第 1 章

証明

自己情報量が $-\log P(x)$ の形になることの証明. 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. 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。

□

なぜ $J(\boldsymbol{w}) = \mathrm{MSE}(\boldsymbol{w}) + \lambda \boldsymbol{w}^\top \boldsymbol{w}$ **か？.** 目的関数を多項式とするときの係数のベクトル $\boldsymbol{w} = [w_0, w_1, \ldots, w_n]^\mathsf{T}$ の成分が大きいと、目的関数の曲がり方が激しくなって、過学習に陥ってしまう。そこで t をある定数として、$\|\boldsymbol{w}\|_2^2 = \boldsymbol{w}^\top \boldsymbol{w} \leq t$ という制約のもとでの平均二乗誤差 $\mathrm{MSE}(\boldsymbol{w}) = \frac{1}{m}\|X\boldsymbol{w} - \boldsymbol{y}\|_2^2 = \frac{1}{m}(X\boldsymbol{w} - \boldsymbol{y})^\mathsf{T}(X\boldsymbol{w} - \boldsymbol{y})$ の最小化を考える。これは KKT 法を用いれば解ける。一般化ラグランジュ関数を

$$L(\boldsymbol{w}, \lambda) = \frac{1}{m}(X\boldsymbol{w} - \boldsymbol{y})^\mathsf{T}(X\boldsymbol{w} - \boldsymbol{y}) + \lambda\big(\boldsymbol{w}^\top \boldsymbol{w} - t\big)$$

として KKT 条件は

$$\begin{cases} \dfrac{\partial L}{\partial \boldsymbol{w}} = \dfrac{2}{m}\big(X^{\mathsf{T}}X\boldsymbol{w} - X^{\mathsf{T}}\boldsymbol{y}\big) + 2\lambda\boldsymbol{w} = \boldsymbol{0}, \\ \dfrac{\partial L}{\partial \lambda} = \boldsymbol{w}^{\top}\boldsymbol{w} - t \le 0, \\ \quad \lambda\big(\boldsymbol{w}^{\top}\boldsymbol{w} - t\big) = 0, \\ \quad \lambda \ge 0 \end{cases}$$

となる。$\lambda = 0$ の場合は最適な解 $\boldsymbol{w}$ がもともと $\boldsymbol{w}^{\top}\boldsymbol{w} \le t$ を満たしていたことを表す。もし $\lambda > 0$ ならば $\boldsymbol{w}^{\top}\boldsymbol{w} = t$ だが、$\boldsymbol{w}$ は第 1 式より $\boldsymbol{w} = \big(X^{\top}X + m\lambda I\big)^{-1}X^{\top}\boldsymbol{y}$ である。したがって

$$t = \boldsymbol{w}^{\top}\boldsymbol{w} = \Big[\big(X^{\top}X + m\lambda I\big)^{-1}X^{\top}\boldsymbol{y}\Big]^{\top}\big(X^{\top}X + m\lambda I\big)^{-1}X^{\top}\boldsymbol{y}.$$

これより λ の値を決めれば t の値は自動的に決まることがわかる。したがって、λ をパラメータとして最初から一般化ラグランジュ関数 $L(\boldsymbol{w}, \lambda)$ から $-\lambda t$ の項を省いた $J(\boldsymbol{w}) = \mathrm{MSE}(\boldsymbol{w}) + \lambda\boldsymbol{w}^{\top}\boldsymbol{w}$ という関数の最小化をすればよい。

□