P-6

ベイズ推測における正則モデルに対する考察
Consideration on regular models in Bayesian estimation

○塚本雄基 [1], 青柳美輝 [2]
Yuki Tsukamoto, Miki Aoyagi

Abstract: The Bayesian estimation in learning theory is to estimate a true probability density function from samples by using learning models and *a priori* probability density functions. The free energy, the Bayes generalization loss and the Bayes training loss play an important role in analysing such estimation. In this paper, we consider these stochastic variables in regular models.

1. 導入

与えられたサンプルから真の分布を推測することを統計的推測という．ベイズ推測では，真の分布から発生したサンプルと確率モデル及び事前分布から定義される予測分布を用いて，統計的推測が行われる．この際に，推測の精度を考えるための指標として，自由エネルギー，汎化損失及び経験損失が挙げられる．ベイズ推測のなかでも，事後分布が正規分布で近似できる特別な場合に限り成り立つ理論を正則理論と呼ぶ．本稿では，この正則理論における自由エネルギーの挙動及び経験損失から汎化損失を推測する方法について考察する．

2. ベイズ推測の定義

$x_i \in \mathbb{R}^N$ $(i = 1, \ldots, n)$ を真の分布 $q(x)$ から得られた n 個のサンプルとし，$x^n = (x_1, x_2, \ldots, x_n)$ とする．関数 $f(x^n)$ が与えられたとき，平均値をとる操作 $\mathbb{E}[\ ]$ を

$$\mathbb{E}[f(x^n)] = \int\int \cdots \int f(x_1, x_2, \ldots, x_n) \prod_{i=1}^{n} q(x_i) dx_i$$

と表記する．$p(x|w)$ を確率モデル，$\varphi(w)$ を事前分布とすると，パラメータ w の事後分布は

$$p(w|x^n) = \frac{1}{Z_n(\beta)} \varphi(w) \prod_{i=1}^{n} p(x_i|w)^{\beta}$$

と定義される．ここで，β は $0<\beta<\infty$ を満たす実数で，逆温度と呼ばれる．また，$Z_n(\beta)$ は正規化定数である．
すなわち，

$$Z_n(\beta) = \int \varphi(w) \prod_{i=1}^{n} p(x_i|w)^{\beta} dw$$

を満たす．この値 $Z_n(\beta)$ を分配関数という．パラメータ w の関数 $g(w)$ が与えられたとき，事後分布 $p(w|x^n)$ による平均を

$$\mathbb{E}_w[g(w)] = \int g(w)\ p(w|x^n) dw$$

と表記し，また確率変数 x の関数 $h(x)$ についての平均を

$$\mathbb{E}_X[h(x)] = \int h(x)\ q(x) dx$$

と表記する．事後分布によって確率モデル $p(x|w)$ を平均したもの

$$p^*(x) = \mathbb{E}_w[p(x|w)] = \int p(x|w)\ p(w|x^n) dw$$

を予測分布という．ベイズ推測とは，「真の分布 $q(x)$ は，おおよそ $p^*(x)$ であろう」と推測することである．

3. 正則理論

挙動を考察する統計量について紹介する．

定義 1 (自由エネルギー)

$$F_n(\beta) = -\frac{1}{\beta} \log Z_n(\beta)$$

を自由エネルギーという．

真の分布 $q(x)$ のエントロピー S を

$$S = -\int q(x) \log q(x) dx$$

と定義すると，$\beta = 1$ の自由エネルギー $F_n(1)$ に対して，

$$\mathbb{E}[F_n(1)] = nS + \int q(x^n) \log \frac{q(x^n)}{p(x^n)} dx^n \qquad (1)$$

が成り立つ．ここで, $q(x^n) = \prod_{i=1}^{n} q(x_i)$ であり, $p(x^n) = \int \varphi(w) \prod_{i=1}^{n} p(x_i|w) dw$ である．

式 (1) について, 右辺第 1 項は確率モデルと事前分布に依存しない．また，右辺第 2 項は真の分布と推測された分布の Kullback Leibler 距離で, 零になるのは $q(x^n) = p(x^n)$ のときに限る．従って，$F_n(1)$ が小さい程，推測された分布 $p(x^n)$ が真の分布 $q(x^n)$ を平均的によく近似していると考えることができる．

1: 日大理工・院 (前)・数学　2: 日大理工・教員・数学

定義 2 (汎化損失，経験損失)

$$G_n = -\int q(x)\log p^*(x)dx,$$
$$T_n = -\frac{1}{n}\sum_{i=1}^{n}\log p^*(x_i)$$

をそれぞれ，汎化損失，経験損失という．

汎化損失 G_n は

$$G_n = S + \int q(x)\log\frac{q(x)}{p^*(x)}dx \tag{2}$$

と表すことができる．式 (2) についても同じく, 右辺第 2 項は真の分布と予測分布の Kullback Leibler 距離である．従って，G_n が小さい程，$p^*(x)$ が $q(x)$ を精度よく推測していることが分かる．しかし，現実の問題では真の分布は不明であるため，汎化損失 G_n を直接に算出することはできない．一方，経験損失 T_n は，組 (p,φ) とサンプル x^n の実現値が与えられれば数値として確定する．よって，T_n から G_n を推測することができると極めて有用である．

定義 3 (平均対数損失関数 , 経験対数損失関数)

$$L(w) = -\mathbb{E}_X[\log p(x|w)],$$
$$L_n(w) = -\frac{1}{n}\sum_{i=1}^{n}\log p(x_i|w)$$

をそれぞれ，平均対数損失関数, 経験対数損失関数という．この定義より, $\mathbb{E}[L_n(w)] = L(w)$ である．

定義 4 $L(w)$ を最小にするパラメータを w_0 とする. パラメータ w について, 確率過程 $\eta_n(w)$ を

$$\eta_n(w) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(K(w) - f(x_i,w))$$

と定義する. また確率変数 ξ_n は $\xi_n = J^{-1/2}\nabla\eta_n(w_0)$ と定義する. ここで, $J = (J_{ij}) := \dfrac{\partial^2 L}{\partial w_i \partial w_j}(w_0)$ である.

