■S3 群（脳・知能・人間）- 4 編（ソフトコンピューティングとニューラルネットワーク）

1 章 ニューラルネットワーク

（執筆者：石井　信）［2010 年 11 月 受領］

■概要■

ヒトの脳は 100 億以上ともいわれる数の神経細胞からなる回路であり，そこでの情報処理が人間の知的活動の源である．神経細胞の応答はしばしば非線形であるため，脳は全体として非線形回路である．また，神経細胞及び回路は外部からの刺激と自己の応答に応じて変化，すなわち，「学習」する．ニューラルネットワーク研究とは，神経細胞及び回路の非線形応答を模した人工的な情報処理器と，その学習方式について，性質を理論的に究明し，工学的に応用しようとするものである．歴史的には，サイバネティクス分野から興り，人工知能，統計物理，数学（統計科学），ロボティクスなど多くの分野を巻き込んでいった融合領域分野である．今や，ニューラルネットワーク自体を研究する研究者は少なくなったが，機械学習と計算神経科学など新しい融合領域研究の母体になった．こうしたニューラルネットワーク研究の歴史については，1-1 節を参照されたい．

本章では，しばしばニューラルネットワークのアーキテクチャを用いて説明されることの多い機械学習の手法について紹介する．神経細胞に相当する回路素子が閉ループを許すように相互に結合したアーキテクチャを相互結合型という（1-2 節）．こうした回路の性質はスピン系とのアナロジーから統計物理の手法を用いて研究されてきた．近年，更に情報理論と融合して情報統計力学という新しい潮流に成長した（1-3 節）．一方で，多くの場合において閉ループを許さず，素子をフィードフォワード型に並べるアーキテクチャを階層型といい，多層パーセプトロンはその中で最も一般的なニューラルネットワークである（1-4 節）．階層型回路はしばしば回路の最下層の細胞への入力から最上層の細胞における出力との入出力関係を表現でき，この入出力関係の学習を関数近似という（1-5 節）．また，パーセプトロンの線形結合系（分類や回帰に用いる）の学習にスパース性とカーネルいう概念を導入したのが，サポートベクトルマシンである（1-6 節）．多層パーセプトロンの学習の挙動を解析するのに，微分幾何と代数幾何の手法が使えることが分かり，そこから発展して，一般の機械学習器の性能に対する理解が進んだ（1-7 節）．多層パーセプトロンやサポートベクトルマシンは教師あり学習（関数近似やパターン認識）に用いられるものであるが，階層型回路であって教師なし学習に用いられるニューラルネットワークのアーキテクチャとして，自己組織化写像がある（1-8 節）．また，古くからデータ解析に用いられてきた主成分分析や因子分析が，線形確率モデルの推定（教師なし学習）にスパース性を導入したものとして定式化できることが発見された．それらの手法を一般化して低ランク行列因子化という（1-9 節）．一方で，単一のニューラルネットワークによるよりも，複数のネットワークを用いる方がしばしば性能が良いということが分かった．それをアンサンブル法と総称する（1-10 節）．このアンサンブル法とベイズ推定とは関係が深いが，推定変数を階層化しベイズ法により推定することで予測精度の高い学習器が構成できることが分かり，自然言語処理などのデータマイニング分野で研究が進められている（1-11 節）．

ニューラルネットワークのアーキテクチャに基づいた機械学習の分野は現在も急激な発展を続けている．本章の内容は 2010 年春の時点で最新のものであるが，本章の読者は，本分野が今も研究の最前線にあり，最新の知識は最新の論文によってもフォローされるべきであ

ることに注意されたい．

【本章の構成】

まずニューラルネットワークモデルの過去と現在（1-1 節）について紹介する．続いて，相互結合型ニューラルネットワーク（1-2 節），情報統計力学の方法（1-3 節），パーセプトロンと多層パーセプトロン（1-4 節），関数近似法（1-5 節），サポートベクトルマシン（1-6 節），微分幾何・代数幾何の方法（1-7 節），自己組織化写像（1-8 節），低ランク行列因子化（1-9 節），アンサンブル法（1-10 節），階層ベイズモデリング（1-11 節）に関して，ニューラルネットワークに根ざしたモデルと理論，応用について解説する．

■S3 群 - 4 編 - 1 章

1-1 ニューラルネットワークモデルの過去と現在

（執筆者：甘利俊一）[2008 年 6 月 受領]

1-1-1 はじめに

脳は精妙にして極めて複雑な情報処理器官である．これは多数のニューロンをつないだ回路網，ニューラルネットワークからなるシステムである．そこで知的な情報処理が営まれることから，ニューラルネットワークは工学の研究対象として注目された．脳の仕組みに学び，これをヒントに高度の情報システムを構築したいという願いと，逆に工学の手法を用いて脳の情報処理の原理を解明したいという願いが融合して，ニューラルネットワークと呼ぶ分野を作った．

コンピュータの基本原理は論理計算であり，その基礎は計算可能性，アルゴリズム，データベースなどの理論で明らかにされている．脳は，これとは違った原理に基づいて知的な情報を処理する．脳の仕組みに学ぶものは，一つは並列計算である．脳は，多くの部位で情報を同時並列に処理している．これによって，高度な知的計算がどのように実現できるのかが興味の対象である．

もう一つは，学習と自己組織化の能力である．脳は優れた記憶と学習の能力をもっている．更に，外界の構造を自動的に学び，外界のモデルを自己の内部に作り上げる．これを自己組織化という．並列計算の原理，学習と自己組織化の原理を求めて，ニューラルネットワークの研究が始まったといってもよい．

1-1-2 ニューラルネットワーク研究小史

脳の仕組みを理論的に研究する試みはずっと以前にさかのぼる．1943 年に，McCullock と Pitts は，ニューロンのモデルを提案した．これは万能な論理素子であり，これと遅延とを組み合わせれば，チューリング機械が構成できる．これにより，脳に対する興味が工学の世界に呼び起こされた．これは，オートマトン理論の源流となった．

1950 年代には，学習認識機械パーセプトロンが Rosenblatt により提唱され，一世を風靡した．また，自己組織化やホメオシュタット，神経の場における興奮パターンの伝播と自己再生などが研究されていた．しかし，こうした研究の勢いは，1970 年代に入って熱気がさめる．脳の秘密はなかなか解き明かすことはできないし，工学技術としてこれを実現するのは難しい．他方，コンピュータの急速な性能向上とあいまって，脳の仕組みに頼らない人工知能研究が急浮上したことにもよる．

1970 年代は，ニューラルネットワークの冬の時代と呼ばれる．多くの研究者がこの分野から去り，コンピュータサイエンスと人工知能研究が本格化したからである．しかし，その中で全世界で少数の研究者によって，しっかりとした体系的な研究が積み重ねられていった．それは，アメリカ，日本，ドイツ，イギリス，フィンランド，ロシアなどにまたがる．

1980 年代に入って，様相が一変した．脳の研究が再び脚光を浴びる．一つは，これまで人工知能と協調していた認知科学分野から，人の認知機能の解明には人工知能による論理計算だけではなくて，並列のダイナミックスに着目することが必要であるとして，コネクショニズムが提唱される[1)]．これに，物理学のスピングラス研究者が加わり，更に脳のモデル研究者も中心に躍り出て，いわゆるニューロブームを迎える．半導体研究者もこれに加わって，

ニューロチップの研究が盛んになった．

しかし，機はいまだに熟していなかった．こうしたブームは 10 年ほどで終焉する．しかし，その残したものは大きい．一つは人工知能分野において，論理的な研究，確率推論，並列学習処理などが融合し，視野を大きく拡大したことである．大規模なデータをもとに，データマイニングとも融合する機械学習と呼ぶ新しい分野も登場した．

逆に，脳科学の分野でも，計算論的神経科学と呼ぶ理論脳科学が定着した[2)3)]．複雑な脳の仕組み，特にその情報処理原理を探求するには，実験だけではなくてそれを主導する理論が必要不可欠である．こうして脳科学が，分子生物学，細胞生物学，システム脳科学に理論脳科学を加えて大きく発展しつつある．

21 世紀に入り，科学技術の分野で学問分野の大きな変動が起こっている．分野を超えた交流であり，方法論の融合である．計算機科学，統計科学，物理学，数理科学はもとより，生命科学，更に人文科学を加えて，知的情報処理分野を統合する新しい動きが興っている．

1-1-3 ニューラルネットワークの並列のダイナミックス

ニューロンを多数結合した回路網においては，現在の状態から結合による相互作用によって並列の計算が行われ，次の状態が決まる状態遷移が刻一刻と行われていく．これが並列計算である．離散時間と連続時間のモデル，それに離散変数と連続変数のモデルがあるが，最もよく使われる連続時間連続変数のモデルを述べよう．

各ニューロンの状態，例えば仮想的な平均膜電位を変数とし，出力であるパルス頻度をその成分ごとの非線形の関数とする．いま各ニューロンの膜電位を成分とするベクトルを $u = (u_1, \cdots, u_n)$，出力を成分とするベクトルを $z = (z_1, \cdots, z_n)$ とする．ニューロン間の結合を行列 $W = (w_{ij})$ で表せば，状態遷移の方程式は

$$\tau \frac{du(t)}{dt} = -u + Wf(u) + s \tag{1·1}$$

のように書ける．s は入力ベクトルである．

一般に，非線形のダイナミカル方程式 (1·1) は，その挙動として，どこから出発しても単一の状態に収束する単安定回路，いくつかの安定平衡状態を有する多安定回路，安定状態が連続につながったもの，例えば 1 次元状につながるラインアトラクター，振動回路，それにカオスの振る舞いをもつもの（ストレンジアトラクター），などいろいろなものがある．ニューラルネットワークは飽和型非線形をもつので発散する解はないが，それ以外にはこれらすべての動作が可能であり，それぞれに並列計算の役割を果たしている．

全体的に一様性と乱雑性の双方をもった回路の場合には，ランダムな結合を考える統計神経力学が有用である[2)]．これにより，こうした回路における多安定性，振動回路，そしてカオスなどが解明されてきた．この種の回路を Hopfield 回路と呼ぶが，Amit はこれを Amari-Hopfield 回路と呼ぶべきであると主張している．

一方，ニューロンが空間に一様に並んだ神経場では，その結合のかたちから，興奮状態の形成，保持，運動などの解析ができる．これは場の理論として，作業記憶のモデルや認知の心理モデルとしても使われている．これには，Wilson-Cowan のモデルと Amari による解析が使われる．

具体的な情報処理としては，連想記憶モデルがある．コンピュータにおける記憶は，記憶事項が各番地に収まっているのが通常である．脳では，記憶事項そのものが書かれているのではなくて，記憶事項を指定する鍵が相互関係として回路の結合の重みに記され，多重に記銘されていて，鍵から相互作用の並列のダイナミックスによって記憶した事項が新たに創られていくのが想起であると考えられる．このとき，記憶事項が回路の多数のアトラクターに対応する．

連想記憶の場合，Hopfield は記憶事項がランダムに作られるとして，その記憶容量を求めた．記憶事項がランダムであれば，これに誘起されて回路の結合の行列は複雑なランダム性を有する．このため，統計神経力学における解析が複雑になる．これは，事項の想起，系列の想起などがあり，今でも理論研究の対象になっている．また，現実の脳において，海馬という部位でこの種の記憶が行われていると考えられていて，研究が進んでいる．

記憶の想起だけでなく，状況に応じて最適解を探索する並列のモデルにおいても多安定回路が用いられる．このとき，Aihara や Tuda による，カオス性を有する回路の特性が優れているという研究があり，脳研究におけるカオスの振る舞いに注目が集まっている[4]．

1-1-4 学習神経回路網

ニューロンは情報変換素子である．これを層状に並べれば，一層の情報変換回路ができ上がる．その能力はたいしたものではない．しかし，これを多層につなげれば，万能の情報変換装置ができ上がる．ここに学習能力を付け加えれば，目的に応じ，環境の変化に適応できる情報変換装置ができ上がる．

1950 年代に，Rosenblatt が提案したパーセプトロンはこのような装置であった．しかし，この時代ではコンピュータによるシミュレーションも思うに任せず，単純パーセプトロンと呼ぶ，最終層の素子のみが学習能力をもつモデルの解析が行われ，これが用いられたに過ぎなかった．

しかし，1967 年にはアナログモデルを用いた多層パーセプトロンの学習の提案が Amari によりなされている．勾配を利用した確率降下法である．こうした方法はその後も提案されたが，これを決定的にしたのが，この方法が誤差逆伝播法（バックプロパゲーション）の名で Rumelhart らによって提案されたことである．これは，コネクショニズムの勃興期に提案され，ニューロブームの花形となった[2]．

誤差逆伝播法を用いると万能の学習機械がつくれる．これにより，パターン認識や関数回帰などで，例題は与えられるもののその正解を計算する仕組みが明らかでない問題に対して，学習によってうまく動作するシステムをつくれる．これは，簡便で万能であることから，学習機械の原型となり，今でもよく使われるツールの一つである．

しかし，その学習動作は遅く，プラトーと呼ぶ平坦領域にとらえられる．また，学習の平衡状態（局所解）は無数にあり，これが障害となる．これを克服すべく，多くの研究がその後に行われた．学習の遅滞に対しては，これが中間層のニューロンの対称構造（交換に関する群不変性）にあること，このためにパーセプトロンの空間のリーマン幾何構造が縮退し，ここで遅滞が生ずることが確かめられた．情報幾何による自然勾配学習はこの難点がないことが明らかにあり，このような縮退が多様体に特異構造をもたらすことが Amari らや Fukumizu により明らかになった．

多様体の特異構造と統計的な推論の関係，その学習への影響，Bayes 学習との関係など，微分幾何学や代数幾何学を駆使した研究が Fukumizu や Watanbe らにより開発されている[5]．また，多数の局所解があるこのようなモデルにおける，弱学習機械の統合法，特に Boosting と呼ぶ学習法が提案されて，機械学習の分野の大きな話題となっている[6]．

機械学習は，例題に学んで自己を調整するが，それだけでは例題に特化した機械にすぎない．例題に特化しすぎることは過適応とよばれる．学習機械が例題から学ぶことによりどこまで一般性を獲得できるのか，その汎化能力が問われる．訓練誤差と汎化誤差との関係が解析された．また，局所解のない線形の学習手法であって，信号の非線形変換を許容するカーネル法の計算手法が見出され，万能の装置として大きな注目を集めている．

機械学習は，このほかにも信念伝播法と呼ばれるグラフィカルモデル（ベイジアンネットともいう）における確率推論の手法が注目を集めている[6]．これは，極めて多数のデータからそこに隠された構造を発見するデータマイニングにとって有力な手段となる．

1-1-5 自己組織化回路網

神経回路網は，外界の情報構造に合わせて，最適な情報処理ができるように自己を組織化する．例えば，外界に顕著な特徴があれば，これを認知し処理するのに適した構造を自己の内部に作り上げる．このための学習回路網が提案され，外界の信号の特徴を抽出する回路網，更に概念の形成などのモデルが提案されている．この種の学習は，出力を伴わず外界からの入力信号のみで学習するため，教師信号なし学習とも呼ばれる．外界の信号集団からその主成分を抽出するモデルなども提案された．

更に，網膜から大脳皮質にいたる経路では，左右眼の網膜上に投射される外界の 2 次元の信号を外側膝状体から大脳皮質にいたるまでそのトポロジー構造を保った写像ができている．これには遺伝情報に基づく化学物質親近性を手がかりとする初期の配線の形成と，それを自己組織化によって精密化する機構が働く．von der Malsburg は，自己組織化モデルを提唱し，後に Willshaw と協力して自己組織マップの形成のモデルをつくった．Amari はこれに神経場のダイナミクスを合わせて，コラム構造形成のモデルをつくっている．

Kohonen は，これらのモデルを単純化し，工学モデルとして強力な自己組織化マップモデルを提唱した[7]．これは 2 次元に限らず，信号空間における信号の親近性に応じてその表現細胞を神経場の空間につくるもので，膨大なデータからその表現を自動的につくるモデルとして，データマイニングの分野でも注目を浴びている．ロボットの地図学習などにも利用できる．

このほか，神経回路モデルに示唆を受けたものとして，外界の信号からその独立な成分を抽出し，信号を独立成分ごとに分解する独立成分分析が注目を集めた．これは主成分分析とは異なり，ガウス性を仮定せずに高次の統計量を用いて学習するもので，音声分離などの信号源分離に利用できる．ここでは情報幾何が使われた．最近は，これを更に発展させて，疎信号解析，正値信号解析などの手法が信号処理の分野で研究されている[8]．

Barto と Sutton は，一連の手順の後に報酬が与えられるシステムの学習の神経機構に興味を持った．これは強化学習と呼ばれ，多段決定過程の学習であり，各過程で直接の教師信号がないという意味で教師なし学習であるが，最終段階では教師信号が与えられる．これは，現実の脳の仕組みの解明に威力を発揮するとともに，ロボットの学習に用いられている．

1-1-6 今後の展望

ニューラルネットワークは，情報処理にかかわる工学的な手法としての側面と，数理的な手法を用いた脳の情報処理原理を解明するための方法という二面性をもつことを述べた．前者は，人工知能や知識処理などの工学分野と融合して発展していくであろう．

後者は，計算論的神経科学として，今はスパイク時系列の解析，学習におけるパルスタイミングの効果など，ミクロな回路構造のモデルの精密化が主流になっている．これは，脳の実験科学と結びついてその有効性を高めていく．しかし，それにとどまらず，思考の原理や概念形成，記憶の自己組織化などの高次の脳機能に迫る必要がある．理論こそがこうした高次機能に迫れるのであり，このために数理脳科学の確立が望まれる．

■参考文献

1) D.E. Rumelhart and D.E. McClleland, “Parallel Distributed Processing,” vol.I, II, MIT Press, 1986.
2) 甘利俊一, “神経回路網の数理,” 産業図書, 1978.
3) P. Dayan and L.F. Abbott, “Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems,” MIT Press, 2001.
4) 津田一郎, “カオス的脳観,” サイエンス社, 1990.
5) 福水健次, 栗木　哲, “特異モデルの統計学,” 岩波書店, 2004.
6) 麻生英樹, 津田宏治, 村田　昇, “パターン認識と学習の統計学,” 岩波書店, 2003.
7) T. Kohonen, “Self-Organizing Maps,” Springer, 2001.
8) A. Cichocki and S. Amari, “Adaptive Blind Signal and Image Processing,” Wiley, 2002.

