一般/特殊ジレンマ，規則/例外ジレンマ，一貫/非一貫ジレンマ
Dilemma between Generalization and Specialization, Regularity and Irregularity, and, Consistancy and Inconsistency

浅川伸一
Shin Asakawa

東京女子大学
Tokyo Woman's Christian University
asakawa@ieee.org

Abstract

Human intelligence seems to process external environment adequately. In other words, human beings can deal with the dilemma between regularity and specialty. In the area in machine learning, there exist many algorithm proposed so far, in order to deal with data complexity. Here, we tried to model this kind of optimization problem in order both to learn generalization and special events in life. Contribution of emotional system, such as Amygdala in limbic system, was implemented as coefficients corresponded to each life event. It could also be regarded as regularization parameters in Ridge regression. Although these parameters might introduce ill-posed problem, it would be worth considering as the first step to an extension toward a human valuable optimization or decision processes.

Keywords — Regularization, Optimization, Overfitting, Parameter Estimation, Decision Problem

1. 問題

機械学習であれ，人間における学習であれ，一般の学習において，一般解と特殊解，規則と例外事例，一貫性と非一貫性，どのように呼んでも構わないが与えられた事例から規則を抽出することと，規則に従わない例外を処理することを適切に按配しないと環境に対して適切に振る舞うことはできない。知的な学習システムを構築する場合，あるいは，人間を知的な学習機械と見なす場合，一般則とそれに当てはまらない例外側とをどのように処理させるべきなのか，あるいは，人間はどのように，規則/例外のトレードオフを処理しているのかを問うのは，興味深いテーマである。そして，この古典的な問題に立ち返ってみることは，知的学習機械の構築，あるいは，知的情報処理機械としての人間の性質を考える上で価値のあることと考える。

話を拡張して考えれば，科学における法則の一般性と特殊性との関連も視野に入れて考えることができる。科学法則においては，適用範囲の広い一般性を有する法則を発見することが目指される。しかし，抽象化された一般的な法則は，その法則の運用まで視野に入れるとなると，特殊解の方が実用的である場合も存在する。例えば，ケプラーによる惑星の運動法則は，ニュートンの運動法則において，太陽の質量に比べて惑星の質量が十分小さいとみなせる場合に相当する。この意味でニュートンの運動法則は一般性が高い。しかし，船乗りが星々の動きから自船の位置を知る方法としては，ケプラーの法則は実用的であり，ニュートンの運動法則を知らなくても十分である。このように，規則，法則を適用する場面において，一般解と特殊解との関係は実用的な側面まで含めて考えた場合，取捨選択のジレンマが存在する。科学法則一般のような大きな話でなく，統計的学習，機械学習の文脈においても，規則を抽出することと，規則に当てはまらない例外を処理することとの間に，トレードオフが存在する。人間であれ，機械であれ，知的に振る舞うシステムにおいては，与えられた事例から規則を抽出すること，抽出した規則に従わない例外を適切に処理する機構を備えている必要がある。

機械学習の文脈では，データの持つ複雑さの問題から過学習を防止し，かつデータを良く再現する手法が種々提案されている[2, 参照]。一般に過学習を防止するために，訓練データとテストデータを分離し，テスト誤差を最小にする手法が採用されることがある。頻度主義の立場からはバイアス–バリアンス分解によって得られる期待損失を最小化する値となるが，一回しか観測データが得られない状況，従って人間のモデルとしては日々観察される人間の認知モデルとして考えるのであれば，未知なる将来観測するかもしれない汎化誤差を如何にして最小化するのかというジレンマとなる。このように考えれば，予め知ることができな汎化誤差を想定するのは適切ではないと考える。

