ポアソン・ガンマ階層モデルのまとめ

酒匂貴市

平成 27 年 1 月 29 日

目 次

第 I 部
本編

1　ポアソン回帰

1.1　ポアソン分布

件数（カウントデータ）の確率分布 X を考えるとき、単位時間の平均発生回数が λ であるものに対しては、ポアソン分布 $Po(\lambda)$ が適当であることが多い。このことは、次の議論からおおよそつかむことができる。

観測期間を $h > 0$ としたとき、観測期間内の平均発生回数が h によらず $h\lambda$ であるとする。このとき、h を微小な期間とすれば、ベルヌーイ試行と考えてもよさそうである。そこで、単位時間を n 分割し、$h = \frac{1}{n}$ とすると、単位時間での件数の確率分布は、二項分布 $B(n, h\lambda)$ に従う [1]。このとき、確率変数は

$$\begin{aligned} p(x) &= \frac{n!}{(n-x)!x!}(h\lambda)^x(1-h\lambda)^{(n-x)} \\ &= \frac{n(n-1)\cdots(n-x+1)}{x!}(\lambda)^x(n-\lambda)^{-x}(1-h\lambda)^n \\ &= \frac{n(n-1)\cdots(n-x+1)}{(n-\lambda)^x}\frac{\lambda^x}{x!}(1-h\lambda)^n \\ &\to 1\ \frac{\lambda^x}{x!}e^{-\lambda} = e^{-\lambda}\frac{\lambda^x}{x!} \end{aligned}$$

であり、ポアソン分布となる。

ポアソン分布は再生性を持つことが重要である。ポアソン分布において再生性とは、$X \sim Po(\lambda_1), Y \sim Po(\lambda_2)$ で X, Y が独立であるとき、$(X+Y) \sim Po(\lambda_1 + \lambda_2)$ となることをいう。

1.2　ポアソン分布における推定・分析

次のような問題設定を考える。データは区分されており、各区分を添字 j で表すことにする。各区分ごとに、エクスポージャー [2] N_j に対して対象とする事象の発生件数が n_j であるとする。n_j はポアソン分布に従っており、発生率のパラメータ λ_j によって

$$n_j \sim Po(N_j\lambda_j)$$

と表されるとする。単に λ_j を推定するだけならば、$\frac{n_j}{N_j}$ がよい推定量（最尤推定量・不偏推定量）である。

1.2.1　ポアソン回帰

実際のデータの分析では、区分自体がほかのファクターの組み合わせ [3] であったりする。このような場合、結果を利用しやすいことから次のように乗法モデルを設定することがよく行われる。

$$\lambda_j = b\ f_1(j)\cdots f_M(j)$$

b は基準となる発生率（以下、ベースライン発生率とする。）であり、$f_1\ \cdots f_M$ は各ファクターごとの調整係数である。f_m はベースラインに対応するとき $f_m = 1$ であるよう正規化されている必要がある。

---

[1] 平均 p で試行回数 n の二項分布を $B(n,p)$ で表すものとする。
[2] 保険データならば経過契約件数など。
[3] 例えば、死亡数であれば、区分は性・年齢といったファクターに基づいて決められていることが多いだろう。

ファクターを変数 $\{y_1, \cdots, y_m\}$ によって $f_m = e^{\gamma_m y_m(j)}$ と表現しておけば、

$$\log(\lambda_j) = \log(b) + \gamma_1 y_1(j) + \cdots + \gamma_M y_M(j)$$

という形で表現できる。この表現に基づく分析は、一般的にポアソン回帰と呼ばれているものである。ベースライン発生率は $y_1 = \cdots = y_M = 0$ に対応するものとなる。

1.2.2 ポアソン回帰の課題

ポアソン回帰は、発生率 λ_j が既知のファクター $\{y_1, \cdots, y_m\}$ の乗法モデルにより完全に決定されると仮定している。しかし、実際に分析をするうえでは、発生率に影響するファクターをすべて把握することが難しいうえ、乗法モデルで表されるとも限らない。また、これらの課題のため、未知のファクターによる変動を既知のファクターで無理やり合わせにいこうとすることにより、過学習となる懸念もある。

