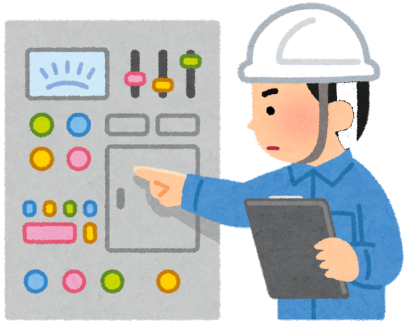
まとめ

- 全体
- 識別
- 回帰
- モデル推定
- パターンマイニング
- 推薦図書等

全体のまとめ

- 機械学習の位置づけ（ p.5, 図 1.3 ）



観測データ
(134.1, 34.6, 12.9)
(135.5, 30.1, 43.0)
…

行動ログ
(検索 = ハブ、購入 = ルータ)
(クリック = メモリ、購入 =USB メモリ)
…

分類ログ
(記事 1, yes)
(記事 2, no)
…

機械学習

異常値の発見　関数　推薦　識別器

現実世界の
複雑な現象

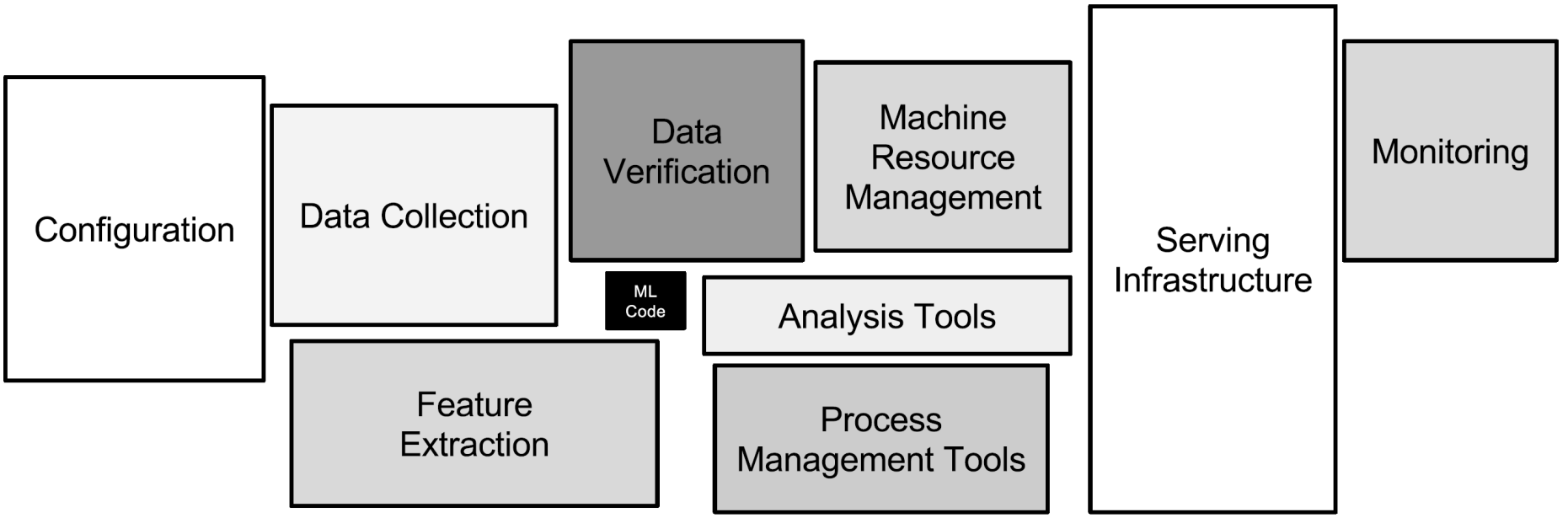
観測

モデルの作成

本講座の範囲

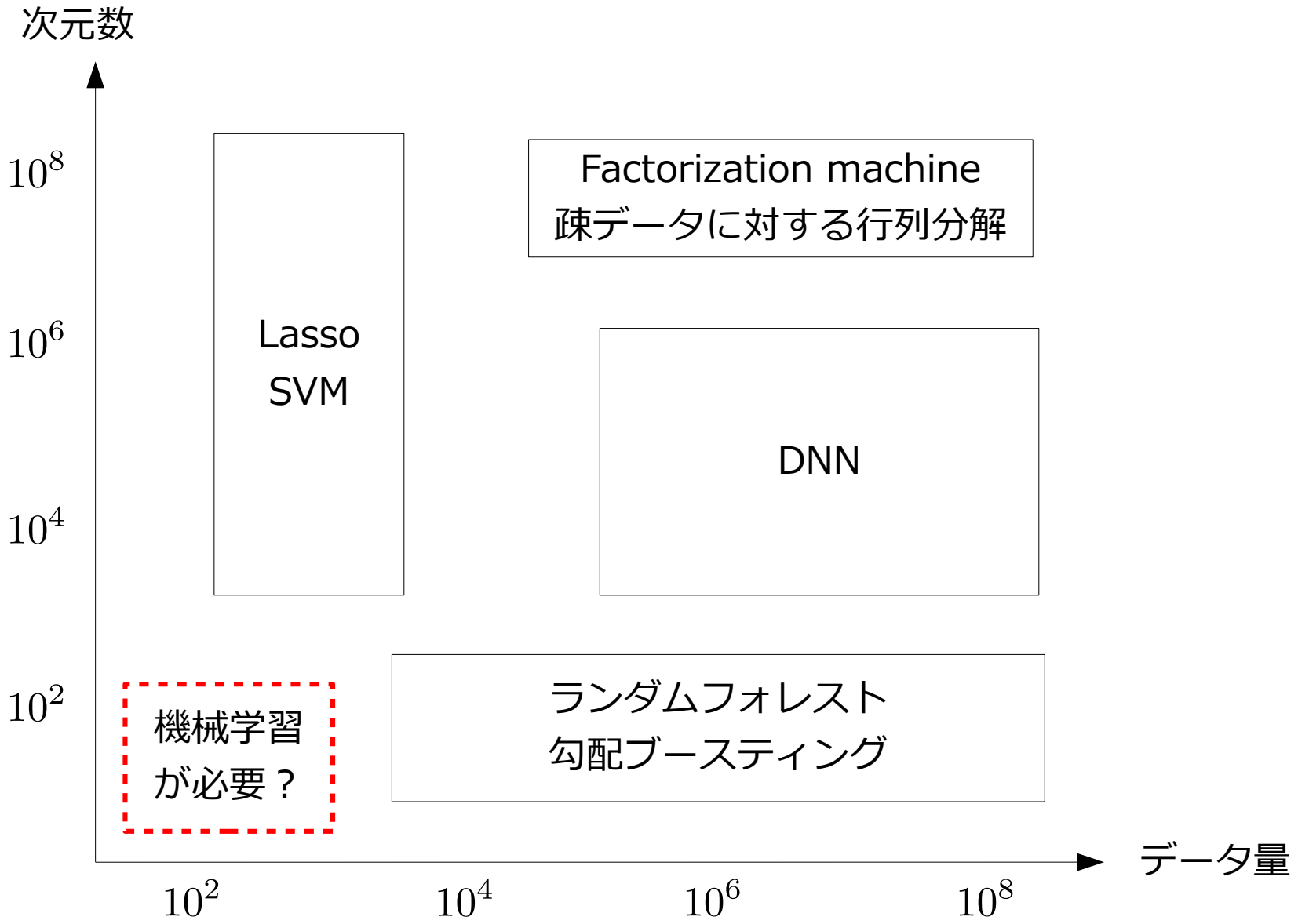
モデルの適用

機械学習システムを実運用する技術



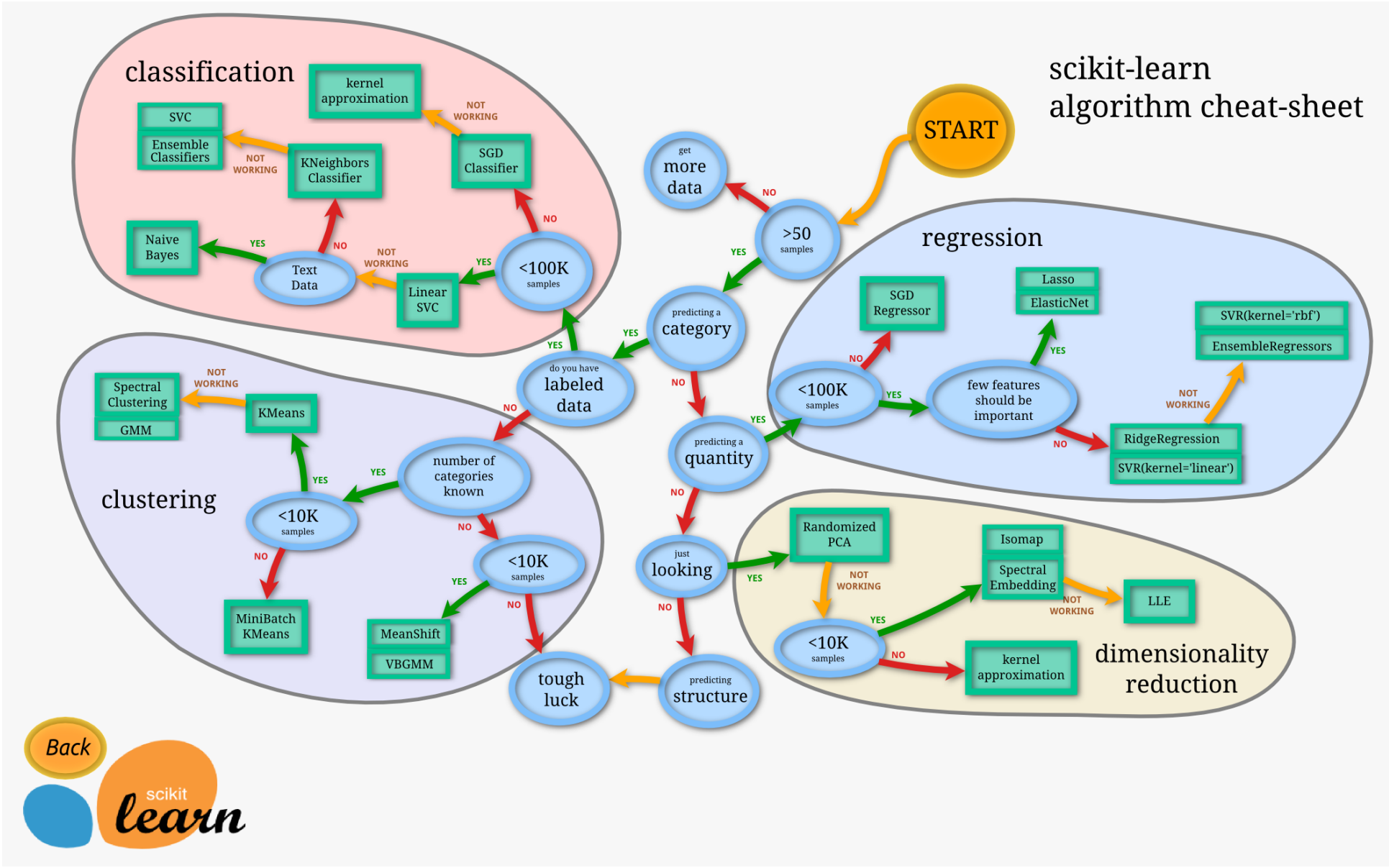
- Sculley, D. et.al.: Hidden Technical Debt in Machine Learning Systems, NIPS 2015.
  - https://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.pdf

