

(ひとりで学べる) 実践 R ケモ・マテリアル・データサイエンス

～付録 R スクリプト付き～

Practical Chemo&Material Data Science with R Script

- ▶ 奈良先端大 計算システムズ生物学研究室 (金谷研究室) メンバーによる執筆!
- ▶ 金谷研究室で作成した、自由に使える R スクリプト付き!
- ▶ 機械学習・統計学をいかに習得するかに焦点を当て、化学・マテリアル化学の具体的データを取り入れ、R 言語におけるプログラム例を基に解説!
- ▶ 公開された R パッケージを活用するための 4 つのステップに沿って説明!
- ▶ 一連の流れをひとりでも学べるように編集! 研究室に 1 冊必携の書!

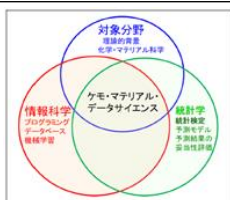
<発行要項>

- 発行: 2020 年 12 月 30 日
- 著者: 金谷 重彦 他
- 定価: 110,000 円 (税込) (付録 CD 付)
※CD 中の書籍 PDF は著作権で保護
- 体裁: A4 判・並製・308 頁・カラー
- 編集・発行: (株)シーエムシー・リサーチ
- ISBN 978-4-904482-95-7

= 刊行にあたって =

第1のパラダイムでは「仮説検定」の発想で経験科学が進化した。そして、ニュートンの法則に代表される第2のパラダイムでは、定量性を考慮した理論科学、計算機の発明により第3のパラダイムとして複雑な現象をシミュレーションに再現する科学へ、そして、現在、第4のパラダイム、として豊富なデータを活用したビッグ・データ・サイエンスが生まれた。ここでは、さらに社会実装も考慮されるようになった。一方、2012年、ハーバード・ビジネス・レビュー誌がデータサイエンスを「21世紀で最もカッコいい仕事」と位置づけたことから、注目を集めるようになった。では、これを具体的に化学・マテリアル科学の領域で進めるにはどうしたいだろうか。実践として必要とされることは、ターゲットとする分野の知識、プログラミングとして必要とされるデータ解析技術としては、データの収集、機械学習と解析結果の評価法(統計学)である。いままで、これらは地道に情報科学の各分野が基盤技術を確立してきた。しかし、オープンサイエンスの時代になり、これらのプログラムおよびデータについてもデータベースとして公開されるようになってきた。ではユーザーとしてこれらのプログラムとデータを活用し、新たな知見を獲得し、さらに社会実装することを目指すことになる。社会実装というと大きさに聞こえるが、企業であれば新たな製品を開発するということへつなげることであり、結局のところ、いま世の中に定着し始めた Sustainable Development Goals (SDGs) に向けた取り組みということへ帰着する。

本書では、このような背景を考慮しながらも、まず「機械学習、統計学をいかに実践的に習得するか?」に焦点を当て、化学・マテリアル科学の具体的データを取り入れ、R 言語におけるプログラム例(約 75 スクリプト)をもとに解説した。R 言語にはさまざまな解析用途に応じたパッケージが公開されている。そこで、本書では、< 1. データの入力 >、ファイルからのデータを入力、< 2. データの整形 >、目的に応じて入力データから必要な項目の抽出、< 3. データ解析 >、抽出されたデータを目的にあった関数・パッケージ(統計、多変量解析、機械学習などの関数)に入力し、解析結果を得る。< 4. 解析結果の表示・出力 > 解析結果をもとにグラフに表示する、あるいは、ファイルへ出力する。という4つのステップに沿って、R スクリプトを作成することを説明した。特に、データの整形について本書全体を通して説明を加えた。また、さまざまなパッケージの活用も習得できるように配慮した。エディタ RStudio、R のインストール、プログラミングの基礎、化学構造からの特徴表現(分子記述子)による多変量データ解析(機械学習により回帰モデル、分類モデル)、妥当性・汎化性能評価という一連の流れを、ひとりでも学べるように本テキスト「(ひとりで学べる) 実践 R ケモ・マテリアル・データサイエンス～付録 R スクリプト付き～」を作成した。また、(株)シーエムシー・リサーチでは、講習会も企画しているそうであるので、これらもご活用いただくと、さらに理解が深まり実践的活用への自信もつくと思う。 金谷 重彦



(c) ファイルから読み込む?

