

Contents

はじめに	III
本書の概要とPythonの基礎	001
1-1 データサイエンティストの仕事	002
1-1-1 データサイエンティストの仕事	002
1-1-2 データ分析のプロセス	003
1-1-3 本書の構成	004
1-1-4 本書を読み進めるのに役立つ文献	005
1-1-5 手を動かして習得しよう	005
1-2 Pythonの基礎	007
1-2-1 Jupyter Notebookの使い方	007
1-2-2 Pythonの基礎	011
1-2-3 リストと辞書型	015
1-2-4 条件分岐とループ	017
Column format記法と%記法	019
1-2-5 関数	023
Practice 練習問題 1-1	026
練習問題 1-2	026
1-2-6 クラスとインスタンス	026
Practice 1章 総合問題	030
科学計算、データ加工、 グラフ描画ライブラリの使い方の基礎	031
2-1 データ分析で使うライブラリ	032
2-1-1 ライブラリの読み込み	032
2-1-2 マジックコマンド	033
2-1-3 この章で使うライブラリのインポート	034

2-2	Numpyの基礎	035
2-2-1	Numpyのインポート.....	035
2-2-2	配列操作.....	035
2-2-3	乱数.....	039
Column	Numpyは高速.....	041
2-2-4	行列.....	042
Practice	練習問題2-1.....	044
	練習問題2-2.....	044
	練習問題2-3.....	044
2-3	Scipyの基礎	045
2-3-1	Scipyのライブラリのインポート.....	045
2-3-2	行列計算.....	045
2-3-3	ニュートン法.....	047
Practice	練習問題2-4.....	048
	練習問題2-5.....	048
	練習問題2-6.....	048
2-4	Pandasの基礎	049
2-4-1	Pandasのライブラリのインポート.....	049
2-4-2	Seriesの使い方.....	049
2-4-3	DataFrameの使い方.....	051
2-4-4	行列操作.....	052
2-4-5	データの抽出.....	053
2-4-6	データの削除と結合.....	054
2-4-7	集計.....	055
2-4-8	値のソート.....	056
2-4-9	nan (null) の判定.....	057
Practice	練習問題2-7.....	058
	練習問題2-8.....	058
	練習問題2-9.....	058
2-5	Matplotlibの基礎	059
2-5-1	Matplotlibを使うための準備.....	059
2-5-2	散布図.....	059

2-5-3	グラフの分割.....	061
2-5-4	関数グラフの描画.....	062
2-5-5	ヒストグラム.....	063
Column	さまざまなデータのビジュアル化.....	064
Practice	練習問題2-10.....	065
	練習問題2-11.....	065
	練習問題2-12.....	065
Practice	2章 総合問題.....	066

Chapter 3 記述統計と単回帰分析..... 067

3-1	統計解析の種類	068
3-3-1	記述統計と推論統計.....	068
3-3-2	この章で使うライブラリのインポート.....	068
3-2	データの読み込みと対話	070
3-2-1	インターネットなどで配布されている対象データの読み込み.....	070
3-2-2	データの読み込みと確認.....	072
3-2-3	データの性質を確認する.....	074
Column	「変数」という用語について.....	075
3-2-4	量的データと質的データ.....	077
3-3	記述統計	079
3-3-1	ヒストグラム.....	079
3-3-2	平均、中央値、最頻値.....	079
3-3-3	分散と標準偏差.....	080
3-3-4	要約統計量とパーセンタイル値.....	081
3-3-5	箱ひげ図.....	083
3-3-6	変動係数.....	084
3-3-7	散布図と相関係数.....	085
3-3-8	すべての変数のヒストグラムや散布図を描く.....	087
Practice	練習問題3-1.....	089
	練習問題3-2.....	089
	練習問題3-3.....	089

3-4	単回帰分析	090
3-4-1	線形単回帰分析	091
3-4-2	決定係数	092
	Practice 練習問題3-4	093
	練習問題3-5	093
	練習問題3-6	093
	Practice 3章 総合問題	094

Chapter 4 確率と統計の基礎

4-1	確率と統計を学ぶ準備	096
4-1-1	この章の前提知識	096
4-1-2	この章で使うライブラリのインポート	096
4-2	確率	097
4-2-1	数学的確率	097
4-2-2	統計的確率	099
4-2-3	条件付き確率と乗法定理	100
4-2-4	独立と従属	101
4-2-5	ベイズの定理	101
	Practice 練習問題4-1	102
	練習問題4-2	102
	練習問題4-3	102
4-3	確率変数と確率分布	103
4-3-1	確率変数、確率関数、分布関数、期待値	103
4-3-2	さまざまな分布関数	104
4-3-3	カーネル密度関数	107
	Practice 練習問題4-4	108
	練習問題4-5	108
	練習問題4-6	108
4-4	応用：多次元確率分布	109
4-4-1	同時確率関数と周辺確率関数	109

