R & Machine Leaning

July 6 2015

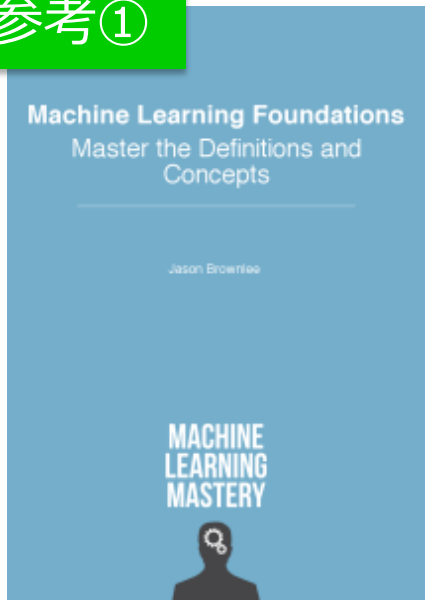
Yoshiharu Ikutani

@ NNCT 勉強会

今回の目標

- 「機械学習とは何か」を理解する
- Rの初歩を理解する
- Rで基本的な機械学習プログラムを実行する

参考①



参考②

アジェンダ

- 機械学習って何？
- Rって何？ どうやって使う？
- 実習：Rで機械学習アルゴリズム

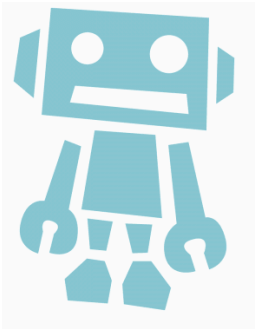
機械学習とは？

- まず質問します
  「機械学習って何ですか？」
  ※さらさら答えられる人は帰って良し

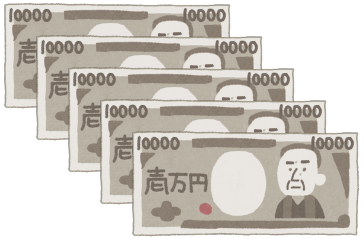
- 分からないなら偉い人に聞きましょう
- Tom Mitchell
  カーネギーメロン大学教授
  人工知能・機械学習の権威

Tom Mitchell による定義

“機械学習では経験により自動的に改善する
プログラムをどう作るかという問題を考える”



プログラム



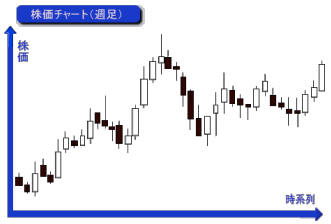
経験

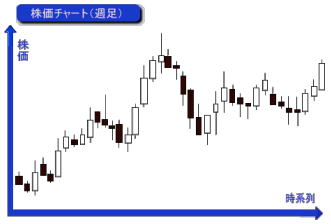
改善（学習）

このプロセスを自動でやるにはどうすれば良い？

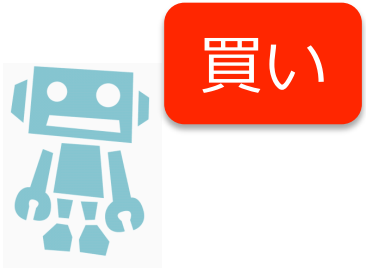
Tom Mitchell による詳細な定義

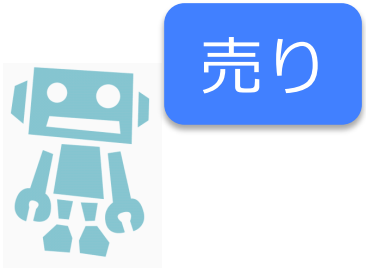
“プログラムはタスクTとパフォーマンス測定Pに関連する経験Eから学習する”



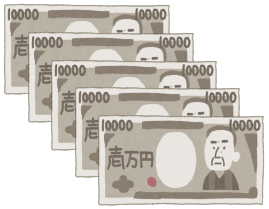


株価チャート
（経験E)





