統計的学習におけるモデル選択への情報量尺度にもとづくアプローチ

田 中 利 幸　　松 井 秀 往 *

An Information-Measure-Based Approach
to Model Selection in Statistical Learning

Toshiyuki TANAKA and Hideyuki MATSUI *

We study an approach to statistical learning via information-theoretic framework, in which we regard a problem of statistical learning as a problem of communication, by viewing model parameters and output data as channel input and output, respectively, and elucidate possibility of utilizing concepts of information theory, such as mutual information, in dealing with the statistical learning problem. We consider a problem of model selection in statistical learning, and propose use of mutual information between model parameters and output data as a criterion of model selection. In order to circumvent computational difficulty in evaluating the mutual information, we consider an approximation scheme that is based on the naïve mean field approximation.

1. はじめに

容易に得られる情報から本当に知りたい情報までの道のりは，往々にして遠い．両者の距離を埋めるために様々な情報処理がなされる．例えば，無線通信では，ノイズや干渉などの影響を受けて劣化した受信信号から送信情報を推理する必要があるし，スパムフィルタにおけるパターン認識では，メールの様々な特徴にもとづいてそれがスパムメールであるかないかを識別する必要がある．このように捉えると，通信の問題もパターン認識の問題も，その本質においては同一の数理的構造を有していることがわかる．

本稿では，両者に共通する数理的構造に注目し，パターン認識の問題を通信の枠組みで捉えるアプローチについて検討する．具体的には，確率的システムの入出力例の集合からシステムのパラメータを推測する統計的学習の問題を，システムのパラメータを送信情報とし出力を受信信号とする通信の問題として捉え，情報エントロピー等の情報理論の概念を統計的学習の問題設定に対して適用していく可能性について基礎的な考察を加えることを目的とする．

2. 基礎的事項

2.1 統計的学習とモデル選択

統計的学習の基本的な枠組みは，以下のように表される：N 個の入出力データ対からなる集合（訓練データ集合）$T_N = \{(x_i, y_i) \mid i = 1, \ldots, N\}$ が与えられたとき，パラメータ θ をもつシステム $p(y \mid x, \theta)$ が T_N を生成したものと考えて未知パラメータ θ を求める．y が 0 または 1 の値をとるものとすればパターン分類，識別の問題であるし，y が実数値をとるものとすれば関数近似や回帰の問題となる．

学習者は，未知パラメータ θ を確率変数として扱うものとし，θ の事前分布を $p_\zeta(\theta)$ とおく．ζ は事前分布のパラメータ（ハイパーパラメータ）である．学習者はまた，システムもハイパーパラメータ ξ を有するとみなすものとする．これを明示するために，システムの入出力関係を $p_\xi(y \mid x, \theta)$ と表記する．

学習者は，ハイパーパラメータ ξ および ζ の値を何らかの手段で定める必要があり，本稿ではこの問題を考察する．ξ, ζ の値を定めることに対応してシステムのモデルが定まると考えれば，ハイパーパラメータの決定の問

2007 年 03 月 31 日　受理
*京都大学大学院情報学研究科

題はモデル選択の問題だとみなすことができるので，以下ではこの問題をモデル選択問題と呼ぶ．

2.2 相互情報量にもとづくモデル選択基準

ハイパーパラメータξ,ζの値を決めれば，入力データ$x^N=\{x_1,\ldots,x_N\}$が与えられたときの出力データに対応する確率変数$Y^N=\{Y_1,\ldots,Y_N\}$とパラメータθとの間の相互情報量$I(Y^N;\theta)$が定まる．相互情報量$I(Y^N;\theta)$はY^Nが未知パラメータθについてどれだけの情報を有しているかを定量的に表わす尺度である．相互情報量$I(Y^N;\theta)$はパラメータξおよびζの関数となるから，モデル選択問題に対する一つのアプローチとして，$I(Y^N;\theta)$をなるべく大きくするようにパラメータξ,ζを定めるというやり方が考えられる．

