はしがき

パターン解析とはデータ内に潜む関係を発見する過程のことをいい, 構造パターン認識や統計パターン認識ともいう. パターン解析は多くの分野において中心的な話題の 1 つと考えられており, ニューラルネットワーク, 機械学習, データマイニングといった領域において研究されている. 応用事例はバイオインフォマティクスから情報検索まで実に幅広い.

本書で取り上げるカーネル法は, このパターン解析の強力なフレームワークの 1 つであり, これらすべての分野において大きな関心を集めている. カーネル法のアルゴリズムは一般のデータ型 (文字列, ベクトル, テキストなど) を入力とし, これらのデータ内において一般の関係 (ランクづけ, クラス分類, 回帰, クラスタなど) を検知する. 本書は 2 つの役割をもつ. 1 つ目は, 手元にパターン解析の具体的な目標 (バイオインフォマティクス, テキスト解析, 画像解析などのタスク) をお持ちの読者に対して, それを解くアルゴリズムとカーネルのツールキットを提供することである. 多くは MATLAB のコードも提供する. 2 つ目は, カーネル法によるパターン解析という近年急速に拡大したフレームワークを身近に学べる入門書を提供することである. 学生や研究者を対象として必要な概念や数学を必要十分に説明しながら, アルゴリズムやカーネルを読者の手元の応用事例に適用できるようにさまざまな例を示す.

本書は 3 部で構成した. 第 1 部はパターン解析やカーネル法の概念的な基礎を説明する. 入門者向けの例を解説しながら, このアプローチの主要な理論的根拠を押さえることをねらいとする. 第 2 部はカーネル法のアルゴリズムを説明する. 簡単なものから高度なものまで, 多数のアルゴリズムを取り上げる. 具体的にはカーネル部分最小2乗(PLS)法, カーネル正準相関(CCA)分析, サポー

トベクターマシン, カーネル主成分分析（PCA）などである. 第 3 部はカーネル関数を説明し, 基本的なカーネルから先進的なカーネルまで, 多数のカーネルを解説する. 再帰を用いるカーネル, HMM などの生成モデルに基づくカーネル, 動的計画法による文字列カーネル, テキストカーネルなどの詳細を取り上げる.

本書はパターン解析, 機械学習, ニューラルネットワーク, バイオインフォマティクス, テキスト分析などの分野に必携であろう.

序文

データ内に潜むパターンを検知する研究は,「科学」と同じだけの歴史をもつ. たとえばヨハネス・ケプラーの 3 つの法則である. 天文学を飛躍的に発展させたこれらの法則は言うまでもなく惑星の運動に関するものだが, そもそもはティコ・ブラーエが集めた大量の観測データの中にケプラーが法則という形で関係を発見したことによる.

データ内に潜むパターンを自動的に検知する研究は,「コンピュータ」と同じだけの歴史をもつ. 科学と工学分野において, 統計学, 機械学習, データマイニングなど, さまざまな方法が研究されてきた.

パターン解析とはデータ内に潜む関係を (自動的に) 探知して特徴化することである. パターン解析のやり方には, 統計的パターン認識と構文パターン認識 (構造化パターン認識) の 2 つの分派が存在する. 統計的パターン認識においては, 多くの場合ベクトル形式の「データ」が与えられ, クラス分類規則, 回帰関数, クラスタ構造という形で「関係」を検知する. 一方, 構文パターン認識においては, 文字列などのベクトル形式ではない「データ」が与えられ, 文法 (やこれと同等なレベルの抽象) という形でこの文字列の間に潜む「関係」(法則) を検知する.

パターン解析のアルゴリズムの自動化の歴史には 3 つの革命が存在する. 第一の革命は, ベクトル集合内に潜む線形の関係を検知する効率的なアルゴリズムの発見である. 1957 年にパーセプトロンのアルゴリズムが発表され, 1960 年代にいくつかが続いた. これらのアルゴリズムは計算量的な解析と統計的な動作の解析を行う. また, 非線形の関係を検知する方法が, 当時の主要な研究目標となった. しかし, 結果は芳しくなく, 線形アルゴリズムと同水準の効率と統計的な保証をもつ非線形アルゴリズムを開発することは実現性の

乏しい目標であることが示された.