定義 5 $d\times d$ 行列 $I(w)$ を

$$I(w) = \mathbb{E}_X[\nabla f(x,w)(\nabla f(x,w))^T] - \nabla K(w)(\nabla K(w))^T$$

と定義し, $I = I(w_0)$ とおく. なお, $w = w_0$ のとき, この式の第 2 項は 0 である. $I(w)$ は対称行列で固有値はすべて 0 以上である.

正則理論が成り立つ，すなわち事後分布を正規分布で近似するためには, 以下の 2 つの条件が必要である．

(条件 1) 平均対数損失関数 $L(w)$ を最小にするパラメータ w_0 が 1 つだけであること．

(条件 2) 行列 J の固有値がすべて正であること．

これら 2 つの条件が満たされている正則理論においては, 以下の 2 つの定理が成り立つ．

定理 6 自由エネルギーは次の挙動をもつ．

$$F_n(\beta) = nL_n(w_0) + \frac{d}{2\beta}\log n + \frac{1}{2\beta}\log\det J$$
$$+\frac{d}{2\beta}\log\left(\frac{\beta}{2\pi}\right) - \frac{1}{2}\|\xi_n\|^2 - \frac{1}{\beta}\log\varphi(w_0) + o_p(1).$$

また，平均値は

$$\mathbb{E}[F_n(\beta)] = nL(w_0) + \frac{d}{2\beta}\log n + \frac{1}{2\beta}\log\det J$$
$$+\frac{d}{2\beta}\log\left(\frac{\beta}{2\pi}\right) - \frac{1}{2}\mathrm{tr}(IJ^{-1}) - \frac{1}{\beta}\log\varphi(w_0) + o(1).$$

定理 7 確率変数としての汎化損失 G_n と経験損失 T_n は次の挙動をもつ．

$$G_n = L(w_0) + \frac{1}{n}\left(\frac{d}{2\beta} + \frac{1}{2}\|\xi_n\|^2 - \frac{1}{2\beta}\mathrm{tr}(IJ^{-1})\right) + o_p\left(\frac{1}{n}\right)$$
$$T_n = L_n(w_0) + \frac{1}{n}\left(\frac{d}{2\beta} - \frac{1}{2}\|\xi_n\|^2 - \frac{1}{2\beta}\mathrm{tr}(IJ^{-1})\right) + o_p\left(\frac{1}{n}\right)$$

また，$\mathbb{E}[\|\xi_n\|^2] = \mathrm{tr}(IJ^{-1}) + o(1)$ である．

定理 7 において,

$$\lambda = \frac{d}{2}\ ,\ \nu = \frac{1}{2}\mathrm{tr}(IJ^{-1})$$

と表すと，$\beta = 1$ のとき,

$$\left(G_n - L(w_0)\right) + \left(T_n + \frac{2\nu}{n} - L_n(w_0)\right) = \frac{2\lambda}{n} + o_p\left(\frac{1}{n}\right)$$

である．また平均値としては，一般の β で

$$\mathbb{E}[G_n] = L(w_0) + \frac{1}{n}\left(\frac{\lambda-\nu}{\beta} + \nu\right) + o\left(\frac{1}{n}\right),$$
$$\mathbb{E}[T_n] = L(w_0) + \frac{1}{n}\left(\frac{\lambda-\nu}{\beta} - \nu\right) + o\left(\frac{1}{n}\right)$$

が成り立つ．これより，$\beta = 1$ のとき，$G_n - L(w_0)$ と $T_n + 2\nu/n - L_n(w_0)$ は平均，分散共に等しいことが分かる．また，$\beta = 1$ で真の分布が確率モデルで実現可能，すなわち，あるパラメータ w が存在して，$q(x) = p(x|w)$ とできる場合，$L(w_0) = S, \nu = d/2$ であるから,

$$\mathbb{E}[G_n] = S + \frac{d}{2n} + o\left(\frac{1}{n}\right),\ \mathbb{E}[T_n] = S - \frac{d}{2n} + o\left(\frac{1}{n}\right)$$

となる．このとき，汎化損失も経験損失も先頭の 2 つの項は β に依存しない．一方，$\beta = 1$ であっても真の分布が確率モデルで実現可能とは限らない場合には,

$$\mathbb{E}[G_n] = L(w_0) + \frac{\lambda}{n} + o\left(\frac{1}{n}\right),\ \mathbb{E}[T_n] = L(w_0) + \frac{\lambda - 2\nu}{n} + o\left(\frac{1}{n}\right)$$

が成り立つ．

4. 参考文献

[1] 渡辺澄夫: ”ベイズ統計の理論と方法”, コロナ社,2012

[2] 小谷眞一: ”測度と確率”, 岩波書店,2005