■S3 群 - 4 編 - 1 章

1-2 相互結合型ニューラルネットワーク

（執筆者：田中利幸）[2009 年 11 月 受領]

相互結合型ニューラルネットワークは，多数の形式ニューロンを相互に結合させてネットワークを構成したものである．リカレントニューラルネットワークとも呼ばれる．ネットワークの中の情報の流れが一方向であり，フィードバックがないようなものは，一般に階層型（あるいはフィードフォワード）ニューラルネットワークと呼ばれる．本節では，相互結合型ネットワークに関する基礎的な事項を整理する．

1-2-1 基礎的事項

(1) 形式ニューロン

生体の神経系を構成するニューロンの振る舞いを数理的にモデル化したものが形式ニューロンである．形式ニューロンは多入力－出力のシステムとしてモデル化される．形式ニューロンの入出力関係は，入力を $\boldsymbol{x} = (x_1, \ldots, x_N)^T \in \mathbb{R}^N$，出力を $y \in \mathbb{R}$ としたとき，

$$y = f(u), \quad u = \sum_{i=1}^{N} w_i x_i = \boldsymbol{w} \cdot \boldsymbol{x} \tag{1·2}$$

というかたちでのモデル化が基本的である．$\boldsymbol{w} \in \mathbb{R}^N$ の各成分はシナプス荷重，結合荷重などと呼ばれ，入力 $\boldsymbol{x}$ の対応する成分に対する重みを表している．入力の重み付き和 $u = \boldsymbol{w} \cdot \boldsymbol{x}$ は，現実の神経細胞においてシナプスからの入力が引き起こす膜電位変化の線形加算を表現している．また，関数 f は出力関数などと呼ばれ，入力の重み付き和 u に対して出力 y がどのように定まるかを記述する．出力関数としては，ステップ関数やシグモイド関数（tanh などの，グラフが S 字形をした関数）などがよく使われる．

$u = \boldsymbol{w} \cdot \boldsymbol{x} + \theta$ のように，u を定める式に定数項 θ を含める場合もあるが，$\tilde{\boldsymbol{w}} = (\boldsymbol{w}^T, \theta)^T$，$\tilde{\boldsymbol{x}} = (\boldsymbol{x}^T, 1)^T$ のように次元を一つ拡張したベクトル $\tilde{\boldsymbol{w}}$, $\tilde{\boldsymbol{x}}$ を考えて $u = \tilde{\boldsymbol{w}} \cdot \tilde{\boldsymbol{x}}$ として式 (1·2) に帰着させたり，あるいは出力関数に θ の効果を取り込んだりして議論することもできる．

多数の形式ニューロンを相互に結合させることで，相互結合型ニューラルネットワークを構成することができる．相互結合型ニューラルネットワークでは情報のフィードバックが存在するため，上記の形式ニューロンのモデルを何らかのかたちで拡張する必要がある．いくつかの代表的な場合について以下で説明する．

(2) ネットワーク

形式ニューロンに連続時間ダイナミクスを組み込むことで，相互結合型ニューラルネットワークを定義できる．具体的には，膜電位変化が入力の線形荷重和に指数的に漸近していくものとみなし，時間スケールを適切に選ぶことで

$$y = f(u), \quad \frac{du}{dt} = -u + \boldsymbol{w} \cdot \boldsymbol{x} \tag{1·3}$$

という時間発展規則にしたがう形式ニューロンを考え，それらを N 個相互に結合させて，

$$\boldsymbol{x} = f(\boldsymbol{u}), \quad \frac{d\boldsymbol{u}}{dt} = -\boldsymbol{u} + W\boldsymbol{x} \tag{1·4}$$

とする．ここで，$\boldsymbol{x} = (x_1, \ldots, x_N)^T$, $\boldsymbol{u} = (u_1, \ldots, u_N)^T$ であり，出力関数 f はベクトル $\boldsymbol{u}$ に対して成分ごとに作用するものとする．行列 $W = (w_{ij})$ は結合荷重行列であり，その ij 成分 w_{ij} は形式ニューロン j から i への結合荷重を表す．それぞれの形式ニューロンから自分自身への結合（自己結合）を考えないこととする場合も多く，その場合は W の対角成分はすべて 0 とみなす．

相互結合型ニューラルネットワークを離散写像系として表現することもできる．具体的には，$t = 0, 1, \ldots$ を離散時刻を表す変数として，

$$\boldsymbol{x}_{t+1} = f(\boldsymbol{u}_t), \quad \boldsymbol{u}_t = W\boldsymbol{x}_t \tag{1·5}$$

と定義する．

こうして定義される相互結合型ニューラルネットワークは，非線形力学系としてとらえることができ，そのダイナミクスは一般には大自由度カオスを含む極めて複雑な挙動を示しうる．出力関数 f を定めたとき，ネットワークが安定となるために W がみたすべき条件を議論する安定性解析に関しては，多くの研究がなされている．

相互結合型ニューラルネットワークを構成する今，一つのアプローチは，形式ニューロンの動作に確率性を持ち込むことである．具体的には例えば，形式ニューロンの出力 y は ± 1 いずれかの値をとるものとして

$$P(y = 1) = f(u), \quad P(y = -1) = 1 - f(u); \quad u = \boldsymbol{w} \cdot \boldsymbol{x} \tag{1·6}$$

といった確率的状態更新規則を定め，それらを N 個相互に組み合わせて確率的に動作する相互結合型ニューラルネットワークを定義する．

こうして定義される確率的ネットワークは，個々の形式ニューロンに対応する「局所的」な条件つき確率分布 $P(x_i|\boldsymbol{x})$ の組合せによって $\boldsymbol{x}$ の「大域的」な確率分布を表しているものと考えることもできる．このようなネットワークは，形式ニューロンをノード，それらの間の結合を有向辺とする有向グラフで表現することもでき，グラフィカルモデルもしくはベイジアンネットワークと呼ばれる確率モデルの一例を与えている．

1-2-2 結合荷重が対称な場合

出力関数 f が有界な単調増加関数であり，かつ結合荷重行列 W が対称である，すなわち $W^T = W$ である場合には，式 (1·4) や (1·5) の系はリャプノフ関数をもち，したがって漸近安定であることが示される[1)]．系のある状態が安定点となるように結合荷重行列 W をうまく選ぶことができれば，その状態を「記憶」し，系のダイナミクスを使って記憶させた状態を「想起」することができる．

また，$f(u) = (1 + \tanh \beta u)/2$ であるとき，式 (1·6) の確率的ニューロンを対称な結合荷重行列で相互結合させた確率的ネットワークは，N 個のニューロンが非同期的に状態更新する場合，$H(\boldsymbol{x}) = -\boldsymbol{x}^T W \boldsymbol{x}/2$ から導かれるボルツマン・ギブス分布

$$P(\boldsymbol{x}) \propto \exp[-\beta H(\boldsymbol{x})] \tag{1·7}$$

のかたちの平衡状態分布をもつことが知られている．このような確率的ネットワークはボルツマンマシン[2]とも呼ばれる．ボルツマンマシンも，結合荷重行列 W をうまく選ぶことによって確率的な記憶と想起のモデルとして使うことができる．

1-2-3 連想記憶モデル

記憶させたいベクトルを p 個用意したとき，系のダイナミクスを使ってこれらのベクトル $\{\boldsymbol{\xi}^1, \boldsymbol{\xi}^2, \ldots, \boldsymbol{\xi}^p | \boldsymbol{\xi}^i \in \mathbb{R}^N\}$ を正確にあるいは近似的にでも想起できるようにするために，これらのベクトルをいかにして結合荷重行列 W に記憶させたらよいか，という問題に関しては，非常に多くの研究がなされている．上述のような記憶と想起のためのモデルは，特に連想記憶モデル[3, 4, 5] あるいはホップフィールドモデル[6]などと呼ばれる．結合荷重行列 W をいかに定めるか，という問題は，いわば連想記憶モデルの学習法を問うことに相当する．基本的な学習則として，相関学習

$$W = \frac{1}{p}\sum_{i=1}^{p} \boldsymbol{\xi}^i (\boldsymbol{\xi}^i)^T \tag{1·8}$$

を挙げることができる．

記憶パターン $\boldsymbol{\xi}^i$ の各成分を独立に確率 $1/2$ で ± 1 とするランダムパターンを p 個用意し，それをヘブ則で記憶させたときに，想起が可能な記憶パターン数 p の上限を問う問題は，連想記憶モデルの記憶容量の問題として多くの研究がなされている．情報統計力学の手法を使った Amit らの結果[7]がよく知られており，それによれば，パターンの次元 N が十分大きいとき，記憶パターン数 p の上限はおよそ $0.14N$ 程度である．

1-2-4 ボルツマンマシンの学習

ボルツマンマシンの学習の問題は，ある確率分布 $q(\boldsymbol{x})$ が与えられたときに，その分布をボルツマン・ギブス分布 (1·7) で最もよく近似するための W を学習によって獲得する問題として定式化される．カルバック・ライブラのダイバージェンス $D(q\|p) = \sum_{\boldsymbol{x}} q(\boldsymbol{x}) \log[q(\boldsymbol{x})/p(\boldsymbol{x})]$ の W に関する勾配降下を考えることによって，ボルツマンマシン学習則[2]

$$\frac{dW}{dt} = \langle \boldsymbol{x}\boldsymbol{x}^T \rangle_q - \langle \boldsymbol{x}\boldsymbol{x}^T \rangle_p \tag{1·9}$$

を導くことができる．ここで $\langle\cdot\rangle_p$ は分布 p に関する期待値を表す．

「隠れ素子」をもつボルツマンマシンについても，同様の学習則を導くことができる．隠れ素子の状態を表すベクトルを $\boldsymbol{z}$ とおき，$\tilde{\boldsymbol{x}} = (\boldsymbol{x}^T, \boldsymbol{z}^T)^T$ と表記すると，隠れ素子をもつボルツマンマシンが表現する平衡状態分布は

$$p(\boldsymbol{x}) = \sum_{\boldsymbol{z}} p(\tilde{\boldsymbol{x}}), \quad p(\tilde{\boldsymbol{x}}) \propto e^{-\beta H(\tilde{\boldsymbol{x}})}, \quad H(\tilde{\boldsymbol{x}}) = -\frac{1}{2}\tilde{\boldsymbol{x}}^T W \tilde{\boldsymbol{x}} \tag{1·10}$$

というかたちで与えられる．対応する学習則は条件つき確率 $p(\boldsymbol{z}|\boldsymbol{x}) = p(\tilde{\boldsymbol{x}})/p(\boldsymbol{x})$ を使って

$$\frac{dW}{dt} = \langle \tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^T \rangle_{q(\boldsymbol{x})p(\boldsymbol{z}|\boldsymbol{x})} - \langle \tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^T \rangle_{p(\tilde{\boldsymbol{x}})} \tag{1・11}$$

と導かれる．

ボルツマンマシンを実際に使用する際には，学習則における期待値の評価に必要な計算量の多さが問題となり，平均場近似などの近似的な期待値評価のための手法が数多く研究されている[8]．また，ボルツマンマシンの構造にある種の制約を課し，それを基本モジュールとして階層的な構成を考えることによって，大規模な確率モデルに対する効率的な学習則を得ようという方向の研究も近年多くなされている[9]．

■参考文献

1) J.J. Hopfield, “Neurons with graded response have collective computational properties like those of two-state neurons,” Proc. Natl. Acad. Sci. USA, vol.81, pp.3088–3092, May 1984.
2) D.H. Ackley, G.E. Hinton, and T.J. Sejnowski, “A learning algorithm for Boltzmann machines,” Cognitive Science, vol.9, pp.147–169, Jan.-March 1985.
3) K. Nakano, “Associatron–a model of associative memory,” IEEE Trans. Systems, Man, and Cybernetics, vol.SMC-12, pp.380–388, July 1972.
4) T. Kohonen, “Correlation matrix memories,” IEEE Trans. Computers, vol.C-21, pp.353–359, April 1972.
5) J.A. Anderson, “A simple neural network generating an interactive memory,” Mathematical Biosciences, vol.14, pp.197–220, Aug. 1972.
6) J.J. Hopfield, “Neural networks and physical systems with emergent collective computational abilities,” Proc. Natl. Acad. Sci. USA, vol.79, pp.2554–2558, April 1982.
7) D.J. Amit, H. Gutfreund, and H. Sompolinsky, “Storing infinite numbers of patterns in a spin-glass model of neural networks,” Physical Review Letters, vol.55, pp.1530–1533, Sep. 1985.
8) D. Saad and M. Opper (eds.), “Advanced Mean Field Methods: Theory and Practice,” MIT Press, Cambridge, 2001.
9) G.E. Hinton and R.R. Salakhutdinov, “Reducing the dimensionality of data with neural networks,” Science, vol.313, pp.504–507, July 2006.

■S3 群 - 4 編 - 1 章

1-3 情報統計力学の方法

（執筆者：樺島祥介）[2009 年 5 月 受領]

ボルツマンマシン，ベイジアンネットワークなどソフトコンピューティングで用いられる知識表現の多くは多変数の確率モデルとみなすことができる．これらのモデルに基づく推論や学習では，観測されたデータに対して未観測のデータやパラメータを推定することが必要となる．一般にこの種の課題に必要な計算量はモデルの大きさに対して指数関数的に増大する．よって，実際的時間で実行可能な近似アルゴリズムの開発が実用上重要になる．

気体や磁性体などの物理系に現れる多体問題を解析するために発展した統計力学では大自由度システムを近似的に取り扱う手法が数多く知られている．それらの手法を“近似アルゴリズム”として情報の問題に広く活用する情報統計力学の研究が近年活発化している．この節では情報統計力学における代表的な近似法とそれに関連した計算テクニックについて背後にある数理に焦点を当てながら説明する．

1-3-1 平均場近似

例として，N 個のニューロンからなるボルツマンマシン

$$P(\boldsymbol{S}|\boldsymbol{w},\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{w},\boldsymbol{\theta})} \exp\left[\sum_{i>j} w_{ij}S_iS_j + \sum_i \theta_i S_i\right] \tag{1·12}$$

($S_i = \pm 1$, $i = 1, 2, \ldots, N$, $Z(\boldsymbol{w},\boldsymbol{\theta}) = \sum_{\boldsymbol{S}} \exp\left[\sum_{i>j} w_{ij}S_iS_j + \sum_i \theta_i S_i\right]$) を考える．このモデルに対し，$\boldsymbol{S}$ の関数 $\boldsymbol{f}(\boldsymbol{S})$ の期待値 $\langle \boldsymbol{f}(\boldsymbol{S})\rangle = \sum_{\boldsymbol{S}} f(\boldsymbol{S})P(\boldsymbol{S}|\boldsymbol{w},\boldsymbol{\theta})$ を評価することは，N が大きくなると一般に計算量的に難しくなる．平均場近似では現実的な計算量で評価できる近似値により期待値を代替することで，この困難に対する実際的解決を図る．

簡単のため，以下 $\boldsymbol{f}(\boldsymbol{S}) = \boldsymbol{S}$ とし，i 番目のニューロン S_i に着目する．S_i の期待値を具体的に求めることは計算量的に難しいが，式 (1·12) に対してキャレン（Callen）の恒等式と呼ばれる等式

$$\langle S_i\rangle = \left\langle \tanh\left(\sum_{j\neq i} w_{ij}S_j + \theta_i\right)\right\rangle \tag{1·13}$$

($w_{ij} = w_{ji}$) が厳密に成立することは容易に確かめることができる．右辺に現われる $\sum_{j\neq i} w_{ij}S_j + \theta_i$ はしばしば（S_i に関する）局所場（local field）と呼ばれる．

キャレンの恒等式において期待値評価 $\langle\cdots\rangle$ と $\tanh(\cdots)$ の順番を入れ替えると近似式

$$\langle S_i\rangle \simeq \tanh\left(\sum_{j\neq i} w_{ij}\langle S_j\rangle + \theta_i\right) \tag{1·14}$$

が得られる．$\sum_{j\neq i} w_{ij}\langle S_j\rangle + \theta_i$ は局所場の平均を意味するので平均場（mean field）と呼ばれる．S_i 以外のニューロン S_j に対しても同様の近似式を与えるとこれらは期待値 $\langle S_i\rangle$ ($i = 1, 2, \ldots, N$) に関する非線形連立方程式を構成する．これらの方程式を解析的に解くこ

とは一般に難しいが，多くの場合，逐次（反復）代入法などにより実際的な計算量で数値解を得ることが可能である．

求めるべき期待値をこの数値解により代替する近似法が最も素朴な平均場近似（mean field approximation）である．しばしば分子場近似とも呼ばれる．以下に説明するクラスタ変分法など，与えられた分布を期待値計算が容易な分布で近似する手法一般を平均場近似と総称することもある．

1-3-2 クラスタ変分法

分子場近似は実際的な計算量で解を構成できるものの，得られる解の近似誤差はしばしば許容できる範囲を越えてしまう．そこで，近似精度を系統的に改良する方法がいくつか提案されている．クラスタ変分法（cluster variation method）はその代表的な枠組みである[1)]．

系統的に近似法を構成するために，式 (1･12) に対して変分自由エネルギー

$$\mathcal{F}(Q) = -\sum_{\boldsymbol{S}} Q(\boldsymbol{S}) \left(\sum_{i>j} w_{ij} S_i S_j + \sum_i \theta_i S_i \right) + \sum_{\boldsymbol{S}} Q(\boldsymbol{S}) \ln Q(\boldsymbol{S}) \quad (1\cdot15)$$

