機械学習実践入門

pythonで誰でもカンタン機械学習

by Yoshiyuki Kato, 2015/10/22@IMELab

目次

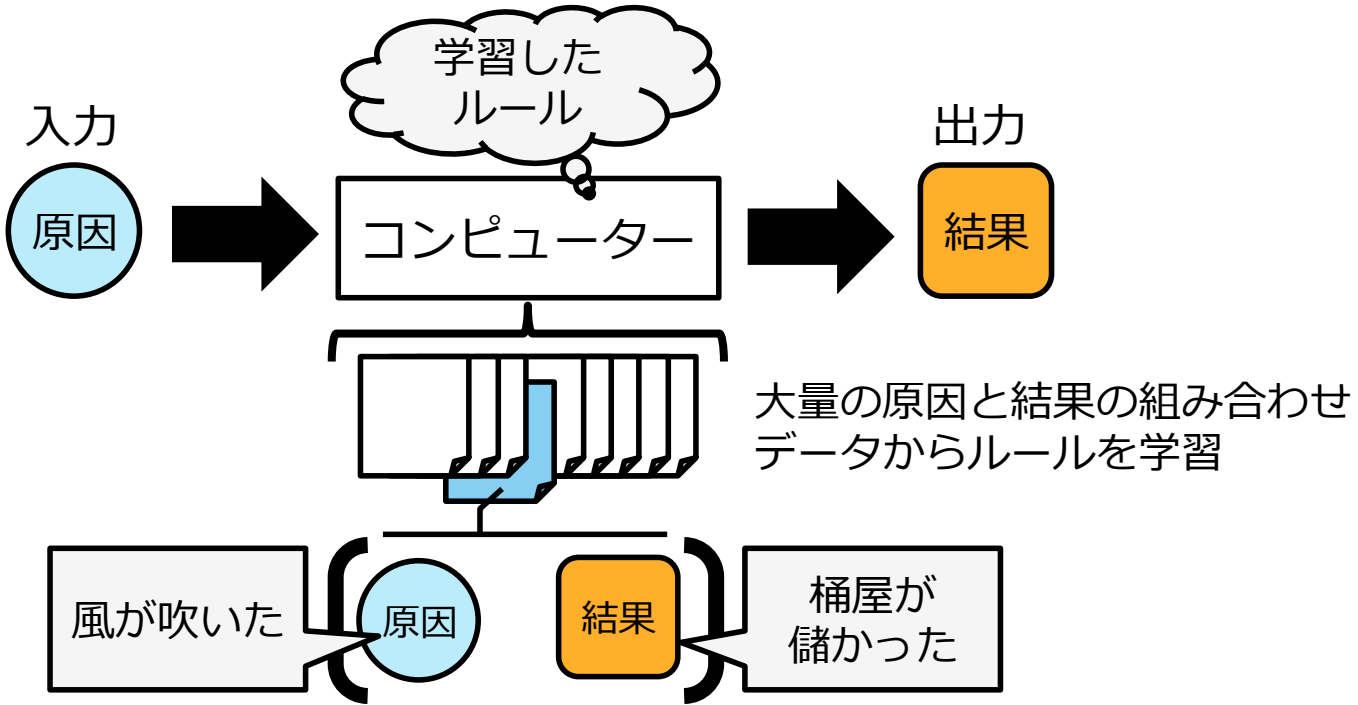
- 基礎編：機械学習について
  - 機械学習とは
  - 機械学習の仕組み
- 実践編：Scikit-learnを用いた機械学習
  - 環境の整備
  - サンプルの実行
- おまけ：Chainerを用いた深層学習
  - 深層学習について
  - サンプルの実行
- 補足とまとめ
- 書籍紹介

基礎編：機械学習について

難しそうに見えますか？実はそうでもないんです

機械学習ってなに？

- 機械学習とは、**原因と結果の間にある関係性**を、統計的に明らかにする手法のことである
  - 大量の原因と結果の組み合わせデータからルールを抽出すれば、コンピューターは入力された原因に応じて結果を出力できる
  - データからルールを抽出するプロセスのことを、**学習**と呼ぶ



