Relationship between different types of iris features and iris species

Author A(a, b)

Contents

Institutions:

a) Department. Center. Country;
b) Research Institute. Country;

Disclosures: Authors declare that there are the following conflicts of interest related to the results of this article.

Funding: This study had the following sources of funding.

Correspondence:

- Author A. Address 1. E-mail address 1. Telephone 1.

ABSTRACT

Background and Objectives:

The *iris* dataset is a well known collection of flower features classified in three species: setosa, versicolor and virginica. Based on Fisher's linear discriminant model, this data set became a typical test case for many statistical classification techniques in machine learning such as support vector machines. The objective os

this study is to describe the iris dataset as an example of semi-automated document creation using R and Rmarkdown.

Methods:

The dataset contains a set of 150 records under five attributes - petal length, petal width, sepal length, sepal width and species.

Results:

The number of samples is 150, being divided in 50 for each species: *setosa*, *versicolor* and *virginica*. The mean sepal length for setosa, versicolor and virginica is of 5.006, 5.936, 6.588 respectively. The mean petal length for setosa, versicolor and virginica is of 1.462, 4.26, 5.552 respectively.

Conclusion:

The iris dataset is adequate for teaching a variety of statistical classification methods, and can be used as a tool to learn to make scientific articles in Rmarkdown.

Keywords: Iris, Rmarkdown.

Abbreviations:

PL: petal length

PW: petal width

SL: sepal length

SW: sepal width

INTRODUCTION

The Iris flower data set or Fisher's Iris data set is a multivariable data set introduced by the British statistician and biologist Ronald Fisher in his 1936 paper *The use of multiple measurements in taxonomic problems* as an example of linear discriminant analysis[1]. It is sometimes called *Anderson's Iris data set* because Edgar Anderson collected the data to quantify the morphologic variation of Iris flowers of three related species. Based on Fisher's linear discriminant model, this data set became a typical test case for many statistical classification techniques in machine learning such as support vector machines.

METHODS

Data collection

The dataset contains a set of 150 records under five attributes - petal length, petal width, sepal length, sepal width and species. Two of the three species were collected in the Gaspé Peninsula "all from the same pasture, and picked on the same day and measured at the same time by the same person with the same apparatus".

Statistical analysis

The iris data set is widely used as a beginner's dataset for teaching classification methods. All analyses were performed using R v.3.4 (R Core Team, Vienna, Austria) with the 'ggplot2'[2] and 'dplyr'[3] packages.

RESULTS

The number of samples is 150, being divided in 50 for each species: *setosa*, *versicolor* and *virginica*. The mean sepal length for setosa, versicolor and virginica is of 5.006, 5.936, 6.588 respectively. The mean petal length for setosa, versicolor and virginica is of 1.462, 4.26, 5.552 respectively (**Table 1**). Medians are represented in **Table 2**. The relationship between sepal length and width and petal length is depicted in **Figure 1**. The cluster centers are represented in **Figure 2**.

DISCUSSION

Based on Fisher's linear discriminant model, this data set became a typical test case for many statistical classification techniques in machine learning such as support vector machines[4,5].

The use of this data set in cluster analysis however is not common, since the data set only contains two clusters with rather obvious separation. One of the clusters contains *Iris setosa*, while the other cluster contains both *Iris virginica* and *Iris versicolor* and is not separable without the species information Fisher used. This makes the data set a good example to explain the difference between supervised and unsupervised techniques in data mining: Fisher's linear discriminant model can only be obtained when the object species are known: class labels and clusters are not necessarily the same.[4,6].

Clustering

Nevertheless, all three species of Iris are separable in the projection on the nonlinear branching principal component. The data set is approximated by the closest tree with some penalty for the excessive number of nodes, bending and stretching. Then the so-called "metro map" is constructed.[6] The data points are projected into the closest node. For each node the pie diagram of the projected points is prepared. The area of the pie is proportional to the number of the projected points. It is clear from the diagram (left) that the absolute majority of the samples of the different Iris species belong to the different nodes. Only a small fraction of Iris-virginica is mixed with Iris-versicolor (the mixed blue-green nodes in the diagram).

Therefore, the three species of Iris (Iris setosa, Iris virginica and Iris versicolor) are separable by the unsupervising procedures of nonlinear principal component analysis. To discriminate them, it is sufficient just to select the corresponding nodes on the principal tree.