従って，データの複雑性に基いてパラメータ推定，モデル選択，およびアルゴリズム選択を行う場合，統計的意思決定理論では説明しきれない人

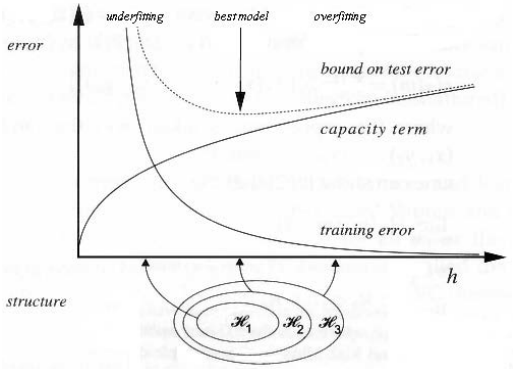

図 1 Structual Risk Minimization:http://www.svms.org/srm/

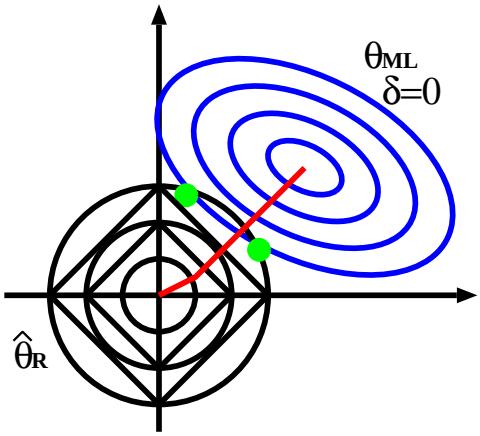

図 2 正則化項と最尤推定量

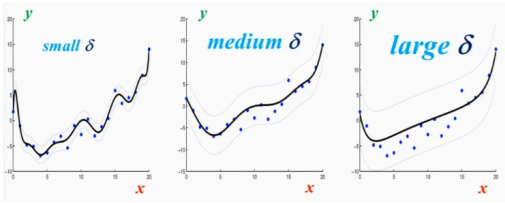

図 3 正則化項の効果

間独自の基準が存在するものと考える。既存の知識と現在の環境が与えられた時にオンラインで意思決定を行うことを想定すればベイズ流の意思決定モデルは有力な候補となると考えられる。さらに，結婚（あるいは離別），入学等，人生に一度きりの体験が忘れがたい重要な決定因となりうることを考慮したい。人生において重大な事件は情動を伴うので，扁桃体などの情動系の関与を仮定して，モデルを構成することにすると機械学習を拡張して最適化問題を定義することを考える。

2. 正則化

機械学習や統計学において，媒介変数を減らすのではなく，誤差関数に正則化項を追加して，モデルの複雑度や自由度に抑制を加え，過学習を防ぐ方法がある。

データ D が与えられたときの経験損失 $L_{\mathrm{emp}}(f, D)$ だけを最適化する関数 f を求めても，過適合などのため汎化誤差は最小にならない。こうした不良設定 (ill-posed) 問題な場合，平準化などを行う罰則項 $p(f)$ を持ちいて

- L1 ノルム $\|\theta\|$ の最小化 LASSO
- L2 ノルムの 2 乗 $\|\theta\|^2$ を最小化 Tikhonov 正則化 (Tikhonov Regularization) する。あるいはリッジ回帰 (Ridge Regression)

入力データを $\boldsymbol{X}$，出力を $\boldsymbol{y}$，推定すべきパラメータを $\boldsymbol{\theta}$ とする。このとき線形モデルでは，

$$\hat{\boldsymbol{\theta}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}, \tag{1}$$

なる形式をとる。このとき，$\boldsymbol{X}^T\boldsymbol{X}$ の逆行列が求められない場合が多い。このとき，対角行列に δ を導入し，

$$\hat{\boldsymbol{\theta}} = \left(\boldsymbol{X}^T\boldsymbol{X} + \delta^2\boldsymbol{I}_d\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}. \tag{2}$$

上式はリッジ回帰に一致する。この時目標関数を次式することに等しい。

$$J(\boldsymbol{\theta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}) + \delta^2\boldsymbol{\theta}^T\boldsymbol{\theta}. \tag{3}$$

