2019 年 11 月 27 日
慶應義塾大学理工学部物理学科
岡崎健人

卒研ゼミ「深層学習」

3　確率と情報理論

3.13　情報理論

「地球は自転している」という情報よりも「明日地球に隕石が衝突する」という情報の方が意外性が高く、より大きな価値をもつ気がする。このような直感的な、「情報の価値」的なものを数学的に表現したい。具体的には

- 起こりやすい事象の情報量は少なく、確実に起こる事象の情報量はないとする
- 起こるのが珍しい事象ほど情報量は大きい
- 独立な事象については情報量は足し算で表される。つまり「コイントスを 2 回行ったところ表が 2 回出た」という事象の情報は、「コイントスを 1 回行ったところ表が出た」という事象のそれよりも、2 倍の大きさをもつ

というようなものを数式で表現する。この 3 つの性質を満たすためには次のような量を作ればよい。

$$I(x) = -\log P(x).$$

これを事象 $\mathrm{x} = x$ の自己情報量といい、対数の底は e で、単位はナット（nats）という。底を 2 にしたものの単位はビット（bits）あるいはシャノン（shannons）である。

自己情報量は 1 つの事象のみについての情報量であるが、全体の情報量の平均をシャノン・エントロピーあるいは平均情報量という。

$$H(\mathrm{x}) = H(P) = \mathbb{E}_{\mathrm{x}\sim P}[I(x)] = -\sum_i P(x_i)\log P(x_i).$$

たとえば確率 p で表が出るコインでコイントスをするときのシャノン・エントロピーは $H(P) = -p\log p - (1-p)\log(1-p)$ となる。この関数を図示すると図 1（原本の図 3.5）のようになり、$p = 0, 1$（結果が確実にわかっている）ときに $H(P) = 0,\ p = 1/2$（結果は不確実）のときに最大値を取ることがわかる。一般に離散型確率変数に対する確率分布のシャノン・エントロピーが最大になるのは一様分布のときである。

同じ確率変数 x に対して異なる確率分布 $P(\mathrm{x})$ と $Q(\mathrm{x})$ があったとき、それらがどれほど異なるのかを表す量として KL ダイバージェンス（Kullback–Leibler divergence）がある。

$$D_{\mathrm{KL}}(P\|Q) = \mathbb{E}_{\mathrm{x}\sim P}\left[\log\frac{P(x)}{Q(x)}\right] = \sum_i P(x_i)\log\frac{P(x_i)}{Q(x_i)}.$$

KL ダイバージェンスは非負の量であり、等号成立は P と Q が同じときである。よってこれは P と Q の距離のような概念と考えられるが、一般に $D_{\mathrm{KL}}(P\|Q) \neq D_{\mathrm{KL}}(Q\|P)$ であるため距離の公理の 1 つ $d(x,y) = d(y,x)$ を満たさず、距離と呼ぶのは正しくない。

図 2（原本の図 3.6）に KL ダイバージェンスの非対称性 $D_{\mathrm{KL}}(p\|q) \neq D_{\mathrm{KL}}(q\|p)$ による効果を示す。$p(x)$ は 2 つのガウス分布の重ね合わせで、$q(x)$ はある 1 つのガウス分布である。イメージとしてはそれぞれ

$$p(x) = \frac{1}{2\sqrt{2\pi}\sigma_0}\left[\exp\left(-\frac{(x-\mu_0)^2}{2{\sigma_0}^2}\right) + \exp\left(-\frac{(x+\mu_0)^2}{2{\sigma_0}^2}\right)\right],$$
$$q(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

のような関数としている。ここで $D_{\mathrm{KL}}(p\|q)$ と $D_{\mathrm{KL}}(q\|p)$ をそれぞれ計算するとどちらも定数 μ_0, σ_0 を含む μ, σ の関数になるが、その関数形は異なっているはずである。したがって $D_{\mathrm{KL}}(p\|q)$ と $D_{\mathrm{KL}}(q\|p)$ を最小にする μ, σ の値も異なり、結果として q^* の形は図 2（原本の図 3.6）のように違いが生じる。ただし図 2 の右側のグラフでは、q^* の平均 μ が $-\mu_0$ に一致している状況を描いていて、それは μ_0 としても $D_{\mathrm{KL}}(q\|p)$ の値は変わらない。ここでは高さの同じ 2 つのガウス関数が混じった分布を、KL ダイバージェンスを尺度として、1 つのガウス分布で近似するということをしている。高さが同じであったので少しわかりにくかったが、たとえば $p(x)$ の片方のピークが十分小さいときには「ちょっとイビツなガウス分布」を「完璧なガウス分布」で近似することができるため、役に立つ方法といえる。そのときに $D_{\mathrm{KL}}(p\|q)$ と $D_{\mathrm{KL}}(q\|p)$ のどちらを選ぶかが重要になる。

