

O'REILLY®  
オライリー・ジャパン



# Pythonではじめる 機械学習

scikit-learnで学ぶ特徴量エンジニアリングと機械学習の基礎

Andreas C. Müller 著  
Sarah Guido 著  
中田 秀基 訳

# 目次

まえがき	v
<b>1章 はじめに</b>	<b>1</b>
1.1 なぜ機械学習なのか?	1
1.1.1 機械学習で解決可能な問題	2
1.1.2 タスクを知り、データを知る	4
1.2 なぜPythonなのか?	5
1.3 scikit-learn	5
1.3.1 scikit-learnのインストール	6
1.4 必要なライブラリとツール	7
1.4.1 Jupyter Notebook	7
1.4.2 NumPy	7
1.4.3 SciPy	8
1.4.4 matplotlib	9
1.4.5 pandas	10
1.4.6 mglearn	11
1.5 Python 2 vs. Python 3	12
1.6 本書で用いているバージョン	12
1.7 最初のアプリケーション：アイリスのクラス分類	13
1.7.1 データを読む	14
1.7.2 成功率合いの測定：訓練データとテストデータ	17
1.7.3 最初にすべきこと：データをよく観察する	19
1.7.4 最初のモデル：k-最近傍法	20
1.7.5 予測を行う	22

1.7.6	モデルの評価	23
1.8	まとめと今後の展望	24
<b>2章</b>	<b>教師あり学習</b>	<b>27</b>
2.1	クラス分類と回帰	27
2.2	汎化、過剰適合、適合不足	28
2.2.1	モデルの複雑さとデータセットの大きさ	31
2.3	教師あり機械学習アルゴリズム	31
2.3.1	サンプルデータセット	31
2.3.2	k-最近傍法	36
2.3.3	線形モデル	46
2.3.4	ナイーブベイズクラス分類器	68
2.3.5	決定木	70
2.3.6	決定木のアンサンブル法	82
2.3.7	カーネル法を用いたサポートベクタマシン	90
2.3.8	ニューラルネットワーク (ディープラーニング)	102
2.4	クラス分類器の不確実性推定	115
2.4.1	決定関数 (Decision Function)	116
2.4.2	確率の予測	119
2.4.3	多クラス分類の不確実性	122
2.5	まとめと展望	124
<b>3章</b>	<b>教師なし学習と前処理</b>	<b>127</b>
3.1	教師なし学習の種類	127
3.2	教師なし学習の難しさ	128
3.3	前処理とスケール変換	128
3.3.1	さまざまな前処理	129
3.3.2	データ変換の適用	130
3.3.3	訓練データとテストデータを同じように変換する	132
3.3.4	教師あり学習における前処理の効果	135
3.4	次元削減、特徴量抽出、多様体学習	137
3.4.1	主成分分析 (PCA)	137
3.4.2	非負値行列因子分解 (NMF)	152
3.4.3	t-SNEを用いた多様体学習	159
3.5	クラスタリング	164
3.5.1	k-meansクラスタリング	164
3.5.2	凝集型クラスタリング	177



3.5.3	DBSCAN	182
3.5.4	クラスタリングアルゴリズムの比較と評価	186
3.5.5	クラスタリング手法のまとめ	202
3.6	まとめと展望	203
<b>4章</b>	<b>データの表現と特徴量エンジニアリング</b>	<b>205</b>
4.1	カテゴリ変数	206
4.1.1	ワンホットエンコーディング (ダミー変数)	207
4.1.2	数値でエンコードされているカテゴリ	211
4.2	ビンニング、離散化、線形モデル、決定木	213
4.3	交互作用と多項式	217
4.4	単変量非線形変換	225
4.5	自動特徴量選択	229
4.5.1	単変量統計	229
4.5.2	モデルベース特徴量選択	232
4.5.3	反復特徴量選択	234
4.6	専門家知識の利用	235
4.7	まとめと展望	244
<b>5章</b>	<b>モデルの評価と改良</b>	<b>245</b>
5.1	交差検証	246
5.1.1	scikit-learnでの交差検証	247
5.1.2	交差検証の利点	248
5.1.3	層化k分割交差検証と他の戦略	248
5.2	グリッドサーチ	254
5.2.1	単純なグリッドサーチ	255
5.2.2	パラメータの過剰適合の危険性と検証セット	256
5.2.3	交差検証を用いたグリッドサーチ	258
5.3	評価基準とスコア	270
5.3.1	最終的な目標を見失わないこと	270
5.3.2	2クラス分類における基準	271
5.3.3	多クラス分類の基準	292
5.3.4	回帰の基準	295
5.3.5	評価基準を用いたモデル選択	295
5.4	まとめと展望	298

<b>6章</b>	<b>アルゴリズムチェーンとパイプライン</b>	<b>299</b>
6.1	前処理を行う際のパラメータ選択	300
6.2	パイプラインの構築	302
6.3	パイプラインを用いたグリッドサーチ	303
6.4	汎用パイプラインインターフェイス	306
6.4.1	make_pipelineによる簡便なパイプライン生成	308
6.4.2	ステップ属性へのアクセス	309
6.4.3	GridSearchCV内のパイプラインの属性へのアクセス	309
6.5	前処理ステップとモデルパラメータに対するグリッドサーチ	311
6.6	グリッドサーチによるモデルの選択	314
6.7	まとめと展望	315
<b>7章</b>	<b>テキストデータの処理</b>	<b>317</b>
7.1	文字列として表現されているデータのタイプ	317
7.2	例題アプリケーション：映画レビューのセンチメント分析	319
7.3	Bag of Wordsによるテキスト表現	322
7.3.1	トイデータセットに対するBoW	323
7.3.2	映画レビューのBoW	324
7.4	ストップワード	329
7.5	tf-idfを用いたデータのスケール変換	330
7.6	モデル係数の調査	333
7.7	1単語よりも大きい単位のBag-of-Words (n-グラム)	334
7.8	より進んだトークン分割、語幹処理、見出し語化	339
7.9	トピックモデリングと文書クラスタリング	343
7.9.1	LDA (Latent Dirichlet Allocation)	343
7.10	まとめと展望	351
<b>8章</b>	<b>おわりに</b>	<b>353</b>
8.1	機械学習問題へのアプローチ	353
8.1.1	人間をループに組み込む	354
8.2	プロトタイプから運用システムへ	354
8.3	運用システムのテスト	355
8.4	独自Estimatorの構築	356
8.5	ここからどこへ行くのか	357
8.5.1	理論	357
8.5.2	他の機械学習フレームワークとパッケージ	357
8.5.3	ランキング、推薦システム、その他の学習	358

---

8.5.4	確率モデル、推論、確率プログラミング	359
8.5.5	ニューラルネットワーク	359
8.5.6	大規模データセットへのスケール	360
8.5.7	名誉を得る	361
8.6	結論	361
	索引	363