Project 52167 - Iris Dataset Invesitgation

David Crowley

April 28, 2018

1 Introduction

The Iris data set was originally created in 1935 by the American botanist Edgar Anderson who examined the geographic distribution of Iris flowers on the Gaspé peninsula in Quebec (Canada) [1]. Fisher [2] used Anderson's Iris data set for multivariate discriminant analysis. Discriminant analysis is a form of classification problem, where two or more groups or clusters or populations are known and one or more new observations are classified into one of the known populations based on the measured characteristics [3]. The data features from [4] available to download at [5] in the data set are as follows

1. sepal length in cm.
2. sepal width in cm
3. petal length in cm
4. petal width in cm
5. Class:
    - Iris Setosa
    - Iris Versicolour
    - Iris Virginica

2 Data set Summary

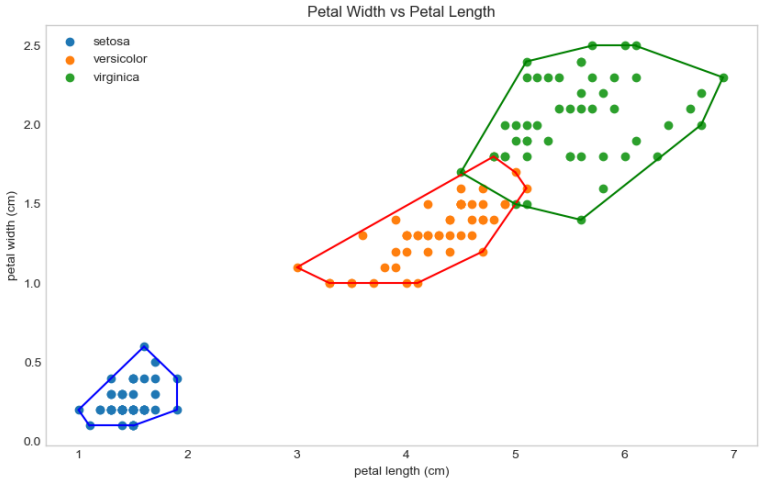
The data set has 150 instances and has no missing values. The data is multivariate which means it involves two or more variable quantities. These quantities are described above with sepal length, sepal width, petal length, petal width which are described as attributes in [5]. The last attribute is the class of Iris plant and this is the predicted attribute. [5] state that the data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant.

One class is linearly separable from the other 2; the latter are not linearly separable from each other. Linearly separable is if two sets S1 and S2 (classes in

Listing 1: Sample Output 1

```
Sepal Width - Descriptive Stats
Mean from stats module: 3.054
Median: 3.0
Mode: 3.0
Standard Deviation: 0.4321465800705435
Max: 4.4
Min: 2.0
******************
Setosa Sepal Width - Descriptive Stats
Mean from stats module: 3.418
Median: 3.4
Mode: 3.4
Standard Deviation: 0.37719490982779713
Max: 4.4
Min: 2.3
******************
Versicolor Sepal Width - Descriptive Stats
Mean from stats module: 2.77
Median: 2.8
Mode: 3.0
Standard Deviation: 0.31064449134018135
Max: 3.4
Min: 2.0
******************
Virginica Sepal Width - Descriptive Stats
Mean from stats module: 2.974
Median: 3.0
Mode: 3.0
Standard Deviation: 0.3192553836664309
Max: 3.8
Min: 2.2
```

Figure 1: Testing Linear Separability - Petal



this data set) are linear separable if there exists at least one line (in Euclidean space) with all of one set S1 on one side of the line and all of the other set S2 on the other side of the line. This can extended beyond two dimensions by replacing the line with a hyerplane [6]. Figure 1 and Figure 2 display that the Iris Setosa class is linearly separable from the other two classes. The other two classes Iris Virginica and Iris Versicolour are shown to be not linearly separable (they overlap - intersect) in both sepal width/length and petal width/length.

