Kushal Singh

CS 5875 - Applied Machine Learning

Homework 0

Team Member: Andrew Palmer

5 September 2019

Iris Flowers

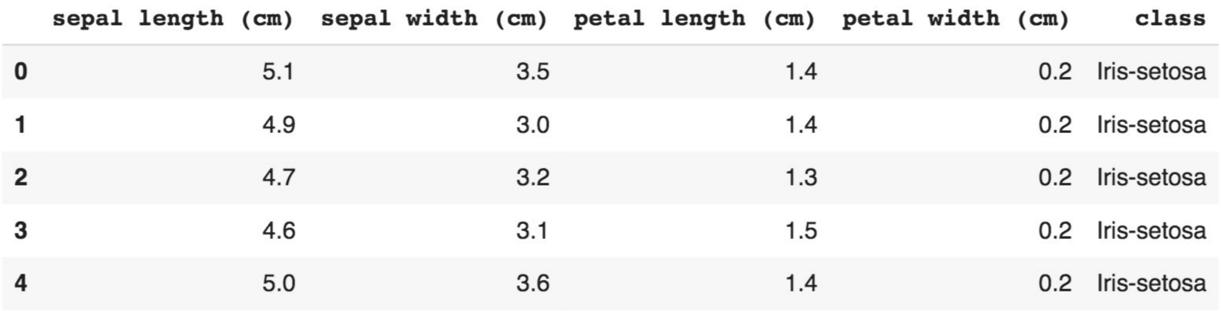
1.1 Downloading the dataset

In order to download the dataset, we called the **pd.read_csv**[1] function on the iris dataset URL, which contains CSV data on the iris plants database.

```
url = 'http://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data'
df = pd.read_csv(url, header=None)
df.columns = ['sepal length (cm)', 'sepal width (cm)', 'petal length (cm)', 'petal width (cm)',
'class']
```

Taking a look at the iris.names file, we can see that there are 150 total samples/instances, with 5 features/attributes for each sample—sepal length (cm), sepal width (cm), petal length (cm), petal width (cm), species. We also observe that there are 3 different species types (*iris setosa*, *iris versicolour*, *iris virginica*), and 50 samples of each species.

1.2 Parsing the dataset

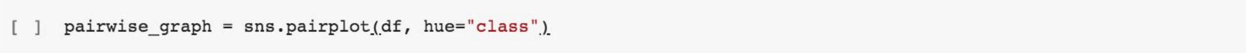


	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	class
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa

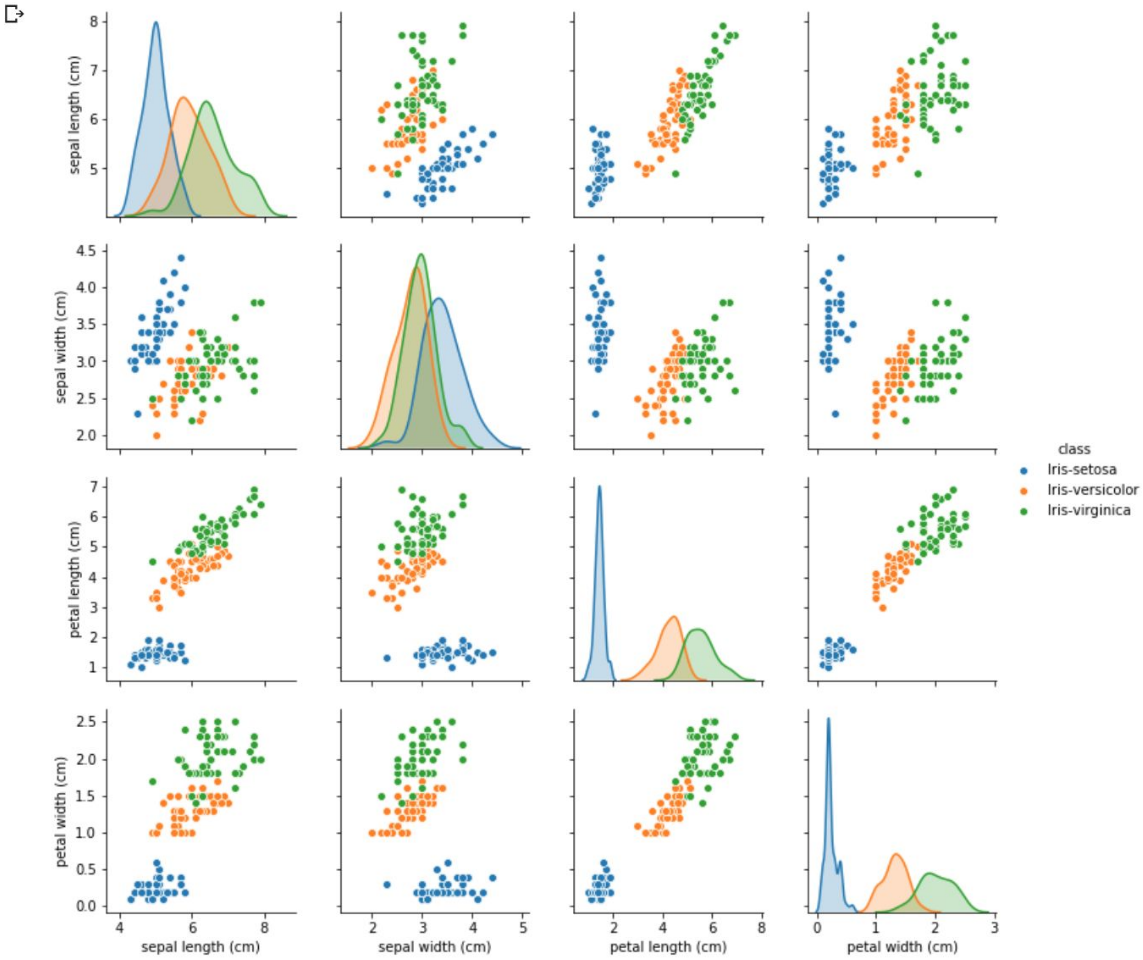
The above screenshot illustrates a sample of the N * p dimensional data frame.
In the code, the line label_vec = df[‘class’].to_numpy()[2], creates the *N*-dimensional array of the class type.

[1] https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.read_csv.html
[2] https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.to_numpy.html

1.3 Visualizing the dataset



```
[ ]  pairwise_graph = sns.pairplot(df, hue="class")
```



From the above scatterplots, generated using the seaborn library[3], we can see that petal length and petal width provide the most information about the class/species type. We can also verify this with the high class correlation coefficients found in iris.names (0.9490 for petal length, 0.9565 for petal width).

[3] http://seaborn.pydata.org/examples/scatterplot_matrix.html