『社会科学のためのベイズ統計モデリング』第 6 章　6.7–6.10

Hiroshi Hamada
Tohoku University

Jun, 2020
at Tohoku University

交差エントロピー

定義 (交差エントロピー)

$A \subset \mathbb{R}$ 上で定義された確率密度関数 $q(x), p(x) > 0$ について，$q(x)$ と $p(x)$ の交差エントロピーを，

$$H_q(p) = -\int_A q(x) \log p(x) dx$$

と定義する.

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

真の分布 $q(x)$ からのモデル $p(x)$ の近さ $D(q||p)$ は，未知の定数部分 $H(q)$ をのぞくと，交差エントロピー $H_q(p)$ の大きさと完全に一致する.

交差エントロピーをデータから推定できれば，真の分布へのモデルの近さを評価できる

汎化損失

真の分布 $q(x)$ と予測分布 $p^*(x)$ の交差エントロピーを，
汎化損失 (generalization loss)G_n とよぶ．

$$
\begin{aligned}
G_n &= -\mathbb{E}_{q(X)}[\log p^*(X)] \\
&= -\int \log p^*(x) q(x) dx \\
&= \underbrace{H(q)}_{\text{真の分布のエントロピー}} + \underbrace{D(q||p^*)}_{\text{真の分布と予測分布の KL 情報量}}
\end{aligned}
$$

G_n は $q(x)$ のエントロピーで $-\log q(x)$ の代わりに，
$-\log p^*(x)$ を使った式

経験損失

予測分布 $p^*(x)$ の情報量として，すでに予測分布の導出に用いたデータ $x^n = (x_1, x_2, \cdots, x_n)$ を入れて，相加平均をとったものを経験損失 (training loss) とよぶ

$$T_n = -\frac{1}{n}\sum_{i=1}^{n} \log p^*(x_i).$$

経験損失 T_n を使って，汎化損失 G_n をデータから推定する方法

最尤推定予測分布 $\rightarrow$ AIC
ベイズ予測分布 $\rightarrow$ WAIC

AIC

仮定

- 真の分布：$q(x) = p(x|\theta)$，仮定した確率モデルのなかに真の分布が含まれる．これを実現可能という
- サンプル：$X^n \sim q(x^n) = \prod q(x)$，サンプルは独立かつ同一の真の分布にしたがう
- 確率モデル（尤度）：$p(x^n|\theta)$
- $q(x)$ は確率モデルについて正則

正則 (regular)

- 真の分布 $q(x)$ と確率モデル $p(x|w)$ の交差エントロピー $H_q(p)$ をパラメータの関数と見なし $L(w)$ と書く（平均対数損失関数 log loss function とも呼ぶ）

$$H_q(p) = -\int \log p(x|w) q(x) dx = L(w)$$

- $L(w) \geq H(q)$ なので $L(w)$ は下に有界.
- $L(w)$ の最小値を与えるパラメータ w_0 がユニークで，その点でヘッセ行列 $\nabla^2 L(w_0)$ が正定値（固有値が全て正）であるとき，真の分布 $q(x)$ に対して確率モデル $p(x|w)$ は 正則 (regular) であるという.

「正則」とは

ちょっとなにいってるか，わからない

直感的なイメージ：確率モデルが『正則』とは，交差エントロピーを最小化するパラメータがユニークで，微分を使ってその点を特定できること

「正則」とは

小西・北川（2004）

- $\log p(x|w)$ は $w \in W \subset R^d$ に関して 3 階連続微分可能
- R 上で積分可能な $F_1(x), F_2(x)$ および適当な実数 M に対して

$$\int_{-\infty}^{\infty} H(x) f(x|w) dx < M$$

となる $H(x)$ が存在して（たとえば期待値が有限），

$$\left|\frac{\partial dp}{\partial w_i}\right| < F_1(x), \left|\frac{\partial^2 dp}{\partial w_i \partial w_j}\right| < F_2(x), \left|\frac{\partial^3 dp}{\partial w_i \partial w_j \partial w_k}\right| < H(x)$$

が任意の $w \in W$ について成り立つ

- $\forall w \in W$

$$0 < \int_{-\infty}^{\infty} p(x|w) \frac{\partial dp}{\partial w_i} \frac{\partial dp}{\partial w_j} dx < \infty$$

このとき最尤推定量が一致性（真の分布のパラメータに確率収束すること）と漸近正規性（最尤推定量が正規分布に漸近近似すること）をもつ

特異モデルでは最尤推定量は漸近正規でも漸近不偏ではなく，AIC は意味を持たない

注意！！

社会学でよく使われる統計モデル（GLM）が正則とは限らない（GLMM は特異）.

社会学でよく使われる統計モデル（GLM）によって未知の分布が実現可能であるとは限らない.