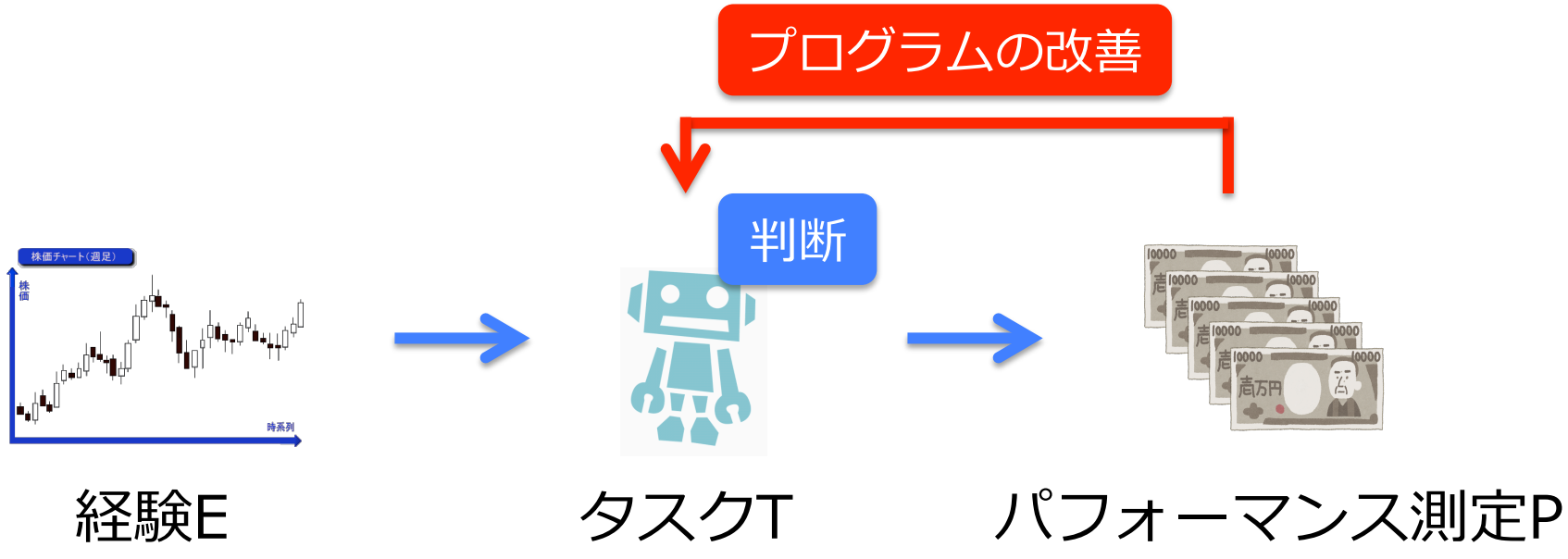
判断
（タスクT）



売買結果
(パフォーマンス測定P)

