汎化を考慮した確率モデルの学習について†

2C－3

後藤正幸‡　松嶋敏泰‡　平澤茂一‡

早稲田大学 理工学部 工業経営学科

1 はじめに

学習データから情報源の確率構造を学習する場合，その結果は学習データに依存するので，何らかの方法で学習データから真の構造に近い確率モデルを推定する必要がある．そのための方法として，適切な次数のモデルを選択する方法があり，AIC や MDL などの情報量規準が提案されている．

AIC は期待平均対数尤度最大のモデルを選択することにより，KL 情報量の意味で真の構造に近い確率モデルを選択しようとする．MDL は，データの記述長とモデルの記述長の和を最小にするモデルを選択するという規準である．しかし，AIC では評価できない項を無視していることや一致性がないことが指摘されている．MDL ではパラメータを量子化する部分が粗い議論であるために，記述長最小という評価規準は厳密には達成されていない．

本稿では，情報源符号化において最近注目されているベイズ符号をモデル選択に応用することを考える．そして，ベイズ決定理論に基づき，次のシンボルを予測符号化したときのベイズ符号長を最小にするという意味で，漸近的に最適なモデルを選択する情報量規準を提案する．

2 従来の情報量規準

2.1 Akaike Information Criterion

AIC は対数尤度の不偏推定量を求めることにより，導出された [Akaike,74]．x_i を情報源アルファベットのシンボルとし，情報源からの系列（学習データ）を $x^n : x_1, x_2, \cdots, x_n$ ，あるモデル M の次元を k，k次元パラメータをθ^kと表すと，AIC は，

$$AIC(x^n) = -\log p_M^n(x^n|\hat{\theta}^k) + k \tag{1}$$

となる．ただし，$\hat{\theta}^k$ は最尤推定量を表す．

2.2 Minimum Description Length

MDL は情報源圧縮の面から確率モデルを選択する規準として導かれた [Rissanen,81]．

$$MDL(x^n) = -\log p_M^n(x^n|\hat{\theta}^k) + \frac{k}{2}\log n + \log|M| \tag{2}$$

となる．ただし，$|M|$ はモデルの数である．しかし，符号長の最小化という目的から見れば，パラメータ空間を直方体で量子化して符号化するという操作が必ずしも最良ではない．情報源符号化の立場からも，ベイズ決定理論に基づいたベイズ符号が提案されており [Matsushima,et.al,91]，有限時点での圧縮性能としては最も強力な符号化法となっている [Clarke et.al,90]．

†"A Note on Learning Theory for Probabilistic Models"，本研究の一部は，文部省科学研究費試験研究 B 07558168，早稲田大学特定課題研究 95A-267 の助成による．

‡M.GOTOH, T.MATSUSHIMA, and S.HIRASAWA
School of Science and Engineering, Department of Industrial Engineering and Management, Waseda University, 3-4-1 Okubo Shinjuku-ku TOKYO, 169 JAPAN

2.3 ベイズ的情報量規準

ベイズ的モデル選択規準の先駆けは，Schwarz によって提案された BIC である [Schwarz,78]．

$$BIC(x^n) = -\log p_M^n(x^n|\hat{\theta}^k) + \frac{k}{2}\log n \tag{3}$$

これは，漸近的に事後確率最大のモデルを選択しようとする規準であったが，これに対して Poskitt はモデル選択の期待効用として，期待 Information Gain を採用し，次の定理を用いて積分を計算した [Poskitt,87]．

[補題 1] [Poskitt,86] $\theta^k \in \Theta^k$ に対して，適当な正則条件のもとで，

$$\phi_M(\theta^k|x^n) = \left\{\left(\frac{n}{2\pi}\right)^k \det E_M(\hat{\theta}^k|x^n)\right\}^{1/2} \cdot \exp[n\{L_M(\theta^k|x^n) - L_M(\hat{\theta}^k|x^n)\}] \tag{4}$$

とおくと，これは $n \to \infty$ でデルタ関数へ確率収束する．ただし，

$$E_M(\theta^k|x^n) = -\frac{\partial^2}{\partial\theta^k \partial\theta^{kT}} L(\theta^k|M, x^n) \tag{5}$$