入力データx^Nが与えられたときの相互情報量$I(Y^N;\theta)$は，Y^Nとθとの結合分布$p_\xi(Y^N,\theta|x^N)$によって

$$I(Y^N;\theta)$$
$$=D[p(Y^N,\theta|x^N)\|p(Y^N|x^N)p_\zeta(\theta)] \quad (1)$$

と与えられる．ただし，

$$p(Y^N|x^N)=\int p(Y^N,\theta|x^N)d\theta \quad (2)$$

である．また，$D[p(z)\|q(z)]$は確率分布$p(z)$から$q(z)$へのKullback-Leibler(KL)ダイバージェンスであり，以下のように定義される．

$$D[p\|q]=\int p(z)\ln\frac{p(z)}{q(z)}dz$$

結合分布$p(Y^N,\theta|x^N)$は

$$p(Y^N,\theta|x^N)=\prod_{i=1}^{N}p_\xi(Y_i|x_i,\theta)p_\zeta(\theta)$$

で与えられるから，相互情報量$I(Y^N;\theta)$は事前確率$p_\zeta(\theta)$，システム$p_\xi(y|x,\theta)$，および入力データx^Nが与えられれば原理的には計算可能な量であり，モデル選択基準として使うことができると考えられる．

3. 相互情報量の近似的評価法

上述のように相互情報量$I(Y^N;\theta)$は原理的には計算可能であるが，実際の問題においてはパラメータθは高次元ベクトルであることが多く，$I(Y^N;\theta)$を評価しようとすると式(2)が多次元積分となって，これを数値的に実行するとなると高い計算複雑度をもつ．また，KLダイバージェンスの計算の際に現れるY^Nに関する積分も数値的に行おうとするとやはり高い計算複雑度を有する．このため，$I(Y^N;\theta)$の評価は実際には困難であることが多い．本節では，訓練データ集合T_Nのなかの出力データy^Nを利用して相互情報量$I(Y^N;\theta)$の評価の困難さを回避する近似の枠組みを定式化し検討する．

まず，以下の分布を定義する．

$$P_e(Y^N|x^N)=\prod_{i=1}^{N}\delta(Y_i-y_i)$$

$\delta(\cdot)$はデルタ関数であり，$P_e(Y^N|x^N)$は訓練データ集合T_Nから決まる出力データ$y^N=\{y_1,\ldots,y_N\}$の経験分布である．この分布を使って相互情報量$I(Y^N;\theta)$を

$$I(Y^N;\theta)$$
$$=D[p(Y^N,\theta|x^N)\|P_e(Y^N|x^N)p_\zeta(\theta)]$$
$$-D[p(Y^N|x^N)\|P_e(Y^N|x^N)] \quad (3)$$

と分解することができる．次に，KLダイバージェンスの二つの引数を入れ替えることによって相互情報量を近似する：

$$I(Y^N;\theta)$$
$$\approx D[P_e(Y^N|x^N)p_\zeta(\theta)\|p(Y^N,\theta|x^N)]$$
$$-D[P_e(Y^N|x^N)\|p(Y^N|x^N)] \quad (4)$$

一般に$D[p\|q]\neq D[q\|p]$であるから，ここでの式変形は相互情報量に対するひとつの近似を与えることになる．$D[p\|q]$を$D[q\|p]$で置き換える近似は，統計力学では

ナイーブ平均場近似と呼ばれているものに相当し，統計的学習の分野でも多くの研究がなされている[1]．式(4)右辺を具体的に計算すると，

$$I(Y^N;\theta) \approx D[p_\zeta(\theta) \| p(\theta|T_N)] \tag{5}$$

となる．$p(\theta|T_N)$は訓練データ集合T_Nが得られたときのパラメータθの事後分布である．したがって，式(5)の近似のもとで相互情報量の最大化を考えることは，直観的にはパラメータθの事後分布$p(\theta|T_N)$が事前分布$p_\zeta(\theta)$からみてなるべくかけ離れた分布となるようにハイパーパラメータξ,ζを決めることに対応している．

式(5)の右辺はまた，

$$D[p_\zeta(\theta) \| p(\theta|T_N)] = \ln \int p(y^N,\theta|x^N)d\theta - \sum_{i=1}^{N} \int p_\zeta(\theta) \ln p_\xi(y_i|x_i,\theta)d\theta \tag{6}$$

