『社会科学のためのベイズ統計モデリング』第 2 章

Hiroshi Hamada
Tohoku University

April, 2020
at Tohoku University

2.1　確率モデル

Table: 10 人分の保険加入記録データ

個人	1	2	3	4	5	6	7	8	9	10
行動	0	1	0	0	0	0	1	0	0	0

このデータを未知の分布から生成された実現値と考える

- データは未知の分布 $q(x)$ から生成される
- 真の分布の推測用にベルヌーイ分布を使う．このときベルヌーイ分布を確率モデル（probabilistic model）と呼ぶ
- 確率モデルは真の分布ではない．推測用の仮のモデル
- 確率モデルにサンプルの実現値をあてはめて，現象の理解と予測を促す一連の手続きを統計モデリング（statistical modeling）という

2.2 最尤法

ベルヌーイ分布の同時確率質量関数

$$
\begin{aligned}
p(y^n) &= p(y_1, y_2, \ldots, y_n) \\
&= p(y_1)p(y_2)\cdots p(y_n) \qquad \text{独立性の定義より} \\
&= q^{y_1}(1-q)^{1-y_1} \times q^{y_2}(1-q)^{1-y_2} \times \cdots \\
&\quad \times q^{y_n}(1-q)^{1-y_n}
\end{aligned}
$$

サンプルの実現値

$$
\begin{aligned}
y^n &= (y_1, y_2, \ldots, y_n) \\
&= \underbrace{(0, 1, 0, 0, 0, 0, 1, 0, 0, 0)}_{\text{この中には 1 が 2 個}}
\end{aligned}
$$

を同時確率関数 $p(y^n)$ に代入.

$$
\begin{aligned}
p(y^n) &= q^{y_1}(1-q)^{1-y_1} \times q^{y_2}(1-q)^{1-y_2} \times \cdots \times q^{y_n}(1-q)^{1} \\
&= q^0(1-q)^{1-0} \times q^1(1-q)^{1-1} \times \cdots \times q^0(1-q)^{1-0} \\
&= q^2(1-q)^8.
\end{aligned}
$$

関数 $q^2(1-q)^8$ を尤度関数（likelihood function）とよぶ

尤度関数

パラメータ q の関数

$$L(q \mid y^n) = \prod_{i=1}^{n} p(y_i)$$

を尤度関数という.

尤度関数の計算例

$$L(q|y^n) = q^2(1-q)^8$$

$q = 0.3$ の場合

$$\begin{aligned} L(0.3|y^n) &= 0.3^2(1-0.3)^8 \\ &\approx 0.00518832 \end{aligned}$$

$q = 0.4$ の場合

$$\begin{aligned} L(0.4|y^n) &= 0.4^2(1-0.4)^8 \\ &\approx 0.00268739. \end{aligned}$$

計算結果を比較すると，

$$L(0.3|y^n) > L(0.4|y^n)$$

$q = 0.3$ でサンプルの実現値 y^n を得る同時確率が，
$q = 0.4$ の場合の同時確率よりも大きい

$\rightarrow q = 0.3$ のほうが，確率モデルのパラメータとしてよさそう

定義 (最尤法)

尤度関数 $L(q \mid y^n)$ が最大となるパラメータを推定する方法を最尤法（maximum likelihood estimation）という．
尤度関数を最大にするパラメータ

$$\hat{q} = \arg\max_{q} L(q \mid y^n)$$

を最尤推定値（maximum likelihood estimate）と呼ぶ

$$\log L(q \,|\, y^n) = 2\log q + 8\log(1-q)$$

とおいて，関数 $\log L(q \,|\, y^n)$ を q で微分

$$\begin{aligned}\frac{d}{dq}\log L(q \,|\, y^n) &= 2\cdot\frac{1}{q} + 8\cdot\frac{1}{1-q}\cdot -1 \\ &= \frac{2(1-q)-8q}{q(1-q)} = \frac{2(1-5q)}{q(1-q)}.\end{aligned}$$

尤度関数が極値を持つ必要条件

$$\frac{2(1-5q)}{q(1-q)} = 0$$

より $q = 0.2$（最尤推定値）.

定義 (実現可能)

$S \subset \mathbb{R}^d$ をパラメータがとりうる値の集合とする．ある $\theta \in S$ により $q(x) = p(x|\theta)$ となるとき，真の分布 $q(x)$ は確率モデル $p(x|\theta)$ により実現可能（realizable）であるという

定義 (最尤推定量)

尤度関数 $L(\theta \,|\, Y_1, Y_2, \ldots, Y_n)$ を最大化する $\hat{\Theta}$ を最尤推定量（maximum likelihood estimator）という．すなわち

$$\hat{\Theta} = \arg\max_{\theta} L(\theta \,|\, Y^n)$$

を満たす確率変数 $\hat{\Theta}$ を最尤推定量とよぶ．最尤推定量はサンプル Y^n の関数である．

$$\hat{\Theta} = f(Y^n)$$

《最尤推定値》は確率変数の実現値

《最尤推定量》は確率変数

確率モデルが i.i.d. ベルヌーイ分布

確率質量関数は

$$p(y^n \mid q) = q^{\sum_{i=1}^n y_i}(1-q)^{n-\sum_{i=1}^n y_i}$$

なので，尤度関数は

$$L(q \mid y^n) = q^{\sum_{i=1}^n y_i}(1-q)^{n-\sum_{i=1}^n y_i}$$

実現値 $y_1, y_2, \ldots, y_n$ を確率変数 $Y_1, Y_2, \ldots, Y_n$ に置き換えれば

$$L(q \mid Y^n) = q^{\sum_{i=1}^n Y_i}(1-q)^{n-\sum_{i=1}^n Y_i}.$$

微分しやすいように，対数をとる（対数尤度関数）

$$\log L(q \mid Y^n) = \sum_{i=1}^n Y_i \log q + \left(n - \sum_{i=1}^n Y_i\right)\log(1-q)$$

