第2版に寄せて

R で学ぶデータサイエンスのシリーズとして「マシンラーニング」を 2009 年に出版してから早くも 5 年が経過した．フリーソフト R の飛躍的な発展とともにマシンラーニングは益々脚光を浴びている．蓄積された大量のデータから最大限の情報を引き出し，予測や制御の点で優れたモデルを生み出すことの重要性が，画像処理，自動翻訳，音声認識，自動制御，テキストマイニング，医療情報などの分野で一層強く認識されるようになった．

マシンラーニングは，そうした目的を実現するための基盤となる技術の集積なので，その基本的な概念と技巧を身につけることはデータを有効に活用するための必須条件となっている．このような時代の変遷に呼応するため，本書もこのたび第 2 版を発刊する運びとなった．主な改訂点は下記のとおりである．

1. 第 3 章では旧版におけるノンパラメトリック回帰を関数データ解析へと発展させた．そこでは，ノンパラメトリック回帰の基礎的な概念を概観した後，ノンパラメトリック回帰が生み出してきた多くの応用分野の中で特に重要性の高い分野である関数データ解析の概要を，ライブラリ {fda version 2.4.0} を用いた R プログラムの例とその実行結果とともに示した．それにより，関数データ解析が幅広い応用の可能性が広がる方法であることが実感でき，ライブラリ {fda version 2.4.0} の利用には高度な知識や技巧は必要ではなく，いくつかの単純な概念を把握すれば十分であることが容易に理解できる．
2. 第 7 章では新たにライブラリ {neuralnet} を活用し，構築されるニューラルネットワークの可視化を試みた．その結果，結合荷重の推定値が直接，ネットワークの図に書き込まれるようになった．AIC や BIC もモデル文で指定できるようになった．さらに目的関数として尤度関数や最小 2 乗誤差の指定ができるようになった．
3. 第 9 章の生存時間解析を大幅に加筆・修正した．時間依存型の特殊なケースとして，多重状態 (multistate) モデルに基づくイベントヒストリー解析を取り上げた．観測開始から目標事象（死亡あるいは打切り）発生までに種々のイベント（すなわち，多重状態）が起こり，その時間と共変量の値が記録される．イベントヒストリーデータを R により解析し，読者の理解の向上を目指した．さらに，競合リスクモデルも加筆した．具体的には，周辺モデル，層別比例ハザードモデル，累積発生関数を取り上げた．累積発生関数についての解説は他書にはあまり見られない．具体例を通じて R の魅力を明示した．

2015 年 1 月

辻谷將明・竹澤邦夫

まえがき

近年におけるパソコンとインターネットの高機能化と廉価化によって，データの取得と蓄積と処理が飛躍的に容易になった．その結果，高い計算能力と洗練されたグラフ化能力を駆使して，大量のデータから有益な情報を引き出す技術が数多く開発されてきた．それらは，データに対する，整理，分類，順序づけ，位置づけ，関連づけ，図式化，グラフ化，体系化，モデル化を行う．それによって，重要なデータの抽出，データ間の関連性の析出，概念の創出，法則の発見，既存の知識の検証，データに立脚した推定・予測・制御，対策技術の立案などが実現する．そうした技術は，マシンラーニング (machine learning) と総称され，情報化社会の要である．マシンラーニングの的確で迅速な活用なしには立ちゆかない分野も多い．

本書では，マシンラーニングにおいて主要な役割を果たしている技術を取り上げる．それは，データに基づいて回帰式を作成することを目的とするものである．データの特徴を把握し実用に資するために回帰式の作成は有力な手段なので，マシンラーニングの中核の 1 つとして揺るぎない地位にある．回帰式に関する理論や手法は長い歴史をもつにもかかわらず，近年，飛躍的な進歩を遂げ，現在もその勢いを弱めていない．そのため，多様な概念や技法が多層に渡って分岐しつつ拡大し，時には混迷するという事態にある．そうした中で，実用的にも理論的にも価値の高い手法を選択し，理解を深め，洗練させ，適切に利用するためには，それぞれの手法を信頼できる視点に立脚して習得することと，主要な手法を自ら試行してその意義や特性を把握することが不可欠になる．本書において，「信頼できる視点」とは統計学的な方法論であり，「主要な手法を自ら試行」するための手段が R というソフトウエアである．

「信頼できる視点」として統計学的な方法論を採用したのは，大量のデータを用いたマシンラーニングを行う場合においては，常に統計学的な正当性を意識しなければならないからである．確率的な不確実性を伴う事象を取り扱うにあたって，統計学的な方法論が唯一の信じるに足る体系であるゆえ，マシンラーニングにおけるその重要性は明らかといえる．実用性の高いマシンラーニングを実現するためには，解決すべき問題を統計学的な視点に基づいて理解し，それぞれの手法が拠って立つ数学的な仮説の意味とその能力と限界を知り，仮説が成り立たない場合の対処法にまで考えが及ばなければならない．

「主要な手法を自ら試行」する手段として R を選択したのは，現実の問題に対処することを最終的な目的とする手法の利用において，それらのアルゴリズムを実際に機能させてその様子を実感することも重要だからである．理論的な内容が明らかになっている事柄であっても，結果の解釈や利用においては複眼的な配慮が必要になることも多い．また，アルゴリズムが進行する中で，数値がどのような様相を呈しながら結果に至るかを知ることによって，新たな側面が明らかになったり，これまでに知られていなかった問題点が浮かび上がったりする．そうした取り組みによっ

て，マシンラーニングの実務的な能力が向上する．この目的のための道具として R は最高の条件を備えている．

これらの目的意識に沿って本書は，重回帰，関数データ解析，樹形モデル，判別分析，一般化加法モデル (generalized additive models)，ニューラルネットワーク (neural network)，サポートベクターマシン (support vector machine)，生存時間解析について解説し，R を使ってこれらの手法を利用する方法を述べている．本書で取り上げた手法は，それ自体が有意義なマシンラーニングを実現するための有力な道具であるとともに，より高度な手法を理解したり構築したりする際の礎にもなる．それぞれの章の担当は，序論（辻谷），重回帰（竹澤），ノンパラメトリック回帰（竹澤），Fisher の判別分析（辻谷），一般化加法モデルによる判別（辻谷），樹形モデルと MARS（竹澤），ニューラルネットワーク（辻谷），サポートベクターマシン (SVM)（辻谷），生存時間解析（辻谷）である．本書が，マシンラーニングの各手法の統計学的な意義に対する理解を深め，R によってその意義を体感し，実用への足がかりを築くために役立つことを期待する．

本書の執筆に至る研究活動において，諸先輩，同僚諸氏から受けた影響は計り知れないものがある．この機会に心から感謝を申し上げたい．大阪府立成人病センターの左近賢人病院長，住宅金融支援機構の外山信夫氏，大塚製薬(株）の伊庭克拓氏，（株）NTT データ数理システムの中園美香氏，イーピーエス(株）の田中祐輔氏には，専門知識の提供やプログラムの開発で多大なご教示を賜った．また，本書の執筆を勧めてくださり，原稿を読んで数々のご指摘をいただいた同志社大学文化情報学部 金明哲教授に感謝の念を表したい．最後に，本書の出版に当たり共立出版の横田穂波氏と北由美子氏にはひとかたならぬお世話になった．

なお，本書に掲載されている R コマンド，正誤表などは共立出版のホームページ
http://www.kyoritsu-pub.co.jp/service/service.html#019263 からダウンロードできる．

2009 年 5 月

辻谷將明・竹澤邦夫