3-4 データマネジメント，データ分析

（情報工学応用演習 II）

金子邦彦

ここで行うこと

- 自動でのグラフ作成
  - 円グラフ
  - 棒グラフ
  - 散布図
  - 線形近似
  - ヒストグラム

- データ分析
  - 分類（クラスタリング）

何の役に立つのか

- **データの管理や分析**では、さまざまなグラフ（**棒グラフ、円グラフ、散布図、ヒストグラム**）や分析法が**役に立つ**

- データに**外れ値**があるか？　の判定でも、さまざまなグラフ（線形近似）や分析法（クラスタリング）が役に立つ

外れ値とは、計測ミス、大きな誤差、装置の異常等により、大きく外れた値を持つデータのこと。

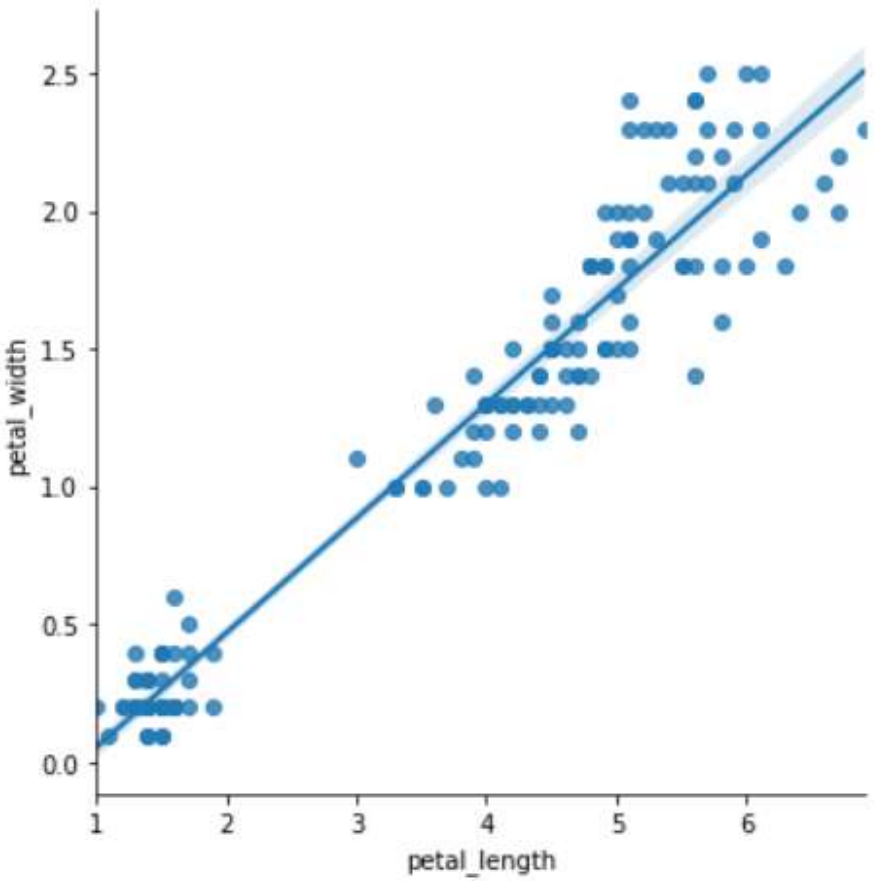
この実習の特徴

- Python で**自動化**する
- Excel よりも、**きれいなグラフ**をめざす
- Excel でできないこと（クラスタリング）も行う
- **Python の簡単なプログラム**で行えるようにする
  matplotlib, seaborn, scikit-learn を使用

