Project 52167 - Iris Dataset Invesitgation

David Crowley

April 25, 2018

1 Introduction

Code available at [1]

2 Data set Summary

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) [2]. Fisher [3] 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 [4]. The data features from [5] available to download at [6] 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

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 [6]. The last attribute is the class of Iris plant and this is the predicted attribute. [6] state that the data set contains 3 classes of 50 instances each, where each class refers to a type of iris

Figure 1: Testing Linear Separability - Petal

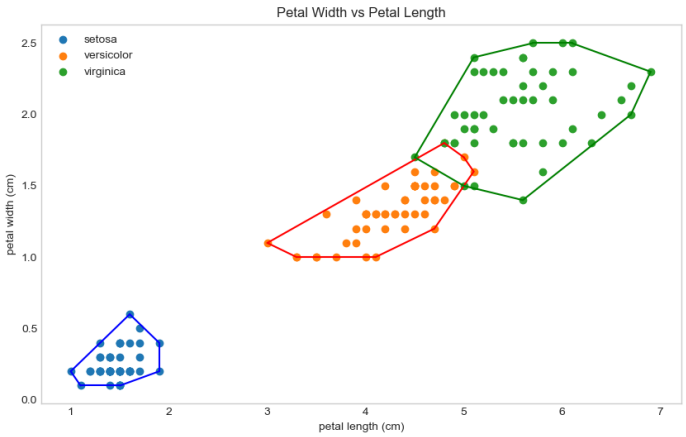


Table 1: Sepal Length - three classes and individually

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

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 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 [7].Figure 1 displays 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).

Setosa Sepal Length - Descriptive Stats Mean from stats module: 5.006 Median: 5.0 Standard Deviation: 0.3489469873777391 Max: 5.8 Min: 4.3

3 Findings

[8]

Table 2: Sepal Length - Descriptive Stats for all 3 classes for Sepal Width

Sepal Length - Descriptive Stats	
Mean	5.843333333333334
Median	5.8
Mode	5.0
Standard Deviation	0.8253012917851409
Max	7.9
Min	4.3

Table 3: Sepal Width - Descriptive Stats for all 3 classes

Sepal Width - Descriptive Stats	
Mean	3.054
Median	3.0
Mode	3.0
Standard Deviation	0.4321465800705435
Max	4.4
Min	2.0

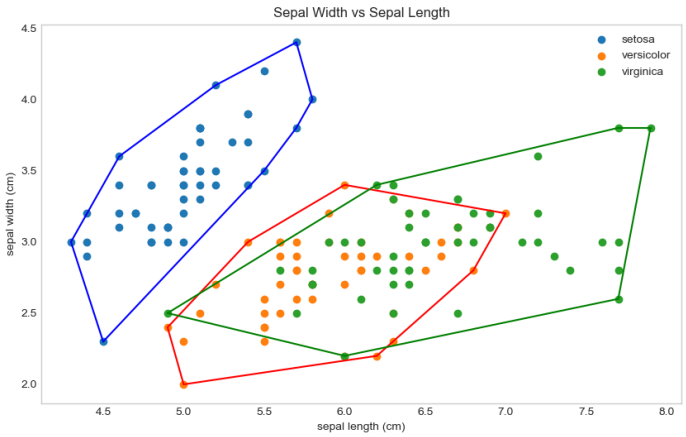
Table 4: Petal Length - Descriptive Stats for all 3 classes

Petal Length - Descriptive Stats	
Mean	3.7586666666666666
Median	4.35
Mode	1.5
Standard Deviation	1.7585291834055212
Max	6.9
Min	1.0

Table 5: Petal Width - Descriptive Stats for all 3 classes

Petal Width - Descriptive Stats	
Mean	1.1986666666666668
Median	1.3
Mode	0.2
Standard Deviation	0.7606126185881718
Max	2.5
Min	0.1

Figure 2: Testing Linear Separability - Sepal



4 Conclusion

Horse

References

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

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

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

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

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

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

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

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