統計的機械学習ことはじめ

データ分析のセンスを磨くケーススタディと数値例

廣野元久［著］

日科技連

ま え が き

最近のデータ分析の進歩はめざましいものがあり，AI(人工知能)や機械学習・ディープラーニングといった言葉が踊っている．ここでいう進歩とは，「実務への応用が指数関数的な速度で展開されている」という意味である．ビジネスの世界で市民権を得た機械学習であるが，中身を知りたければ理屈を学ぶ必要がある．それには数理的な素養が不可欠であり，演算プロセスをトレースできる計算環境が必要となる．成書では演算プロセスの説明に Python や R などのコマンドを使った短いプログラムが示されているものが多い．例題のプログラムを参考にしてキーボードを打てば何かしらの計算結果が返ってくる．それが正しい処理かどうかは学習しないとわからない．正しい処理であっても実務に役立つとは限らない．機械に何を実現させるかは対峙している問題に対する深い洞察と専門性が必要で，データの分析ではどのような方法を選択すれば良いかを見極めることが大切である．このため，問題を解決しようとする専門家とデータ分析の専門家，加えてソフトウェア技術者の協力が欠かせない．

機械学習では主にビッグデータを対象とするが，ビッグデータとは単にデータ量が大きい(**V**olume)だけでなく，動画や会話・文章などさまざまな種類・形式(**V**ariety)が含まれる非構造化データ・非定型的データを対象にしたものであり，さらに，日々膨大に生成・記録される時系列性・リアルタイム性のあるようなもの(**V**elocity)を指すことが多い．この性質(3V)から，ビッグデータの多くは観察データである．ビッグデータを分析する目的は，今までは管理しきれなかったために見過ごされてきた情報を記録・保管して即座に分析することで，ビジネスや社会に有用な知見を得たり，これまでにないような新たな仕組みやシステムを生み出したりして問題解決を行うことである．この目的を達成するための方法の一つが機械学習である．

機械学習では，最初にさまざまなデータから数値化する作業である**特**

徴抽出が行われることが多い．特徴抽出には最新のソフトウェア技術が使われる．数値化されたデータは**特徴量**と呼ばれ，得られた無数の特徴量からデータ分析に必要とされる複数の特徴量を選択するプロセスは**特徴量選択**と呼ばれる．

こうして選ばれた特徴量を使ってデータ分析を行う．データ分析では目的に合わせてデータの分類や予測などが行われる．機械学習で得られたモデルはアルゴリズムと呼ばれるが，アルゴリズムや推定されたパラメータの解釈は行われない．統計モデルとは異なり，得られたアルゴリズムは人が解釈できるような式の形では求まらない．

ところで，問題解決には「原因を究明して，そこに手を打ち悪い結果を発生させない」ようにする原因追求型と「現象を深く観察して結果を予測することで，悪い結果を回避する」結果予測型の活動がある．

機械学習で扱う特徴量の多くは研究対象の状態を表す結果を数値化したものであるから，そもそも機械学習は結果に影響を与える主要な要因を探し出し，要因を制御することを目的にしていない．機械学習の目的は分類や予測に役立つアルゴリズムを使って，リアルタイムに演算を行い，意思決定を支援する学習システムを構築することである．

例えば，万引き防止を考えると，万引き犯の心理を解明して万引きという行為を撲滅するのではなく，店舗に来る人の行動や表情から特徴量を見つけ，学習を繰り返して万引き犯を事前に特定して，万引きという行為を留まらせる予測システムを構築するのである．

本書は機械学習全体についての概要を説明することを目的にしていない．機械学習で利用される基本的なアルゴリズムについて，統計学，特に多変量解析の延長線上にある非線形な手法であると位置づけて，その概要と使い方を数値例やケーススタディを使って説明するものである．

執筆にあたり，読者層を QC 検定(品質管理検定)で 2 級合格を目指す技術者(実務経験があり統計学の基礎知識を有している)とした．その理由は，次世代のものづくりの中核をなす方々がもつ疑問，例えば，「機械学習で何ができるのか」「どのようなことが伝統的な統計的方法に比べて新しいのか」「統計的方法が苦手とする問題にどのように適用でき

るのか」といった点に答えることが必要と考えたからである．多くの図表(カラーの図は日科技連出版社のウェブサイトからダウンロードできる)を使って，読者が以下の①～③を達成できるよう，配慮した．

① 17 の数値例を使って機械学習の前提条件や考え方の基本を理解できる(表 1)

② 32 のケーススタディを通じて機械学習が何を叶えてくれるのかを擬似体験できる(本書ではデータファイルを《……》で，特徴量(変数)を『……』で表している)(表 1)

