4E−2　MDL原理によるLaplace型推定量の導出と確率的PAC学習モデルによるその性能評価

竹内純一
NEC C&C情報研究所

1　はじめに

二項確率$p^* \in [0,1]$を，有限回の独立試行をもとに推定する問題を考える．二項確率とは，標本空間$X=\{0,1\}$上の確率測度のもとで，1が起こる確率(値)を指すものとする．ここでは，p^*を定めればX上の確率測度が定まるので，p^*自体を確率測度またはモデルと呼ぶ．また，$H=[0,1]$と書く．

いま，未知のモデルp^*のもとで，N回の独立試行を行って得られる標本を$\xi_N=\{x_1,x_2,\cdots,x_N\}$とし，このなかの1(正例)の数を$k$とする(以下$\xi_N$は全て同じものを指すものとする)．このとき，$p^*$に対する推定量$\frac{k+1}{N+2}$を，Laplaceの推定量と呼ぶ．この推定量は通常Bayesianの立場から導かれる．すなわち，モデルpの事前分布(H上の確率測度)を一様分布として事後分布を求め，そのもとで得られるpの期待値が$\frac{k+1}{N+2}$である．しかし，実際的な場面で良い結果が得られることが多いため，多くの実験家(非Bayesianを含めて)がこれを用いている．

本稿では，$\frac{k+a}{N+2a}(a>0)$の形の推定量をLaplace型推定量と呼ぶことにし，この形の推定量がなぜ良い結果をもたらすのかを非Bayesianの立場から考察する．

具体的には，

1. MDL原理に基づいて，Laplace型推定量を導き，
2. KL情報量を距離基準とする確率的PAC学習モデルにおいて，Laplace型推定量の収束速度が迅速であることを示す．

確率的PAC(Probably Approximately Correct)学習モデル[1, 7]は，概念学習に関するPAC学習モデル[5]の確率的学習対象への拡張である．そこでは，ある距離基準d(本稿ではKL情報量を用いる)を設け，任意の$p^* \in H$，ε(精度パラメータ)，δ(信頼度パラメータ)に対し，少なくとも$1-\delta$以上の確率で，$d(p^*|q)<\varepsilon$なる仮説qを出力するために十分な標本数を，$\frac{1}{\varepsilon}$と$\frac{1}{\delta}$の関数として表したものを必要最小事例数の上界と呼ぶ．ここで，p^*によらない一様な上界であることに注意されたい．

KL情報量を基準としたときの必要最小事例数の上界は，$O(\frac{1}{\varepsilon^2})$なる結果が[1, 2]で示されている．そこでは，KL情報量の発散(確率値が0と1に近いところでおきる)を避けるために，仮説の集合から，0に近い確率をもった仮説を除くという手法を用いており，本質的にはLaplace的な推定を行っている．すなわち，多項式数の必要最小事例数を得るために，すでにLaplace的な推定は不可欠であったといえる．本稿では，推定量として真にLaplace型を用い，この場合に必要最小事例数の上界が$O(\frac{1}{\varepsilon}\log^2\frac{1}{\delta})$となることを示す．

2　MDL原理によるLaplace型推定量の導出

MDL(Minimum Description Length)原理について簡単に復習しよう([7]参照)．MDL原理とは，一言でいうと「与えられた標本を，モデル自身の記述長を含めて，最も短い記述長で記述できるモデルを選択せよ．」と言うことである．

ここで，標本の記述長とはある仮説qの下での理想符号を用いて標本を記述した場合の符号長である．この符号長が短かいほどモデルが標本に適合していることを意味する．符号を復号するためには，どのように符号化されているのかを知らねばならず，従ってどのモデルを仮定しているのかを記述する必要がある．そのための符号長がモデル記述長であり，モデルの複雑さを表す．MDL原理とは，これらの和を最小とするモデルを最良とする原理である．

