# Statistic for fools design doc

情報科学類二年 江畑 拓哉 (201611350)

## Contents

# 1 目的

機械学習を学習し始める人々に対して基礎的な定義式からボトムアップ的に統計・機械学習理論を理解するためのツールを作成することである。ツールは Web アプリとして制作する他、視覚的な要素があることや、使用したそれぞれの統計・機械学習手法についての解説がわかりやすく示されていることが望ましい。

# 2 背景

機械学習ツールを作成する、という初期目標を掲げ学習を行った結果多大な苦労があったため、学習成果を公開することで他の初学者の学習コストを低減させたいと考えたことにある。ここで言う多大な苦労とは、主に言語の壁の問題や参考文献の膨大さなどが挙げられる。

# 3 対応項目

今回は時系列解析手法の一つである、ARIMA モデル (ここには SARIMA、AR、MA、ARMA モデルを含む）と回帰木についてのランダムフォレストについて解説・実行するための Web アプリを作成する。
以下に対応項目をリスト形式で示す。尚、作成済みのものに関してはチェックをつける。また実装する機械学習に関する関数については以降に記載する仕様にて詳細に記述する。

- □ 画面制御
- □ アカウント制御
- [%] 機械学習制御
  - □ 実行・パラメータ設定
  - □ 実行する機械学習についての解説の表示

  - □ 実行結果表示

- □ 機械学習の実行
- □ 実行結果出力制御
- □ アカウントに紐付いた機械学習履歴の保存・呼び出し
- [%] 統計・機械学習に関する関数
  - □ ARIMA
  - □ ランダムフォレスト

## 4 非対応項目

対応項目と同様にリスト形式で示す。

- 多人数での実行結果の共有

## 5 スケジュール

期末試験期間終了後である 2017 年 12 月 29 日から情報特別演習最終発表である 2018 年 1 月 19 日未明までを実装期間とする。
詳細な予定としては 2017 年 12 月 31 日までに統計・機械学習（対応項目で言う”統計・機械学習に関する関数”）に関しない実装を終わらせ、その後時間の許す限り統計・機械学習についての関数を実装していく予定である。

## 6 仕様

以下に詳細な仕様を、対応項目ごとに示していく。

## 6.1 画面制御

画面制御の関係図を以下に示した後、それぞれの画面の機能についての説明を行う。

```
・ホーム画面
 <=>（(・サインイン未済=>サインイン・サインアップ画面
       ・サインイン済み=>（アカウント制御画面||
                           機械学習履歴画面||
                           機械学習画面）)||
     統計・機械学習解説インデックス画面）
・サインイン・サインアップ画面
 <=>(サインイン画面||
     サインアップ画面)
・アカウント制御画面
 <=>（パスワード再設定画面||
     アカウント削除画面||
     Quandl キー設定画面）
・機械学習画面
 <=>(パラメータ等設定画面||
     機械学習実行画面)
・統計・機械学習解説インデックス画面
 <=>（各々の解説画面）
・パラメーター等設定画面
 <=>（データ設定画面||
     機械学習手法選択画面）
・機械学習実行画面
 <=>  実行結果画面
・実行結果画面
  （・実行結果出力ボタンが押された=>  実行結果出力画面）
 =>   ホーム画面）
・機械学習履歴画面
 <=>（各々の機械学習画面）
```

- ホーム画面
  ホーム画面である。この Web アプリの概要に関する簡単な説明を行う。ルートはここに設定される。

- サインイン・サインアップ画面
  サインイン･サインアップ画面などへの遷移を行うための画面。プルダウンメニューなどで実装する可能性もある。

- アカウント制御画面
  アカウント制御に関する画面への遷移を行うための画面。プルダウンメニューなどで実装する可能性もある。

- 機械学習履歴画面
  アカウントに紐付けられた機械学習履歴を一覧表示する。表示する内容は、実行日時とデータ、パラメータである。

- 機械学習画面
  機械学習のパラメータ設定や実行に関する画面をまとめたもの。ここで ARIMA やランダムフォレストなどの機械学習のタイプを選択する。

- 統計・機械学習解説インデックス画面
  蓄積してある統計・機械学習解説のインデックスを表示する。

- サインイン画面
  アカウント名とパスワードを認証してサインインをする。

- サインアップ画面
  一意になるアカウント名とパスワードを入れてアカウントを登録する。

- パスワード再設定画面
  以前のパスワードと新しいパスワードを用いてパスワードを再設定する。

- アカウント削除画面
  アカウントとパスワードを用いてアカウントを削除する。この際にこのアカウントに紐付けられた機械学習履歴も削除する。

- Quandl キー設定画面
  Quandl の API キーを設定する。テストとして例えば日経平均株価を Quandl から入手できるかを検査する。

- パラメータ等設定画面
  機械学習のデータなどのパラメータを設定するためのルートページである。ここでパラメータの確認を行うことができる。

- 機械学習実行画面
  機械学習を実行するための確認などを行うための画面、チェックボックスリストを作成し、例えば統計・機械学習に関する説明を追記するなどのオプションの有効・無効を設定できるようにする。