である．□

補題 1 を用い，Information Gain を期待効用とすると，次のモデル選択基準が導かれる．

$$\delta_M(x^n) = -\log p_M^n(x^n|\hat{\theta}^k) + \frac{k}{2}(1 + \log n) \tag{6}$$

3 ベイズ符号

情報源符号化の分野では，MDL が暗に事前分布を仮定しているのにも関わらず，パラメータを符号化する部分の議論が粗いことが指摘されていた．これに対し，事前分布に対して最適な冗長度を達成するベイズ符号が提案され，注目を集めている．

ベイズ符号では，事前分布 $p_M(\theta^k)$ で平均化したベイズ平均符号長 $BL(x^n|P_M(\theta^k), AP)$ の最小化により定式化される．

$$BL(x^n|P_M(\theta^k), AP) = -\int_{\theta^k \in \Theta^k} p_M(\theta^k) \sum_{x^n} AP(x^n|\theta^k) \log P_M(x^n|\theta^k) d\theta^k \tag{7}$$

[補題 2] [Matsushima,et.al,91] ベイズ平均符号長 $BL(x^n|P(\theta^k), AP)$ は，決定関数 $AP^*(x_t|x^{t-1})$

$$AP^*(x_t|x^{t-1}) = \int_{\theta^k \in \Theta^k} P_M(x_t|x^{t-1}, \theta^k) p_M(\theta^k|x^{t-1}) d\theta^k \tag{8}$$

による予測符号化により最小化される．□

4 ベイズ決定理論に基づく最小記述長モデルの選択基準

ここでは，最適なベイズ符号に基づいて，モデル選択規準を導出する．今，i.i.d 系列を仮定して，n 個の系列 x^n を観測したもとで，次の $z = x_{n+1}$ を予測符号化することを考える．このとき，次の z の符号長を最小化するモデルが，ベイズ符号による情報源符号化の面から，最適なモデルと考えられる．したがって，モデル選択の事後期待損失 $BM(z|x^n, AP)$ を以下のように定義する．

$$BM(z|x^n, AP) =$$

$$-\int_{\theta^k \in \Theta^k} p_M(\theta^k|x^n) \sum_z P_M(z|\theta^k) \log AP(z|x^n) d\theta^k \quad (9)$$

この事後期待損失の最小化は，ベイズ符号の意味で次に出てくる情報を最も圧縮できるモデルを最適と考えていることになる．以下では，この規準を漸近的に評価する．

補題 2 より，最適な決定関数 $AP^*_M(z|x^n)$ は，

$$AP^*_M(z|x^n) = \int_{\theta^k \in \Theta^k} p_M(\theta^k|x^n) P_M(z|\theta^k) d\theta^k \quad (10)$$

で表される．事後分布を，

$$p_M(\theta^k|x^n) =$$

$$\left\{ \left(\frac{n}{2\pi}\right)^k \det E_M(\hat{\theta}^k|x^n) \right\}^{-1/2} P_M(x^n|\hat{\theta}^k) p_M(\theta^k) \phi_M(\theta^k) \quad (11)$$

と変形する．

モデル M の事前分布は等確率とし，モデル M が与えられたもとでの，パラメータの事前分布 $p_M(\theta)$ は事前情報がないことを想定して，Jeffreys prior を用いることにする．

$$p_M(\theta^k) \propto \left\{ \det I_M(\theta^k) \right\}^{1/2} \quad (12)$$

このとき，$\det E_M(\hat{\theta}^k|x^n) \to \det I_M(\theta^k)$ より，

$$p_M(\theta^k|x^n) = C^{-1} \left(\frac{n}{2\pi}\right)^{-k/2} P_M(x^n|\hat{\theta}^k) \phi_M(\theta) \quad (13)$$

となる．C^{-1} は規準化定数である．補題 1 より，$\phi_M(\theta)$ の部分はデルタ関数に確率収束するから，漸近的に，

$$AP^*_M(z|x^n) = \int_{\theta^k \in \Theta^k} p_M(\theta^k|x^n) P_M(z|\theta^k) d\theta^k$$

$$= \int_{\theta^k \in \Theta^k} C^{-1} \left(\frac{n}{2\pi}\right)^{-k/2} P_M(x^n|\hat{\theta}^k) P_M(z|\theta^k) \phi_M(\theta^k) d\theta^k$$

$$= \left(\frac{n}{2\pi}\right)^{-k/2} C^{-1} p(\hat{\theta}^k|x^n) P_M(z|\hat{\theta}^k) \quad (14)$$