③ 統計的方法と機械学習の両者を比較することで各方法の特色を明らかにする

ただし，本書を執筆するにあたり意識したことが 2 つある．1 つは「Python や R などを使った計算例を示さない(成書やウェブサイトの記事に譲る)」ことであり，もう 1 つは「詳しい数理的説明をしない(研究者の成書や論文に譲る)」ことである．これらは機械学習を提供する研究者の領域であり，すでに良質な文献が多く出回っている(インターネット上には無料で読めるものさえある)からである．なお，Python や R などの使い手であれば，日科技連出版社のウェブサイトからケーススタディのデータ(Excel ファイル)をダウンロードしてデータ分析を楽しむことができる．

本書では SAS Institute の JMP Pro 15 と日本科学技術研修所の JUSE-StatWorksV5 機械学習編の 2 つを使ってデータ分析を行っている．手法名やハイパーパラメータの設定，扱える手法は 2 つのソフトウェアで異なるところがある(表 2)．詳しくはソフトウェア開発先が主催するセミナーやウェブページを参照してほしい．

最後に，早稲田大学の小島隆矢先生には草稿の段階から多くのアイデアを頂戴した．日本科学技術研修所の機械学習セミナー講師をされている立教大学の山口和範先生，青山学院大学の西垣貴央先生の貴重な資料からは多くを勉強させていただき，アドバイスも数多く頂戴した．友人であるデンソーの吉野睦氏と産業技術総合研究所の遠藤幸一氏には内容を詳細に添削していただき，多くの助言をいただいた．また，日本科学

表 1　本書のケーススタディ(32 個)と数値例(17 個)の一覧表

章	題名	章	題名
1	数値例①：ヒストグラムに潜むクラスの発見	6	数値例⑩：判別分析の弱点
	数値例②：正規混合分布を使った 2 クラス分類		数値例⑪：必要な個体だけを使った分類
	ケーススタディ①：誘発磁場の分布		ケーススタディ⑰：鮭と鱸の線形 SVM
	数値例③：時系列データに潜むパターンの発見		数値例⑫：２つの特徴量の高次元化
	ケーススタディ②：誘発磁場の傾向		ケーススタディ⑱：鮭と鱸の非線形 SVM
	数値例④：散布図に潜むクラスの発見		ケーススタディ⑲：誘発磁場の非線形 SVM
	ケーススタディ③：誘発磁場の等高線図	7	ケーススタディ⑳：統計的な見方・考え方
	ケーススタディ④：投手成績の等高線図		ケーススタディ㉑：良否判定を予測するロジスティック判別
	ケーススタディ⑤：投手成績のカラーマップ		ケーススタディ㉒：ROC 曲線を使ったカットオフ値の算出
	ケーススタディ⑥：投手成績の相関係数のカラーマップ	8	数値例⑬：2 次ロジスティック判別
	ケーススタディ⑦：誘発磁場のカラーマップ		ケーススタディ㉓：原油成分の隠れ層とノード数の設定
2	ケーススタディ⑧：杉花粉の重回帰分析		数値例⑭：グラフを見ればわかるニューロ判別境界
	ケーススタディ⑨：投手成績のホールアウト検証		ケーススタディ㉔：誘発磁場のニューロ判別
	ケーススタディ⑩：杉花粉データの正則化		ケーススタディ㉕：プリンタのローラ径のニューロ判別
3	数値例⑤：特徴量の次元圧縮		ケーススタディ㉖：カラー画像のニューロ判別
	数値例⑥：カーネル主成分によるクラス発見	9	数値例⑮：簡単な予測
	数値例⑦：重要なカーネル主成分の探索		ケーススタディ㉗：住宅価格予測のニューラルネットワーク
	ケーススタディ⑪：誘発磁場の主成分分析		ケーススタディ㉘：時系列データを予測するニューラルネットワーク
4	数値例⑧：非階層クラスター分析によるクラス発見	10	数値例⑯：木モデルを使った判別境界
	ケーススタディ⑫：誘発磁場計測の k-平均法		数値例⑰：決定木を使った予測
	数値例⑨：出現可能性を使ったクラス発見		ケーススタディ㉙：糖尿病患者の症状進行予測
	ケーススタディ⑬：死因の分類		ケーススタディ㉚：クレジットリスクを求めるランダムフォレスト
5	ケーススタディ⑭：鮭と鱸(すずき)の判別分析		ケーススタディ㉛：アルゴリズムのコンテスト
	ケーススタディ⑮：鮭と鱸の重判別分析		ケーススタディ㉜：アンサンブル学習による体脂肪率の予測
	ケーススタディ⑯：金属部品の重判別分析		——