を導入する．ここで $Q(\boldsymbol{S})$ は真の分布 (1･12) を近似するテスト分布である．$Q(\boldsymbol{S})$ に関する汎関数である式 (1･15) は真の分布 $Q(\boldsymbol{S}) = P(\boldsymbol{S}|\boldsymbol{w}, \boldsymbol{\theta})$ に対して最小値 $-\ln Z(\boldsymbol{w}, \boldsymbol{\theta})$ をとる．このことは，テスト分布 $Q(\boldsymbol{S})$ を実際的な計算量で期待値評価が可能な分布族に制限し，式 (1･15) を最小化することによって様々な近似法が導出できることを意味する．期待値 $\langle S_i \rangle$ 自体をパラメータとする因数分解可能なテスト分布 $Q(\boldsymbol{S}) = 2^{-N} \prod_{i=1}^{N} (1 + \langle S_i \rangle S_i)$ に対して式 (1･15) を最小化すると，平均場近似 (1･14) が得られる．

クラスタ変分法では $\boldsymbol{S}$ を基本的な部分要素（クリーク）の組に分割し，各クリークに関する結合確率の集まりによってテスト分布を構成する．クリークへの分割は式 (1･15) に複雑な変数間の依存関係をもたらし，近似解の構成は一般に困難になる．この困難はクリークへの分割方法に応じて変分自由エネルギー (1･15) 自体を適切に近似することで回避される．

相互作用の基本単位である S_i と S_j の組をクリークとした場合について説明する．この近似はペア近似あるいはベーテ（Bethe）近似と呼ばれる．ペア近似では各ペア S_i，S_j に関する結合確率並びに各ニューロン S_i に関する周辺確率に対応するテスト分布 $b_{ij}(S_i, S_j)$ 並びに $b_i(S_i)$ を用意する．$b_{ij}(S_i, S_j)$，$b_i(S_i)$ はしばしばビリーフ（belief）と呼ばれる．これらに対し変分自由エネルギー (1･15) を

$$\begin{aligned} \mathcal{F}_{\mathrm{Bethe}}(\{b_{ij}, b_i\}) \quad = \quad & \sum_{(ij)} \sum_{S_i, S_j} b_{ij}(S_i, S_j) \ln \frac{b_{ij}(S_i, S_j)}{\exp\left[w_{ij} S_i S_j + \theta_i S_i\right]} \\ & + \sum_i (1 - c_i) \sum_{S_i} b_i(S_i) \ln \frac{b_i(S_i)}{\exp\left[\theta_i S_i\right]} \end{aligned} \quad (1\cdot16)$$

によって近似する．ここで (ij) は結合 $w_{ij} = w_{ji}$ がゼロでないクリーク（ペア）を意味し，c_i はそのようなクリークのうちでニューロン S_i が関係するものの個数を表す．

近似解は $\mathcal{F}_{\mathrm{Bethe}}(\{b_{ij}, b_i\})$ を最小化することにより構成される．ただし，ここで結合確率 $b_{ij}(S_i, S_j)$ と周辺確率 $b_i(S_i)$ は独立ではなく可約（reducibility）条件

$$\sum_{S_j} b_{ij}(S_i, S_j) = b_i(S_i) \tag{1·17}$$

を満たす必要があることに注意しなければならない．そのため，クラスタ変分法ではこの拘束条件付最小化問題を効率よく解くための数値計算上の工夫が重要な研究課題となる．

1-3-3 確率伝搬法

ペア近似では拘束条件 (1·17) の下で目的関数 (1·16) を最小化する必要がある．ラグランジュの未定乗数法を適用すると，この問題は目的関数

$$\mathcal{F}_{\rm Bethe}(\{b_{ij}, b_i\}) + \sum_i \sum_{S_i} \sum_{j \in \mathcal{N}(i)} \lambda_{i\to j}(S_i) \left(\sum_{S_j} b_{ij}(S_i, S_j) - b_i(S_i) \right) \tag{1·18}$$

に関する拘束条件のない極値問題に還元される．ただし，$\lambda_{i\to j}(S_i)$ は拘束条件 (1·17) に対応して導入されたラグランジュ未定乗数である．簡潔に表現するためビリーフ $b_{ij}(S_i, S_j)$，$b_i(S_i)$ の規格化に関する拘束条件は省略している．$\mathcal{N}(i)$ は S_i とゼロでない w_{ij} で結合しているニューロンの添え字の集合を表す．

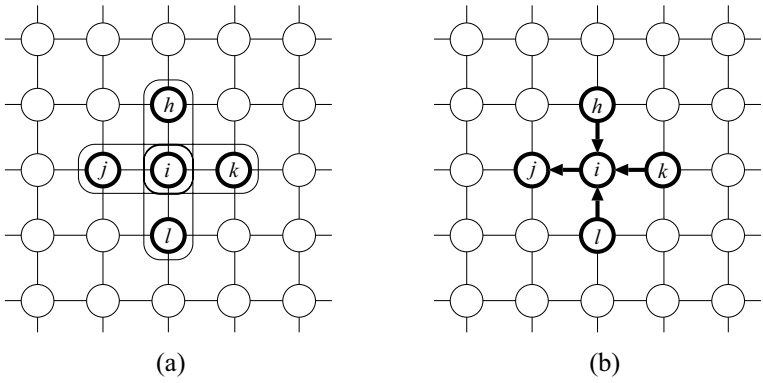


図 1·1　2 次元正方格子に対するペア近似．(a):**最近接ペアでクリークを形成する．**S_i **は四つのクリークと関係しているので** $c_i = 4$ **である．**(b):**確率伝搬法** (1·21) **の動作の様子．**i **から** j **に送るメッセージ** $m_{i\to j}$ **は** j **以外から** i **に流入するメッセージ** $m_{h\to i}$**，**$m_{k\to i}$ **及び** $m_{l\to i}$ **に基づいて計算される．**

ビリーフを消去し，式 (1·18) の極値条件をラグランジュ乗数を用いて表現すると

$$e^{-\frac{1}{c_i - 1}\sum_{k\in\mathcal{N}(i)} \lambda_{i\to k}(S_i) + \lambda_{i\to j}(S_i)} \propto \sum_{S_j} e^{w_{ij} S_i S_j} \left(e^{\theta_j S_j} e^{-\lambda_{j\to i}(S_j)} \right) \tag{1·19}$$

が得られる．ここで $S_i = \pm 1$ についてはメッセージ変数 $m_{i\to j}$ を用いて必ず $e^{\theta_i S_i - \lambda_{i\to j}(S_i)} \propto (1 + m_{i\to j} S_i)/2$ のように表現できることを用いると，ラグランジュ未定乗数 $\lambda_{i\to j}(S_i)$ に

関する関数方程式 (1･19) は $m_{i\to j}$ に関する連立非線形方程式

$$m_{i\to j} = \tanh\left(\theta_i + \sum_{k\in\mathcal{N}(i)\backslash j} \tanh^{-1}\left(\tanh(w_{ik})m_{k\to i}\right)\right) \tag{1･20}$$

に帰着される（図 1･1(b)）．ただし，$\mathcal{N}(i)\backslash j$ は $\mathcal{N}(i)$ から j を除いた添え字の集合を表す．$\langle S_i\rangle$ の近似値は式 (1･20) の解を用いて

$$\langle S_i\rangle \simeq \sum_{S_i} S_i b_i(S_i) = \tanh\left(\theta_i + \sum_{j\in\mathcal{N}(i)} \tanh^{-1}\left(\tanh(w_{ij})m_{j\to i}\right)\right) \tag{1･21}$$

と評価される．

非線形方程式 (1･21) の解を解析的に求めることは容易ではない．しかしながら，適当な条件の下では平均場近似の場合と同様に式 (1･21) の反復代入を繰り返すことで数値解を効率的に求めることが可能である．この解法は確率推論の分野で確率伝搬法（probability propagation）あるいは信念伝搬法（belief propagation）として知られているアルゴリズムにほかならない[2)]．

確率伝搬法はペア近似に対応する極値条件の反復代入を意味している．クラスタ変分法の枠組みではペアよりも大きなクリークを基本単位とした近似法を構成することもできる．この枠組みに沿って確率伝搬法を一般化したアルゴリズムを導出することも可能である[3)]．

1-3-4 レプリカ法

平均場近似やクラスタ変分法は，分布を定めるパラメータの組 $\boldsymbol{w}$ や $\boldsymbol{\theta}$ が具体的に与えられた状況で，分布 (1･12) からの期待値評価に関する計算量的な難しさを近似的に解決するための方法である．しかしながら，情報に関する研究課題はこのようなアルゴリズム開発にかかわるものだけではない．情報理論や通信工学には $\boldsymbol{w}$，$\boldsymbol{\theta}$ が確率的に与えられる場合に分布 (1･12) に含まれる情報量の評価が必要となる問題が数多く存在する．不規則系の統計力学で発達したレプリカ法（replica method）はこれら情報科学における性能評価問題に対する強力な解析法として注目されつつある．これまでにレプリカ法によって画期的な成果が得られた情報科学の問題例としては，連想記憶模型・パーセプトロンの容量評価，学習曲線の解析，低密度パリティ検査符号・線形ベクトル通信路模型の性能評価，ランダム K-SAT 問題に関する SAT/UNSAT 転移の評価などが挙げられる[4)]．

例として，式 (1･12) に関しパラメータ $\boldsymbol{w}$，θ が確率分布 $P(\boldsymbol{w}) = \prod_{i>j} P(w_{ij})$ 及び $P(\boldsymbol{\theta}) = \prod_i P(\theta_i)$ にしたがって生成される場合に，1 自由度当たりの自由エネルギーの期待値

$$-\frac{1}{N}\left[\ln Z(\boldsymbol{w},\boldsymbol{\theta})\right] = -\frac{1}{N}\int \prod_{i>j} dw_{ij} P(w_{ij}) \prod_i d\theta_i P(\theta_i) \ln Z(\boldsymbol{w},\boldsymbol{\theta}) \tag{1･22}$$

を評価する問題を考える．統計力学の知見を用いると，式 (1･22) に基づき相互情報量や条件付エントロピーなどボルツマンマシン (1･12) に関する様々な性能指標の評価が可能になる．

一般に $\ln Z(\boldsymbol{w},\boldsymbol{\theta})$ の $\boldsymbol{w}$，$\boldsymbol{\theta}$ に関する依存性は複雑であるため式 (1･22) の評価は技術的

に難しい．ところで，$n = 1, 2, \ldots$ に対しては展開式

$$Z^n(\boldsymbol{w}, \boldsymbol{\theta}) = \sum_{\boldsymbol{S}^1, \boldsymbol{S}^2, \ldots, \boldsymbol{S}^n} \exp\left[\sum_{a=1}^{n}\left(\sum_{i>j} w_{ij} S_i^a S_j^a + \sum_i \theta_i S_i^a\right)\right] \tag{1·23}$$

が成立する．これを利用すると $\boldsymbol{w}$，$\boldsymbol{\theta}$ に関する期待値を成分ごとに独立に評価することができる．そのため，$n = 1, 2, \ldots$ に関するモーメント $[Z^n(\boldsymbol{w}, \boldsymbol{\theta})]$ については式 (1·22) の評価を阻む技術的な困難は生じない．レプリカ法ではこの性質に着目し，$n = 1, 2, \ldots$ に関して $[Z^n(\boldsymbol{w}, \boldsymbol{\theta})]$ を n の関数として解析的に評価した後，その関数形を実数の n へ解析接続することで，恒等式

$$-\frac{1}{N}\left[\ln Z(\boldsymbol{w}, \boldsymbol{\theta})\right] = -\lim_{n \to 0} \frac{1}{N} \frac{\partial}{\partial n} \ln\left[Z^n(\boldsymbol{w}, \boldsymbol{\theta})\right] \tag{1·24}$$

を利用した式 (1·22) の評価を行う．式 (1·23) の右辺に現われる $\boldsymbol{S}^1, \boldsymbol{S}^2, \ldots, \boldsymbol{S}^n$ は同一のシステムパラメータ $\boldsymbol{w}$，$\boldsymbol{\theta}$ を有する n 個の複製系（レプリカ）の状態変数と解釈することができる．これがレプリカ法と呼ばれる名前のゆえんである．レプリカ法により $\boldsymbol{w}$，$\boldsymbol{\theta}$ に関する期待値計算の技術的困難の回避は可能になるものの，モーメント評価に関する計算量的な難しさは依然残されたままである．多くの場合，この困難は平均場近似やクラスタ変分法を用いることにより近似的に解決される．

現時点において，レプリカ法の基礎となる自然数から実数への解析接続については数学的な厳密性が完全に保証されているわけではない．一方，しばしば無限レンジ模型あるいは平均場模型と総称される一群のモデルについては，平均場近似あるいはクラスタ変分法による近似解が大システム極限で厳密解に漸近すると考えられている．そのような状況に対応するいくつかの例については，レプリカ法により得られる結果の数学的厳密性が証明されている[5)]．

■参考文献

1) R. Kikuchi, “A theory of cooperative phenomena,” Physical Review, vol.81, no.6, pp.988–1003, March 1951.
2) Y. Kabashima and D. Saad, “Belief propagation vs. TAP for decoding corrupted messages,” Europhysics Letters, vol.44, no.5, pp.668–674, Dec. 1998.
3) J.S. Yedidia, W.T. Freeman and Y. Weiss, “Constructing Free Energy Approximations and Generalized Belief Propagation Algorithms,” IEEE Trans. Information Theory, vo.l.51, no.7, pp.2282–2312, July 2005.
4) 西森秀稔, “スピングラス理論と情報統計力学,” 岩波書店, 東京, p.206, 1999.
5) M. Talagrand, “Spin Glasses: A Challenge for Mathematicians. Cavity and Mean Field Models,” Springer-Verlag, Berlin, p.586, 2003.

■S3 群 - 4 編 - 1 章

1-4 パーセプトロンと多層パーセプトロン

（執筆者：麻生英樹）[2009 年 3 月 受領]

パーセプトロン（perceptron）は，フランク・ローゼンブラット（Frank Rosenblatt）が 1958 年頃に提案した人間や動物の脳神経系における情報処理のモデルである[1]．いくつかの層（layer）に分かれた複数のニューロンモデル（ユニット）が相互に結合する構造をもち，結合を修正することで入力信号の識別を学習する．ローゼンブラットは様々な構造のパーセプトロンを提案・解析したが，それらの中で，基本パーセプトロン（elementary perceptron）と呼ばれた，入力にあたる S（sensory）層，一つの A（association）層，1 個のユニットからなる R（response）層（出力層）をもち，入力から出力に向かう前向き結合のみをもつネットワークが最も盛んに研究されたこともあり，現在では「パーセプトロン」という言葉は，前向き結合の階層型ニューラルネットワークモデル（図 1･2）という意味で使われることが多い．入力層と出力層以外の層は中間層，隠れ層（hidden layer）などと呼ばれ，隠れ層が存在することを明示するために「多層パーセプトロン（multi-layer perceptron, MLP）」という言葉も使われる．

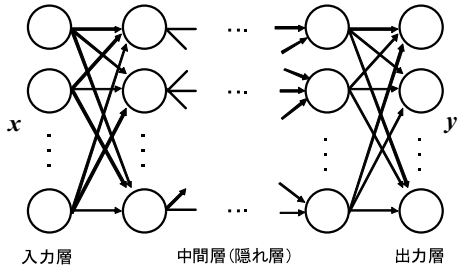


図 1･2　前向き結合階層型ニューラルネットワークモデル（多層パーセプトロン）

1-4-1 単純パーセプトロンと誤り訂正学習

最も簡単なモデルとして，入力層と出力層のニューロン 1 個だけからなるモデルを考える．これは単純パーセプトロン (simple perceptron) と呼ばれることがある．ニューロン 1 個の情報処理は，一般に多次元の入力 $\boldsymbol{x} = (x_1, \cdots, x_n)$ を出力 y に変換する関数であるが，入力信号の重み付き和を取り，その値を変換して出力するかたちにモデル化されることが多い．数式では，$y = f(\sum_i w_i x_i - \theta)$ と書ける．w_i は入力 x_i の結合荷重（connection weight），θ はしきい値（threshold value）と呼ばれる．関数 f としては，ヘヴィサイド関数，単位ステップ関数などの階段形のしきい関数，ロジスティック関数や tanh などのシグモイド関数が使われることが多い．恒等関数が使われる場合には出力は入力の線形和になるため，線形パーセプトロン（linear perceptron）と呼ばれる．

f として単位ステップ関数を用いた場合には，上記の処理は，入力信号 x_i の重みつきの和を取り，しきい値 θ と比較して入力の総和が大きければ 1 を出力し，小さければ 0 を出力することになる．これは入力 $\boldsymbol{x}$ の空間を超平面 $\sum_i w_i x_i - \theta = 0$ で二つに切断し，その片側の入力については 1 を，もう一方の側の入力については 0 を出力することになる．この

ように，単純パーセプトロンによって入力信号は二つのクラスに分類される．クラス 1 とクラス 2 に所属する信号が，入力信号の空間の超平面によって分離できる場合，クラス 1 とクラス 2 は線形分離可能（linearly separable）であるという．定義から明らかに，単純パーセプトロンは入力信号のクラスが線形分離可能な場合にのみ，それらを正しく識別することができる．なお，入力信号に常に－ 1 の値を取るダミーの信号を追加することで，しきい値 θ は結合荷重の一つと見なすことができるため，以下ではダミーの信号を加えたものを改めて $\boldsymbol{x}^{(i)}$ とし，θ と $\boldsymbol{w}$ をまとめたものを $\boldsymbol{w}$ とする．

単純パーセプトロンが入力信号の空間をどのように分離するかは，結合荷重 $\boldsymbol{w}$ によって決まる．入力信号 $\boldsymbol{x}^{(i)}$ とそれに対する正しい識別出力 $t^{(i)}$（1 か 0 の値を取る）のペア N 組 $(\boldsymbol{x}^{(i)}, t^{(i)})(i = 1, ..., N)$ が学習データとして与えられているとする．このとき，これらの入力信号を正しく識別するような $\boldsymbol{w}$ を得る方法はいろいろ考えられるが，ローゼンブラットは次のような単純な手続きを繰り返す方法を提案した．

