F1 score for iris flower datasets.

Ravi R. Gorasiya
AITS

***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.**

I. 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.

II. Graphical representation.

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.

III. Classify dataset.

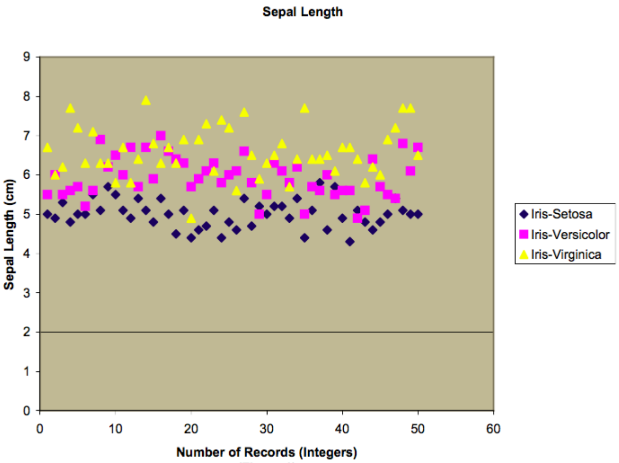
In looking at the Iris Dataset, the patterns that I wanted to draw from the dataset are related to how the three types of classes of Iris differ. I wanted to see how the classes of Iris-Setosa, Iris- Versicolor and Iris-Virginica related to each other when compared with their commons dimensions of Sepal 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

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 four(4) total variables. We want to compare only two (2) variables at one time, so this would give us a total of ten (8) possible graphs. Since we want to compare the classifications directly and not compare say Sepal to Petal, we can remove four (4) of the possible graphs. In total, there should be four (4) relevant graphs from this dataset.

here, only conclude this four most relevent graph for classification iris flower problem

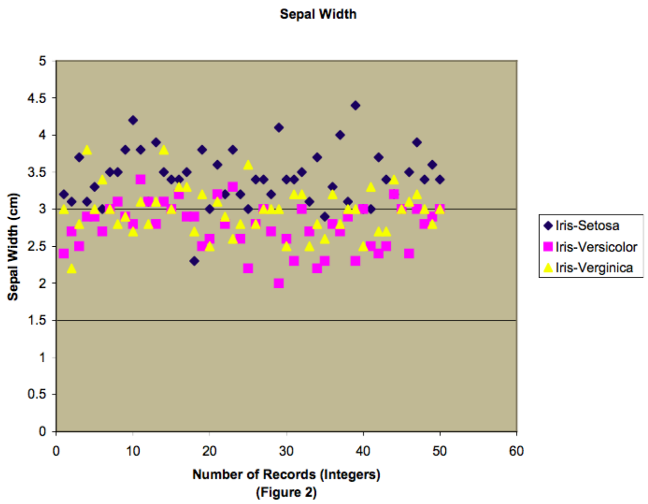
IV. Define Graph

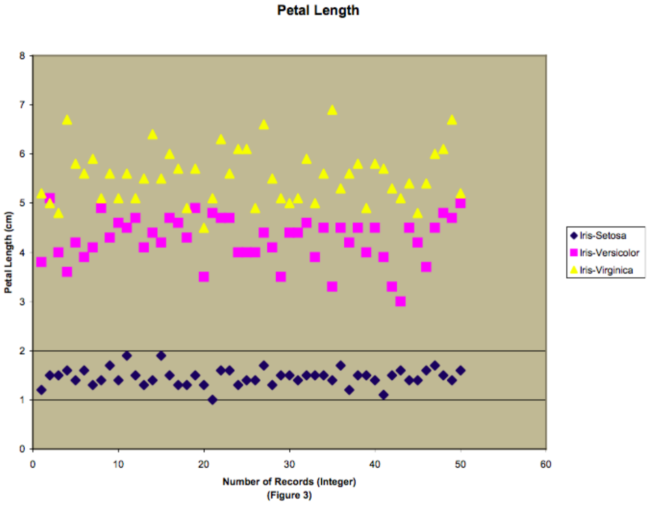
here, define most understandable graph

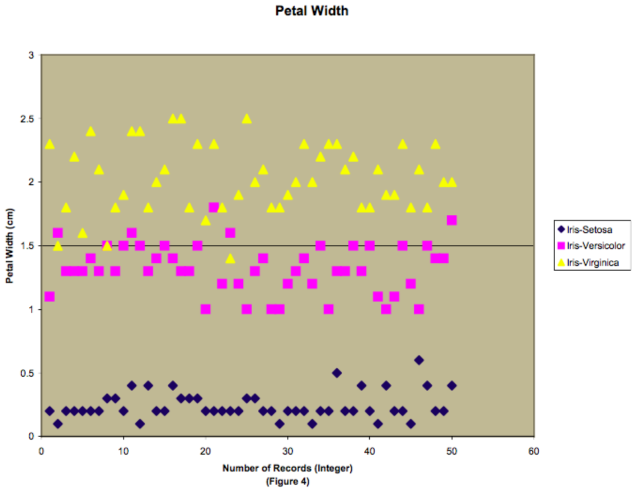


(Figure 1)

Ravi R. Gorasiya (e-mail: admin@earn-online.tech)







V. F1-Score analysis.

for finding f1-score of predicted data first we declared metrix that's define below.

First we difine some terms.

T.P.=true positive.

F.P=false positive.

T.N=true negetive.

F.N=false negetive

F1 score= 2*((precision*recall)/(precision+recall))

		Predicted	
		Negative	Positive
Actual	Negative	True Negative	False Positive
	Positive	False Negative	True Positive

using this matrix first we define **Precision** and **Recall.**

Appendix A
Precision.

Precision= T.P/(T.P + F.P)

		Predicted	
		Negative	Positive
Actual	Negative	True Negative	False Positive
	Positive	False Negative	True Positive

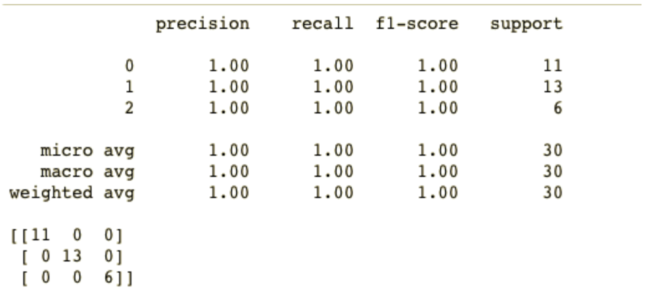
Appendix B
Recall

Recall=T.P/(T.P + F.N)

In iris flower classification problem real implementation and find F1 score.using sklearn library.

```
#convert into numpy array for sklearn
    classification report to get f1
    score
y_test_class = np.argmax(y_test,axis=1)
y_pred_class = np.argmax(y_pred,axis=1)
    #Accuracy of the predicted
    valuesfrom sklearn.metrics import
    classification_report,
    confusion_matrix
print(classification_report(
    y_test_class,y_pred_class))
print(confusion_matrix(y_test_class,
    y_pred_class))
```

Code Snippet 1.



```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        11
           1       1.00      1.00      1.00        13
           2       1.00      1.00      1.00         6

   micro avg       1.00      1.00      1.00        30
   macro avg       1.00      1.00      1.00        30
weighted avg       1.00      1.00      1.00        30

[[11  0  0]
 [ 0 13  0]
 [ 0  0  6]]
```

Reference

[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/Fis her’sIrises.html

[3] “Variance” http://www.ruf.rice.edu/~lane/hyperstat/A16252.html

[4]"f1 score" https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9