表 2　本書の表記と使用したソフトソフトウェアの表記の対応表

章	本書の表記	JMP の表記	JUSE-StatWorks の表記
1	正規混合分布	正規混合	密度プロット
	AICc	—	—
	カーネル平滑化	カーネル平滑化	—
	等高線図	密度等高線図	等高線図
	メッシュプロット	メッシュプロット	—
	カラーマップ	カラーマップ	—
2	重回帰分析	標準最小 2 乗 ステップワイズ法	重回帰分析
	予測判定グラフ	予測値と実測値のプロット	予測判定グラフ
	評価(用)データ	テスト(用)データ	評価(用)データ
	正則化回帰分析	一般化回帰	正則化回帰
3	主成分分析	主成分分析	主成分分析
	カーネル主成分分析	—	カーネル主成分分析
4	*k*-平均法	*k* means クラスター分析	*k*-means 法 (マハラノビス汎距離)
	正規混合法	正規混合	混合ガウス分布
5	判別分析	判別分析	判別分析
6	サポートベクターマシン	サポートベクトルマシン	サポートベクターマシン
7	ロジスティック判別	名義ロジスティック	ロジスティック回帰分析
	混同行列	混同行列	誤判別表
	的中率	—	再現率
	感度/特異度	混同行列	適合率
8	ニューロ判別	ニューラル	—
9	ニューラルネット	ニューラル	—
10	決定木分析	パーティション[1)] (G^2を表示)	多段層別分析
	ランダムフォレスト	ブートストラップ森	ランダムフォレスト (誤分類率を表示)

1)　パーティションは厳密には CHAID のアルゴリズムではない.

技術研究所の犬伏秀生氏と SAS Institute Japan の岡田雅一氏にはソフトウェアを提供する立場からソフトウェアの使い方などを指導していただいた．ここでは名前を挙げなかったが，筆者と縁ある多くの方々の叱咤激励がなければ到底，完成に辿り着くことはできなかったと思う．心より感謝を申し上げたい．

出版に際しては，予定より大幅に遅れても励まし続けていただいた日科技連出版社の田中延志氏にこの場を借りて御礼を申し上げたい．また，PC を前に渋い顔をして原稿を書いている筆者を終始和ませてくれた妻，峰子にはいつもながら頭の下がる思いである．

本書は伝統的な統計的データ分析の視点から対岸にある機械学習を眺めた内容になっているが，統計的品質管理と機械学習の架け橋になればという思いとともに読者が機械学習の知識を知恵に変えて，ビジネスに活用していただく一助となれば幸いである．

2021 年 8 月　コロナ禍の中だからこそ，愛で世界が満たされますように

廣野　元久

目　次

コラム一覧

第6章　サポートベクターマシン

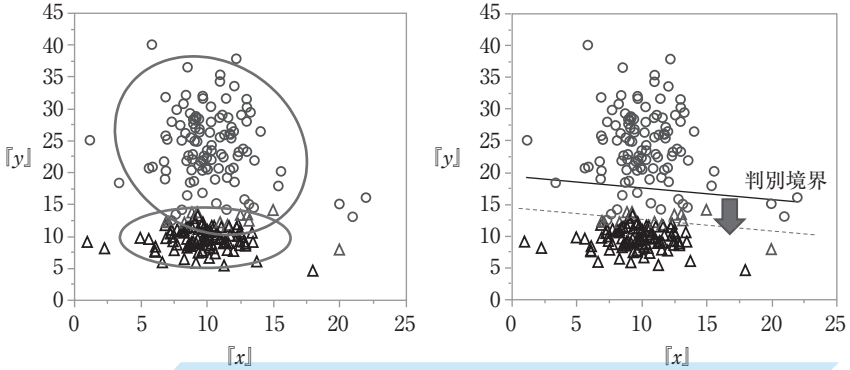
判別の目的は境界を決めることでもあるから，境界付近にある点を重視し，境界から遠い点を無視するほうが都合のよい場合がある．その要求に答えてくれるのが本章で紹介する**サポートベクターマシン**(**SVM**)である．SVM は機械学習の主要なアルゴリズムの１つで，線形な方法とカーネル法を用いる非線形な方法とがある．

6.1　線形 SVM の基礎

判別分析ではすべての個体を使って判別境界を求めている．その際，観測値は多変量正規分布に従う確率変数であること，また，各群の母分散と母共分散は等しいという前提にもとづく．この前提はある意味で行儀が良すぎるものである．**線形 SVM** は判別に役立つ個体を数理的に選んで，少数の観測点(**サポートベクター**)で判別境界を定める方法である．

