CS5785: Assignment #0

Due on Wednesday, August 30, 2017

Prof. Serge Belongie

Noshin Anjum Nisa

August 30, 2017

Abstract

An introductory homework to machine learning using a simple dataset to visua In this assignment, we have set up a working Python development environment using Anaconda for Machine Learning. The code is written and compiled in the "IPython Notebook". We have cited all the code and references used for this assignment.

1 Introduction

In this assignment a machine learning development environment has been set up using python on Anaconda [1]. The intention is to download a simple dataset and parse it using either Panda and visualize it in 2d using Matplotlib. The 2d visualization is simple and will help us understand the data.

2 Questions

(i) *How many features/attributes are there per sample?*
There are five attributes per sample: sepal length, sepal width, petal length, petal width and species.

(ii) *How many different species are there?*
There are three different species: Iris Setosa, Iris Versicolour and Iris Virginica.

(iii) *How many samples of each species did Anderson record*
There are 50 samples of each species, all together there are 150 samples.

(iv) *Load the samples into an Nxp array where N is the number of samples and p is the number of attributes per sample. Additionally, create a N-dimensional vector containing each sample's label(species)*
The samples code 1 contains the Nxp array and the N-dimensional vector containing the species.

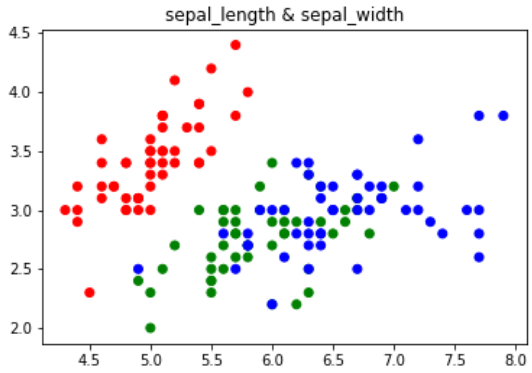
3 Procedures

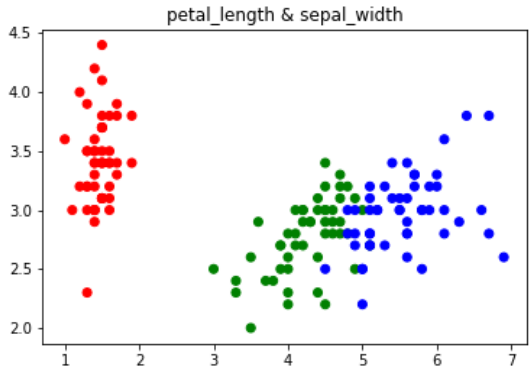
The first step was to install Anaconda [1], which is a Python data science platform that allows us to set up a machine learning environment. I decided to use the Jupyter Notebook that comes with Anaconda because a notebook can include descriptions, comments, graphs, and figures in-line with the source code that generated them. To begin the notebook I ran 'jupyter notebook' on terminal, which started a Notebook server with a web browser link where I created a new notebook for my project.
I imported all the necessary libraries that I can possibly use for the project. Later, I loaded the data from a url and parsed it using panda. The dataset was then divided according attribute and assigned one color for each class (species).

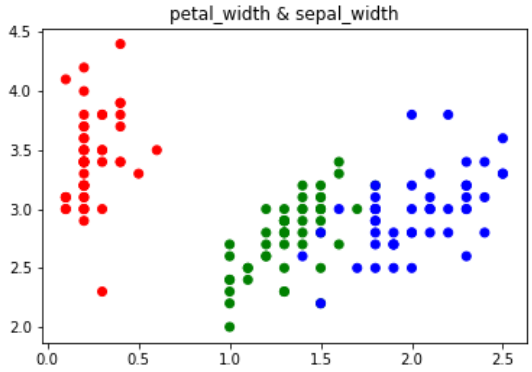
These divided datasets are now used to plot 2d graphs. The red color is assigned to Iris-setosa, green to Iris-versicolor and blue to Iris-Virginica.

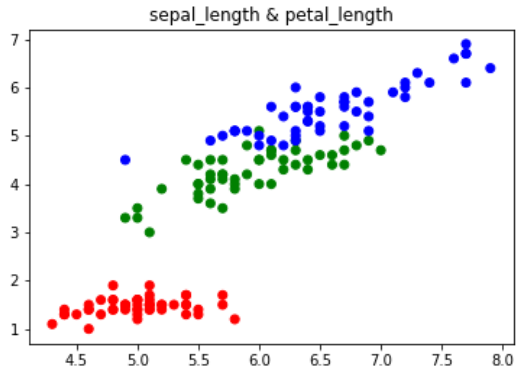
4 Scatter Plots

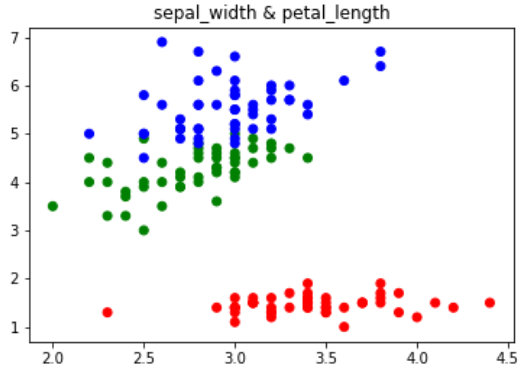
Red - Iris Setosa, Green - Iris Vesicolour, Blue - Iris Virginica