1) 入力 $\boldsymbol{x}^{(i)}$ が与えられるたびにそれに対する出力 $y^{(i)}$ を求める．

2) 結合荷重 $\boldsymbol{w}$ を $\boldsymbol{w} + \eta(y^{(i)} - t^{(i)})\boldsymbol{x}^{(i)}$ に修正する．

η は正の定数で, 学習の速度を決めるため学習係数, 学習率などと呼ばれる．誤差信号 $y^{(i)} - t^{(i)}$ が出力が正解の場合には 0 となることに注意すれば，この学習規則では，パーセプトロンが誤った出力を出したときだけ結合荷重を修正することが分かる．この性質から，誤り訂正学習（error-correcting leaning）と呼ばれる．また，入力と正解が与えられるたびに結合荷重を修正し，修正結果のみを保存して入力と正解のデータは保存しておく必要がないため，オンラインの学習と呼ばれる．これを繰り返すことで，学習データが線形分離可能である場合には，最終的にすべての入力 $\boldsymbol{x}^{(i)}$ に対して正解を出力するようになることが証明できる（パーセプトロン学習の収束定理）．

1-4-2 多層パーセプトロンと誤差逆伝播学習

単純パーセプトロンの情報処理能力は限られているが，入力層と出力層の間に中間層を追加することによって，その能力を拡大することができる．例えば，好きなだけ多くのユニットをもつ中間層を使えば，任意の識別関数を構成することができるのは明らかである．ミンスキー（Minsky）とパパート（Papert）は，中間層のユニットが限られた範囲の入力層ユニットからのみ結合を受けている（有限次元）などの制約を入れて，多層パーセプトロンの情報処理能力を詳しく検討している．入力層に提示された 1/0 のパターンを図形として見たときに，その連結性を判定するようなパーセプトロンは有限次元ではない，という結果がよく知られている．このほかにも，結合荷重の値の範囲が非常に大きくなるような場合があることも示された[2)]．

誤り訂正学習は，正解が与えられる出力層ユニットの荷重の修正規則を与えているが，中間層のユニットの結合荷重の修正規則は与えられていない．ラメルハート（Rumelhart）らが提案した誤差逆伝播学習（error back-propagation learning）はこの問題に解を与えて，多層パーセプトロンの学習を可能にした[3)]．そのための一つの鍵は，ユニットの出力関数 f としてロジスティック関数などの微分可能な関数を用いることであった．これによって，出力値と正解値の誤差 $R(\boldsymbol{w})$ を特定の結合荷重で偏微分することが可能になり，最急降下法などの連続関数の最適化法を適用できるようになった．ここで， $\boldsymbol{w}$ はすべてのユニットの結

合荷重をまとめたベクトルである．

多層パーセプトロン全体を一つの関数とみなして $f(\boldsymbol{x}, \boldsymbol{w})$ と書く．誤差逆伝播学習では $\boldsymbol{w}$ を $\boldsymbol{w} - \eta \nabla_{\boldsymbol{w}} R(\boldsymbol{w})$ のように最急降下方向に修正する．上記の偏微分係数が，出力層で計算される誤差を入力層の方向へ後ろ向きに伝播してゆくような過程で効率よく計算できることが学習法の名前の由来である．多層のネットワークを最急降下法などの最適化手法で学習させるというアイディアはラメルハートらの提案以前にも研究されていた．例えば甘利は 1967 年の論文で中間層ユニットをもつネットワークの学習のアルゴリズムとシミュレーション結果を示している．

入力信号と正解のペアが与えられるたびに結合荷重の修正を行うことは，それぞれの入力信号に対する誤差を減らすことにはなるが，必ずしも全入力信号に対する誤差を減らすことにはならない．しかし，各回の修正量が小さければ，平均的には全入力信号に対する誤差を減らす方向へと学習が進む．更に確率的降下法の理論から，学習係数の適切な制御を行うことで学習が収束することを示せる．ただし，多層パーセプトロンにおいては，結合荷重に対する誤差の関数は，一般に複数の極小値をもつため，学習が誤差を最小にする解に到達するか否かは初期値に依存することになる．

1-4-3 自然勾配法

誤差逆伝播学習は最急降下法をベースとしているため, 特に学習の終盤で収束速度が遅いという問題があった．この点を補うために, 共役勾配法やニュートン法, 準ニュートン法などのより高速な最適化技法を適用することが有効である[4]．一方, 甘利らは, 情報幾何学（information geometry）の観点から自然勾配法（natural gradient method）を提案した[5]．多層パーセプトロンは非線形の回帰式とみなすことができる．したがって適当な雑音モデルを付加すれば条件付確率分布 $p(y|\boldsymbol{x}; \boldsymbol{w})$ とみなすことができる．この確率分布のフィッシャー情報行列は，$q(\boldsymbol{x})$ を入力信号 $\boldsymbol{x}$ の分布として，$G(\boldsymbol{w}) = \int\int (\nabla_{\boldsymbol{w}} \ln p)(\nabla_{\boldsymbol{w}} \ln p)^T p(y|\boldsymbol{x}; \boldsymbol{w}) q(\boldsymbol{x}) dy d\boldsymbol{x}$ となる．このとき，自然勾配法では $\boldsymbol{w}$ を $\boldsymbol{w} - \eta G^{-1}(\boldsymbol{w}) \nabla_{\boldsymbol{w}} l(\boldsymbol{x}, y, \boldsymbol{w})$ と更新する．ここで $l(\boldsymbol{x}, y, \boldsymbol{w}) = -\ln p(y|\boldsymbol{x}; \boldsymbol{w}) - \ln q(\boldsymbol{x})$．これによって，$p(y|\boldsymbol{x}; \boldsymbol{w})$ の多様体上の自然な座標系で降下方向が決まるため，学習が途中で停滞するプラトー現象が減り，効率良い学習が実現される．フィッシャー情報行列の逆行列を求めることは一般に計算コストが高いが，それをオンラインで適応的に近似しながら計算する適応自然勾配法（adaptive natural gradient method）も提案され，様々なモデルでの有効性が示されている[6]．

■参考文献

1) F. Rosenblatt, “Principles of Neurodynamics, Perceptrons and the Theory of Brain Mechanisms,” Spartan Books, Washington, 1962.
2) M.A. Minsky and S.A. Papert, “Perceptrons Expanded Edition,” MIT Press, Cambridge, 1988. ※中野　馨, 阪口　豊 訳, “パーセプトロン,” パーソナルメディア, 東京, 1993.
3) D.E. Rumelhart, J.L. McClelland, and the PDP Research Group, “Parallel Distributed Processing, Explorations in the Microstructure of Cognition, Volume 1: Foundations,” MIT Press, Cambridge, 1986. ※甘利俊一　監訳, “PDP モデル,” 産業図書, 東京, 1989.

4) C.M. Bishop, “Pattern Recognition and Machine Learning,” Springer-Verlag, New York, 2006. ※元田　浩, 他監訳, “パターン認識と機械学習 (上),” シュプリンガー・ジャパン, 東京, 2007.
5) S. Amari, “Natural gradient works efficiently in learning,” Neural Computation, vol.10, no.2, pp.251–276, Feb. 1998.
6) S. Amari, H. Park, and K. Fukumizu, “Adaptive method of realizing natural gradient learning for multilayer perceptrons,” Neural Computation, vol.12, no.6, pp.1399–1409, June 2000.

■S3 群 - 4 編 - 1 章

1-5 関数近似法

（執筆者：杉山　将）[2008 年 11 月 受領]

教師付き学習（supervised learning）とは，入力 x と出力 y の組からなる n 個の訓練データ $\{(x_i, y_i)\}_{i=1}^n$ を用いて，その背後に潜んでいる入出力関係を学習する問題である[1, 2]．入出力関係をうまく学習することができれば，学習していない入力 x に対する出力 y を予測できるようになる．すなわち，未知の状況に適応する汎化能力（generalization ability）が獲得できる．与えられた訓練データからできるだけ高い汎化能力を獲得することが教師付き学習の目標である．ここでは，訓練データ $\{(x_i, y_i)\}_{i=1}^n$ が同時確率密度 $p(x, y)$ に独立同一分布（independent and identically distributed; i.i.d.）に従うと仮定し，出力 y の条件付き期待値 $\mathrm{E}_{p(y|x)}[y]$ を推定する問題を考える．出力 y が実数値を取るとき回帰（regression）問題と呼び，y がカテゴリ値を取るとき分類（classification）問題と呼ぶ．

1-5-1 線形モデルによる回帰

回帰問題における最も基礎的な学習法は，線形モデル（linear model）を用いた最小二乗法（least-squares）であろう．線形モデルは，基底関数 $\{\varphi_j(x)\}_{j=1}^t$ の線形和によって関数を近似するモデルである．

$$f_{\mathrm{linear}}(x) = \sum_{j=1}^{t} \theta_j \varphi_j(x)$$

最小二乗法は，二乗誤差基準のもとでパラメータ $\{\theta_j\}_{j=1}^t$ を訓練データに適合させる方法である．

$$\min_{\{\theta_j\}_{j=1}^t} \sum_{i=1}^{n} (y_i - f_{\mathrm{linear}}(x_i))^2$$

これは，以下のガウスモデルの最尤推定法（maximum likelihood estimation）に対応している．

$$q_{\mathrm{Gauss}}(y|x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y - f_{\mathrm{linear}}(x))^2}{2\sigma^2}\right)$$

線形モデルに対する最小二乗法の最適化問題は凸であり，大域的最適解が解析的に求まるという長所がある．しかし，あらかじめ基底関数を固定しておく必要があるため，柔軟性に欠ける．

1-5-2 様々な回帰モデル

基底関数にもパラメータを含めることによって，より柔軟な関数近似を行うことができる．そのような非線形モデルの代表例は，動径基底関数（radial basis function; RBF）モデルである．RBF モデルの最小二乗解は次式で求められる．

$$\min_{\{\theta_j,\mu_j,\Sigma_j\}_{j=1}^{t}} \sum_{i=1}^{n} (y_i - f_{\mathrm{RBF}}(x_i))^2$$

$$f_{\mathrm{RBF}}(x) = \sum_{j=1}^{t} \theta_j \exp\left(-\frac{1}{2}(x-\mu_j)^\top \Sigma_j^{-1}(x-\mu_j)\right)$$

このように RBF モデルでは，線形結合係数 $\{\theta_j\}_{j=1}^{t}$ だけでなくガウス基底関数の中心 $\{\mu_j\}_{j=1}^{t}$ と共分散行列 $\{\Sigma_j\}_{j=1}^{t}$ も訓練データを使って適応的に決める．したがって，非常に柔軟なモデリングが可能である．しかし，最適化問題が非凸であるため大域的最適解を求めることは困難であり，勾配降下法などにより局所的最適解を求めるのが一般的である．

線形モデルと RBF モデルの中間に位置づけられるのが，カーネルモデル（kernel model）である．ガウスカーネルモデルの最小二乗解は次式で求められる．

$$\min_{\{\theta_j\}_{j=1}^{n}} \sum_{i=1}^{n} (y_i - f_{\mathrm{kernel}}(x_i))^2$$

$$f_{\mathrm{kernel}}(x) = \sum_{j=1}^{n} \theta_j \exp\left(-\frac{(x-x_j)^\top (x-x_j)}{2\sigma^2}\right)$$

これは凸最適化問題であり，線形モデルのときと同様にして大域的最適解を解析的に求めることができる．更に，パラメータ数が訓練データ数とともに増加することから，線形モデルよりも柔軟に関数近似を行うことができる．

分類問題でも最小二乗法を利用することはできるが，ロジスティック回帰（logistic regression）を用いれば確率的な出力が得られ便利である．ここでは，出力が $y = \pm 1$ の二値分類問題を考える．ロジスティック回帰では，出力 y の条件付き確率 $p(y|x)$ を次のようにモデル化する．

$$q_{\mathrm{logistic}}(y|x) = \frac{1}{1 + \exp\left(-y f_{\mathrm{linear}}(x)\right)}$$

線形モデル $f_{\mathrm{linear}}(x)$ のパラメータ $\{\theta_j\}_{j=1}^{t}$ の最尤推定量は次式で求められる．

$$\min_{\{\theta_j\}_{j=1}^{t}} \sum_{i=1}^{n} \log\left(1 + \exp\left(-y_i f_{\mathrm{linear}}(x_i)\right)\right)$$

これは凸最適化問題であり，勾配降下法や準ニュートン法によって大域的最適解を求めることができる．カーネルモデル $f_{\mathrm{kernel}}(x)$ に対するロジスティック回帰も同様に定義することができ，やはり凸最適化問題として定式化される．

1-5-3 情報量規準

最尤推定法の近似性能は，基底関数の選び方に依存する．学習結果の汎化性能は，カルバック・ライブラー（KL）情報量を使って測るのが一般的である．真の分布 $p(x,y)$ から，学習結果 $q(y|x)p(x)$ への KL 情報量は次式で与えられる．

$$\mathrm{KL}[p(x,y)||q(y|x)p(x)] = \int p(x,y) \log \frac{p(x,y)}{q(y|x)p(x)} dxdy$$

KL 情報量がゼロになることと学習結果 $q(y|x)$ が真の条件付き分布 $p(y|x)$ と一致することは等価である．情報量規準（information criterion）とは KL 情報量の推定量を指し，例えば赤池情報量規準（Akaike information criterion; AIC）[3]は次式で定義される．

$$\mathrm{AIC} = -2 \sum_{i=1}^{n} \log q(y_i|x_i) + 2t$$

ここで t は，モデルに含まれるパラメータの次元（パラメータ数）である．AIC は，適当な条件の下で KL 情報量のよい推定量になっている．したがって，AIC を最小にするようにモデルを決定すれば，高い汎化能力が得られると期待される．しかし，カーネルモデルのようにパラメータ数が訓練データ数とともに増加するモデルや，RBF モデルのように特異性をもつモデルに対しては AIC の近似精度は良くないことが知られている[4]．

■参考文献

1) 元田　浩, 栗田多喜夫, 樋口知之, 松本裕治, 村田　昇（編）, “パターン認識と機械学習（上）：ベイズ理論による統計的予測,” シュプリンガー・ジャパン, 東京, 2007.
2) 元田　浩, 栗田多喜夫, 樋口知之, 松本裕治, 村田　昇（編）, “パターン認識と機械学習（下）：ベイズ理論による統計的予測,” シュプリンガー・ジャパン, 東京, 2008.
3) H. Akaike, “A new look at the statistical model identification,” IEEE Trans. Automatic Control, vol.AC-19, no.6, pp.716–723, Dec. 1974.
4) S. Watanabe, “Algebraic analysis for nonidentifiable learning machines,” Neural Computation, vol.13, no.4, pp.899–933, April 2001.

■S3 群 - 4 編 - 1 章

1-6 サポートベクトルマシン

（執筆者：池田和司）［2008 年 12 月 受領］

1-6-1 マージン最大化

（1）統計的学習理論

二分類問題を考えよう．パーセプトロン学習は与えられた例題を分離できる超平面を見つけると停止するアルゴリズムである．また，多層パーセプトロンにおける誤差逆伝播学習のような二乗誤差を用いる方法は，出力に正規分布に従うノイズが加算されると仮定して最尤推定を行うことに相当する．それでは，与えられた例題の分布が全く未知の場合にはどのように扱えばよいのだろうか．

例題の分布が完全に未知であっても，例えばチェビシェフ不等式など，分布の性質には一定の制限がある．この性質を利用して汎化誤差の上限を定式化したのが PAC 学習である[1)]．PAC（probably approximately correct）とは，与えられた例題数が N のときに誤って分離される確率が ϵ 以下になる確率が $1-\delta$ 以上であるとし，N と ϵ, δ の関係を議論するものである．ここでは学習機械の複雑さを表す VC（Vapnik-Chervonenkis）次元が重要であり，またクラス間の距離を表す“マージン”が陽に現れる．

（2）線形分離におけるマージン最大化

統計的学習理論の枠組みでは，一般にマージンが大きいほど汎化性能はよい．したがって，線形二分機械においては，マージンを最大化する分離超平面が望ましい．これを実現したものがサポートベクトルマシン（SVM）であり，以下のように定式化される[2)]．

入力ベクトル $\boldsymbol{x} \in R^m$ に対し，線形二分機械は重みベクトル $\boldsymbol{w} \in R^m$ としきい値 b を用いて

$$y = \mathrm{sgn}\ \left[\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b\right] \tag{1·25}$$

によって出力 $y \in \{+1, -1\}$ を定める．ここで sgn は符号関数である．線形分離可能，すなわちすべての例題を正しく分離する超平面が存在するような N 個の例題 $(\boldsymbol{x}_n, y_n)$, $n = 1, \ldots, N$ が与えられたとき，マージンは例題と超平面との距離の最小値として定義される．

$$\min_n \frac{y_n(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_n + b)}{\|\boldsymbol{w}\|} \tag{1·26}$$

したがって，SVM は (1·26) を最大にする $\boldsymbol{w}$, b を選ぶ．具体的には，(1·26) が $\boldsymbol{w}$ 及び b の定数倍について不変であることを利用し，最も超平面に近い例題が $y_n(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_n + b) = 1$ を満たす，すなわち

$$y_n(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_n + b) \geq 1 \tag{1·27}$$

という制約条件の下で (1·26) を最大化する．このとき，最大化すべき関数は $1/\|\boldsymbol{w}\|$ となるが，これは $\|\boldsymbol{w}\|^2/2$ を最小化することと等価である．したがって，SVM は

$$\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2 \qquad \text{s.t. } y_n(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_n + b) \geq 1 \tag{1･28}$$

という線形不等式制約付凸 2 次計画問題に帰着される．

(3) 主問題と双対問題

式 (1･28) は SVM の主問題と呼ばれる．通常，不等式制約付凸計画問題は，ラグランジュ関数を用いて双対問題に変換できる．SVM の双対問題は，以下のように導出される．

n 番目の制約式のラグランジュ乗数を $\alpha_n \geq 0$ とし，$\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)$ とすれば，ラグランジュ関数 $L(\boldsymbol{w}, b, \boldsymbol{\alpha})$ は

$$L(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{n=1}^{N} \alpha_n [y_n(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_n + b) - 1] \tag{1･29}$$

