重回帰分析の変数選択におけるベイズ的方法
– 平均場近似に基づくアプローチ –
ISM Research Memo. No.624

伊庭幸人 (統計数理研究所)*

Abstract— Variable selection for multiple regression models (noisy linear perceptrons) is studied with a mean field approximation. Variable selection is formulated as estimation of discrete parameters that indicate a subset of the regression coefficients and a mean field approximation is introduced for the calculation of the posterior marginals over the space of submodels. Application to an real world example is shown.

Keywords—Mean Field Approximation, Model Selection, Multiple Regression, Bayes

1 背景

1.1 Neural Net における generalization

多層パーセプトロンなどにおいて、パラメータの数が多すぎると、例をそのまま覚えてしまって規則を抽出できなくなるという問題は早くから知られていた (generalization の困難)。これは統計学的にいえば当然のことであり、多層パーセプトロンでなくても、重回帰モデル ("Noisy Linear Perceptron") などの線形モデルですでにみられる。パーセプトロンの場合には、学習を途中で止めるという方法もあるが、現在では変数選択や罰金項の追加など、通常の統計学で overfitting を押える方法を採り入れる方向で話が進んでいるように思われる。ここで、両極端の方針として、

- 変数選択、あるいは、ネットワークの最適サイズの選択を行なう。
- 線形の罰金項の付加 (ここでいう "線形" は weight decay での 1 次式、エネルギーでの 2 次式を意味する。重回帰分析ではリッジ回帰に相当)。

があるが、より興味深いのはその中間、すなわち "やわらかな選択" とでもいうべき方法である。

やわらかな選択を実現するための方法としては次の 2 つが考えられる。

1. 非線形の罰金項の使用 (石黒 (1985) [1] , Weigend and Rumelhart (1991), Nowlan and Hinton (1992), Williams (1995), Ishikawa (1996))
2. モデル選択 (変数選択) の対象となるモデル空間の上の事後分布を考える[2]。(重回帰分析については、 Mitchell and Beauchamp (1988), George and McCulloch (1993), Kuo and Mallick (1996)、より一般的には、 Phillips and Smith (1996) を参照。)

なお、完全にベイズ的に考えた場合は、 1. と 2. はほぼ同じ土俵で論じられることに注意しておく。パラメータについての事前分布がパラメータが値 0(あるいは何らかの "デフォルト値") を取る場所で δ 分布的な重みを持つ場合が 2. に相当する。

1.2 重回帰分析と共線性

重回帰分析で特に興味深いのは、いわゆる共線性の効果である。説明変数の間に相関がある場合、回帰係数の推定値に大きく揺らぎが生じる。たとえば、説明変数の和がつねに 0 だとすると、各回帰係数を一定だけずらしても、回帰式のあてはまりは変わらないことになる。共線性はきわめて普通にみられる現象である[3]。

共線性のある場合に起こる現象は overfitting を押える方法によって異なる。リッジ回帰 (線形の罰金項) の場合には、罰金項をふくめたエネルギー関数が 2 次式であるから、 local optima は存在しない。これに対して、変数選択の場合には、最適化のアルゴリズムに依存する local optima が存在しうる。両者の中間である "やわらかい" 選択の場合にどうなるかは、各方法の特質を示すものとして興味深い。

共線性がある場合の local optima の存在は、与えられたデータに対して、与えられた認識の枠組の中で、複数の本質的にことなる解釈が存在するということに相当している。罰金項の強さや "温度" などのパラメータを変化させていくと、 local optima(あるいは準安定状態) は生成 / 消滅したり、分岐したりすると考えられる。データがしだいに増加した場合も同じである。この様子は、データの性質を理解する上で、また理論的研究の対象として興味深い。

* 統計数理研究所, 〒 106 港区 南麻布 4-6-7 The Institute of Statistical Mathematics, 4-6-7 Minami Azabu, Tokyo 106, Japan. E-mail: iba@ism.ac.jp TEL: 03-5421-8759 FAX: 03-3446-1695

