

Statistical Analysis of the Iris Flower Dataset

Patrick S. Hoey
University of Massachusetts At Lowell

Abstract

The Iris Flower Dataset is a popular multivariate dataset that was introduced by R.A. Fisher as an example for discriminant analysis. The Iris dataset has been analyzed via two distinct methods. First, plotted the dataset onto scatterplots to determine patterns in the data in relation to the Iris classifications. Second, developed an application in Java that will run a series of methods on the dataset to extract relevant statistical information from the dataset. With these two methods, I can make concrete predictors about the dataset.

1 Introduction

The Iris Flower Dataset is a popular multivariate dataset that was introduced by R.A. Fisher as an example for discriminant analysis. The data reports on four characteristics of the three species of the Iris Flower, sepal length, sepal width, petal length, and petal width. The goal of a discriminant analysis is to produce a simple function that, given the four measurements, will classify a flower correctly. This is the beginning of creating “predictors” in order to try to make a more educated guess on a record in a dataset. This article will attempt to analyze this dataset to try to draw some conclusions from the model.

2 Graphical Plots

2.1 Overview of Patterns in Dataset

When comparing variables in a multidimensional or multivariate dataset, some conclusions must be drawn from the patterns in the dataset. In comparing different variables it is good practice to first make an educated assumption on what type of patterns you wish to find. An example is with a dataset involving the sleeping habits, sleep duration etc. amongst all the animals known. If one was to run an analysis on the dataset, comparing sleeping duration to sleep time, the plots could get very confusing with all the animals partitioned into their individual bins. This type of granularity in dividing the data

could make pattern matching very difficult. The easier technique in order to be able to derive common patterns across the entire animal kingdom would be to partition the animals into bins according to classification, such as Reptiles, Mammals and Birds. We want to aggregate these animals into their classifications in order to pull out patterns.

2.2 Classification of Iris Dataset

In looking at the Iris Dataset, the patterns that I wanted to draw from the dataset are related to how the three types of classes of Iris differ. I wanted to see how the classes of Iris-Setosa, Iris-Versicolor and Iris-Virginica related to each other when compared with their common dimensions of Sepal Length, Sepal Width, Petal Length and Petal Width. On an Iris, the sepal is larger, lower petal and the petal is the upper petal. I will try to draw patterns out of this Iris dataset. In analyzing a multivariate dataset, there is a simple formula to evaluate the total permutations necessary to compare variables:

$$\frac{n(n-1)}{Di}$$

Where n is the total number of variables or dimensions in the dataset and Di is the total number of variables or dimensions you wish to compare in a graph. With the Iris dataset, there are five (5) total variables. We want to compare only two (2) variables at one time, so this would give us a total of ten (10) possible graphs. Since we want to compare the classifications directly and not compare say Sepal to Petal, we can remove six (6) of the possible graphs. In total, there should be four (4) relevant graphs from this dataset.

2.3 Graphs of Iris Dataset

The first scatterplot graph (Fig. 1) compares the sepal length with the classification of flower. The second scatterplot graph (Fig. 2) compares the sepal width with the classification of flower. The third scatterplot graph (Fig. 3) compares the petal length with the classification of flower. The

fourth scatterplot graph (Fig. 4) compares the petal width with the classification of flower. From these four scatterplots we can determine a pattern and therefore create a possible predictor.

3 Statistical Analysis

3.1 Statistical Analysis of Scatterplots

Lets first examine Figure 1 that compares the Sepal length to the classification. In looking at the clustering of points, there is overlapping for Sepal length across all classifications for the Iris. In comparing the classifications, some general observations can be made, such as 1) The Iris-Virginica has the longest sepals compared to the other classes, 2) The Iris-Versicolor has the medium length for sepals compared to the other classes and 3) The Iris-Setosa has the shortest sepals compared to the other classes.

In Figure 2, the sepal width is compared to the classification. Conclusions drawn from the scatterplot, 1) The Iris-Setosa has the widest sepals compared to the other classes, 2) The Iris-Virginica has the medium width in comparison to the other classes and 3) The Iris-Versicolor has the narrowest of Sepals compared to the other classes.