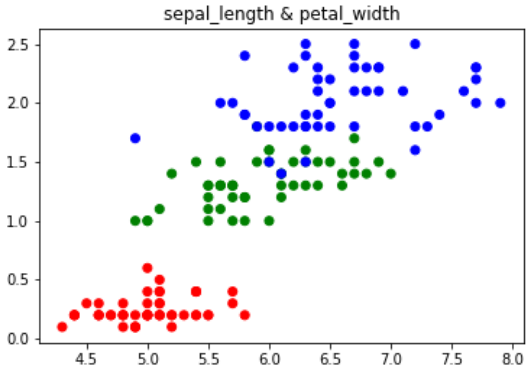


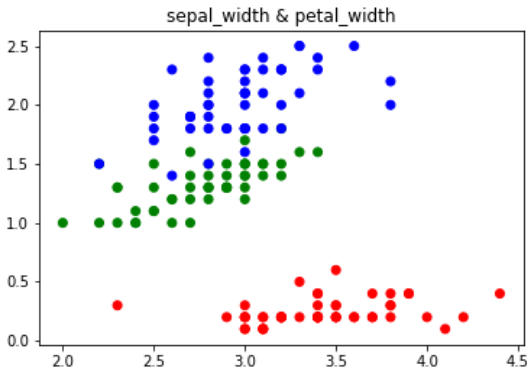


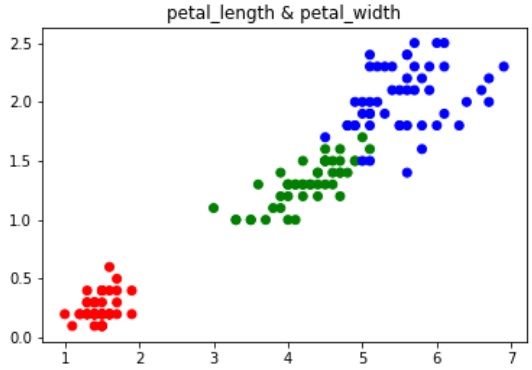


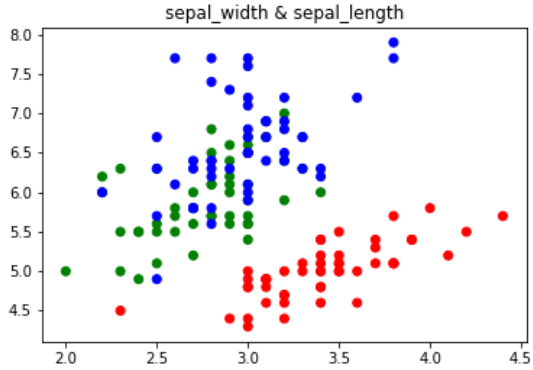

sepal_length & petal_length
7
6
5
4
3
2
1
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0

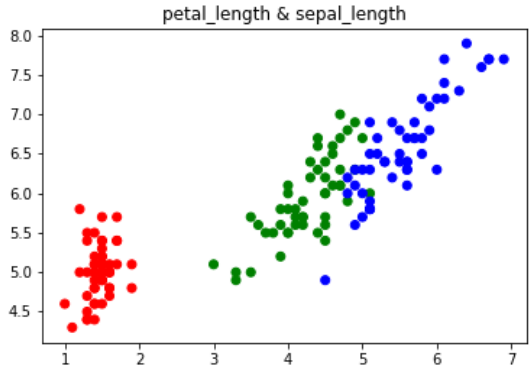

sepal_width & petal_length
7
6
5
4
3
2
1
2.0
2.5
3.0
3.5
4.0
4.5

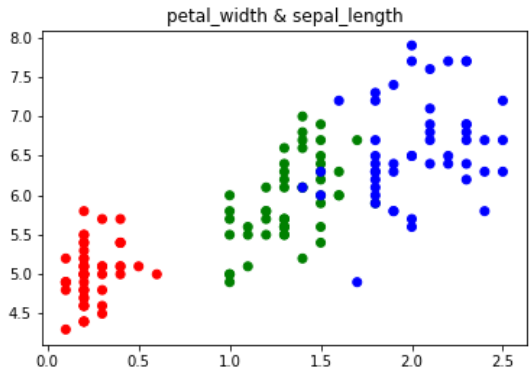

sepal_length & petal_width
2.5
2.0
1.5
1.0
0.5
0.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0