[1] 以前に石黒の "準ガウス事前分布" が石川や Williams の Laplace prior と同じだと書いたことがあるが、これは誤りのようである。しかし、本質的なアイディア、すなわち、リッジ回帰と変数選択の "補間" をすること、及び、階層ベイズ的な枠組によって罰金項の強さと種類を制御することがこの短い論文にすべて含まれていることは注目に値する。

[2] 非ベイズ的な接近としては、ひとつのモデルの代わりにモデルの "信頼集合" を選ぶというのが、この立場にもっとも近いと思われる (下平 (1993))。

[3] たとえば、ある病気になる割合を食事の中の成分と関係づけようとした場合、食事の中の脂肪の量とタンパク質の量が無相関であるとは考えにくい。

2　この論文の内容

本論文では、上記の 2.(モデルの混合にもとづくアプローチ) において、事後分布の周辺分布を近似するために、平均場近似を適用することを提案し、簡単な適用例を示す。一般にこの種の問題では、

- どのような規準にもとづいて変数選択を行なうか
- 多くのモデルの中から最適なモデルを選ぶ、あるいは、良いモデルの集まり (モデル空間での事後分布) についての期待値を計算するにはどうするか

の両者が問題になるが、本論文では主に後者を論じる。

3　重回帰分析における変数選択

以下では重回帰分析における変数選択を、説明変数の存否を指定する離散変数を含んだ問題として定式化する。このような定式化自体は新しいものではない。

重回帰分析の尤度関数は、

$$-E_L = -\frac{1}{2\sigma^2}\sum_{\alpha}(y^{\alpha} - \sum_i a_i x_i^{\alpha})^2 \tag{1}$$

$$L(y|\{a_i\}) = \frac{1}{(2\pi\sigma^2)^{1/2}}\exp(-E_L) \tag{2}$$

と書ける。変数 i が含まれているかどうかを indicator $\widetilde{S}_i \in \{0,1\}$ であらわすと、部分モデルの尤度は、

$$-\widetilde{E}_L = -\frac{1}{2\sigma^2}\sum_{\alpha}(y^{\alpha} - \sum_i a_i x_i^{\alpha}\widetilde{S}_i)^2 \tag{3}$$

$$\widetilde{L}(y|\{a_i\},\{\widetilde{S}_i\}) = \frac{1}{(2\pi\sigma^2)^{N/2}}\exp(-\widetilde{E_L}) \tag{4}$$

となる (N はデータ数)。$\widetilde{S}_i^2 = \widetilde{S}_i$ であることを利用し、$S_i \in \{\pm 1\}$ を $\widetilde{S}_i = \frac{S_i+1}{2}$ で定義すると、(3) は次のように書ける。

$$-\widetilde{E_L} = \frac{1}{2\sigma^2}(C + \sum_i H_i S_i + \sum_i\sum_{j>i} J_{ij}S_iS_j) \tag{5}$$

$$C = -V_{y^2} + \sum_i a_i V_{x_i y} - \frac{1}{2}\sum_i\sum_{j>i} a_i a_j V_{x_i x_j} - \frac{1}{2}\sum_i a_i^2 V_{x_i x_i} \tag{6}$$

$$H_i = a_i V_{x_i y} - \frac{1}{2}\sum_{j\neq i} a_i a_j V_{x_i x_j} - \frac{1}{2}a_i^2 V_{x_i x_i} \tag{7}$$

$$J_{ij} = -\frac{1}{2}a_i a_j V_{x_i x_j} \tag{8}$$

ただし、十分統計量 $V_{y^2}, V_{x_i y}, V_{x_i x_j}$ を $V_{y^2} = \sum_\alpha (y^\alpha)^2$, $V_{x_i y} = \sum_\alpha (x_i^\alpha y^\alpha)^2$, $V_{x_i x_j} = \sum_\alpha (x_i^\alpha x_j^\alpha)^2$ のように定義した。

