

## 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)  $\chi^2$  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群の順位検定
    - (a) Wilcoxon 符号つき順位検定：対応がある2群の検定
    - (b) ウィルコクソン順位検定 (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
- まとめ

#### 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、距離行列、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次元クラスタリング
  - (c) 分割法
    - (c1) K平均
    - (c2) ギャップ統計量
- まとめ

## 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

## 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
- 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
- まとめ

- (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