データ量・特徴次元数・学習手法の関係

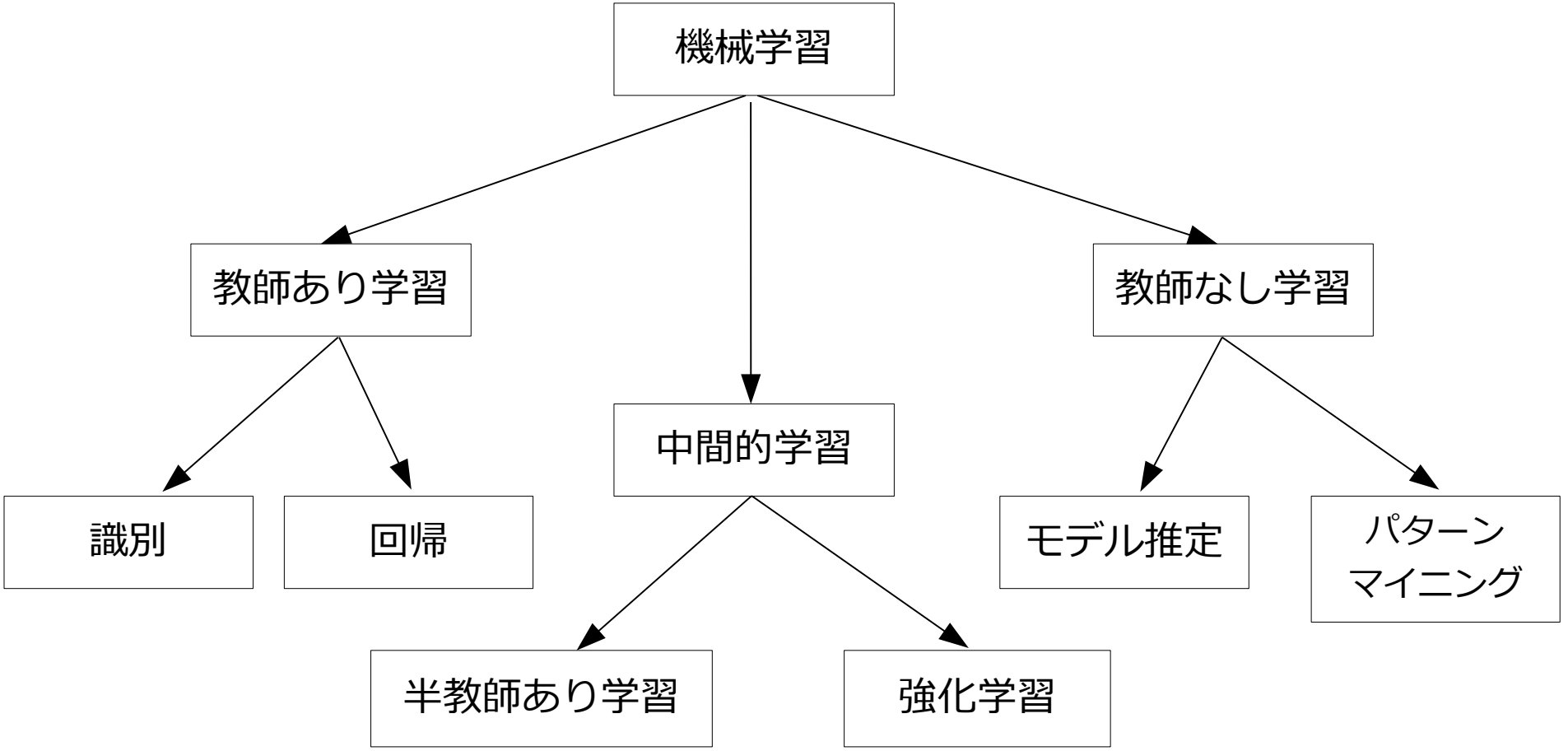


データ量・特徴次元数・学習手法の関係

http://scikit-learn.org/stable/tutorial/machine_learning_map/index.html



1.3 機械学習の分類


機械学習
教師あり学習
中間的学習
教師なし学習
識別
回帰
モデル推定
パターン
マイニング
半教師あり学習
強化学習

識別のまとめ

- Step1: データの性質を知る
  - 1-1: 主成分分析で 2 次元に変換し、プロット
    - 累積寄与率の確認が必要
  - 1-2: ベースライン性能の見当をつける
    - 用いる手法： k-NN 法、単純ベイズ、ロジステック識別
    - スコアの評価：

生成モデル　　識別モデル

      - すべて高い：良質のデータ。 Step2 へ
      - すべて低い：質の悪いデータ。特徴の見直しを
      - 極端に違う：データ数が少なすぎる可能性あり

識別のまとめ

- Step2: 識別器の作成
  - SVM
  - ランダムフォレスト
  - 勾配ブースティング