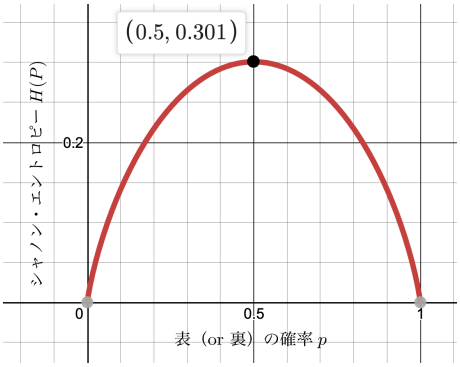

図 1　ベルヌーイ分布のシャノン・エントロピー

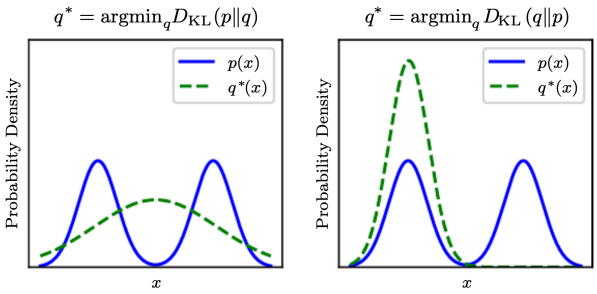

図 2　KL ダイバージェンスの非対称性

また

$$H(P,Q) = H(P) + D_{\mathrm{KL}}(P\|Q)$$

という量を交差エントロピーという。これを少し式変形をすると

$$\begin{aligned} H(P,Q) &= H(P) + D_{\mathrm{KL}}(P\|Q) \\ &= -\sum_i P(x_i)\log P(x_i) + \sum_i P(x_i)\log\frac{P(x_i)}{Q(x_i)} \\ &= -\sum_i P(x_i)\log Q(x_i) \end{aligned}$$

となり、これは $\mathrm{x} \sim P$ のもとでの $Q(x)$ のシャノン・エントロピー $\mathbb{E}_{\mathrm{x}\sim P}[-\log Q(x)]$ を表していることがわかる。KL ダイバージェンスと比較して取り除かれている部分 $\sum_i P(x_i)\log P(x_i)$ は Q に依存しないため、

交差エントロピーを Q に関して最小化することは KL ダイバージェンスを最小化することと等価である。

計算をするなかで $0\log 0$ の形の式が現れるが、これは $\lim_{x\to 0^+} x\log x = 0$ と解釈する。

証明

自己情報量が $-\log P(x)$ の形になることの証明 [1]. 情報量 f に要求される性質は P と Q を $[0,1]$ 内の変数として

$$\begin{cases} f(PQ) = f(P) + f(Q), \\ f(1) = 0, \\ f(P) > 0 \quad \text{for } P \in [0,1) \end{cases}$$

と書き表わせる。ϵ を正の微小定数として、$Q = 1-\epsilon$ とする。$f(P(1-\epsilon)) = f(P-\epsilon P)$ を $\epsilon = 0$ まわりでテイラー展開すると

$$f(P-\epsilon P) = f(P) - \epsilon P\frac{\mathrm{d}f(P)}{\mathrm{d}P} + \mathcal{O}\big(\epsilon^2\big)$$

となることから、$f(P(1-\epsilon)) = f(P) + f(1-\epsilon)$ より

$$f(1-\epsilon) \approx -\epsilon P\frac{\mathrm{d}f(P)}{\mathrm{d}P}. \qquad \therefore\ \frac{\mathrm{d}f(P)}{\mathrm{d}P} \approx -\frac{1}{P}\frac{f(1-\epsilon)}{\epsilon} \xrightarrow{\epsilon\to 0} -\frac{k}{P},$$

