まえがき

異常検知と変化検知は，統計学においてほとんど一世紀近い歴史を持つ伝統ある分野です．しかしここ数年，機械学習の技術がデータ解析の現場に浸透するにつれ，その様相が一変しつつあるように感じています．それはまるで，仮説検定の理論に象徴されるような厳粛な学問の世界から，実データの荒波にもまれることを喜びとするような自由で活気ある世界に理論が解放されたかのようです．

本書の目的は，統計学における伝統的な仮説検定の枠にとらわれず，最新の機械学習の技術に基づいて，異常検知・変化検知の実用的な技術を体系的に解説することです．データ解析の現場においては，異常検知と変化検知は実用上極めて重要な位置を占めていますが，実務家の問題意識に役立つような資料は著者らの知る限り非常に乏しく，各問題に対してばらばらの技術が個別的に適用され，あるときはうまくいき，あるときはそうではない，というような知見が蓄積されている現状だと思います．そのような問題意識から，著者のひとりは最近，現代的観点に基づく異常検知の入門書を出版しました [*1]．本書はその続編として位置づけられます．

本書では，前著に比べてより発展的な内容を体系的に解説しています．実数値が観測される状況においては，前著とあわせて読むことで，現在知られている異常検知・変化検知の手法の大半がカバーされることを期待しています．本書の構成を図 1 に描きました．読み方としては，1 章で本書の前提事項をまとめていますので，まずこれを通読していただければと思います．加えて，次の 2 章と 3 章にざっと目を通すと，1 章で抽象的に述べた考え方がどう具体化されるのかイメージがつかめるかと思います．その上で，必要に応じて他の章に目を通せばよいと思います．各章は 10 ページほどで短く，なるべく独立に読めるように書きました．各章の冒頭には，これから何を説

*1 井手剛，入門 機械学習による異常検知 —— R による実践ガイド，コロナ社，2015 年.

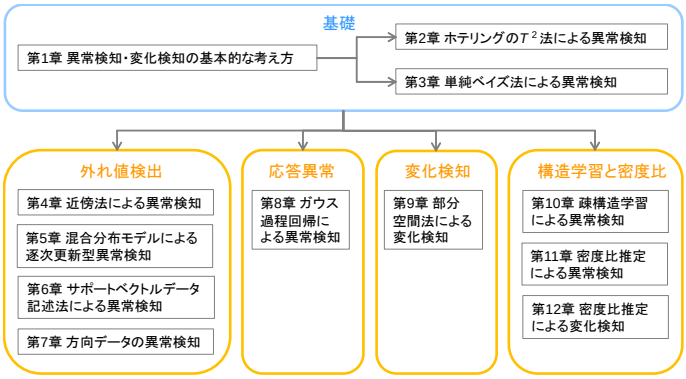

図 1　本書の構成.

明するかについての簡潔なまとめが書かれていますので，どの章が自分の目的意識に合うかを把握するためにご活用くだされればと思います．

執筆にあたっては，理論のための理論を他人事のように語るのではなく，異常検知と変化検知に実世界で従事する当事者の観点から，心を込めて説明するように努力しました．その結果として本書が，最大事後確率推定，計量学習，逐次更新型学習，非線形回帰，グラフィカルモデル，密度比の直接推定，などなど，機械学習の現代的諸概念についての，実世界でデータ解析に立ち向かう人々の観点から見た有用なテキストになっていることを期待しています．また，たとえば二値分類問題と異常検知問題の違いなど，基礎に属するにも関わらず研究者にも実務家にもあまり知られていない内容を体系的に伝える有用なテキストとなることを期待しています．

本書は第 1 章から 10 章までは井手が，第 11 章と 12 章を杉山が担当し，全体の調整は井手が行いました．本書の執筆に際し，京都大学 加納学 先生と，大阪大学 河原吉伸 先生からは内容に関する本質的なコメントを多く頂きました．IBM 東京基礎研究所の吉田一星氏からも記述の改善に関するコメントを頂きました．この場を借りて御礼申し上げます．

2015 年 4 月
井手 剛・杉山 将

目　次

Chapter 1

異常検知・変化検知の基本的な考え方

この章では，異常検知と変化検知の問題設定をまとめます．特に，確率分布を使って，異常の度合い・変化の度合いがどのように一般的に表せるのかを説明します．

1.1 知識の要約としての確率分布

あらゆるビジネスの現場で，変化あるいは異常の兆候を捉えることは大変重要な課題です．売り上げの変化を捉えることでいち早く次の一手を打てるかもしれませんし，稼働中の化学プラントの異常の兆候を見つけることで，重大な事故を事前に避けられるかもしれません．現場の職人芸に頼らず，客観的にこういうことを行いたいというのは大昔からある問題意識です．伝統的にはこの問題は，過去の事例を「ルール」という形で蓄えることで対処されてきました．たとえばこういう感じです．

IF (気温 $\geq 28°C$) AND (湿度 $\geq 75\%$) THEN 不快.