  - ハイパーパラメータの調整も行う
    - Grid サーチ　or　ランダムサーチ
    - 連続値をとるハイパーパラメータは桁を変えて試す
    - 良い性能になりそうなところをさらに細かくサーチする

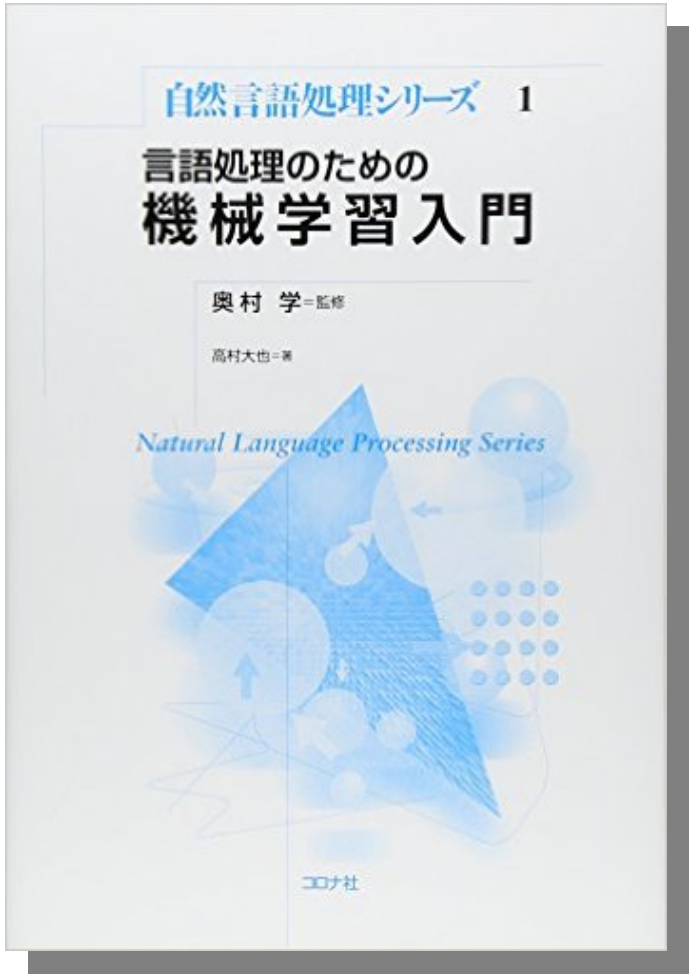
識別のまとめ

- Step3: 評価
  - データが少ない場合はひとつ抜きまたは 10-fold CV
    - 正確な性能評価にはデータをシャッフルして CV を繰り返す
  - データが多い場合は分割学習法
    - 学習データ・検証データ・評価データ
      - 学習データで学習、検証データでハイパーパラメータ調整を繰り返す
      - 最後に評価データで実稼働時の性能を予測

識別のまとめ

- Step4: 解釈
  - 必ず混同行列を見て結果を解釈する
    - どの性能に着目すべきか
      - 正解率／精度／再現率／ F 値
      - クラス間でデータ数に大きな違いがあるときはマクロ平均
  - 最初に目標を設定して、それがクリアできれば OK
    - スコア 100% はありえない
    - どの程度のスコアが達成できれば、どのような効果があるか、最初に見積もっておく

