ASSIGNMENT 0

September 5, 2019

Bar Kadosh (bk497), Ben Kadosh (bk499)
CS5785: Applied Machine Learning
Instructor: Nathan Kallus, Teaching Assistants: Xiaojie Mao, Yichun Hu

Contents

0.1 Summary

Our general approach to this assignment was to first organize the data into an easy to use and readable format. To do so, we utilized the pandas and numpy libraries to create a dataframe and the necessary array structures. We then created vectors for each of the individual attributes so that we could plot the necessary attributes against each other. Furthermore, in order to create relevant and insightful plots, we assigned each unique class a different color – red, green, and blue for Iris Setosa, Iris Versicolor, and Iris Virginica, respectively. We accomplished this by creating a new array that replaced the class names with their respective colors and then used this new array of colors during the plotting phase. Finally, we proceeded to plot each of the four numeric, predictive attributes against the other three. We split these plots into two figures with each figure containing two rows of three plots (six plots per figure).

After analyzing the plots, it can be seen that Iris Setosa has a clear separation from the other two species in each plot. In addition, while Iris Versicolor and Iris Virginica are more similar to each other than either is to Iris Setosa (as it relates to Sepal Length, Sepal Width, Petal Length, and Petal Width), each class generally has noticeable clusters that distinguish it from the other classes (although there is some overlap).

0.2 Problem 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?

The five features/attributes for each sample are:

- Sepal Length in cm (numeric, predictive)
- Sepal Width in cm (numeric, predictive)
- Petal Length in cm (numeric, predictive)
- Petal Width in cm (numeric, predictive)
- Class/Species

In the dataset, samples exist for three different species (Iris Setosa, Iris Versicolour, Iris Virginica). For each species, Anderson recorded 50 samples.

0.3 Problem 2

Parse the dataset you downloaded and load the samples into an N x p array, where N is the number of samples and p is the number of attributes per sample. Aditionally, create an N-dimensional vector containing each sample's label.

In order to parse the dataset, we utilized the pandas library to read the dataset into a pandas dataframe. We then leveraged the numpy library to convert the dataframe into a numpy array in the shape of an N x p array.

```python
# read data into a pandas dataframe and define the names of the columns
iris_df = pd.read_csv("iris.data", header = None, names = ["Sepal Length", "Sepal Width", "Petal Length", "Petal Width", "Class/Species"])

# convert the pandas dataframe to a two-dimensional array (N by p)
iris_data_array = iris_df.to_numpy()
```

Listing 1: Python Listing 1

After creating the N x p array, we used python slicing to create an N-dimensional vector containing each sample's label (species), as well as N-dimensional vectors for each of the other attributes in the dataset.

```python
# extract the four numerical attributes from the two dimensional array and create a vector for each attribute
sepal_length = iris_data_array[:,0]
sepal_width = iris_data_array[:,1]
petal_length = iris_data_array[:,2]
petal_width = iris_data_array[:,3]

# extract the descriptive class attribute from the two dimensional array and create a vector for it
iris_class_labels = iris_data_array[:,4]
```

Listing 2: Python Listing 2

0.4 Problem 3

Create every possible scatterplot from all pairs of two attributes (for example, one scatterplot would graph petal length vs sepal width, another would graph petal length vs. sepal length, and so on). Within each scatterplot, the color of each dot should correspond with the sample species.

In order to plot each combination of the pairs of attributes against each other, we first created N-dimensional vectors for each of the four numeric attributes (shown in **Listing 2**).

We then created an additional N-dimensional vector 'iris_class_colors', in which we replaced the existing 'Iris-Setosa', 'Iris-Versicolor', and 'Iris-Virginica' class values with the color values 'r', 'g', and 'b', respectively (where 'r' is red, 'g' is green, and 'b' is blue).

```python
iris_class_colors = iris_class_labels
iris_class_colors = np.where(iris_class_colors == "Iris-setosa", "r", iris_class_colors)
iris_class_colors = np.where(iris_class_colors == "Iris-versicolor", "g", iris_class_colors)
iris_class_colors = np.where(iris_class_colors == "Iris-virginica", "b", iris_class_colors)
```

Listing 3: Python Listing 3

Once we had the underlying data prepared, we proceeded with plotting the necessary data. In order to do so, we created two figures, each containing 6 scatter plots.