このようにして、列を指定することにより 4 検定を実行できる。ところで、DataBeforeAver2.csv は、行数が同じデータであるので問題なく実行できた。しかし、行数が同じでない 2 つのデータについては、一方に空白行ができてしまう。そこで、「CSV プログラムに直接データを読み込む」では、サンプルサイズが異なる場合でも読めるようにスクリプトを改良しよう(01test14.R)。

スクリプト名: 01test14.R

(入力ファイル: DataBeforeAver2.csv)

```
l1=scan("<file>DataBeforeAver2.csv")
df1=read.csv(l1,header=TRUE,as.is=TRUE)
df2=scan("<file>DataBeforeAver2.csv")
```

第 4-3 集合演算に関する関数

関数	説明
setdiff(A, B)	集合 A の中で集合 B と異なるものを列挙する。
union(A, B)	集合 A と B の和集合
intersect(A, B)	集合 A と B の積集合
setdiff(A, B)	集合 A について集合 B と異なる要素からなるベクトルをつくる。集合 A の要素について、集合 B に含まれる要素には TRUE、含まれない要素には FALSE を返す。
A[A%in%B]	集合 A について、A%in%B が TRUE の要素からなるベクトルを作成する。

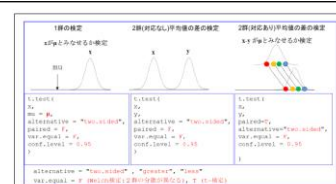


図 5.3 3 種の 1 検定

付録 CD ご利用にあたって: 付録 CD には、書籍内容の PDF と金谷研で開発した R スクリプトのフォルダが入っています。R スクリプトは著作権フリーですが、書籍の PDF は著作権で保護されておりますので、ご購入いただきました研究室や部署・部門の範囲内のご利用に限定させていただきます。無断転載・シェアは固くお断りいたします。

注文書		メルマガ会員の登録	登録済み / 登録希望	お申込み・お問合せ
品名	(ひとりで学べる) 実践 R ケモ・マテリアル・データサイエンス	価格	書籍(付録 CD 付): 100,000 円(税込 110,000 円) ※メルマガ会員は定価の 10%OFF	
会社名		TEL		
部課名		FAX		
お名前		E-mail		
住所	〒			

編集発行:
(株)シーエムシー・リサーチ
101-0054
東京都千代田区神田錦町
2-7 東和錦町ビル 3F

TEL: 03 (3293) 7053
FAX: 03 (3291) 5789
URL: https://cmcre.com
E-mail: re@cmcre.com

* 書籍はご注文を受けた翌営業日に納品書・請求書とともに送付します。* お支払いは請求書指定口座に納品日の翌月末日までに振り込みをお願いします。

第1章 ケモ・マテリアル・データサイエンス

第I部 プログラミング基礎編

第2章 RStudioの使い方

- 2.1 インストール
- 2.2 RStudioの使い方
 - (a) RStudioの立ち上げ (b) 新規スクリプトの作成
 - (c) プログラムの作成と実行 (d) Rスクリプトの保存

第3章 Rプログラミング入門

- 3.1 はじめに
- 3.2 t検定によるプログラミングの例
 - (a) プログラムに直接データを組み込む
 - (b) ファイルから読み込む1 (c) ファイルから読み込む2
- 3.3 ファイルの入出力
- 3.4 typeof()とclass()

第4章 データ構造

- 4.1 ベクトル
 - (a) ベクトルの定義 (b) ベクトルの長さ、並べ替えなど
 - (c) 集合にかかわる演算
- 4.2 リスト
 - (a) リストの定義 (b) 空リストの作り方とリストの名前、要素の名前のつけ方、呼び出し方
- 4.3 行列
 - (a) 行列の定義 (b) 行列の行と列に名前をつける
 - (c) 空行列を作成する
 - (d) データ解析で役に立つデータ成型法
 - (d1) NA(欠落値)を含む行を削除する (d2) NA(欠落値)を含む列を削除する (d3) 同一の要素からなる行の重複を削除する
- 4.4 apply()系関数
 - (a) apply 関数の使い方
 - (b) 同一の値のみからなる行、あるいは列を削除したい。
 - (c) 行ごとにパーセントに変換する
 - (d) tapply() 関数によるデータの分

第II部 データマイニング入門

第5章 統計検定

- 5.1 統計検定とは 5.2 正規分布との適合性
- 5.3 パラメトリック統計学
 - 5.3.1 2群の平均値の差の検定
 - (a) t検定(Welch検定を含む)
 - (b) t.test()
 - (b1) 1群の検定 (b2) 対応がない2群の平均値の差の検定
 - (b3) 対応がある2群の平均値の差の検定
 - (c) ボックスプロット
 - 5.3.2 分割表の統計学
 - (a) 統計学でいう複数の因子が独立とは (b) χ^2 独立性の検定
 - 5.3.3 分散分析
 - (a) 2群のグループの等分散性の検定
 - (b) 一元配置の分散分析(one-way analysis of variance, one-way ANOVA)
 - (c) 多群の検定(Turkey-Kramer検定) (d) 確率プロット
 - (e) 分散分析: 二元配置
- 5.4 ノンパラメトリック検定法
 - 5.4.1 2群の順位の検定
 - (a) Wilcoxon符号つき順位和検定: 対応がある2群の検定
 - (b) ウィルコクソン順位和検定(対応がとれない場合の順位検定)
 - (c) Fisher's Exact Test (Fisherの直接確率計算法)
 - (d) 1要因のクロス集計 (e) 正規分布を用いた符号検定

