

詳細目次

第1章	ケモ・マテリアルズ・インフォマティクス事始め	2
1.1	ケモ・マテリアルズ・インフォマティクスとは？	2
1.2	Rプログラミングを活用できる 11 の場面	9
1.2.1	利用場面 1：関数電卓の代用	10
1.2.2	利用場面 2：グラフ作成ツール	10
1.2.3	利用場面 3：離散型データ（計数データ）の分析	11
1.2.4	利用場面 4：連続型データ（計量データ）の分析	12
1.2.5	利用場面 5：相関分析	13
1.2.6	利用場面 6：主成分分析・多次元尺度構成法・自己組織化マップ	13
1.2.7	利用場面 7：クラスター分析	14
1.2.8	利用場面 8：回帰手法	14
1.2.9	利用場面 9：分類手法	16
1.2.10	利用場面 10：化学構造を利用した材料特性の解析	17
1.2.11	利用場面 11：ディープラーニングを利用した画像認識・グラフ表現の解析	17
1.3	R言語とプラットフォームのインストール	18
1.3.1	R言語とは何だろうか？	18
1.3.2	RおよびRStudioのインストールと使い方	19
第2章	データハンドリング～Rプログラミングの基礎事項～	23
2.1	事前学習	23
2.2	外部ファイルとの間のデータの入出力	27
2.2.1	Excelファイル(CSV形式)との間のデータの入出力	27
	[プログラム 2_2_1] RInputOutput.R	27
2.3	ベクトル・行列・リストの操作と応用	30
2.3.1	ベクトル	30
2.3.1.1	異なる型のデータを結合してベクトル作成	30
	[プログラム 2_3_1_1] Rvector01.R	30
2.3.1.2	繰り返し文で利用するための保存用ベクトルの作成	31
	[プログラム 2_3_1_2] Rvector02.R	31
2.3.1.3	ベクトルの作成で使われる関数	32
	[プログラム 2_3_1_3] Rvector03.R	32
2.3.1.4	複数のベクトルの成分についての集合演算	33
	[プログラム 2_3_1_4] Rvector04.R	33

2.3.1.5	ベクトルの中から特定の値を持つ成分番号の抽出	34
	[プログラム 2_3_1_5] Rvector05.R	34
2.3.2	行列	34
2.3.2.1	行列と部分行列の作成	34
	[プログラム 2_3_2_1] Rmatrix01.R	34
2.3.2.2	行列の行と列にラベル(名前)付加	36
	[プログラム 2_3_2_2] Rmatrix02.R	36
2.3.2.3	行列の操作および四則演算	37
	[プログラム 2_3_2_3] Rmatrix03.R	37
2.3.2.4	欠損値のある行列の操作	41
	[プログラム 2_3_2_4] Rmatrix04.R	41
2.3.2.5	重複するデータの削除	43
	[プログラム 2_3_2_5] Rmatrix05.R	43
2.3.2.6	統計量に基づくデータの抽出	44
	[プログラム 2_3_2_6] Rmatrix06.R	45
2.3.2.7	特定の列あるいは行のデータ大小の順番に沿って 行列全体の成分並べ替え	47
	[プログラム 2_3_2_7] Rmatrix07.R	47
2.3.2.8	データの整形(1)	48
	[プログラム 2_3_2_8] Rmatrix08.R	48
2.3.2.9	データの整形(2)	51
	[プログラム 2_3_2_9] Rmatrix09.R	51
2.3.3	リスト	53
2.3.3.1	ベクトルとの相互変換・演算	54
	[プログラム 2_3_3_1] Rlist01.R	54
2.3.3.2	グループごとにデータを分類したリスト作成	55
	[プログラム 2_3_3_2] Rlist02.R	55
2.3.4	関数	58
2.3.4.1	独自関数を同じプログラムの中で定義・利用	58
	[プログラム 2_3_4_1] Rfunc01.R	58
2.3.4.2	別ファイルで定義した独自関数の呼び出し・利用	59
	[プログラム 2_3_4_2] Rfunc02a.R と Rfunc02b.R	59
2.3.5	行列の処理に便利な apply 系関数	60
2.3.5.1	行あるいは列単位で演算させる apply 関数(1)	60
	[プログラム 2_3_5_1] Rapply01.R	61
2.3.5.2	行あるいは列単位で演算させる apply 関数(2)	63

