# 目次


はじめに ..... 清水秀幸 3
コード・データのダウンロードについて ..... 12

第1章 機械学習の概要とライフサイエンス研究への応用 ..... 清水秀幸 14

1.1 AlphaFold2の衝撃 ..... 14
1.2 機械学習速習 ..... 15
1.2.1 機械学習とは何か？ ..... 15
1.2.2 機械学習が行うこと ..... 17
1.2.3 データの「学習」を紐解く ..... 19
1.2.4 データを丸暗記してはいけない ..... 20
1.2.5 機械学習の性能指標 ..... 22
1.2.6 教師なし学習 ..... 27
1.3 深層学習ことはじめ ..... 27
1.3.1 ニューラルネットワークの基本構造 ..... 27
1.3.2 ニューラルネットワークの学習 ..... 29
1.3.3 さまざまなニューラルネットワーク ..... 29
1.3.4 転移学習のパラダイム ..... 30
1.4 生命医科学への機械学習の応用 ..... 31
1.4.1 ゲノム・トランスクリプトームへの応用 ..... 31
1.4.2 タンパク質・創薬への応用 ..... 31
1.4.3 バイオテクノロジーへの応用 ..... 32
1.5 おわりに ..... 32

第2章 Google Colaboratory, Pandas, Matplotlib, NumPyの基礎 ..... 保住英希，清水秀幸 34

2.1 本章で扱うデータ ..... 34
2.2 Google Colaboratoryの使い方 ..... 35
2.2.1 ノートブックの作成 ..... 35
2.2.2 ファイルのアップロード ..... 36
2.2.3 Google Driveからの読み込み ..... 36
2.2.4 GPUの使用 ..... 38

2.3 データを知る ── Pandas の基礎 1 ..... 39
2.3.1 データの読み込み ..... 39
2.3.2 データの概要の把握 ..... 40
2.4 データを見る ── Matplotlib の基礎 ..... 44
2.4.1 figure と subplot の関係 ..... 44
2.4.2 微調整 ..... 45
2.4.3 実践課題 1 ..... 46
2.4.4 補足：Seaborn ..... 49
2.4.5 実践課題 2 ..... 51
2.4.6 補足：pandas_profiling ..... 53
2.5 データを整形する ── Pandas の基礎 2 ..... 54
2.5.1 データの前処理 ..... 54
2.5.2 データの操作 ..... 55
2.5.3 表の操作 ..... 56
2.6 解析の実行 ── NumPy の基礎 ..... 59
2.6.1 NumPy による計算 ..... 60
2.7 おわりに ..... 61

第 3 章 教師あり学習のためのデータ前処理 澤田高志，清水秀幸 63

3.1 機械学習の概説 ..... 63
3.1.1 機械学習とは何か？ ..... 63
3.1.2 機械学習とそのライブラリ ..... 65
3.1.3 機械学習のワークフロー ..... 65
3.1.4 マイクロアレイデータの解析 ..... 66
3.2 データの前処理 ..... 66
3.2.1 GEOparse を用いたデータの読み込みと欠損値処理 ..... 68
3.2.2 遺伝子発現量データの可視化 ..... 80
3.2.3 遺伝子発現量データの重複の処理 ..... 88
3.2.4 重要な遺伝子発現量データの抽出と欠損値の処理 ..... 94
3.2.5 選ばれた 13 種の mRNA の図示 ..... 99
3.3 おわりに ..... 105

第 4 章 scikit-learn を用いた トランスクリプトームデータの分類 澤田高志，清水秀幸 106

4.1 機械学習：サポートベクトルマシンの実行 ..... 106
4.1.1 訓練データセットとテストデータセット ..... 108
4.1.2 サポートベクトルマシンの導入 ..... 113

4.1.3 カーネルトリックによるサポートベクトルマシンの拡張 ..... 121
4.1.4 グリッドサーチによるハイパーパラメータの最適化 ..... 123
4.1.5 検証データセットの導入 ..... 124
4.1.6 ベイズ最適化によるハイパーパラメータの調整 ..... 133
4.2 おわりに ..... 139

