CS 5785 Applied Machine Learning
Homework 0

Ruoyu Wang (rmw252), Cyrus Ghazanfar (cg595)

August 31, 2018

1. Setting up Python

1.1 Install Python

Download and install Python 2.7 release for OSX from official site:

$$https://www.python.org/download/releases/2.7.7/ \tag{1.1}$$

1.2 Install pip

```
curl https://bootstrap.pypa.io/get-pip.py -o get-pip.py
```
(1.2)

```
python get-pip.py
```
(1.3)

1.3 Install Jupyter Notebook

```
pip install jupyter
```
(1.4)

2. Iris Flowers

2.1 Get Iris Flowers data

Running:

```
wget http://archive.ics.uci.edu/ml/machine-learning-databases/iris/bezdekIris.data
```
(2.5)

will download the CSV data file.
Running:

```
wget http://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.names
```
(2.6)

will download the classification descriptions.

2.2 Read data

Python code to read Iris data from csv file:

```
# import libs
from matplotlib import pyplot as plt
import numpy as np
import csv
import itertools
```

```
# Read data from CSV
raw_data = np.array(list(csv.reader(open('iris.data'), delimiter=',')))
color_map = {
    'Iris-setosa': "r",
    'Iris-versicolor': "g",
    'Iris-virginica': "b"
}
data = [raw_data[:, i].astype(np.float) for i in range(4)]
colors = list(map(lambda x: color_map[x], raw_data[:,4]))
```

2.3 Plot scatter graph

```
# Plot the scatter graph
features = ("Sepal Length", "Sepal Width", "Pedal Length", "Pedal Width")
permutations = list(itertools.permutations([0,1,2,3], 2))

plt.subplots(4,4,figsize=(13,13))

for p in permutations:
    position = p[0]*4 + p[1] + 1
    plt.subplot(4, 4, position)
    xs = data[p[1]]
    ys = data[p[0]]
    plt.scatter(xs, ys, c=colors)
    plt.xlabel(fetures[p[0]])
    plt.ylabel(fetures[p[1]])
```

2.4 Insights from the plot

1. The iris flowers can be clustered by combinations of any two of the four features: sepal length, sepal width, pedal length and pedal width.
2. Some pair of features show much better clustering results, such as pedal width + pedal length, sepal width + pedal width, etc. Some are not as good, such as pedal width + sepal length. This indicates that pedal width + sepal length might not be a good basis for describing an iris flow.
3. Iris Setosa is obviously easier to be classified using the features. This implies the four feature we are observing work better for this flower.

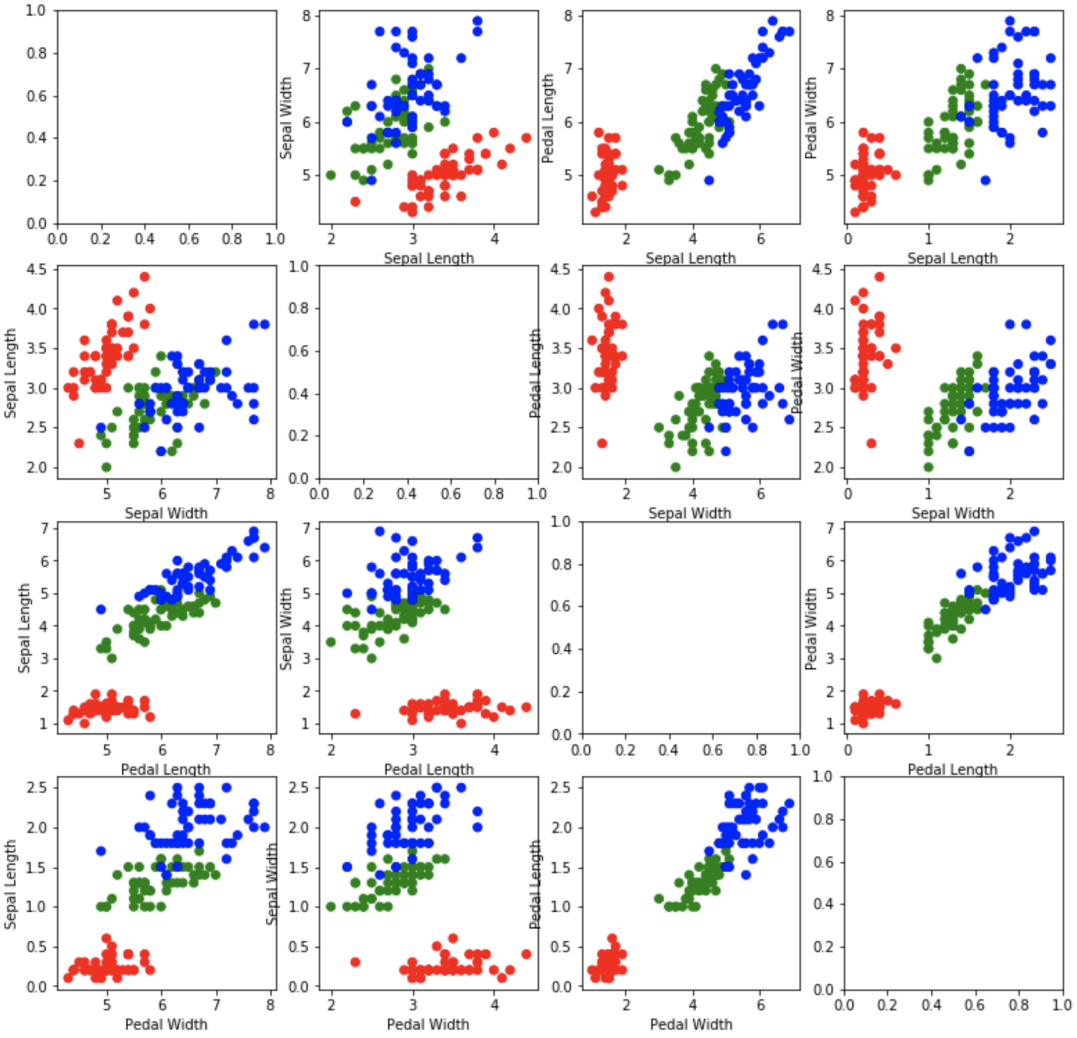

Figure 1: Iris Flowers