推薦図書



高村著 コロナ社 , 2010

1. 必要な数学的知識
2. 文書および単語の数学的表現
3. クラスタリング
4. 分類
5. 系列ラベリング
6. 実験の仕方など

推薦図書

平井著 森北出版 , 2012

第 1 章 はじめに

第 2 章 識別規則と学習法の概要

第 3 章 ベイズの識別規則

第 4 章 確率モデルと識別関数

第 5 章 k 最近傍法（ kNN 法）

第 6 章 線形識別関数

第 7 章 パーセプトロン型学習規則

第 8 章 サポートベクトルマシン

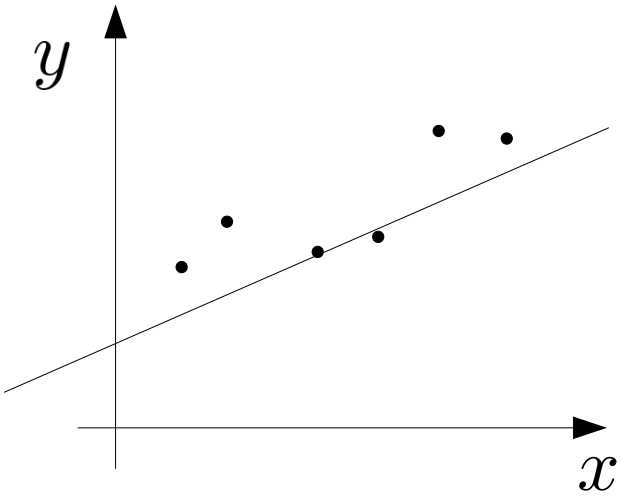
第 9 章 部分空間法

第 10 章 クラスタリング

第 11 章 識別器の組み合わせによる性能強化

回帰のまとめ

- Step1: データの性質を知る
  - 線形回帰の性能で見当をつける



$$\hat{c}(\boldsymbol{x}) = \sum_{i=0}^{d} w_i x_i$$

  - 入力 $\boldsymbol{x}$ から出力 y を求める回帰式を 1 次式に限定
  - 解析的に係数 $\boldsymbol{w}$ が求まる

回帰のまとめ

- Step2: 基底関数・正則化項を導入し、性能の向上を試みる
  - 基底関数 $\boldsymbol{\phi}(\boldsymbol{x}) = (\phi_1(\boldsymbol{x}), \ldots, \phi_b(\boldsymbol{x}))$ を考える