以下に，二項確率の推定問題にMDL原理を適用する．モデルqのもとでの標本ξ_Nの記述長を$l(\xi_N|q)$と書くと，

$$l(\xi_N|q)=-k\log q+(N-k)\log(1-q) \tag{1}$$

である(logの底は2とし，lnは自然対数とする)．モデルを記述する場合，一意に復号できるならばどんな符号を用いてもよいが，我々は最も自然と思える符号を用いる．すなわち，qの値そのものの二進小数展開を用いる．ここで，もしqの値が無限小数であったら，無限の符号長になってしまうが，その様な値のqは記述しないことにする．すなわち，ある有限の長さで展開を打ち切る(丸める)ことにする．例えば，一様に$-\log\delta$桁の小数を用いると，これはδの間隔の量子点を記述することになる．ここでは，最小の記述長で記述するために，qの値とデータ数に応じて符号長を変えることにする．今，$[0,\frac{1}{2}]\subset H$上の量子点を小さい順に$\bar{q}_1,\bar{q}_2,\cdots,\bar{q}_s$と書く．また，$\delta_i=\bar{q}_i-\bar{q}_{i-1}$とする．このとき，$\bar{q}_i$の符号長$l(\bar{q}_i)$について，$l(\bar{q}_i)\geq-\log\delta_i$ならば，一意に復号できる．よって，$l(\bar{q}_i)=-\log\delta_i$なる符号を用いる．全記述長は，

$$l(\xi_N|\bar{q}_i)+l(\bar{q}_i)=-k\log\bar{q}_i-(N-k)\log(1-\bar{q}_i)-\log\delta_i \tag{2}$$

となる．標本記述長については，$\bar{q}_i=\frac{k}{N}$とすれば最小になるが，丸め誤差のため一般にはそれは出来ない．δ_iを小さくすればするほど，標本記述長は理想的な値に近くできるが，それはモデル記述長を大きくすることを意味する．従って，標本記述長とモデル記述長の間にはトレードオフがある．$\hat{q}=\frac{k}{N}$とおくと，$\bar{q}_i-\hat{q}\cong\delta_i$である．ここで，標本記述長を一次の係数が0であることに注意して，$\hat{q}$のまわりでTaylor展開すると，

$$l(\xi_N|\bar{q}_i)+l(\bar{q}_i)\cong l(\xi_N|\hat{q})+N\frac{\delta_i^2}{\hat{q}(1-\hat{q})\ln 2}-\log\delta_i \tag{3}$$

となる．δ_iで微分すると，$2N\frac{\delta_i}{\hat{q}(1-\hat{q})\ln 2}-\frac{1}{\delta_i\ln 2}$となるから，全記述長は，

$$\delta_i=\sqrt{\frac{\hat{q}(1-\hat{q})}{2N}}\cong\sqrt{\frac{\bar{q}_i(1-\bar{q}_i)}{2N}} \tag{4}$$

のとき最小になる．よって，$\delta_i=\sqrt{\frac{\bar{q}_i(1-\bar{q}_i)}{2N}}$と定める．また，$\bar{q}_1=\frac{1}{2N+2}$とする．$(0.5,1]$に属するモデルについても，対称に量子点をとれば良い．このとき，

$$l(\bar{q}_i)=-\frac{1}{2}(\log\bar{q}_i+\log(1-\bar{q}_i))+\log 2N \tag{5}$$

となる．従って，全記述長$l(\xi_N|\bar{q}_i)+l(\bar{q}_i)$は

$$-\left(k+\frac{1}{2}\right)\log\bar{q}_i-\left(N-k+\frac{1}{2}\right)\log(1-\bar{q}_i)+\log 2N \tag{6}$$

となる．これを最小にするモデルは，

$$\bar{q}_i=\frac{k+\frac{1}{2}}{N+1}+\Delta \tag{7}$$

で与えられる．ここで，Δは丸め誤差であり，その絶対

Derivation of a Laplace-like Estimator based on the MDL principle and its Evaluation in the Probabilistic PAC Learning Model. Jun-ichi Takeuchi. C&C Information Technology Research Laboratories, NEC Corporation.

