カルバック-ライブラー情報量についてのノート*

浜田　宏

東北大学

1　はじめに

このノートは,『社会科学のためのベイズ統計モデリング』朝倉出版の第 6 章に登場したカルバック-ライブラー情報量の意味を補足説明するためのノートです.

このノートは、黒木玄さんがウェブ上に公開しているノート「Kullback-Leibler 情報量と Sanov の定理」(`https://genkuroki.github.io/documents/20160616KullbackLeibler.pdf` 以下，黒木 2016 と表記）および渡辺・村田（2005:91-93）を参考にして，その内容を浜田がパラフレーズしたものです.

2　かんたんな具体例

出目 $\{1,2,3\}$ が確率

$$q=(q_1,q_2,q_3)=(1/2,1/4,1/4)$$

で生じる 3 面のサイコロがあると仮定します．この q を真の分布と呼ぶことにします.

サイコロを 1 回ふると，$\{1,2,3\}$ の中からどれか一つの目が出ます．この試行を n 回繰り返して観察した後，各出目の回数をカウントして k_1,k_2,k_3 と呼ぶことにします.

たとえばこのサイコロを 5 回ふって出目が

$$3,1,2,1,2$$

だったとしましょう．するとそれぞれの出目の回数は

$$k_1=2,k_2=2,k_3=1$$

*ver.1.0., 2020 年 2 月 28 日に公開.

です．この試行の回数 n が大きくなると各目の相対頻度

$$\frac{k_1}{n}, \frac{k_2}{n}, \frac{k_3}{n}$$

は大数の法則によって

$$\frac{1}{2}, \frac{1}{4}, \frac{1}{4}$$

に近づくと予想できます．

いま，サイコロの真の分布が $q = (1/2, 1/4, 1/4)$ であることを知らず，出目は均等に $(1/3, 1/3, 1/3)$ で出ると予測したと仮定します．この予測は

$$p = (p_1, p_2, p_3) = (1/3, 1/3, 1/3)$$

で表せます．以下 p を予測分布と呼びます．

真の分布が $q = (1/2, 1/4, 1/4)$ である 3 面サイコロを 10 回振ったときの各出目の出現頻度 $k_i/10$ を計算してみます．たとえば Mathematica なら次のようなコードで 10 回試行を実験できます．

```
sim[n_, q1_, q2_, q3_] := Module[{x},
x = Table[
RandomVariate[MultinomialDistribution[1, {q1, q2, q3}]], {n}];
N[Sum[x[[i]], {i, 1, n}]/n]
];

sim[10, 1/2, 1/4, 1/4]
```

結果は

$$(0.45, 0.25, 0.3)$$

でした．試行回数が 10 回程度だと，観察結果は予測分布 $p = (1/3, 1/3, 1/3)$ に近いように見えます．

そこで試行回数を 1000 回に増やしてみます．

```
sim[1000, 1/2, 1/4, 1/4]
```

すると観察結果は

$$(0.504, 0.261, 0.235)$$

でした．この観察結果から得た経験分布は、予測分布 $p = (1/3, 1/3, 1/3)$ からはズレていることが分かります．

以上の単純例は、

> 未知の分布からランダムサンプリングでサイズ n のデータを抽出すると、n が大きくなればその相対頻度 k_i/n は大数の法則で真の分布 q に近づく．ゆえに、予測分布 p と相対頻度 k_i/n が一致する確率は（p が q と一致していない場合は）小さくなる

ことを示唆しています．

この具体的な例を，もう少し一般的に表現してみましょう．

3 多項分布とカルバック-ライブラー情報量

n 回独立試行における各出目の回数を確率変数 K_1, K_2, K_3 とおけば、実現値が k_1, k_2, k_3 である確率は多項分布の確率質量関数

$$P(K_1 = k_1, K_2 = k_2, K_3 = k_3) = \frac{n!}{k_1!k_2!k_3!} q_1^{k_1} q_2^{k_2} q_3^{k_3}, \quad \sum k_i = n \tag{1}$$

で与えられます[1]．

いま，未知の真の分布 $q = (q_1, q_2, q_3)$ に対する予測分布として $p = (p_1, p_2, p_3)$ を仮定します．

$n \to \infty$ のとき，出目 i が出た割合 k_i/n がほぼ p_i になる確率について考えます．

まず $n \to \infty$ のとき

$$k_i = np_i + O(\log n)$$

を満たすと仮定します．この仮定は予測分布 p_i は真の分布 q_i と一致するとは限らないので，そのズレを $O(\log n)$ で見積もるというアイデアを表現しています．

するとスターリングの公式と上記の仮定から

$$\begin{aligned} \log\left(\frac{n!}{k_1!k_2!k_3!} q_1^{k_1} q_2^{k_2} q_3^{k_3}\right) &= -n \sum_{i=1}^{3} p_i \log \frac{p_i}{q_i} + O(\log n) \\ &= -nD(p||q) + O(\log n) \end{aligned}$$

が成立します（計算の詳細は黒木 (2016) を参照してください）．ここでカルバック-ライブラー情報量 $D(p||q)$ が出てきました．この式を指数変換すると

$$\frac{n!}{k_1!k_2!k_3!} q_1^{k_1} q_2^{k_2} q_3^{k_3} = \exp\{-nD(p||q) + O(\log n)\} \tag{2}$$

[1] 一般に、実現値が r 種類あるような離散分布からランダムサンプリングで得た各状態の回数を，この多項分布で表現できることに注意してください．つまり以下の話は多項分布という特殊な確率分布に限定して成立する話ではありません．任意の離散分布の経験分布が多項分布で表現できるため，任意の離散分布について成立します．

です.

