = 1$ で最小値 0 をとる。したがって

$$\alpha - 1 - \log\left(\alpha\right) \geq 0$$

となり、不等式 (15) を得る。■

不等式 (15) と $\int p(x)dx = \int q(x)dx = 1$ を用いると

$$
\begin{aligned}
D(q||p) &= \int q(x)\log\frac{q(x)}{p(x)}\,dx\\
&= -\int q(x)\log\frac{p(x)}{q(x)}\,dx\\
&\geq -\int q(x)\left(\frac{p(x)}{q(x)} - 1\right)dx\\
&= -\int \left(p(x) - q(x)\right)dx\\
&= 0 \qquad (16)
\end{aligned}
$$

となり、**KL ダイバージェンスの非負性** ： $D(q||p) \geq 0$ が示された。

式 (13) と $D(q||p) \geq 0$ より、クロスエントロピー $H(q,\, p)$ はエントロピー $H(p)$ よりも大きいか等しいことが分かる。

2 つの確率分布 p と q が完全に一致するとき、KL ダイバージェンス $D(q||p)$ は 0 となり、異なれば異なるほど大きくなっていく。これは相違度として好ましい性質である。このため機械学習においては、正解との相違度を表わす損失関数として KL ダイバージェンスが用いられることが多い。

参考文献

1) 手塚 太郎 『しくみがわかる深層学習』 朝倉書店 (2018/6/25)

2）瀧 保夫 『情報理論 I -情報伝達の理論-』 岩波全書 306 (1982/1/10)