[プログラム 2_3_5_2] Rapply02.R	63
2.3.5.3 分類に基づくデータ整形に利用できる tapply 関数	64
[プログラム 2_3_5_3] Rtapply.R	64
2.3.5.4 ベクトルあるいは行列の成分全体に関数を適用する lapply 関数、 sapply 関数	67
[プログラム 2_3_5_4] Rlapply.R	67
2.3.5.5 グローバル変数とローカル変数	69
[プログラム 2_3_5_5] Rmapply.R	69
2.4 グラフィックスを用いたデータの可視化	70
2.4.1 折れ線グラフ	71
2.4.1.1 折れ線グラフと関数プロット	72
[プログラム 2_4_1_1] Rsimpleplot.R	72
2.4.2 箱ヒゲ図	74
2.4.2.1 箱ヒゲ図(1)	74
[プログラム 2_4_2_1] Rboxplot.R	74
2.4.2.2 箱ヒゲ図(2)	75
[プログラム 2_4_2_2] RDataProcessing01.R	75
2.4.2.3 箱ヒゲ図(3)	78
[プログラム 2_4_2_3] RDataProcessing02.R	78
2.4.3 ヒストグラム	78
2.4.3.1 ヒストグラム(1)	79
[プログラム 2_4_3_1] Rhist01.R	79
2.4.3.2 ヒストグラム(2)	79
[プログラム 2_4_3_2] Rhist02.R	79
2.4.4 散布図	81
2.4.4.1 散布図(1)	81
[プログラム 2_4_4_1] Rscatplot01.R	81
2.4.4.2 散布図(2)	83
[プログラム 2_4_4_2] Rscatplot02.R	84
2.5 まとめ	86
コラム 2.1 特殊なデータ型 (NULL, NA, NaN, Inf)	87
コラム 2.2 プログラム時のクイック・レファレンス	88
 第3章 離散型データ (計数データ) の分析	 90
3.1 事前学習	90
3.2 二項分布とそれを利用した検定	93

3.2.1	目の出方に偏りがないようにサイコロが作られているかどうかの検定	93
	[プログラム 3_2_1A] Rdice01.R	94
	[プログラム 3_2_1B] Rdice02.R	95
3.2.2	ド・メレの 2 つのダイス (サイコロ)	95
	[プログラム 3_2_2] Rdice03.R	96
3.2.3	二項分布の確率密度関数と累積確率密度関数	97
	[プログラム 3_2_3] Rbinom01.R	97
3.2.4	二項分布を使った統計検定	98
	[プログラム 3_2_4] Rbinom02.R	99
3.2.5	二項分布を用いた符号検定 (サンプル数 $n \leq 25$)	100
	[プログラム 3_2_5] Rbinom03.R	101
3.2.6	正規分布を用いた符号検定 (サンプル数 $n > 25$)	102
	[プログラム 3_2_6] Rbinom04.R	102
3.3	超幾何分布とそれを利用した検定	103
3.3.1	再度釣り上げた魚のタグ数の確率 (ケース 1)	104
	[プログラム 3_3_1] Rhypergeo01.R	104
3.3.2	検査済製品のタグ数の確率 (ケース 2)	105
	[プログラム 3_3_2] Rhypergeo02.R	105
3.3.3	2x2 分割表 (クロス分割表) を用いた「独立性の検定」	106
	[プログラム 3_3_3A] RFisherExact01.R	108
	[プログラム 3_3_3B] RFisherExact02.R	108
3.4	ポアソン分布	110
3.4.1	二項分布とポアソン分布の比較	110
	[プログラム 3_4_1] RPoisson01.R	110
3.5	まとめ	111
第4章	連続型データ (計量データ) の分析	112
4.1	事前学習	112
4.2	正規分布の特性と応用	113
4.2.1	1 次関数から 5 次関数までの描画	114
	[プログラム 4_2_1] Rpolynom.R	114
4.2.2	平均値が変化したときの正規分布の確率密度関数の描画	115
4.2.3	標準偏差を変化させたときの正規分布の確率密度関数の描画	115
	[プログラム 4_2_3] Rnorm_dist02.R	115
4.2.4	正規分布 $N(0, 1)$ の確率密度関数と累積分布関数の比較	116
	[プログラム 4_2_4] Rnorm_dist03.R	116

