CS5785 Applied Machine Learning

Homework 0

Student Names: Chen Zhang, Chao Suo
Contact: cz365@cornell.edu, cs2326@cornell.edu

Due Date: 08/31/2016

Table of Contents

Problem

The whole Iris Flower problem can be break down into three parts:

First, analyze the dataset and identify the whole structure of data:
a) How many attributes are there per sample?
b) How many different species are there and how many samples of each species did Anderson record?

As the structure of the data has been determined, the second step will be to parse and process data so that we can better use them to generate visualized outcome.

Thirdly, we will use the data arrays prepared in the previous step to plot graphs, which will allow us to better analyze the relationship among the four attributes.

Instructions

In the command line, run: python hw0.py
As result of running this code, six graphs will be generated as output with appropriate names.

Methodology

The first part of the problem can be easily detected from the records shown Iris data file such as:

5.1,3.5,1.4,0.2,Iris-setosa

Each row as above contains 4 attributes describing the sepal and petal dimensions and 1 species name. Since there are three species within total 150 records, each species includes 50 samples.

Known that one row of data includes attributes and species name, we need to separate them so that we can use the number part to achieve visualization, which is what we did for the second problem. Details can be listed as follows:

a) Load the data from data file fetched from the web. In this step, we named the file as txt type and used numpy.genfromtxt [1] command built in numpy library [2] to transfer the data into Python because it is an easy to execute code that can reduces errors that may happen in traditional methods.

b) During that process, we added constraints (usecols=[0,1,2,3] & usecols=[4]) to parse the data to be first four columns as an *N x p* array and the fifth column as an *N*-dimensional vector. In this homework case, N=150, p=4. Please refer to data_set and data_symbols respectively in our original code.

With the data parsed, *N x p* array and *N*-dimensional vector ready, we are going to visualize the dataset by plotting the scatters. Before that, we need to prepare the colors for the three species to differentiate them in the graph to be plotted. By using a for loop and colors.append code, we detected each type of species and assigned them with red, blue and green color respectively.

As for the *N x p* array, we furtherly divided it into four separate attribute vectors with name sepal length, sepal width, petal length and petal width. These vectors are going to be plotted on the graph to visualize the data. There will be six 2D displays in total among the attributes against each other, and twelve if considering the order as well.

Results

After running the program, we have the following six scatter graphs visualizing the relationship between each two attributes of Iris Flowers. Please see figure 1.1 to 1.6.