4-4-2	条件付き確率関数と条件付き期待値	109
4-4-3	独立の定義と連続分布	110
4-5	推計統計学	112
4-5-1	大数の法則	112
4-5-2	中心極限定理	113
4-5-3	標本分布	114
	Practice 練習問題4-7	116
	練習問題4-8	116
	練習問題4-9	116
4-6	統計的推定	117
4-6-1	推定量と点推定	117
4-6-2	不偏性と一致性	117
4-6-3	区間推定	118
4-6-4	推定量を求める	118
	Practice 練習問題4-10	119
	練習問題4-11	119
	練習問題4-12	119
4-7	統計的検定	120
4-7-1	検定	120
4-7-2	第1種の過誤と第2種の過誤	121
4-7-3	ビッグデータに対する検定の注意	122
	Practice 練習問題4-13	122
	Practice 4章 総合問題	122

Chapter 5 Pythonによる科学計算 (NumpyとScipy)

5-1	概要と事前準備	124
5-1-1	この章の概要	124
5-1-2	この章で使うライブラリのインポート	124
5-2	Numpyを使った計算の応用	125
5-2-1	インデックス参照	125

Practice	練習問題5-1	128
	練習問題5-2	128
	練習問題5-3	128
5-2-2	Numpyの演算処理	129
Practice	練習問題5-4	131
	練習問題5-5	131
	練習問題5-6	131
5-2-3	配列操作とブロードキャスト	132
Practice	練習問題5-7	136
	練習問題5-8	136
	練習問題5-9	136
5-3	Scipyを使った計算の応用	137
5-3-1	補間	137
5-3-2	線形代数：行列の分解	139
Practice	練習問題5-10	139
	練習問題5-11	139
	練習問題5-12	139
	練習問題5-13	142
	練習問題5-14	142
5-3-3	積分と微分方程式	143
Practice	練習問題5-15	146
	練習問題5-16	146
5-3-4	最適化	146
Practice	練習問題5-17	149
	練習問題5-18	149
Practice	5章 総合問題	150

Chapter 6 Pandasを使ったデータ加工処理 151

6-1	概要と事前準備	152
6-1-1	この章で使うライブラリのインポート	153

6-2	Pandasの基本的なデータ操作	154
6-2-1	階層型インデックス	154
Practice	練習問題6-1	156
	練習問題6-2	156
	練習問題6-3	156
6-2-2	データの結合	157
Practice	練習問題6-4	162
	練習問題6-5	163
	練習問題6-6	163
6-2-3	データの操作と変換	163
Practice	練習問題6-7	169
	練習問題6-8	169
	練習問題6-9	169
6-2-4	データの集約とグループ演算	169
Practice	練習問題6-10	172
	練習問題6-11	172
	練習問題6-12	172
6-3	欠損データと異常値の取り扱いの基礎	173
6-3-1	欠損データの扱い方	173
Practice	練習問題6-13	176
	練習問題6-14	176
	練習問題6-15	176
6-3-2	異常データの扱い方	177
6-4	時系列データの取り扱いの基礎	178
6-4-1	時系列データの処理と変換	178
Practice	練習問題6-16	181
6-4-2	移動平均	181
Practice	練習問題6-17	182
Practice	6章 総合問題	182

Chapter 7 Matplotlibを使ったデータ可視化 183

7-1	データの可視化	184
7-1-1	データの可視化について	184
7-1-2	この章で使うライブラリのインポート	184
7-2	データ可視化の基礎	185
7-2-1	棒グラフ	185
7-2-2	円グラフ	187
	Practice 練習問題7-1	190
	練習問題7-2	190
	練習問題7-3	190
7-3	応用：金融データの可視化	191
7-3-1	可視化する金融データ	191
7-3-2	ローソクチャートを表示するライブラリ	192
7-4	応用：分析結果の見せ方を考えよう	193
7-4-1	資料作成のポイントについて	193
	Practice 7章 総合問題	194
	Column 移動平均時系列データと対数時系列データ	196

Chapter 8 機械学習の基礎 (教師あり学習) 197

8-1	機械学習の全体像	198
8-1-1	機械学習とは	198
8-1-2	教師あり学習	200
8-1-3	教師なし学習	201
8-1-4	強化学習	201
8-1-5	この章で使うライブラリのインポート	202
8-2	重回帰	203
8-2-1	自動車価格データの取り込み	203
8-2-2	データの整理	205
8-2-3	モデル構築と評価	206