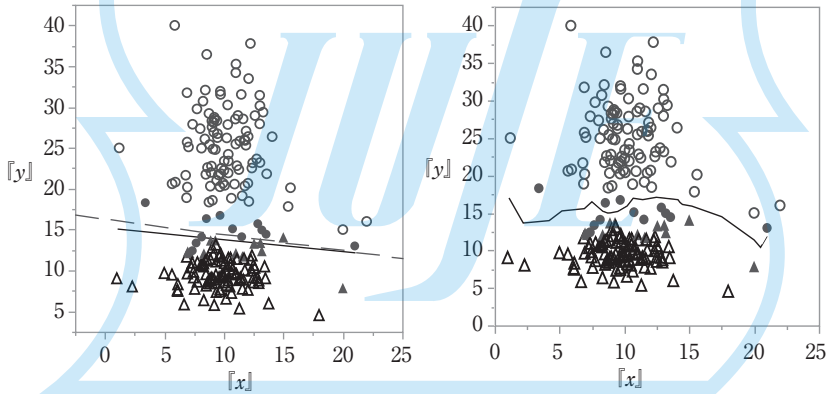
6.1.1　【数値例⑩：判別分析の弱点】

データの背後に正規性・等分散共分散性といった統計仮説が成り立っている条件で判別分析が行われる．もし，その仮説が崩れた場合にはどのようなことが起きるのかを数値例で確認する．**図 6.1**-Ⓐは《判別分析の弱点》のデータファイルから『x』と『y』を『クラス』で分けた確率楕円を追記した層別散布図である．『クラス』は C1 がマーカーの△で，C2 がマーカーの○で識別されている．２つの確率の様子(長軸の方向や楕円の短軸方向の幅)から，「２つのクラスに等分散共分散性が成り立っている」とグラフからは感じられない．統計仮説が崩れている状況で判別境界を求めると何が起こるであろうか．判別分析で得られた判別境界を追記したグラフを**図 6.1**-Ⓑに示す．C1 を C2 と誤分類することはないが，C2 を C1 と誤分類することが多い結果となった．グラフ



Ⓐクラスで層別した確率楕円　　Ⓑ2群の判別境界線の追記

図6.1　《判別分析の弱点》の『x』と『y』での判別(△はC1，○はC2を表す)



Ⓐ恣意的な境界と線形SVMの境界　　Ⓑ非線形SVMの境界

図6.2　線形判別分析とは異なる方法で求めた判別境界

から直感的に，「判別境界線を垂直方向にもっと下げたほうがよい」と感じるであろう．

そこで，主観的に境界に近い観測点だけを使って判別分析を行い，新たな境界線を引くことを考える．**図6.2**-Ⓐの実線は●と▲以外の観測点の重みを0にした場合の判別境界である．すべての観測点を使った判別境界線よりもよい判別結果が得られている．しかし，このような恣意的な方法は分析者の主観が色濃く反映されるので，客観性に乏しく好ま

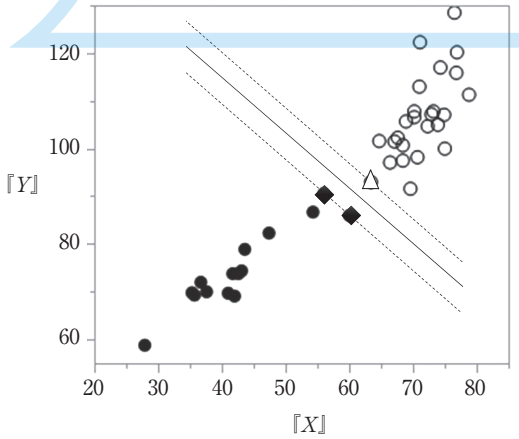
しい方法とはいえない．

ところで，図中の破線は線形 SVM により求めた判別境界線である．こちらの直線は客観的な基準で選定された観測点を使って得られたものである．また，**図 6.2**-Ⓑの複雑な境界は非線形 SVM により求めた判別境界線である．非線形 SVM のほうがよい判別ができているように感じられるであろう．以降では，判別境界を求めるためにどのような基準で観測点が選ばれるのか，および，SVM の概念を説明する．

6.1.2　【数値例⑪：必要な個体だけを使った分類】

SVM は広く使われている学習アルゴリズムの 1 つである．判別分析は，すべての個体の情報を使って判別境界を計算するが，SVM では分類に必要な個体の情報(分類をサポートするベクトルとして)を使って判別境界を計算する．また，カーネル法と組み合わせることで，非線形な判別を行うことができ，高い判別性能が得られる．