- 各々の解説画面
  指定されたキーワードと一致する単語の解説を表示する。解説文中のキーワードに関する単語の解説も表示できるようにする。

- データ設定画面
  機械学習のタイプに応じた個数のデータを設定する。
  すでにアカウントの所持者がそのデータを取得している場合は、データベースからデータを読みだすようにする。つまりデータベースから呼び出し可能なキーを保持する。
  データを取得しておらず、設定された Quandl キーがある場合は、そのキーを用いてデータを取得し、アカウントに紐付けられたデータベースに保存する。そしてデータベースに保存した際に用いたキーを保持する。
  データを取得しておらず、Quandl キーが設定されていない場合は、そのままデータを取得し、アカウントに紐付けられたデータベースに保存する。そしてデータベースに保存した際に用いたキーを保持する。
  データを取得する際に取得できなかった場合にはエラー内容を記述したメッセージを表示する。

- 機械学習手法選択画面
  選択した機械学習のタイプに合わせたスロットを用意しておき、そこに任意の選択可能な検定やパラメータなどの値をセットできるようにする。“Exit” ボタンなどを作ることで、機械学習の処理を中断して、例えばある地点の検定結果までを実行できるようにする。

- 実行結果画面
  選択された機械学習のタイプ、パラメータを用いて実行した結果を表示する画面。タイプ、パラメータ、実行結果の PDF をアカウントに紐付けられたデータベースに保存する。

- 実行結果出力画面
  PDF に実行結果画面を出力した画面。これは別のタブやウィンドウなどで開かれるものとする。

- 各々の機械学習画面
  呼びだされたキーに対応した実行結果の PDF を表示する。

## 6.2 アカウント制御画面

サーバとの通信を行い、データベースにそれぞれのアカウントに関するデータを保存する。
扱うデータは以下の通りである。

- アカウント名 (id)
  ユーザごとに一意の値を取る識別子である。使用可能な文字は英数字のみとする。

- パスワード (password)
  長さは 8 文字以上 16 文字未満とする。使用可能な文字は英数字のみとする。

- Quandl キー (q-key)
  Quandl の API キーである。入力されたキーが正しいかどうかをサーバ側でテストする。このキーはいつでも変更可能なものとする。

利用する HTTP リクエストメソッドを以下に示す。但しここで示される parameter はあくまで目安であり、実際の parameter には暗号化などの処理が行われる。

| HTTP Request | method name | parameter | response |
|---|---|---|---|
| POST | /register | id, password | 未定 |
| POST | /re-password | id, password, password | 未定 |
| POST | /login | | 未定 |
| POST | /logout | | 未定 |
| POST | /set-q-key | q-key | 未定 |

## 6.3 機械学習制御

### 6.3.1 実行・パラメータ等設定

機械学習のタイプ、パラメータ、データを設定し、実行命令をサーバに送信する。
ARIMA、ランダムフォレストが選択可能な機械学習のタイプである。
処理は途中までの結果のみを出力できる。つまり各処理は次の処理を行うかそこで処理を終了するかを選択することができる。
それぞれのタイプにおけるデータ、パラメータを以下に示す。

1. ARIMA
   - データ

a) 単一時系列データ
　単一の時系列データを用いる。時系列の長さもここで設定する。

- パラメータ
　パラメータは処理と検定を合わせて自動的に設定できる部分を作成する。

a) 定常化処理リスト
　定常化を行う際の工程リストを示す。例えば、対数化や n 次階差、季節階差などがここに該当する。

b) 定常化後検定パラメータ
　定常化が行われているかを検定するためのパラメータ、例えば KPSS や ADF などの検定内容、有意水準などがここに該当する。

c) AR(p) モデル値設定
　AR(p) モデルの p の値を設定する。

d) 推定
　上で設定した p の値を使って推定を行う。ここには最尤推定などの推定手法をパラメータとして設定する。

e) 検定
　AIC や BIC などを用いて分析がどの程度元データに近づけることができたかを検定する。

f) MA(q) モデル値設定
　MA(q) モデルの q の値を設定する。また、3,4,5 工程を行わずにここに飛ぶことができるようにする。

g) 推定
　上で設定した q の値を使って推定を行う。ここには最尤推定などの推定手法をパラメータとして設定する。

h) 検定
　AIC や BIC などを用いて分析がどの程度元データに近づけることができたかを検定する。

i) 予測
　作成したモデルを使って時系列予測を行う。どこまで予測するかをパラメータとして設定する。

2. ランダムフォレスト
   - データ

     a) 複数時系列データ
        複数の時系列データを用いる。説明変数 (複数) にあたる部分は目的変数 (単一) を時系列的に広義に包含している必要がある。

