データ解析論・データ解析演習

講義資料は http://nip.info.kogakuin.ac.jp/lectures/ で公開する

授業中のアンケートに Google Form を利用したクリッカーを使います
http://goo.gl/forms/ifqOCEgRQB

第１回（2016-04-07）概論・準備

授業の狙い

現在得られているデータから背景にある構造を推定し未来あるいは未知のデータに関する予測を行う方法として、確率モデルと各種の推定手法を学ぶ。特に、データの変数間の関係性について調べる基本的な考え方である多変量正規分布の理解を軸に、主成分分析、因子分析、判別分析、クラスタ分析の利用法について学ぶ。

受講に当たっての前提条件

「データ解析論」と「データ解析演習」を同時に履修すること。

具体的な到達目標

- ➔ 分散・共分散行列と多変量正規分布について理解する
- ➔ 重回帰分析と因子分析について理解し、実際にデータに適用できる
- ➔ 判別分析による分類学習について理解し、実際にデータに適用できる
- ➔ いくつかのクラスタ分析の手法について理解し、実際にデータに適用できる
- ➔ 分類学習やクラスタ分析に関して、より実用性の高い代表的な方法を知る

評価

データ解析論

- ➔ 授業内試験と期末試験の結果を３：７の割合で評価する

データ解析演習

- ➔ １～１２回の各回は、演習時間内に提出された課題の回答を各回５点満点で評価する
- ➔ １３回には総合的な課題を課し、提出されたレポートを期末試験として４０点満点で評価
- ➔ 各回評価の合計を総合評価とし、A+ ~ F の 6 段階評価で D 以上を合格とする。

教科書

★ 足立浩平「多変量データ解析法」
ISBN　978-4-7795-0057-2
出版社　ナカニシヤ出版

参考書

★ Bruce Frey 「STATISTICS HACKS」
ISBN　978-4-87311-335-7
出版社　オライリー・ジャパン
具体的な例で実践的に統計を理解したい場合の参考に

★ Wes McKinney「Pythonによるデータ分析入門」
ISBN　978-4-87311-655-6
出版社　オライリー・ジャパン
演習で利用する Python - Pandas の詳細な利用方法を理解したい場合の参考に

授業計画（案）

1. 多変量データ
   多変量データの扱い方について学ぶ
2. 分散・共分散行列と多変量正規分布
   多変量データの分散・共分散行列とその性質について学ぶ
3. 多変量正規分布と主成分分析
   主成分分析の考え方と利用法について学ぶ

4. 重回帰分析とパス解析
   重回帰分析とその一般化であるパス解析について学ぶ
5. 確認的因子分析
   因子分析の基本的な考え方について学ぶ
6. 探索的因子分析
   実践的な因子分析の利用法を学ぶ

7. 習熟度の確認
   １～６回の範囲について試験を行う（講義）
   １～６回の範囲について総合的な課題を行う（演習）

8. 判別分析
   ２群のデータが与えられた時に未知のデータの分類を行う方法を学ぶ
9. サポートベクトルマシンの基礎
   汎化性能の高い分類器であるサポートベクトルマシンについて学ぶ
10. サポートベクトルマシンの利用
    ソフトマージンの考え方とカーネル法について学ぶ

11. 階層的クラスタリング
    階層的クラスタリングについて学ぶ
12. k-平均法
    k-平均法によるクラスタリングついて学ぶ
13. 確率モデルに基づいたクラスタリング手法
    混合正規分布モデルによるクラスタリングについて学ぶ

期末試験（講義）・最終レポート（演習）

14. 学習内容の振り返り

準備学習

・教科書の該当する章をあらかじめ読んでおくこと

オフィスアワー

竹川高志 <jt13456@ns.kogakuin.ac.jp>
場所： 新宿高層棟A1516
時間： 月曜日３限

今日の講義＆演習内容

平均／分散／標準偏差／共分散／相関係数の復習

定義

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad , \quad s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad , \quad s_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \quad , \quad r_{xy} = \frac{s_{sy}}{s_x s_y}$$

参考：　分散についての計算上の性質

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$$

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}$$

データ形式

ファイル形式（表計算，CSV）

- ➢ 表計算ソフト
  - ○ Microsoft Office の Excel など
- ➢ CSV (camma「,」separated values)
  - ○ テキストファイル
  - ○ 一行に１つのデータ
  - ○ 要素の間は「,」で区切る
  - ○ 先頭行をデータの種類を示すヘッダとする場合が多い

表計算や CSV を利用する際のデータ形式

- ➢ 1行にすべての関連する情報を入れる
  - ○ 階層型のデータと扱いと比較して考えると重複が多い
  - ○ すべての要素を対等に扱えるので処理の際わかりやすい
- ➢ 具体例

実験番号	条件1	条件2	結果	CSV ファイル
1	a	x	2.0	**実験番号,条件1,条件2,結果,** **1,a,x,2.0,** **2,a,y,3.0,** **3,a,z,4.0,** **4,b,x,2.5,** **5,b,y,3.5,** **6,b,z,4.5,**
2	a	y	3.0	
3	a	z	4.0	
4	b	x	2.5	
5	b	y	3.5	
6	b	z	4.5	

利用する計算機環境について
（Python + 各種ライブラリ + Jupyter Notebook）

➔ 演習室にインストールされている
➔ 各自でインストールするのも比較的容易である（講義ページの「環境設定」参照）

➢ WinPython version 3 http://winpython.sourceforge.net/
  - Python プログラミング言語 https://python.rog/ http://www.python.jp/
    - numpy, scipy 数値計算，科学技術ライブラリ http://www.numpy.org/
    - pandas 統計データ解析用ライブラリ http://pandas.pydata.org/
    - matplotlib グラフ作成ライブラリ http://matplotlib.org/
    - seaborn matplotlib を用いたさらに高度なグラフ作成ライブラリ https://stanford.edu/~mwaskom/software/seaborn/
  - Jupyter Notebook
    - python の統合実行環境の一つ

演習準備

WinPython (Jupyter Notebook) の起動（既存のファイルを利用）

➔ http://nip.info.kogakuin.ac.jp/lectures/ から Lecture01.ipynb をダウンロード
➔ Notebook ファイル（拡張子 .ipynb）を「Jupyter Notebook,exe」を選択して実行
➔ localhost にサーバが起動
➔ Web ブラウザ上で実行される（http://localhost:8888）

Jupyter 環境について

➔ 「Code」ブロック
  ◆ プログラムを記述 Shift + Return で実行
➔ 「Markdown」ブロック
  ◆ 文章を Markdown 記法で記述（ヘルプや web の情報を参照すること）

課題

1. http://nip.info.kogakuin.ac.jp/lectures/ から data01.csv をダウンロード
2. Lecture01.ipynb と同じ作業ディレクトリに置く
3. Lecture01.ipynb の最初の Markdown ブロックに学籍番号／氏名などを記入
4. 続くブロックの指示に従って，プログラムを順に実行する
5. 最後の Markdown ブロックに質問／感想を記入
6. ファイルを保存する
7. 作業済みの ipynb ファイルを提出
   http://nip.info.kogakuin.ac.jp/lectures/2016/data_analysis

参考資料