が得られる．このときの最適な事後損失 $BM^*(x^n)$ は，

$$BM^*(x^n) = \min_{AP} BM(z|x^n, AP)$$

$$= -\sum_z \left(\frac{n}{2\pi}\right)^{-k/2} C^{-1} P_M(x^n|\hat{\theta}^k) P_M(z|\hat{\theta}^k)$$

$$\cdot \log \left\{ \left(\frac{n}{2\pi}\right)^{-k/2} C^{-1} P_M(x^n|\hat{\theta}^k) P_M(z|\hat{\theta}^k) \right\}$$

$$= n^{-k/2} C^{-1} P_M(x^n|\hat{\theta}^k)$$

$$\cdot \left\{ H_{\hat{\theta}^k}(z) - \log P_M(x^n|\hat{\theta}^k) + \frac{k}{2} \log \frac{n}{2\pi} + \log C \right\} \quad (15)$$

と展開される．従って，漸近的な符号長は (15) 式で表されるので，これを最小にするモデル M を選べば，それが事後期待損失に対する最適なベイズ決定である．以上により，新たに次式のベイズ符号をもとにしたモデル選択規準 BDL (Bayes Decision Theoretic Description Length) が導出できた．

[定理] パラメータに関する適当な正則条件，及び適当な事前分布と x^n が i.i.d であることを仮定すると，事後損失関数 (9) 式を最小化する確率モデルの選択は，漸近的に次の規準の最小化と等価になる．

$$BDL = n^{-k/2} P_M(x^n|\hat{\theta}^k)$$

$$\cdot \left\{ H_{\hat{\theta}^k}(z) - \log P_M(x^n|\hat{\theta}^k) + \frac{k}{2} \log \frac{n}{2\pi} \right\} \quad (16)$$

ここで，$H_{\hat{\theta}^k}(z)$ は，

$$H_{\hat{\theta}^k}(z) = \sum_z P_M(z|\hat{\theta}^k) \log P_M(z|\hat{\theta}^k) \quad (17)$$

であり，$\theta^k = \hat{\theta}^k$ のときの確率モデルのエントロピーである．□

ここで，モデルの事前分布 $P(M)$ を，

$$P(M) \propto (2\pi)^{-k/2} \quad (18)$$

とおけば，規準の $\frac{k}{2} \log 2\pi$ がキャンセルされる．これは，次数の低いモデルほど選ばれやすい，けちの原理を反映している．

5 まとめ

本稿では，新にベイズ符号に基づく漸近的に最適なモデル選択規準を導出した．この規準の性質（一致性など）を理論的に調べること，及び実験的な検証が今後の課題である．

参考文献

[1] H. Akaike: "A new look at the Statistical Model Identification", *IEEE Trans. on Auto. Contr.* AC-19, No.6, pp.716-722,(1974)

[2] B.S.Clarke, and A.R.Barron: "Information-Theoretic Asymptotics of Bayes Methods", *IEEE Trans. on Information Theory*, Vol.36, No.3, pp.453-471,(1990)

[3] T.Matsushima, H.Inazumi, and S.Hirasawa: "A Class of Distortionless Codes Designed by Bayes Decision Theory", *IEEE Trans. on Information Theory*, Vol.37, No.5, pp.1288-1293, (1991)

[4] D.S.Poskitt: "A Bayes Procedure for the Identification of Univariate Time Series Models", *The Annals of Statistics*, Vol.14, No.2, pp.502 - 516, (1986)

[5] D.S.Poskitt: "Precision, Complexity and Bayesian Model Determination" , *J. R. Statist. Soc.* B, 49, No.2, pp.199-208, (1987)

[6] J.Rissanen: "Stochastic Complexity and Modeling", *Annals of Statistics*, Vol.14, No.3, pp.1080-1100, (1986)

[7] C.Schwarz: "Estimating the dimension of a model", *Ann. Statist.*, 6, pp.461-464, (1978)