4.2.5	正規分布 $N(0, 1)$ および $N(0, 2)$ の分位数関数の描画	117
	[プログラム 4_2_5] Rnorm_dist04.R	117
4.2.6	正規分布 $N(0, 1)$ と信頼区間の関係	117
	[プログラム 4_2_6] Rnorm_dist05.R	118
4.3	カイ二乗分布の特性と応用	119
4.3.1	自由度 n のカイ二乗分布の確率密度関数	120
	[プログラム 4_3_1] Rchisq_dist01.R	120
4.3.2	カイ二乗分布を利用した母集団の分散の推定	121
	[プログラム 4_3_2] Rchisq_dist02.R	121
4.4	t 分布の特性と応用	122
4.4.1	自由度 n の t 分布と標準正規分布の比較	123
	[プログラム 4_4_1] Rt_dist01.R	123
4.4.2	t 分布を利用した母集団の平均(期待値)の区間推定	124
	[プログラム 4_4_2] Rt_dist02.R	124
4.5	t 分布を利用した 2 組のデータの比較	124
4.5.1	2 種類の触媒違いによる反応率増減に関する有意差検定	127
4.6	ノンパラメトリック統計検定	129
4.6.1	2 種類の処理違いによる特性値変化について Wilcoxon 順位和検定	130
4.6.2	同順位がある場合の Wilcoxon 順位和検定	131
4.6.3	対応のある 2 群のデータについて平均値の差の検定	132
	[プログラム 4.6.3] RWilcoxon01.R	132
4.6.4	2 種類の材料について反応時間の有意差の検定	134
	[プログラム 4_6_4A] RWilcoxon02.R	135
	[プログラム 4_6_4B] RWilcoxon03.R	136
4.7	分割表を利用した独立性の検定・適合性の検定	137
4.7.1	故障件数の独立性の検定	138
	[プログラム 4_7_1] Rindepend01.R	138
4.7.2	アレルギー症状発症と製品との間の独立性の検定	139
	[プログラム 4_7_2] Rindepend02.R	140
4.8	サンプル数に応じた検定手法の選択	140
4.9	まとめ	141
第5章	データに潜む類似度・距離の分析—相関・距離・クラスターの視点から	146
5.1	事前学習	146
5.2	相関分析	148
5.2.1	各国の ABO・Rh 式血液型割合の相関分析	149

[プログラム 5_2_1] Rcorrelation01.R	150
5.2.2 各国の ABO・Rh 式血液型割合の相関分析 (5.1.1 の続き)	153
[プログラム 5_2_2] Rcorrelation02.R	153
5.2.3 ポリマーの物性値間の相関分析	154
[プログラム 5_2_3] Rcorrelation03.R	155
5.3 主成分分析・多次元尺度構成法・自己組織化マップ	159
5.3.1 主成分分析 (PCA)	159
5.3.1.1 ガスクロマトグラフィーよりワインの種類を識別する主成分分析	161
[プログラム 5_3_1_1] RPCA01.R	161
5.3.2 多次元尺度構成法 (MDS)	167
5.3.2.1 ワインの種類を識別する計量 MDS : 標準装備の関数利用	168
[プログラム 5_3_2_1] RMDS01.R	168
5.3.2.2 ワインの種類を識別する計量 MDS : 自作プログラミング	170
[プログラム 5_3_2_2] RMDS02.R	170
5.3.3 自己組織化マップ (SOM)	171
5.3.3.1 ガスクロマトグラフィーよりワインの種類を識別する自己組織化マップ	173
[プログラム 5_3_3_1] RSOM01.R	173
5.4 クラスタ分析	176
5.4.1 階層法 (凝集法)	177
5.4.1.1 「最長距離法」による樹形図の描画 (1)	180
[プログラム 5_4_1_1] RclusterHierarch01.R	180
5.4.1.2 「最長距離法」による樹形図の描画 (2)	182
[プログラム 5_4_1_2] RclusterHierarch02.R	182
5.4.1.3 「ヒートマップ」描画によるクラスタ分析	184
[プログラム 5_4_1_3] RclusterHierarch03.R	184
5.4.2 分割法 (k 平均法)	187
5.4.2.1 Gap 統計量と k 平均法を用いたクラスタ分析	189
[プログラム 5_4_2_1] RClusteGapKmeans01.R	189
5.2 まとめ	192
第6章 データに潜む変数間の関係をモデル化する手法—回帰分析の視点から—	194
6.1 事前学習	194
6.2 線形重回帰分析	196
6.2.1 線形重回帰分析を理解しよう!	198
[プログラム 6_2_1] RMultipleRegression01.R	198
6.2.2 線形重回帰モデルの妥当性を評価しよう!	204