媒介変数が十分に多ければ，目標関数を任意の精度で近似することが可能となるが，このことはモデルの複雑度を増し，自由度を高めるがゆえに過学習に陥る。さらに，与えられた事例だけに特化して学習させると，未学習の事例において対処できなくなる過学習が起きる。逆に，学習が不十分であれば事例に対して正解できなくなる。統計モデルのおいては，モデルの対数尤度と自由パラメータ数とを用いて AIC [1] などを使って最適なモデルを定めるような手法が考案されてきた。ベイズ流の推論も可能である [2]。

$$\mathrm{AIC} = -2\ln L(\theta) + 2p \tag{4}$$

$$\mathrm{BIC} = -2\ln L(\theta) + p\ln(p) \tag{5}$$

$$\mathrm{MDL} = -\ln L(\theta) + \frac{p}{2}\ln(p) \tag{6}$$

$$\mathrm{NIC} = D(p(\theta)) + \frac{1}{t}\mathrm{tr}\left(G(\theta)\,Q(\theta)^{-1}\right) \tag{7}$$

サポートベクターマシンにおいては，媒介変数（パラメータ）を減らすのではなく，マージンを最大化することにより，過学習を防いでいて，これも，Tikhonov 正則化と同じような手法に基づいている。

Deep learning においては dropout [5] によってユニットを間引くことによって，汎化を改善する努力がなされる。しかし，これは RBMが微分不能で

あるための苦肉の策と言える。微分可能な出力関数を持つバックプロパゲーション[12]においては，種々の枝刈り法[10, 4]が使えたが，バイナリユニットを仮定するディープラーニングにおいてはこのような従来から手法を活用することができない。

微分可能な出力関数を持つバックプロパゲーション[12]においては，種々の枝刈り法[10, 4]が考案されてきた。一方で，ディープラーニングにおいてはdropout[5]によってユニットを間引くことによって，汎化を改善する努力がなされる。これはRBMが微分不能であるための苦肉の策と言える。が，バイナリユニットを仮定するディープラーニングにおいてはこのような従来から手法を活用することができない。

ベイズ的な枠組みで尤度も事前分布もガウシアンで与えらている場合 $\mathcal{N}\left(y\middle|\boldsymbol{X\theta}, \sigma^2\boldsymbol{I}_n\right)$

$$p\left(\theta\middle|\boldsymbol{X}, \boldsymbol{y}, \sigma^2\right) \propto \mathcal{N}\left(\boldsymbol{\theta}|\boldsymbol{\theta}_0, \boldsymbol{V}_0\right)\ \mathcal{N}\left(y\middle|\boldsymbol{X\theta}, \sigma^2\boldsymbol{I}_n\right), \quad (8)$$

であるから，

$$\boldsymbol{\theta}_n = \boldsymbol{V}_n\boldsymbol{V}_0^{-1}\boldsymbol{\theta}_0 - \frac{1}{\sigma^2}\boldsymbol{V}_n\boldsymbol{X}^T\boldsymbol{y}_n \quad (9)$$

$$\boldsymbol{V}_n^{-1} = \boldsymbol{V}_0^{-1} - \frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{X} \quad (10)$$

を得る。ここで初期値を $\boldsymbol{\theta}_0 = \boldsymbol{0}$, かつ $\boldsymbol{V}_0 = \tau_0^2\boldsymbol{I}_d$ とすれば事後確率は，

$$\boldsymbol{\theta}_n = \left(\lambda\boldsymbol{I}_d + \boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}_n \quad (11)$$

$$\boldsymbol{V}_n^{-1} = \left(\lambda\boldsymbol{I}_d + \boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\sigma^2, \quad (12)$$