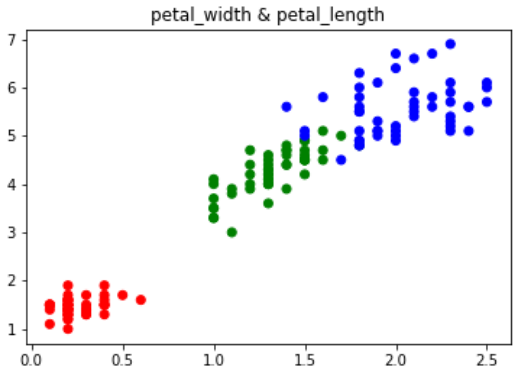

sepal_width & petal_width
2.5
2.0
1.5
1.0
0.5
0.0
2.0
2.5
3.0
3.5
4.0
4.5


petal_length & petal_width
2.5
2.0
1.5
1.0
0.5
0.0
1
2
3
4
5
6
7


sepal_width & sepal_length
8.0
7.5
7.0
6.5
6.0
5.5
5.0
4.5
2.0
2.5
3.0
3.5
4.0
4.5


petal_length & sepal_length
8.0
7.5
7.0
6.5
6.0
5.5
5.0
4.5
1
2
3
4
5
6
7


petal_width & sepal_length
8.0
7.5
7.0
6.5
6.0
5.5
5.0
4.5
0.0
0.5
1.0
1.5
2.0
2.5



5 Source Code

```
1
2  #Importing necessary libraries
3  import pandas as pd
4  import numpy as np
5  #Library to plot graphs
6  from matplotlib import pyplot as plt
7  #To ensure inline plotting
8  %matplotlib inline
9
10
11 #Loading the data from a url and setting heading names for each data column
12 dataSetUrl = "https://archive.ics.uci.edu/ml/machine-learning-databases/iris/
       iris.data"
13 dataSet = pd.read_csv(dataSetUrl, names=['sepal_length', 'sepal_width', '
       petal_length', 'petal_width', 'class'])
14
15 #Once the data is loaded, they are divided according to their header names,
       which makes it easier to plot and assign color
16 sepal_length = dataSet['sepal_length']
17 sepal_width = dataSet['sepal_width']
18 petal_length = dataSet['petal_length']
19 petal_width = dataSet['petal_width']
20
21 #each column gets a color based on which class they belong to
22 flower_class = dataSet['class']
23 colors = []
24 for i in flower_class:
25     if i == 'Iris-setosa':
26         colors.append('r');
27     elif i == 'Iris-versicolor':
28         colors.append('g');
29     else:
30         colors.append('b');
31
32
33 #In the code below we are plotting 2d graph for two attributes per graph
```

```
plt.scatter(sepal_length, sepal_width, c=colors)
plt.savefig("sepal_length&sepal_width.png")

plt.figure()
plt.scatter(sepal_length, petal_length, c=colors)
plt.savefig("sepal_length&petal_length.png")

plt.figure()
plt.scatter(sepal_length, petal_width, c=colors)
plt.savefig("sepal_length&petal_width.png")

plt.figure()
plt.scatter(sepal_width, sepal_length, c=colors)
plt.savefig("sepal_width&sepal_length.png")

plt.figure()
plt.scatter(sepal_width, petal_length, c=colors)
plt.savefig("sepal_width&petal_length.png")

plt.figure()
plt.scatter(sepal_width, petal_width, c=colors)
plt.savefig("sepal_width&petal_width.png")

plt.figure()
plt.scatter(petal_length, sepal_length, c=colors)
plt.savefig("petal_length&sepal_length.png")

plt.figure()
plt.scatter(petal_length, sepal_width, c=colors)
plt.savefig("petal_length&sepal_width.png")

plt.figure()
plt.scatter(petal_length, petal_width, c=colors)
plt.savefig("petal_length&petal_width.png")

plt.figure()
plt.scatter(petal_width, sepal_length, c=colors)
plt.savefig("petal_width&sepal_length.png")

plt.figure()
plt.scatter(petal_width, sepal_width, c=colors)
plt.savefig("petal_width&sepal_width.png")

plt.figure()
plt.scatter(petal_width, petal_length, c=colors)
plt.savefig("petal_width&petal_length.png")
```

Listing 1: Source Code

6 Analysis

From the data visualization we can analyze the differences and similarities among the different species. Such as, Iris Versicolor and Iris Virginica flowers tend to have longer petals

compared to Iris Setosa.

References

[1] *Anaconda*, available at https://www.anaconda.com/download/

[2] *Dataset*, available at http://archive.ics.uci.edu/ml/datasets/Iris

[3] *To read dataset directly from a url and name their columns*, https://stackoverflow.com/questions/32400867/pandas-read-csv-from-url

[4] *To label graphs*, www.matplotlib.org/users/pyplot_tutorial.html/