機械学習の数理100問シリーズ
統計的機械学習の数理100問 with R

鈴木 譲

平成32 年2 月20 日

シリーズ序文

機械学習の書籍としておびただしい数の書籍が出版されているが，ななめ読みで終わる，もしくは難しすぎて読めないものが多く，「身につける」という視点で書かれたものは非常に少ないと言ってよい。本シリーズは，100 の問題を解くという，演習というスタイルをとりながら，数式を導き，R 言語もしくは Python のソースプログラムを追い，具体的に手を動かしてみて，読者が自分のスキルにしていくことを目的としている。

各巻では，各章でまず解説があり，そのあとに問題を掲載している。解説を読んでから問題を解くこともできるが，すぐに問題から取り組む読み方もできる。その場合，数学の問題において導出の細部がわからなくても，解説に戻ればわかるようになっている。

「機械学習の数理 100 問シリーズ」は，2018 年以降に大阪大学基礎工学部情報科学科数理科学コース，大学院基礎工学研究科の講義でも使われ，また公開講座「機械学習・データ科学スプリングキャンプ」2018, 2019 でも多くの参加者に解かれ，高い評価を得ている。また，その間に改良を重ねている。講義やセミナーでフィードバックを受け，洗練されたものだけを書籍のかたちにしている。

プログラム言語も，大学やデータサイエンスで用いられている R 言語と企業や機械学習で用いられている Python の 2 種類のバージョンを出す。これも本シリーズの特徴の一つである。

本シリーズのそれぞれの書籍を読むことで，機械学習に関する知識が得られることはもちろんだが，脳裏に数学的ロジックを構築し，プログラムを構成して具体的に検証していくという，データサイエンス業界で活躍するための資質が得られる。「数理」「情報」「データ」といった人工知能時代を勝ち抜くための，必須のスキルを身につけるためにうってつけのシリーズ，それが本シリーズである。

まえがき

人工知能時代をエンジョイするのか，人工知能のエジキになるのか

2016 年に，大阪大学理学部数学科の学生を対象に（統計的）機械学習に関する講義をはじめたことが，本書を執筆するきっかけになったと思います。当時，テキストとして適当なものがなく，2018 年に日本語に翻訳された『R による統計的学習入門』*"Introduction to Statistical Learning with Application in R"* を参考に講義を組み立てました。説明が丁寧で，わかりやすく，非常に気に入り，日本語の翻訳をしてみたいと思うぐらいになりました（その時点で他の方が権利をすでに獲得していて，翻訳は実現しませんでした)。本書の扱っている単元や，章立てもその著述と似ているのも，そうしたところによっています。

2017 年に現在の所属，基礎工学部情報科学科数理科学コースに異動となりました。早速，3 年生後期の計算数理 B という科目で，2016 年と同様の内容の講義を試みました。数理科学コースは 2 年生で R 言語を習熟していて，当初は違和感なく首尾よく講義を行えました。しかし，よりどころとしていた『R による統計的学習入門』については，2018 年，2019 年というように講義を重ねるごとに，問題意識を抱くようになりました。初学者の方が機械学習の概要を把握するのには，最適な書籍とは思われますが，

1. 本質的な理解というよりは，感覚的な理解だけで十分
2. R 言語の処理のステップを見ないで，パッケージにデータを放り込めば十分

という姿勢には，どうしても合点がいきませんでした。

現在は，人工知能の時代と言われます。言うまでもありませんが，インターネットのおかげで，必要な情報を即座に得ることができ，職場の業務が効率化され，生活も豊かになりました。その一方で，人間が行ってきた業務の多くが人工知能に置き換えられるのではないかという危惧が生じています。私自身は，データサイエンスや機械学習の業務に携わる人との付き合いが多いのですが，「業務でどのような資質が求められますか」と聞くと，知識や経験というよりは，「ロジック」という答える人が多く，活躍している人ほどその傾向が強いように思われます。情報の真偽を吟味する，人が見えない本質やチャンスを見る，制約にとらわれない発想などが，「ロジック」の結果として生まれているという視点です。逆に，そういう「ロジック」が欠如していて，ヤマカンに頼るというのであれば，人工知能のエジキになる可能性が高くなるように思われます。

もっとも，そうした「ロジック」が，数学やプログラミングをやらないと身につかない，というのは真実ではないかもしれませんが，大学教員 30 年間の中で，多くの学生をみてきた経験からす

ると，それらには非常に強い相関があるように思われます。本書は，機械学習に関する知識も提供しますが，それと同時に，数学的に本質を理解して，プログラミングで処理を構成して，検証するという経験を通じて，読者の方々の「ロジック」を脳裏に構築することを，目標の一つに掲げています。

『R による統計的学習入門』は，講義を組み立てる上で大変参考にはなりましたが，人工知能時代をエンジョイするために不可欠な「ロジック」を構築するという視点が十分ではなく，どうしても本書を執筆せざるを得なかった，というのが私の正直な気持ちです。

また，数年前に日本語に翻訳された *"Elements of Statistical Learning"*『統計的学習の基礎』(共立出版）は，分量が多く，輪講して挫折したという話をよく聞きます。困ったときに参考にする百科事典として使う方が多いように思われます（私は頻繁に利用しています)。『統計的学習の基礎』を大学の講義半期に圧縮できたら，という気持ちも本書を執筆するきっかけとなりました。また，『統計的学習の基礎』は信頼できる知識を提供していますが，ソースコードをおいたり，読者が具体的にスキルを身につけるような誘導があってもよいように思いました。