と表される．(1･28) の解は (1･29) の鞍点となるので，$L(\boldsymbol{w}, b, \boldsymbol{\alpha})$ を $\boldsymbol{w}$ と b について微分したものは 0 に等しく，

$$\boldsymbol{w} = \sum_{n=1}^{N} \alpha_n y_n \boldsymbol{x}_n, \qquad 0 = \sum_{n=1}^{N} \alpha_n y_n \tag{1･30}$$

が成り立つ．(1･30) は，SVM の分離超平面の法線ベクトル $\boldsymbol{w}$ が例題 $\boldsymbol{x}_n$ の重み付線形和で表されることを示している．

上式を利用して $L(\boldsymbol{w}, b, \boldsymbol{\alpha})$ の $\boldsymbol{w}$ と b を消去すると，(1･28) の双対問題が得られ，

$$\max_{\alpha_n \geq 0} \sum_{n=1}^{N} \alpha_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{n'=1}^{N} \alpha_n \alpha_{n'} y_n y_{n'} {\boldsymbol{x}_n}^{\mathrm{T}} \boldsymbol{x}_{n'}$$
$$\text{s.t. } \sum_{n=1}^{N} \alpha_n y_n = 0 \tag{1･31}$$

となる．これは再び凸 2 次計画問題になっている．(1･31) の解を $\hat{\alpha}_n$, $n = 1, \ldots, N$ とすると，SVM による分離超平面は

$$y = \sum_{n=1}^{N} \hat{\alpha}_n y_n {\boldsymbol{x}_n}^{\mathrm{T}} \boldsymbol{x} + b \tag{1･32}$$

と表される．ここで b は，$\hat{\alpha}_n > 0$ を満たす n 番目の例題を (1･32) に代入して得る．

$\hat{\alpha}_n > 0$ となる例題はサポートベクトルと呼ばれ，一般に例題数 N に比べ少数となる．PAC 学習の枠組みでは，サポートベクトルが少ないほど汎化能力が高いことが知られている．

1-6-2 カーネル法

(1) 特徴空間の導入

入力ベクトルをそのまま利用する SVM では，例題が線形分離不可能な場合には解が存在しない．そこで多層パーセプトロンのように階層化することを考えよう．具体的には，入力

ベクトル $\boldsymbol{x}$ を特徴写像と呼ばれる非線形関数 $f(\cdot)$ により $\boldsymbol{f} = f(\boldsymbol{x})$ に変換し，$\boldsymbol{f}$ の空間においてマージン最大化を行う．$\boldsymbol{f}$ を特徴ベクトルと呼び，$\boldsymbol{f}$ の空間を特徴空間と呼ぶ．

特徴空間は特徴写像 $f(\cdot)$ によって定まる．多層パーセプトロンではこの特徴写像も学習によって得るが，SVM の場合には事前に決めて固定しておくことに注意する．このとき，SVM の双対問題は $\boldsymbol{f}_n = f(\boldsymbol{x}_n)$ を用いて

$$\max_{\alpha_n \geq 0} \sum_{n=1}^{N} \alpha_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{n'=1}^{N} \alpha_n \alpha_{n'} y_n y_{n'} \boldsymbol{f}_n{}^{\mathrm{T}} \boldsymbol{f}_{n'}$$
$$\text{s.t.} \sum_{n=1}^{N} \alpha_n y_n = 0 \tag{1·33}$$

と表され，分離超平面は

$$y = \sum_{n=1}^{N} \hat{\alpha}_n y_n \boldsymbol{f}_n{}^{\mathrm{T}} \boldsymbol{f} + b \tag{1·34}$$

となる．

(2) カーネル関数の利用

式 (1·33) 及び (1·34) には特徴ベクトルの内積が現れている．したがって，特徴空間の次元が高くなると内積の計算コストが大きくなる．しかし，特徴ベクトルは内積のかたちでしか現れず，また特徴ベクトルはもともと $\boldsymbol{x}$ の関数であることに注意すると，$K(\boldsymbol{x}, \boldsymbol{x}') = f^{\mathrm{T}}(\boldsymbol{x}) f(\boldsymbol{x}')$ という関数を事前に定義しておけば，特徴ベクトルの明示的な表現は必要ないことが分かる．この関数 $K(\cdot, \cdot)$ をカーネル関数と呼び，カーネル関数を利用した計算量削減をカーネルトリックと呼ぶ．

カーネル関数を利用する利点は計算量削減だけではない．上で述べたように特徴ベクトルの明示的な表現が必要ないので，特徴ベクトルを意識せずにカーネル関数を設計することができる．実際，カーネル関数 $K(\cdot, \cdot)$ が非負定値，すなわち任意の $\boldsymbol{x}_n, \boldsymbol{x}_{n'} \in R^m$, $c_n, c_{n'} \in R$ について

$$\sum_{n,n'} K(\boldsymbol{x}_n, \boldsymbol{x}_{n'}) c_n c_{n'} \geq 0 \tag{1·35}$$

であれば，特徴空間がヒルベルト空間になるような特徴写像 $f(\cdot)$ が存在することが示されている．これはマーサーの定理と呼ばれている．

また，一種の類似度が定義されていればいいので，$\boldsymbol{x}$ は必ずしもベクトルでなくてよい．実際，文字列やグラフなどを扱えるカーネルが数多く提案されている．

1-6-3 SVM の拡張

(1) C-SVM

特徴空間をうまく選べばどんな例題集合でも線形分離可能にできるが，そのような SVM は必ずしも高い汎化性能をもたない．なぜなら SVM はサポートベクトルをなすごく少数の

例題によって分離超平面を構成するため，ノイズや外れ値に大きく影響を受けるからである．

この問題を克服するため，スラック変数 ξ_n を用いて拘束条件を緩めるテクニックが提案されている[2)]．マージンを ξ_n だけ割り込む例題の存在を許す代わりに，ξ_n をコスト関数に組み込むものであり，

$$\min_{\boldsymbol{w},b,\xi_n} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{n=1}^{N}\xi_n$$
$$\text{s.t. } y_n(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_n + b) \geq 1 - \xi_n, \quad \xi_n \geq 0 \tag{1·36}$$

と定式化される．その双対問題は

$$\max_{0\leq\alpha_n\leq C} \sum_{n=1}^{N}\alpha_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{n'=1}^{N}\alpha_n\alpha_{n'}y_ny_{n'}\boldsymbol{x}_n{}^{\mathrm{T}}\boldsymbol{x}_{n'}$$
$$\text{s.t. } \sum_{n=1}^{N}\alpha_n y_n = 0 \tag{1·37}$$

となり，α_n の上限 C の存在を除いて 式 (1·31) と同じである．この手法はソフトマージンと呼ばれるが，次に紹介する ν-SVM と区別するため，しばしば C-SVM と称される．なお，ソフトマージンを導入しない SVM はハードマージンと呼ばれる．

(2) ν-SVM

C-SVM に現れるパラメータ C はソフトマージンの “柔らかさ” を表しているが，その設定が難しい．ν-SVM はマージンを 1 に固定する代わりに変数 β とし，$-\beta$ をコスト関数に加えるもので，その主問題，双対問題は以下のように表される[3)]．

$$\min_{\boldsymbol{w},b,\xi_n} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{n=1}^{N}\xi_n - \beta$$
$$\text{s.t. } y_n(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_n + b) \geq \beta, \quad \xi_n \geq 0 \tag{1·38}$$

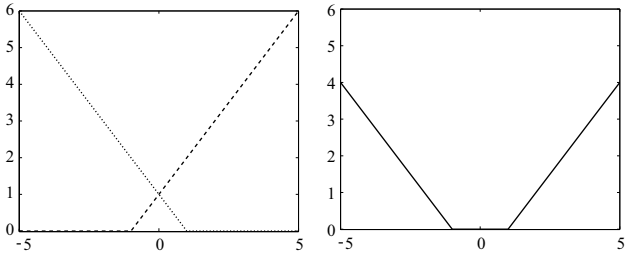
$$\max_{0\leq\alpha_n\leq C} -\frac{1}{2}\sum_{n=1}^{N}\sum_{n'=1}^{N}\alpha_n\alpha_{n'}y_ny_{n'}\boldsymbol{x}_n{}^{\mathrm{T}}\boldsymbol{x}_{n'}$$
$$\text{s.t. } \sum_{n=1}^{N}\alpha_n y_n = 0, \quad \sum_{n=1}^{N}\alpha_n = 1 \tag{1·39}$$

C-SVM は主問題においてマージン最大化という幾何学的な意味をもっているが，ソフトマージン，特に定数 C の意味ははっきりしない．一方，ν-SVM の双対問題は，α_n の和が一定という条件から，例題がつくる二つの縮小凸包の最短線分条件に対応するという幾何学的な意味を持っており，ν は凸包の縮小の度合いを表すことが知られている[4, 5)]．そのため，例題の性質をパラメータ設定に反映させるのが容易である．

(3) サポートベクトル回帰

これまでは二分類問題を考えてきたが，SVM のアイディアは回帰問題にも適用できる．ソ

フトマージンを用いた SVM では，正例に対して下図 (a) の点線の誤差関数，負例に対して破線の誤差関数を適用していた．回帰問題では，ϵ 許容誤差関数と呼ばれる (b) のような誤差関数 $E_\epsilon(\cdot)$ を用いればよい[2]．この回帰のための SVM はサポートベクトル回帰（SVR）と呼ばれる．



（a）二分類問題の誤差関数　　（b）回帰問題の誤差関数

図 1·3

結局，SVR は以下の誤差関数を最小化することになる．

$$\frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{n=1}^{N} E_\epsilon(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_n + b - y_n) \tag{1·40}$$

二分類問題と同様にスラック変数を導入すると，上の双対問題は

$$\begin{aligned}\max_{0\le\alpha_n,\alpha'_n\le C} & -\frac{1}{2}\sum_{n=1}^{N}\sum_{n'=1}^{N}(\alpha_n-\alpha'_n)(\alpha_{n'}-\alpha'_{n'})\boldsymbol{x}_n{}^{\mathrm{T}}\boldsymbol{x}_{n'} \\ & -\epsilon\sum_{n=1}^{N}(\alpha_n+\alpha'_n)+\sum_{n=1}^{N}(\alpha_n-\alpha'_n)y_n\end{aligned} \tag{1·41}$$

という凸 2 次計画問題として表される．

(4) 多クラス SVM

SVM は二分類問題に特化した手法であるため，多クラス分類問題に直接応用するのは困難である．そのため，2 クラス SVM を組み合わせて多クラス問題に応用するのが一般的であり，一つのクラスとそれ以外のクラスに分ける SVM を組み合わせる 1 対他方式，すべてのクラスの組合せについて SVM を用いる 1 対 1 方式，誤り訂正符号を用いた手法などが提案されている．しかしいずれも二分類器が SVM であることは本質的ではない．

また，確率密度分布推定に関連する 1 クラス SVM も提案されている[6]．これは SVM を利用して，特徴空間をデータがある領域とない領域に分けるというものであり，ν-SVM を応用している．

■参考文献

1) L.G. Valiant, “A Theory of the Learnable,” Commun. ACM, vol.27, pp.1134–1142, Nov. 1984.

2) V.N. Vapnik, “The Nature of Statistical Learning Theory,” Springer-Verlag, New York, 1995.
3) B. Schölkopf, et al., “New Support Vector Algorithms,” Neural Computation, vol.12, no.5, pp. 1207–1245, May 2000.
4) K.P. Bennett and E.J. Bredensteiner, “Duality and Geometry in SVM Classifiers,” Proceedings of the 17th International Conference on Machine Learning, pp.57–64, 2000.
5) K. Ikeda and T. Aoishi, “An Asymptotic Statistical Analysis of Support Vector Machines with Soft Margins,” Neural Networks, vol.18, pp.251–259, April 2005.
6) B. Schölkopf, et al., “Estimating the Support of a High-Dimensional Distribution,” Neural Computation, vol.13, no.7, pp.1443–1471, July 2001.

■S3 群 - 4 編 - 1 章

1-7 微分幾何・代数幾何の方法

（執筆者：渡辺澄夫）[2009 年 9 月 受領]

情報学における理論の構築及びアルゴリズムの創出において，確率分布のつくる集合の幾何学が重要な役割を果たすことが知られている．ここでは，微分幾何[1)2)]と代数幾何[3)4)5)]の方法について分かりやすく説明する．厳密な記述を必要とする読者は，参考文献を参照されたい．

1-7-1 微分幾何の方法

例えば，東京からニューヨークへ飛行機で行くとき，長方形に描かれた世界地図（メルカトル図法）の上を直線で結んだ経路は最短距離を与えない．最短距離になる経路を求めるためには，地球儀のように考察する対象の真の形を表すことのできるものが必要である．情報学において理論やアルゴリズムの研究を行うとき，確率分布がつくる集合の真の姿を描き出す必要がある．例えば「平均値 m と標準偏差 σ をパラメータとしてもつ正規分布全体の集合」は，集合としては 2 次元ユークリッド空間の部分集合 $\{(m,\sigma) \in \mathbf{R}^2; \sigma > 0\}$ と同じものと考えてよいが，(m,σ) を変化させたときの学習・認識・予測への影響は線形的ではないから，情報学の意味で自然な観点からは曲がった空間である．

ある集合 W において，その各点を含む十分小さい開集合（近傍という）がとれて，d 次元ユークリッド空間の開集合と同等とみなせるとき，W を d 次元多様体，あるいは単に多様体という．輪ゴムは 1 次元多様体であり球面やドーナツの表面は 2 次元多様体である．以下では，d 次元のパラメータ w によって定まる情報 x の確率密度関数 $p(x|w)$ の全体がつくる集合 $\{p(x|w); w\}$ が多様体である場合を考える．任意の多様体は十分高い次元のユークリッド空間の中に埋め込むことができるが，それでも地球儀のように人間が直感的に扱えるものを制作することはできない．そこで地球儀の代わりに用いられる数学的方法が微分幾何である．

微分幾何では，接空間・計量・接続が基礎的な概念である．第一に接空間について述べる．多様体の点 p の十分小さい近傍はユークリッド空間と同等とみなせるから，「点 p の近傍上に定義された関数を微分するという操作全体」は d 次元のベクトル空間になる．このベクトル空間は点 p で多様体に接している集合と自然に同一視できるのでこれを接空間といい T_p と書く．第二に計量について述べる．多様体の点 p の接空間 T_p の要素であるベクトル $v = (v_i)$ の長さの 2 乗をある正定値行列 g_{ij} を用いて $\sum_{i,j} g_{ij} v_i v_j$ と定めることにする．ここで g_{ij} は点 p に依存する関数である．この g_{ij} が正定値行列であるとき，すなわち，その固有値がすべて 0 より大きく有限の値であるとき g_{ij} を計量という．第三に接続について述べる．多様体の 2 点 p と q が与えられたとき，接空間 T_p から T_q への全単射な線形写像を定義して接続という．接続を定義することは異なる接空間の間に平行移動を定義することであると考えてよい．平行移動を定めると接空間 T_p から T_q への関数を「平行移動してから微分する」という操作が可能になる．これを共変微分 ∇ という．逆に共変微分が定義されれば接続が定まるので ∇ のことも接続と呼ぶ．

ある多様体には無限に多くの異なる計量と接続を定義することができる．計量と接続を一つ定めることは地球儀を一つ定めることであると考えてよい．地球儀を定めれば，与えられ

た 2 点間の距離を最小にする曲線を求めることや，与えられた点の近傍の曲がっている程度や平坦であるかどうかを調べることが可能になる．確率分布がつくる集合として定義される多様体にも無限に多くの異なる計量と接続が定義されうるが，統計学的に自然な要請である「十分統計量をもつ確率モデルにおいては計量と接続は十分統計量だけで表される」を満たすものはフィッシャー計量と α-接続だけである．すなわち統計学の観点から自然な地球儀が決定される．確率モデルが与えられたときフィッシャー計量はフィッシャー情報行列からユニークに定まるが，α-接続は α の値に応じて様々なものがある．指数型分布は $\alpha = 1$ の接続に対して平坦であり，混合型分布は $\alpha = -1$ の接続に対して平坦である．この二つの平坦性は種々のアルゴリズムの構築において重要である．

さて，確率分布がつくる集合の微分幾何を考えることによって少なくとも以下の四つのことが可能になる．第一に，統計学や情報理論の方法や定理を，微分幾何に基づいて情報学的に理解し解釈することができる．例えば，正定値計量をもつ確率分布の集合は統計的正則モデルであると理解できる．クラメル・ラオの不等式は「不偏推定量の共分散の下限がフィッシャー計量によって与えられる」ということであると解釈できる．第二に，統計学や情報理論において計算が困難な問題を見通しよく実行でき計算結果を確認できる．例えば，統計的推測における高次漸近論や検定における検出力関数など計算が複雑であるため解決できていなかった問題において，微分幾何の観点から見通しのよい計算を行うことが可能になり，また結果の正しさや意味を確認することができる．第三に，統計学や情報理論に現れる問題に対して幾何学的に自然な理論やアルゴリズムを創造することができる．例えば，最尤推定量の探索に用いられる最急降下法は，微分幾何の立場からは座標系に依存するため自然な方法ではない．そこで座標系に依存しない方法（微分幾何の意味での最急降下法）を用いるというアルゴリズムを提案することができる．第四に，統計学や情報理論の中から新しい数学的構造が発見され，数学と情報学の間に新しい研究分野が開拓されることがある．例えば，上記で述べた α-接続の概念は確率分布がつくる集合の微分幾何を考察する過程で初めて発見されたものであり純粋数学の研究の対象になっている．

近年，情報学は広く発展を遂げている．量子情報学・脳神経情報学・遺伝子情報学などにおいて，情報の構造を考える場合に確率分布の集合の微分幾何を考えることがしばしば重要な解決を与える．特に量子情報学は観測の概念が古典論とは異なるため，推定や検定の研究において幾何学的方法も刷新されつつある．遺伝子情報学や脳神経情報学においても，対象とする現象の解析や理解において微分幾何の方法は多岐に渡り発展を遂げつつある．

