CS 5785 Homework 0

Scatterplots of Iris Data

Serge Belongie

Partner: Harrison Zhao, Haochen Jia

Aug.26.2017

Statement of Objective

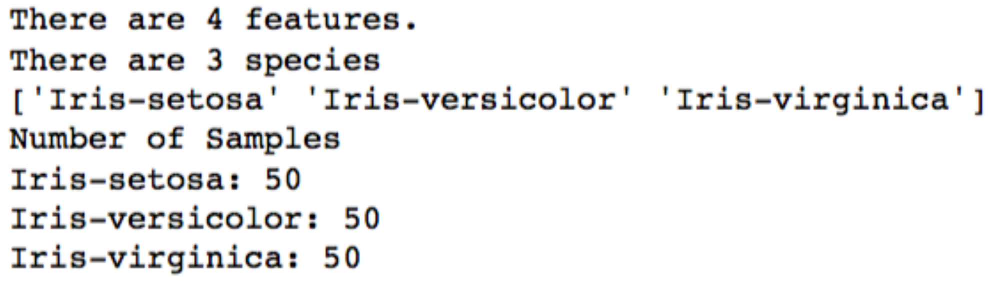
Edgar Anderson's iris flower data sets, include the length and width of the sepals and petals on several flowers in the field, are used as "sanity-check" for Python environment and plotting libraries in this project. We have 3 missions on this project: first, we need to download Python 3 environment, data sets of iris, and then figure out the properties of the data sets, such as the number of sample, species, and the number of features in each data. Second, we have to import the data set into python as data frames. Finally, we are requested to visualize the dataset in p-dimensional scatterplot by mapping it into 2D displays which plots two attributes of the data against another and repeat for each pair of attributes.

Procedure

We wrote a Python program for achieving goals. The first method we use is "pandas" to import the iris data downloaded from the internet. The second method we use is "matplotlib.pyplot" to print the all 2D scatterplot as the injection of high dimensional scatterplot map, respectively.