と書き直すことができる．式(6)右辺第一項はハイパーパラメータの対数周辺尤度に相当し，ハイパーパラメータの最尤推定はこの項を最大化するようにハイパーパラメータを選ぶことに相当する．本節で導入した相互情報量の近似的評価法は，ハイパーパラメータの対数周辺尤度に対して式(6)右辺第二項を付加したものだと理解することもできる．この付加項の是非については，次節で数値シミュレーションにもとづいて考察する．

相互情報量の近似を数値的に実際に評価する際にも，式(6)右辺の形で表しておくと都合がよいことが多い．既に述べたように，θに関する積分は高次元積分となって高い計算複雑度をもつことが多く，そのような場合には式(6)右辺を直接数値的に評価するという形で提案手法を適用することはやはり実用的でない．しかし一般に，意味のある統計的学習を行うにはパラメータθの値をある程度特定できるだけの訓練データ数が必要であり，そのような状況では式(6)右辺第一項に現れる被積分関数$p(y^N,\theta|x^N)$はθのもっともらしい値付近に局在した関数となっているであろうことが期待できる．我々はSollich[2]に倣い，被積分関数$p(y^N,\theta|x^N)$をθのガウス形関数で近似したうえで式(6)右辺第一項の積分を解析的に行う近似手法を検討する．これはラプラス近似と呼ばれている．上述の状況のもとでは，ラプラス近似はよい近似を与えると期待することができる．一方，式(6)右辺第二項の積分については，具体的なシステムの性質に応じた工夫が必要であると考えられる．これについては次節で再び議論する．

4. 数値シミュレーション

4.1 パーセプトロン

本節では，提案手法を具体的にパーセプトロン学習の問題に適用した例を示す．

入力データはd次元実ベクトル空間に値をとるものとし，$x,\theta \in \Re^d, y \in \{-1,1\}$とする．パーセプトロンは，以下で定義される．

$$p(y|x,\theta) = \frac{e^{y\theta\cdot x}}{e^{\theta\cdot x}+e^{-\theta\cdot x}}$$

ハイパーパラメータξは考えないものとする．パラメータθの事前分布はハイパーパラメータζをもつd次元正規分布

$$p_\zeta(\theta) = \frac{1}{(2\pi\zeta^2/d)^{d/2}} e^{-d|\theta|^2/2\zeta^2}$$

であるものとする．

以上の定義のもとで，式(6)右辺第二項の積分は一変数の積分に帰着させることができる．また，右辺第一項の積分については，θの事後平均

$$\hat{\theta} = \frac{\int \theta\, p(y^N,\theta|x^N)d\theta}{\int p(y^N,\theta|x^N)d\theta}$$

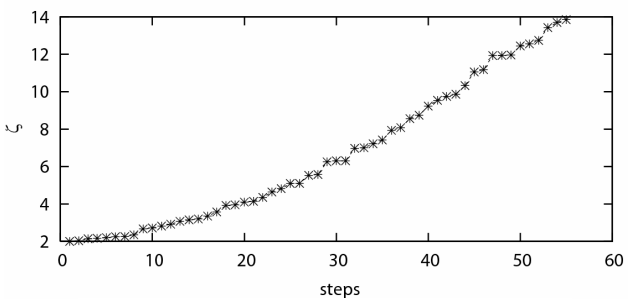
を平均とし，被積分関数のヘシアンによって定まる分散共分散行列をもつd次元正規分布の確率密度関数で被積分関数を近似する（ラプラス近似）ことによって解析的に評価できる．また，θの事後平均$\hat{\theta}$は確率伝搬法[3]によって近似的に評価できる．

4.2 数値シミュレーションの条件

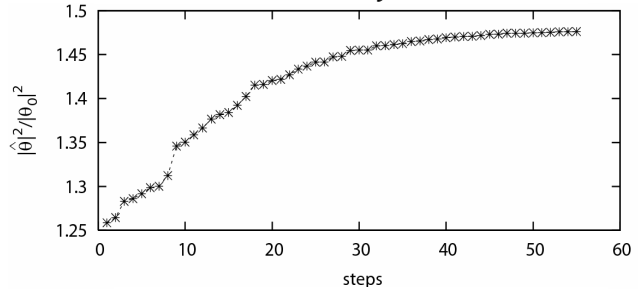
学習に使用した訓練データ集合は以下の手続きによって構成した：ハイパーパラメータζを1とし，パラメータθを事前分布$p_{\zeta=1}(\theta)$に従って定めた．これをθ_0とおく．$\{x_1,\ldots,x_N\}$をd次元正規分布