In Figure 3, the petal length is compared to the classification. The points are partitioned pretty cleanly with little overlap. This is an important distinction for these classifications. Conclusions drawn from the scatterplot, 1) The Iris-Virginica has the longest petals compared to the other classes, 2) The Iris-Versicolor has the medium length in comparison with the other classes and 3) The Iris-Setosa has the shortest petals compared to the other classes.

In Figure 4, we see another distinction in petal width amongst the different classes. Conclusions drawn from the scatterplot, 1) The Iris-Virginica has the widest petals when compared to the other classes, 2) The Iris-Versicolor has the medium width of petals in comparison to the other classes and 3) The Iris-Setosa has the narrowest petals when compared to the other classes.

In summary of these results, I can build an objective predictor for a particular classification of Iris flowers. If the Iris flower has a long sepal (6-8cm), long petals (5-7cm) and wide petals (1.5-2.5cm) then the Iris is most likely an Iris-

Virginica. If the Iris flower has a short sepal (4.5-5.5cm), short petals(1-2cm) and very narrow petals (.1-.5cm) then the Iris is most likely an Iris-Setosa. Any Iris flower that falls in between these two classifications is most likely an Iris-Versicolor.

3.2 Statistical Analysis of Variables

In analyzing the data for the Iris dataset, along with the scatterplots, I wanted to develop an application to parse the dataset and try to extract out relevant analysis from the data. I chose to write the application in Java since the clean design and extensive libraries with the language makes rapid implementation possible. The approach was to build a framework interface in order to implement similar functionality on any dataset that is comma delimited like the Iris dataset. I then implemented specific functions based on these abstract functions for the dataset I wanted to analyze. Currently though, I only created general functions that take vectors and analyze the data within the vectors. The next step will be to create an interface with these general functions. When these functions are applied to the data parsed from the dataset, the output is the following for the Iris dataset:

Total Records: 150
Total Dimensions: 5
The Sepal Length Mean: 5.8433337
The Sepal Length Mode: 5.0
The Sepal Length Median: 5.8
The Sepal Length Range: 3.5
The Sepal Length Variance: 0.65975
The Sepal Length Standard Deviation: 0.81224996
The Sepal Length Minimum Value: 4.4
The Sepal Length Maximum Value: 7.9
The Sepal Width Mean: 3.0540001
The Sepal Width Mode: 3.0
The Sepal Width Median: 3.0
The Sepal Width Range: 2.2
The Sepal Width Variance: 0.1766741
The Sepal Width Standard Deviation: 0.42032617
The Sepal Width Minimum Value: 2.2
The Sepal Width Maximum Value: 4.4
The Petal Length Mean: 3.7586663
The Petal Length Mode: 1.5
The Petal Length Median: 4.4
The Petal Length Range: 5.8
The Petal Length Variance: 3.0336936
The Petal Length Standard Deviation: 1.7417501
The Petal Length Minimum Value: 1.2

The Petal Length Maximum Value: 6.9
The Petal Width Mean: 1.1986667
The Petal Width Mode: 0.2
The Petal Width Median: 1.3
The Petal Width Range: 2.4
The Petal Width Variance: 0.56981766
The Petal Width Standard Deviation:
0.75486267
The Petal Width Minimum Value: 0.1
The Petal Width Maximum Value: 2.5

The functions find the mean, mode, median, range, variance, standard deviation, minimum value and maximum values.

From this analysis, we can determine statistical properties contained within a dataset. For instance, in looking at the Petal Length data, we can see that for a mean value of 3.76, the variance (spread of the distribution) is large at 3.0. That would explain why on the graph (Fig 3) the clusters are very defined with little overlap between the three classes. Another example is with the Sepal Width. With a mean value of 3.0, the variance is very small at .18. You would expect the clusters of data points to be centered around the mean, and in looking at Figure 2, you can see that the points overlap each other considerably especially around the mean.