第5章 PyTorchを用いたトランスクリプトームデータの分類 澤田高志，清水秀幸 140

5.1 機械学習：ニューラルネットワークの基礎 ..... 142
5.1.1 基本的なテンソル計算 ..... 142
5.1.2 深層学習の基本知識 ..... 147
5.1.3 PyTorchでニューラルネットワークを構築する ..... 159
5.1.4 PyTorchのハイパーパラメータを最適化する ..... 179
5.2 おわりに ..... 193

第6章 実践編①：生命科学・医歯学分野の画像を用いた機械学習 安齋達彦，髙橋邦彦 194

6.1 はじめに ..... 194
6.2 畳み込みニューラルネットワークを用いた画像判別 ..... 195
6.2.1 プログラムを動作させるための準備と実行手順 ..... 195
6.2.2 画像の読み込み：画像処理の基本 ..... 198
6.2.3 畳み込みニューラルネットワークモデルの構築 ..... 201
6.2.4 学習の実施とその評価 ..... 203
6.2.5 テストデータに対する予測・判別性能の評価 ..... 206
6.2.6 その他のチューニングについて ..... 207
6.3 Grad CAMによる画像分類の判断根拠 ..... 208
6.4 転移学習による判別モデルの構築 ..... 209
6.5 画像セグメンテーションモデルの構築 ..... 211
6.6 おわりに ..... 218

第7章 実践編②：腫瘍特異的ネオ抗原の機械学習を用いた予測 長谷川嵩矩 219

7.1 はじめに ..... 219
7.1.1 ゲノム解析とがん免疫療法 ..... 219
7.1.2 ライブラリのインストール ..... 220
7.1.3 解析の対象とする変異ペプチド候補の作成 ..... 221

7.2 Python を用いたネオ抗原予測 ..... 223
7.3 おわりに ..... 240

第 8 章 実践編③：シングルセル解析と VAE
水越周良，小嶋泰弘，島村徹平 241

8.1 背景と準備 ..... 241
8.1.1 シングルセル解析における課題 ..... 241
8.1.2 VAE とシングルセル解析 ..... 242
8.1.3 VAE のシングルセル解析への応用例 ..... 242
8.1.4 ライブラリとデータセットの用意 ..... 243
8.2 エンコーダとデコーダの構造 ..... 246
8.3 VAE の学習方法 ..... 248
8.3.1 VAE の理論 ..... 248
8.3.2 デコーダの尤度関数 ..... 249
8.3.3 VAE クラスの実装 ..... 251
8.4 その他の部分の実装 ..... 252
8.4.1 early stopping の実装 ..... 252
8.4.2 データの振り分け ..... 253
8.4.3 学習の実行箇所の実装 ..... 254
8.5 学習の実行とモデルの評価 ..... 256
8.6 おわりに ..... 260

第 9 章 実践編④：エピジェネティクスを含む多階層の統合によるがん研究
浅田 健，浜本隆二 261

9.1 はじめに ..... 261
9.2 オートエンコーダを利用したマルチオミクス解析 ..... 263
9.3 オートエンコーダのための PyTorch コード解説 ..... 265
9.4 コード全体を .py ファイルとして書き出す ..... 273
9.5 書き出した .py ファイルの実行 ..... 281
9.6 オプションの使用例 ..... 282
9.7 追記：Anaconda 仮想環境 ..... 284
9.8 おわりに ..... 285

第10章 実践編⑤：タンパク質の「言語」の法則を解き明かす
アミノ酸配列からのタンパク質局在の予測 清水秀幸 287

10.1 生命科学研究に応用されつつある自然言語処理AI ..... 287
10.2 アミノ酸配列のみからタンパク質の局在を予測する：事前学習済みモデルの利用 ..... 288
10.2.1 必要になるライブラリの準備 ..... 289
10.2.2 タンパク質局在データのダウンロードと探索 ..... 291
10.2.3 事前学習済みタンパク質言語モデルのダウンロード ..... 298
10.2.4 アミノ酸配列の事前学習済みモデルによる数値化 ..... 299
10.3 アミノ酸配列のみからタンパク質の局在を予測する：タンパク質局在データによるfine-tuning ..... 302
10.3.1 ニューラルネットワークの設定 ..... 303
10.3.2 タンパク質局在の学習 ..... 304
10.3.3 学習済みモデルのテストデータに対する性能評価 ..... 306
10.4 おわりに ..... 310