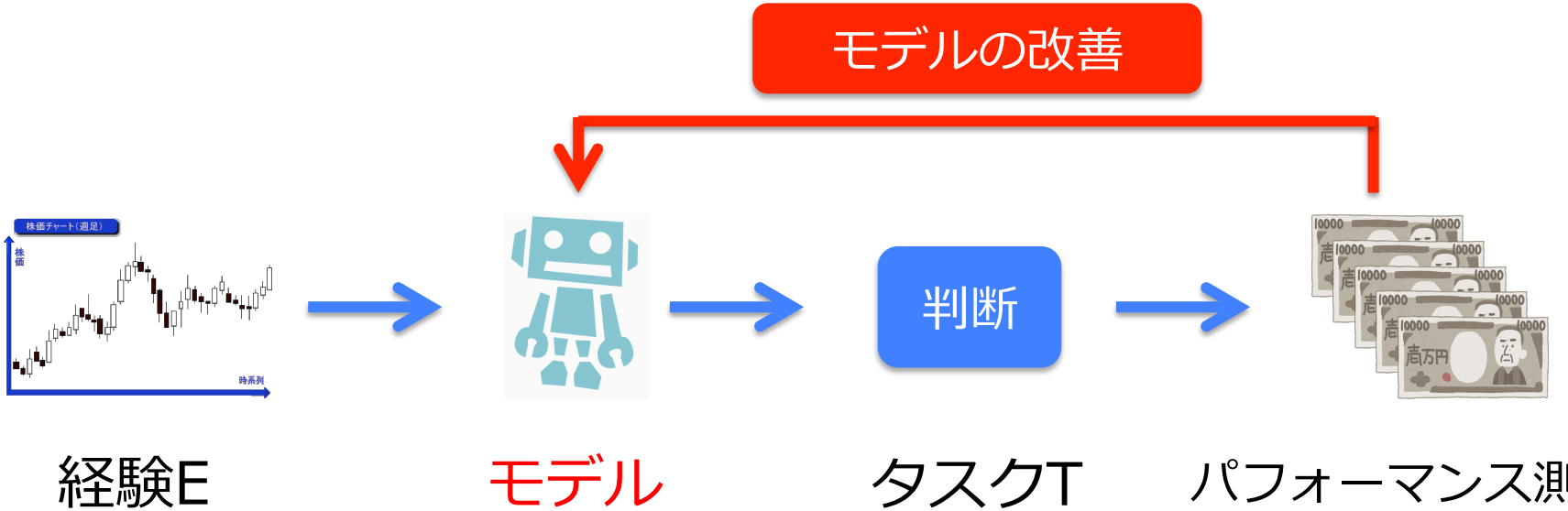
Tom Mitchell による詳細な定義

“タスクTでパフォーマンスした場合
パフォーマンス測定Pにより評価され
経験Eにより改善されていく”



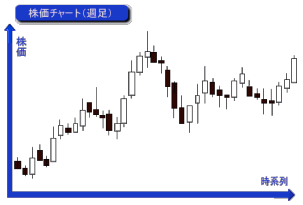
けっきょく機械学習とは？

“パフォーマンス測定Pに対する
判断(タスクT)を一般化するため
データ(経験E)からモデルを訓練すること“

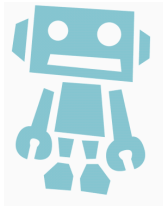


機械学習ができると何が嬉しいの？

- 過去のデータから
  未来の現象を予測できる
- もうすこし厳密に言うと・・・
  過去のデータによるモデルの訓練から
  未来の現象への適した判断を予測できる

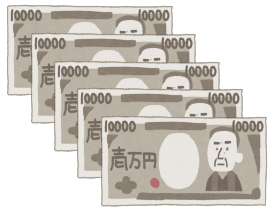


未知の経験



訓練された
モデル

最適な
判断



高パフォーマンス

機械学習で解ける問題

- 代表的な問題は以下の4つ
  1. Classification (分類)
  2. Regression (回帰)
  3. Clustering (基準なし分類)
  4. Rule Extraction (ルール抽出)

- 発表の目的範囲を超えるので説明は割愛
  ※各自，上のキーワードで調べてみてください

アジェンダ

- 機械学習って何？
- Rって何？ どうやって使う？
- 実習：Rで機械学習アルゴリズム

Rとは？

- オープンソース&フリーの
  統計解析向けプログラミング言語
- (私見では）データいじり特化型言語

読書き・操作
グラフ出力が容易
（だいたい1行）

データ操作以外苦手
(Text処理すら微妙)

膨大な数の解析手法
がパッケージで提供

ベクトルベースの
変わった処理体系

とりあえず触ってみる

- 準備

1. Rのインストール

   http://cran.r-project.org/bin/macosx/

2. Rstuidoのインストール

   http://www.rstudio.com/products/rstudio/

3. GitHubレポジトリのクローン

   https://github.com/Yoshiharu-Ikutani/R_machine

CSV の読込と表示

- WorkingDirectoryをR_machineに設定

Rstudio上で Ctrl+Shift+H

- CSVを読み込む

```
> data <- read.csv("data_pca.csv")
```

- data の中身を表示する

```
> data
```

data の部分表示

- data の1行目を表示

```
> data[1,]
```

- data の1列目を表示

```
> data[,1]
```

- data の1-3行目の2-3列目を表示

```
> data[1:3,2:3]
```

data のグラフ出力

- dataの1列目を棒グラフで出力

```
> barplot(data[,1])
```

- dataの列ごとの分布を箱ヒゲ図で出力

```
> boxplot(data)
```

- dataの3列目を線グラフで出力

```
> plot(data[,3],type="l")
```

R まとめ

- データ処理なら簡単に何でもできる
  - 統計的検定，信号処理 etc.
- 競合としてはPythonが熱い
  - Scipy, Numpyでの数学処理
  - Pandasによるデータフレーム
  - 分析以外もできる (Rより上?)