2 階層モデル

上述のポアソン回帰の課題に対応するひとつの方法として、発生率 λ_j に確率的な変動を認める方法がある。まず、λ_j の条件付においては、ポアソン分布に従うものとする。

$$n_j|\lambda_j \sim Po(N_j\lambda_j)$$

その上で、λ_j を $\mathrm{E}[\lambda_j] = b\ f_1(j) \cdots f_M(j)$ なる非負確率変数として考えるというものである。λ_j の密度関数を g_j とすると n_j の確率関数は

$$p(x) = \int_0^\infty d\lambda\ e^{-N_j\lambda} \frac{(N_j\lambda)^x}{x!} g_j(\lambda)$$

である [4]。

これは、階層ベイズモデルの一種でもある。問題は、λ_j の分布として何を用いるかである。何らかの事前情報がある場合には、それを反映することがよりよい推定につながるだろう。特に事前情報が無い場合が問題となる。

2.1 対数正規分布

ポアソン回帰における λ_j の対数に関する関係式

$$\log(\lambda_j) = \log(b) + \gamma_1 y_1(j) + \cdots + \gamma_M y_M(j)$$

は線形回帰の形状に似ており、そのアナロジーから $\log(\lambda_j)$ に対して正規分布を適用することは自然な方法のひとつである。この場合は、つまり、λ_j の分布を対数正規分布

$$LN\left(\log(b) + \gamma_1 y_1(j) + \cdots + \gamma_M y_M(j) - \frac{\sigma^2}{2}, \sigma^2\right)$$

とおくことに相当する。この場合の n_j の確率関数は簡単な形では表現できない。対数正規分布を用いた場合は、MCMC（マルコフ連鎖モンテカルロ法）によって分析を行うことになるだろう。

[4] n_j と λ_j の同時分布として考えないことにより、λ_j に変動性を与えている。

3　エントロピーとガンマ分布

3.1　ガンマ分布の微分エントロピー

ガンマ分布の微分エントロピーは $Y \sim \Gamma(\alpha, \beta)$ として

$$\begin{aligned} H(Y) &= -\mathrm{E}\left[\alpha \log \beta - \log \Gamma(\alpha) - \beta Y + (\alpha - 1) \log Y\right] \\ &= -\alpha \log \beta + \log \Gamma(\alpha) + \beta \mathrm{E}\left[Y\right] - (\alpha - 1) \mathrm{E}\left[\log Y\right] \\ &= -\alpha \log \beta + \log \Gamma(\alpha) + \alpha - (\alpha - 1) \mathrm{E}\left[\log Y\right] \end{aligned}$$

である。$\mathrm{E}\left[\log Y\right]$ については、$Z \equiv \log Y$ として、Z のモーメント母関数 $M(\theta)$ を考える。

$$\begin{aligned} M(\theta) &= E\left[e^{\theta Z}\right] \\ &= E\left[Y^{\theta}\right] \\ &= \int dy y^{\theta} \frac{\beta^{\alpha}}{\Gamma(\alpha)} e^{-\beta y} y^{\alpha - 1} \\ &= \int dy \frac{\beta^{\alpha}}{\Gamma(\alpha)} e^{-\beta y} y^{\alpha + \theta - 1} \\ &= \frac{\beta^{\alpha}}{\Gamma(\alpha)} \frac{\Gamma(\alpha + \theta)}{\beta^{\alpha + \theta}} \\ &= \frac{\Gamma(\alpha + \theta)}{\Gamma(\alpha) \beta^{\theta}} \end{aligned}$$

より

$$M'(\theta) = \frac{\Gamma'(\alpha + \theta) - \Gamma(\alpha + \theta) \log \beta}{\Gamma(\alpha) \beta^{\theta}}$$

なので

$$\mathrm{E}\left[\log Y\right] = M'(0) = \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} - \log \beta$$

である。よってガンマ分布の微分エントロピーは

$$\begin{aligned} H(Y) &= -\alpha \log \beta + \log \Gamma(\alpha) + \alpha - (\alpha - 1) \mathrm{E}\left[\log Y\right] \\ &= \log \Gamma(\alpha) + \alpha - (\alpha - 1) \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} - \log \beta \end{aligned}$$

である。

ここで、非負連続確率変数 X を考え、ガンマ分布 Y との Kullback Leibler divergence を計算する。

$$\begin{aligned} D(X \| Y) &= -H(X) - \mathrm{E}\left[\log \left(\frac{\beta^{\alpha}}{\Gamma(\alpha)} e^{-\beta X} X^{\alpha - 1}\right)\right] \\ &= -H(X) - \alpha \log \beta + \log \Gamma(\alpha) + \beta \mathrm{E}\left[X\right] - (\alpha - 1) \mathrm{E}\left[\log X\right] \end{aligned}$$