ベイズ的な扱いのためには事前分布を決める必要がある。モデルの indicator については、h を定数 (ハイパーパラメータ) として、

$$\pi(\{S_i\}) = \frac{\exp(-h\sum_i S_i)}{\cosh(h)} \tag{9}$$

とする。回帰係数 a_i については難しいが、とりあえず、一般的に

$$\pi_a(\{a_i\}) = \frac{\exp(\sum_i -E_a(a_i))}{Z_a} \tag{10}$$

とおこう (Z_a は分布の規格化定数)。すると、事後分布 $P_{pos}(\{a_i\},\{S_i\})$ は、"事後エネルギー"

$$-\widetilde{E}_{pos} = \frac{1}{2\sigma^2}(C + \sum_i H_i S_i + \sum_i\sum_{j>i} J_{ij}S_iS_j) - h\sum_i S_i - \sum_i E_a(a_i) \tag{11}$$

から定まるギブス分布になる。

$\{\widetilde{S}_i\}$ を与えたとき、事後分布 $P_{pos}(\{a_i\},\{S_i\})$ を最大にするような $a_i = a_i^*$ の値は $\widetilde{S}_i = 1$ のとき、正規方程式

$$V_{x_i x_i}a_i^* + \sum_j V_{x_i x_j}a_j^*\widetilde{S}_j = V_{x_i y} \tag{12}$$

から定まる。$\widetilde{S}_i = 0$ のときは不定である。

4　平均場近似

平均場近似の概念は統計物理では古くから用いられてきたが、統計的情報処理において話題になったのは比較的最近である。"ボルツマン・マシン" に関しては Peterson and Hartman (1989) があり、画像処理 に関しては Geiger and Girosi (1991) をはじめとして相当数の論文が発表されている[4]。また、Opper and Winther (preprint) はパーセプトロンのベイズ的取り扱いについて平均場近似を論じている。ただし、Opper らのいう "平均場近似" はこの論文でいうものとはかなり違っている[5]。ここで考える平均場近似は離散変数に関するものである。

平均場近似は組合わせ的最適化における Hopfield-Tank の方法 (Hopfield and Tank (1985)) と深い関係にある。実際、Hopfield-Tank の方法は、simulated annealing 法で有限温度の状態を動的モンテカルロ法で生成する部分を、平均場近似で置き換えたものにほかならない。ただし、Hopfield-Tank 法では local minima を避けるために温度を導入するのに対し、前の段落で述べた研究、ないし、それに続く研究では事後分布の周辺分布に対する近似という面が強調されてきており、その点で問題意識に違いがある。Hopfield-Tank の方法でモデル選択を行なう研究としては、たとえば、Sakai (1989) がある。

5　変数選択問題についての平均場近似

ここでは indicator $\widetilde{S}_i$ についての平均場近似を考える。すなわち、$\{a_i^*\}$ を与えたとき、S_i について和をとった

[4] 画像処理については、ほぼ同じ内容の近似が統計学者によってそれ以前に議論されている (Kay and Titterington (1986))。

[5] もともと "平均場近似" という用語は非常に広い意味で使われるので注意が必要である。

量

$$l_{a_i=a_i^*}(\sigma^2, h) = \log \sum_{\{S_i\}} \widetilde{L}(y|\{a_i^*\}, \{S_i\})\pi(\{S_i\}) \quad (13)$$

を次のように近似する (定数は省略)。

$$l_{MF}(\sigma^2, h) = -F_{MF} - \log\cosh(h) - \frac{N}{2}\log 2\sigma^2 \quad (14)$$

$$\begin{aligned} -F_{MF} = & \\ & \tfrac{1}{2\sigma^2}(C + \textstyle\sum_i H_i m_i + \sum_i \sum_{j>i} J_{ij} m_i m_j) - h \sum_i m_i \\ & - \textstyle\sum_i \frac{1+m_i}{2} \log \frac{1+m_i}{2} - \sum_i \frac{1-m_i}{2} \log \frac{1-m_i}{2} \end{aligned} \quad (15)$$