$$\hat{c}(\boldsymbol{x}) = \sum_{j=0}^{b} w_j \phi_j(\boldsymbol{x})$$

例）高次式による回帰
サポートベクトル回帰

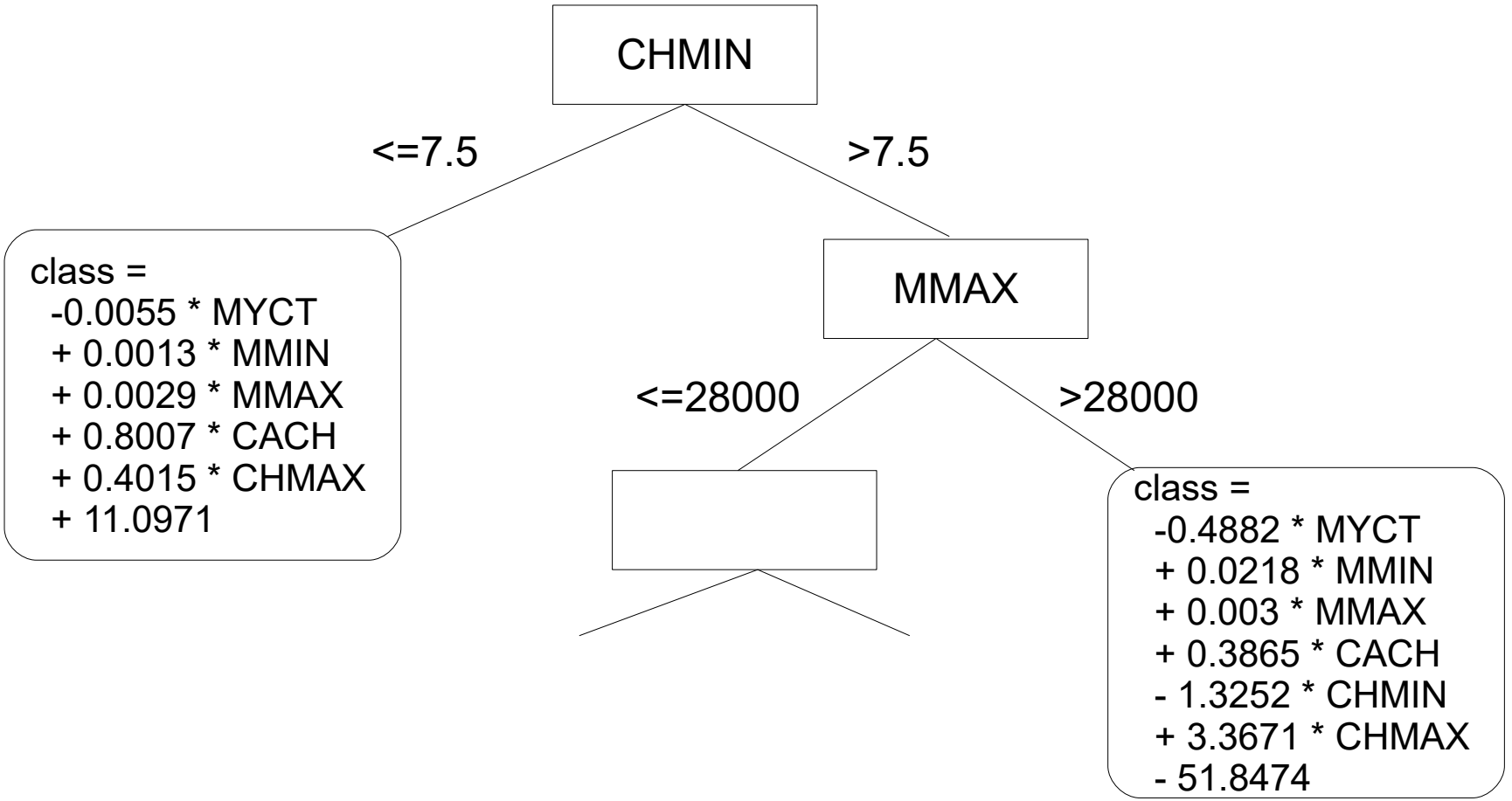
  - 正則化項の導入

    → 複雑なパラメータ $\boldsymbol{w}$ （過学習）の回避

    - L1 ノルム $|\boldsymbol{w}|$ ： 0 となるパラメータが多くなる — Lasso
    - L2 ノルム $\|\boldsymbol{w}\|^2$ ： パラメータを 0 に近づける — Ridge

回帰のまとめ

- Step3: モデル木やアンサンブル学習を試す



回帰のまとめ

- Step4: 評価
  - 誤差の二乗和：手法間の評価に有効
  - 相関係数：出力と正解とがどの程度似ているか
  - 決定係数：相関係数の 2 乗

Weka の結果表示例

```
=== Cross-validation ===
=== Summary ===

Correlation coefficient          0.9012
Mean absolute error             41.0886
Root mean squared error         69.556
Relative absolute error         42.6943 %
Root relative squared error     43.2421 %
Total Number of Instances       209
```

決定係数の式

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{c}(x_i))^2}{\sum_{i=1}^{N}(y_i - \tilde{y})^2}$$

$\tilde{y}$： y の平均

推薦図書

髙橋著 オーム社 , 2005

プロローグ ノルンへようこそ！

第 1 章 基礎知識

第 2 章 回帰分析

第 3 章 重回帰分析

第 4 章 ロジスティック回帰分析

付録　Excel で計算してみよう！

モデル推定のまとめ

- クラスタリング
  - 教師なしデータから、まとまりを発見する
  - 階層的手法
    - ボトムアップに小さなまとまりを結合
  - 分割最適化手法
    - k-means： 分割数を予め与える
    - X-means, label propagation： 分割数を自動的に決定
  - 確率密度推定
    - EM アルゴリズム

モデル推定のまとめ

- 異常検出とは
  - 正常クラスのデータと、それ以外のデータとのクラスタリング
  - 外れ値検知、変化点検出、異常状態検出など
- 局所異常因子による外れ値検知
  - LOF: 対象データの周辺密度と、近くの k 個のデータの周辺密度の平均との比
- One-class SVM
  - RBF カーネルを使うと、元の空間の孤立した点は変換後の空間の原点付近に対応する

推薦図書

石井・上田著 オーム社 , 2014

第 1 章　ベイズ統計学
第 2 章　事前確率と事後確率
第 3 章　ベイズ決定則
第 4 章　パラメータ推定
第 5 章　教師付き学習と教師なし学習
第 6 章　 EM アルゴリズム
第 7 章　マルコフモデル
第 8 章　隠れマルコフモデル
第 9 章　混合分布のパラメータ推定
第 10 章　クラスタリング
第 11 章　ノンパラメトリックベイズモデル
第 12 章　ディリクレ過程混合モデルによるクラスタリング
第 13 章　共クラスタリング

パターンマイニングのまとめ

- バスケット分析
  - 支持度を基準に頻出項目集合を抽出

$$support(A) = \frac{T_A}{T}$$

$P(A)$ に対応

  - 確信度またはリフト値の高い規則を抽出

$$\mathrm{confidence}(\mathrm{A} \rightarrow B) = \frac{support(A \cup B)}{support(A)} = \frac{T_{A \cup B}}{T_A}$$