本書の 100 問は，2017〜2019 年の講義で学生に課した演習問題，特に『R による統計的学習入門』を漫然と読んでいる学生にツッコミを入れるために作成したものです。数式を正しく導出し，プログラムを組んで実行結果を見るなど，手を動かして自分のものにすべきと伝えました。ただ，問題だけを与えても，優秀な学生以外は自力では解けないので，全員が課題を提出できるよう，講義で解答に限りなく近い丁寧なヒントを与えました。それが本書の本文です。問題 1〜100 と本文で重複している記述が若干あるのは，そのためです。また，講義で提供している 10〜15 分間の復習ビデオも提供しています。

本シリーズの特徴

本書というよりは，本シリーズの特徴を以下のようにまとめてみました。

1. **身につける：ロジックを構築する**
   数学で本質を把握し，プログラムで処理を構成して，データを処理していきます。読者の皆さんの脳裏に，「ロジック」を構築していきます。機械学習の知識だけではなく，視点が身につきますので，新しい機械学習の技術が出現しても追従できます。100 問を解いてから，「大変勉強になりました」と言う学生がほとんどです。

2. **お話だけで終わらない：コードがあるのですぐにコードー（行動）に移せる**
   機械学習の書籍でソースプログラムがないと，非常に不便です。また，パッケージがあっても，ソースプログラムがないとアルゴリズムの改良ができません。Git などでソースが公開されている場合もありますが，MATLAB や Python しかなかったり，十分でない場合もあります。本書では，ほとんどの処理にプログラムのコードが書かれていて，数学がわからなくても，それが何を意味するかを理解できます。

3. 使い方だけで終わらない：大学教授が書いた学術書
   パッケージの使い方，実行例ばかりからなる書籍も，よく知らない人がきっかけを掴めるなど，存在価値はありますが，手順にしたがって機械学習の処理を実行できても，どのような動作をしているかを理解できないので，満足感として限界があります。本書では，機械学習の各処理の数学的原理とそれを実現するコードを提示しているので，疑問の生じる余地がありません。本書はどちらかというと，アカデミックで本格的な書籍に属します。

4. 100 問を解く：学生からのフィードバックで改善を重ねた大学の演習問題
   本書の演習問題は，大学の講義で使われ，学生からのフィードバックで改良を重ね，選びぬかれた最適な 100 問になっています。そして，各章の本文はその解説になっていて，本文を読めば，演習問題はすべて解けるようになっています。

5. 書籍内で話が閉じている (self-contained)
   定理の証明などで，詳細は文献○○を参照してください，というように書いてあって落胆した経験はないでしょうか。よほど興味のある読者（研究者など）でない限り，その参考文献をたどって調査する人はいないと思います。本書では，外部の文献を引用するような状況を避けるように，題材の選び方を工夫しています。また，証明は平易な導出にし，難しい証明は各章末の付録においています。本書では，付録まで含めれば，すべての議論が完結しています。

6. 読者ページ：章ごとのビデオ，オンラインの質疑応答，プログラムファイル
   大学の講義では，slack で 24/365 体制で学生からの質問に回答していますが，本書では Disqus という質疑応答のシステムを採用しています。また，講義で利用した各章 10〜15 分のビデオを公開しています。また，本書にあるプログラムをすべて手で打ち込むのは難しいので，プログラムのリストのファイルをサイトにおいています。

7. 線形代数
   機械学習や統計学を学習するうえでネックになるのが，線形代数です。研究者向きのものを除くと，線形代数の知識を仮定しているものは少なく，本質に踏み込めない書籍がほとんどです。そのため，シリーズ第 1 号の『統計的機械学習 100 問 with R』と第 2 号の『同 with Python』では，第 0 章として，線形代数という章を用意しています。14 ページしかありませんが，例だけでなく，証明もすべて掲載しています。ご存知の方はスキップしていただいて結構ですが，自信のない方は休みの日を 1 日使って読まれてもよいかと思います。

本書の使い方

各章は，問題，その解説（本文），付録（証明，プログラム）からなっています。問題を解いてみて，わからないときだけ本文を読める方（上級者）もいらっしゃるでしょうが，本文から読み始めて最後に問題をとくという形でも問題ありません。最後まで読破して，身につけることを優先してください。

講義で使われる場合，第1章を3回，第4章を1〜2回，第6章を2回，それ以外の章で各1回で，合計12〜13回程度の講義が組めるかと思います。線形代数の章（第0章）に2〜3回かけて15回としてもよいですが，第1章〜第9章をやりながら適宜戻ってくるという進め方でもよいかと思います。かなりできる学生なら，最初に100問を課してもよいでしょう。本文をじっくり読めば，回答できるようになっています。輪講（担当を決めて交代で読み続けていく）でも，同じ程度の回数で読めるかと思います。

謝辞

機械学習の数理100問の執筆をご提案いただいた共立出版の皆様，特に本書の担当編集者で，シリーズ化を受け入れ，本書の出版に関して細かい点までチェックしていただいた大谷早紀氏に感謝します。また，大阪大学大学院の学生である稲岡雄介君には，提出前の原稿に目を通して，プログラムや数式の誤りを指摘してもらいました。2018〜2019年の講義の学生（大阪大学基礎工学部）からのフィードバックが，本書にとって有益な情報となったことも，特記すべきと思っています。

目　次

以下では，実数全体を $\mathbb{R}$ とかき（複素数全体は $\mathbb{C}$），$n \times m$ の実数成分の行列の集合を $\mathbb{R}^{n \times m}$，$n \times 1$ の実数成分の行列（列ベクトル）の集合を $\mathbb{R}^n$ と書くものとする。また，行列やベクトルの転置は，A^T, b^T のように右上に T を上付きで表記する。