左辺は確率関数ですから，それとイコールで結ばれた右辺は, なんらかの確率を表しています．n が大きいときに $O(\log n)$ の項を無視できると仮定すると (2) は，ある確率の大小がカルバック-ライブラー情報量の大きさによってほぼ決まることを表しています.

(1) は確率 $P(K_1 = k_1, K_2 = k_2, K_3 = k_3)$ なので，状態 $(1, 2, 3)$ が生じる回数が (k_1, k_2, k_3) となる確率を表しています.

つまり

> n 回の独立試行で状態 $(1, 2, 3)$ が生じる回数が，(k_1, k_2, k_3) となる確率

を表しています.

一方， (2) の右辺はこの確率 (1) を $n \to \infty$ のとき $k_i = np_i + O(\log n)$ を満たすと仮定して，k_i を p_i の関数に置き換えて近似したものです. したがって確率 (2) は

> 十分に大きな n 回の独立試行で状態 $(1, 2, 3)$ が生じる回数が，$(np_1 + O(\log n), np_2 + O(\log n), np_3 + O(\log n))$ となる確率

を表しています．$np_i + O(\log n)$ を《ほぼ np_i》と呼ぶことにすれば、(2) の右辺は

> 十分に大きな n 回の独立試行で状態 $(1, 2, 3)$ が生じる回数が，ほぼ (np_1, np_2, np_3) となる確率

を表しています.

つまり KL 情報量は，真の分布が q であるとき，そこからランダム・サンプリングで得たデータの経験分布がほぼ予測分布 p となる確率を決める量であることが分かります.

4 計算例

上記の KL 情報量の意味を計算によって確認してみましょう.

真の分布 $q = (1/2, 1/4, 1/4)$ および予測分布 $p = (1/3, 1/3, 1/3)$ を仮定して《q の経験分布がほぼ p となる確率》

$$\exp\{-nD(p||q)\}$$

を計算します.

```
kld[q_, p_, n_] :=
Exp[-n Sum[p[[i]] Log[p[[i]]/q[[i]]], {i, 1, 3}]];

ListPlot[Table[{i, kld[{1/2, 1/4, 1/4}, {1/3, 1/3, 1/3}, i]}, {i, 1,
80}], PlotRange -> {0, 1.1}]
```

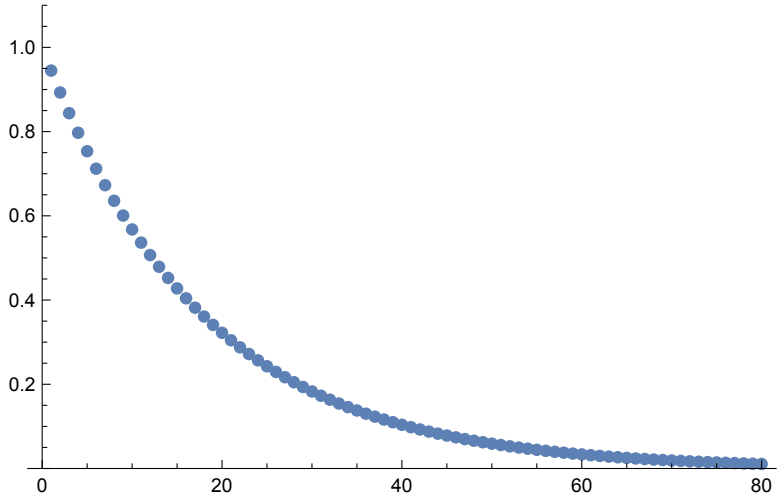

図 1: 予測分布 $p = (1/3, 1/3, 1/3)$ が $q = (1/2, 1/4, 1/4)$ の経験分布を近似できる確率．横軸は n

この計算例から、n が大きくなるにつれて，予測分布 p によって真の分布 q の経験分布を近似できる確率が 0 に近づく様子が分かります．

予測分布 p が真の分布 q に相対的に近い場合は，この確率の減少スピードがゆるやかになります．

つぎに、真の分布 $q = (1/2, 1/4, 1/4)$ および予測分布 $p = (1/2 + 0.05, 1/4, 1/4 - 0.05)$ を仮定して《q の経験分布がほぼ p となる確率》

$$\exp\{-nD(p||q)\}$$

を計算して見ましょう．

```
ListPlot[Table[{i,
kld[{1/2, 1/4, 1/4}, {1/2 + 0.05, 1/4, 1/4 - 0.05}, i]}, {i, 1,
80}], PlotRange -> {0, 1.1}]
```