GLM で実現可能性と正則性を暗に仮定してパラメータの因果的効果を解釈するのは注意が必要（観察データから因果的効果を推定したいなら，条件付き独立の仮定を満たす条件付き期待値回帰を考えたほうがよい）

Quiz

交差エントロピーを最小化する，正則な確率モデルの例を考えよ

AIC の導出

最尤推定値をもとにした予測分布

$$p^*(x) = p(x|\hat{\theta}), \quad \hat{\theta} = \arg\max_{\theta} p(x^n|\theta)$$

最尤推定の汎化損失

$$\begin{aligned} G_n &= -\mathbb{E}_{q(X)}[\log p^*(X)] \\ &= -\mathbb{E}_{q(X)}[\log p(X|\hat{\theta})] \\ &= -\int \log p(x|\hat{\theta}) q(x) dx \end{aligned}$$

最尤推定の経験損失

$$
\begin{aligned}
T_n &= -\frac{1}{n} \log p(x^n | \hat{\theta}) \\
&= -\frac{1}{n} \sum_{i=1}^{n} \log p(x_i | \hat{\theta}).
\end{aligned}
$$

意味：観察データのもとでの最大対数尤度 (maximum log likelihood) の $-1/n$ 倍.

汎化損失（知りたい）を経験損失（分かる）から推定

$b(x^n)$: 汎化損失と経験損失のズレ

データ x^n のもとでの偏り $b(x^n)$

$$
\begin{aligned}
b(x^n) &= G_n - T_n \\
&= -\mathbb{E}_{q(X)}[\log p(X|\hat{\theta})] + \frac{1}{n}\log p(x^n|\hat{\theta})
\end{aligned}
$$

バイアスの期待値

$$\begin{aligned}\mathbb{E}_{q(X^n)}[b(X^n)] &= \mathbb{E}_{q(X^n)}[G_n - T_n] \\ &= \mathbb{E}_{q(X^n)}\Big[-\mathbb{E}_{q(Z)}[\log p(Z|\hat{\Theta})] \\ &\quad + \frac{1}{n}\log p(X^n|\hat{\Theta})\Big]\end{aligned}$$

Z: 予測する新たな確率変数

バイアスの期待値 $\mathbb{E}_{q(X^n)}[b(X^n)]$ は仮定のもとで d/n に漸近的に一致

AIC

$$\mathrm{AIC} = T_n + \frac{d}{n}$$

$$\mathbb{E}_{q(X^n)}[\mathrm{AIC}] = \mathbb{E}_{q(X^n)}[T_n] + \frac{d}{n}$$

AIC

$$\mathrm{AIC} = T_n + \frac{d}{n}$$

パラメータ数が出てくる直感的な理由:真の分布が実現可能であるとき，フィッシャー情報行列 $I(\theta_0)$ とヘッセ行列の期待値 $J(\theta_0)$ が一致する．その結果パラメータ次元数の正方行列のトレースがパラメータ数となる（小西・北川 2004）.

対数尤度のバイアス

$$= \mathbb{E}_{q(X^n)}\left[\sum_{i=1}^{n} \log p(X_i|\hat{\theta}) - n\mathbb{E}_{q(z)}[\log p(z|\hat{\theta})]\right]$$

$$= tr\{I(\theta_0)J(\theta_0)^{-1}\} = tr(I_d) = d$$

$$
\begin{aligned}
E[b(x^n)] &= E[G_n] - E[T_n] \\
\frac{d}{n} + o(1/n) &= E[G_n] - E[T_n] \\
E[T_n] + \frac{d}{n} + o(1/n) &= E[G_n] \\
E[AIC] + o(1/n) &= E[G_n] \quad \text{AIC の定義より} \\
E[AIC] &= E[G_n] + o(1/n)
\end{aligned}
$$

$$\mathbb{E}_{q(X^n)}[\mathrm{AIC}] = \mathbb{E}_{q(X^n)}[G_n] + o(1/n).$$

AIC は漸近的に平均的に汎化損失に一致

AIC は確率変数！！

ランダウ記号

$o(1/n)$ は

$$\lim_{n\to\infty} \frac{o(1/n)}{1/n} = 0$$

$o(1/n)$ は $n \to \infty$ のとき $1/n$ よりも先に 0 に収束

WAIC

WAIC を定義するための仮定

- 真の分布：$q(x) = p(x|\theta)$，仮定した確率モデルのなかに真の分布が含まれなくてもよい
- サンプル：$X^n \sim q(x^n) = \prod q(x)$，サンプルは独立かつ同一の真の分布にしたがう
- 確率モデル（尤度）：$p(x^n|\theta)$
- $q(x)$ は確率モデルについて正則でなくてもよい

予測分布

ベイズ推定をもとにした予測分布 (確率モデルの事後分布による期待値)

$$
\begin{aligned}
p^*(x) &= \mathbb{E}_{p(\theta|x^n)}[p(x|\theta)] \\
&= \int p(x|\theta)p(\theta|x^n)d\theta
\end{aligned}
$$

