WAIC/WBICの定義

Amuta (Twitter: @Amuta151)

1 概要

Bayes 統計学における重要な量である WAIC と WBIC の定義を測度論的確率論の下で整理することが本稿の目的である.

[1] でも十分に正当化可能な定義が与えられているものの, 著者が同書の中で述べているように測度論をはじめとする数学的な知識は仮定されていない. そのため, 何が確率変数で何がそうでないのかといった確率論的な意味での定義がやや見えにくくなっている部分がある.

実際, 標語的に「Bayes 統計学は母数を確率変数とみなす」や「Bayes 統計学は Bayes の定理を基礎とする」といった説明がなされることがある. しかし, 現代確率論の標準的基礎である測度論的確率論の下で定式化する場合, どちらの考えも必要としない.

そこで, 本稿では Bayes 統計学の諸概念を確率論的に定義し, 数学的な見通しをよくすることを目標とした. 主部は諸定義を与えた第 2 節であり, それ以降の内容は [1], [2], [3], [4] の内容を適宜まとめたものである.

2 諸定義

(Ω, P) を確率空間, $X : \Omega \to \mathbb{R}$ を確率変数とする. また, X によって $\mathbb{R}$ 上に誘導される測度が確率密度関数 $q(x) \in L^1(\mathbb{R})$ を持つとする. すなわち, $\mathbb{R}$ 上の可測集合 S に対して $P_X(S) = P(X^{-1}(S))$ と表したとき,

$$dP_X = q(x)dx,$$

が成り立つ[*1].

[*1] $dP_X = q(x)dx$ とは任意の $\mathbb{R}$ 上の可測集合 S に対して $P(S) = \int_S q(x)dx$ が成り立つことである.

さらに W を $\mathbb{R}^d$ のコンパクト部分集合する. $\mathbb{R}\times W$ 上の関数 $p(x,w)\in L^1(\mathbb{R}\times W)$ であって, 各 $w\in W$ に対して,

$$\int_{\mathbb{R}} p(x,w)dx = 1,$$

となるような関数 $p(x,w)$ を任意に選び**確率モデル**という. $p(x,w)$ はパラメータ w を持つ確率密度関数を意図したものであり, 条件付き確率の記法を模倣して $p(x|w)$ と表す[*2].

X と独立同分布である確率変数 $X_1, X_2, \cdots, X_n$ を考え, その組を $X^n = (X_1, X_2, \cdots, X_n)$ と表す. また, W 上の確率密度関数 φ, すなわち $\varphi\in L^1(W)$ かつ $\int_W \varphi = 1$ であるものを任意に取り**事前分布**という.

上記の表記法の下で, 事後分布と予測分布を次で定義する.

定義 1. (事後分布)

$\beta\in(0,\infty)$ に対し,

$$p(w|X^n) = \frac{1}{Z_n(\beta)}\varphi(w)\prod_{i=1}^{n} p(X_i|w)^\beta,$$

を**事後分布**という. ただし,

$$Z_n(\beta) = \int_W \varphi(w)\prod_{i=1}^{n} p(X_i|w)^\beta dw,$$

とする.

また, 正規化定数 $Z_n(\beta)$ により事後分布の W 上の積分は 1 になるから, 事後分布を確率密度関数として見たときの平均[*3]を,

$$\mathbb{E}_w^\beta[\,\cdot\,] = \int \cdot\, p(w|X^n)dw,$$

と表す. $\beta = 1$ のとき, $\mathbb{E}_w^1[\,\cdot\,]$ を単に $\mathbb{E}_w[\,\cdot\,]$ と書く.

事後分布は確率変数の関数であるからそれ自身, 確率変数である ($w\in W$ を固定するごとに $\Omega\to\mathbb{R}$ なる確率変数となる).

[*2] しかし, これは単なる記法であり $p(x|w)$ の実体は $L^1(\mathbb{R}\times W)$ に属する関数である. いま, w は確率変数の値域ではないから x, w の同時確率密度関数は考えられず, したがって $p(x|w)$ は条件付き確率密度関数ではない.

[*3] W は確率変数の値域ではないから正確には $\mathbb{E}_w[\,\cdot\,]$ は平均とは言えない. 平均の定義は確率変数の Ω 上での測度 P による積分である. しかし, 全空間での積分が 1 になる可積分関数 $p(w|X^n)$ との積の積分はまさに確率密度関数による平均の類似物である.

定義 2. (予測分布)
事後分布による確率モデルの平均,

$$
\begin{aligned}
p^*(x) &= \mathbb{E}_w[p(x|w)], \\
&= \int p(x|w)p(w|X^n)dw,
\end{aligned}
$$

を**予測分布**という.