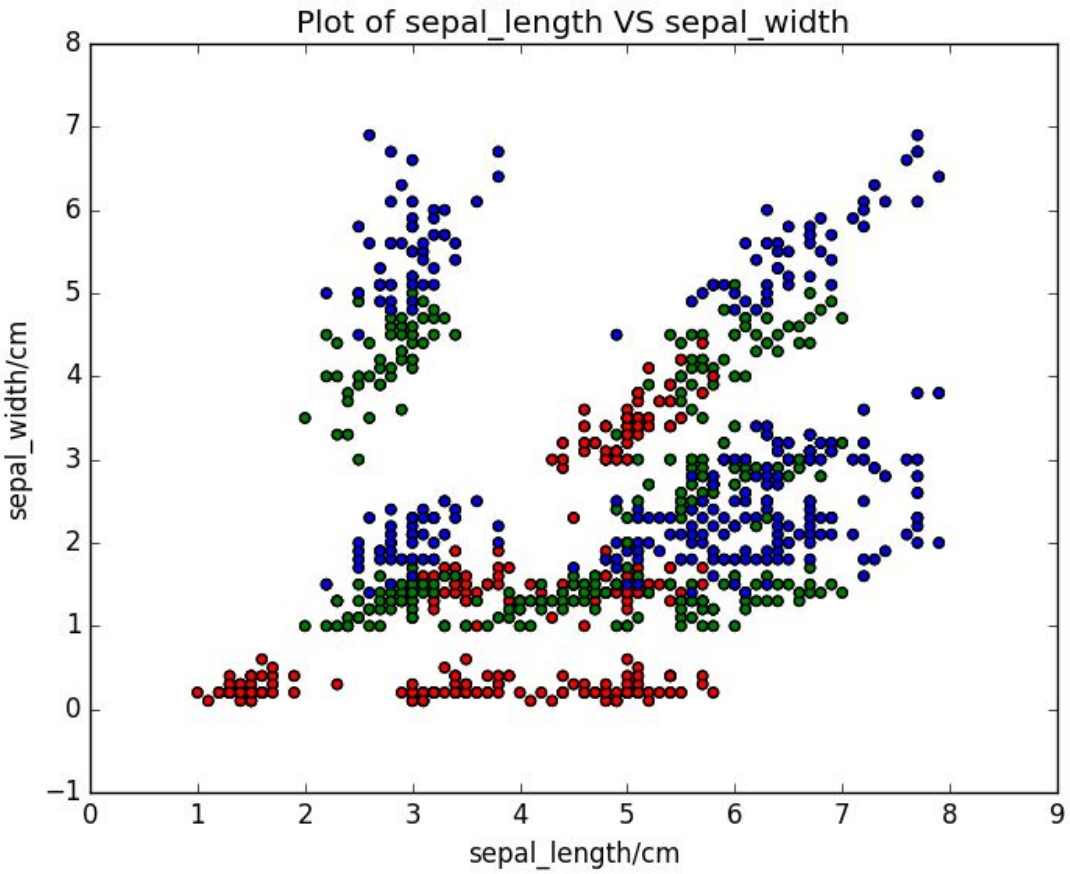

Figure 1.1 sepal length vs sepal width

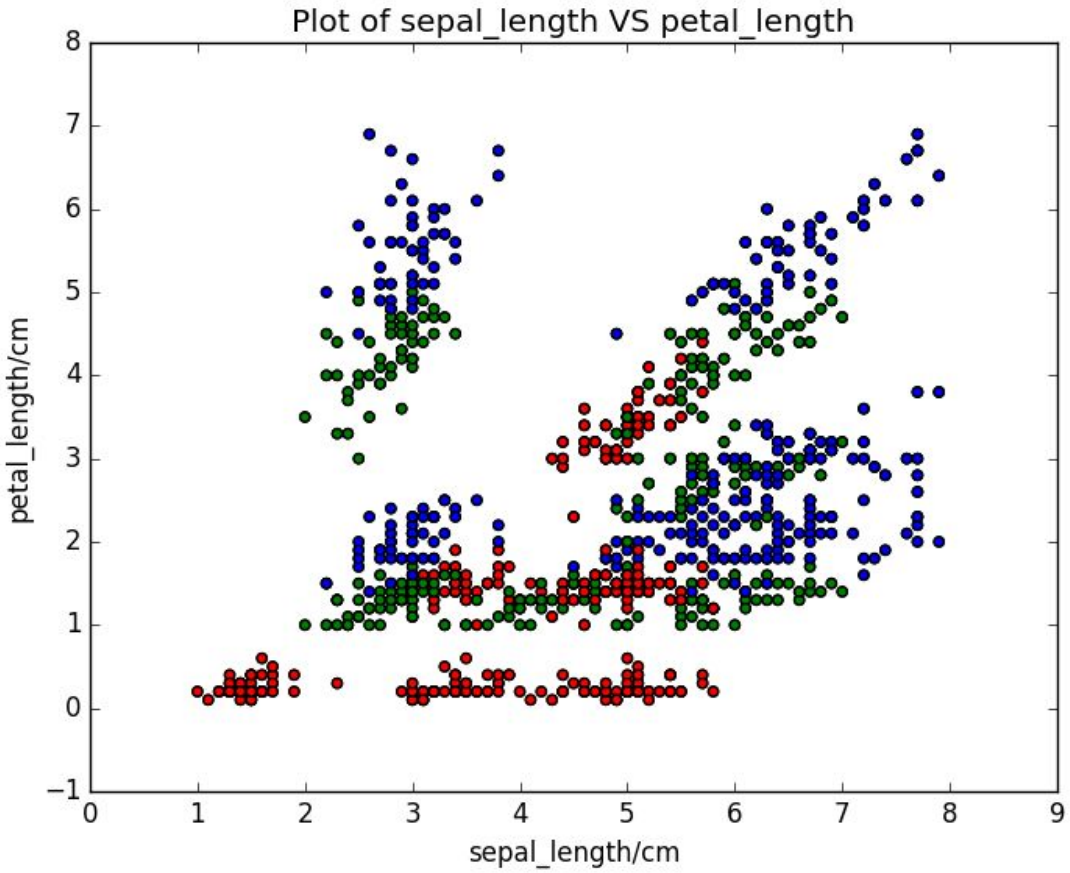

Figure 1.2 sepal length vs petal length