vs

python™

アジェンダ

- 機械学習って何？
- Rって何？ どうやって使う？
- 実習：Rで機械学習アルゴリズム

Rで機械学習アルゴリズム

- できそうな気がしてきましたか？
- 基本は以下のフロー通り
  今日は特にRを使うところだけ実習

取り上げる機械学習

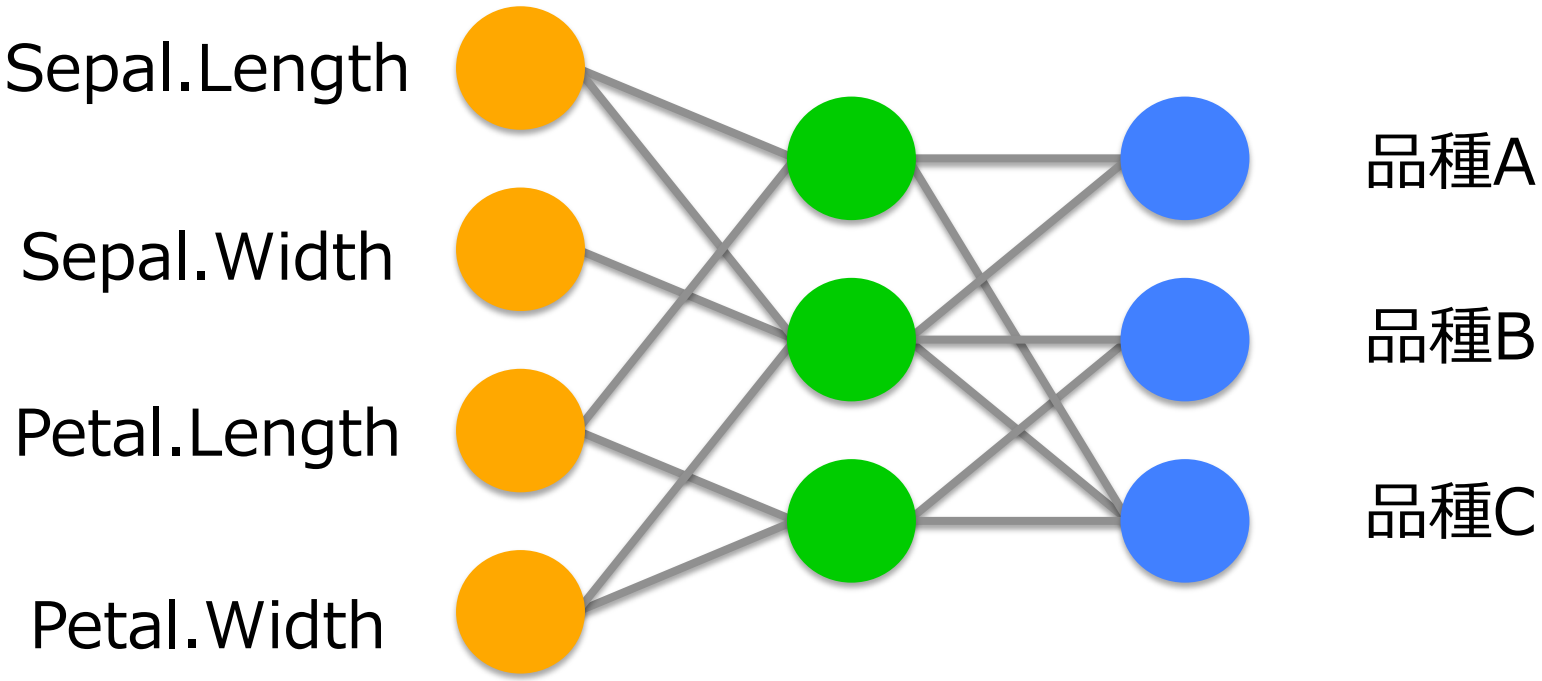
- 4種類のアルゴリズムのRコードを用意
  1. ニューラルネット (NeuralNetwork.R)
  2. 線形回帰 (LinerRegression.R)
  3. k-means法 (kmeans.R)
  4. 主成分分析 (PCA.R)
- 今回はニューラルネットだけ解説
  ※残りは自由に試してください