実習の見どころ

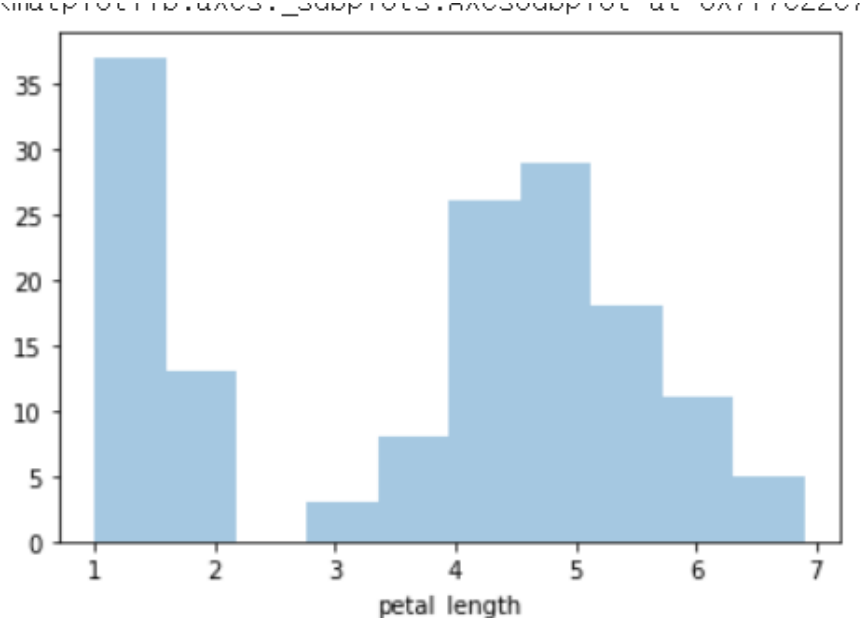
- 線形近似
- カーネル密度分布
- 自動分類（クラスタリング）

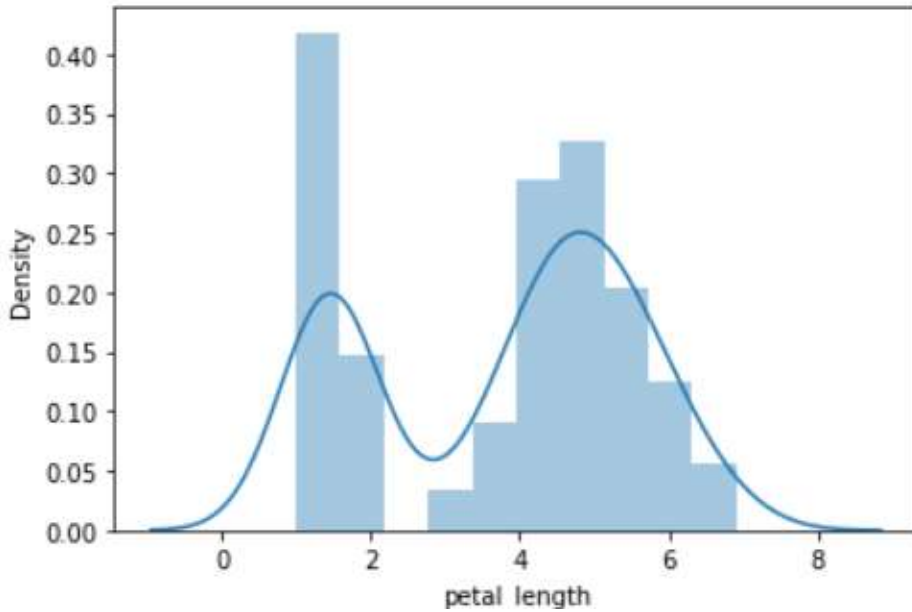
線形近似



- 互いの関係性（増加、減少）を直線で近似する
- その直線を自動で得ることが可能

カーネル密度分布



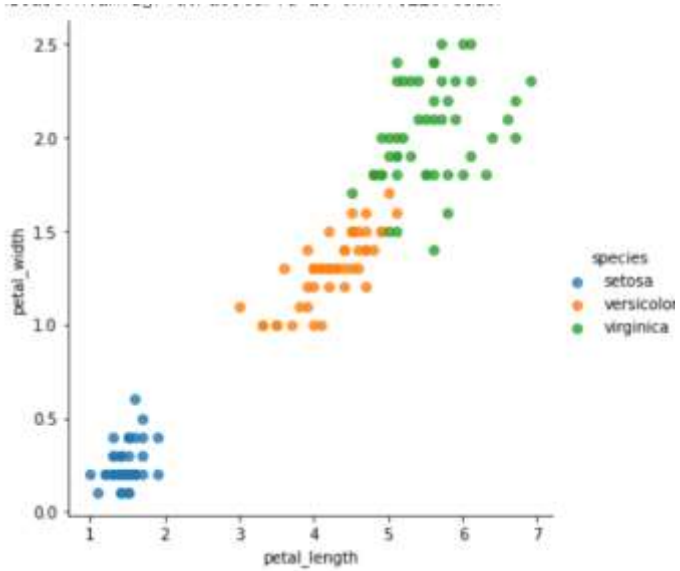


元のヒストグラム

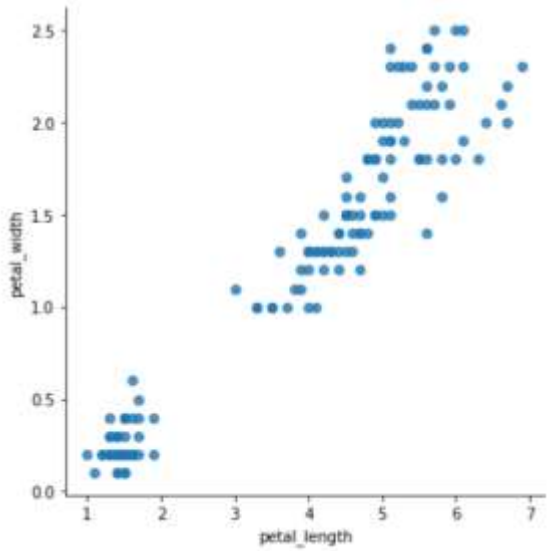
カーネル密度分布付き

- ヒストグラムをなめらかにすることにより、密度／確率を推定する

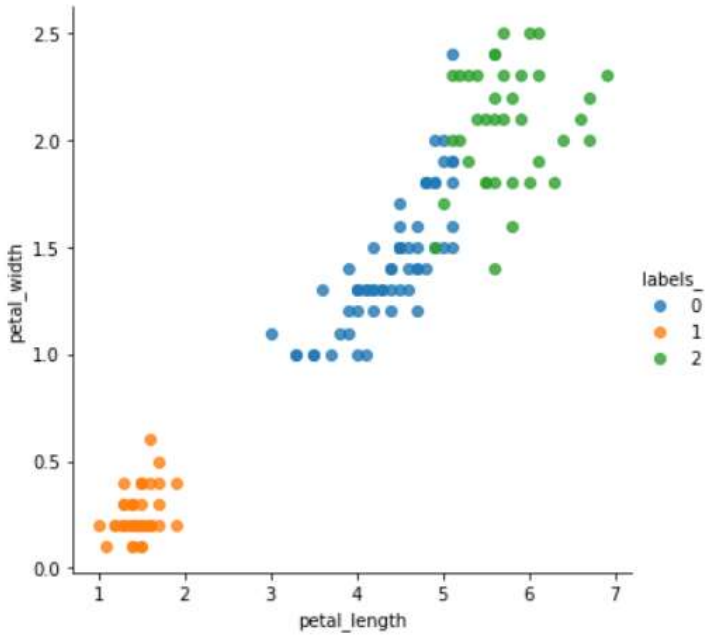
自動分類（クラスタリング）



元データ
（3種類の花のデータ）



自動分類
（クラスタリング）



種類情報なしのデータ

実習で気を付けて欲しいこと

前準備

Google アカウントの取得が必要

- 次のページを使用

https://accounts.google.com/SignUp

- 次の情報を登録する

氏名

自分が希望するメールアドレス

＜ユーザー名＞@gmail.com

パスワード

生年月日，性別

実習で気を付けて欲しいこと ①

実験で使う Google Colab のページ

URL は次の通り
https://colab.research.google.com/drive/1hY4O7yUV0zqcmHypRst1RnX2mt8_zatE?hl=ja#scrollTo=gIuquwzlcOck

URL は、**セレッソ**の**レポート**の「**6. 実験手順**」にも記載

6．実習手順
次の YouTube ビデオで説明している。
同じ内容の PDF ファイルおよびパワーポイントを添付している。

https://www.youtube.com/embed/

実習
https://colab.research.google.com/drive/1S55yEFiQpdIRdjWbdH0zzEYD5VAfklHd?hl=ja#scrollTo=7JdSy61xJGBv

実習で気を付けて欲しいこと②

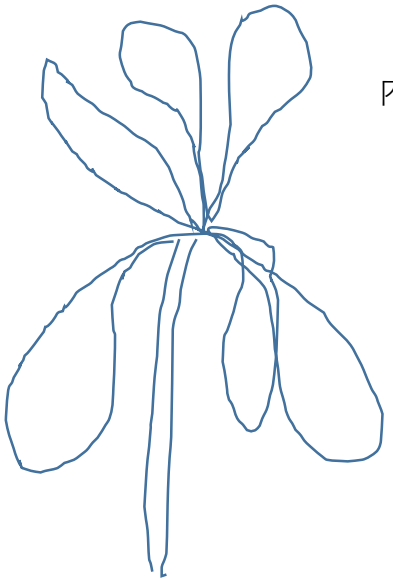
- **Google Colab を使う手順で説明**している

- Windows などのパソコンで動かすのは簡単
(興味のある人は各自で調べて試してみること)

1. Python のインストール
2. Windows のコマンドプロンプトを開き、次を実行

pip install matplotlib searborn scikit-learn

アヤメ属 (Iris)



◆多年草
