Assignment 0
CS5785 Applied Machine Learning

Yixuan Li (Rylee) yl2557
Ningran Song (Sunny) ns632

08/31/2021

Summary

In this assignment, we analyzed the Iris Flower dataset. We first figured out the structure and features of the dataset. Then we loaded the samples into a 150 x 4 matrix with a 150-dimensional vector containing each sample's label. Finally we visualized this dataset by creating all possible scatterplots with all pairs of two attributes.

Questions

- How many features/attributes are there per sample? 4 numeric, predictive attributes and the class.
- How many different species are there? 3 species.
- How many samples of each species did Anderson record? 50 instances in each of three species.

Solutions

The solution works well as the figure we plotted (Figure 1) is the same as the figure showed in the assignment requirement. Below is the source code that generated the figure.

Load Data

```
import matplotlib.pyplot as plt
import numpy as np

# Load the samples into an 150 * 4 array
# where N is the number of samples and p is the number of attributes per sample
data = np.loadtxt("iris.data", delimiter=',', usecols=range(0,4));
print (data.shape);

# create a N-dimensional vector containing each sample's label (species)
vector = np.loadtxt("iris.data", delimiter=',', usecols=range(4,5), dtype=str);
print (vector.shape);
```

Visualization

```
# map color from species
colorMap = {
    'Iris-setosa': "r",
    'Iris-versicolor': "g",
    'Iris-virginica': "b"
}
colors = list(map(lambda x: colorMap[x], vector));
# print (colors);

# plot data points
fig, axs = plt.subplots(4, 4, figsize=(12,12));

attributes = ["Sepal.Length", "Sepal.Width", "Petal.Length", "Petal.Width"];

for i in range(len(attributes)):
    for j in range(len(attributes)):
```

```
        if i == j:
            axs[i, j].text(0, 0, attributes[i],
                           horizontalalignment='left', verticalalignment='bottom');
        else:
            axs[j, i].scatter(data[:, i], data[:, j], c = colors, s = 7);

# save pics
plt.savefig ("hw0_iris_flower_plot.png");
```

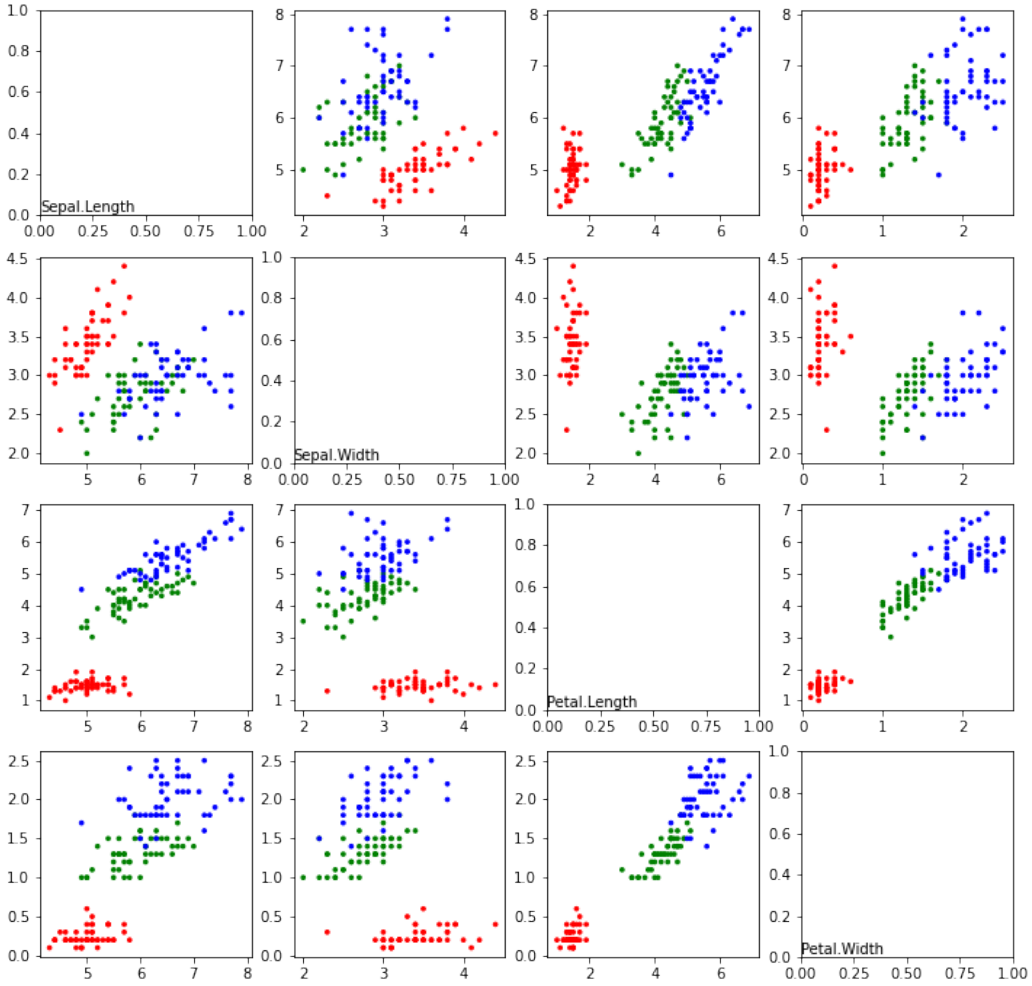

Figure 1: Solution

Insights

From the above figure, we can see that there are two major clusters. One contains the Iris-setosa (red points), and the other contains Iris-versicolor green points) and Iris-virginica (blue points). Iris-setosa can be easily identified, but the other two species have some overlap. We think we can further train a machine learning model using a multi-class classification algorithm based on this dataset in order to make species predictions.