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２回（2016-04-14）多変量正規分布

前回の復習

平均／分散／標準偏差／共分散／相関係数の復習

定義

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

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

データ形式

- CSV (camma「,」separated values)
  - テキストファイルで一行に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
- 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

相関と因果関係について

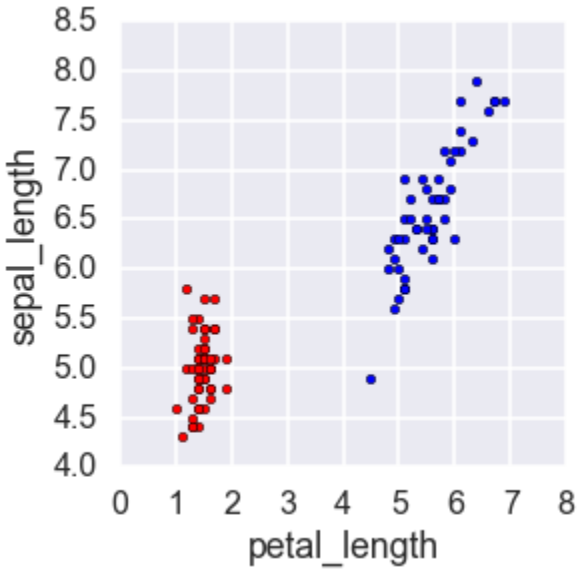
- 相関係数の数字だけで判断してはいけない
  - 外れ値の影響を受けやすい
  - 相関係数の値を評価することは難しい
- A と B に相関がある場合には，幾つかの可能性がある
  - A によって B が引き起こされる　（因果関係）
  - B によって A が引き起こされる　（因果関係）
  - A と B を引き起こす共通の要素がある　（共通要素）
  - 上記の組み合わせ
- 関係性は相関だけではない　＝　相関係数が 0 でも無関係とは限らない
  - 相関　＝　線形な関係性
  - 非線形な関係
  - 独立　⊂　無相関