$$p(x) = \frac{1}{(2\pi)^{d/2}} e^{-|x|^2/2}$$

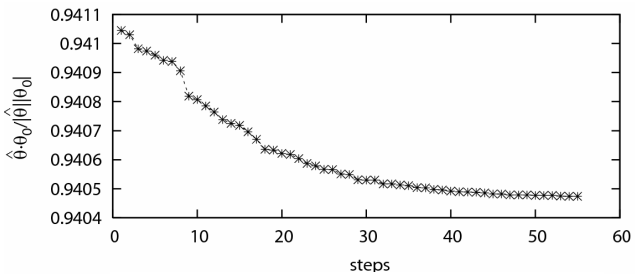
に従って独立に定め，パラメータθ_0をもつパーセプトロン$p(y|x,\theta_0)$によって入力データx^Nに対応する出力デ



(a) ハイパーパラメータ ζ の推定値の推移

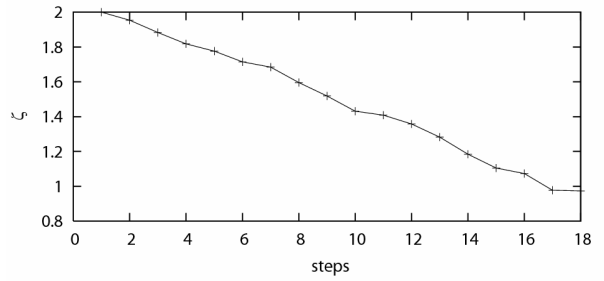


(b) $|\hat{\theta}|^2/|\theta_0|^2$ の値の推移

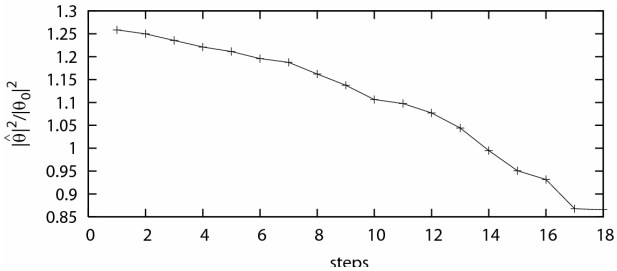


(c) $\hat{\theta}\cdot\theta_0/|\hat{\theta}||\theta_0|$ の値の推移

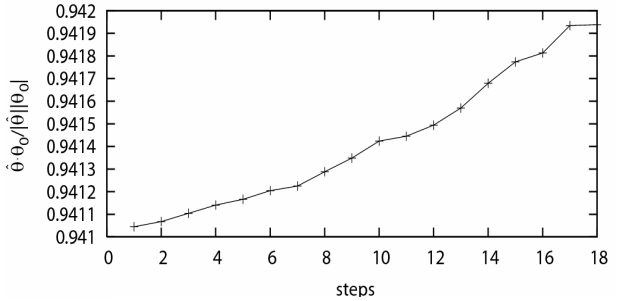
図 1　相互情報量を近似的に評価したものをモデル選択基準としたときの結果



(a) ハイパーパラメータ ζ の推定値の推移



(b) $|\hat{\theta}|^2/|\theta_0|^2$ の値の推移



(c) $\hat{\theta}\cdot\theta_0/|\hat{\theta}||\theta_0|$ の値の推移

図 2　対数周辺尤度をモデル選択基準としたときの結果

ータ y^N を定めた．$T_N=\{(x_i,y_i)\,|\,i=1,\ldots,N\}$ とした．

学習者のハイパーパラメータ ζ の推定には「貪欲法」による推定値の反復更新アルゴリズムを使った．すなわち，$\Delta\zeta$ をランダムに生成し，ζ に対する相互情報量と $\zeta+\Delta\zeta$ に対する相互情報量とをそれぞれ近似的に評価して，後者が前者より大きければハイパーパラメータの推定値を $\zeta+\Delta\zeta$ に更新する，という手順を反復した．