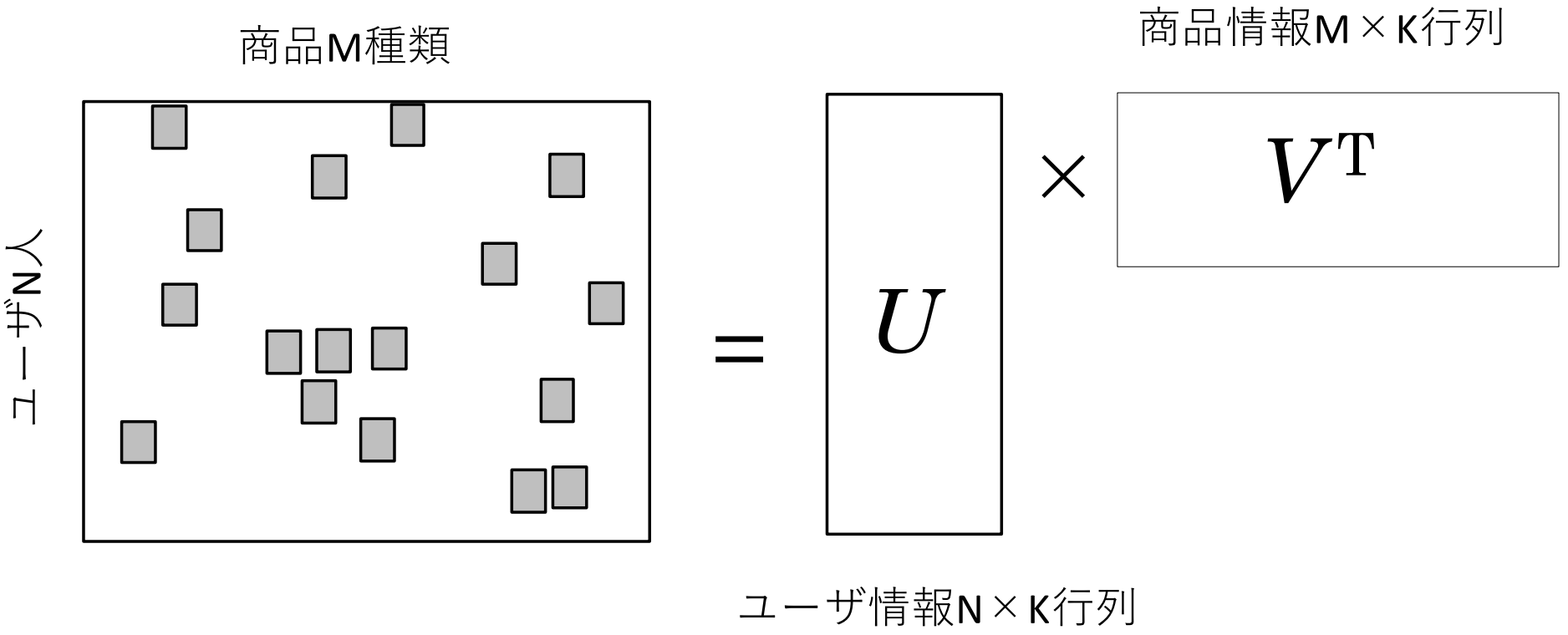
$$\frac{P(A,B)}{P(A)} = \frac{P(B|A)P(A)}{P(A)} = P(B|A)$$

$$\mathrm{lift}(\mathrm{A} \rightarrow B) = \frac{confidence(A \rightarrow B)}{support(B)}$$