第11章 実践編⑥：AI創薬へのはじめの一歩 清水秀幸 311

11.1 従来の創薬が抱える2つの難題と機械学習への期待 ..... 311
11.2 環境の準備 ..... 312
11.2.1 RDKitのインストール ..... 313
11.3 プロジェクト1：csvファイルを読み込み，水への溶解度を予測する線形モデルを作る ..... 313
11.3.1 RDKitの使い方とSMILES表記 ..... 313
11.3.2 SMILESからの分子記述子の抽出 ..... 317
11.3.3 初めてのQSPR解析 ..... 322
11.4 プロジェクト2：アンサンブル学習による水溶解度予測 ..... 326
11.4.1 データのダウンロード ..... 327
11.4.2 アンサンブル学習による溶解度予測 ..... 332
11.5 プロジェクト3：グラフ畳み込みニューラルネットワークによる水溶解性予測 ..... 335
11.5.1 グラフとは何か？ ..... 335
11.5.2 ライブラリのインストールとデータの確認 ..... 336
11.5.3 深層学習モデルの構築 ..... 338
11.5.4 グラフ畳み込みニューラルネットワークの学習 ..... 344
11.6 プロジェクト4：コロナウイルス治療薬探索 ..... 348
11.6.1 コロナウイルスに関するデータの収集 ..... 348
11.6.2 特徴量の抽出 ..... 358

11.6.3 部分的最小二乗回帰モデルの作成 ..... 360
11.7 おわりに ..... 363

第12章 発展編①：機械学習を用いたアプタマー配列の解析と創薬 岩野夏樹，浜田道昭 364

12.1 はじめに ..... 364
12.1.1 アプタマー創薬 ..... 364
12.1.2 アプタマー創薬と機械学習 ..... 366
12.1.3 本章で取り扱う内容 ..... 367
12.2 RaptGen を用いた配列解析 ..... 368
12.2.1 配列解析の準備 ..... 368
12.2.2 モデルの学習 ..... 372
12.2.3 学習結果の描画 ..... 375
12.2.4 本配列生成モデルの応用 ..... 380
12.3 おわりに ..... 386

第13章 発展編②：機械学習によるマイクロバイオームと機能未知遺伝子の解析
メタゲノム・対偶遺伝学・近傍遺伝子解析 西村祐貴，綿野桂人，岩崎 渉 387

13.1 準備 ..... 387
13.2 はじめに ..... 391
13.3 ヒト腸内のメタゲノム解析 ..... 392
13.3.1 アセンブリ ..... 393
13.3.2 ビニング ..... 394
13.3.3 系統プレイスメント ..... 398
13.4 機能未知遺伝子の機能解析と対偶遺伝学的解析・近傍遺伝子解析 ... 400
13.4.1 系統解析 ..... 401
13.4.2 オルソログクラスタリング ..... 403
13.4.3 機能アノテーション ..... 404
13.4.4 機能未知遺伝子解析 ..... 408
13.5 おわりに ..... 420

第14章 終章：さらなる学習のためのリソース 清水秀幸 423
14.1 注目されつつあるノーコード・ローコードAI ..... 423
14.2 本書で扱えなかった重要トピックス ..... 424
14.2.1 強化学習 ..... 424
14.2.2 教師あり学習，教師なし学習の境界の曖昧化 ..... 425
14.2.3 グラフ・ネットワークへの応用 ..... 426
14.2.4 生成モデル ..... 427
14.2.5 説明可能なAI ..... 428
14.2.6 蒸留 ..... 428
14.2.7 連合学習と群学習 ..... 429
14.3 より優れた計算リソースを求めて ..... 430
14.4 生命医科学領域のデータサイエンス・機械学習をさらに勉強するために ..... 431
14.4.1 Pythonを習得する ..... 431
14.4.2 機械学習を理解する ..... 431
14.4.3 機械学習を実践する ..... 433
14.4.4 機械学習のメディア・学会をチェックする ..... 433
14.4.5 生命医科学への応用を実例を通じて学ぶ ..... 434
14.5 おわりに ..... 435

索引 ..... 437
執筆者一覧 ..... 444