取得できるデータ数があまり多くなくて，データの性質について十分な知識がある場合は，このように手作業的にルールを作っても十分だと思いますが，実用上のほとんどの場合，人間の経験を直接ルール化するというようなやり方はうまくいきません．なぜなら，人間が明示的に認識できるルールの数は，

実世界の多様性に比べて桁違いに乏しいからです．この点，すなわち「誰がルールを作ってくれるのか」という問題は，人工知能研究の長い歴史の中で，**知識獲得のボトルネック**（knowledge acquisition bottleneck）と特に呼ばれています．

かつての人工知能研究はこの問題に対して有効な方策を提示できず，そのため長い冬の時代を過ごすことになりました．しかし最近，統計的機械学習の技術を使うことで，実用的な要請に十分耐える異常検知・変化検知の仕組みを構築できるようになってきました．従来の人工知能の（あるいは現代のソフトウェア工学の）考え方との根本的な違いは，人間の言語に近い形で知識を記述することをあきらめて，確率の言葉で知識を表現するということです．もう少し具体的に言えば，観測量を $\boldsymbol{x}$ と表したとき [*1]，$\boldsymbol{x}$ の取りうる値についての確率分布 $p(\boldsymbol{x})$ を使って数式で異常や変化の条件を記述するというやり方です．

数式が苦手な人，確率・統計が苦手な人は「変化とか異常とかわかりきったことを見るために，なぜ p とか $\boldsymbol{x}$ とかの面倒で難しい数式が必要なのか」と怪訝に思うかもしれません．しかしそれは，何十年にもわたる努力の結果，そうするのが実用的には最善であると人類が到達した現時点での結論なのです．

1.2 異常検知と変化検知のいろいろな問題

確率分布の話に入る前に，本書で取り扱う問題について雰囲気をつかんでおきましょう．例として，図 1.1 に 1 変数の時系列データにおけるいくつかの典型的な異常パターンを挙げます [*2]．赤で異常個所を示しています．上の段の 2 つは「仲間から値が外れている」というタイプの異常で，この手の異常標本を見つける問題を**外れ値検出**（outlier detection）と呼びます．右上は横軸の順序をシャッフルすると検知できなくなりますから，時系列的な外れ値ということができるでしょう．

∗1 本書では $\boldsymbol{x}$ のような太字のイタリックで列ベクトルを表します．行列は A のようなサンセリフ体を使い区別します．

∗2 心電図データは，Keogh ら [10] により研究されたもので，2015 年 1 月の時点で http://www.cs.ucr.edu/~eamonn/discords/ からダウンロードできます．

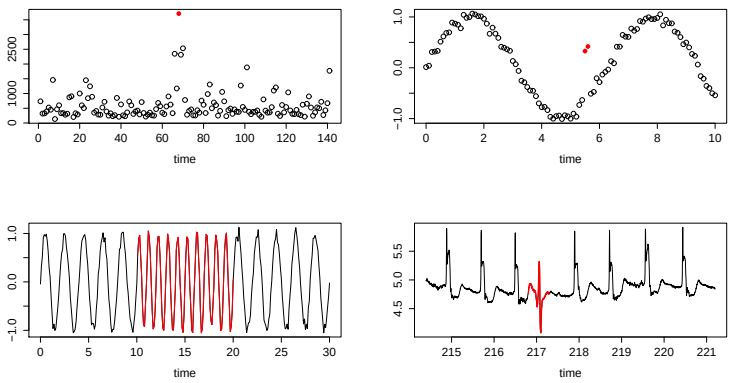

図 1.1 時系列データのさまざまな異常の例．上段: 外れ値（左），時系列的外れ値（右）．下段: 変化点（左: 周波数変化データ），変化点または異常部位（右: 心電図データ）．

一方，下の段は，値がずれているというよりは，観測値のふるまいが変化したタイプの異常です．周波数変化データでは，横軸が 10 と 20 のところで周波数の変化が生じています．これらの変化を見つける問題を**変化検知**（change detection）または**変化点検知**（change-point detection）と呼びます．心電図データの異常は，外れ値と変化点が同時に起きているとも見ることができます．見方を変えれば，異常を呈している部位を見つける問題とも取れますので，これを**異常部位検出**（discord discovery）と呼ぶこともあります．

図 1.1 のような物理的な数値データ以外にも，異常検知，変化検知の問題は定義できます．たとえば，有名なものとしては，スパムメール（広告メール）の判定問題があります（第 3 章参照）．この場合，どの語がいくつ出てきたかという長い数値ベクトルを定義します．図 1.1 のようなグラフには描きにくくなりますが，確率分布を考えることにより理論上は同じ枠組みで取り扱うことができます．

統計的機械学習に基づく異常検知・変化検知の問題は，データの性質に応じて確率分布をどう学習するか（「データから求める」ことを機械学習の用語では「**学習する**（learn）」と言います），そして異常ないし変化の度合いをど