The first figure has two rows of three plots. The first row plots Sepal Length on the x-axis and the other three attributes on the y-axis. The second row plots Sepal Width on the x-axis and the other three attributes on the y-axis.

The second figure also has two rows of three plots. The first row plots Petal Length on the x-axis and the other three attributes on the y-axis. The second row plots Petal Width on the x-axis and the other three attributes on the y-axis.

Each figure has a title and each plot within the figures has x and y labels.

```python
fig = plt.figure(figsize=(18,12))
fig.suptitle("Iris Data (Red = Iris-Setosa, Green = Iris-Versicolor, Blue = Iris-Virginica)", fontsize=24, y = 0.95)

plt.subplot(2, 3, 1)
plt.scatter(sepal_length, sepal_width , c = iris_class_colors)
plt.xlabel("Sepal Length in cm", fontsize=12)
plt.ylabel("Sepal Width in cm", fontsize=12)
```

Listing 4: Python Listing 4

Listing 4 shows the general code we wrote to build each figure and each subplot. The source code for the rest of the plots/figures can be found in the attached Jupyter notebook.

The two figures containing the 12 plots we created can be found in **Figure 1** and **Figure 2** below:

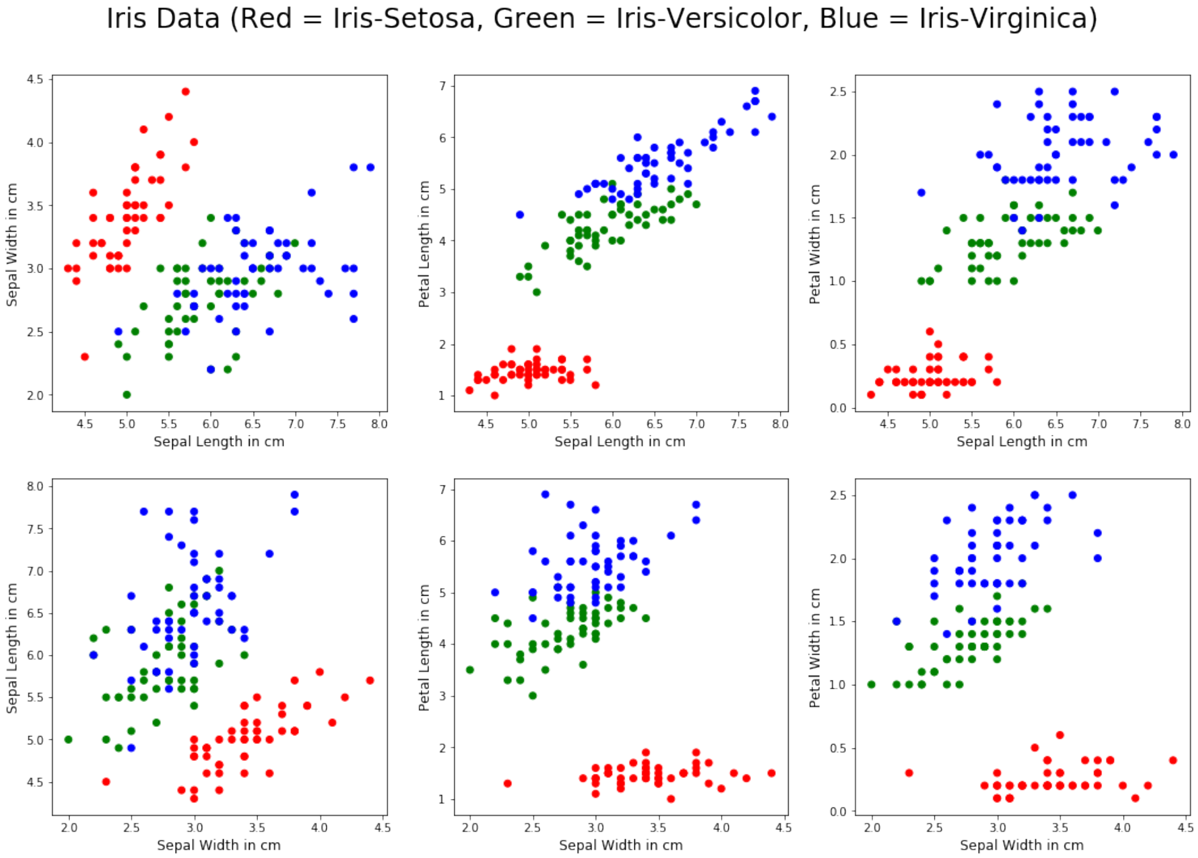

Figure 1: 6 of the 2-attribute combinations where Sepal Length is on the x-axis for the first row and Sepal Width is on the x-axis for the second row

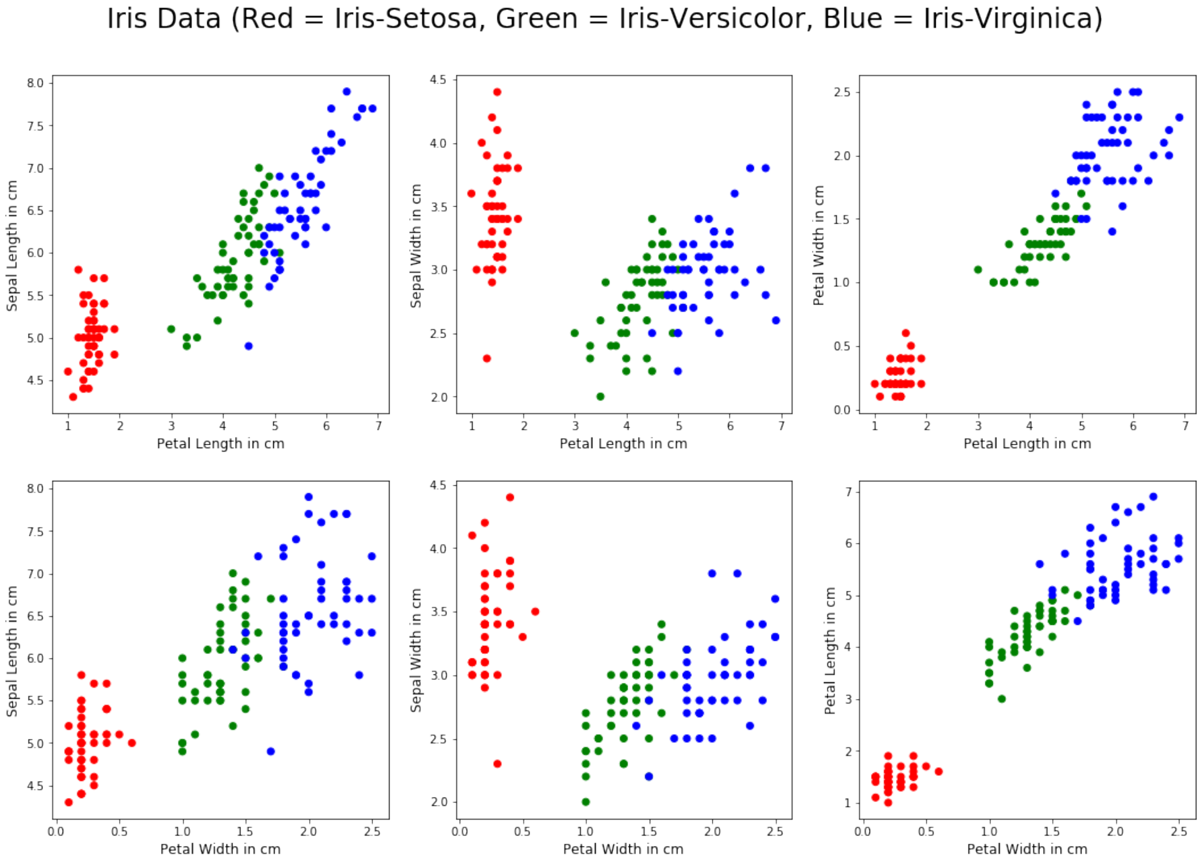

Figure 2: 6 of the 2-attribute combinations where Petal Length is on the x-axis for the first row and Petal Width is on the x-axis for the second row