4 Summary

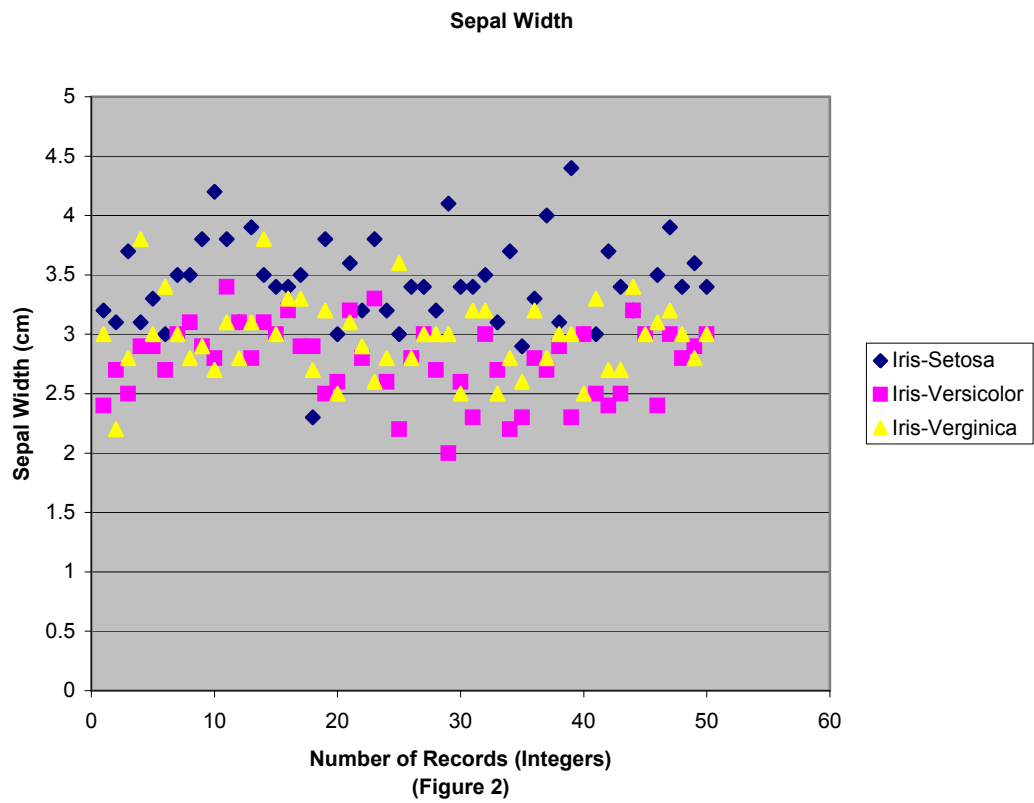
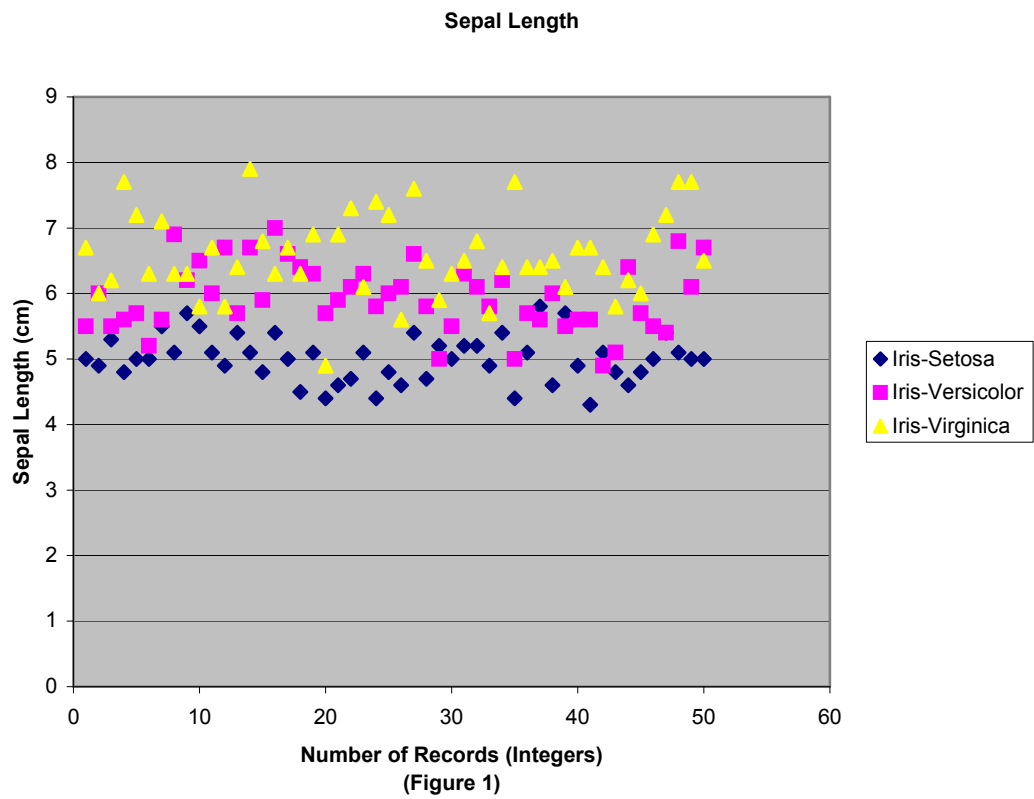
In evaluating the Iris dataset, I found it important to be able to reference graphs of the data with the statistical analysis run from the application. In the future, it would be a nice feature to be able to add visualization with the analysis, instead of static graphs generated in Excel. Along with the framework mentioned earlier in the article, having the data stored as XML and using VTK to render the data would be a nice step in creating interactive visualizations of the data. This would be the next step on the journey to create an interface for interactive visualizations of data.