- まとめ
- 1群の差の検定
- 2群の差の検定(独立2群)の場合
- クロス集計
- 1要因のクロス集計
- 2要因のクロス集計

第6章 行列データを作ろう

- 6.1 はじめに
- 6.2 正規化テーブルの作り方
 - (a) reshapeパッケージの活用 (b) reshape2パッケージの活用
- 6.3 部分行列の取得法
 - (a) 行列[c(xxx), c(yyy)]あるいは行列[-c(xxx), -c(yyy)]として部分行列を定義する
 - (b) 同一の数値のみから構成される列を削除する (c) 行の削除

第7章 教師なし学習: 多変量データの視覚化、クラスター分析など

- 7.1 はじめに
- 7.2 相関係数

- (a) ピアソン相関係数 (b) スピアマン相関係数 (c) ケンドール相関係数
- (d) 相関係数の検定 (e) 多様なpairs()を活用した関数群
- (f) pairs()では視覚化できない多くの変数間の相関を列挙する
- 7.3 データ行列、相関行列、相関行列、スケーリング
 - (a) スケーリング (b) 対数変換
- 7.4 欠損値(欠落値)の対応
 - (a) 距離行列
- 7.5 多次元尺度構成法、主成分分析
 - (a) 多次元尺度構成法 (b) 主成分分析
- 7.6 自己組織化マップ: Self-Organizing Mapping (SOM)
- 7.7 クラスター分析法
 - (a) 階層法(凝集法) (a1) 最小距離法 (a2) 重心距離法
 - (b) 2次元クラスターリング
 - (c) 分割法 (c1) K平均 (c2) ギャップ統計量

第8章 多変量回帰モデル

- 8.1 はじめに
- 8.2 重回帰分析
 - (a) 10種競技データ (b) 重回帰分析 (c) 線形回帰モデルの妥当性の評価法 (d) 重回帰モデルの係数bの求め方 (e) 多重共線性
- 8.3 PLS: 部分最小二乗法
 - (a) PLS回帰モデル
 - (b) 重回帰モデルとPLSモデルのどちらを選ぶべきか?
- 8.4 スパースモデリング
 - (a) リッジ解析 (b) ラッソ解析

第9章 機械学習

- 9.1 はじめに 9.2 教師あり学習 9.3 データセット
- 9.4 caretパッケージ
 - (a) caretパッケージとは (b) インストール (c) caretマニュアル
- 9.5 アヤメデータの教師なし学習
- 9.6 アヤメデータの教師あり学習
 - (a) 線形判別分析 (b) 2次判別関数法(mmethod='qda')
 - (c) k最近隣法(kNN法) (d) NaiveBayes法 (e) 決定木(Decion Tree)
 - (f) ニューラルネットワーク
 - (g) カーネルサポートベクトルマシーン
 - (h) アンサンブル学習: バギング, ランダムダムフォレスト、ブースティング
- 9.7 アヤメデータ解析のまとめ

第10章 化学構造処理

- 10.1 はじめに
- 10.2 化学構造のデジタル処理
- 10.3 SMILES 10.4 rcdkパッケージ
- 10.5 rcdk
 - 10.5.1 SMILESから化合物構造の描画1
 - (a) SMILESから化学構造を描画する (b) SMILESから化学構造を描画する2
 - 10.5.2 モルファイルからSMILESへの変換
 - (a) 多数のモルファイルをSMILESに変換し、表データをマージする
 - 10.5.3 SMILESによる物性値の推算
 - (a) 種々の分子特性を計算しよう! (RcdkSmilesToMP01.R)
 - (b) 分子記述子 (c) 分子フィンガープリント

第III部 化学データによるデータサイエンス実践

第11章 データサイエンスによる化学・マテリアル化学の課題解決の実践

- 11.1 はじめに
- 11.2 プラスチックパーツの引張強度
- 11.3 ホモポリマーの物性相関
 - (a) 2Dクラスター分析
 - (b) モノマーの分子記述子によるポリマーの物性予測のための回帰モデルの開発
- 11.4 L-Aspartyl Dipeptidesの苦味と甘味の分子記述子による識別
- 11.5 農薬添加回収率のケモインフォマティクス
 - (a) 説明変数と目的変数の相関解析
 - (b) 説明変数と目的変数の相関データを視覚化する
 - (c) グラフから次の作業を考える (d) 多変量回帰モデルを作成する
 - (e) 回帰モデルを選択する

第12章 おわりに: さらなる展開

謝辞

付録1: caretパッケージの方法とmethodの定義

著者一覧

黄 銘	奈良先端科学技術大学院大学	先端科学技術研究科	情報科学領域	計算システムズ生物学研究室	助教
小野 直亮	奈良先端科学技術大学院大学	先端科学技術研究科	情報科学領域	計算システムズ生物学研究室	准教授
モハマド アルタフル アミン	奈良先端科学技術大学院大学	先端科学技術研究科	データサイエンス創造センター		准教授
	奈良先端科学技術大学院大学	先端科学技術研究科	情報科学領域	計算システムズ生物学研究室	准教授
金谷 重彦	奈良先端科学技術大学院大学	先端科学技術研究科	情報科学領域	計算システムズ生物学研究室	教授