Limitations

The dataset offers a detailed description with high quality data; however, the low number of flower features difficults a more precise classification.

CONCLUSION

The iris dataset is adequate for teaching a variety of statistical classification methods, and can be used as a tool to learn to make scientific articles in Rmarkdown.

TABLES

Table 1: means of the sepal and petal measures by species.

Species	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
setosa	5	3.4	1.5	0.25
versicolor	5.9	2.8	4.3	1.3
virginica	6.6	3	5.6	2

Table 2: medians of the sepal and petal measures by species.

Species	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
setosa	5	3.4	1.5	0.2
versicolor	5.9	2.8	4.3	1.3
virginica	6.5	3	5.5	2

FIGURES

Figure 1: iris species classification.

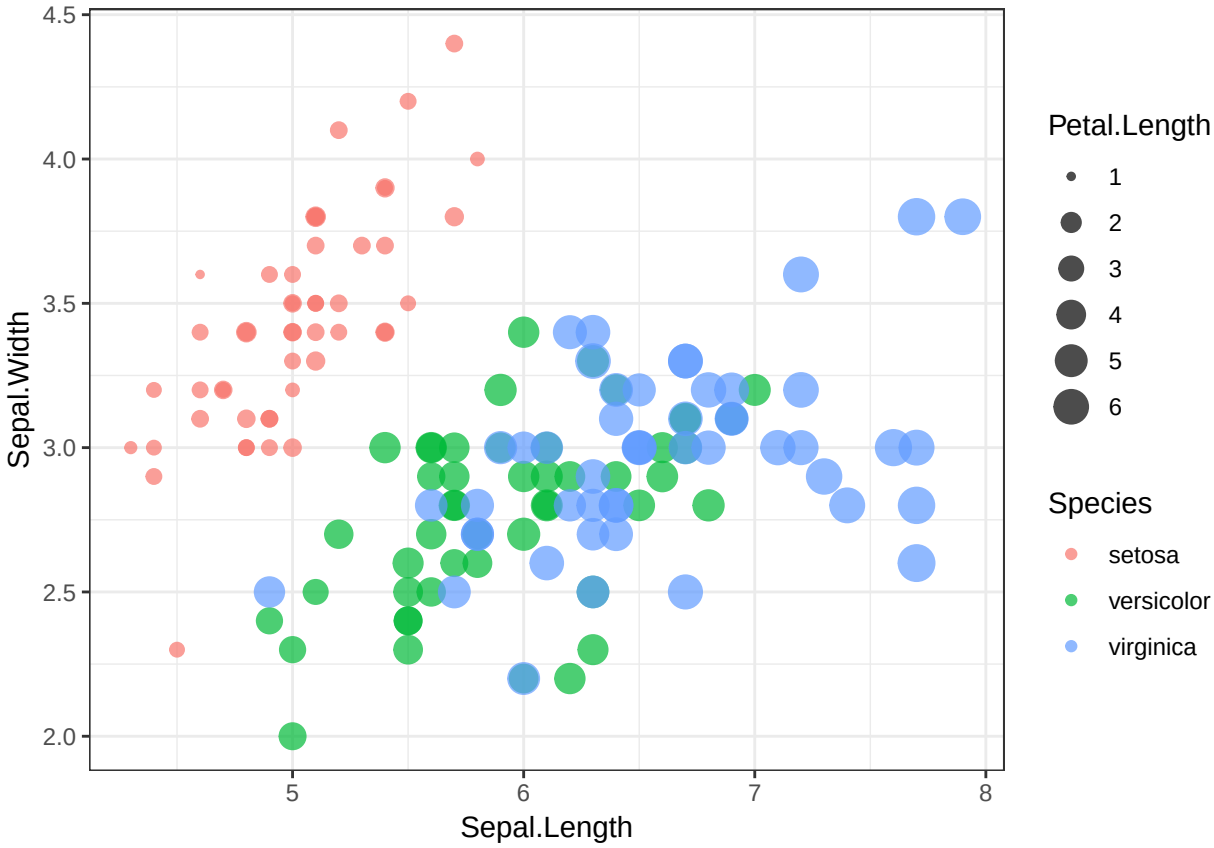