であり、$H(Y) = -\alpha \log \beta + \log \Gamma(\alpha) + \beta \mathrm{E}\left[Y\right] - (\alpha - 1) \mathrm{E}\left[\log Y\right]$ なので

$$\begin{aligned} D(X \| Y) = -H(X) + H(Y) + \beta \left(\mathrm{E}\left[X\right] - \mathrm{E}\left[Y\right]\right) - (\alpha - 1) \left(\mathrm{E}\left[\log X\right] - \mathrm{E}\left[\log Y\right]\right) \geq 0 \\ H(Y) + \beta \left(\mathrm{E}\left[X\right] - \mathrm{E}\left[Y\right]\right) - (\alpha - 1) \left(\mathrm{E}\left[\log X\right] - \mathrm{E}\left[\log Y\right]\right) \geq H(X) \end{aligned}$$

が成立する。

まず、$\alpha = 1$ の場合、つまり指数分布について考えると、X と Y の平均が等しいとき

$$H(Y) \geq H(X)$$

となる。つまり、非負連続確率分布で、平均を固定した場合に、微分エントロピーを最大化するものは指数分布である。したがって、非負連続確率分布であること以外にとくに情報が無い場合、指数分布を採用することは、ひとつの考え方である。ただし、指数分布は 0 に近いほど確率が大きいという特徴的な形状をしており、これが適当であるかはひとつの観点として考えるべきである。特に、現在考えている階層モデルにおいては、0 に近いほど確率が大きいという形状は適当でないと考えられる。

$\alpha \neq 1$ の場合には、平均値に加えて、対数の平均値が等しい（$\mathrm{E}[\log X] = \mathrm{E}[\log Y]$）場合に

$$H(Y) \geq H(X)$$

となる。つまり、平均値と対数の平均値を固定した場合、ガンマ分布が微分エントロピーを最大化する。この意味を考えるにおいては、対数変換した確率分布を検討することがよい。

3.2 ガンマ分布の対数変換と乗法モデル

定理 3.1 $Z = \log X$ のとき、$H(Z) = H(X) - \mathrm{E}[\log X]$ である。

(proof)
X の密度関数を p とするとき、Z の密度関数を $g(z)$ とすると $g(z) = p(e^z)e^z$ である。よって

$$\begin{aligned} H(Z) &= -\mathrm{E}\left[\log\left(p(e^Z)e^Z\right)\right] \\ &= -\mathrm{E}\left[\log\left(p(X)X\right)\right] \\ &= -\mathrm{E}\left[\log p(X) + \log X\right] \\ &= H(X) - \mathrm{E}[\log X] \end{aligned}$$

である。 証明終

よって、ガンマ分布 $Y \sim \Gamma(\alpha, \beta)$ に対して、対数変換した $Z = \log Y$ のエントロピーは

$$H(Z) = -\alpha \log \beta + \log \Gamma(\alpha) + \alpha - \alpha \mathrm{E}\left[Z\right]$$

である。また、$\mathrm{E}[X] = \mathrm{E}[Z]$ かつ $\mathrm{E}[e^X] = \mathrm{E}[e^Z]$ なる任意の連続確率変数 X について

$$\begin{aligned} H(X) &= H(e^X) - \mathrm{E}[X] \\ &\leq H(Y) - \mathrm{E}[X] \quad \because \text{ガンマ分布が微分エントロピーを最大化する} \\ &= H(Z) + \mathrm{E}[Z] - \mathrm{E}[X] \\ &= H(Z) \end{aligned}$$

である。つまり、平均と指数をとった平均を固定したときには、ガンマ分布の対数変換が微分エントロピーを最大化する。

これに対して、正規分布は、分散を固定したときに微分エントロピーを最大化するものである。つまり、平均を固定し、さらに分散を固定したときに微分エントロピーを最大化するのが正規分布であり、**分散の代わりに $\mathrm{E}[e^X]$ を固定したときに微分エントロピーを最大化するのがガンマ分布の対数変換である。**

乗法モデルを e^X の形で表現するのは自然であることから、乗法モデルにおいては、このときの X について分散が同じものを考えるよりも、e^X の平均が同じ分布の中で微分エントロピーが最大となる分布を考えるほうが合理的という考え方もあり得る。したがって、乗法モデルで採用すべき分布について情報がない場合に、ガンマ分布を採用することはエントロピー（微分エントロピー）の観点からは自然・合理的であるといえよう。