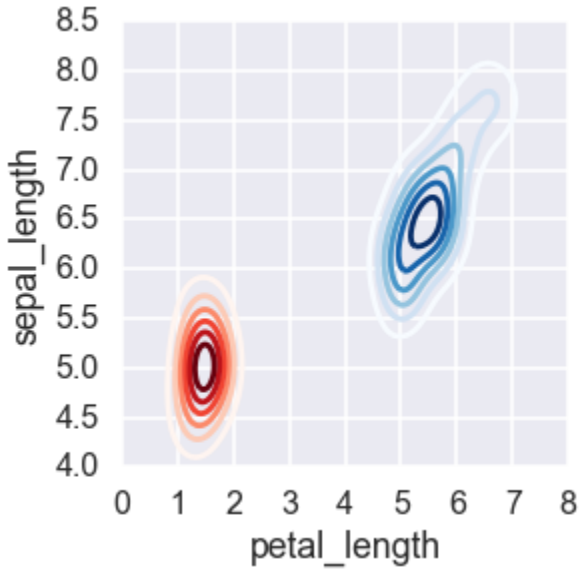
今日の講義

データ：　あやめ（Iris）の種類と花びらとがくの大きさ

http://archive.ics.uci.edu/ml/datasets/Iris

散布図と等高線図

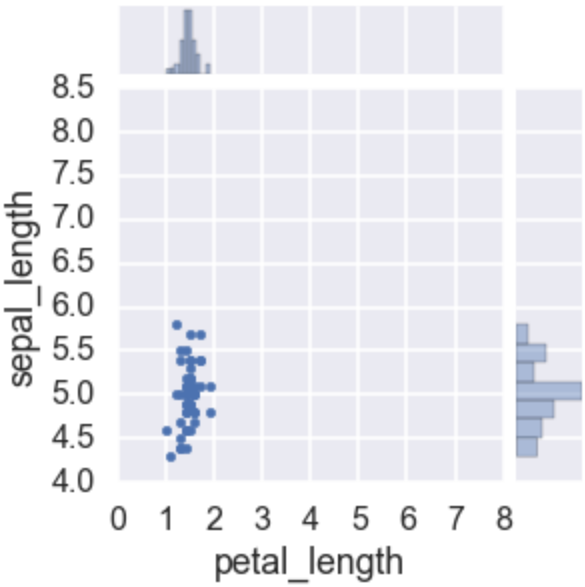


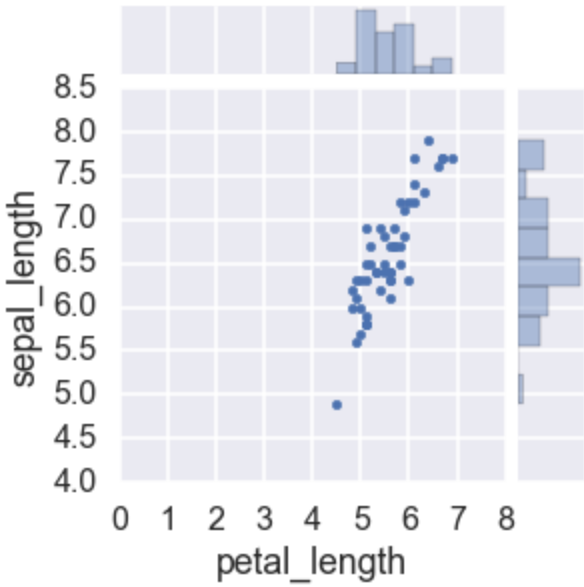


分散・共分散行列

- 平均＝分布の中心，　行列の対角成分＝各成分の分散，　共分散＝分布の傾きと関連

setosa	petal_length	sepal_length	**virginica**	petal_length	sepal_length
mean	1.462	5.006		5.552	6.588
petal_length	0.030159	0.016355		0.304588	0.303290
sepal_length	0.016355	0.124249		0.303290	0.404343





マハラノビス距離

- 定義

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \qquad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \qquad \Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix}$$

- 1変数の場合の z 値

$$z = \frac{x - \mu}{\sigma}$$

- 多変量の場合の z 値　＝　マハラノビス距離

$$z = \sqrt{(x-\mu)^\top \Sigma^{-1} (x-\mu)}$$

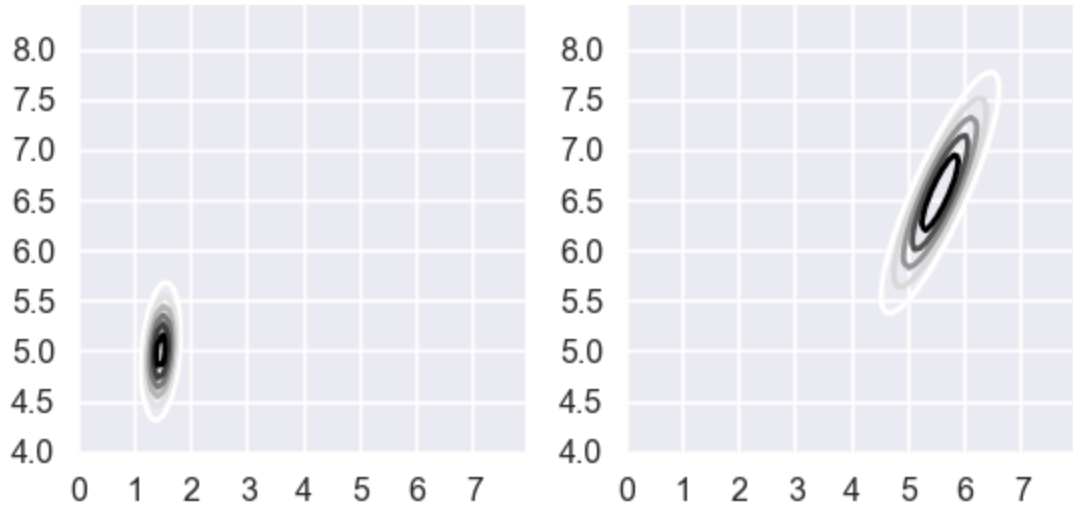
- 共分散が 0 なら簡単

$$z^2 = \begin{bmatrix} x_1 - \mu_1 & x_2 - \mu_2 \end{bmatrix} \begin{bmatrix} \sigma_1^{-2} & 0 \\ 0 & \sigma_2^{-2} \end{bmatrix} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}$$

$$= z_1^2 + z_2^2 = \frac{(x_1 - \mu_1)^2}{\sigma_1^2} + \frac{(x_2 - \mu_2)^2}{\sigma_2^2}$$

多変量正規分布

- マハラノビス距離 z が等しいところの "尤度" が等しい



演習内容

環境設定

Jupyter では様々な環境設定ができるので，少しずつ紹介していく．

起動時の挙動

- 前回は ipynb ファイルを開くやり方で起動した．
- C:\WinPython-64bit-3.5.1.1\Jupyter Notebook.exe を直接ダブルクリックすると
  フォルダ C:\WinPython-64bit-3.5.1.1\notebooks のファイルリストが表示される．
- S:\Documents の方が都合が良いのでその方法を考える．

方法（応急処置的な方法）

- S:\Documents に C:\WinPython-64bit-3.5.1.1\Jupyter Notebook.exe のショートカット作成
- 作成したショートカットのプロパティで，リンク先を
  "C:\WinPython-64bit-3.5.1.1\Jupyter Notebook.exe" --notebook-dir="S:\Documents" とする

今後の予定

- 設定ファイルを書き換えて，自分の使いやすい環境にする
  - code 部分のフォント
  - グラフの日本語表示
  - 拡張機能の追加

課題

http://nip.info.kogakuin.ac.jp/lectures/2016/data_analysis

1. 講義資料サイトから Lecture02.ipynb と data02.csv をダウンロード
2. ファイルを同じ作業ディレクトリに置く
3. Lecture02.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