1-7-2 代数幾何の方法

情報学において扱われる対象が複雑な構造を有するようになり，階層構造・モジュール構造・隠れた構造を含む確率分布が研究されるようになった．例えば，混合正規分布・混合 2 項分布・多層パーセプトロン・縮小ランク回帰・ボルツマンマシン・隠れマルコフモデル・確率文脈自由文法などを挙げることができる．これらの確率分布が作る集合はフィッシャー情報行列が固有値 0 を含むので計量を与えない．このようなモデルを総称して特異モデルと呼ぶ．

統計的正則モデルにおいては最尤推定量の漸近分布もベイズ事後分布の漸近形もフィッシャー計量により定まる正規分布になるが，特異モデルにおいてはどちらの分布もフィッ

表 1·1　正則と特異

統計学	正則	特異
確率分布の集合	多様体	解析的集合
代数	線形代数	環とイデアル
幾何	微分幾何	代数幾何
解析	実数に値をとる関数	関数に値をとる関数
確率論	中心極限定理	関数空間上の中心極限定理
最尤推定量	漸近正規かつ漸近有効	発散あるいは非有効
ベイズ事後分布	漸近正規かつ漸近有効	特異分布
AIC 補正値	パラメータの次元	特異ゆらぎ
BIC 補正値	パラメータの次元	対数閾値
例	正規分布，線形回帰	混合正規分布，神経回路網

シャー情報行列では定まらず，正規分布にもならない．このような場合の解析に有効な方法として代数幾何の方法がある．

代数幾何における基礎的な概念に代数と幾何の同等性と双有理同値の概念がある．第一に代数と幾何の同等性について説明する．代数多様体や解析的集合は非常に複雑な特異点を含むため真の形状を幾何学的に把握することは難しい．しかしながら代数多様体の幾何学的な性質は代数多様体を定義するイデアルの代数的な性質に反映されているから，代入や消去といった代数計算によってその特性を知ることができる．これが代数と幾何の等価性である．例えば，代数閉体上では代数的集合の集合と定義イデアルの集合は一対一かつ全射に対応する．また代数多様体上の点が特異点であるかどうかはヤコービ行列のランクについての代数計算で判定できる．更に代数多様体を別の代数多様体に変換すること（ブローアップ，トーリック改変）は代数計算を行うことと等価である．第二に双有理同値の概念について説明する．微分幾何においては，二つの図形はそれらのヤコービ行列の行列式が非零の写像で写り合えるときに等価であると考え，その変換によって不変な特徴が研究される．一方，代数幾何においては，二つの図形はそれらのヤコービ行列の行列式が 0 や無限大になりうる写像で写り合えるときに双有理同値であるという．双有理同値な図形に対して不変な特徴を与える数学的な量を双有理不変量という．このとき直感的にまったく違った形の図形も双有理同値になるが，この概念を導入することで任意の代数多様体が特異点をもたない図形と双有理同値になる（特異点解消定理）．情報学上は，特異点をもつ任意の確率モデルに対して双有理同値な空間をうまく選ぶことでカルバック情報量を正規交差特異点だけをもつようにできるという点が重要である．

以上の二つの考え方を用いると，例えば次のことが可能になる．統計的正則モデルでは任意の確率モデルの対数尤度の漸近形はパラメータの 2 次形式で与えられた．特異モデルでは任意の確率モデルの対数尤度は正規交差因子と法則収束する確率過程を用いて与えられる．統計的正則モデルでは，平均対数尤度の漸近形が AIC であり，ベイズ対数周辺尤度の漸近形が BIC であって，確率モデルの妥当性を数値的に表す情報量規準を定義している．統計的正則モデルの AIC と BIC にはパラメータ空間の次元 d が経験対数尤度に対する補正量

として現れる．特異モデルでは，AIC に相当する概念では d の代わりに特異ゆらぎという双有理不変量が現れる．また，BIC に相当する概念では d の代わりに実対数閾値という双有理不変量が現れる．これらの双有理不変量は与えられたサンプルから不偏推定することができる．統計的正則モデルと特異モデルの対応を表 1･1 にあげる．

■参考文献

1) 甘利俊一，長岡浩司，“情報幾何の方法，” 岩波書店, 東京, 1993.
2) S. Amari, K. Nagaoka, “Methods of Information Geometry,” Oxford University Press, Oxford, 2000.
3) 福水健次，栗木　哲，竹内　啓，赤平昌文，“特異モデルの統計学，” 岩波書店, 東京, 2004.
4) M. Drton, B, Sturmfels, S.Sullivant, “ Lectures on Algebraic Statistics,” Birkhäuser, Basel, 2008.
5) S. Watanabe, “Algebraic Geometry and Statistical Learning Theory,” Cambridge University Press, Cambridge, 2009.

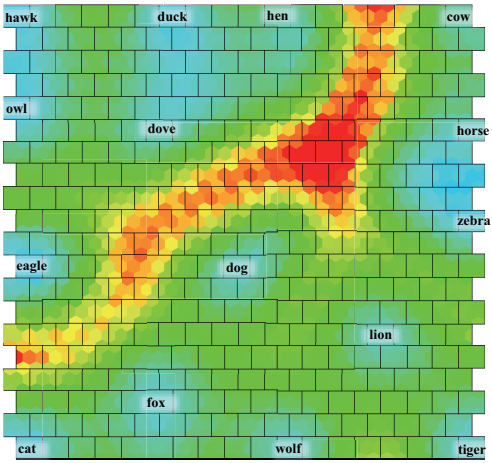
■S3 群 - 4 編 - 1 章

1-8 自己組織化写像

（執筆者：古川徹生）[2008 年 10 月 受領]

1-8-1 自己組織化写像（SOM）の概要

自己組織化写像（Self-Organizing Map: 以下 SOM と表記）はコホネン（Kohonen）が考案した教師なし学習アルゴリズムであり，高次元データの可視化や解析などに用いられる[1)]．SOM の目的は，高次元ベクトルのデータ集合を，データ分布の位相的構造をなるべく保持したまま低次元（多くの場合は 2 次元）の特徴空間へ写像することである．すなわち SOM とは高次元データ空間から低次元特徴空間へのトポロジー保存写像を自己組織的に生成する学習装置である．SOM によってデータ分布は特徴空間内の地図のように可視化される．これを特徴マップと呼ぶ（「マップ」の本来の意味は「写像」であるが，しばしば地図に比喩される）．



	owl	dog	cow	···
小型	1	0	0	···
中型	0	1	0	···
大型	0	0	1	···
夜行性	1	0	0	···
2 本足	1	0	0	···
4 本足	0	1	1	···
毛を持つ	0	1	1	···
蹄を持つ	0	0	1	···
鬣を持つ	0	0	0	···
羽を持つ	1	0	0	···
縞がある	0	0	0	···
狩猟する	1	0	0	···
速く走る	0	1	0	···
空を飛ぶ	1	0	0	···
泳ぐ	0	0	0	···
草食	0	0	1	···

図 1･4　自己組織化写像 (SOM) による動物マップとそのためのデータベクトル

図 1･4 は SOM による「動物マップ」の例である．この例では，様々な動物の特徴が 16 次元のベクトルデータとして表現されており，これらの動物データを SOM により 2 次元特徴空間へ写像したものが「動物マップ」である．似た特徴をもつ動物（例えば哺乳類や鳥類，草食動物や肉食動物）はマップ上でも近い位置に写像されている．このように高次元データを低次元空間へ写像することで，データ分布の位相的構造を可視化するのが SOM の働きである．得られた特徴マップからはデータどうしの類似関係を読み取れるだけでなく，U-matrix と呼ばれる手法で特徴マップに濃淡を付加し，クラスタ境界を可視化することもできる．図 1･4 の場合は哺乳類と鳥類に対応する二つのクラスタが見て取れる．

SOM は特徴マップを活かしたデータ分析やデータマイニングに使われるだけでなく，写像学習装置，多様体学習装置として信号処理やパターン認識，ロボティクスなど幅広い分野で応用されている．

1-8-2 自己組織化写像 (SOM) のアルゴリズム

SOM のアーキテクチャはユニットの集合体でできており，各々のユニットはデータと同じ次元のベクトル（参照ベクトル）値を記憶する．図 1･4 の例では六角格子上に並んだ 18×20 個のマスがユニットを表す．このユニット配置が特徴空間のトポロジーを決定する．この例ではデータが 16 次元ベクトルなので，各ユニットは 16 次元のベクトル値を記憶する．第 k 番ユニットが記憶する参照ベクトルを $\boldsymbol{w}_k$ とすれば，そのユニットは 16 次元データ空間内の一点 $\boldsymbol{w}_k$ を指し示すとともに，特徴空間上の固定位置 $\boldsymbol{y}_k$ に縛られている（図 1･5）．SOM の学習とは，データ集合 $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ を与えることで参照ベクトル集合 $\{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_M\}$ を逐次的に更新し決定することである．

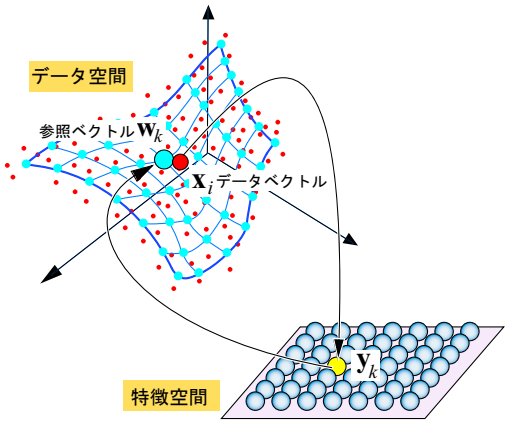


図 1･5　SOM のアーキテクチャ概念図

SOM の学習アルゴリズムの本質は，最近傍則によるユニット間競合と，近傍関数によるユニット間協調の組合せである．学習アルゴリズムにはいくつかのバリエーションが存在し，しばしば使われるのがオンライン型とバッチ型である．SOM の歴史を知るうえでは内積型 SOM も重要であるが，現在では使われることが少ない．ここでは基本となるバッチ型を解説する．

バッチ型アルゴリズムでは，以下の Step 0〜3 を 1 ループとして繰り返し計算する．

Step 0　計算に先立って各ユニットの参照ベクトルを乱数などで初期化し，計算時刻を $t = 0$ とする．

Step 1　すべてのデータベクトルに対して，それぞれ最も近い参照ベクトルをもつユニットを「勝者」（Winner もしくは Best Matching Unit）とする．すなわちデータ $\boldsymbol{x}_i$ に

対する勝者 $k^*(i)$ を最近傍則 $k^*(i) = \arg\min_k \|\boldsymbol{w}_k - \boldsymbol{x}_i\|$ で決定する．

Step 2 近傍関数を用いて，各ユニットがそれぞれのデータを学習する量を決定する．第 k 番ユニットが第 i 番データを学習する度合いは

$$\alpha_{k,i} = \frac{h\left(\|\boldsymbol{y}_k - \boldsymbol{y}_{k^*(i)}\|; \sigma(t)\right)}{\sum_{i'=1}^{N} h\left(\|\boldsymbol{y}_k - \boldsymbol{y}_{k^*(i')}\|; \sigma(t)\right)} \tag{1･42}$$

で与えられる．$h(d;\sigma)$ が近傍関数であり，特徴空間内の距離 d について単調減少関数である．すなわち特徴空間において勝者からの距離が近いユニットほど多くの学習量が与えられる．近傍関数としてはガウス関数が一般的である．$\sigma(t)$ は近傍半径とも呼ばれ，近傍関数の広がりを決める．近傍半径は学習時刻 t とともに単調減少する．

Step 3 各ユニットの参照ベクトルを次式で更新する．

$$\boldsymbol{w}_k(t) = (1-\varepsilon)\,\boldsymbol{w}_k(t-1) + \varepsilon \sum_{i=1}^{I} \alpha_{k,i}\,\boldsymbol{x}_i \tag{1･43}$$

ε は学習係数であり，$0 < \varepsilon \leq 1$ の範囲を取る．バッチ型の場合，$\varepsilon = 1$ とすることが多い．各ユニットの参照ベクトルを更新し終えたら，$t := t+1$ として Step 1 に戻る．

オンライン型の場合は各ループについて 1 個（もしくは少数個）のデータを SOM に与える．この場合，学習係数 ε を小さく取り，更に計算時刻 t とともに ε を 0 に近づける．また式 (1･42) は，分子のみで評価されることもある．

1-8-3 利用上の注意点

SOM は教師なし学習であるため，データを与えれば何らかの特徴マップが得られるが，それがユーザの望む結果である保証はない．例えばデータ空間と特徴空間の位相構造が一致しない場合，特徴マップはデータ分布を適切に反映しない．また特徴マップはデータのスケール変換などに対して影響を受けるため，事前のデータ処理にも注意を要する．したがって，得られた特徴マップの妥当性は十分に検証するべきである．ほかの手法，例えばサモンのマップなどと比較することは有効な検証方法である．

基本的なアルゴリズムでは，参照ベクトルは乱数により初期化される．しかし実用目的の場合は，まずデータに対し主成分分析を行い，その結果を利用して初期化する方が好ましい．もし特徴空間が 2 次元ならば，第 1，第 2 主成分を用いて 2 次元線形部分空間にデータ分布を射影し，この部分空間上に参照ベクトルを初期配置する．

SOM をクラスタリングやベクトル量子化の目的で使う場合は，ほかの類似手法も検討するとよい．特徴マップを必要としない場合は，後述するニューラルガスの方が良い性能を示す．

SOM の特徴マップについて等確率性や誤差最小性などの原理で最適性を示すことはできない．すなわち各ユニットが勝者になる率は等しくならないし，エネルギー関数を用いて学

習アルゴリズムを導くことも（アルゴリズムに修正を施さない限り）できない[2]．更に学習の収束性についても，簡単な場合以外は示されていない．こうした理論的不完全さを補うため，修正された SOM アルゴリズムや SOM に類似したアルゴリズムがいくつか提案されている．生成位相マップ (Generative Topographic Map: GTM) [3]やカーネルベース最大エントロピー学習 (Kernel-based Maximum Entropy Learning: kMER)[2]などが代表例である．ただし計算負荷は SOM より大きい．

1-8-4 SOM のバリエーションおよび関連手法

SOM には多くのバリエーションがある．アーキテクチャを変えたものとして，特徴空間のトポロジーを可変にした成長型 SOM や，階層構造を取り入れた木構造 SOM などがある．扱う対象をベクトルデータ以外に拡張したものとして，適応部分空間 SOM（Adaptive Subspace SOM: ASSOM）や演算子マップ（Self-Organizing Operator Map），メジアン SOM などがある．

SOM をベクトル量子化の一手法として見た場合，k-means 法やニューラルガスなどが類似アルゴリズムとして挙げられる．特にニューラルガスは位相構造の制約がなく，ベクトル量子化としての性能は SOM に優る[4]．

SOM は多様体学習法や部分空間法の一種でもあり，非線形主成分分析や主曲線法などが類似手法として挙げられる．主曲線法は理論的に見ても SOM と同等であるなど，SOM のアルゴリズムは様々なかたちで再発見されている．

SOM に関する解説としては，コホネン自身による書籍が詳しい[1]．またヘルシンキ工科大学のウェブサイト `http://www.cis.hut.fi/research/som-research/` では SOM のソフトウェアパッケージ “SOM_PAK” や SOM に関する文献リストなどが公開されている．

■参考文献

1) T. コホネン, “自己組織化マップ・改訂版,” シュプリンガー・フェエラーク東京, 東京, 2005.
2) マーク M. ヴァン・フッレ, “自己組織化マップ—理論・設計・応用,” 海文堂, 東京, 2001.
3) C.M. Bishop, M .Svensén and C.K.I. Williams, “GTM: The generative topographic mapping,” Neural Computation, vol.10, no.1, pp.215–234, Jan. 1998.
4) T.M. Martinetz, S.G. Berkovich and K.J. Schulten, “Neural-gas network for vector quantization and its application to time-series prediction,” IEEE Trans. Neural Networks, vol. 4, no.4, pp.558–569, July 1993.