4　ポアソン・ガンマ階層モデル

4.1　ポアソン・ガンマ階層モデル

上述の階層モデルでは、発生率 λ_j について、その平均 $\mathrm{E}[\lambda_j] = b\ f_1(j) \cdots f_M(j)$ を固定した中で考えている。よって、これまでの議論により、発生率の分布に特段の情報を持たない場合は、ガンマ分布を採用することが適当であろう。このとき、この階層モデルは次のようなものとなる。これを、ポアソン・ガンマ階層モデルと呼ぶものとする。

$$
\begin{aligned}
n_j|\lambda_j &\sim Po(N_j\lambda_j) \\
\lambda_j &\sim \Gamma(\alpha_j, \beta_j) \\
\frac{\alpha_j}{\beta_j} &= b\ f_1(j) \cdots f_M(j)
\end{aligned}
$$

このとき、n_j の分布は負の二項分布に従うことが示される。n_j の確率関数を $p(x)$ とすると

$$
\begin{aligned}
p(x) &= \int_0^\infty d\lambda\ e^{-N_j\lambda} \frac{(N_j\lambda)^x}{x!} \frac{\beta_j^{\alpha_j}}{\Gamma(\alpha_j)} e^{-\beta_j\lambda}\lambda^{\alpha_j-1} \\
&= \frac{\beta_j^{\alpha_j} N_j^x}{\Gamma(\alpha_j)\ x!} \int_0^\infty d\lambda\ e^{-(\beta_j+N_j)\lambda}\lambda^{x+\alpha_j-1} \\
&= \frac{\beta_j^{\alpha_j} N_j^x}{\Gamma(\alpha_j)\ x!} \frac{\Gamma(x+\alpha_j)}{(\beta_j+N_j)^{x+\alpha_j}} \\
&= \frac{\Gamma(x+\alpha_j)}{\Gamma(\alpha_j)\ x!} \left(\frac{\beta_j}{\beta_j+N_j}\right)^{\alpha_j} \left(\frac{N_j}{\beta_j+N_j}\right)^x
\end{aligned}
$$

であり、負の二項分布 $NB\left(\alpha_j, \frac{\beta_j}{\beta_j+N_j}\right)$ にしたがっている。平均は

$$
N_j \frac{\alpha_j}{\beta_j} = N_j b\ f_1(j) \cdots f_M(j)
$$

であり、分散は

$$
N_j \frac{\alpha_j}{\beta_j} \frac{\beta_j+N_j}{\beta_j}
$$

である。発生率 λ に変動を許すことにより、分散が増えている。

4.2　信頼性推定量

n_j と λ_j の同時分布の密度関数を $f(n_j, \lambda_j)$ とすると

$$
f(n_j, \lambda_j) = e^{-N_j\lambda} \frac{(N_j\lambda)^x}{x!} \frac{\beta_j^{\alpha_j}}{\Gamma(\alpha_j)} e^{-\beta_j\lambda}\lambda^{\alpha_j-1}
$$

である。n_j の観測値を得ると、λ_j の事後分布を求めることができる。事後分布の密度関数を $f(\lambda_j|n_j)$ とすると

$$
\begin{aligned}
f(\lambda_j|n_j) &\propto e^{-N_j\lambda}\lambda^{n_j} e^{-\beta_j\lambda}\lambda^{\alpha_j-1} \\
&\propto e^{-(\beta_j+N_j)\lambda}\lambda^{\alpha_j+n_j-1}
\end{aligned}
$$

なので

$$
\lambda_j|n_j \sim \Gamma(\alpha_j+n_j, \beta_j+N_j)
$$

であり、その平均は

$$
\frac{\alpha_j+n_j}{\beta_j+N_j}
$$

である。これは、$Z_j \equiv \frac{N_j}{N_j+\beta_j}$ とすると

$$\frac{\alpha_j + n_j}{\beta_j + N_j} = Z_j \frac{n_j}{N_j} + (1 - Z_j)\frac{\alpha_j}{\beta_j}$$

となる。これは Z_j を信頼度とし、観測値とモデル推定値を混合した信頼性推定量となる。

モデル推定値と信頼性推定量の差は、発生率に変動を認めたことから生じており、織り込まれていないファクターの影響や乗法モデルに伴う誤差などが含まれていると考えられる。

また、問題設定に多少差異があるものの

$$\beta_j = \frac{\mathrm{E}[\lambda_j]}{V[\lambda_j]}$$

であり、ビュールマン (Bühlmann) モデルにおける信頼度 (Bühlmann Credibility Factor) と同様の形になっていることも、注目に値する。