となる。ここで，$\lambda = \frac{\sigma^2}{\tau_0^2}$ である。先見知識が完全無知であれば $\lim_{\tau\to\infty}\lambda = 0$ とみなしうるので，ベイズ回帰においてはリッジ回帰を取り込むことができる[11]。

3. 例外の取り扱い

上記のどのような方法を用いるのであれ，例外事例が極端に少なくなると，システムのパフォーマンスは劣化する。それは例外事例を例外事例として正しく認識し，例外扱いする機構が存在しないからである。学習機械は全学習データを均等に扱おうとするので，例外事例に対する資源が相対的に減少し最終的には例外を正しく処理できなくなってしまう。人間の学習においてはこのようなことが起こらないのであるから，何らかの別の機構が存在すると考えざるを得ない。

例外事例と一般則とを正しく分割し，処理させる機構として混合エキスパート[15, 7, 6, 8, 9, 16]を考える（図4）。混合エキスパートによって，問題空間を分割し，分割された小領域の中では例外的な規則を適用する機構を用意する。概念的なイメージを図5に示した。

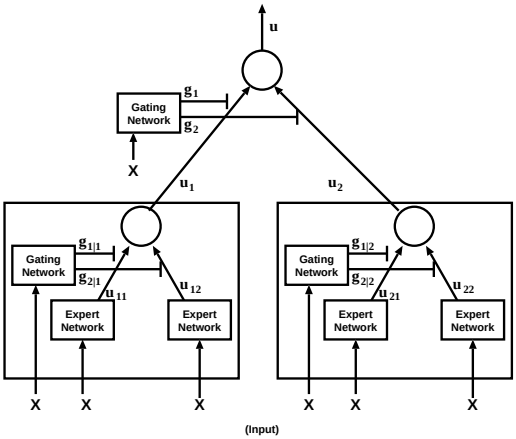

図4 混合エキスパートモデル

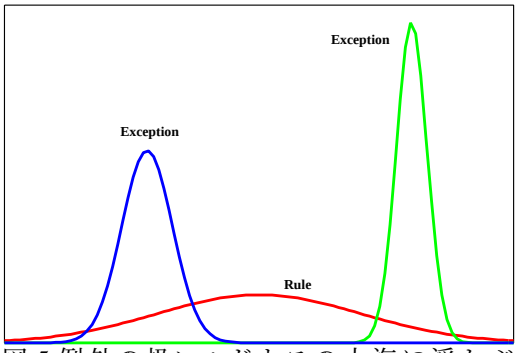

図5 例外の扱い：ガウスの大海に浮かぶディラックの孤島

3.1 人間におけるヒューリステック

[3]がカスケードコリレーションを導入した歳に用いた問題に二重らせん問題がある(図6上)。この図では，2次元上の2つの螺旋からOとXとを判別しなければならない。2値分類課題と捉えた場合，OとXとは線形分離が難しい問題となる。従って学習機械は，この問題を解く場合に困難を生じる場合が多い。ところが，人間がこの問題を解く場合には，OとXとが規則的に配置されていることを利用して図6下のような3次元空間に射影し，XY平面で区切れば容易に線形分離可能である。古くは，ケーラーの洞察学習につながるこのような発見的思考を取り込みたい。

4. モデル

以上のような考察から，課題の持つ複雑さと状況とを同時に考慮し，適応的にパラメータを調節する機構を考えなければならない。そして，パラメータ空間の探索にジレンマが生じた場合，学習を破棄し，一気にヒューリティックなアプローチを取るように仕向ける機構までを実装したい。図6下のような解を見出すことは難しいにしても，