■S3 群 - 4 編 - 1 章

1-9 低ランク行列因子化

（執筆者：赤穂昭太郎）[2009 年 5 月 受領]

高次元の入力データをそのまま学習や推論に用いようとすると二つの問題が起きる．一つは，次元が高いゆえに処理の計算量や必要な記憶領域が大きくなるため，大量のサンプルを扱うことが難しくなる点である．もう一つは，データの次元とともに学習モデルのパラメータも多くなり，いわゆる次元の呪いと呼ばれる現象が起きて，学習における汎化能力が低くなる点である．

高次元のデータといっても，本質的に重要な情報はその一部分に過ぎないことが多いので，適切に情報を抽出することによって情報圧縮することができれば上記の問題点は解決できる．本節では，高次元実数ベクトルを低次元の実数ベクトルに次元圧縮する低ランク行列因子化と呼ばれる手法について説明する*．特に 2 次元ないし 3 次元空間に次元圧縮すれば，データを散布図のかたちで表示することができるため，低ランク行列因子化は可視化の目的でも有効である．

低ランク行列因子化とは，一般に，N 個の n 次元データ $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ を $N \times n$ 行列のかたちに並べ $X = (\boldsymbol{x}_1^T, \ldots, \boldsymbol{x}_N^T)^T$ とするとき，行列 X を $N \times m$ 行列 U と $m \times n$ 行列 V を用いて

$$X \simeq UV \tag{1·44}$$

のかたちで近似することをいう． $m = n$ ならば，厳密に X を再現する U と V を求めることができるが，低次元化では通常 $m < n$ とするため，一般には X を忠実に再現することはできない．この近似によって，X の各行に相当する n 次元データ $\boldsymbol{x}_i^T$ は，対応する U の第 i 行の m 次元ベクトル $\boldsymbol{u}_i^T$ に圧縮されているとみなすことができる．

低ランク因子化を，データの生成のされ方という観点で言い換えると，まず元となる内的な因子 $\boldsymbol{u}_i$ というのがあって，それが何らかの行列 V によって線形変換されて高次元データとして観測されたものが $\boldsymbol{x}_i$ であるというモデル化をしていると解釈することもできる．

さて，このような近似を行うためには，近似のよさを評価する何らかの規準を定めてそれを最適化する必要があり，その規準の定め方によっていろいろな手法が導出される．

また，任意の正則な m 次正方行列 W について $UV = (UW)(W^{-1}V)$ が成り立つので，$U' = UW$, $V' = W^{-1}V$ も同じ近似を与える．したがって U や V に制約を置いて解が一意に求められるようにすることが多い．

1-9-1 特異値分解に基づく方法

(1) 特異値分解

低ランク行列因子化の最も基本的なものは特異値分解（SVD: Singular Value Decomposition）に基づく方法である．任意の $N \times n$ 行列 X は，$N \times n$ 行列 U, n 次対角行列 Λ, $n \times n$ 行列 V によって

* 低ランク行列因子化に対して，高次元実数ベクトルを有限個の点で代表させるのがクラスタリングである

$$X = U\Lambda V^T, \quad U^TU = I, \quad V^TV = I,$$
$$\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n), \quad \lambda_1 \geq \ldots \geq \lambda_n \geq 0 \tag{1·45}$$

のかたちに分解できる．これを X の特異値分解という．上式の $\lambda_1, \ldots, \lambda_n$ を特異値と呼ぶが，特異値のうち小さいものを 0 に置き換えると，以下のように X の低ランク近似が得られる．すなわち，U, V のはじめから m 番目の列を取り出してそれぞれ $\tilde{U}, \tilde{V}$ をつくり，$\tilde{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_m)$ とおき，

$$\tilde{X} = (\tilde{U}\tilde{\Lambda})\tilde{V}^T \tag{1·46}$$

とすることにより，$N \times m$ 行列 $\tilde{U}\tilde{\Lambda}$ と $m \times n$ 行列 $\tilde{V}^T$ の積のかたちの近似が得られる．

特異値分解で作られる $\tilde{X} = (\tilde{X}_{ij})$ は，$X = (X_{ij})$ との差の二乗和（フロベニウスノルム）を最小にする低ランク行列因子化であることが知られている．つまり，ユークリッド距離で測ったときに，X に対する復元誤差を最小にするものとなっている．

（2）主成分分析

特異値分解は，多変量解析の代表的な手法である主成分分析（PCA: Principal Component Analysis）とほぼ等価である．ただし，主成分分析では，前処理として各サンプル $\boldsymbol{x}_i$ から全体の平均 $\mathbf{m} = (\sum_{i=1}^{N} \boldsymbol{x}_i)/N$ を引いておくことが多い．これにより，平均からの残差行列を低ランク因子化することになる．以下の説明ではデータからあらかじめ平均が引かれているものとして，特異値分解による低ランク因子化を主成分分析と呼ぶことにする．

主成分分析では，もとのデータ X から低次元表現 $\tilde{U}\tilde{\Lambda}$ への変換は

$$\tilde{U}\tilde{\Lambda} = X\tilde{V} \tag{1·47}$$

によって与えられる．これは入力データ $\boldsymbol{x}^T$ に $\tilde{V}$ をかけることによって低次元表現が得られることを意味している．$\tilde{V}$ は直交行列 V の一部の列だけを取り出したものであるから，上記の変換は，入力 $\boldsymbol{x}$ を回転させて，一部の軸だけを取り出したものに相当する．その各軸方向のデータの標準偏差は λ_i, $i = 1, 2, \ldots, m$ に比例するため，主成分分析はデータの分散が大きい部分空間を見つけてそこに射影する方法であるとみなすこともできる．

この「分散最大」という性質は，データが多変量正規分布に従っているという仮定のもとで，もとのデータの情報量（エントロピー）を最大限保持するような線形変換になっていると言い換えることもできる．

1-9-2 学習アルゴリズム

主成分分析では，データ行列 X を特異値分解することによって低次元化された $n \times m$ 行列 $\tilde{V}$ を求めることができる．これはすべてのデータがそろってから学習を行うので一般にバッチ型の学習と呼ばれる．一方，データ $\boldsymbol{x}$ が逐次的に与えられ，そのたびに $\tilde{V}$ の値を更新し，$\boldsymbol{x}$ の値そのものは捨ててしまうような学習は一般にオンライン型（逐次型）の学習と呼ばれ，データの性質が時刻とともに変動するような場合や，記憶領域を節約したい場合などに有効な方法である．

主成分分析に対するオンライン型の学習アルゴリズムのうち，最も基本的なものとして以

下のものがある[1)]．

$$W_{t+1} = W_t + \gamma \boldsymbol{x}\boldsymbol{x}^T W_t, \quad \tilde{V}_t = W_t (W_t^T W_t)^{-1/2} \tag{1·48}$$

ここで，W_t は $n \times m$ 行列で，W_0 を適当な値で初期化しておき，上の逐次式を $t = 0, 1, 2, \ldots$ に対して適用していく．第 2 式は，$\tilde{V}_t$ の列ベクトルが互いに直交するという制約を満たすように W_t を修正するためのものである．この学習則は，主成分分析の一つの性質である「分散最大化」を行うオンライン型の最急勾配法である．つまり，低次元化されたデータの分散の総和は $\mathrm{E}[\mathrm{tr}(W^T \boldsymbol{x}\boldsymbol{x}^T W)]$ と書けるので，それを W について微分して $\mathrm{E}[\ ]$ を除くと，式 (1·48) の第 1 式の $\boldsymbol{x}\boldsymbol{x}^T W_t$ が得られる．オンライン学習の収束性については，確率近似法や可積分系についてのいろいろな理論研究がある[2)]．

(1) 次数の決定

低ランク因子化において，次数 m をどこまで取るかというのは重要な問題である．可視化では 2〜3 次元と決まっているし，コンピュータビジョンの 3 次元復元法の一種である因子分解法[3)] などのように，あらかじめ想定される次元数が物理的に決まっている場合もある．

しかしながら一般には m をデータに基づいて決めなければならない．主成分分析では，寄与率 $\sum_{i=1}^{m} \lambda_i^2 / \sum_{j=1}^{n} \lambda_j^2$ がある閾値を越えるかどうかで m を決めることが多い．また，次項で述べる確率モデルに基づく方法で低ランク因子化を行う際にはいろいろな統計的モデル選択法（AIC, MDL, 交差検証法など）を適用することにより m を決めることができる．

1-9-3 確率モデルに基づく方法

ここでは，データの生成過程を確率的にモデル化することによって低ランク行列因子化を行う手法について説明する．

観測される n 次元ベクトル $\boldsymbol{x}$ は m 次元の内的な因子ベクトル $\boldsymbol{u}$ に $n \times m$ 行列 $\tilde{V}$ をかけて高次元にしたものに，ランダムなノイズが付加されたものとみなす． すなわち，

$$\boldsymbol{x} = \tilde{V}\boldsymbol{u} + \boldsymbol{n} \tag{1·49}$$

と仮定する．このように観測データの生成過程を確率的にモデル化することにより，低ランク行列因子化の問題を統計的な推定の問題として扱うことができる．実際に $\boldsymbol{u}$ や $\boldsymbol{v}$ の分布をどのようにモデル化するかによって様々な手法が導かれる．特に，指数分布族と呼ばれる確率モデルについては情報幾何学の観点から統一的に議論できることが知られている[4)]．また，$\tilde{V}$ にも事前確率を導入すればベイズ推定の枠組みに拡張することができる．

パラメータ推定のアルゴリズムについても，最急勾配法や Newton 法といった一般的な最適化法のほか，EM アルゴリズムや変分ベイズ法などの統計的推定アルゴリズムを用いることが可能である．

(1) 因子分析

古くから因子分析（Factor Analysis）と呼ばれているのは，$\boldsymbol{u}$ を等方的な標準正規分布，$\boldsymbol{n}$ を各成分ごとに独立な正規分布としてモデル化して，パラメータを最尤推定によって定めるものである．また，確率的主成分分析と呼ばれているものも基本的には因子分析の一種とみなすことができる[5)]．

さて，主成分分析と因子分析の関係について述べておく．式 (1･46) の行列分解を，$\tilde{U}$ と $\tilde{\Lambda}\tilde{V}$ の積と考えると，$\tilde{U}$ は等方的な分散をもつため，ノイズの効果を無視すれば因子分析における $\boldsymbol{u}$ を並べたものと同じになる．ただし，ノイズがある場合には，ノイズモデルを陽に仮定して統計的なモデル化を行った因子分析と主成分分析の結果は異なる．

因子分析のパラメータ推定のアルゴリズムは，特異値分解を行うだけで済む主成分分析と異なり，一般に繰り返し計算で求める．これにはいろいろな方法があるが，例えば，モデルから定義される尤度を目的関数として，適当な初期値（典型的には主成分分析による解）を出発点として，EM アルゴリズムや最急勾配法によって最適化を行う．

因子分析で問題となるのは，$\tilde{V}$ が一意的に定まらないことである．なぜなら因子に等方的な正規分布を仮定しているため，因子ベクトルに任意の直交行列 W をかけた $W\boldsymbol{u}$ も同じく等方的な正規分布となるからである． つまり，元々の生成過程の式 (1･49) と

$$\boldsymbol{x} = (\tilde{V}W^T)(W\boldsymbol{u}) + \boldsymbol{n} \tag{1･50}$$

とが同じ尤度となる．この余った自由度を消すためには，新たに別の最適化規準を設定する必要があり，近年注目されているのが以下に説明する独立成分分析である．

(2) 独立成分分析

因子分析で余っていた回転の自由度を消すために，独立成分分析（ICA: Independent Component Analysis）では因子の成分間の統計的独立性という規準を課す．独立性は無相関よりも強い概念である． そもそも因子分析で得られた成分間は無相関になっているため，独立成分分析ではより高次の統計量を見る必要がある． ここでは簡単のためノイズを含まない場合の独立成分分析について解説する．

因子の成分 $u_1, u_2, \ldots, u_m$ があったとき，それらが独立であるとは $p(u_1, u_2, \ldots, u_m) = p(u_1)\ p(u_2) \cdots\ p(u_m)$ が成り立つことである．ただし，通常独立性を厳密に満たすような因子は求められないし，m 変数の同時確率分布を有限個のサンプルから推定することも困難であるため，独立性を測るいくつかの近似的な評価規準により最適化することで直交行列 W を求める手法が提案されている． その中でも一般性の高いのは，$p(u_1, u_2, \ldots, u_m)$ と $p(u_1)p(u_2)\cdots p(u_m)$ の間の Kullback-Leibler ダイバージェンスを損失関数として取ったもので，最急勾配法は

$$W_{t+1} = W_t + \gamma(I - E[\boldsymbol{v}(\boldsymbol{u})\boldsymbol{u}^T])W, \quad \boldsymbol{v}(\boldsymbol{u}) = \left(\frac{\partial \log p(u_i)}{u_i}\right)^T_{i=1,\ldots,m} \tag{1･51}$$

というかたちで与えられる． ここで，周辺分布 $p(u_i)$ は何らかのかたちでモデル化して推定し，u_i で微分可能なかたちにしておく必要がある．$E[\cdot]$ はサンプルに関する平均を表しており，これを取り去ることによりオンライン学習則が導ける．また，上式では省略したが，このままのかたちで W_t を更新すると直交行列ではなくなるので，主成分分析のオンライン学習と同様に直交化の操作が必要である． なお，単純に損失関数を偏微分して得られる最急勾配法は W^{-1} という行列を計算する必要があるが，行列を群として扱う Lie 群の不変な計量に基づいてそれを修正して得られたのが上式の更新式であり，逆行列の計算が不要となっている． このように，最適化パラメータの微分幾何学的構造に着目して自然な計量のもとで

導いた最急勾配法を自然勾配法と呼び，いろいろなメリットがある[6]．

なお，独立性の規準として具体的には正規分布からのずれを測ることが多い．なぜなら，中心極限定理により独立な成分が混じり合うと正規分布に近づくし，もともと正規分布であるような因子に対しては直交行列の自由度を消すことができないからである． 正規分布からのずれを表すものとして 4 次のキュムラント $\kappa_4 = E[u_i^4] - 3E[u_i^2]^2$ などがある（正規分布では 0 となる）．このほか，成分がスパース（多くの成分が 0）となるような規準も提案されている．これらの規準に基づく場合でも基本的には式 (1･51) のバリエーションとみなすことができる．

1-9-4 その他の話題

低ランク行列因子化に関する重要な話題について簡単に言及しておく．

（1）非負行列因子化

データ行列を $X \simeq UV$ と分解したときに，因子 U や係数行列 V の物理的な性質から，それらが正の値だけを取ることが分かっていることがある．そのような制約条件のもとで低ランク因子化を行う枠組みを非負行列因子化（NMF: Nonnegative Matrix Factorization）と呼ぶ[7]．基本的には非負を取るポアソン分布などの確率モデルに基づいた推定問題に帰着させる．

（2）カーネル法による非線形化

低ランク行列因子化は，因子と観測値との線形関係を求めるものであるが，観測データ $\boldsymbol{x}$ を n' 次元の実ベクトルに非線形変換し，$\mathbf{s}(\boldsymbol{x})$ としたものを特徴ベクトルとして，特徴ベクトルに対して低ランク行列因子化を施すことが考えられる．この際，n' を大きく取ればそれだけ分解の自由度を増やすことになる．

カーネル法は，n' がサンプル数よりも大きい場合に有効な枠組みを与える[8]．二つの $\boldsymbol{x}$, $\boldsymbol{x}'$ に対する特徴ベクトル間の内積をカーネル関数 $k(\boldsymbol{x}, \boldsymbol{x}') = \mathbf{s}(\boldsymbol{x}) \cdot \mathbf{s}(\boldsymbol{x}')$ として定義し，N 個のサンプル $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ 間のカーネル関数の値を並べた行列 $K = (k(\boldsymbol{x}_i, \boldsymbol{x}_j))_{i,j=1,\ldots,N}$ をグラム行列と呼ぶ．カーネル関数は直観的にはサンプル間の類似度を表す．

特徴ベクトルに対する主成分分析はグラム行列の主成分分析と等価であることが知られている．また，カーネル関数は正定値性という性質を満たせば，陽に $\mathbf{s}(\boldsymbol{x})$ を計算する必要がないことが知られており，n' が無限に大きい場合などには計算量が大きく減らせるため，カーネルトリックと呼ばれている．

カーネル法を使うもう一つのメリットは，文字列やグラフ構造など実数値でないデータ観測に対しても，カーネル関数さえ定義できれば適用可能であるということである．ただし，カーネル法ではカーネル関数の定義の仕方によって結果が大きく変わることがあるので，適切に定義することが重要となる．

■参考文献

1) E. Oja, “Principal Components, Minor Components, and Linear Neural Networks,” Neural Networks, vol.5, pp.927–935, November-December 1992.
2) 中村佳正，“アルゴリズムと可積分系，” システム制御情報学会誌, vol.43, no.11, pp.584–592, Nov. 1999.

3) 藤木　淳, “点対応を用いた複数の 2 次元画像からの 3 次元形状復元―因子分解法の数理―,” 統計数理, vol.49, no.1, pp.77-107, 2001.
4) S. Akaho, “The e-PCA and m-PCA: dimension reduction of parameters by information geometry,” Proceedings. 2004 IEEE International Joint Conference on Neural Networks, vol.1, pp.129–134, July 2004.
5) C. Bishop, “Pattern Recognition and Machine Learning,” Springer-Verlag, New York, 2006（元田　浩 他監訳, “パターン認識と機械学習（上・下）,” シュプリンガー・ジャパン, 東京, 2007, 2008.）
6) Y. Nishimori and S. Akaho, “Learning Algorithms Utilizing Quasi-Geodesic Flows on the Stiefel Manifold,” Neurocomputing, vol.67, pp.106–135, August 2005.
7) D.D. Lee and H.S. Seung, “Algorithms for Non-negative Matrix Factorization,” Advances in Neural Information Processing Systems, vol.13, pp.556–562, MIT Press, Cambridge, 2001.
8) 赤穂昭太郎, “カーネル多変量解析 ―非線形データ解析の新しい展開,” 岩波書店, 東京, 2008.
9) 村田昇, “入門独立成分分析,” 東京電機大学出版局, 東京, 2004.