但しここで $E_a \equiv 0$ とした。$l_{MF}(\sigma^2, h)$ を最大にする値 $\{m_i^*\}$ は、

$$m_i^* = \tanh\left(\frac{\sum_{j\neq i} J_{ij} m_j^* + H_i}{2\sigma^2} - h\right) \quad (16)$$

の解であるから、適当な定数 $0 < \delta \leq 1$ について、

$$m_i^{t+1} := (1-\delta)\cdot m_i^t + \delta \cdot \tanh\left(\frac{\sum_{j\neq i} J_{ij} m_j^t + H_i}{2\sigma^2} - h\right) \quad (17)$$

を iterate することで、(local optima が) 計算できる。

逆に $\{m_i^*\}$ を与えれば、(14) を最大にする $\{a_i^*\}$ を求めることができる。これは、(12) 式の代わりに

$$V_{x_i x_i} a_i^* + \sum_j V_{x_i x_j} a_j^* p_j = V_{x_i y} \quad (18)$$

を解くことに相当する。但し、p_i は項 i が選択される事後確率の近似値 $p_i = (1+m_i)/2$ である。そこで、$\{m_i^*\}$ に適当な初期値を与えて、

> $\{m_i^*\}$ から $\{a_i^*\}$ を求め、$\{a_i^*\}$ から $\{m_i^*\}$ を求め、収束するまで繰り返す

ことで両者を同時に求めることが考えられる。

実際の計算では σ^2 もデータから推定することが望ましい。すぐあとに示す数値例では、$\{a_i^*\}$ 及び $\{m_i^*\}$ の推定と並行して (14) を最大にする σ^2 を代入することで、σ^2 を推定している。h の推定につては本稿の最後で論じる。

(16) 式は、S_i の事後分布での期待値 m_i に関する自己無矛盾 (self-consistent) 方程式とみなせる。本来 $S_j (j \neq i)$ の取る値の確率分布は S_i の値に依存しており、またお互いに独立ではない。この効果 (期待値のまわりの揺らぎの相関の効果) を無視して、期待値の間の関係式で話をすませてしまうのが平均場近似の特徴である[6]。(18) 式を (12) 式と比較すると、a_i^* に対する他の変数 $a_j^* (j \neq i)$ の存否の影響は $\widetilde{S}_j$ の期待値の近似値 p_j を通して取り込まれているが、実際は変数 a_i が存在するかどうかによって $\widetilde{S}_j$ の期待値は変化するのに、その影響は無視されている。

われわれの近似は a_i を "点推定" a_i^* で置き換えているために、通常の平均場近似にくらべてさらに粗い近似になっている。a_i^* は "i 番目の変数 a_i が存在するという条件のもとでの a_i の推定値" と解釈するのが最も妥当ではないかと思われるが、その揺らぎは無視されている (a_i^* は "存在しさえすれば" つねに一定の値をとると仮定されている)。

変数の間に強い共線性がある場合には "変数 a_i と a_j はどちらも同じくらい重要であるが、変数 a_i が選択されるなら、変数 a_j は選択されにくい。逆に、変数 a_j が選択されるなら、変数 a_i は選択されにくい。" といった排反的な状況が出現しうる。このような傾向が強いと、平均場近似は破綻する。この場合、近似の破綻は local optima の存在として現れるだろう。逆に、"ある変数 a_i が存在するかどうかは微妙だが、その判断はほかの変数によらない" といった場合は、local optima は生ぜずに、むしろ、$p_i = (1+m_i)/2$ が 0 と 1 の間の値をとることが予想される。

6 例

ここでは、提案した方法を、Belsley et al. (1980) の p.244 にある、ボストンの住宅に関するデータ (データ数 506, 説明変数の候補 13 個) に適用する。このデータは、下平 (1993) の 6.3 節で全く違う観点から解析されている。いくつかの h についての結果を図 1 に、h による変化を図 2 に、それぞれ示した。

