まとめ

- 全体
- 識別
- 回帰
- モデル推定
- パターンマイニング

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

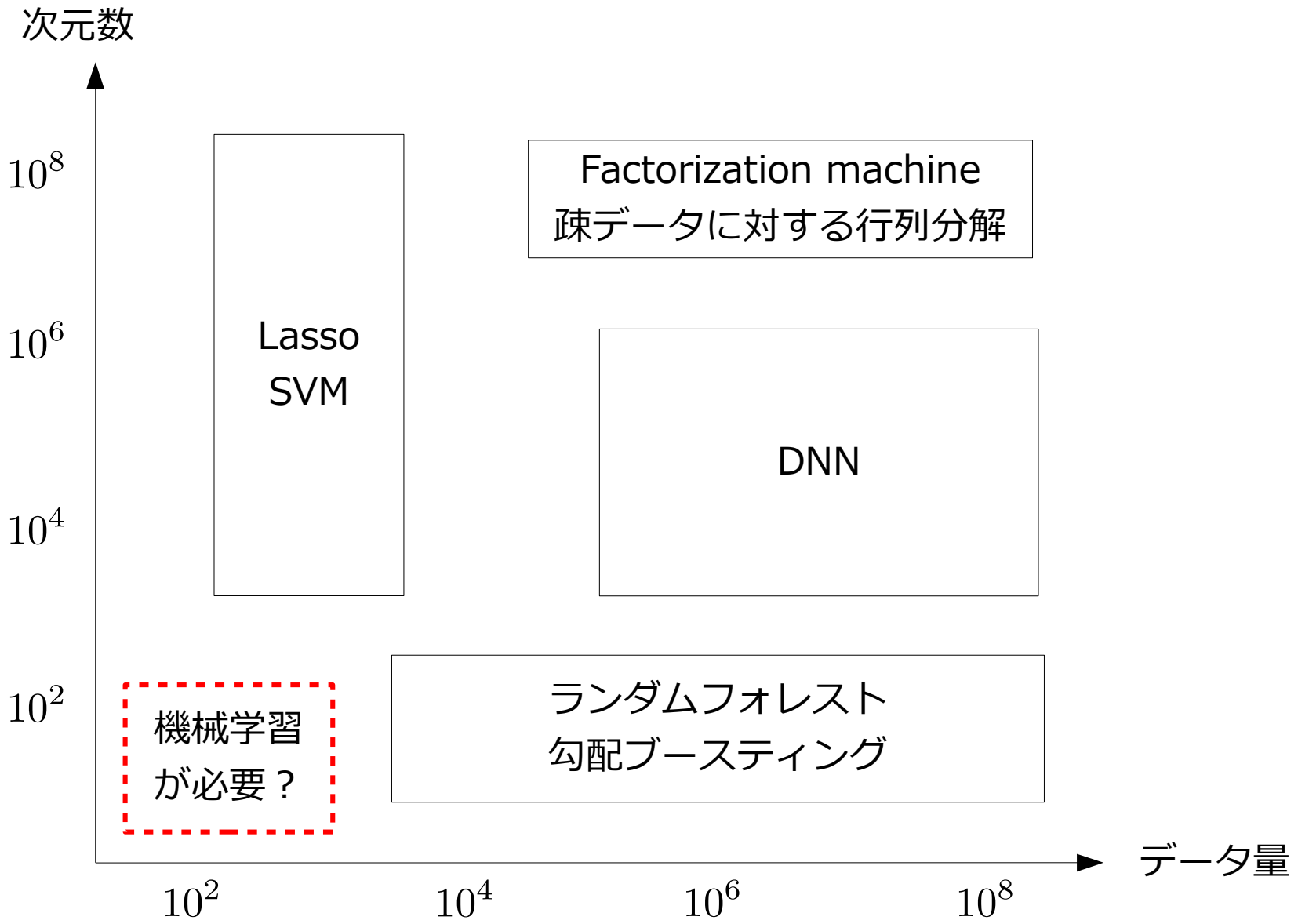
観測

モデルの作成

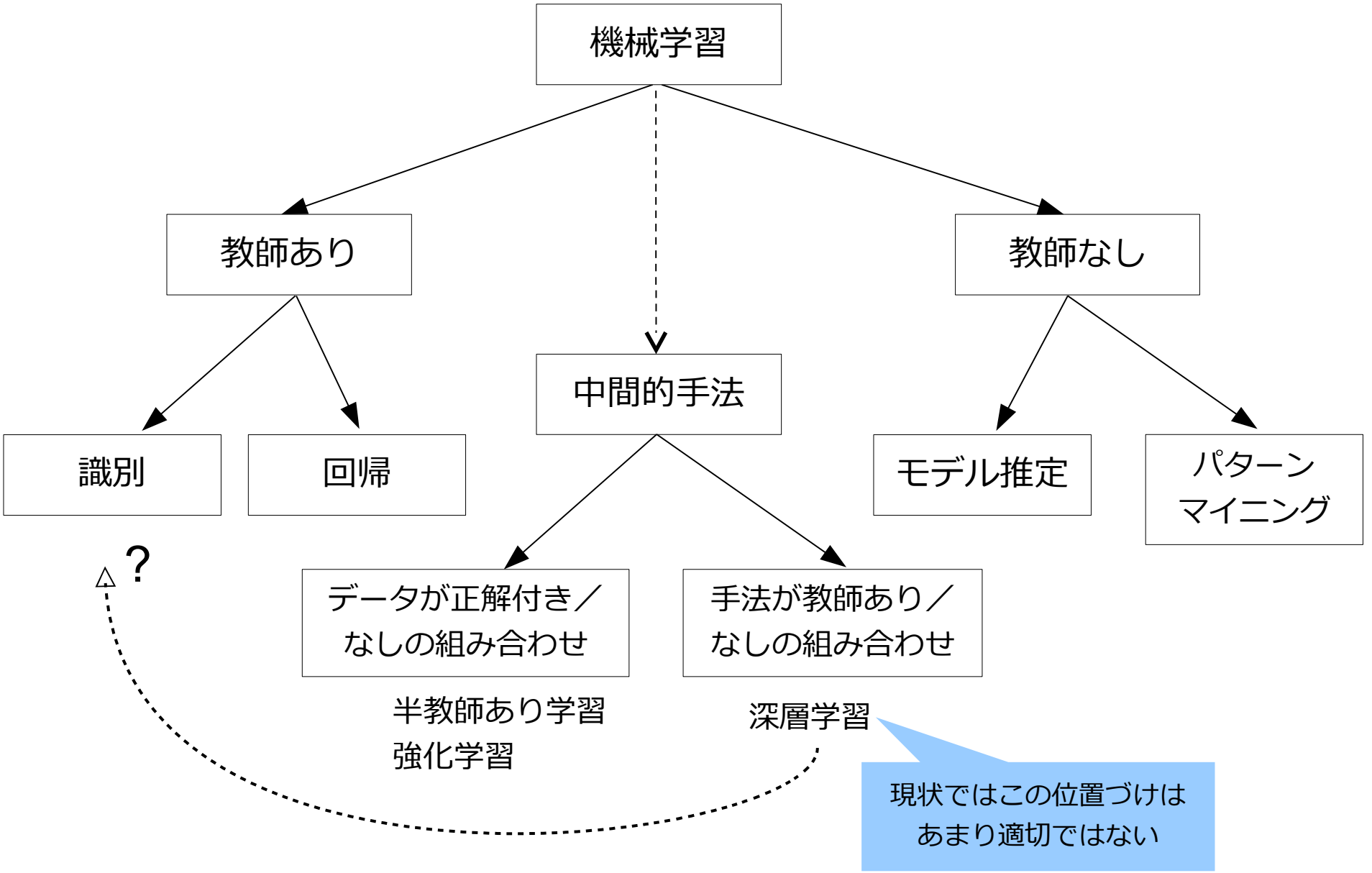
本講座の範囲

モデルの適用

データ量・特徴次元数・学習手法の関係



1.3 機械学習の分類


機械学習
教師あり
教師なし
中間的手法
識別
回帰
モデル推定
パターン
マイニング
?
データが正解付き／
なしの組み合わせ
手法が教師あり／
なしの組み合わせ
半教師あり学習
強化学習
深層学習
現状ではこの位置づけは
あまり適切ではない

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
    - 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