Sample output of running the code available at [7] is shown in Sample Output 1 using Sepal Width (each run different calls to functions were commented out - preferably each time the code was run a file with all the data could be created and stored and charts created). Some entries have n/a as there was no unique mode for the given feature.

Figure 3 explores four different dimensions of the data: the (x, y) location of each point corresponds to the sepal length and width, the size of the point is related to the petal width, and the colour is related to the particular species of flower [8]. Table 1 shows the descriptive stats for Sepal Width. Table 2 displays the descriptive stats for Sepal length. Table 3 shows the descriptive stats for Petal width. Table 4 shows the descriptive stats for Petal length.

3 Findings

From the beginning the project was planned to be in two components or aims, one was an exploration of using Python to write a program to examine the descriptive statistics generally used in basic analysis (mean, mode, median, stan-

Table 1: Sepal Width - three classes and individually

Description	All 3	Setosa	Versicolor	Virginica
Mean	3.054	3.418	2.77	2.974
Median	3.0	3.4	2.8	3.0
Mode	3	3.4	3.0	3.0
Standard Deviation	0.432	0.377	0.311	0.319
Max	4.4	4.4	3.4	3.8
Min	2.0	2.3	2.0	2.2

Table 2: Sepal Length - three classes and individually

Description	All 3	Setosa	Versicolor	Virginica
Mean	5.843	5.006	5.936	6.588
Median	5.8	5.0	5.9	6.5
Mode	5.0	n/a	n/a	6.3
Standard Deviation	0.825	0.349	0.511	0.629
Max	7.9	5.8	7.0	7.9
Min	4.3	4.3	4.9	4.9

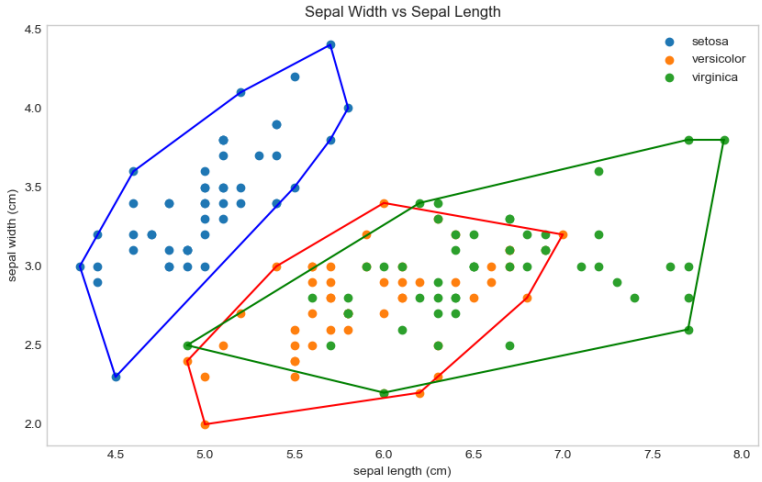
Table 3: Petal Width - three classes and individually

Description	All 3	Setosa	Versicolor	Virginica
Mean	1.197	0.244	1.326	2.026
Median	1.3	0.2	1.3	2.0
Mode	0.2	0.2	1.3	1.8
Standard Deviation	0.761	0.106	0.196	0.272
Max	2.5	0.6	1.8	2.5
Min	0.1	0.1	1	1.4

Table 4: Petal Length - three classes and individually

Description	All 3	Setosa	Versicolor	Virginica
Mean	3.7586	1.464	4.26	5.552
Median	4.35	1.5	4.35	5.55
Mode	1.5	1.5	4.5	5.1
Standard Deviation	1.759	0.172	0.465	0.546
Max	6.9	1.9	5.1	6.9
Min	1.0	1.0	3.0	4.5