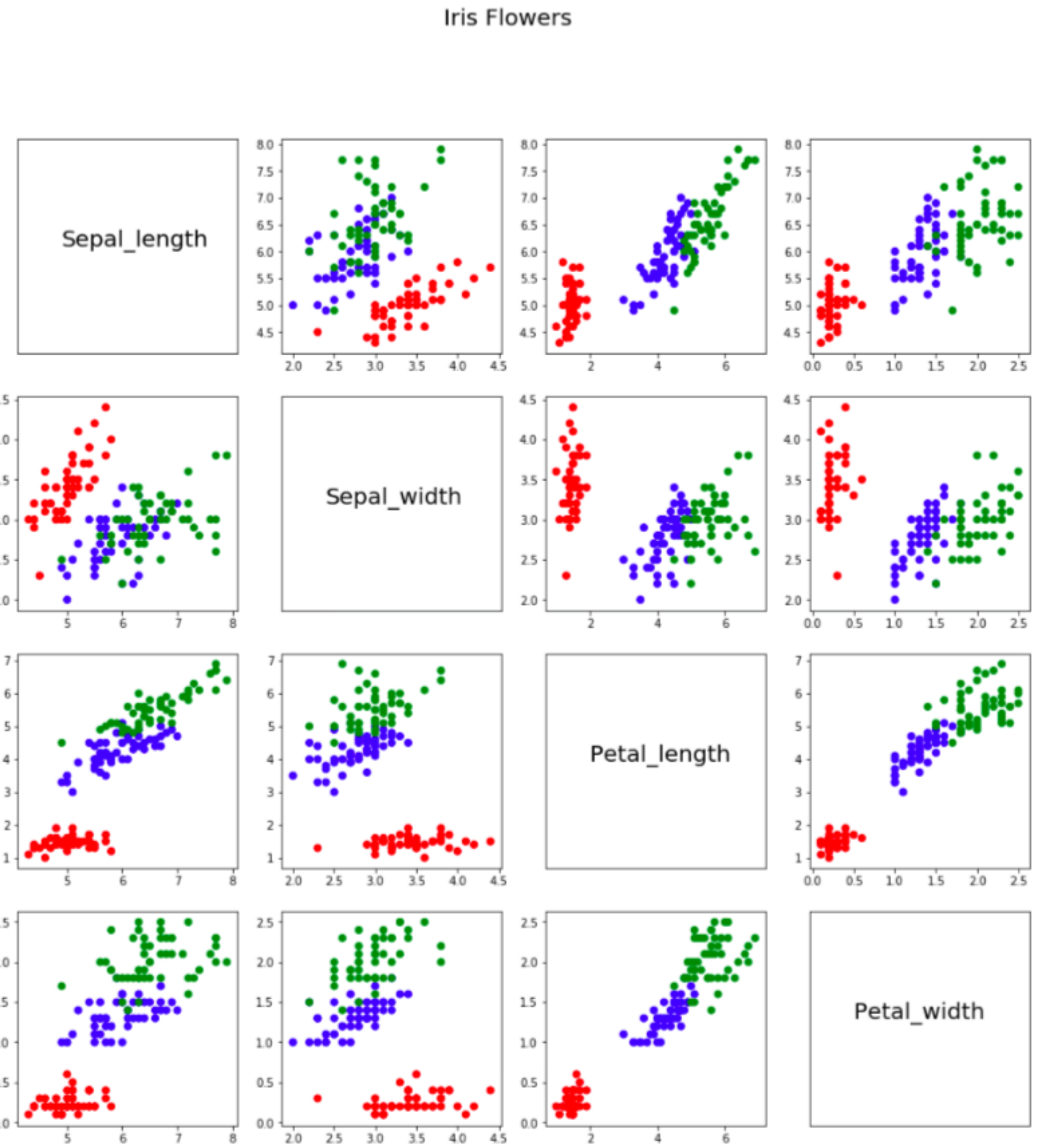
Data

We run the program to read the number of flower species, the number of each species, and the number of features for each sample.

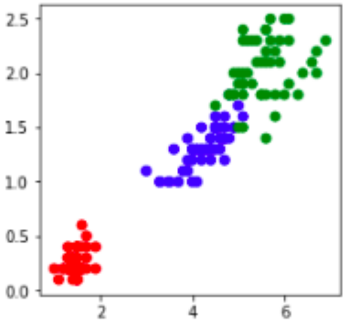
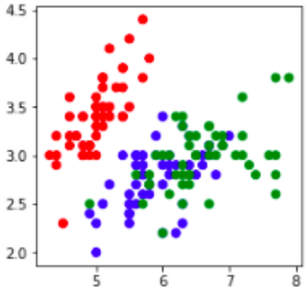


```
There are 4 features.
There are 3 species
['Iris-setosa' 'Iris-versicolor' 'Iris-virginica']
Number of Samples
Iris-setosa: 50
Iris-versicolor: 50
Iris-virginica: 50
```

Due to there are 4 features in each sample, that means 12 different 2D scatterplot maps are expected:



Analysis of data

	The clustering in graph “Petal_length VS Petal_width” is more clear than other graphs.
	The clusters in other graphs such as “Sepal_length VS Sepal_width” is not that clear.

Appendices

Here is the code for our program.

```
import pandas as pd
names = ['Sepal_length', 'Sepal_width',
        'Petal_length', 'Petal_width',
        'Class']
iris = pd.read_csv("iris.data", names = names)
iris.head()


# In[5]:

print( "There are 4 features.")
print( "There are " + str(len(iris.Class.unique()))
      + " species")
print(iris.Class.unique())
print("Number of Samples")
print("Iris-setosa: " +
    str(len(iris[iris.Class == 'Iris-setosa'])))
print("Iris-versicolor: " +
    str(len(iris[iris.Class == 'Iris-versicolor'])))
print("Iris-virginica: " +
    str(len(iris[iris.Class == 'Iris-virginica'])))


# In[6]:

import matplotlib.pyplot as plt
import matplotlib
get_ipython().magic('matplotlib inline')


# In[17]:

plt.figure(figsize = (16, 16))
features = [iris.Sepal_length, iris.Sepal_width,
        iris.Petal_length, iris.Petal_width]

for i, a in enumerate(features):
  for j, b in enumerate(features):
     if i != j:
        plt.subplot(4, 4, i+ 1 + j * 4)
        colors = iris.Class.replace({'Iris-setosa': 'r',
                                     'Iris-versicolor': 'b',
                                     'Iris-virginica': 'g'})
        plt.scatter(a, b, c=colors)
     else:
        fig = plt.subplot(4, 4, i + 1 + j * 4)
        plt.text(0.20, 0.5, names[i], fontsize=20)
        fig.axes.get_xaxis().set_visible(False)
        fig.axes.get_yaxis().set_visible(False)
plt.suptitle('Iris Flowers', fontsize=20)
```