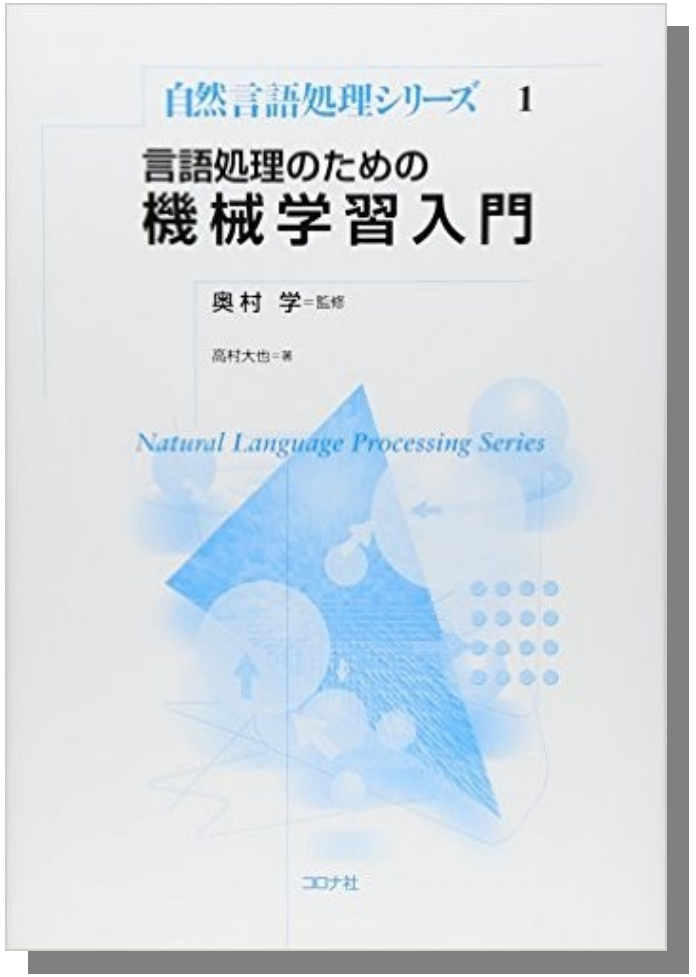
識別のまとめ

- Step3: 評価
  - データが少ない場合はひとつ抜きまたは 10-fold CV
  - 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
      - 特にクラス間でデータ数に大きな違いがあるときは要注意
  - 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