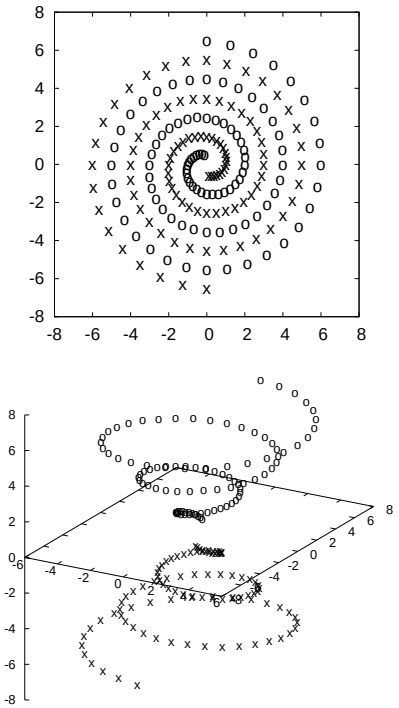

図 6 Two spirals

現在のアプローチを諦めるという判断までは，できる知的なエージェントの作成を目指す。一つのトリヴィアルなモデルとしては，解くべきデータ集合と文脈情報とを学習機械に与え，例えば正則化パラメータ λ を自動調節させるようなモデルを考える。一つには，効用関数を事例ごとに割り当てることによって期待効用を最大化するというような機構を考えることができるかも知れない。効用は報酬に基づくと考えれば，脳内の神経伝達物質であるドーパミンが報酬に関与している可能性は指摘されてきている。D1，D2 を促進，抑制系のドーパミン下位物質だとすれば D1, D2 を制御することによって効用を調整し，そのことによって分類，回帰システムを柔軟に運用する系を考えることができよう。ここにで $D1/D2$ を改めて λ ととらえ直すことにより，パラメータ過剰により不良設定問題に逆戻りすことになる。$\boldsymbol{\theta}^T = (1, 1, \ldots, 1)$ であれば解が求まる可能性があるが多くの場合局所最小に陥る。データがモデルに提示された毎に，オンライン学習で $\boldsymbol{\theta}$ を $\boldsymbol{\theta} + \boldsymbol{\sigma}$ で置き換えることを考える。σ を決めるために，線形近似により，

$$f(\boldsymbol{x}_i, \boldsymbol{\theta} + \boldsymbol{\delta}) \sim f(\boldsymbol{x}_i, \boldsymbol{\theta}) + J_i \boldsymbol{\delta}, \tag{13}$$

ここで，

$$J_i(\boldsymbol{\theta}) = \frac{\partial f(\boldsymbol{x}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}, \tag{14}$$

は，$\boldsymbol{\theta}$ の勾配ベクトルである。

近似誤差 $S(\boldsymbol{\theta})$ は，

$$S(\boldsymbol{\theta} + \boldsymbol{\delta}) \sim \|\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{\theta}) - \boldsymbol{J}\boldsymbol{\delta}\|^2, \tag{15}$$

であり，これを $\boldsymbol{\sigma}$ で結果を 0 とおけば，

$$\left(\boldsymbol{J}^T \boldsymbol{J}\right)\boldsymbol{\delta} = \boldsymbol{J}^T \left[\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{\theta})\right], \tag{16}$$

ここで $\boldsymbol{J}$ is ヤコビアンである。Levenberg によればこの式は以下ように，

$$\left(\boldsymbol{J}^T \boldsymbol{J} + \lambda \boldsymbol{I}\right)\boldsymbol{\delta} = \boldsymbol{J}^T \left[\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{\theta}))\right], \tag{17}$$

書くことができる。ここで $\boldsymbol{I}$ は単位行例つであり，$\boldsymbol{\delta}$ はパラメータベクトル $\boldsymbol{\theta}$ の推定値となる。ここで $\boldsymbol{I}$ は単位行列であり，$\boldsymbol{\delta}$ はパラメータベクトル $\boldsymbol{\theta}.$ の推定値である。このとき $\boldsymbol{J}^T\boldsymbol{J} + \lambda\boldsymbol{I}$ の逆行列が求まるか否かだが，Marquardt によれば $\boldsymbol{I}$ を $\boldsymbol{J}^T\boldsymbol{J}$ の対角行列で置き換えて，

$$\left(\boldsymbol{J}^T \boldsymbol{J} + \lambda \operatorname{diag}\left(\boldsymbol{J}^T \boldsymbol{J}\right)\right). \tag{18}$$

