Homework 0 for AML

by Hanfu Liao@ORIE5750, Yuxiang Zhu@CS5785

IRIS FLOWERS

0. Overview of the problem

Iris data set is perhaps the best known database to be found in the pattern recognition literature. Thus we aim to find a pattern from data by visualization. We write a python script to visualize the attributes of samples. Through the result of visualization, we find out a high correlation between the species of flowers and their petal length, and a high correlation between the species of flowers and their petal width.

1. How many features/attributes are there per sample? How many different species are there, and how many samples of each species did Anderson record?

- Number of attributes: 5 features including 4 numeric, predictive attributes and the species of the samples.
- Number of species: 3 species that are Iris Setosa, Iris Versicolour, and Iris Virginica.
- Number of Instances in each species: 50
- Attribute Information in detail:
    - 4 numeric:
        1. sepal length in cm
        2. sepal width in cm
        3. petal length in cm
        4. petal width in cm
    - Specie of the sample: Iris Setosa/Iris Versicolour/Iris Virginica

2. Data parse

We create two arrays to hold data: `data` and `labels`. `data` is a 150 x 4 array contains 4 arrays stand respectively 4 numeric, predictive attributes in each sample: sepal length, sepal width, petal length and petal width. The element number of each array is 150 which is the total sample number. `labels` is a 1D array with 150 elements recording species in each sample.

```
data = [[],[],[],[]]
```

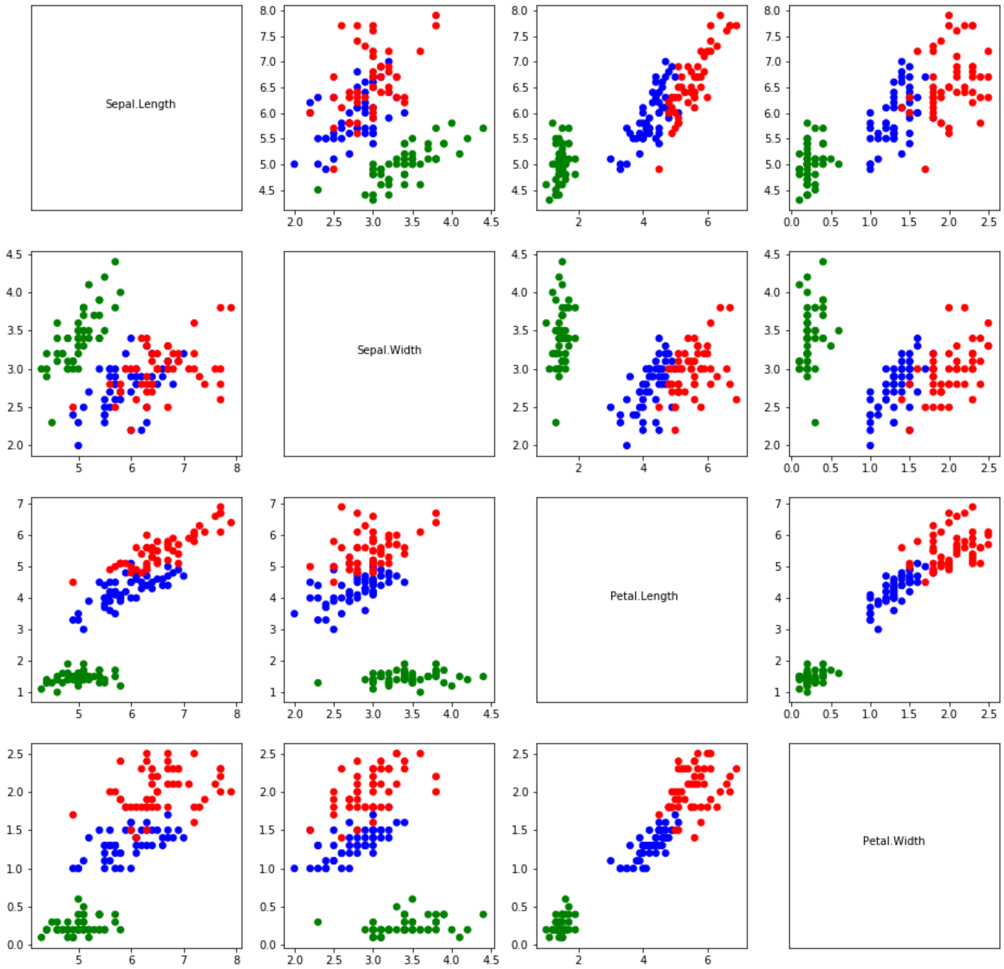
```
labels = []
for line in open("iris.data.txt"):
    if line.strip() == '': continue
    origin = line.strip().split(',')
    for i in range(4):
        data[i].append(origin[i])
    labels.append(origin[-1])
h = [data[0] ,data[1],data[2],data[3]]
```

3. Visualization

In order to visualize, we use `matplotlib` to create a 4 by 4 matrix. We plot each group of two attributes out of four attributes in a subplot and the attribute names are marked on the diagonal subplot. With respect to the different entries of the data, we put subplots accordingly. Precisely, we map three species of the sample to red, blue and green using a function called `choose_color`, where the species the first read in was mapped to red color, the second one was mapped to blue color, and the third one was mapped to green color. In this case, Iris-setosa is in red, Iris-versicolor is in blue, and Iris-virginica is in green.

```
plt.figure(figsize = (16,16))
for i in range(4):
    for j in range(4):
        if i == j:
            fig = plt.subplot(4,4,i+4*j+1)
            fig.axes.get_xaxis().set_visible(False)
            fig.axes.get_yaxis().set_visible(False)
            cap = [”Sepal.Length”,”Sepal.Width”,”Petal.Length”,”Petal.Width”]
            plt.text(0.35, 0.5, cap[i])
        else:
            plt.subplot(4,4,i+4*j+1)
            plt.scatter(h[i], h[j], c=colors) from matplotlib import pyplot as plt
def choose_color(x):
    color = ["r","b","g"]
    i = list(set(labels)).index(x)
    return color[i]
colors = list(map(choose_color, labels))

plt.figure(figsize = (16,16))
for i in range(4):
    for j in range(4):
        if i == j:
            fig = plt.subplot(4,4,i+4*j+1)
            fig.axes.get_xaxis().set_visible(False)
            fig.axes.get_yaxis().set_visible(False)
            cap = ["Sepal.Length","Sepal.Width","Petal.Length","Petal.Width"]
            plt.text(0.35, 0.5, cap[i])
        else:
            plt.subplot(4,4,i+4*j+1)
            plt.scatter(h[i], h[j], c=colors)
```



4. Insight

We can distinguish more precisely the species of a flower given the length of petal or the width of petal.