参考図書



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

参考図書

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

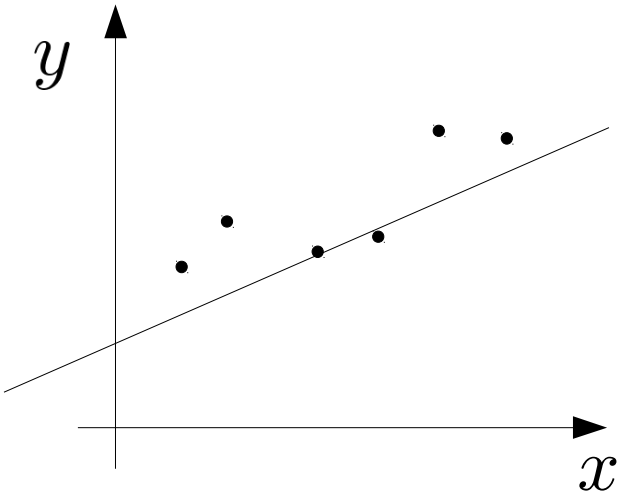
第 11 章 識別器の組み合わせによる性能強化

参考図書

杉山著 講談社 , 2013

回帰のまとめ

- Step1: データの性質を知る
  - 線形回帰の性能で検討をつける



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

  - 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

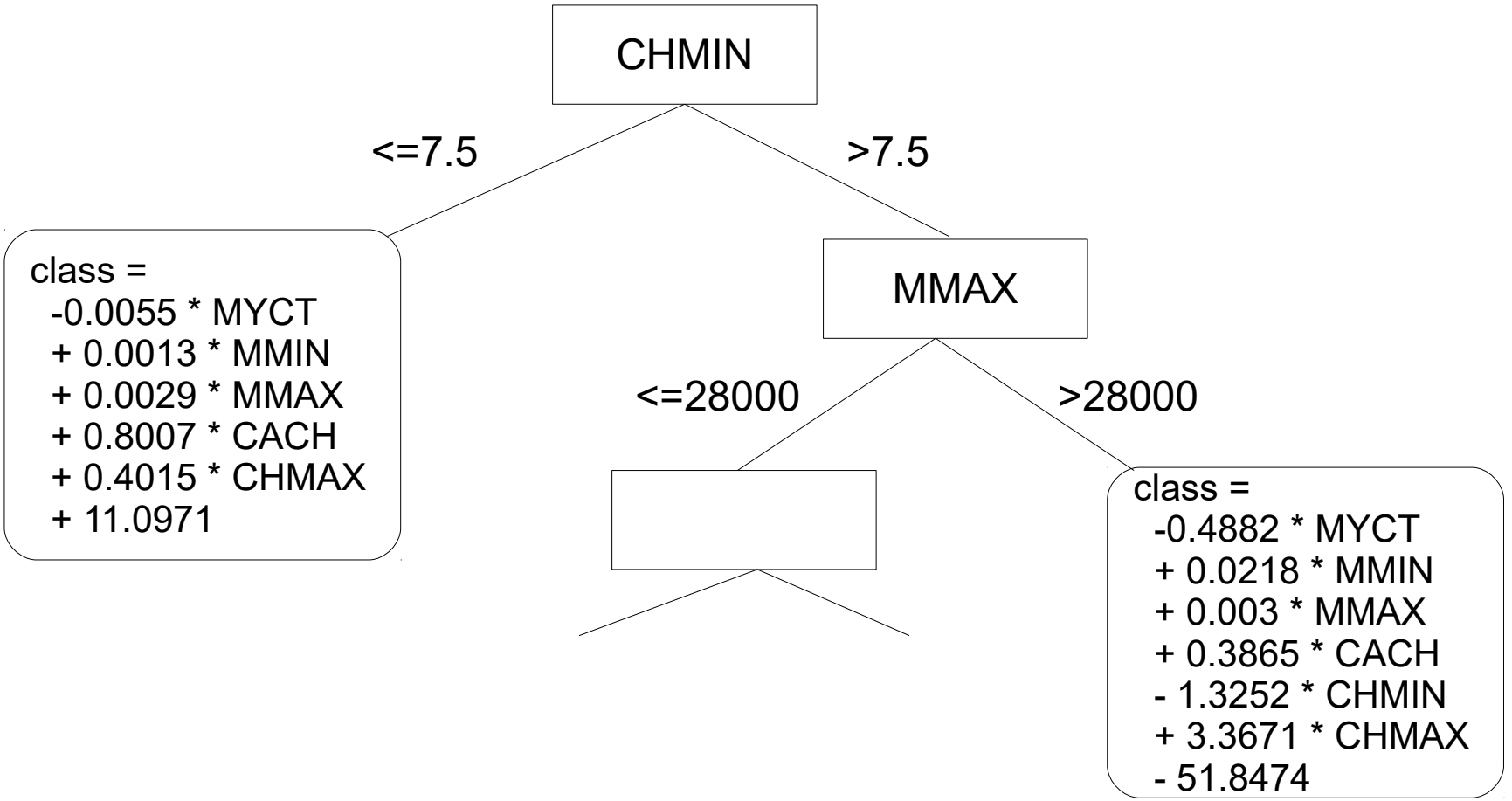
  - 正則化項の導入

    → 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

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

$\tilde{y}$： y の平均

参考図書

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
- 外れ値検知（静的異常検出）
  - データの分布から大きく離れている値を見つける
    - 近くにデータがないか、あるいは極端に少ないものを外れ値とみなす

参考図書

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

$$support(items) = \frac{T_{items}}{T}$$

  - 確信度またはリフト値の高い規則を抽出

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

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

パターンマイニングのまとめ

- 協調フィルタリング　応用例）通販サイトでの推薦
  - メモリベース法
    - k-NN 法と似たアイディアで、振る舞いが似ているトランザクションを探す
  - モデルベース法
    - 事前に似たような振る舞いをするユーザの集合をクラスタリングで得る
    - 入力がどのクラスタに入るかを判別し、そのクラスタの購買履歴から推薦を行う

