FDS Assignment 2 Report
Working on Fisher-Iris dataset

Jasnoor Singh
2018eeb1154@iitrpr.ac.in

Indian Institute of Technology
Ropar, PB

Abstract

This report includes basic insights and representations of the Fisher-Iris dataset, also containing analysis of some classifiers namely: Gaussian Naive Bayes, Logistic Regression and K-Means. Classifier analysis is done by constructing confusion matrices as well as classification reports. All classifiers were analysed using 5-fold cross validation with shuffling enabled in order to get a good estimate of the effectiveness of the model.

1 Introduction

In this assignment, both supervised as well as unsupervised learning algorithms were applied on the dataset. For supervised algorithms, the dataset was divided into training and test set in the ratio 8:2. For unsupervised part(K-means), the model was trained for the entire dataset.

2 Introduction to dataset used

The Fisher-Iris dataset is a multivariate data set introduced by the British statistician and biologist Ronald Fisher in 1936. The Fischer-Iris dataset has 150 rows and 4 attributes namely: the length and the width of the sepals and petals, in centimeters. These were measured for three classes of flowers namely setosa(class: 0), versicolor(class: 1) and virginica(class: 2). Following are the visualizations of the distributions of Sepal Width, grouping Sepal length into 10 bins. Similarly the distributions of Petal Width and Petal length are also visualized using a scatter plot.

2.1 Distribution of Sepal Width and Sepal Length

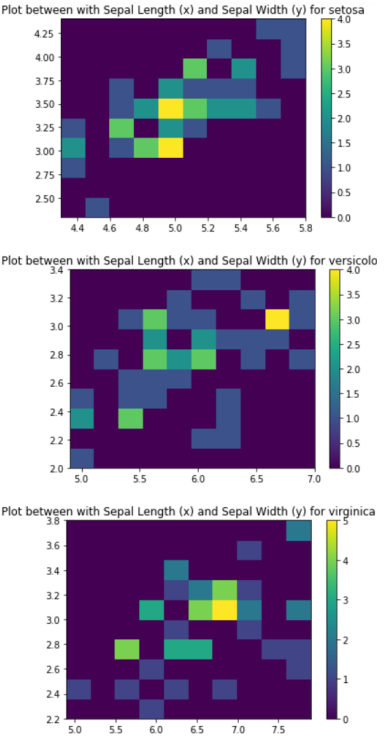

Figure 1: 2D Histogram of sepal, length vs width distributions.

Inferences:

- The correlation between Sepal length is much more defined for setosa as compared to versicolor or virginica.
- The correlation between Sepal lengths and widths is positive for all three types for flowers.

2.2 Distribution of Petal Width and Petal Length using Scatter Plot

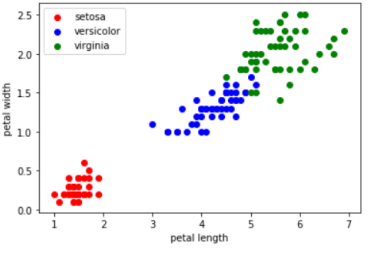

Figure 2: Scatter plot between with Petal Length (x) and Petal Width (y)

Inferences:

- Setosa has petal length between 1 an 2 cm, and petal width between 0 to 0.5 cm with no overlapping region with any other class.
- Versicolor has petal length between 3 and 5 cm, and petal width between 1 and 1.75 cm, which is intersecting with the virginica class.
- Virginica has petal length between 4.5 and 7 cm, and petal width between 1.25 and 2.5 cm, which is intersecting with the virginica class.
- The model might run into some problems if trying to predict between class 1 and 2 using petal parameters, whereas class 0 should be predicted accurately.