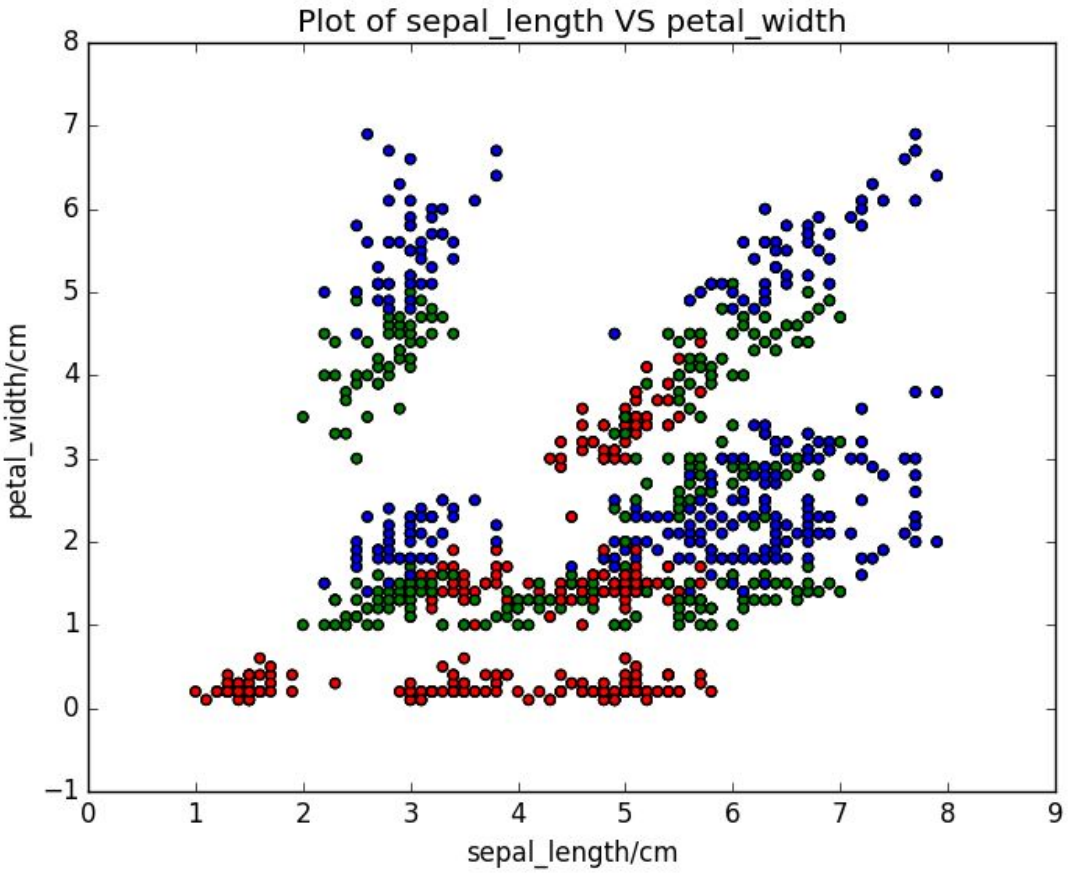

Figure 1.3 sepal length vs petal width

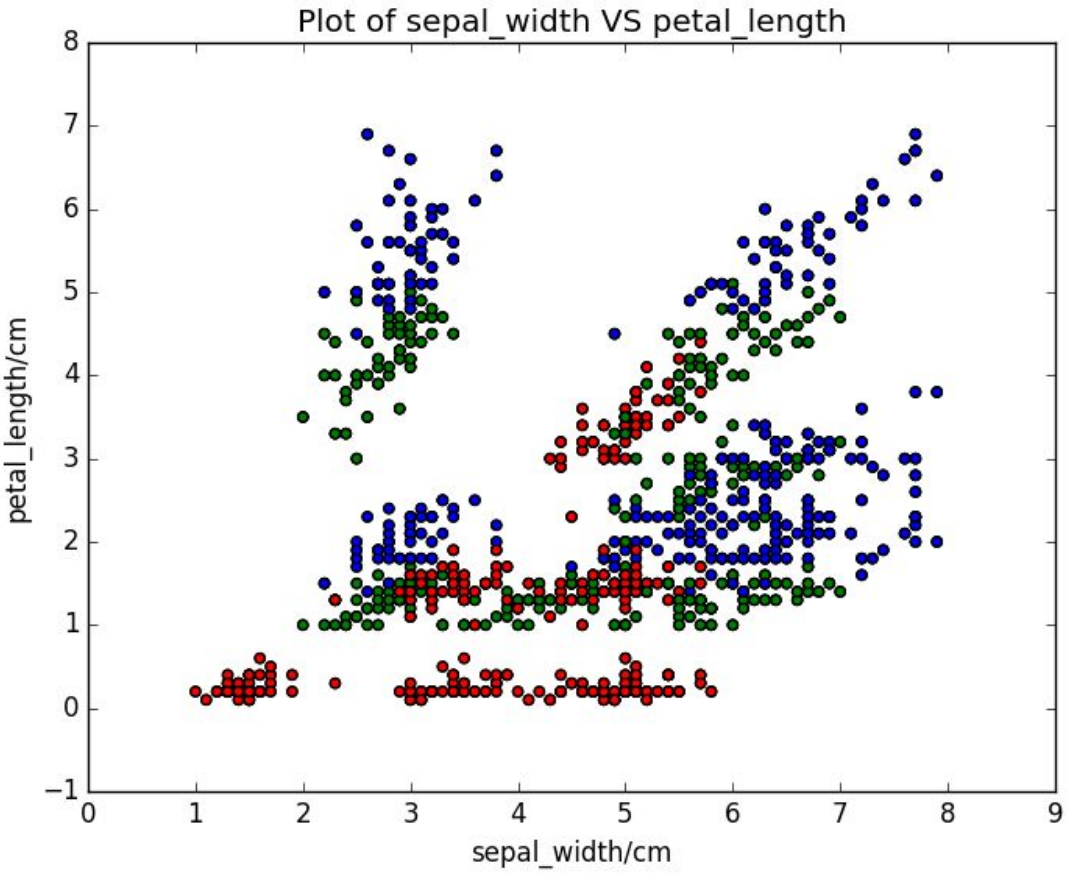

Figure 1.4 sepal width vs petal length