問題設定

- アヤメ（花の一種）の"がく"と"花びら"の情報から品種を推測したい
- irisにはアヤメの情報が格納
  - Sepal.Length & Width：がくの長さ・幅
  - Petal.Length & Width：花びらの長さ・幅
  - Species：品種

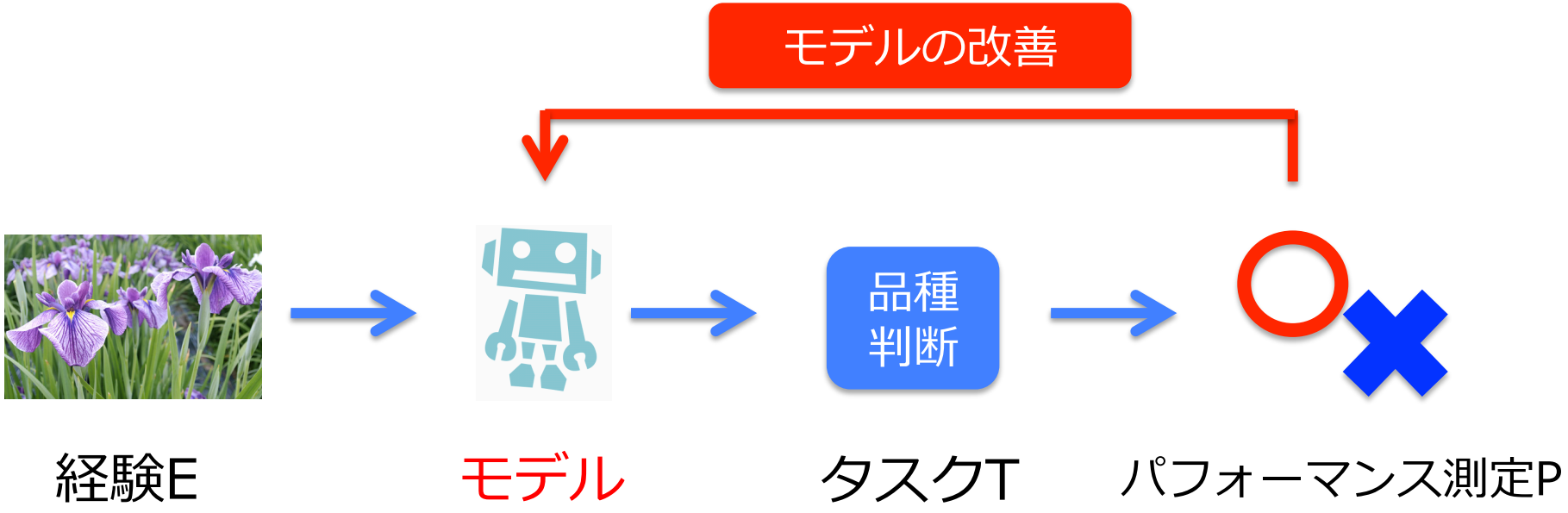
ニューラルネットワーク

- 問題解決のためニューラルネットを利用
  - 原理についての説明は割愛



学習条件の整理

- irisには150個のデータが格納
  - 75個のデータを訓練用 (iris.train)
  - 残り75個のデータをテスト用 (iris.test)



実際に試してみる

- NeuralNetwork.Rを動作させ結果を確認

まとめ

- 機械学習は4つの要素から構成
  1. 経験E：アヤメの情報
  2. タスクT：品種の判断
  3. パフォーマンス測定P：判断の正否
  4. 自動で改善可能なモデル：ニューラルネット

- 機械学習とは：

“未来の現象への適した判断を予測するための過去のデータを使ったモデルの訓練”