■S3 群 - 4 編 - 1 章

1-10 アンサンブル法

（執筆者：村田　昇）[2008 年 10 月 受領]

数理科学においてアンサンブル（ensemble）とは系の集合，あるいは集団を表す言葉として用いられるが，機械学習の文脈においては単なる無秩序な集団ではなく，統制・調和のとれた集団を積極的に表す．効率的な学習則をもつ単純な学習器（弱学習器と呼ばれることがある）を組み合わせることによって，より複雑なタスクをこなすことができる複合体的な学習器を構成する方法を一般にアンサンブル法と呼ぶ（図 1･6）．

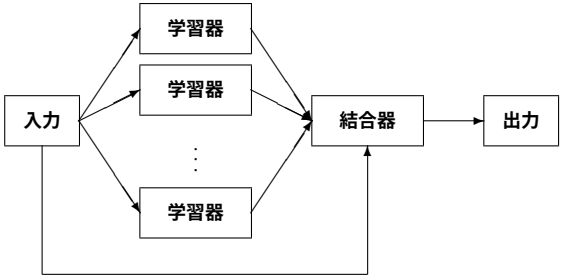


図 1･6　アンサンブル法の概念図．結合器の入力に依存して学習器の出力の重み付けを変化させる動的な構成法と，入力によらず一定の重み付けを行う静的な構成法がある．

学習器の能力を高めるためには，可変なパラメータの数を増やし自由度を高くする必要があるが，一方で自由度の増加とともに学習に必要な計算量は爆発的に増大し，またデータに過剰に適合してしまう過学習（over-training, over-fitting）という現象が顕著に現れるため汎化能力が低下していく．アンサンブル法では集団を大きくすることによって結合した学習器の能力を高めているが，学習に必要な計算量は一般に集団の大きさに比例するため指数的な爆発を避けることができ，また学習器として比較的単純なものを選ぶことによって集団としても過学習しにくい性質をもつようになるため，上記の大自由度の学習器における汎化能力が低下する問題を回避している．

アンサンブルを構成する方法には様々なものが提案されているが，最終的に個々の学習器をどのように組み合わせるかという観点からは大きく二つの考え方に分類される．一つは入力に応じて用いる学習器の重み付けを変える動的な結合で，もう一つは入力によらず一定の重み付けで学習器を組み合わせる静的な結合である．

動的な結合では，実質的に入力空間のクラスタリング（clustering）を行っていることになる．このクラスタリングは更に，明確な境界を定めて不連続に分割するハードクラスタリング（hard clustering）と，クラスタに属する確率を定義してこれを連続的に変化させるソフトクラスタリング（soft clustering）に分類することができる．具体的には k-平均法（k-means clustering method），LVQ（learning vector quantization）や SOM（self-organizing map）などの競合学習，あるいは混合分布を用いたクラスタリング法が用いられる．動的なモデルの代表例は後に詳述する MoE（Mixture of Experts）であるが，決定木により複数

の線形判別や線形回帰を組み合わせる CART（classification and regression tree）などもこの一例であると考えられる．

一方，静的な結合では多数の学習器を並列・独立に構成するのか，あるいはほかの学習器の情報を使って逐次的に構成するのかによって大きく方法が異なる．前者の典型例はバギング（bagging）と呼ばれるブートストラップ法を応用したもので，与えられたデータからリサンプリングしたデータを用いて多様な学習器をつくり，その出力の多数決を用いるという方法である．また，後者の典型例はブースティング（boosting）で，後述するようにデータの一つひとつに重みを設定し，逐次的に重みを更新しながら異なる学習器をつくっていくという方法を用いる．

以下では，MoE とブースティングをアンサンブル法の代表例として取り上げ，その基本的な考え方を紹介する．そのほかの方法に関しては文献 1, 2) などを参照されたい．

1-10-1 Mixture of Experts

MoE（Mixture of Experts）は動的なアンサンブル法の典型例であり，文献 3) において基本的なアイディアが提案され，文献 4) で EM アルゴリズム（EM algorithm）を用いて学習則が整理された．ここでは，与えられた例題 $\mathcal{D} = \{(x_i, y_i); i = 1, \ldots, n\}$ を用いて入力 x に対する適切な出力 y を学習する問題を例に取り，MoE の考え方を説明する．

まず，学習器のパラメータを θ で表し，複数の学習器のパラメータを $\{\theta_k\}$ で記述することにする．入力 x が与えられたとき，k 番目の学習器の出力と y との関係を表す確率モデルを考え，これを $P(y|x;\theta_k)$ と書くことにする．例えば学習器がニューラルネットワークの場合には，出力に加法的な正規ノイズを想定すれば，学習器の出力と目標出力 y の差がガウス分布に従う確率モデルとして考えることができる．あるいは GLM（generalized linear model）のような確率モデルを想定してもよい．次に，どの学習器をどのくらいの重み付けで用いるのかを決定する結合器の出力は，正値で総和が 1 であると制約し，この結合器のパラメータを ξ で表す．この制約によって結合器の出力は入力 x が与えられたときに学習器 k が選ばれる確率としてとらえることができる．この確率を $P(k|x;\xi)$ と書くことにする．

学習器と結合器をそれぞれ確率モデルと考えることによって，MoE 全体は x が与えられたときに y の条件付き確率を表す混合モデル

$$P(y|x;\{\theta_k\},\xi) = \sum_k P(k|x;\xi)P(y|x;\theta_k)$$

としてとらえることができる．このとき与えられた例題 $\mathcal{D}$ に対するパラメータ $\{\theta_k\},\xi$ の最尤推定量は EM アルゴリズムを用いて求めることができる．具体的には以下の計算をパラメータが収束するまで繰り返せばよい．

$$P(k|x,y) = \frac{P(k|x;\xi^{(t)})P(y|x;\theta_k^{(t)})}{\sum_k P(k|x;\xi^{(t)})P(y|x;\theta_k^{(t)})}$$

$$\theta_k^{(t+1)} = \arg\max_{\theta_k} \sum_{(x,y)\in\mathcal{D}} P(k|x,y)\log P(y|x;\theta_k)$$

$$\xi^{(t+1)} = \arg\max_{\xi} \sum_{(x,y)\in\mathcal{D}} P(k|x,y) \log P(k|x;\xi)$$

なお上添字 (t) は繰り返しの回数を表すものとする．

MoE は様々なものを学習器として利用でき，複数の MoE を組み合わせることによって更に階層化することも可能である．

1-10-2 ブースティング

ブースティング（boosting）の理論的な背景は文献 5) に遡るが，最初の実用的なアルゴリズムとして知られる AdaBoost が提案されたのは文献 6) である．ここでは 2 値判別問題を対象として，AdaBoost のアルゴリズムを説明する．

2 値判別問題では，対象とする空間 $\mathcal{X}$ の点 x に 2 値のラベル $y \in \{+1, -1\}$ が割り当てられている状況を考え，与えられた例題 $\mathcal{D}\{(x_i, y_i); i = 1, \ldots, n\}$ から空間全体の判別ルールを推定する．学習器として x を入力すると ± 1 を出力する関数 $h(x)$ を考え，これを用いて判別ルールを最も良く記述する関数を構成することが目的となる．ブースティングアルゴリズムの特徴は各例題 (x_i, y_i) に重み D_i が与えられていると考えて学習を行う点にあり，重みを用いると AdaBoost アルゴリズムは以下のように非常に単純なかたちで書き下せる．

重みの初期値を $D_i^{(1)} = 1/n$ として以下の計算を繰り返し，T 個の学習器を得る．

- 学習器 h が間違えた例題 (x_i, y_i) の添字の集合を $\mathcal{F}(h)$ とする．

  以下で定義される重み付き誤り率をできるだけ小さくする学習器 $h^{(t)}$ を選ぶ．

$$\epsilon^{(t)}(h) = \sum_{i\in\mathcal{F}(h)} D_i^{(t)} \quad \text{（重み付き誤り率）}$$

- 以下で定義される信頼度を $h^{(t)}$ の重み付き誤り率 $\epsilon^{(t)} = \epsilon^{(t)}(h^{(t)})$ を用いて計算する．

$$\alpha^{(t)} = \frac{1}{2} \ln\left(\frac{1-\epsilon^{(t)}}{\epsilon^{(t)}}\right) \quad \text{（信頼度）}$$

- 例題の重みを更新する．

$$D_i^{(t+1)} = \frac{D_i^{(t)} \exp(-\alpha^{(t)} y_i h^{(t)}(x_i))}{Z} \quad \text{（}Z \text{ は } \textstyle\sum D_i = 1 \text{ とする規格化因子）}$$

最終的な判別は，以下の信頼度で重み付けた多数決により行う．

$$H(x) = \mathrm{sign}\left(\sum_{t=1}^{T} \alpha^{(t)} h^{(t)}(x)\right)$$

上記の重みの更新においては，直前の学習器が正解した例題（$h(x)$ と y が同符号）は重みが小さく，逆に誤った例題（$h(x)$ と y が異符号）は重みが指数的に大きくなっていく．こ

のため次の学習では前の学習器が誤った例題を間違えない学習器が選ばれやすくなっていく．これが逐次的に異なる学習器を効率良くつくっていく原理となっている．また，学習器としては線形判別器，スタンプ（decision stump），ニューラルネットワーク（多層パーセプトロン）など様々なものが利用でき，その学習則は誤り率が 50 %未満となることさえ保証されれば計算量の少ない簡便なものを用いて構わない．誤り率 50 %はランダムゲス（random guess）と呼ばれ，当てずっぽうと同等になるため，不可である．このように非常に適用範囲が広いため，応用上は様々な学習問題の増強（boost）に利用されている．

■参考文献

1) 麻生英樹, 津田宏治, 村田　昇, “パターン認識と学習の統計学,” 岩波書店, 東京, 2003.
2) C.M. Bishop, “Pattern Recognition and Machine Learning,” Springer, Berlin, 2006.
3) R.A. Jacobs, M.I. Jordan, S.J. Norlan and G.E. Hinton, “Adaptive Mixtures of Local Experts,” Neural Computation, vol.3, no.1, pp.79–87, Spring 1991.
4) M.I. Jordan and R.A. Jacobs, “Hierarchical Mixtures of Experts and the EM Algorithm,” Neural Computation, vol.6, no.2, pp.181–214, March 1994.
5) R.E. Schapire, “The strength of weak learnability,” Machine Learning, vol.5, pp.197–227, June 1990.
6) Y. Freund and R.E. Schapire, “A decision-theoretic generalization of on-line learning and an application to boosting,” J. Comput. Syst. Sci., vol.55, no.1, pp.119–139, August 1997.

■S3 群 - 4 編 - 1 章

1-11 階層ベイズモデリング

（執筆者：上田修功）[2009 年 3 月 受領]

統計的学習では, 観測データの背後にあるデータ生成過程を確率分布のかたちでモデル化する. ベイズモデリングでは, モデルパラメータ θ の事前分布 $p(\theta)$ を考え, データ $D = \{x_1, .., x_N\}$ を観測した後, パラメータの事後分布 $p(\theta|D)$ をベイズ則より,

$$p(\theta|D) \propto p(D|\theta)p(\theta) \tag{1·52}$$

として計算する. ここに $p(D|\theta)$ は観測データ D のモデルに対する尤度である. そして, 未知データ x については, この事後分布を用いて次式の事後予測分布

$$p(x|D) = \int p(x|\theta)p(\theta|D)d\theta \tag{1·53}$$

のかたちで予測する. 一つの値として予測したい場合は, 期待値を求めればよい.

協調競合ネットワークのように, 複数の統計モデルを想定する場合, 各統計モデルのパラメータを互いに独立とすると, ある統計モデルに対する観測データ数が少ない場合, そのパラメータの事後分布の推定値の信頼性が低くなる. この問題に対し, 階層ベイズモデリングでは, K 個の統計モデル $p(x|\theta_{(1)}), ..., p(x|\theta_{(K)})$ に対し, 各モデルパラメータの事前分布として, 共通の事前分布 $p(\theta_{(j)}|\alpha)$,$j = 1, ..., K$ を仮定する. α はハイパーパラメータで $p(\alpha)$ をその事前分布とする. そして, $\theta = (\theta_{(1)}, \ldots, \theta_{(K)})$ 及び, D_j をモデル j からの観測データとすると, 事後分布は次式で計算される.

$$p(\theta, \alpha|D) \propto \prod_{j=1}^{K} p(D_j|\theta_{(j)})p(\theta_{(j)}|\alpha)p(\alpha) \tag{1·54}$$

すなわち, 階層ベイズモデリングでは, K 個のモデルパラメータの事後分布を独立に計算するのではなく, 階層を一段加え, 上位階層においてモデル間に穏やかな拘束を導入することで事後分布推定の信頼度を向上させている.

協調競合ネットワークのような混合モデルの共通の問題として, モデル数 K の最適決定の問題がある. この問題に対する有力なアプローチとしてディリクレ過程（Dirichlet Process:DP）がある[1)]. DP は確率分布に対する分布（distribution over distributions）で, 直観的には, 分割の生成モデルといえる. 標本空間のいかなる有限個（K 個）の分割 $A_1, \ldots, A_K$ に対し, 各分割の生起確率を $P(A_1), \ldots, P(A_K)$ とするとき, K 個の確率分布の同時分布 $G = (P(A_1), \ldots, P(A_K))$ が

$$G \sim \text{Dirichlet}(G;\ \gamma G_0(A_1), \ldots,\ \gamma G_0(A_K)) \tag{1·55}$$

のように書けるとき, G を DP と呼び, 基底分布 G_0 と正のパラメータ γ を用いて $G \sim \text{DP}(\gamma, G_0)$ と表記される. つまり, G, すなわち, K 次元単体 $(P(A_1) + \cdots + P(A_K) = 1)$ 上の $P(A_1), \ldots, P(A_K)$ の同時分布は, $\{\gamma G_0(A_k)\}_{k=1}^{K}$ をパラメータとするディリクレ分布に従う.

モデルパラメータ θ が G から生成され，そのパラメータから観測データが生成されるモデルの総称をディリクレ混合（Dirichlet Process Mixture: DPM）モデルと呼ぶ．DPM モデルでは，モデルの数 K はあらかじめ固定されず，データの生成過程で確率的に増大する．原理的には無限個のモデルが生成できることから，無限混合モデルとも呼ばれる．

x_i を生成するモデルパラメータを θ_i とすると，$i = 1, 2, \ldots, i$ に対し，DP では以下のような過程で θ_i が生成される．これらの導出の詳細は文献 2) を参照．

1. $\theta_1 \sim G_0(\theta)$
2. $\theta_2 \sim \frac{1}{\gamma+1}\delta_{\theta_1}(\theta) + \frac{\gamma}{\gamma+1}G_0(\theta)$

$\vdots$

i. $\theta_i \sim \frac{1}{\gamma+i-1}\delta_{\theta_1}(\theta) + \cdots + \frac{1}{\gamma+i-1}\delta_{\theta_{i-1}}(\theta) + \frac{\gamma}{\gamma+i-1}G_0(\theta)$

表記 $x \sim p(x)$ は x が分布 $p(x)$ から生成されることを表す．また，$\delta_x(y)$ は $x = y$ のとき 1，それ以外は 0 となるディラックのデルタ関数を表す．上記過程は，θ_i が，既に生成された $\theta_1, \theta_2, \ldots, \theta_{i-1}$ から選択されるか，あるいは，事前分布 G_0 から新規に生成され，かつ，$\theta_j,\ j = 1, \ldots, i-1$ については確率 $1/(\gamma + i - 1)$ で生成され，新規の θ については確率 $\gamma/(\gamma + i - 1)$ で生成されることを意味する．$i-1$ までに生成された θ_j，$(j < i)$ の種類が $\theta_{(1)}, \ldots, \theta_{(K)}$，すなわち，異なり数が K のとき，θ_i は，各パラメータの出現個数に比例した確率で $\theta_{(i)}$ を生成するが，それだけではなく，γ に比例した確率で新規の $\theta_{(K+1)}$ を生成し得るプロセスとなっている．これまで多く出た種類のパラメータ程，以後もよく出現し，パラメータのクラスタリングが自然に定義される．つまり，DP では，データが観測されるたびに要素数が必要に応じて増える柔軟なデータ生成過程となっている．また，i が十分大きいとき，それまでに得られる θ_i の値の異なり数（クラスタ数）は，ほぼ $\log i$ のオーダで増加することが知られている．上記表記で，θ_i と $\theta_{(k)}$ の表記の意味の違いに注意．前者は x_i を生成するモデルパラメータを意味し，$\theta_i \in \{\theta_{(1)}, \ldots, \theta_{(K)}\}$（ただし，$\theta_{(k)} \neq \theta_{(l)}$（for $k \neq l$））で，$i \neq j$ でも $\theta_i = \theta_j$ となり得る．

z_i をデータ x_i に対する潜在変数（K モデルの混合モデルの場合，z_i はモデル指標に相当し，$z_i \in \{1, \ldots, K\}$）とすると，上記 DP でのパラメータのクラスタリング過程は z_i を用いて以下のように等価表現できる．すなわち，$i-1$ 番目までの潜在変数の値が決定されたとき，i 番目のモデル指標が $z_i = k$ となる確率は以下となる．下記は上記 θ_i で説明した DP と等価であることに注意されたい．

$$P(z_i = k | z_1, \ldots, z_{i-1}) = \begin{cases} m_k/(\gamma + i - 1) & \text{if } m_k > 0 \\ \gamma/(\gamma + i - 1) & \text{if } m_k = 0 \end{cases} \quad (1\cdot56)$$

m_k は第 k クラスタに属すデータ数を表し，$m_k = 0$ は k が新規クラスタに相当し，その場合，K は $K+1$ に更新されることになる．i をレストランでの i 番目の客とし，$\theta_{(j)}$ をテーブルとすると，中国のレストランでは，客は会話を楽しみたいので既に多く座っているテーブルを好むという比喩で，式 (1·56) は Chinese Restaurant Process（CRP）と呼ばれる[2)]．

CRP の重要な性質として，交換可能性（exchangability）がある．すなわち，CRP に従って生成した $z_1, \ldots, z_N$ に対する同時分布は

$$P(z_1, \ldots, z_N) = P(z_1)P(z_2|z_1) \cdots P(z_N|z_1, z_2, \ldots, z_{N-1}) \tag{1·57}$$

として計算できるが，その結果は，z_i の順序によらず同じ値となる．この交換可能性は，DPM における Gibbs サンプリング学習アルゴリズムの正当性の根拠となっている．DPM モデルでの学習アルゴリズムとは，例えば，潜在変数の事後分布推定の場合，$P(z_1, \ldots, z_N|D)$ の推定を意味する．DP では，$Z = (z_1, \ldots, z_N)$ を生成する過程ではクラスタ数が単調増加するが，観測データ D を得た後での事後分布の Gibbs サンプリング学習では，モデルの尤度項のために適切なクラスタ数に収束する．DPM の学習アルゴリズムや応用例については文献 2) を参照されたい．

DP は確率分布上の確率分布であったが，関数空間上の確率分布を定義する確率過程としてガウス過程（Gaussian Process: GP）がある．DP とは応用分野が異なるが，若干関連する話題でもあるのでその定義のみを説明する．詳細は文献 3) を参照のこと．GP とは，確率変数の集合であり，かつ，どの有限個（K 個）の結合分布も K 次元ガウス分布となっているようなものである[3)]．ガウス分布が平均値ベクトルと共分散行列で表現されたように，GP は平均値関数と共分散関数で表現される．

■参考文献

1) T.S. Ferguson, “A Bayesian analysis of some nonparametric problems,” The Annals of Statistics, vol.1, no.2, pp.209–230, March 1973.
2) 上田修功, 山田武士, “ノンパラメトリックベイズモデル,” 応用数理, vol.17, no.3, pp.196–214, Sep. 2007.
3) C.E. Rasmussen and C.K. Williams, “Gaussian Processes for Machine Learning,” MIT Press, Cambridge, 2006.