2.3 Boxplots

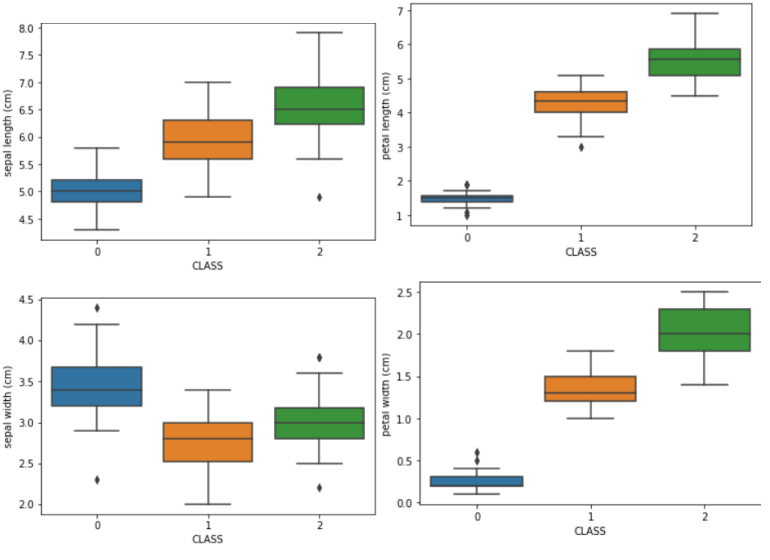

Figure 3: Box plots for Sepal length and width, Petal length and width

Inferences:

- The Petal lengths and widths from this fig. confirm our inferences from fig. 3.
- Class 1 and 2 are overlapping in every parameter, so there are bound to be some errors while classifying.
- It will be very easy to classify class 0.

3 Supervised Learning Classifiers

3.1 Gaussian Naive Bayes

```
[[10  0  0]
 [ 0  9  1]
 [ 0  1  9]]
              precision    recall  f1-score   support

      setosa       1.00      1.00      1.00        10
  versicolor       0.90      0.90      0.90        10
   virginica       0.90      0.90      0.90        10

    accuracy                           0.93        30
   macro avg       0.93      0.93      0.93        30
weighted avg       0.93      0.93      0.93        30
```

Figure 4: Confusion matrix and classification report for Gaussian Naive Bayes Model

Inference: As observed in the data visualization part, here also we observe that class 0 is classified accurately whereas there is some error while differentiating classes 1 and 2. Accuracy of this model is 93 percent.

3.2 Logistic Regression

```
[[10  0  0]
 [ 0 10  1]
 [ 0  0  9]]
              precision    recall  f1-score   support

      setosa       1.00      1.00      1.00        10
  versicolor       0.91      1.00      0.95        10
   virginica       1.00      0.90      0.95        10

    accuracy                           0.97        30
   macro avg       0.97      0.97      0.97        30
weighted avg       0.97      0.97      0.97        30
```

Figure 5: Confusion matrix and classification report for Logistic Regression Model

Inference: A high accuracy of 97 percent was achieved using the logistic regression, with perfect classification of class 0 but still some errors in the 1st and 2nd class due to overlap of parameters.
Still, logistic regression was able to classify most of the class 1 and 2 datapoints correctly.

3.3 KMeans

Inference: Kmeans gave a rather low accuracy with a lot of misclassifications of class 1 and 2. But it still managed to get complete accuracy in classifying the 0th class.
The 5-fold accuracy turmed out to be 89percent.

4 Learnings and outcomes

- We can conclude that for this application, Logistic Regression is the best way to classify the datapoints, as along with correctly classifying the class 0, it also managed to efficiently predict class 1 and 2.

```
[[50  0  0]
 [ 0 48 14]
 [ 0  2 36]]
              precision    recall  f1-score   support

      setosa       1.00      1.00      1.00        50
  versicolor       0.77      0.96      0.86        50
   virginica       0.95      0.72      0.82        50

    accuracy                           0.89       150
   macro avg       0.91      0.89      0.89       150
weighted avg       0.91      0.89      0.89       150
```

Figure 6: Confusion matrix and classification report for KMeans Model

- Kmeans classified this dataset quite poorly, not even being able to achieve an accuracy of 90 percent.
- We came to know about the possible reasons for misclassification in a classifier.(ex. parameter overlap).
- This exercise taught us about the importance of data visualization before applying any classifier.
- We also came to visually learn about importance of selecting only the relevant features for some of our applications.(eg.only Petal length was enough to classify class 0.)