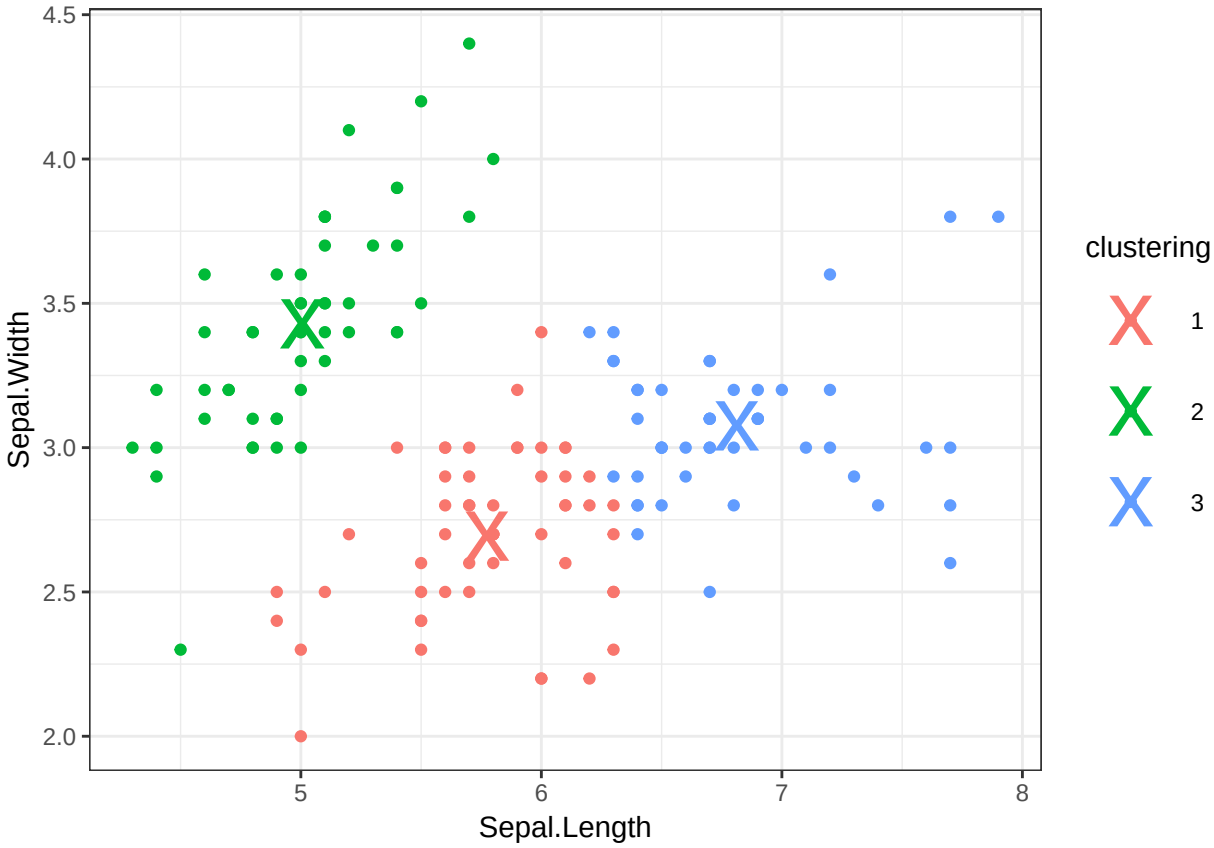


Figure 2: iris K-means classification.



REFERENCES

1. Fisher RA: The Use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics 1936; 7:179–88

2. Wickham H: Ggplot2: Elegant Graphics for Data Analysis, 2nd edition. Springer International Publishing, 2016 at <//www.springer.com/it/book/9783319242750>

3. Grolemund HW Garrett: R for Data Science. at <http://shop.oreilly.com/product/0636920034407.do>

4. *Iris* flower data set 2019 at <https://en.wikipedia.org/w/index.php?title=Iris_flower_data_set&oldid=906180779>

5. UCI Machine Learning Repository: Iris Data Set at <https://archive.ics.uci.edu/ml/datasets/iris>

6. Gorban AN, Sumner NR, Zinovyev AY: Topological grammars for data approximation. Applied Mathematics Letters 2007; 20:382–6