Figure 2: Testing Linear Separability - Sepal



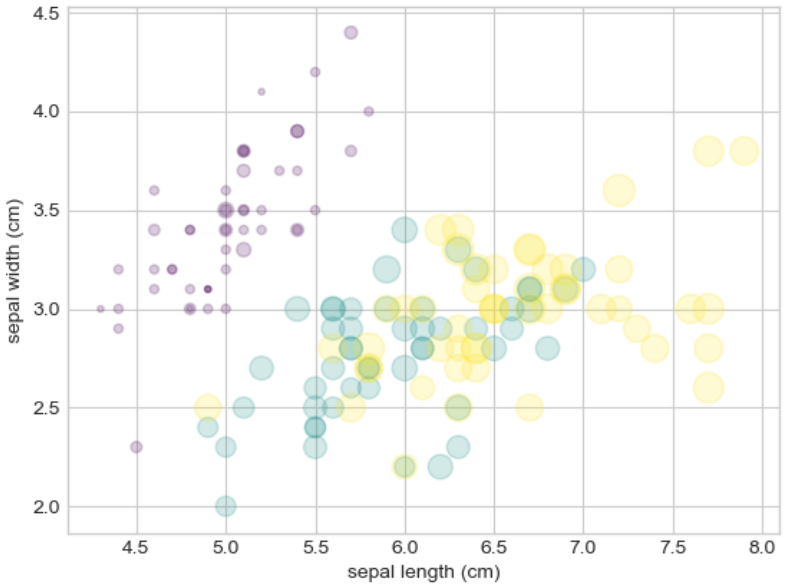
dard deviation, and general descriptive measures like maximum and minimum. After initial coding to create a simple mean calculation and an exploration of writing code for a median and mode calculation, the Statistics module of Python was discovered [9]. This allowed a very quick implementation of mean, mode, median and standard deviation. This also led to findings that if the mode was not unique then an error was thrown, so code was written (a simple if statement) to work around this.

The second aim of the project was to examine the more data science modules in Python. Modules including Matplotlib [10],NumPy [11] and pandas [12] for data analysis and visualisation. These packages were used to create the charts shown in this document and to examine linear separability. More work needs to be undertaken to explore these modules as their scope and features are daunting and powerful compared to "hand-coding" simple calcuations.

4 Future Work

- Build more robust error handling
- Build on knowledge of numpy, pandas and matplotlib packages
- Explore the data more - look at classification approaches to define the classes
- Examine more visualisation methods. Histograms were developed as part of the project (code is available in Github repository but seemed very poor

Figure 3: Iris Data Visualisation



in showing anything of value

References

[1] E. Anderson, "The irises of the gaspe peninsula," *Bulletin of the American Iris Society*, vol. 59, pp. 2–5, 1935.

[2] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of human genetics*, vol. 7, no. 2, pp. 179–188, 1936.

[3] "Lesson 10: Discriminant analysis." [Online]. Available: https://onlinecourses.science.psu.edu/stat505/node/89

[4] D. Dheeru and E. Karra Taniskidou, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[5] "UCI machine learning repository - iris data set," 2018. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/iris

[6] "Linear separability," Mar 2018. [Online]. Available: https://en.wikipedia.org/wiki/Linear_separability

[7] D. Crowley, "GMIT - 52167 Project Repository," 2018. [Online]. Available: https://github.com/kingcrowley/gmit_52167_project

[8] J. VanderPlas, "Python data science handbook - github," 2018. [Online]. Available: https://github.com/jakevdp/PythonDataScienceHandbook/blob/master/notebooks/04.02-Simple-Scatter-Plots.ipynb

[9] "Python online documentation - statistics - mathematical statistics functions," 2018. [Online]. Available: https://docs.python.org/3/library/statistics.html

[10] "Matplotlib homepage," 2018. [Online]. Available: https://matplotlib.org/

[11] "Numpy homepage," 2018. [Online]. Available: http://www.numpy.org/

[12] "pandas homepage," 2018. [Online]. Available: https://pandas.pydata.org/