とすれば解が求まる可能性がある。

5. まとめと考察

例外事例を特別扱いして処理する機構を仮定した。人間が知的であることの一つの理由は，そのような規則と例外とを柔軟に処理する能力にあるのだろう。そのためには最適化関数として何を設定すればよいのかという問題設定が重要になってくる。このように考えれば，正則化は ill-posed な問題を解くための制約条件を与えてはくれるが，それだけでは人間の知的振る舞いを記述できてるとは言い難い。更に付加条件を加えて適応的な最適化システムとして捉え直す必要があるのだと考える。かつて Tversky と Kahneman は人間が以下に不合理な判断をするのかを示したが，見方を変えれば，脳内の報酬系からもたらされる効用を含めて最適化するシステムとして捉え直すことが可能になるかも知れない。

引用文献

[1]Hirotsugu Akaike. A new look at the statistical model identification. *IEEE Transaction of Autom. Control*, AC-19:716–723, 1974.

[2]Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer–Verlag, New York, 2006.

[3]Scott E. Fahlman and Christian Lebiere. The cascade-correlation learning architecture. In D.S. Touretzky, editor, *Advances in Neural Information Processing Systems*, volume 2, pages 524–532. Morgan-Kaufman, 1990.

[4]B. Hassibi, D. G. Stork, and G. Wolff. Optimal brain surgeon. In S. J. Hanson, J. D. Cowan, and C. L. Giles, editors, *Advances in Neural Information Processing Systems (Denver)*, volume 5, pages 164–171, San Mateo, 1993. Morgan Kaufmann.

[5]Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *The Computing Research Repository (CoRR)*, abs/1207.0580, 2012.

[6]Robert A. Jacobs and Michael I. Jordan. A competitive modular connectionist architecture. In *Advances in Neural Information Processing Systems*, volume 3, pages 767–773. Morgan-Kaufmann, San Fransisco, CA, USA, 1991.

[7]Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3:79–87, 1991.

[8]Michael I. Jordan and Robert A. Jacobs. Hierarchies of adaptive experts. In Richard Lippmann John Moody, Steven Hanson, editor, *Advances in Neural Information Processing Systems*, volume 4, pages 985–992. Morgan-Kaufmann, San Fransisco, CA, 1992.

[9]Michael I. Jordan and Robert A. Jacobs. Hierarchical mixtures of experts and the em algorithm. *Neural Computation*, 6:181–214, 1994.

[10]Yann LeCun, John S. Denker, and Sara A. Solla. Optimal brain damage. In D. S. Touretzky, editor, *Advances in Neural Information Processing Systems*, pages 589–605, Denver, WS, USA, 1990. Morgan Kaufmann.

[11]Kevin Patrick Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, Cambridge, MA, UK, 2012.

[12]David E. Rumelhart, Geoffery E. Hinton, and Ronald J. Williams. Learning internal representations by error propagation. In David E. Rumelhart and James L. McClelland, editors, *Parallel Distributed Porcessing: Explorations in the Microstructures of Cognition*, volume 1, chapter 8, pages 318–362. MIT Press, Cambridge, MA, USA, 1986.

[13]Albert Tarantola. *Inverse Problem Theory*. the Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2005.

[14]Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, Series B, 58(1):267–288, 1996.

[15]S. R. Waterhouse and A. J. Robinson. Classification using hierarchical mixtures of experts. In *Proceedings of the 1994 IEEE Workshop on Neural Networks for Signal Processing IV*, pages 177–186, Long Beach, CA, USA, 1994. IEEE Press.

[16]Seniha Esen Yuksel, Joseph N. Wilson, and Paul D. Gader. Twenty years of mixture of experts. *IEEE Transactions on Neural Networks and learning Systems*, 23(8):1177–1193, 2012.