[プログラム 6_2_2] RMultipleRegression02.R	204
6.2.3 線形重回帰モデルの誤差に着目しよう！	205
[プログラム 6_2_3] RMultipleRegression03.R	205
6.2.4 線形重回帰モデルの結果をもとに別の選手の総合得点を予測しよう！	207
[プログラム 6_2_4] RMultipleRegression04.R	207
6.2.5 多重共線性の問題	209
[プログラム 6_2_5] RMultipleRegression05.R	209
6.3 部分最小二乗法 (PLS)	212
6.3.1 部分最小二乗法による解析	213
[プログラム 6_3_1] RPLS01.R	213
6.4 正則化を利用した回帰 (正則化最小二乗法)	223
6.4.1 リッジ回帰による解析	225
[プログラム 6_4_1] RRidge01.R	225
6.4.2 ラッソ回帰による解析	229
[プログラム 6_4_2] RLasso01.R	229
6.5 まとめ	234
第7章 識別・分類・認識に役立つモデル化手法—教師あり機械学習の視点から—	236
7.1 事前学習	236
7.2 教師無し学習	239
7.2.1 主成分分析 (教師無し学習) を利用した 3 種のアヤメへの分類	240
[プログラム 7_2_1] RIrisPCA01.R	241
7.3 判別分析	245
7.3.1 線形判別分析 (教師無し学習) を利用した 3 種のアヤメへの分類	248
[プログラム 7_3_1] RML_LDA01.R	248
7.3.2 2 次判別分析 (教師あり学習) を利用した 3 種のアヤメへの分類	255
[プログラム 7_3_2] RML_QDA01.R	255
7.4 k 最近傍法 (kNN 法)	256
7.4.1 k 最近傍法 (教師あり学習) を利用した 3 種のアヤメへの分類	257
[プログラム 7_4_1] RML_KNN01.R	257
7.5 ナイーブベイズモデル	260
7.5.1 ナイーブベイズモデル (教師あり学習) を利用した 3 種のアヤメへの分類	261
[プログラム 7_5_1] RML_NBA01.R	261
7.5.2 ナイーブベイズモデル (教師あり学習) を利用した 3 種のアヤメへの分類	264