いろいろな機械学習

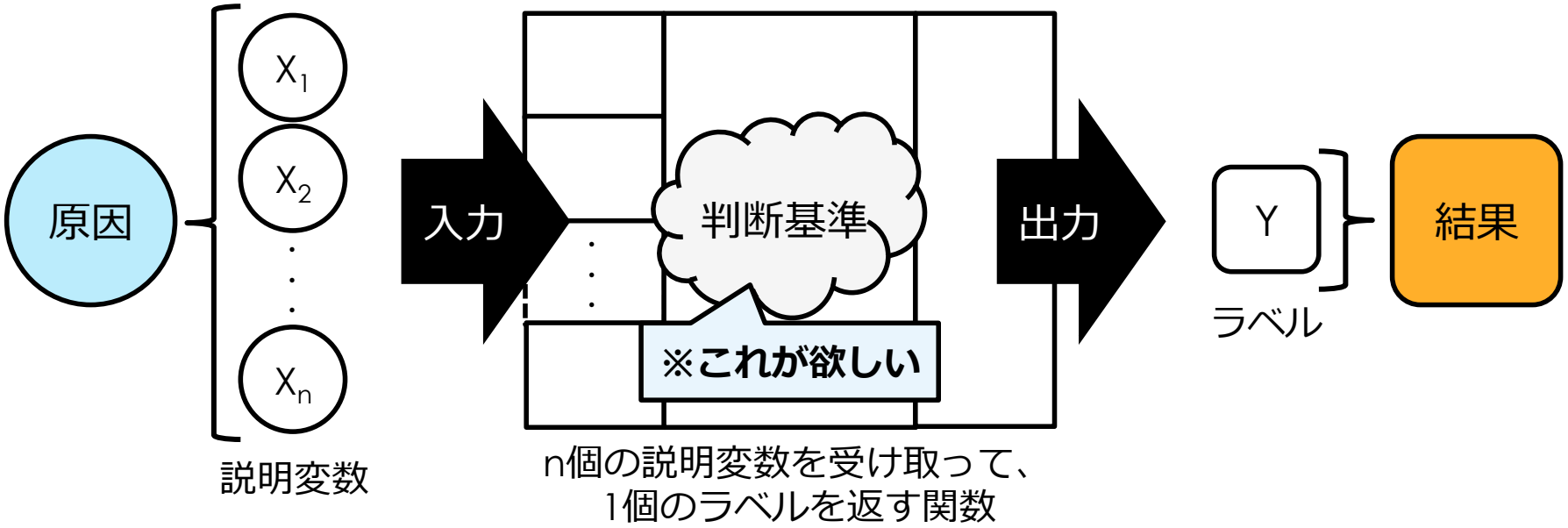
- 教師あり学習
  - 最もポピュラーで強力な機械学習手法。学習用の原因-結果データから、データの母集団におけるルールの推定を行う。
  - スパム検知、レコメンド、画像認識、音声認識
- 教師なし学習
  - データマイニングを機械学習的な手法によって行った場合の呼び方。教師あり学習が個々のデータに着目するのに対し、教師なし学習は全体的な傾向に着目する。
  - クラスタリング、特徴選択
- 強化学習
  - エージェントと呼ばれるシステムが、周囲の状況に応じて経験的に次にとるべき行動を決定するための手法。ロボットの動作学習などに用いられる。

いろいろな機械学習

- 教師あり学習 ✓
  - 最もポピュラーで強力な機械学習手法。学習用の原因-結果データから、データの母集団におけるルールの推定を行う。
  - スパム検知、レコメンド、画像認識、音声認識
- 教師なし学習
  - データマイニングを機械学習的な手法によって行った場合の呼び方。教師あり学習が個々のデータに着目するのに対し、教師なし学習は全体的な傾向に着目する。
  - クラスタリング、特徴選択
- 強化学習
  - エージェントと呼ばれるシステムが、周囲の状況に応じて経験的に次にとるべき行動を決定するための手法。ロボットの動作学習などに用いられる。