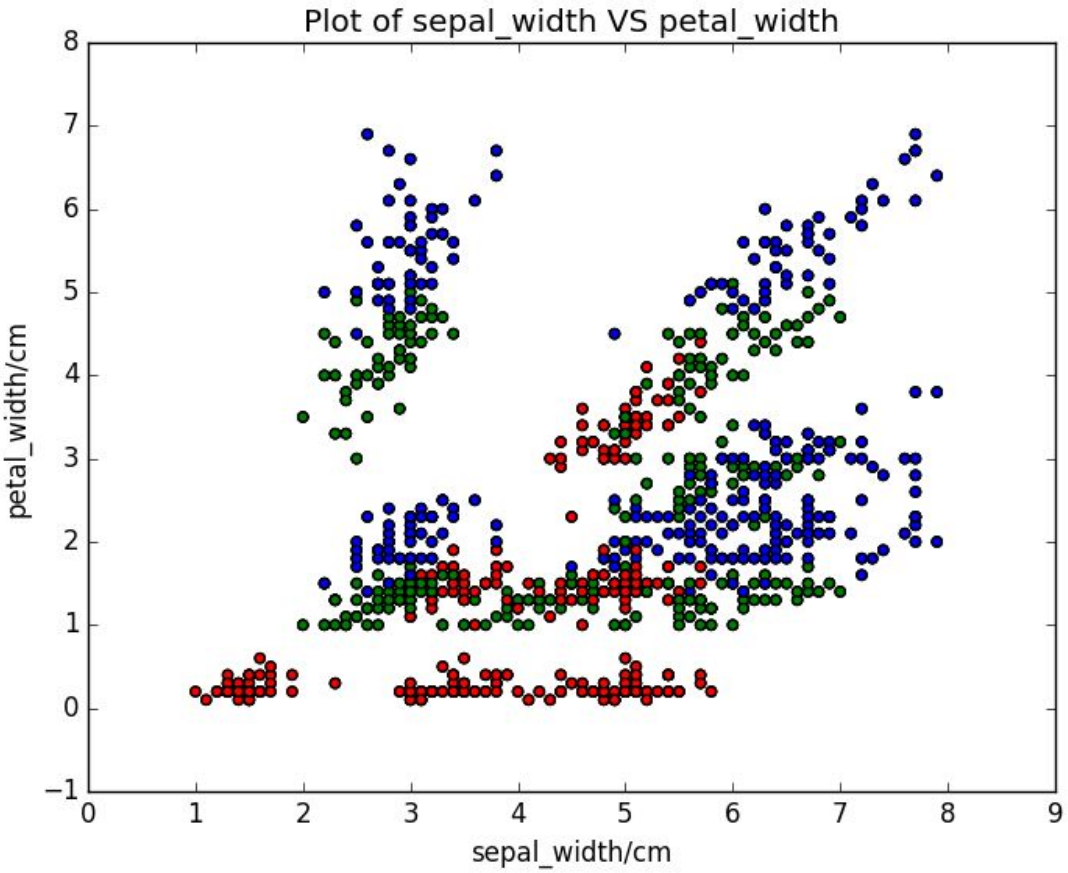

Figure 1.5 sepal width vs petal width

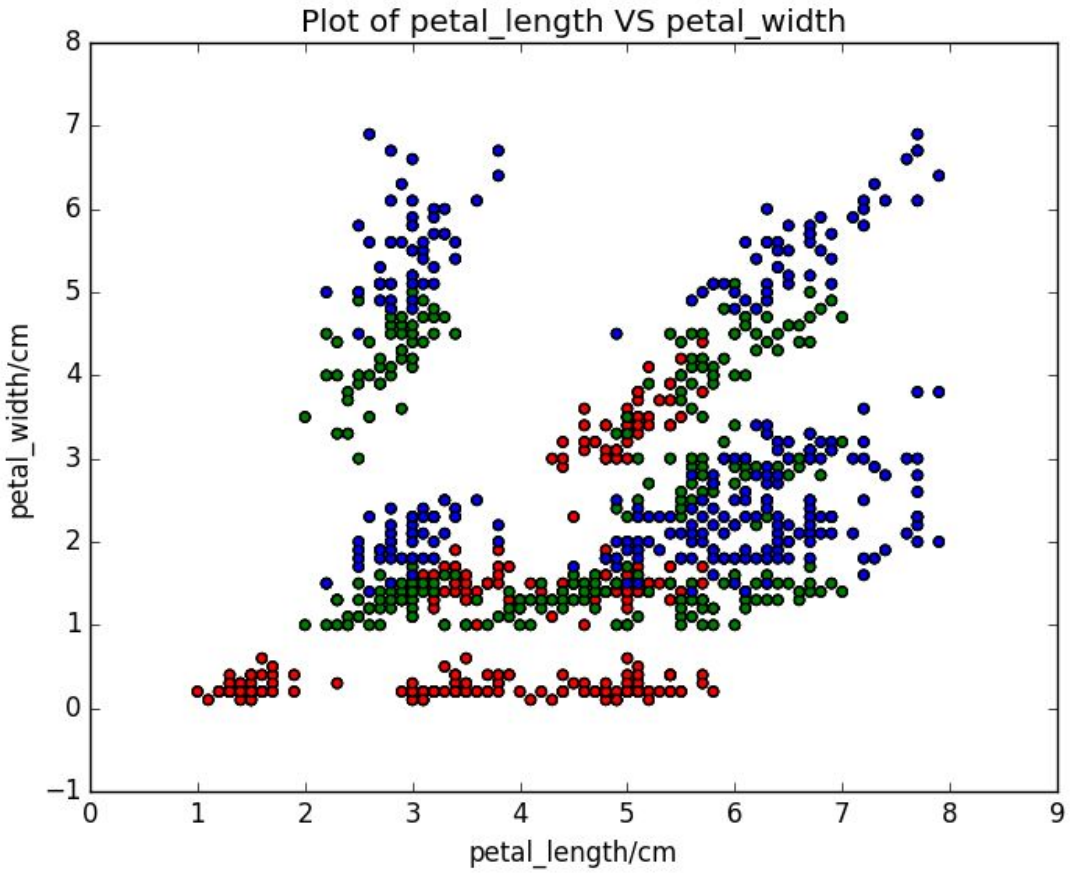

Figure 1.6 petal length vs petal width

Reference

[1] Retrieved August 30, 2016 from http://docs.scipy.org/doc/numpy/user/basics.io.genfromtxt.html

[2] Retrieved August 30, 2016 from http://docs.scipy.org/doc/numpy/reference/