ただし $f(1-\epsilon)/\epsilon$ の $\epsilon \to 0$ における極限を k とおいた。よって f に関する微分方程式が得られたのでこれを解くと $f(P) = -k\log P + C$ となる。$f(1) = 0$ より積分定数はゼロであることがわかるので、$f(P) = -k\log P,\ f(P) \geq 0$ より k は正の定数であればよい。

□

KL ダイバージェンスの非負性 $D_{\mathrm{KL}}(P\|Q) \geq 0$ の証明 [2]. 一般に $y > 0$ に対して $\log y \leq y - 1$ である。y を $Q(x_i)/P(x_i)$ とすれば

$$\log\frac{Q(x_i)}{P(x_i)} \leq \frac{Q(x_i)}{P(x_i)} - 1.$$

両辺に $-P(x_i)$ をかけることにより

$$P(x_i)\log\frac{P(x_i)}{Q(x_i)} \geq P(x_i) - Q(x_i)$$

を得る。両辺の総和をとれば

$$\sum_i P(x_i)\log\frac{P(x_i)}{Q(x_i)} \geq \sum_i P(x_i) - \sum_i Q(x_i) = 1 - 1 = 0$$

より $D_{\mathrm{KL}}(P\|Q) \geq 0$ が示された。なお等号が成立するのは $\log y = y - 1$ を解いて $y = 1$ であるとき、すなわち $P(x_i) = Q(x_i)$ となるときである。 □

$\lim_{x\to 0^+} x\log x = 0$ の証明.

$$\lim_{x\to 0^+} x\log x = -\lim_{x\to 0^+}\frac{-\log x}{1/x} = -\lim_{x\to 0^+}\frac{-1/x}{-1/x^2} = -\lim_{x\to 0^+} x = 0^-.$$

途中で ∞/∞ の不定形が現れるのでロピタルの定理を用いた。 □

「シャノン・エントロピーが最大となるのは一様分布のとき」の証明. 第 4 章で最大化（最小化）問題を扱う場面があるので、ラグランジュの未定乗数法の復習を兼ねて載せておく。離散型確率変数 $i = 1, 2, \ldots, n$ に対する確率を p_i とする。シャノン・エントロピーは

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

ただし確率の総和が 1 であるという制約が付いている。

$$\sum_{i=1}^{n} p_i - 1 = 0.$$

なお確率を並べたベクトルを $\boldsymbol{p} = [p_1, p_2, \ldots, p_n]^{\mathsf{T}}$ とした。この等式制約のもとでの最大化問題はラグランジュの未定乗数法を用いて解くことができる。ラグランジュ乗数を λ として、ラグランジュ関数を

$$L(\boldsymbol{p}, \lambda) = -\sum_{i=1}^{n} p_i \log p_i + \lambda \left(\sum_{i=1}^{n} p_i - 1 \right)$$

と作る。このとき

$$\begin{cases} \dfrac{\partial L}{\partial \boldsymbol{p}} = \boldsymbol{0}, \\ \dfrac{\partial L}{\partial \lambda} = 0 \end{cases}$$

という条件を満たす点 $(\boldsymbol{p}, \lambda)$ が $H(\boldsymbol{p})$ の最大値を与える。具体的には

$$\begin{cases} \dfrac{\partial L}{\partial p_i} = -\log p_i - 1 + \lambda = 0, \qquad i = 1, 2, \ldots, n, \\ \dfrac{\partial L}{\partial \lambda} = \displaystyle\sum_{i=1}^{n} p_i - 1 = 0, \end{cases}$$

となる。第 1 式より $p_i = e^{\lambda - 1}$, これを第 2 式に代入することで $ne^{\lambda - 1} - 1 = 0$ となるので $e^{\lambda - 1} = p_i = 1/n$ が得られる。したがってシャノン・エントロピーを最大にする離散型確率分布は一様分布である。

□

参考文献

[1] 赤間世紀.「情報理論入門」. 株式会社工学社, 2010 年, pp. 9–12.

[2] 堀部安一.「情報エントロピー論」. 第 2 版, 森北出版株式会社, 1997 年. pp. 85–92.