References

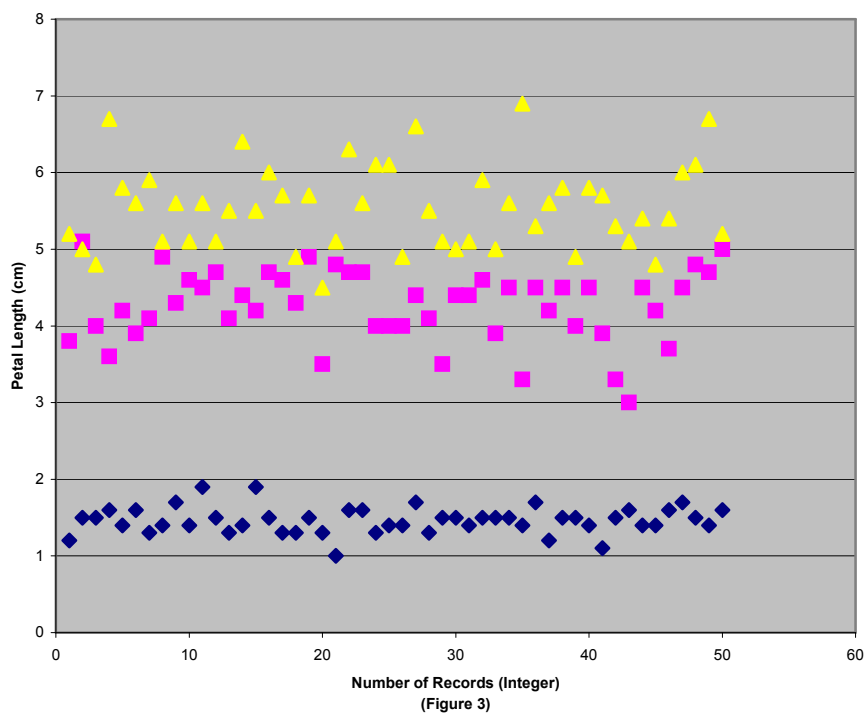
[1] Card, S., Mackinlay, J., and Shneiderman, B. Information Visualization. Readings in Information Visualization: Using Vision to Think, pp.1-34; 1999, Morgan Kaufmann Publishers, Inc., USA.

[2] "Fisher's Irises Story".
<http://lib.stat.cmu.edu/DASL/Stories/Fisher'sIrises.html>

[3] "Variance"
<http://www.ruf.rice.edu/~lane/hyperstat/A16252.html>



Petal Length



Petal Width