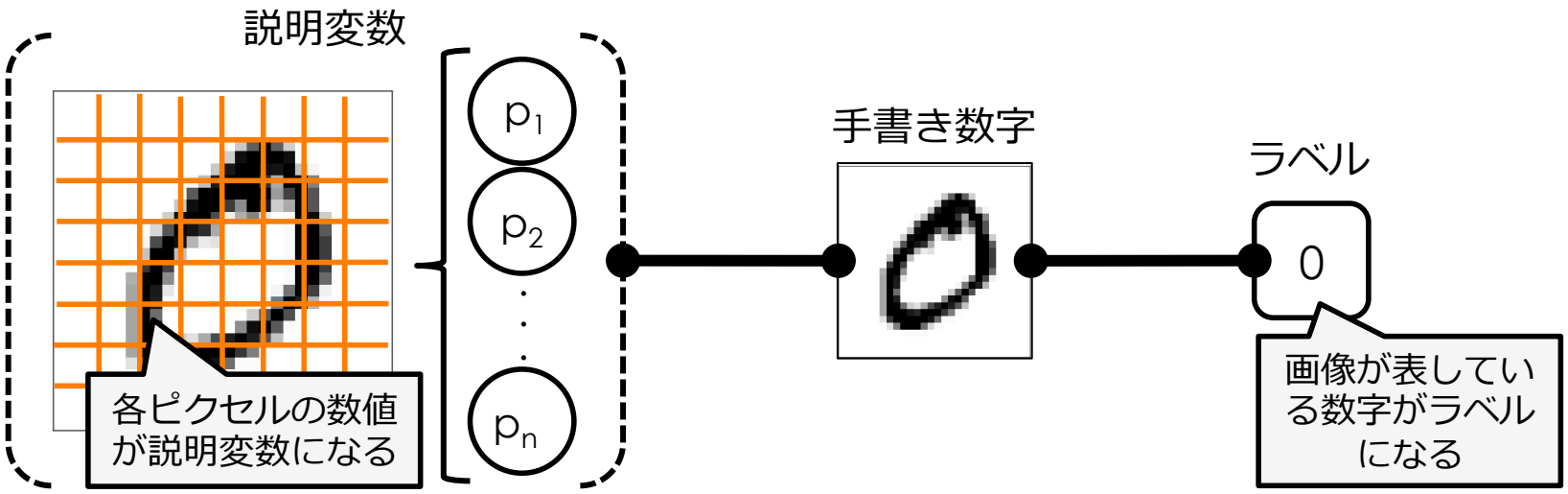
教師あり学習

- 教師あり学習では、原因-結果のデータを用いる。このとき、原因データのことを**説明変数**、結果データのことを**被説明変数(ラベル)**と呼ぶ。
- 教師あり学習が目的とするのは、説明変数を入力したら、ラベルを出力するような関数。つまり、原因に応じて結果を判断する物差しである。