   - パラメータ (1〜4 のパラメータは必須であり途中で処理を止めることはできない)

     a) サンプルサイズ
        ひとつのサンプルにおけるサイズをデータ全体から Q % という形で設定する。

     b) サンプル数
        決定木の数を設定する。

     c) 分岐関数
        分岐に用いる関数を設定する。例えばエントロピーやジニ係数がこれに該当する。

     d) 分岐深度
        分岐を行う深さを設定する。

3. これらのパラメータ元に HTTP リクエストを送りサーバーと通信を行う。利用する HTTP リクエストメソッドを以下に示す。但しここで示される parameters は処理内容やパラメータを保存した json 形式のデータであり、Response は実行結果を保存しているデータベース上のキーベクトルであるとする。

| HTTP Request | method name | parameter | response |
|---|---|---|---|
| GET | /arima/{account} | parameters | Response |
| GET | /rforest/{account} | parameters | Response |

### 6.3.2 実行する機械学習についての解説の表示

インデックスに収められた一意のキーを用いてサーバから解説データを取得し HTML として表示する。内容中の統計・機械学習に関するキーワードにはそれについてへのリンクが含まれている。

### 6.4 実行結果出力制御

渡されたデータベースへのアクセスキーのベクトルを用いて、実行結果を保存しているデータベース上からデータを呼び出し、その形式に応じて表示を行っていく。更にチェックボックス形式のオプション選択に応じてそれぞれの処理についての解説を表示する。

利用する HTTP リクエストメソッドを以下に示す。access-key はデータベースへのアクセスキー、json-data は json 形式のデータである。

| HTTP Request | method name | parameter | response |
|---|---|---|---|
| GET | /get-graph/{account} | access-key | json-data |
| GET | /get-param/{account} | access-key | float |
| GET | /get-param-list/{account} | access-key | json-data |

### 6.5 実行結果表示

実行結果出力制御から得られた設定やデータベースへのアクセスキーのベクトルを元に、サーバから受け取った PDF ファイルを表示する。これはユーザがファイルを保存してブラウザやその他のアプリケーションで開くものとする。

利用する HTTP リクエストメソッドを以下に示す。json-data はデータベースへのアクセスキーと処理内容、その他のオプションなどを含んだ json 形式のデータである。

| HTTP Request | method name | parameter | response |
|---|---|---|---|
| GET | /get-pdf/{account} | json-data | pdf |

### 6.6 アカウントに紐付いた機械学習履歴の保存・呼び出し

機械学習を行った実行結果を記した pdf に対してアカウントに紐付けた保存・呼び出しを行う。

利用する HTTP リクエストメソッドを以下に示す。key はデータベース上の pdf を保存しているデータへのアクセスキーである。

| HTTP Request | method name | parameter | response |
|---|---|---|---|
| GET | /get-past-pdf/{account} | key | pdf |

### 6.7 アカウントに紐付いた元データの保存・呼び出し

機械学習制御で用いるデータを呼び出す際にはサーバ側で元となるデータがそのアカウントを含んでいるかを確認する。そのデータが呼び出したアカウントを含んでいればデータベースからデータを呼び出す。そうでなければ Quandl からデータを呼び出す。

データは毎週更新され、それに応じてデータに紐付けられていたアカウントの情報はリセットされる。更新のタイミングはそのデータがその週最初に呼び出されてた時点である。

利用する HTTP リクエストメソッドを以下に示す。params は Quandl のデータへのアクセスのための名称と、欲しいデータの開始日時と終了日時を示した json 形式のデータ、

key はデータベースに保存された該当データへのキーである。

| HTTP Request | method name | parameter | response |
|---|---|---|---|
| GET | /get-raw-data/{account} | params | key |

### 6.8 統計・機械学習に関する仕様

#### 6.8.1 ARIMA

#### 6.8.2 ランダムフォレスト

## 7 実装

実装すべき主要な関数については以下に示す。全ての関数の詳細は、別紙の実装する関数についての API ドキュメントに示す。

## 8 セキュリティやプライバシーについて

HTTPS 通信は実現可能であるか現状不明（SSL 証明書を入手可能であるのかが不明）であるため、セキュリティの高い Web アプリを作成することは困難であると考えられる。プライバシーについてはアカウントについて独立の内容を提供しているため、プライバシーを侵す心配はないと考えられる。

## 9 リスク

テストに多くの時間を割けないため、セッション維持などに関する問題や、過大なスケールの処理に対する処置が問題になることが考えられる。極力初期実装時に完璧なプログラムを作成する予定である。

## 10 テスト計画

今のところ、テストを行う予定はない。

## 11 参考資料

## 12 リポジトリ

2017 年 12 月 29 日に git@elect000 アカウントに作成する予定である。

# 13 編集履歴

## 13.1 DONE 第 1 回編集 [2017-12-18 月 17:02]

このファイルの作成、大まかな全体の設計に加え、画面制御に関する部分を編集した。

## 13.2 DONE 第 2 回編集 [2017-12-19 火 04:17]

最後に実装する予定である機械学習についての部分を除いて全ての仕様を編集した。これにより、先述の部分と実装を除いた全ての項目を編集したことになり、これを第一計画としてプロジェクトの見直しを行う。尚、これ以上の大幅な編集は期末試験終了である 2017 年 12 月 26 日まで行わない。