図 6.3 は《SVM》の『X』『Y』に線形な判別モデルをあてはめた結果を示したものである．SVM では判別境界をサポートするベクトルは境界に面した個体を利用する．《SVM》では，クラス 1 から 2 つ(◆のマーカー)とクラス 2 から 1 つ(◇のマーカー)を使って判別境界を作る．



クラス	C1	C2
個体数n	15	25
サポートベクター数	2	1
マージン上の数	2	0
マージン内の数	0	1
誤分類数	0	0
カーネル関数	線形カーネル	
正則化(コスト)パラメータc	10.0	

図 6.3　《SVM》の線形な判別境界

図中の実線は判別境界を，破線はマージンを加えた境界線を表している．

6.1.3　SVM の特徴

判別分析も SVM も判別関数の符号で，どちらのクラスに所属するかを決定する．判別分析では**図 6.4**-Ⓐのように正規分布を仮定しており，マハラノビス距離を基準に分類する．判別境界は各クラスの平均からのマハラノビス距離が等しい点になる．SVM では**図 6.4**-Ⓑのように判別境界は各クラスの 1 番近い個体からできるだけ距離をとるように作られる．これを**マージン最大化**という．この判別境界線に 1 番近い観測点を**サポートベクター**という．この名前の由来は，観測点が境界を支える(サポートしている)ベクトルであることから来ている．**図 6.4** はサポートベクターの数が 2 つの例を示している．一般に，サポートベクターの数が少ないと小さな次元で分類できており，判別境界は単純である．一方，サポートベクターの数が多いと誤分類の数が見た目で少なくなるが，高い次元で分類されているので判別境界が複雑になる．サポートベクターの数が必要以上に多い場合は，モデルが**過学習**している可能性が高く，汎化性能が劣っている可能性がある．また，SVM はモデルの再構築が容易であることが知られている．それは判別境界に近い個体と新しい個体だけを使ってモデルの再構築を行うからである．

ところで，SVM にも弱点がある．それは，計算時間が個体数 n に大きく依存することである．後述する**カーネルトリック**によって特徴量の

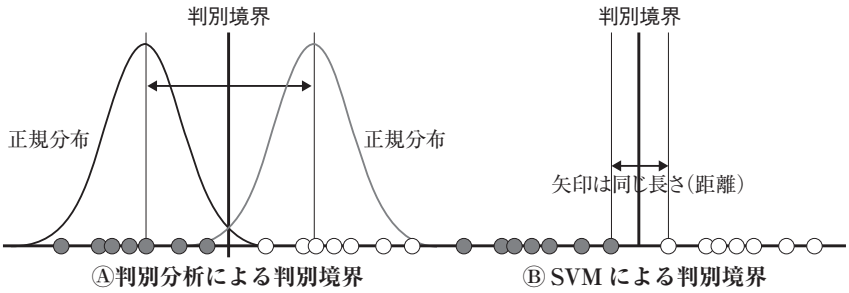


図 6.4　2 つの判別境界の考え方の比較

●著者紹介

廣野元久（ひろの　もとひさ）

1984 年，㈱リコー入社．以来，社内の品質マネジメント・信頼性管理の業務，SQC の啓蒙普及に従事，品質本部 QM 推進室長，NA 事業部 SF 事業センター所長を経て，現在，㈱リコー 倫理審査委員会委員．

東京理科大学工学部経営工学科　非常勤講師（1997～1998 年），慶應義塾大学総合政策学部　非常勤講師（2000～2004 年）．

主な専門分野は SQC，信頼性工学．主著に『グラフィカルモデリングの実際』（共著，日科技連出版社，1999 年），『JMP による多変量データ活用術』（海文堂出版，2004 年），『SEM 因果分析入門』（共著，日科技連出版社，2011 年），『アンスコム的な数値例で学ぶ統計的方法 23 講』（共著，日科技連出版社，2013 年），『目からウロコの統計学』（日科技連出版社，2017 年），『JMP による技術者のための多変量解析』（日本規格協会，2018 年），『目からウロコの多変量解析』（日科技連出版社，2019 年）など．

統計的機械学習ことはじめ
データ分析のセンスを磨くケーススタディと数値例

2021 年 9 月28日　第 1 刷発行

著　者　廣 野 元 久
発行人　戸 羽 節 文

発行所　株式会社 日科技連出版社
〒151−0051　東京都渋谷区千駄ヶ谷 5−15−5
DS ビル
電　話　出版　03−5379−1244
営業　03−5379−1238

検　印
省　略

Printed in Japan

印刷・製本　東港出版印刷㈱

ISBN 978−4−8171−9740−5
URL　https://www.juse-p.co.jp/