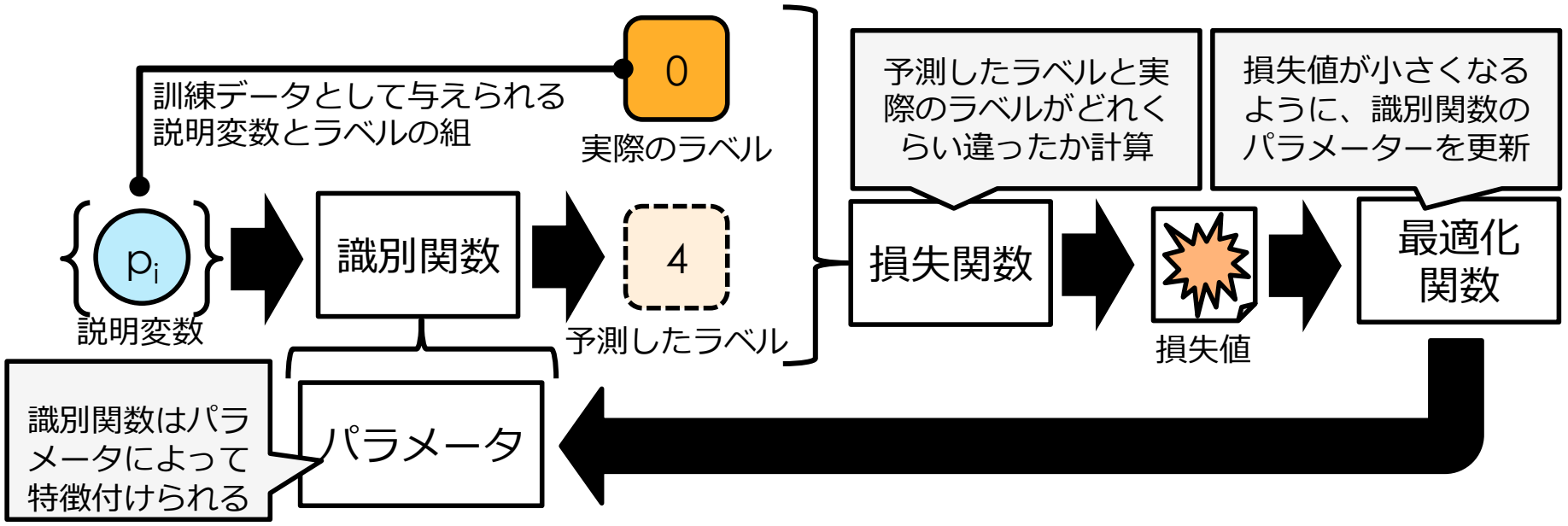
教師あり学習による手書き数字認識

- 教師あり学習を用いて、手書き数字画像が0~9のどの数字を表しているのかを判別する
- 手書き数字画像の各ピクセルが説明変数（原因）、画像が表している数字がラベル（結果）になる
- ピクセルと数字の組み合わせパターンを分析し、どんな手書き数字画像がどの数字を表しているのかを学習する



学習プロセス

- 教師あり学習のプロセスは、**訓練**と**評価**からなる。
  - 訓練とは、説明変数とラベルの組から関係性を学習する過程のこと。このときに用いるデータを訓練データと呼ぶ
  - 評価とは、訓練を経て得られた判断基準を検証する過程のこと
- 訓練は主に**識別関数（学習器）、損失関数、最適化関数**の三つによって実行される



実践編：SCIKIT-LEARNを用いた機械学習

10行以内のプログラムで動く機械学習ツール

環境の確認（Macですでにpythonを使っている人）

1. ターミナルを開く
2. pythonのバージョンを確認する（下記コマンドを打つと確認できます）
   - 2.7以外の場合は、homebrewからインストールするか、2.7系の環境を新しく準備する（次のスライドで説明します）

```
$ python --version
```

環境の準備（pythonを使ったことがない人）

1. Canopyをインストールする
    1. https://store.enthought.com/downloads/#default
    2. SFCのメールアドレスを使ってユーザ登録すると、アカデミックバージョンが使える
    3. デフォルトのpythonをCanopyのpythonに設定させようとしてくるので全力で阻止
2. パッケージマネージャからライブラリをインストール
    1. numpy
    2. scipy
    3. scikit-learn
    4. ipython
3. ToolsからCanopy Terminalを開く

サンプルの起動

- terminal（Canopy Terminal）で、配布資料のディレクトリに移動し、次のコマンドを順に実行

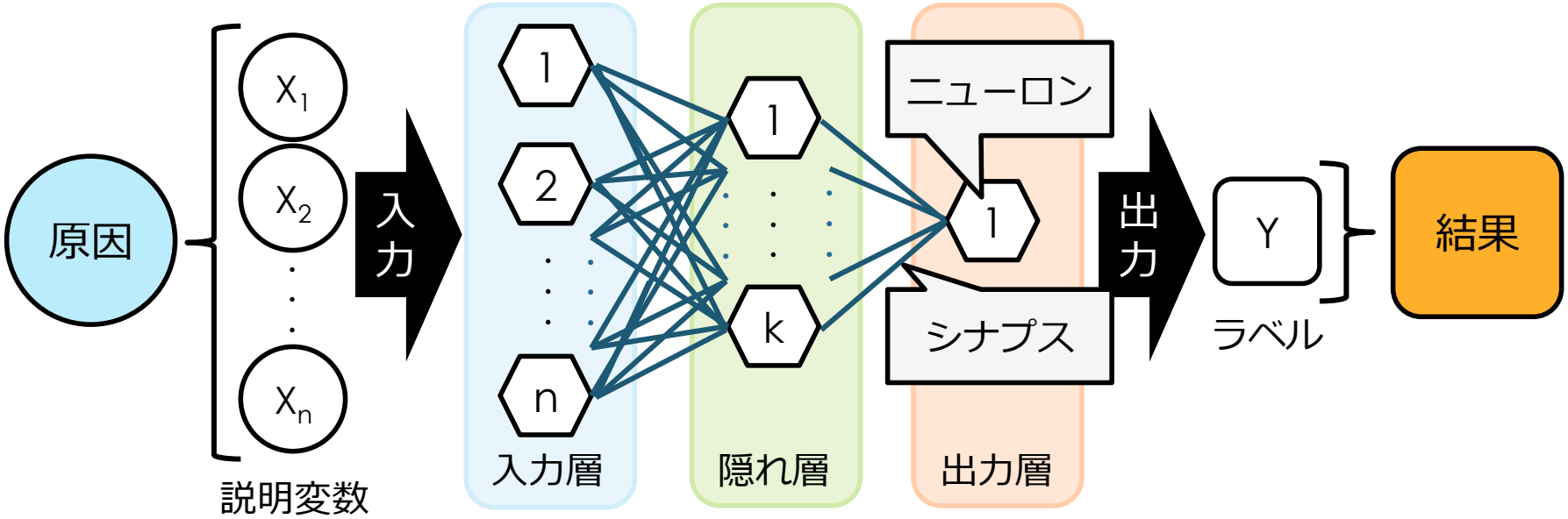
```
$ pip install -r requirements.txt
$ ipython notebook
```

