

統計処理及び機械学習に基づく
データマイニング入門
第2回

宮本 隆志

ナビプラス株式会社

March 9, 2015

この勉強会について

- ▶ データマイニングの入門講座です。
 - ▶ 想定する聴衆は、これからデータを解析してみようという初学者を想定しています。
 - ▶ 専門家や既に実務経験の豊富な方々には物足りない内容かと思います。
- ▶ データマイニング =
 - ▶ 大量のデータを
 - ▶ 統計学や機械学習などの手法を用いて探索・分析して
 - ▶ 意味あるパターンやルールを発見すると考えます。この勉強会では手法の話をしします。
- ▶ ツールは無料のオープンソースのものを使用します
 - ▶ メインに Python を使用します。Anaconda-2.1.0 を用いて説明します。
- ▶ 参考書はイベント Web ページには記載しましたが、あまり準拠しません。
 - ▶ あまり準拠すると著作権的に問題があるので。

自己紹介

名前 宮本 隆志 (@tmiya_)

所属/仕事 ナビプラス株式会社 / データ解析周りの R&D の仕事

ナビプラス マーケティングソリューションツールの開発・提供

- ▶ サイト内検索エンジン・レコメンドエンジン、レビュー投稿エンジンが中心
- ▶ 次世代インターネットサービスの研究・開発
- ▶ 上記に付随する広告商品の販売

前職 ネット広告の入札サーバを開発する会社で似たような仕事

興味 機械学習 / 関数型言語 / 定理証明系

- ▶ Coq という定理証明系の勉強会を毎月開催しています

勉強会の進め方：予定

- ▶ 講義 (30 分 + 30 分) + 実習 (40 分) の形式。
- ▶ 前半は統計処理とか機械学習の手法の講義を中心。
 - ▶ 前回分へのご意見を元に、演習資料の説明時間を増やすことにしました。
- ▶ 後半は Python を用いて簡単なハンズオンを予定。
 - ▶ 興味のない方は講義について質問したり退出したり Python 以外で解析するとかでお願いします。
 - ▶ Python 環境は Anaconda-2.1.0 を推奨しますが、各自に任せます。
 - ▶ ハンズオン用ファイル配布のために、Python の他に git も導入してください。(あるいは USB メモリでファイルを配布します。)
- ▶ 講義形式の勉強会とは別に、読書会 (教科書とか論文とか) とかやりたいので興味のある方、別途相談しましょう。

今回の内容

今回話す内容

- ▶ 検定の基礎
- ▶ χ^2 検定

今回話さない内容 \Rightarrow 次回に回します

- ▶ χ^2 検定の別な例：Bradley-Terry モデル
- ▶ 多腕バンディット問題

参考文献

検定の話は、きちんとした話をするのは大変なので、教科書を読んで勉強したほうが良いと思います。

- ▶ 統計学：各自お好きな教科書で。
 - ▶ 東京大学教養学部統計学教室編 基礎統計学ⅠⅡⅢ「統計学入門」「自然科学の統計学」「人文・社会科学の統計学」
- ▶ 検定
 - ▶ 豊田秀樹「検定力分析入門」：Rを使った入門書です。
 - ▶ Geoff Cumming "Understanding The New Statistics" 評判が良い教科書のようなです。

ハンズオンファイルの取得

ハンズオンファイルの取得方法

- ▶ 初めての方は適当な作業用ディレクトリで (下記を 1 行で)
`$ git clone https://github.com/takashi-miyamoto-naviplus/spml4dm.git`
を行ってください。
- ▶ 既に `git clone` している人は、`spml4dm/` ディレクトリで、
`$ git pull`
して下さい。
- ▶ `git` を導入してない場合は、USB メモリにてファイルを配布します。

IPython 起動方法

- ▶ `spml4dm/2/` ディレクトリに移動して、
`$ ipython notebook`

IPython

spml4dm/2/ ディレクトリに移動 ⇒

\$ ipython notebook

⇒ Home 画面：ノートブック選択 or 新規作成 ⇒

ファイル名：クリックするとrename出来る

実行：スクリプトの実行 or レンダリング

コードと markdown の切り替え

The screenshot shows the IPython Notebook web interface. At the top, there's a browser address bar and a navigation menu. Below that is the notebook title 'IP[y]: Notebook' followed by the filename 'aic (autosaved)'. A menu bar includes 'File', 'Edit', 'View', 'Insert', 'Cell', 'Kernel', and 'Help'. Below the menu is a toolbar with icons for saving, undo, redo, and running. The 'Code' button is highlighted with a red circle. The main content area shows a notebook titled '多項式近似とAICによるモデル選択の例'. It contains a markdown cell with text and a code cell with Python code. Red arrows point from yellow text boxes to specific UI elements: one to the filename 'aic', one to the 'Code' button, one to the 'Cell' menu, and one to the 'Run' button in the toolbar.

多項式近似とAICによるモデル選択の例

ここではアーク点を最小二乗法で多項式近似するに際して、多項式次数をAIC(赤池情報量基準)を用いて最適化してみます。

まず使用する

- matplotlib を使えるようにするおまじない(matplotlibなどで描写したグラフをそのまま表示したりできる)
- NumPy と matplotlib を import

を行います。

```
In [1]: %matplotlib inline
import numpy as np
import matplotlib.pyplot as plt
```

多項式近似される元データを作成します。x として 0 ~ 3 の区間を30等分して両端

Markdown
ダブルクリックで編集可能
LaTeX数式も表示可能

コード
ダブルクリックで選択
▶でコード実行

次回

次回は、 χ^2 検定の話の続きと、A/B テストの話、をしようと思います。