◆世界に 150種. 日本に 9種.
◆花被片は 6個
　外花被片（がいかひへん） Sepal
　　3個（大型で下に垂れる）
　内花被片（ないかひへん） Petal
　　3個（直立する）

Iris Flower データセット

Iris Flower データセットは Python でも利用可能

```
In [1]: import seaborn

In [2]: iris = seaborn.load_dataset('iris')

In [3]: print(iris)
     sepal_length  sepal_width  petal_length  petal_width    species
0             5.1          3.5           1.4          0.2     setosa
1             4.9          3.0           1.4          0.2     setosa
2             4.7          3.2           1.3          0.2     setosa
3             4.6          3.1           1.5          0.2     setosa
4             5.0          3.6           1.4          0.2     setosa
..            ...          ...           ...          ...        ...
145           6.7          3.0           5.2          2.3  virginica
146           6.3          2.5           5.0          1.9  virginica
147           6.5          3.0           5.2          2.0  virginica
148           6.2          3.4           5.4          2.3  virginica
149           5.9          3.0           5.1          1.8  virginica

[150 rows x 5 columns]

In [4]:
```

- Ronald Fisher, 1936年
- 3種のアヤメの外花被辺、内花被片の**幅**と**長さ**を計測したデータセット

  Iris setosa

  Iris versicolor

  Iris virginica
- データ数は 50 × 3 種類