微分で最大点を求める

$\log L(q \mid Y^n)$ を q で微分

$$\frac{d \log L(q \mid Y^n)}{dq} = \left(\sum_{i=1}^{n} Y_i\right)\frac{1}{q} - \left(n - \sum_{i=1}^{n} Y_i\right)\frac{1}{1-q}$$

$d \log L(q \mid Y^n)/dq = 0$ とおいて，$L(q \mid Y^n)$ を最大化する q を求める（簡略化のため $S_n = \sum_{i=1}^{n} Y_i$ とおく）

$$\frac{d \log L(q \mid Y^n)}{dq} = 0$$

$$\iff q = \frac{S_n}{n}$$

最尤推定量

ベルヌーイ分布の最尤推定量 $\hat{\Theta}$ は

$$\hat{\Theta} = \frac{S_n}{n} = \frac{Y_1 + Y_2 + \cdots + Y_n}{n} = \bar{Y}$$

$\hat{\Theta}$ は確率変数 Y^n の関数.

確率変数の関数は確率変数

2.3　最尤法のもとでの予測分布

定義 (予測分布（最尤法）)

最尤推定値

$$\hat{\theta} = \arg\max_{\theta} L(\theta \,|\, y^n)$$

を確率モデル $p(y \,|\, \theta)$ のパラメータに代入した

$$p(y \,|\, \hat{\theta})$$

を最尤法の予測分布という.

例：人気のないブログ

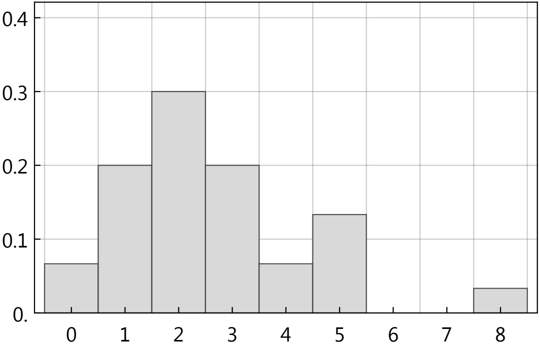

Figure: ブログのアクセス数（横軸はアクセス人数，縦軸はその割合）

確率モデルとしてポアソン分布を仮定する．確率質量関数は

$$\mathrm{Poisson}(x|\lambda) = \frac{\lambda^x}{x!}e^{-\lambda}.$$

データ $(x_1, x_2, \ldots, x_n)$ を観測した場合の尤度関数は

$$L(\lambda|x^n) = \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!}e^{-\lambda}.$$

対数をとると

$$\log L(\lambda|x^n) = \left(\sum x_i\right) \log\lambda - n\lambda - \sum \log x_i!$$

対数尤度関数を λ で微分

$$\frac{d\log L(\lambda|x^n)}{\lambda} = \frac{\sum x_i}{\lambda} - n$$

導関数を 0 とおいて，最大点を探す.

$$\frac{\sum x_i}{\lambda} - n = 0 \iff \lambda = \frac{\sum x_i}{n}$$

ポアソン分布の最尤推定量は標本平均

$$\frac{\sum X_i}{n}$$

ブログアクセスデータを使って計算した最尤推定値

$$\hat{\lambda} = \sum_{i=1}^{n} \frac{x_i}{n} = \frac{78}{30} = 2.6$$

最尤法に基づく予測分布 $p(x|\hat{\lambda})$ は

$$p(x|2.6) = \frac{2.6^x}{x!} e^{-2.6}.$$

予測分布とデータのフィッティング

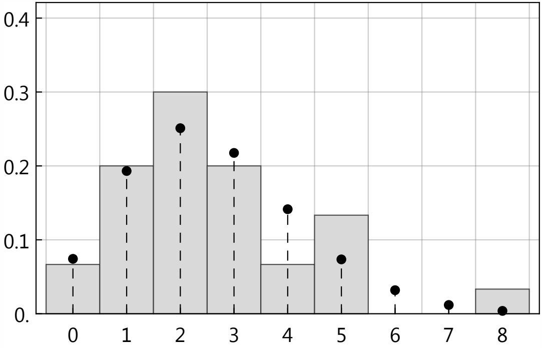

Figure: ブログデータと予測分布

たまたま得たデータに過剰にモデルを適合させてしまうことを過剰適合（overfitting）あるいは過学習という

2 章まとめ

- 未知の分布の推測用に研究者が想定した仮の分布を確率モデルという
- 確率モデルのパラメータをうまく選ぶと確率モデルと真の分布が一致するとき，真の分布は実現可能という
- 確率モデルのパラメータを推定する方法の一つが最尤法である
- 確率モデルの同時確率関数を実現値を固定してパラメータの関数としてみたものを尤度関数という
- 尤度関数が最大になるパラメータを選ぶ手法を最尤法という