おまけ：CHAINERを用いた深層学習

シンプルで軽量なニューラルネットワークライブラリ

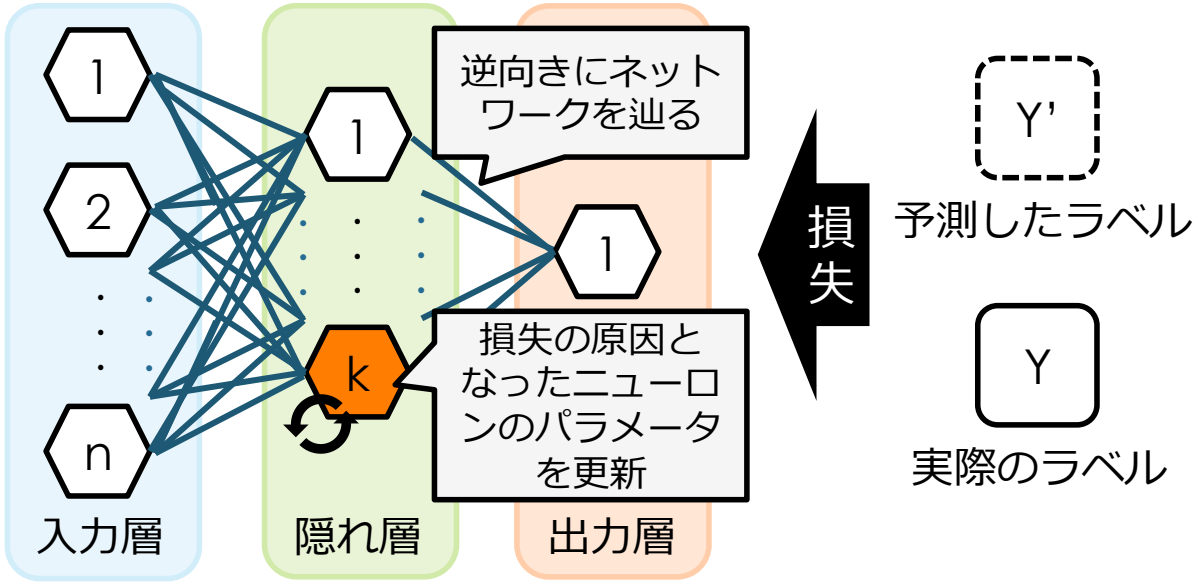
深層学習について

- 深層学習とは、三層以上の階層（入力層、隠れ層、出力層）からなるニューラルネットワークを用いた機械学習手法のこと
- 深層学習では、各層のニューロンをつなぐシナプスの結合度を調節することで、理論上任意の関数を表現することが可能
  - 表現可能であっても、目的とする関数にたどり着けるとはかぎらない。ニューラルネットワークには、表現力が高い反面過学習に陥りやすいという欠点がある