8-2-4	モデル構築とモデル評価の流れのまとめ	208
	Practice 練習問題8-1	208
8-3	ロジスティック回帰	209
8-3-1	ロジスティック回帰の例	209
8-3-2	データの整理	210
8-3-3	モデル構築と評価	211
8-3-4	スケーリングによる予測精度の向上	212
	Practice 練習問題8-2	213
	練習問題8-3	213
8-4	正則化項のある回帰：ラッソ回帰、リッジ回帰	214
8-4-1	ラッソ回帰、リッジ回帰の特徴	214
8-4-2	重回帰とリッジ回帰の比較	215
	Practice 練習問題8-4	215
8-5	決定木	216
8-5-1	キノコデータセット	216
8-5-2	データの整理	218
8-5-3	エントロピー：不純度の指標	219
8-5-4	情報利得：分岐条件の有益さを測る	222
8-5-5	決定木のモデル構築	225
	Practice 練習問題8-5	226
8-6	k-NN (k近傍法)	227
8-6-1	k-NNのモデル構築	227
	Practice 練習問題8-6	229
	練習問題8-7	229
8-7	サポートベクターマシン	230
8-7-1	サポートベクターマシンのモデル構築	230
	Practice 練習問題8-8	232
	Practice 8章 総合問題	232

Chapter 9 機械学習の基礎 (教師なし学習) 233

9-1	教師なし学習	234
9-1-1	教師なしモデルの種類	234
9-1-2	この章で使うライブラリのインポート	235
9-2	クラスタリング	236
9-2-1	k-means法	236
9-2-2	k-means法でクラスタリングする	237
9-2-3	金融マーケティングデータをクラスタリングする	239
9-2-4	エルボー法によるクラスター数の推定	243
9-2-5	クラスタリング結果の解釈	245
9-2-6	k-means法以外の手法	249
	Practice 練習問題9-1	249
9-3	主成分分析	250
9-3-1	主成分分析を試す	250
9-3-2	主成分分析の実例	255
	Practice 練習問題9-2	258
9-4	マーケットバスケット分析とアソシエーションルール	259
9-4-1	マーケットバスケット分析とは	259
9-4-2	マーケットバスケット分析のためのサンプルデータを読み込む	259
9-4-3	アソシエーションルール	261
	Practice 9章 総合問題	264

Chapter 10 モデルの検証方法とチューニング方法 265

10-1	モデルの評価と精度を上げる方法とは	266
10-1-1	機械学習の課題とアプローチ	266
10-1-2	この章で使うライブラリのインポート	267
10-2	モデルの評価とパフォーマンスチューニング	268
10-2-1	ホールドアウト法と交差検証法	268
	Practice 練習問題10-1	270
10-2-2	パフォーマンスチューニング：ハイパーパラメータチューニング	270

	Practice 練習問題10-2	273
10-2-3	パフォーマンスチューニング：特徴量の扱い	274
10-2-4	モデルの種類	275
10-3	モデルの評価指標	276
10-3-1	分類モデルの評価：混同行列と関連指標	276
	Practice 練習問題10-3	279
10-3-2	分類モデルの評価：ROC曲線とAUC	279
	Practice 練習問題10-4	284
10-3-3	回帰モデルの評価指標	284
10-4	アンサンブル学習	288
10-4-1	バギング	288
	Practice 練習問題10-5	290
10-4-2	ブースティング	290
	Practice 練習問題10-6	292
10-4-3	ランダムフォレスト、勾配ブースティング	292
10-4-4	今後の学習に向けて	294
	Practice 練習問題10-7	294
	Practice 10章 総合問題	294

Chapter 11 総合演習問題 295

11-1	総合演習問題	296
11-1-1	総合演習問題 (1)	296
11-1-2	総合演習問題 (2)	297
11-1-3	総合演習問題 (3)	297
11-1-4	総合演習問題 (4)	298
11-1-5	総合演習問題 (5)	299
11-1-6	総合演習問題 (6)	300
11-1-7	参考：今後のデータ分析に向けて	301

A-1	本書の環境構築について	304
A-1-1	Anacondaについて	304
A-1-2	Anacondaのパッケージをダウンロードする	304
A-1-3	Anacondaをインストールする	305
A-1-4	pandas-datareaderおよびPlotlyのインストール	310
A-2	練習問題解答	311
A-2-1	Chapter1 練習問題	311
A-2-2	Chapter2 練習問題	313
A-2-3	Chapter3 練習問題	318
A-2-4	Chapter4 練習問題	327
A-2-5	Chapter5 練習問題	334
A-2-6	Chapter6 練習問題	341
A-2-7	Chapter7 練習問題	348
A-2-8	Chapter8 練習問題	364
A-2-9	Chapter9 練習問題	371
A-2-10	Chapter10 練習問題	375
A-2-11	Chapter11 総合演習問題	381
	Column ダミー変数と多重共線性	394
A-3	参考文献・参考URL	408
A-3-1	参考文献	408
A-3-2	参考URL	413
	おわりに	416
	Index	419