以上の定義の下で, 次の量を定める.

$$
\begin{aligned}
T_n &= -\frac{1}{n}\sum_{i=1}^{n}\log p^*(X_i) : \text{学習損失}, \\
F_n &= -\frac{1}{\beta}\log Z_n(\beta) : \text{自由エネルギー}, \\
G_n &= -\int q(x)\log p^*(x)dx : \text{汎化損失}, \\
V_n &= \sum_{i=1}^{n}\left\{\mathbb{E}_w[(\log p(X_i|w))^2] - \mathbb{E}_w[\log p(X_i|w)]^2\right\} : \text{汎関数分散}, \\
w_0 &= \arg\min_{w\in W}\left\{-\int q(x)\log p(x|w)dx\right\}, \\
L_n(w) &= -\frac{1}{n}\sum_{i=1}^{n}\log p(X_i|w), \\
L(w) &= -\int q(x)\log p(x|w)dx.
\end{aligned}
$$

3 Bayes 統計学の状態方程式

$\mathbb{E}$ は Ω 上の平均を表すとする (確率変数 X についての平均であり, 事後分布の平均ではない).

定義 3. (相対的に有限な分散)
対数尤度比関数 l を,

$$
l(x, w_0, w) = \log\frac{p(x|w_0)}{p(x|w)},
$$

で定める.

ある定数 $c_0 > 0$ が存在して, 任意の $w \in W$ について,

$$\mathbb{E}[l(x, w_0, w)] \geq c_0 \mathbb{E}[l(x, w_0, w)^2],$$

が成り立つとき, 対数尤度比関数は**相対的に有限な分散**を持つという.

定理 4. (Bayes 統計学の状態方程式)
対数尤度比関数が相対的に有限な分散を持つとする. このとき,

$$\mathbb{E}[G_n] = \mathbb{E}\left[T_n + \frac{\beta V_n}{n}\right] + o\left(\frac{1}{n}\right),$$

が成り立つ[*4].

4 WAICの定義

定義 5. (WAIC)
広く使える情報量基準 (Widely Applicable Information Criterion) W_n^A を,

$$W_n^A = T_n + \frac{V_n}{n},$$

と定義する.

定理 6. $\mathbb{E}$ は Ω 上の平均, $\mathbb{V}$ は分散であるとする.

1. 汎化損失 G_n と WAIC は平均平均の差が $1/n^2$ に比例する.

$$\mathbb{E}[G_n] = \mathbb{E}[W_n^A] + O\left(\frac{1}{n^2}\right).$$

2. $G_n - L(w_0)$ と $W_n^A - L_n(w_0)$ は分散の差が $1/n^2$ に比例する.

$$\mathbb{V}[G_n - L(w_0)] = \mathbb{V}[W_n^A - L_n(w_0)] + O\left(\frac{1}{n^2}\right).$$

[*4] 以降で使用する o, O は収束の速さを評価する Landau の記号である. 定義は標準的な微積分学の教科書もしくは [1] p.26-27 を参照のこと. 簡単に述べるならば $n \to \infty$ としたとき $o(1/n)$ は $1/n$ より収束が速い任意の関数のうちの1つを表し, $O(1/n)$ は $1/n$ と収束の速さが同程度である任意の関数のうちの1つを表す.

5　WBICの定義

定義 7. (WBIC)

広く使えるBayes情報量基準 (Widely Applicable Bayesian Information Criterion) W_n^B を,

$$W_n^B = \mathbb{E}_w^\beta[nL_n(w)],$$

と定義する. ただし, $\beta = 1/\log n$ である.

定理 8. 自由エネルギー F_n と WBIC は平均の差が $\log\log n$ に比例する.

$$\mathbb{E}[F_n] = \mathbb{E}[W_n^B] + O\left(\log\log n\right).$$

参考文献

[1] 渡辺澄夫, ベイズ統計学の理論と方法, コロナ社, 2012.

[2] S. Watanabe, "Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory," Journal of Machine Learning Research, Vol.11, (DEC), pp.3571-3591, 2010.

[3] S. Watanabe, "A widely applicable Bayesian information criterion," , Journal of Machine Learning Research, Vol. 14, (Mar), pp.867-897, 2013.

[4] 渡辺澄夫, 広く使える情報量規準 (WAIC), http://watanabe-www.math.dis.titech.ac.jp/users/swatanab/waic2011.html, 2011

[5] 伊藤清, 確率論, 岩波書店, 1991.

[6] Gerald B. Folland, Real Analisys Second Edition, Wiley-Inter Sience, 1999.