第二の革命は 1980 年代中頃の非線形革命である. 逆伝播の多層ニューラルネットワークの考案であり, ほぼ同時期の効率的な決定木学習アルゴリズムの考案である. これら 2 つのアプローチが非線形のパターンを検知する歴史上初めてのアルゴリズムである. ヒューリスティックなアルゴリズムを用いる点と統計的な解析だという点に関しては不完全だが, この非線形の革命の意義は計り知れない. データマイニングやバイオインフォマティクスなどの学問分野は非線形なしには成立しないからである. しかし, 局所的最小に陥るという欠点や, しばしば過学習に陥るという欠点をもつ. 前者はこれらのアルゴリズムが勾配下り法や欲張り法に基づくためであり, 後者はこれらのアルゴリズムの統計的な振る舞いを捉えることが困難であることによる.

第三の革命は 1990 年代中頃のカーネルに基づく学習法 (以下, 略してカーネル法とする) の出現である. このアプローチは線形アルゴリズムの利点である効率性を失わずに非線形関係を検知できる. さらに, 統計解析法の進歩により, 高次元の特徴空間において過学習を避けながらパターンを検知することも可能である. この第 3 世代の非線形パターン解析アルゴリズムは, 計算的, 統計的, 概念的観点から線形アルゴリズムと同程度に効率的であり, 同程度に基盤が強固である. ニューラルネットワークや決定木の典型的な欠点である局所的最小や過学習は解決される. 同時に, ベクトル以外のデータ型に対しても非常に効果的に対処できる. さらに, パターン解析の別の分野 (後に述べる構文パターン解析) との関係が生まれる.

カーネル法の 1 号機はサポートベクターマシンである. サポートベクターマシンは上記の計算的な欠点や統計的な欠点を解消するクラス分類アルゴリズムである. カーネルに基づきクラス分類以外のタスクを解くアルゴリズムもまもなく開発される. 次第に明らかとなったのは, このカーネル法はパターン解析における革命であったということである. 計算効率を保証しながら厳格に理論的な解析をすることをアルゴリズム設計の動機とする, 完全に新たな技術であった.

さらに, カーネル法はパターン認識における分野間のギャップを埋める. なぜならカーネル法は, ベクトル, 文字列だけではなく, さらに複雑なオブジェクトも含めたすべてのデータ型に対して機能しうる統合フレームワークだか

らである. また, カーネル法は相関, ランクづけ, クラスタリングなど, 幅広い種類のパターンを解析できるからである.

本書はこの新たな手法として登場したカーネル法の入門書である. カーネル法の研究者はこの 10 年間にさまざまな研究成果を生み出した. 本書はこれらの研究成果を圧縮して紹介しようと試みる. パターン解析法にいくつかのクラスがあることがこの 10 年間にわかり, これらのクラスは実用ツールキットとして重要となりつつある.

本書の第 II 部では, 伝統的なクラス分類や回帰といったタスクから, ランクづけやクラスタリングという特化したタスク, さらに主成分分析(PCA)や正準相関分析(CCA)などのタスクを取り扱う. これらのパターン解析タスクにおいてさまざまな関係を同定するアルゴリズムを紹介する. さらに, 第 III 部においては, さまざまなカーネルを記述する. これらのカーネルを適用していくつかの典型的なパターン解析タスクを解く. すると, 標準的なベクトルというデータ型のみならず, 複雑な画像やテキスト文書に関係するデータ型や, さらには生物における記号列, グラフ, 文法に関係するより高度なデータ型にいたる, 幅広いデータ解析に適用できる.

カーネル法は数学者, 科学者, 工学技術者に対して新しい強力な技法を提供する. この解析法は視点の置き方次第でさまざまに活用できる. たとえばパターン解析法, 信号解析法, 構文パターン認識法, スプラインからニューラルネットワークにいたるパターン認識法, あるいはこれらの組合せである. カーネル法が提供する新たな視点は幅広く, その可能性の全貌はまだ明らかになってはいない.

著者はこのカーネル法アルゴリズムの開発の一端を担ってきた. 理論, 実装, 応用事例にさまざまな貢献をし, カーネル法を広めてきた. 著者らの *An Introduction to Support Vector Machines*[1]は, 多数の大学で教科書として使われ, また研究者の参考書としても用いられている. 著者は EU コミッションを基金とするワーキンググループである NeuroCOLT (Neural

1. 訳注: Nello Cristianini, John Shawe-Taylor. *An Introduction to Support Vector Machines: And Other Kernel-Based Learning Methods.* Cambridge University Press, 2000. (邦訳) 大北 剛 訳,『サポートベクターマシン入門』, 共立出版, 2005.

and Computational Learning, ニューラル計算学習) にも参加している. このワーキンググループは新たな研究議題を定義する重要な役割を担う. また, KerMIT (Kernel Methods for Images and Text, 画像とテキストのためのカーネル法) というプロジェクトにも参加している. このプロジェクトはテキスト解析の研究を行う.