ベイズ予測分布についての汎化損失 G_n

$$
\begin{aligned}
G_n &= -\mathbb{E}_{q(X)}[\log p^*(X)] \\
&= -\mathbb{E}_{q(X)}[\log \mathbb{E}_{p(\theta|x^n)}[p(X|\theta)]] \\
&= -\int q(x) \log \left(\int p(x|\theta) p(\theta|x^n) d\theta \right) dx
\end{aligned}
$$

ベイズ予測分布についての経験損失 T_n

$$
\begin{aligned}
T_n &= -\frac{1}{n} \sum_{i=1}^{n} \log p^*(x_i) \\
&= -\frac{1}{n} \sum_{i=1}^{n} \log \mathbb{E}_{p(\theta|x^n)}[p(x_i|\theta)]
\end{aligned}
$$

汎化損失 G_n と経験損失 T_n のズレの期待値

$$\mathbb{E}_{q(X^n)}[b(X^n)] = \mathbb{E}_{q(X^n)}\left[G_n(X^n) - T_n(X^n)\right]$$

漸近的に汎関数分散 V_n に一致

$$V_n = \sum_{i=1}^{n} \left\{ \mathbb{E}_{p(\theta|x^n)}[(\log p(x_i|\theta))^2] - \mathbb{E}_{p(\theta|x^n)}[\log p(x_i|\theta)]^2 \right\}.$$

WAIC

$$\mathrm{WAIC} = T_n + \frac{V_n}{n}$$

$$\mathbb{E}_{q(X^n)}[\mathrm{WAIC}] = \mathbb{E}_{q(X^n)}[G_n] + o(1/n).$$

実現可能性，正則条件にかかわらず，一般的に成立

WAIC は確率変数！！

ベイズ自由エネルギー

周辺尤度 (marginal likelihood)

$$p(x^n) = \int p(x^n|\theta)\varphi(\theta)d\theta$$

- ベイズモデルの事後分布 $p(\theta|x^n)$ の分母
- パラメータとデータの同時確率 $p(x^n, \theta) = p(x^n|\theta)\varphi(\theta)$ を θ について周辺化したもの
- 事前分布と確率モデルを前提とした場合のサンプル X^n の確率分布

ベイズ自由エネルギー

ベイズ自由エネルギー F_n

$$F_n = -\log p(x^n)$$
$$= -\log \int p(x^n|\theta)\varphi(\theta)d\theta$$

周辺尤度の対数の符号反転

周辺尤度 $p(x^n)$ と真の分布の同時分布 $q(x^n)$ との交差エントロピー

$$
\begin{aligned}
& H_{q(X^n)}[(p(X^n)] \\
& = \mathbb{E}_{q(X^n)}[F_n] \\
& = -\int q(x^n) \log p(x^n) dx^n \\
& = -\int q(x^n) \log q(x^n) dx^n + \int q(x^n) \log \frac{q(x^n)}{p(x^n)} dx^n \\
& = \underbrace{H[(q(X^n)]}_{\text{真の分布のエントロピー}} + \underbrace{D[(q(X^n)||p(X^n)]}_{\text{真の分布と周辺尤度の KL 情報量}}
\end{aligned}
$$

自由エネルギーの期待値は、周辺尤度と真の分布の近さ

真のモデルが正則な場合

$$\mathrm{BIC} = -\sum_{i=1}^{n} \log p(x_i|\hat{\theta}) + \frac{d}{2} \log n$$

d はモデルのパラメータ数.

$$\mathbb{E}_{q(X^n)}[\mathrm{BIC}] = \mathbb{E}_{q(X^n)}[F_n] + O(1).$$

WBIC

経験対数損失

$$L_n(\theta) = -\frac{1}{n}\sum_{i=1}^{n} \log p(x_i|\theta)$$

の n 倍について，逆温度が $\beta = 1/\log n$ のときの事後分布による期待値

$$\mathrm{WBIC} = \int nL_n(\theta) \left[\frac{\prod_{i=1}^{n} p(x_i|\theta)^{\beta}\varphi(\theta)}{\int \prod_{i=1}^{n} p(x_i|\theta)^{\beta}\varphi(\theta)d\theta} \right] d\theta$$

真のモデルが正則でない場合でも使える

$$\mathbb{E}_{q(X^n)}[\mathrm{WBIC}] = \mathbb{E}_{q(X^n)}[F_n] + O\left(\log\log n\right).$$

WBIC は平均的に自由エネルギーに近似する

まとめ

- 真の分布から別の分布の近さを測る一般的な指標が KL 情報量である.
- KL 情報量のコア部分は交差エントロピーである.
- 真の分布と予測分布の交差エントロピーは，汎化損失とよばれ，最尤推定値による予測分布では AIC，ベイズ推定による予測分布では WAIC が平均的によい近似を与える.
- 周辺尤度の情報量は自由エネルギーとよばれ，真の分布と周辺尤度との交差エントロピーは，自由エネルギーの期待値と等しい.