のように確率分布と結びつけて定義するかが定式化の重要なポイントとなります．次章以降で説明するように，確率分布の学習法に応じてさまざまな異常検知手法が考えられます．

1.3 異常や変化の度合いを確率分布で表す

さて，何らかの観測量に対する確率分布が求まったとして，異常ないし変化の度合いを定量的に表すためにはどうすべきでしょうか．2 つの場合に分けて一般的な枠組みを与えます．

1.3.1 ラベルつきデータの場合

まず考えるのは，異常判定モデルを構築するためのデータとして，M 次元ベクトル $\boldsymbol{x}$ に加えて，異常か正常か（または変化点かそうでないか）を示すラベル y が同時に観測されている場合です．この場合，N 個の標本を含む訓練データとして

$$\mathcal{D} = \{(\boldsymbol{x}^{(1)}, y^{(1)}), (\boldsymbol{x}^{(2)}, y^{(2)}), \ldots, (\boldsymbol{x}^{(N)}, y^{(N)})\} \tag{1.1}$$

のようなものが観測されると想定されます．$y^{(n)}$ は n 番目の標本のラベルで，慣例に従い異常な場合は 1，正常な場合は 0 という値をとると考えます．たとえば身長と体重の値を，1 クラスの 50 人にわたって計測した健康診断データがあったとすれば，$M = 2, N = 50$ となります．$\boldsymbol{x}^{(n)}$ はたとえば出席番号 n 番の人の身長と体重を組にしたもの，$y^{(n)}$ はその人が病気かどうかを表すフラグです．

この場合，異常か正常かにより異なる確率分布，すなわち，ラベル y を与えたときの**条件つき分布**（conditional distribution）$p(\boldsymbol{x} \mid y)$ を考えるのが自然です．今，次章以降で説明する何らかの方法でこの条件つき分布 $p(\boldsymbol{x} \mid y, \mathcal{D})$ を求めたとします [*3]．$p(\boldsymbol{x} \mid y = 0, \mathcal{D})$ よりも $p(\boldsymbol{x} \mid y = 1, \mathcal{D})$ が優勢であれば異常（もしくは変化あり）と判定することになるので，次のように**異常度**（anomaly score）を定義することができます（図 1.2）．

---

*3 $p(\cdot \mid \cdot)$ は条件つき分布を表す記法です．y は条件を決める確率変数ですが，$\mathcal{D}$ の方は「この分布はデータ $\mathcal{D}$ に依存して決められる」という気持ちを表すために入れています．

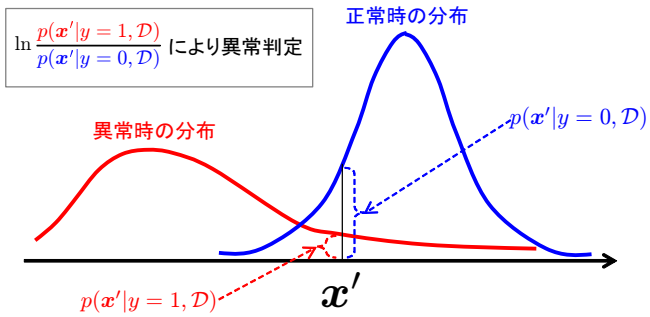

図 1.2 ラベルつきデータについての異常判定の説明．観測値 x' に対する確率密度を双方のクラスについて求め，それを比較する．比較するのが y についての分布ではないことに注意．

$$a(\boldsymbol{x}') = \ln \frac{p(\boldsymbol{x}' \mid y=1,\mathcal{D})}{p(\boldsymbol{x}' \mid y=0,\mathcal{D})} \tag{1.2}$$

ln は自然対数です．上記の異常度に定数を加えても，あるいは単調増加関数で変換しても異常度としての役目は果たすので，自然対数を使うかどうかには任意性があります．たとえば，2 章で論ずるホテリングの T^2 法から派生したマハラノビス・タグチ法という手法では，上記の異常度をある意味で常用対数により変換した量が異常の指標として使われます [33]．

上記の定義で本質的なのは，確率分布の比，すなわち，**密度比**（density ratio）もしくは**尤度比**（likelihood ratio）で異常度を定義するという点です．は 上記の異常度による判別規則を本書では**ネイマン・ピアソン決定則**（Neyman-Pearson decision rule）と呼びます．改めて書くと次の通りです．

定義 1.1（ネイマン・ピアソン決定則）

$\ln \frac{p(\boldsymbol{x}' \mid y=1,\mathcal{D})}{p(\boldsymbol{x}' \mid y=0,\mathcal{D})}$ が所定の閾値を越えたら $y=1$ と判定．

上記において「所定の閾値」の正確な意味は次節で説明します．実はこのネイマン・ピアソン決定則は，次節で定義する性能指標に照らして，最善の判