4.3 乗法モデルにおける β_j の設定

上の信頼性推定量において、$\beta_j = N_j$ のとき、$Z_j = 0.5$ であり、観測値とモデル推定値の信頼性が同等であることになる。このことから、β_j はモデル推定値の信頼性を観測値のエクスポージャーの量（経過契約数など）に換算して表現したものといえる。

いま、発生率の平均値については、$\frac{\alpha_j}{\beta_j} = b\ f_1(j) \cdots f_M(j)$ という乗法モデルの形で置くものとしている。このときに β_j をどのようにすべきかは自明でない。そのまま β_j を各区分に対して全ておくこともちろん可能である。この場合、パラメータ数が多くなりすぎることがデメリットである。

これ以外のひとつの考え方としては、乗法モデルにおける各ファクターに対応させるというものがある。つまり、ベースラインに対応する β_0 と、各ファクターごとに対応する $\beta_1, \cdots, \beta_M$ によって

$$\beta_j = \beta_0\ \beta_1(j) \cdots \beta_M(j)$$

とおく。$\beta_1, \cdots, \beta_M$ はベースラインに対応するとき 1 になるよう正規化されている必要がある。この方法のメリットは、パラメータが少なくてすむことと、各ファクターごとの調整係数 f_m の信頼性がどの程度のものか見える点にある。

4.4 excel における計算

ポアソン・ガンマ階層モデルのパラメータの推定は、非線形最適化問題であるため、一般論としてはそんなに容易でないが、パラメータが多くなりすぎないようにすれば、excel でも十分計算できる。これは、企業内で実務を行う際には小さくないメリットである [5]。

パラメータの推定においては、$q_j \equiv \frac{N_j}{\beta_j+N_j}$ とし、対数尤度のうちパラメータに関係ある部分である

$$l \equiv \sum_j \{\log\Gamma(n_j + \alpha_j) - \log\Gamma(\alpha_j) + \alpha_j \log(1 - q_j) + n_j \log q_j\}$$

を最大化することで最尤推定すればよい。この値自体は、GAMMALN 関数や LN 関数を使うことで、excel の表計算で計算できる。最尤推定値は、solver で l の値を最大化することで求められる。このようにすれば、

[5] 一般的に、システム部門の許可無くソフトウェアをインストールできないことが多いためである。

マクロすら組むことなく計算が可能である。

ただし、solver は計算が速くないため、パラメータはある程度あたりをつけておく必要がある。そのためには、まず最初にポアソン回帰を行い、パラメータの当たりをつけるとともに、有効でないパラメータを取捨選択することが考えられる。ポアソン・ガンマ階層モデルでは、ベースライン発生率 b、各ファクターに対する調整係数 $f_1, \cdots, f_M$ および β_j といったパラメータが必要であるが、ポアソン回帰では、推定すべきパラメータは各ファクターに対する調整係数 $f_1, \cdots, f_M$ のみとなり、計算がだいぶ速くできるためである [6]。

[6] ベースライン発生率は観測値と調整係数から算式で計算できる。詳細は読者確かめよ。

第 II 部
付録

A　エントロピーと関連する情報理論の概略

A.1　符号化

まず、離散的な事象を考え、それぞれの事象に対して Q 進数の符号を割り当てる。例えば $Q = 2$ ならば、次のような符号があり得る。以下 $Q = 2$ の場合で例を記述する。

$$0, 1, 00, 01, 10, 11, 000, 001, \cdots$$

事象 j に割り当てられた符号の長さを l_j とする。

この確率的事象を繰り返して観測し、実現した事象を符号の並びによって表現する。例えば、4 つの事象があり得るとし、それぞれに次のような符号が割り当てられているとする。

$$0, 100, 101, 11$$

3 回の繰り返しで実現した事象が順に $11, 0, 100$ だとしたとき、実現した 3 回の事象の繰り返しを 110100 と表現するものとする。

このような符号化を行う場合に、得られた符号を一意に解釈できるようにするには、Q 進数で表されるすべての符合を用いるわけには行かず、使用する符号を限定する必要がある [7]。