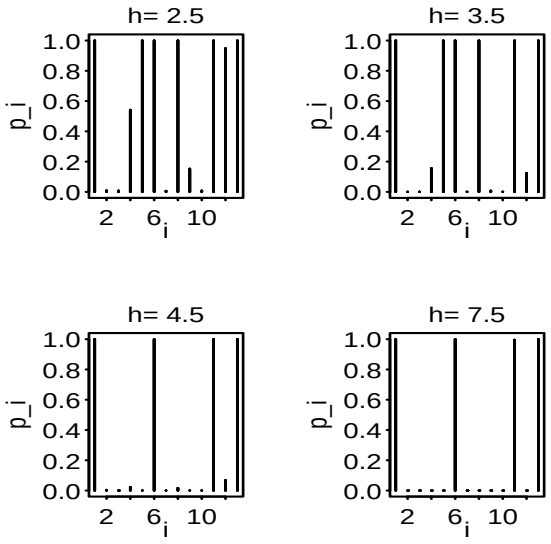

図 1: 横軸は変数の番号 (1-13)、縦軸はその変数が選択される事後確率 $p_i = (1+m_i)/2$ を示す。

[6] Gibbs sampling のような動的なモンテカルロ法では確率的な繰り返し操作によって、この揺らぎを (理論的には完全に) 取り入れている。一方、S_i について最適化した場合は S_i の揺らぎ自体を無視しているので、m_i のような、± 1 のどちらでもない半端な値は出てこない。平均場近似はある意味では両者の中間に相当するわけである。

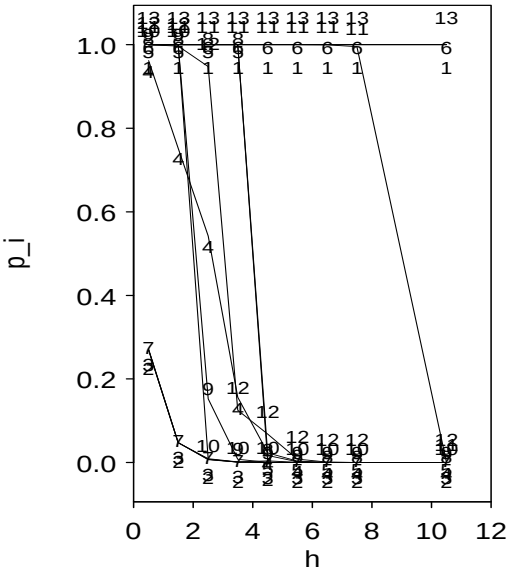

図 2: 横軸は h、番号は変数の番号 (1-13)、縦軸はその変数が選択される事後確率 $p_i = (1 + m_i)/2$ を示す。番号は見やすいように上下にずらしてある。番号をつなぐ線が正しい位置を示す。

図 1, 図 2 を描いたときには、このデータでは local optima は無いと考えていたが、投稿前のチェックで、(比較的小さい確率で) 図とは違った解に収束することがあることがわかった。ただし、この local optima は必ずしも共線性から予想されるような性格のものではなく、むしろ、σ^2 の推定法に関連して生じている可能性がある。このあたりの事情の詳しい考察、及び、複雑で強い共線性を持った模擬データを用いた数値実験の結果については、口頭発表に譲ることにする。

7 問題点と今後の課題

次の段階で問題になるのは、事前分布に含まれるハイパーパラメータ h をどのように決めるかということである。階層ベイズ法の枠組 (たとえば、Akaike (1980), MacKay (1992)) で考えると (14) 式を最大にするような h を選ぶことが考えられる。しかし、本稿の扱いでは、S_i についての和は近似的に計算されているが、a_i は点推定 a_i^* で置き換えられている。このため、h の異なる場合の比較に関しては、大きなバイアスが生じることが予想される。ひとつの解決策は、赤池の "擬ベイズ法" を用いることである。"擬ベイズ法" では (14) 式を尤度と考えたときに、AIC に相当するような補正を加えたものを用いて、h を決定する (詳しくは伊庭 (1991) とその文献を参照)。いまの場合、具体的には、(14) 式の $\cosh(h)$ の部分はそのままにして、F_{MF} の部分に含まれる h を $h + 1/2$ で置き換えることに相当する。