$$\frac{P(B|A)}{P(B)} = \frac{P(A)P(B|A)}{P(A)P(B)} = \frac{P(A,B)}{P(A)P(B)}$$

パターンマイニングのまとめ

- 協調フィルタリング
  - アイデア：疎な行列は低次元の行列の積で近似できる
  - 値のある部分だけで行列分解を行う
  - 空所の値を予測する



推薦文献

本稿の構成より

第 I 部では，推薦システムとは何か，またその設計指針や分類について述べる．

第 II 部では，データの入力，嗜好の予測，そして推薦の提示からなる推薦システムの実行過程について述べる．

第 III 部では，さまざまな嗜好の予測アルゴリズムのを紹介する．

第 IV 部では，推薦システムに関連する話題や，さまざまな状況での推薦を紹介する．

第 V 部は関連資料の紹介とまとめである．

神嶌著　2016
http://www.kamishima.net/archive/recsysdoc.pdf

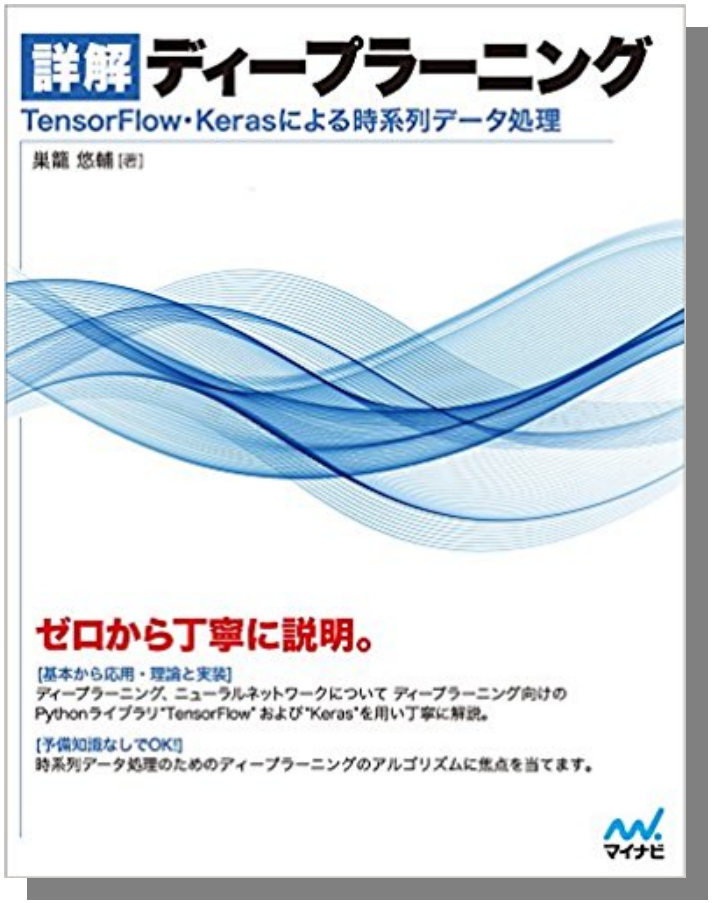
全般的な推薦図書等

推薦図書

杉山著 講談社 , 2013

第 I 部 はじめに
第 1 章 機械学習とは
第 2 章 学習モデル
第 II 部 教師付き回帰
第 3 章 最小二乗学習
第 4 章 制約付き最小二乗学習
第 5 章 スパース学習
第 6 章 ロバスト学習
第 III 部 教師付き分類
第 7 章 最小二乗学習に基づく分類
第 8 章 サポートベクトル分類
第 9 章 アンサンブル分類
第 10 章 確率的分類
第 11 章 系列データの分類
第 IV 部 教師なし学習
第 12 章 異常検出
第 13 章 教師なし次元削減
第 14 章 クラスタリング
第 V 部 発展的話題
第 15 章 オンライン学習
第 16 章 半教師付き学習
第 17 章 教師付き次元削減
第 18 章 転移学習
第 19 章 マルチタスク学習
第 VI 部 おわりに
第 20 章 まとめと今後の展望

推薦図書



巣籠著 マイナビ出版 , 2017

第 1 章 数学の準備

第 2 章　Python の準備

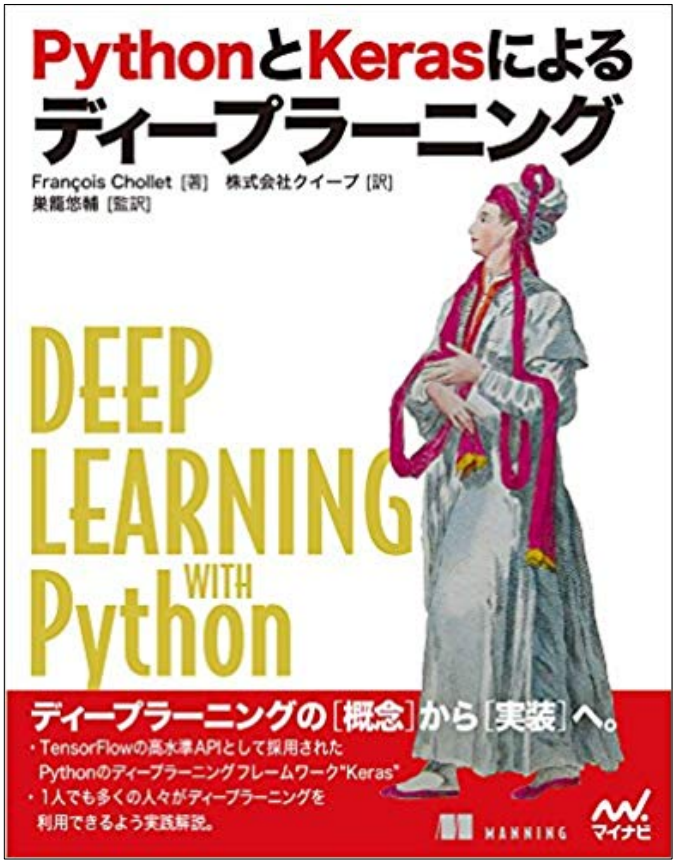
第 3 章 ニューラルネットワーク

第 4 章 ディープニューラルネットワーク

第 5 章 リカレントニューラルネットワーク

第 6 章 リカレントニューラルネットワークの応用

推薦図書



Francois Chollet 著，巣籠 訳
マイナビ出版，2018

1章　ディープラーニングとは何か
2章　予習：ニューラルネットワークの数学的要素
3章　入門：ニューラルネットワーク
4章　機械学習の基礎
5章　コンピュータビジョンのためのディープラーニング
6章　テキストとシーケンスのためのディープラーニング
7章　高度なディープラーニングのベストプラクティス
8章　ジェネレーティブディープラーニング
9章　本書のまとめ

推薦コース

- オンライン学習 (udemy)
  - Python で機械学習： scikit-learn で学ぶ識別入門
    - https://www.udemy.com/python-scikit-learn/
    - 講師：玉木 徹（広島大学 准教授）
    - 合計 9 時間のビデオレクチャー