ここで、実現した m 回の事象の繰り返しを表す符号を考える。長さが L の符号は、Q 進数においては Q^L 個存在する。したがって、一意に解釈できるためには、長さが L の符号は Q^L 個以下であることがまず必要条件となる。ここで

$$r \equiv \sum_j Q^{-l_j}$$

という量を考える。このとき、実現した m 回の事象の繰り返しを表す符号が一意に解釈できる場合には

$$\begin{aligned}
r^m &= \sum_{j(1)} \cdots \sum_{j(m)} Q^{-\left(l_{j(1)}+\cdots+l_{j(m)}\right)} \\
&= \sum_L \sum_{l_{j(1)}+\cdots+l_{j(m)}=L} Q^{-L} \\
&= \sum_L Q^{-L} \left(\sum_{l_{j(1)}+\cdots+l_{j(m)}=L} 1 \right) \\
&\leq \sum_L Q^{-L} Q^L = \sum_L 1 \\
&\leq L \leq m \max_j(l_j) \\
\frac{r^m}{m} &\leq \max_j(l_j)
\end{aligned}$$

が成立する。この式は符号が一意に解釈できるためには任意の自然数 m に対して成立しなければならない。このとき、$r > 1$ であるとすると、左辺はいくらでも大きくできることになり矛盾する。したがって、$r \leq 1$ すなわち

$$\sum_j Q^{-l_j} \leq 1$$

[7] 実は上の例は一意に解釈できるようにとってあり、$1, 00, 01, 10$ などが使われていない

が成立する。

逆に、$\sum_j Q^{-l_j} \leq 1$ が成立するときを考える。$\{l_j\}$ の重複を整理して小さい順に並べたものを $\{L_1, \cdots, L_K\}$ とし、$\{l_j\}$ のうち長さ L_k なるものの数として $N_k \equiv \sum_{l_j=L_k} 1\ (>0)$ を定義しておく。このとき

$$\sum_j Q^{-l_j} = \sum_k N_k Q^{-L_k} \leq 1$$

である。仮に $N_{k'} \geq Q^{L_{k'}}$ なる k' が存在したとすると

$$\sum_k N_k Q^{-L_k} > N_{k'} Q^{-L_{k'}} \geq Q^{L_{k'}} Q^{-L_{k'}} = 1$$

となり矛盾する。したがって、任意の $k(=1, \cdots, K)$ について

$$N_k < Q^{L_k}$$

が成立する。

まず、長さ L_1 の符号については、上に示したとおりその個数 N_1 は、長さ L_1 の Q 進数の数 Q^{L_1} よりも少ないため、そのうちのひとつをより長い符号のために取っておいた上で、残りで N_1 個の符号を割り当てることができる。例えば、$Q=2, L_1=2$ の場合、長さ $L_1=2$ の符号は

$$00, 01, 10, 11$$

の $4=2^2$ 個あるが、$N_1<4$ であるため、00 をより長い符号のために取っておいた上で、残りの $01, 10, 11$ で N_1 個の符号を割り当てることができる。

この手続きを繰り返せば、符号長がそれぞれ $\{l_j\}$ である一意に解釈できる符号を構成することができる。したがって、上で示したこととあわせて、符号長がそれぞれ $\{l_j\}$ である符号が一意に解釈できることについて

$$\sum_j Q^{-l_j} \leq 1$$

が必要十分条件であることが示される。これは McMillan の不等式と呼ばれている関係式である。

A.2 エントロピー

離散的な確率事象の実現値を符号で表現することとし、事象 j の発生確率を p_j、事象 j を表す符号の長さを l_j とする。確率分布の実現値を表現するための、平均的な符号の長さは

$$f = \sum_j p_j l_j$$

である。確率 p_j を所与としたとき、一意に解釈できる符号で f を最小化することを考える。最小化された f は離散確率分布 $\{p_j\}$ の情報量を表す指標となるだろう。

Q 進数の符号の場合、一意に解釈できる符号の条件は上述の McMillan の不等式であり、$\sum_j Q^{-l_j} \leq 1$ の制約条件化で f を最小化する問題を考えることになる。本来 $\{l_j\}$ には整数であるという制約があるが、まずは整数制約を外して考える。

$\sum_j Q^{-l_j} < 1$ の場合、$c \equiv \sum_j Q^{-l_j}$ とおくと、$l'_j \equiv l_j - \log_Q(C)$ によって

$$\sum_j Q^{-l'_j} = \sum_j Q^{-l_j} C^{-1} = 1$$

が成立し

$$\sum_j p_j l'_j = \sum_j p_j l_j - \log_Q(C) < \sum_j p_j l_j$$

となるため、$\sum_j Q^{-l_j} < 1$ の場合に最小化されることは無い。よって $\sum_j Q^{-l_j} = 1$ を制約条件として考えればよい。これを解けば

$$l_j = -\log_Q(p_j) = -\frac{\log(p_j)}{\log(Q)}$$

となる。実際には整数であるという制約が存在するが、つぎのような考え方により、実質的に無視できることがわかる。まず、Q は 2 以上の整数であれば、何進数であろうと本質的ではない。また、複数回の独立事象をまとめてひとつの事象とみなすことにより、p_j を小さくしていく、すなわち $-\log(p_j)$ を大きくしていくことが可能である。したがって、$-\log(p_j)$ を十分に大きくしたうえで、$\log(Q)$ を調整することで、$-\frac{\log(p_j)}{\log(Q)}$ を整数に近づけることができる。