値は $\frac{1}{2}\sqrt{\frac{\hat{q}(1-\hat{q})}{2N}}$ 以下である．以上により，MDL 原理から Laplace 型推定量が導かれることがわかった．

なお，以上の議論は [7] に基づいている．[7] においても，式 5の第一項にあたる項を導いているが，次数選択には影響しないとして無視している．また [6] では，これとほぼ同様の議論で $\frac{k+0.5+a}{N+1+a+b}$ という推定量を導いている．ただし，彼らの立場は Bayesian(的な MDL) であり，事前分布を仮定している (a, b は事前分布のパラメータ)．一方で本稿の議論は符号長のみに基づいている．

3 Laplace 型推定量の性能

標本 ξ_N に基づく p^* の推定量を次の様に定義する．

$$\hat{p}_a = \frac{k+a}{N+2a} \quad (a>0) \tag{8}$$

$$\hat{p}_{MDL} = \frac{k+\frac{1}{2}}{N+1} + \Delta \tag{9}$$

ここで，前者は Laplace 型推定量，後者は MDL 推定量であり，$|\Delta| \leq \frac{1}{2}\sqrt{\frac{\hat{p}(1-\hat{p})}{2N}}$ $(\hat{p}=\frac{k}{N})$ が成り立つ．

我々は，これらの性能を確率的 PAC 学習モデルにおける必要最小事例数によってみることにする．距離基準として用いる KL 情報量を二項確率に限って定義しよう．

定義 1 (KL 情報量) p と q を H の要素とする．p に対する q の *KL* 情報量 $d(p|q)$ は次式で定義される．

$$d(p|q) \equiv p\ln\frac{p}{q} + (1-p)\ln\frac{1-p}{1-q}. \tag{10}$$

ここで，$d(p|q) \geq 0$(等号は $p=q$ のときのみ成り立つ) である．また，$\exp(-Nd(p|q))$ は，p による最も確からしいサイズ N の標本が，q のもとで得られる確率にほぼ比例するという意味をもつ．

上述の推定量の性能は，以下の二つの定理により保証される．

定理 1 (Laplace 型推定量の必要最小事例数) 任意の $p^* \in H$，$\varepsilon \in (0, 0.5)$，$\delta \in (0,1)$ に対し，

$$N > \max\left\{\frac{2048}{9a\varepsilon}\ln^2\frac{4}{\delta}, \frac{4a}{\varepsilon}\right\} \tag{11}$$

ならば，$1-\delta$ 以上の確率で $d(p^*|\hat{p}_a) < \varepsilon$ が成立する．

定理 2 (MDL 推定量の必要最小事例数) 任意の $p^* \in H$, $\varepsilon \in (0, 0.5)$, $\delta \in (0,1)$ に対し，

$$N > \frac{4096}{9\varepsilon}\ln^2\frac{4}{\delta} \tag{12}$$

ならば，$1-\delta$ 以上の確率で $d(p^*|\hat{p}_{MDL}) < \varepsilon$ が成立する．

これらの定理の証明は略すが，次の二つの補題を用いて証明出来る．

補題 1 (Bernstein の不等式) $Y_i, (i=1,2,\cdots,N)$ は，平均が 0 で絶対値の上界が M の，N 個の独立な確率変数であるとし，それぞれの分散を σ_i^2 とする．ここで，$VN \geq \sigma_1^2+\sigma_2^2+\cdots+\sigma_N^2$ とすると，任意の $\beta > 0$ に対し，

$$\Pr\left\{\frac{|Y_1+Y_2+\cdots+Y_N|}{N} > \beta\right\} \leq 2\exp\left(-\frac{\beta^2 N}{2V+\frac{2}{3}M\beta}\right) \tag{13}$$