[プログラム 7_5_2] RML_NBA02. R	264
7.6 決定木モデル	268
7.6.1 決定木モデル(教師あり学習)を利用した 3 種のアヤメへの分類	269
[プログラム 7_6_1] RML_DTR01. R	269
7.7 ニューラルネットワークモデル	271
7.7.1 ニューラルネットワークモデル(教師あり学習)を利用した 3 種のアヤメへの分類	272
[プログラム 7_7_1] RML_NNW01. R	272
7.8 サポートベクトルマシーン	277
7.8.1 サポートベクトルモデル(教師あり学習)を利用した 3 種のアヤメへの分類	279
[プログラム 7_8_1] RML_KSV01. R	279
7.9 アンサンブル学習とランダムフォレストモデル	281
7.9.1 ランダムフォレストモデル(教師あり学習)を利用した 3 種のアヤメへの分類	282
7.10 まとめ	285
7.10.1 3 種のアヤメの分類に関する 8 種類の教師あり学習の結果比較	285
[プログラム 7_10_1] RML_ALL. R	285
 第8章 化学情報処理—化学構造の解析とその応用—	 290
8.1 事前学習	290
8.2 化学構造の表記法	292
8.3 rcdk パッケージの応用	293
8.3.1 SMILES 形式の化学構造データの描画	293
8.3.1.1 SMILES 形式の化学構造の描画(その 1)	293
[プログラム 8_3_1_1] RCDK01. R	293
8.3.1.2 SMILES 形式の化学構造の描画(その 2)	294
[プログラム 8_3_1_2] RCDK02. R	294
8.3.2 モルファイル情報から SMILES 形式のデータ作成	296
[プログラム 8_3_2] RCDK03. R	297
8.3.3 SMILES 形式のデータから分子特性の計算	298
[プログラム 8_3_3] RCDK04. R	298
8.3.4 SMILES 形式のデータから分子特性の計算	300
[プログラム 8_3_4] RCDK05. R	300
8.3.5 SMILES 形式のデータからフィンガープリントの作成と クラスター分析への応用	305

[プログラム 8_3_5] RCDK06. R	306
8.4 ChemmineR・ChemmineOB パッケージの応用	307
8.4.1 ChemmineR への SDF 形式ファイルの読み込み、ファイル出力、 化学構造の描画	308
[プログラム 8_4_1] RChM01. R	308
8.4.2 ChemmineR を利用したファイル形式変換およびファイル統合	310
8.4.2.1 モルファイルの SDF 形式データから SMILES 形式のデータへの変換 および csv ファイルの保存	310
[プログラム 8_4_2_1] RChM02. R	310
8.4.2.2 ChemmineR を用いて SMILES 形式のデータから SDF 形式データへの変換と保存	313
[プログラム 8_4_2_2] RChM03. R	313
8.4.2.3 モルフォルダに格納された複数のモルファイルと 化合物情報のファイルの統合	315
[プログラム 8_4_2_3] RChM04. R	315
8.4.3 ChemmineR を利用した化学フィンガープリントによる 化合物の階層的クラスター分析	318
[プログラム 8_4_3] RChM05. R	318
8.4.4 ChemmineR を利用した分子記述子の計算	323
[プログラム 8_4_4] RChM06. R	323
8.4.5 ChemmineR を利用した原子対類似度の評価と化合物の分類	325
8.4.5.1 化合物の原子対類似度にもとづく化合物の検索と評価	325
[プログラム 8_4_5_1] RChM07. R	325
8.4.5.2 ChemmineR を用いた原子対類似度と 2次元尺度構成法による化合物の分類	327
[プログラム 8_4_5_2] RChM08. R	327
8.5 まとめ	329
 第9章 深層学習 (ディープラーニング)	 330
9.1 事前学習	330
9.2 ニューラルネットワークの基本要素	331
9.3 ニューラルネットワークの構築	333
[プログラム 9_3] RDL01. R	333
9.4 実データによる学習	335
[プログラム 9_4A] RDL02. R	335
[プログラム 9_4B] RDL03. R	337

9.5 畳み込みニューラルネットワーク	340
[プログラム 9_5] RDL04.R	342
9.6 まとめ	346
付録	
A 関連パッケージのインストールおよび環境設定	347
A.1 第8章で利用する化学情報(Cheminformatics) 関連パッケージの インストールと環境設定	348
A.1.1 インストールの基本情報	348
A.2 第9章で利用する Keras のインストールと環境設定	349
A.2.1 インストールの基本情報	349
A.3 利用すると便利な機能	351
A.3.1 Windows での Rtools の自動切替方法	351
A.3.2 R の C/C++/Fortran 設定方法	352
A.3.3 Java 環境変数の設定方法	352
A.3.4 深層学習の GPU での計算方法について	353
A.3.5 Windows 版 R の線形計算ライブラリ更新方法	354
B トラブルシューティング	354
参考文献	357
サポート	
関数一覧 (第8, 9章を除く各章から抜粋)	358
利用パッケージ一覧	368
コラム一覧	368
索引	369