この議論により、f の最小値として $-\frac{1}{\log(Q)} \sum_j p_j \log(p_j)$ という値を考えることができる。Q に何をとるかは本質的ではないため、離散確率事象の情報量を表す指標として

$$-\sum_j p_j \log(p_j)$$

という量を考えることができる。これはエントロピーと呼ばれている。

以上の議論を踏まえると、エントロピーは、確率分布の実現値の情報を最大限圧縮した場合の平均的な圧縮率に相当する。そのため、確率分布に偏りがある（情報が含まれている）と圧縮率が向上し、エントロピーが小さくなる。例えば、ある事象が確率 1 で発生する（つまり、確定的である）場合には、エントロピーは 0 である。逆に、事象数が有限の離散確率分布においては、「同様に確からしい」場合がエントロピーを最大にする [8]。

A.3 エントロピーの拡張

A.3.1 事象が無限にある離散確率分布

事象数が有限と限らない離散確率分布については、符号化の議論は直接適用できないものの、エントロピー自体は自然な拡張として定義することが可能である。そこで、次のように定義しておく。

定義 A.1 実現値が $\mathcal{X} = \{x_1, \cdots, x_n\}$ で、それぞれの事象の確率が $P(X = x_i) = p_i$ である離散確率変数 X について、**エントロピー**を

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i$$

と定義する。ただし、$0 \log 0 = 0$ と扱うとする [9]。また、実現値はエントロピーの値に関係がないので、確率分布を表わす $\boldsymbol{p} = (p_1, \cdots, p_n)$ によって $H(\boldsymbol{p})$ と表すこともある。事象が可算無限個に及ぶ場合は、無限級数を考える。発散する場合にはエントロピーは ∞ であると考える。◂

[8] $\sum_j p_j = 1$ の制約条件化で $\{p_j\}$ を変数と考え、エントロピーを最大化する問題を考えればよい。
[9] $\lim_{x \to 0} x \log x = 0$ だからである。

補題 A.1 定義より明らかに、エントロピーは確率のみで決まり、確率変数のとる値にはよらない。従って、特に、スケール変換と平行移動に対して不変である、つまり、任意の $a \neq 0, b$ に対して

$$H(aX + b) = H(X)$$

である。

A.3.2 連続確率分布

密度関数 p をもつ連続確率変数 Y にエントロピーの定義を拡張する。$\boldsymbol{R}$ を区間 $I_t = (t\delta, (t+1)\delta)$ に分割すると、平均値の定理より

$$\frac{\int_0^{(t+1)\delta} p(y)dy - \int_0^{t\delta} p(y)dy}{(t+1)\delta - t\delta} = \frac{\int_{I_t} p(y)dy}{\delta} = p(y_t) \quad y_t \in I_t$$

なる y_t が存在するので、これより

$$P(Y^\delta = t) = \int_{I_t} p(y)dy = \delta y_t$$

という離散確率変数をつくると

$$\begin{aligned} H(Y^\delta) &= -\sum_t \delta p(y_t) \log(\delta y_t) \\ &= -\sum_t \delta p(y_t) \log p(y_t) - \log(\delta) \end{aligned}$$

となる。$\delta \to 0$ とすると

$$\sum_t \delta p(y_t) \log(y_t) \to \int p(y) \log p(y) dy$$

$$-\log(\delta) \to \infty$$

であり、$H(Y^\delta)$ そのままでは発散する。そこで、$\lim_{\delta \to 0}\{H(Y^\delta) + \log \delta\} = -\int p(y) \log p(y) dy$ を連続確率変数に対するエントロピーとして採用する。