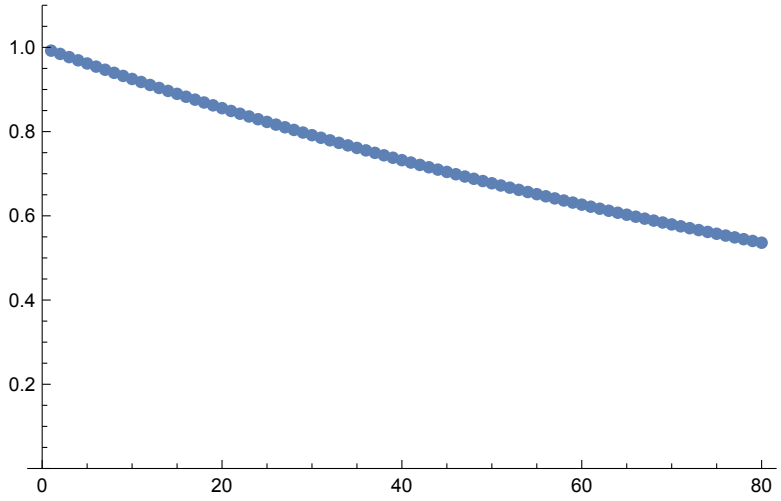

図 2: 予測分布 $p = (1/2 + 0.05, 1/4, 1/4 - 0.05)$ が $q = (1/2, 1/4, 1/4)$ の経験分布を近似できる確率

この計算結果から、予測分布 p が真の分布 q に相対的に近い場合でも,《q の経験分布がほぼ p となる確率》は n の増加と共に減少することが分かります．ただし相対的に近いぶん，減少がすこし緩やかになりました．

5 尤度関数からの KL 情報量の導出

KL 情報量は定義上，分布 q と分布 p の近さを測る指標になっています．しかし 2 つの確率密度関数の違いを定量的に把握するのであれば、KL 情報量でなくてもかまいません．

たとえば，

$$KL(p||q) = \int \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 dx$$

のような関数を考えることも可能です（渡辺・村田 2005:93-94）．

いま真の分布が $q(x)$ であるときに、それを $p(x)$ ではないかと推測したとします．以下 $p(x)$ を予測分布と呼びます．

そこで以下，データが得られた場合の尤度関数から KL 情報量が自然に導出できることを示します．

データ $(x_1, x_2, \ldots, x_n)$ が与えられたとき、予測分布 $p(x)$ の尤度関数 L をつぎのように定義します．

$$L(p) = \prod_{i=1}^{n} p(x_i)$$

両辺対数をとってから $1/n$ 倍すると

$$\log L(p) = \log \left(\prod_{i=1}^{n} p(x_i) \right)$$
$$\log L(p) = \log \sum_{i=1}^{n} p(x_i)$$
$$\frac{1}{n} \log L(p) = \frac{1}{n} \sum_{i=1}^{n} \log p(x_i)$$

ここで n が大きくなった場合について考えます．$1/n$ はデータ点 1 つが出現する相対頻度なので、各データ点 x_i に対する重みを表しています．データ点 x_i を一般的に x で表すと、n が大きくなったとき、x に対する重みは大数の法則によって真の分布の確率密度関数 $q(x)$ に近づくと考えられます．

その結果

$$\lim_{n \to \infty} \frac{1}{n} \log L(p) = \int \log p(x) q(x) dx$$

となります．右辺の総和は、極限が存在する場合はリーマン積分として定義できます．

右辺を変形すると

$$\begin{aligned}
\int \log p(x) q(x) dx &= \int \log p(x) q(x) dx \\
&= \int \log p(x) q(x) + \log q(x) q(x) - \log q(x) q(x) dx \\
&= \int (\log p(x) - \log q(x)) q(x) dx + \int \log q(x) q(x) dx \\
&= - \int \log \frac{q(x)}{p(x)} q(x) dx + \int \log q(x) q(x) dx \\
&= -D(q||p) - H(X)
\end{aligned}$$

です．$-H(X)$ は真の分布 $q(x)$ のエントロピーなので、なんらかの定数です．

ゆえに尤度関数 $L(p)$ を大きくするような予測分布 $p(x)$ は，カルバック-ライブラー情報量 $D(q||p)$ を小さくするような $p(x)$ であることが分かります．

尤度関数 $L(p)$ を大きくする予測分布 p $\iff$ $D(q||p)$ を小さくする予測分布 p

KL 距離情報量が小さくなるような予測分布 $p(x)$ を選ぶことは、データをあてはめたときに尤度関数が大きくなるような予測分布 $p(x)$ を選ぶことと一致します（渡辺・村田 2005:93-94）．

6 なぜKL情報量をモデル評価に使うのか？

ここまでに説明してきたことをまとめます.

1. KL 情報量 $D(p||q)$ は《真の分布 q の経験分布がほぼ予測分布 p となる確率》を決める関数として自然に導出できる
2. KL 距離情報量が小さくなるような予測分布 $p(x)$ を選ぶことは、データをあてはめたときに尤度関数が大きくなるような予測分布 $p(x)$ を選ぶことと一致する

以上が、KL 情報量をモデル評価に使う理由です.