が成り立つ．
証明は [3] にある．

補題 2 (KL 情報量に関する不等式) $p^* \leq \frac{1}{2}$ を仮定する．$\kappa = p-p^*$ とすると，$\frac{1-p^*}{2} \geq \kappa \geq -\frac{p^*}{2}$ ならば，次の二つの式が成り立つ．

$$d(p^*|p) \leq \frac{\kappa^2}{(1-p^*)p^*}, \quad d(p^*|p) \leq 2|\kappa|.$$

証明は略す．

4 おわりに

本稿では，Laplace 型推定量に二つの根拠を与えた．一つは，Laplace 型推定量は MDL 原理からの自然な帰結であるという点であり，もう一つは Laplace 型推定量の (KL 情報量を距離基準とした) 必要最小事例数が $O(\frac{1}{\varepsilon})$ であり，真のモデルへの収束の速さという点で確かに Laplace 型推定量が有効であるということである．これは，MDL 原理がパラメータ推定にも有効に作用することを示したともいえる．

MDL 原理は通常，パラメータ数が異なる (パラメトリック) モデルを選択するのに用いられ，その有効性は Hellinger 距離[1]を基準として [7] に示されている．具体的には，必要最小事例数として，$O(\frac{s}{\varepsilon}\log\frac{1}{\varepsilon}+\frac{1}{\varepsilon}\log\frac{1}{\delta})$ (s は真のモデルのパラメータ数) が示されている．しかし，KL 情報量を基準としたときに，[7] と同様の問題設定で $O(\frac{1}{\varepsilon})$ なるオーダの必要最小事例数の上界は報告されていない[2]．本稿の結果は，上記問題設定においても MDL 基準をパラメータ推定にまで適用すれば，KL 情報量を基準とした場合でも，$O(\frac{1}{\varepsilon})$ の事例数で十分である可能性を示したといえる．しかし，[7] の証明技法が，MDL の性質を本質的に用いて複数のパラメトリックモデルの分離が可能なことを示しているのに対し，本稿の証明技法は MDL の性質を本質的に用いてはいない．このため，本稿の結果をパラメトリックモデルが複数ある場合に拡張する問題は，自明ではない．今後，この結果を MDL の性質と直に結びつけることが望まれる．

謝辞

議論して頂き，貴重な助言を下さった NASA の Buntine 氏，NEC の安倍氏，山西氏，岡村氏に感謝いたします．また，UCSC の Haussler 教授と安倍氏には Bernstein の不等式を教えて頂きました．

参考文献

[1] Abe, N. & Warmuth, M.(1992). On the computational complexity of approximating distributions by probabilistic automata. *Machine Learning, a special issue for COLT '90, 9(2/3).*

[2] Abe, N.,Takeuchi, J., & Warmuth, M.(1991). Polynomial Learnability of Probabilistic Concepts with respect to the Kullback-Leibler Divergence. *Proceedings of COLT '91* (pp. 277-289).

[3] Pollard, D.(1984).Convergence of Stochastic Processes. (pp. 191-193), Springer Verlag.

[4] Takeuchi, J.(1992). Some Improved Sample Complexity Bounds in the Probabilistic PAC Learning Model. *To appear, Proceedings of ALT '92.*

[5] Valiant, L.G.(1984). A theory of the learnable. *Communications of the ACM, 27,* 1134-1142.

[6] Wallace, C.S. & Freeman, P.R.(1987). Estimating and Inference by Compact Coding. *J. of Royal Stat. Soc.(B).* 49. No.3 pp.240-265.

[7] Yamanishi, K.(1992). A learning criterion for stochastic rules. *Machine Learning, a special issue for COLT '90, 9(2/3).*

[1] $d_H(p|q) \equiv |\sqrt{p}-\sqrt{q}|^2 + |\sqrt{1-p}-\sqrt{1-q}|^2$ で定義される．$d \geq d_H$ であるが，どんな自然数 i に対しても，$(\forall p,q \in H) d_H \geq d^i$ は成立しない．

[2] 学習アルゴリズムが入力として ε およびモデルの複雑さの上限をもらえるという，緩い PAC モデルのもとでは [4] に $O(\frac{1}{\varepsilon})$ が示されている．