別の扱い方は、a_i を点推定で置き換える代わりに a_i について先に積分することである。重回帰分析の場合、エネルギーが a_i について 2 次式なので、これは比較的容易に可能である。この操作は S_i の間に複雑な多体相互作用を持ち込むことになるが、この場合でも平均場近似が有効かどうかは技術的に興味深い問題である。ただし、この方法をとった場合、より本質的な問題として、a_i についての事前分布をどう定めるか (E_a をどう定めるか) という問題が生じる ($E_a \equiv 0$ とすると、$\widetilde{S}_i = 0$ のときに a_i に関する積分が発散してしまう)。ひとつの考え方は、事前分布の裾の長さを階層ベイズ的な考えを用いて定めることであろう。

有益な討論をして下さった樺島祥介氏に感謝します。

Akaike,H.(1980) in Bayesian Statistics, Eds. Bernardo,J.M, DeGroot,M.H., Lindley,D.V., and Smith,A.F.M., University press, Valencia.

Belsley,D.A., Kuh,E. and Welsch,R.E. (1980) Regression Diagnostics, Wiley, New York.

Geiger,D. and Girosi,F.(1991) IEEE Transactions on Pattern Analysis and Machine Intelligence 13 401-412.

George,E.I. and McCulloch,R.E. (1993) Journal of the American Statistical Association 88 881-889.

Hopfield,J.J. and Tank,D.W.(1985) Biological Cybernetics 52 141-152.

伊庭幸人 **(1991)** 統計数理 39 No.2 225-244.

石黒真木夫 **(1985)** 統計数理 33 No.1, 特集 – エントロピー最大化原理の展開と統計モデル –, pp.8-11.

Ishikawa,M. (1996) Neural Networks, Vol.9, 509-521.

Kay.J.W. and Titterington D.M. (1986) Proc. 2nd. Int. Conf. Image processing and Applications Conf. Publ. No.265 London Inst. Elec. Engrs. 44-48.

Kuo,L and Mallick,B. (1996) preprint.

MacKay, D.J.C. (1992) Neural Computation 4 448-472.

Mitchell,J. and Beauchamp,J.J. (1988) Journal of the American Statistical Association 83, 1023-1036.

Nowlan,S.J. and Hinton,G.E. (1992) Neural Comutation Vol.4 No.4 pp.473-493.

Opper and Winther preprint.

Peterson,C. and Hartman,E. (1989) Neural Networks 2 475-494.

Phillips,D.B. and Smith,A.F.M. (1996) Chap.13 in Markov Chain Monte Carlo in Practice, Eds. Gilks,W.R., Richardson,S. and Spiegelhalter,D.J., Chapman and Hall.

Sakai,H. (1993) ISM Cooperative Research Report 20.

下平英寿 **(1993)** 統計数理, 41, No.2, 131-147.

Weigend,A.S. and Rumelhart,D.E. (1991) in Computing Science and Statistics: Proc. of the 23rd symposium on the interface(Ed. E.M.Keramides) pp.156-163 Interface Foundation, Fairfax Station, Va.

Williams,P.M.(1995) Neural Computation, Vol.7, No.1 117-143.

表題を含んだ文献目録は
`http://www.ism.ac.jp/~iba/mf.html`
で提供する予定です。そのほか、関連のある情報や訂正もそこに掲載する可能性があります。ただし、これは個人的な試みであって、原著論文としての本論文あるいは SITA96 と公式な関係を持つものではありません。