ハイパーパラメータの対数周辺尤度(式(6)右辺第一項)の最大化を行う手法についても数値シミュレーションを行い，結果を比較した．

4.3 結果および考察

図 1 に結果を示す．図 1(a)はハイパーパラメータの各更新ステップにおいて学習者によって推定されたハイパーパラメータ ζ の推定値の推移を，図 1(b)，(c)はそれぞれ $|\hat{\theta}|^2/|\theta_0|^2$，$\hat{\theta}\cdot\theta_0/|\hat{\theta}||\theta_0|$ の値の学習による推移を表している．$|\hat{\theta}|^2/|\theta_0|^2$，$\hat{\theta}\cdot\theta_0/|\hat{\theta}||\theta_0|$ の値はともに 1 に近ければ近いほどパラメータの学習結果 $\hat{\theta}$ が真のパラメータ θ_0 に近いことを表わす．また，比較のためにハイパーパラメータの対数周辺尤度(式(6)右辺第一項)をモデル選択基準とした手法について対応する結果を示

したのが図 2(a)－(c)である．

ハイパーパラメータの対数周辺尤度をモデル選択基準とした場合の結果(図2)では,パラメータの学習結果$\hat{\theta}$とハイパーパラメータζとがいずれも真の値θ_0，$\zeta=1$に漸近している様子が見られ，モデル選択基準が有効に機能していることがわかる．一方，相互情報量を近似的に評価したものをモデル選択基準とした場合の結果(図1)は,$\hat{\theta}\cdot\theta_0/|\hat{\theta}||\theta_0|$の値は1に近い値に収束している(図1(c))ものの，$|\hat{\theta}|^2/|\theta_0|^2$の値は1より大きい値に収束しており(図1(b))，真のパラメータθ_0を正確に学習できているとは言い難い．また，ハイパーパラメータζの推定値は収束しておらずほぼ単調に増大している(図 1(a))．この結果は，相互情報量を近似的に評価したものがモデル選択基準として有効に機能していないことを示唆している．

対数周辺尤度の最大化によってハイパーパラメータを定める方法は，データ数が十分大きければハイパーパラメータのよい推定値を与えることが期待できる．それに対して，本稿の第3節において定式化した相互情報量の近似的評価法は,式(6)右辺第二項の影響のためにデータ数が十分大きかったとしてもハイパーパラメータ推定に偏りが生じ，そのためによい推定結果が得られないものと考えられる．しかし，以上の結果は直ちにモデル選択基準としての相互情報量の有効性を否定するものではない．相互情報量を近似的に評価するために導入されたナイーブ平均場近似が，モデル選択基準としての相互情報量の有効性を損ねている可能性があると我々は考えている．相互情報量をより適切に評価する近似手法について検討していくことが，今後の重要な課題である．

5. おわりに

統計的学習におけるモデル選択の問題に対して，相互情報量をモデル選択基準として使うアプローチを議論した．実際の統計的学習の問題において相互情報量を数値的に評価するのは困難であると考えて，ナイーブ平均場近似にもとづく相互情報量の近似的評価手法を定式化し,この手法の性質を数値シミュレーションによって検討した．数値シミュレーションの結果からは，本稿で定式化したモデル選択基準の有効性を示すには至らなかった．

本稿での結果は直ちに相互情報量がモデル選択基準として有効でないことを意味するわけではない．相互情報量を評価する際に使用した近似手法がモデル選択基準としての相互情報量の有効性を損ねている可能性があり,相互情報量の評価に際してより適切な近似手法の検討を進めていく必要がある．

引用文献

1) M. Opper and D. Saad (eds.), Advanced Mean Field Methods : Theory and Practice, The MIT Press, 2001.
2) P. Sollich, Bayesian methods for support vector machines : Evidence and predictive class probabilities,” Machine Learning, **46**, pp. 21－52, 2002.
3) T. Tanaka, “Derivation of a belief-propagation-based parallel interference cancellation algorithm for CDMA multiuser detection problem,” Proceedings of 2004 SMAPIP Workshop on Mathematics of Statistical Inference, Sendai, Japan, pp. 18－27, December 2004.