書評

An Introduction to Statistical Learning (with Applications in R)

川野　秀一*

Gareth James,
Daniela Witten,
Trevor Hastie and
Robert Tibshirani 著
Springer
2013 年 8 月，430 pp.
価格　59.99 €
ISBN 978-1-4614-7137-0

本書は，統計的学習 (statistical learning) の入門書であり，特に，統計的学習を用いてデータ解析を行いたいと考えている人々を対象としている．統計的学習とは，古典的な統計学と機械学習を包括するような形で定義された比較的新しい言葉である．本書の特徴は，Trevor Hastie 氏達によって書かれた「The Elements of Statistical Learning」(以下 ESL と略す) の内容をもとに，細部は極力排除して，各手法が作り出された経緯や各手法の長所・短所を中心に話を進めていることである．ここで，細部というのは，推定量の性質や漸近理論はもちろんのこと，微分積分の計算や固有値計算なども含まれる．しかし，数式が無いという訳ではないため，数式アレルギーのある方々は注意して欲しい．

本書の構成，特色を簡単に述べる．まず，全 10 章より構成されている．明確には記述されていないが，線形モデルを扱った 1 章から 6 章までが第 1 部，主に非線形モデルを扱った 7 章から 10 章までが第 2 部のようである．1 章を除いた他の章では，各手法を具体的なデータ解析例を交えながら説明したのち，「Lab」という節で解説した手法を統計ソフトウェア R で実装する方法が説明されている．その後，演習問題となるが，演習問題が「Conceptual」と「Applied」に分かれており，手法部分に対する理解を深めたい場合は「Conceptual」を，実装する技術を磨きたい場合は「Applied」をと工夫がなされている．すべての「Lab」内で使われているデータセットは，R 内のパッケージ ISLR にまとめられている．さらに，http://www.StatLearning.com よりアクセス可能なサポートページには，「Lab」で用いた R のソースコード，データセットや正誤表が置かれておりかなり充実している．そして，本書の PDF 版もこのページよりダウンロード可能となっている．また，本書には参考文献の章がないことに気付く．これは内容が本書で self-contained であ

* 大阪府立大学大学院工学研究科

ることを意味しており，より発展的な内容を勉強したい場合はESLを参照して下さいということのようだ．

それでは，各章の内容をざっと見ていこう．1章では，統計的学習の重要性について様々なデータを例示しながら述べている．2章は，統計的学習の概要である．回帰・判別問題を例にして話を進め，統計的学習では予測と解釈が重要であることを述べている．特に，2.1.3節では，予測と解釈は拮抗関係にあり，後章で紹介される様々な手法がどの位置関係にあるかを図を用いて説明している．3章では，(線形) 回帰モデルを紹介している．モデル構築の考え方，回帰係数の解釈はもちろんのこと，交互作用項を導入したときの階層原理や多重共線性に対する対処法の一つであるvariance inflation factorまでも記述している．また，k近傍回帰モデルと線形回帰モデルを比較している3.5節も特徴的である．4章では，判別問題を扱っている．4.4.3節では，ROC曲線による判別手法の評価方法とともに，偽陰性率や偽陽性率の同義語が簡潔にまとめられている．5章は，リサンプリング法に基づくモデル評価方法が紹介されている．本章全体から見ると，ブートストラップ法に関する節が短く物取りなさを感じた．8章でバギング，ランダムフォレストを紹介するため，もう少し内容を増やして欲しいところである．6章では，より良い予測や解釈を求めて，変数選択，縮小推定，次元圧縮を紹介している．変数選択法から縮小推定への流れは，さすがこの分野の先駆者達だと思わせる書き方である．そして，本章の特筆すべき内容は，6.4節の高次元データ解析の注意点である．高次元データ解析に関する書籍は数多く出版されているが，このように解析を行う，または行った際に着目した記述がある書籍は珍しい．

7章では，ノンパラメトリック回帰を紹介している．7.4節から7.5節にかけての回帰スプラインと平滑化スプラインの関係性は，簡潔に書かれていてわかりやすい．8章は，樹形モデル，バギング，ランダムフォレスト，ブースティングの紹介である．まず，樹形モデルを構築するためのアルゴリズムが書かれているが，この部分が若干冗長ではないかと感じられた．また，ブースティングの紹介を無理矢理入れ込んだ感が否めない．9章は，機械学習分野で発展を遂げたサポートベクトルマシンの紹介である．内容は標準的であるが，最後にロジスティック回帰モデルと結びつけるあたりがいかにも統計学者らしい．10章は，教師なし学習法として主成分分析とクラスタリング法を紹介している．k平均法や階層的クラスタリング法を実際に行う際に重要な事柄が丁寧に記述されており，特に，階層的クラスタリング法での非類似度性についてはかなりのページを割いている．

以上，本書を紹介してきたが，第1部は丁寧に書かれている一方，第2部は駆け足気味のように感じられる．元来，第2部は発展的な内容であるため，第2部については大まかに内容を把握した後は早くESLを参考してくださいということかも知れない．最後に，本文をきっかけに本書を手に取っていただき，少しでも多くの人々が統計的学習に興味を持っていただければ幸いである．