深層学習の仕組み

- 深層学習の学習プロセスは、順伝搬処理と逆伝搬処理からなる。
  - 順伝搬処理では現在のパラメータを用いた予測を、逆伝搬処理では損失値に応じたパラメータの更新を行う。これを繰り返すことで、損失を最小化するような関数に収束させることを目指す。
  - 逆伝搬処理のメカニズムを**Back Propagation**といい、予測値を決定する原因となったニューロン、すなわち損失に責任があるニューロンのパラメータを更新する仕組みを提供する



補足とまとめ

機械学習と人間とこれからについて

機械学習が扱う問題

- 機械学習が扱う問題は、主に分類問題と回帰問題に分けられる
  - 分類問題には、YES or NOを判別する二値分類と、複数のクラスに類別する多値分類がある。分類問題を扱う学習器を分類器という
  - 回帰問題では、離散的なクラスではなく、連続的な値を予測することを目的とする。たとえば、今日の気象情報から明日桶屋が儲かるかどうかを予測するのは分類問題だが、明日の桶屋の儲けがいくらになるのかを予測するのは回帰問題になる
- 手書き数字認識のような、パターン（ピクセル）と数種類のラベル（0~9）との関連を目的とするケースは、多値分類問題に属する
  - 多値分類問題の具体例として、音楽や絵画のジャンル分類がある

機械学習（AI）は人間を超えるか

- この問いはキャッチーだがナンセンス極まりない。なぜならば、最初からコンピューターは人間より計算が得意であり、計算可能な問題ならばコンピューターは人間よりも優秀たりうるから
  - ある問題を解決する際に、コンピューターが人間よりも“結果的に”いい仕事をしたならば、その問題の計算可能性が相応に高かったということであり、それ以上でもそれ以下でもない。
- 機械学習の最も大きな貢献は、人間にしかできない領域を浮き彫りにすることである。機械学習にできることをわざわざ人間がやる必要はないが、それ以外の事柄こそ人間がやるべきこと

まとめ

- 機械学習の目的は原因と結果の関数的関係性を探ること
- 機械学習は強力だが、決して万能ではないことを忘れてはいけない
  - 機械学習にできることとできないことを明確に理解すること
- 人間の仕事は機械学習という仕組みを前提にしてシステムや計算手法を構築すること
  - scikit-learnなどのツールの進歩によって、誰でも機械学習を手軽に扱えるようになってきているが、ただ使うだけでは勿体ない
  - 人間にできて機械学習にできないことを見つけたらチャンス。それができるようなシステムを作ろう

リンク

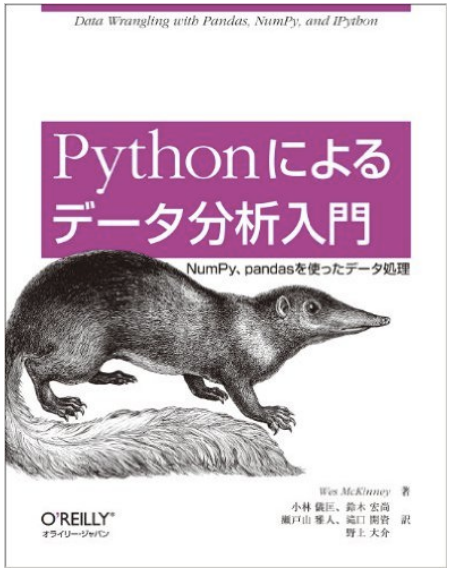
- Scikit-learn 公式ドキュメント（英語）
  - http://scikit-learn.org/stable/index.html
- Chainer 公式ドキュメント（英語）
  - http://docs.chainer.org/en/stable/
- 機械学習 – 朱鷺の杜wiki
  - http://ibisforest.org/index.php?%E6%A9%9F%E6%A2%B0%E5%AD%A6%E7%BF%92