参考）神嶌敏弘：推薦システムのアルゴリズム
http://www.kamishima.net/archive/recsysdoc.pdf

参考図書

杉山著 講談社 , 2015

第 1 章 確率変数と確率分布

第 2 章 離散型確率分布の例

第 3 章 連続型確率分布の例

第 4 章 多次元確率分布の性質

第 5 章 多次元確率分布の例

第 6 章 任意の確率分布に従う標本の生成

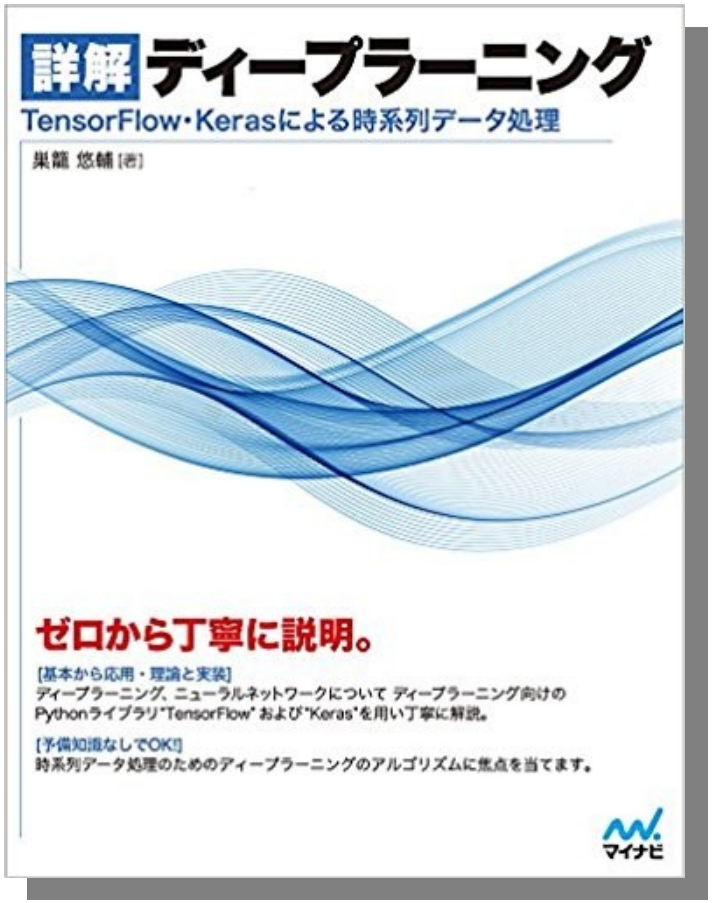
第 7 章 独立な確率変数の和の確率分布

第 8 章 確率不等式

第 9 章 統計的推定

第 10 章 仮説検定

参考図書



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

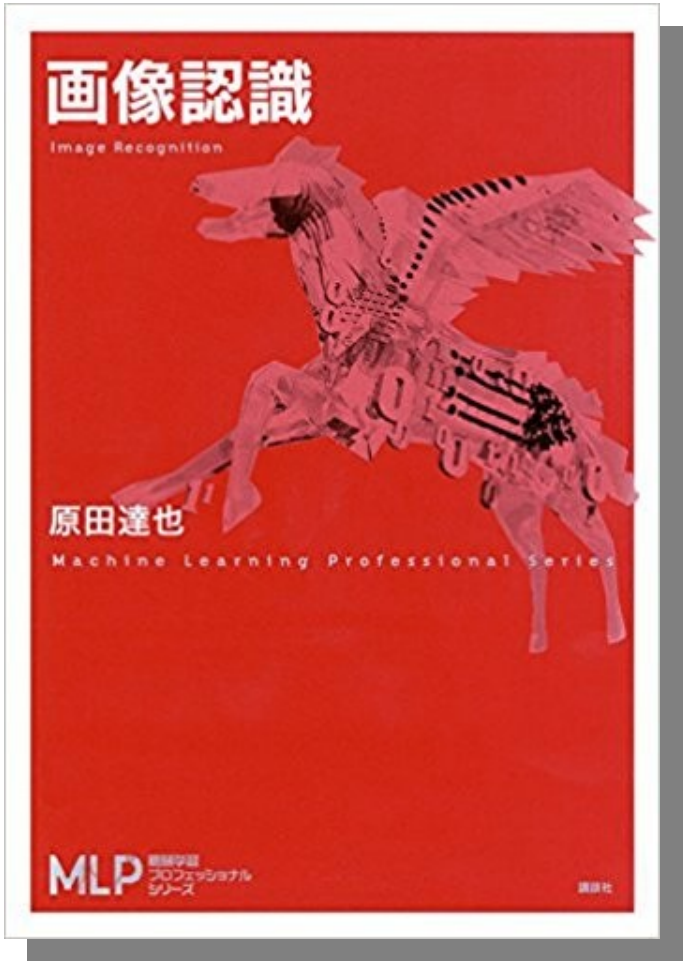
第 3 章 ニューラルネットワーク

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

参考図書



原田著 講談社 , 2017

第 1 章 画像認識の概要

第 2 章 局所特徴

第 3 章 統計的特徴抽出

第 4 章 コーディングとプーリング

第 5 章 分類

第 6 章 畳み込みニューラルネットワーク

第 7 章 物体検出

第 8 章 インスタンス認識と画像検索

第 9 章 さらなる話題

参考図書

黒橋著 放送大学教育振興会 , 2015

自然言語処理の概要と歴史

文字列・テキスト処理の基礎

系列の解析

コーパスに基づく自然言語処理

意味の解析

構文の解析

文脈の解析

情報抽出と知識獲得

情報検索

対話システム

機械翻訳

まとめ