本書に対しては, 多くの人々から議論や提案, そして詳しく啓発的なフィードバックを多数いただいた. 著者はこれらの人々に感謝したい. 特に, Gert Lanckriet, Michinari Momma (門間道也), Kristin Bennett, Tijl DeBie, Roman Rosipal, Christina Leslie, Craig Saunders, Bernhard Schölkopf, Nicolò Cesa-Bianchi, Peter Bartlett, Colin Campbell, William Noble, Prabir Burman, Jean-Philippe Vert, Michael Jordan, Manju Pai, Andrea Frome, Chris Watkins, Juho Rousu, Thore Graepel, Ralf Herbrich, David Hardoon に感謝する. また, われわれのカーネル法の開発に対する研究開発を支援してくれた EU コミッションと, イギリスの EPSRC 基金にも感謝したい.

著者の 1 人, ネロ・クリスティアニーニ (カリフォルニア大学デイビス校准教授) は, 2001〜2002 年に非常勤講師を務めたカリフォルニア大学バークレー校と Mike Jordan, および 2002 年の夏にお世話になった MIT の CBLC と Tommy Poggio に感謝する. また, 本書を執筆する理想的な環境を提供してくれたカリフォルニア大学デイビス校の統計学科にも感謝したい. 本書の多くの内容はカリフォルニア大学バークレー校とデイビス校での授業, そして, いくつかの会議におけるチュートリアルに基づいている.

著者の 1 人, ジョン・ショーテイラー (サザンプトン大学計算機科学科教授) は, 本書の執筆時に勤務していたロンドン大学ロイヤルハロウェイ校計算機科学科の同僚に感謝する.

著　者

訳者序文

本書ではカーネル法によるパターン解析が詳説される. カーネル法はカーネルのモジュール (第 III 部) と線形アルゴリズムのモジュール (第 II 部) に分割できるという視点から本書は構成されている.

第 II 部 (第 5 章から第 8 章) では, 新奇性検知 (1 クラス SVM), クラス分類, カーネル主成分分析 (PCA), カーネル部分最小2乗法 (PLS), カーネル正準相関分析 (CCA), サポートベクターマシン, オンライン学習, クラスタリング, ランクづけ, データの可視化などさまざまなアルゴリズムが説明される. さらにこれらのカーネル法のアルゴリズムを解析する道具として統計学習が導入される. これはラドマッハ複雑性を用いてパターンの安定性を議論するという形で使われる.

第 III 部 (第 9 章から第 12 章) では, 入力データにより 4 つの章に分割される. 第 9 章はベクトル, 第 10 章はテキスト (文書), 第 11 章は文字列, 第 12 章は確率 (生成モデルを考慮にいれる) を入力とする. これらと平行して, カーネルを学習する際に知っておくべき基礎事項も随時説明される. 再帰による解法, 動的計画法, トライ木による計算法, 積和の交換に関する計算法などである.「文字列」と「文字並び」の微妙な違いも容易に学ぶことができる.

本書でのさらなる特徴は, 各々のアルゴリズムが 10 行から 20 行程度の短いアルゴリズムで示される. これは自分でプログラムを組み, これらのさまざまなアルゴリズムを理解する早道となるに違いない. MATLAB については http://www.kernel-methods.net を参照されたい. また, 訳者の python コードは http://www.kyoritsu-pub.co.jp/service/service.html#122505 に置く予定である.

共立出版の石井氏には本書翻訳の機会をいただいて感謝している. また, グラベルロード (株) の伊藤氏には校正で非常にお世話になった. 本書の査読は

統計数理研究所の福水健次氏, 九州大学の畑埜晃平氏, マックスプランク研究所の津田宏治氏, IBM の鹿島久嗣氏にお願いした. python のプログラムの査読は奈良先端科学技術大学院大学の渡邊陽太郎氏にお願いした. 第一線で活躍されている研究者の方々に査読をお引受けいただき, 非常に光栄である.

大北　剛