書籍紹介

機械学習関連の参考書籍を何冊か紹介します

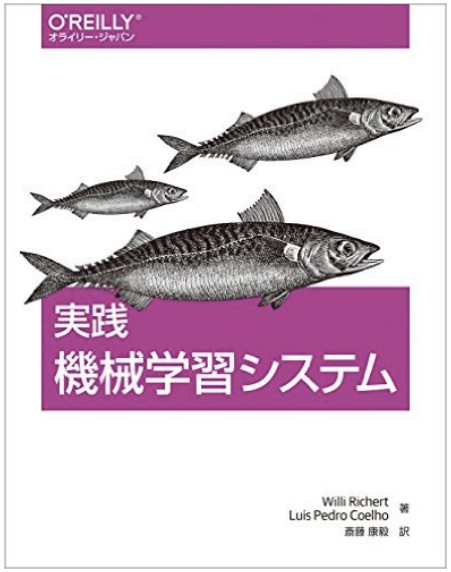
実装系(1/3)

- Pythonによるデータ分析入門（オライリー）
  - Wes McKinney (著)
  - データ分析ってなんだろう？何をどうすればいいの？という人向けの入門書。けっこう分厚いので少し威圧感があるかもしれないが、丁寧な説明の結果として量が増えているだけなのでご安心を。Pythonに慣れてない人でも安心して読める一冊。



実装系(2/3)

- 実践 機械学習システム（オライリー）
  - Willi Richert, Luis Pedro Coelho (著)
  - numpy、scipyの使い方からしっかりと説明してくれている良書。最初はデータの操作から、可視化、クラスタリング、トピックモデル、ナイーブベイズと、章を重ねるごとに容赦なく難易度が上がっていく体育会系書籍。心身ともに強くなりたい人におすすめ。

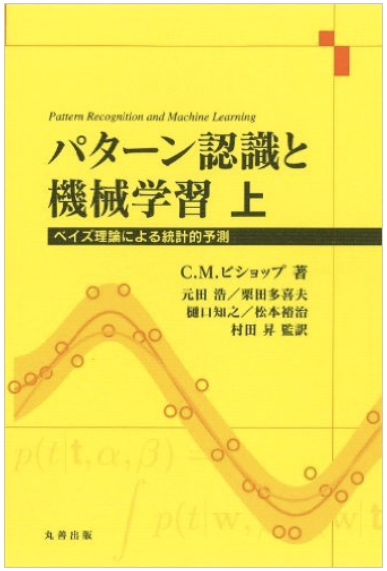


実装系(3/3)

- 集合知プログラミング（オライリー）
  - Toby Segaran (著)
  - 前出の二冊に比べるとやや古い書籍。機械学習というトピックワードが一人歩きし始める前の書籍なので、推薦やクラスタリングなど個々の仕組みそのものにフォーカスして書かれている。ただし、誤植が多いので読むときには注意

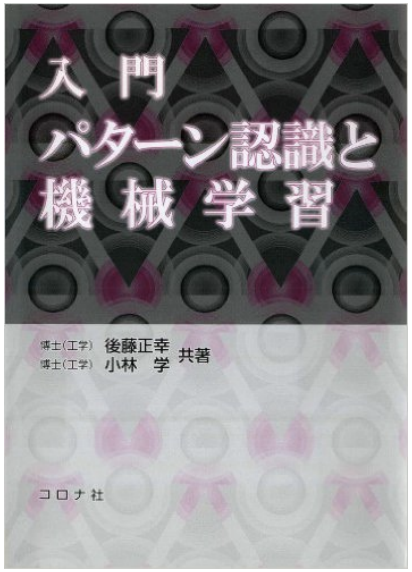
理論系(1/3)

- パターン認識と機械学習 上/下 （丸善出版）
  - C.M. ビショップ (著)
  - 機械学習理論の定番書籍。内容的には比較的難易度が高い方だが、全く曖昧な表現がないので、ある程度機械学習の概要を理解している人にはとても良い参考書。ただし、一定レベル以上の数学の知識が求められるので注意



理論系(2/3)

- 入門パターン認識と機械学習　（コロナ社）
  - 後藤 正幸, 小林 学 (著)
  - パターン認識と機械学習を少し噛み砕いたような内容。ただし、行列演算や統計学の知識は依然として必要とされる。数式が苦手な人は無理して理解しようとせずに、実装から入った方がよい



理論系(3/3)

- 確率的最適化（講談社）
  - 鈴木 大慈 (著)
  - 機械学習のコアとも言える関数最適化にフォーカスした書籍。広範な機械学習の領域を無理して網羅しようとせず、対象を最適化に絞って丁寧に書かれている