定義 A.2 密度関数 p をもつ連続確率変数 Y について

$$H(Y) = H(p) = -\int p(y) \log p(y) dy$$

を**微分エントロピー（*differential entropy*）**という。◂

微分エントロピーは、離散のエントロピーとは少し異なる性質を示す。Lemma 1.3 にあたる部分である。

補題 A.2 微分エントロピーは正負いずれもとり得る。また、$-\infty$ に発散することもある。

補題 A.3 微分エントロピーは確率密度関数のみにより、値にはよらない。したがって、平行移動に対しては不変であるが、スケール変換に対しては密度関数そのものが $Y : p(y)$ に対して $aY : \frac{1}{a} p\left(\frac{y}{a}\right)$ と変化するので、任意の a, b に対して

$$H(aY + b) = H(Y) + \log a$$

となる。

a を大きくしていけば、分布は大きく広がり（より確定していない）、微分エントロピーは低下する。逆に a を 0 に近づけていけば、集積した分布（より確定している）となり、微分エントロピーは大きくなる。連続確率分布においては、スケール要素を調整することにより微分エントロピーをいくらでも変化させられることに注意する必要がある。

A.4　エントロピーの最大化

有限の離散確率分布については、同様に確からしい場合がエントロピーが最大となることを述べたが、ほかにも、エントロピー・微分エントロピーが最大となる分布は、分布について詳しい情報が無い場合に自然な分布である。ここでは、その主な結果について述べておきたい。

A.4.1　Kullback Leibler divergence

定義 A.3 二つの確率分布（もしくは確率変数）X, Y について、確率関数もしくは密度関数を p, q とする。このとき、***Kullback Leibler divergence*** を

$$D(X\|Y) = \mathrm{E}\left[\log \frac{p(X)}{q(X)}\right]$$

と定義する。◂

定理 A.4

$$D(X\|Y) \geq 0$$

(proof)
$\log(x) \leq x - 1$ であるので

$$\begin{aligned} D(X\|Y) &= -\mathrm{E}\left[\log \frac{q(X)}{p(X)}\right] \\ &\geq -\mathrm{E}\left[\frac{q(X)}{p(X)} - 1\right] = 0 \end{aligned}$$

である。 証明終

定理 A.5

$$D(X\|Y) = 0 \;\Leftrightarrow\; p(x) = q(x)\ a.e.$$

(proof)
$\log(x) \leq x - 1$（等号条件は $x = 1$）であるので

$$\begin{aligned} \mathrm{E}\left[\left|\log \frac{q(X)}{p(X)} - \frac{q(X)}{p(X)} + 1\right|\right] &= \mathrm{E}\left[\log \frac{q(X)}{p(X)} - \frac{q(X)}{p(X)} + 1\right] \\ &= \mathrm{E}\left[\log \frac{q(X)}{p(X)}\right] \\ &= -D(X\|Y) = 0 \end{aligned}$$

となる。よって、ほとんどいたるところ [10]

$$\log \frac{q(x)}{p(x)} - \frac{q(x)}{p(x)} + 1 = 0$$

が成立する。これは上の不等式で等号条件が成立している場合に相当するため

$$\frac{q(x)}{p(x)} = 1\ a.e. \Leftrightarrow\ p(x) = q(x)\ a.e.$$

が成立する。逆は容易に示すことができる。 証明終

[10] a.e.(almost everywhere) などと表される。ある命題が成立しない確率が 0 であることを意味している。

A.4.2 正規分布

平均が μ で分散 σ^2 である正規分布 $N(\mu, \sigma^2)$ について、微分エントロピーは $\frac{1}{2}\log 2\pi\sigma^2 + \frac{1}{2}$ である。

定理 A.6 分散 σ^2 を固定したとき、正規分布 $N(\mu, \sigma^2)$ が微分エントロピーを最大化する。

(proof)

X を分散が σ^2 である任意の分布とし、$Y \sim N(\mu, \sigma^2)$ とする。それぞれの確率関数もしくは密度関数をそれぞれ p, q とする。このとき

$$\begin{aligned} D(X\|Y) &= \mathrm{E}\left[\log \frac{p(X)}{q(X)}\right] \\ &= \mathrm{E}\left[\log p(X)\right] - \mathrm{E}\left[\log q(X)\right] \\ &= -H(X) - \mathrm{E}\left[-\log 2\pi\sigma^2 - \frac{(X-\mu)^2}{2\sigma^2}\right] \\ &= -H(X) + \log 2\pi\sigma^2 + \frac{1}{2} \\ &= -H(X) + H(Y) \geq 0 \end{aligned}$$

よって $H(Y) \geq H(X)$ であり、等号が成立するのは $p(x) = q(x)$ *a.e.* つまり X の分布が正規分布のときである。 証明終