Data Science
Assignment 2:Multi Class Classification

Adarsh Kumar
2018CSB1066
Dr.Ramanathan Subramanian
https://sites.google.com/site/raamsubram/

Indian Institute of Technology Ropar
Ropar,Punjab
India

Abstract

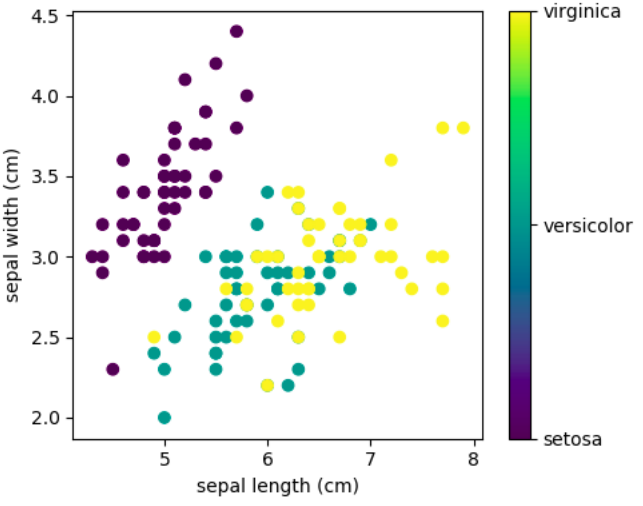
This document contains the results and inferences gathered after applying various classification algorithms like Logistic Regression,Gaussian Naive Bayes classifier and K-Means unsupervised learning algorithm to the Fischer Iris dataset,for multiclass classification of the three species of Iris.Five fold cross validation was employed to ensure that proper conclusions about model accuracies are drawn.

1 Introduction

We have used the Fischer Iris Dataset for multiclass classification using various learning and classification algorithms.We have explored the data set using scatter plots and histograms,and try to find correlation between the features of the dataset.We also made boxplots for each of the features of the dataset,to get an idea of how uniformly they were distributed and if there were many outliers/anomalies.We used Logistic Regression,Guassian Naive Bayes and K Means unsupervised learning algorithms for clssification of the dataset and noted their accuracy and errors.We have also tabulated the confusion matrix and generated the classification reports for each of the models.

2 Dataset Exploration and Inferences

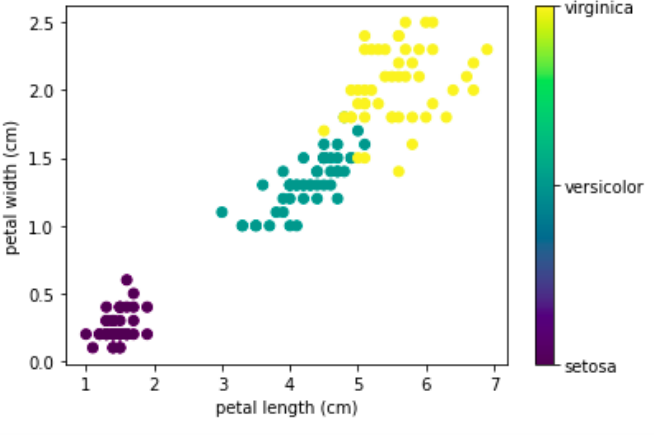
Using the following scatter plot of the whole dataset based on sepal length and sepal width:



1)We can clearly infer that the Setosa species of Iris is linearly separable from the rest of the species based on sepal length and sepal width where as the other two species have some overlap.

2)We can also observe strong correlation between sepal length and sepal width here.

Using the following scatter plot of the whole dataset based on petal length and petal width:



1)We can clearly observe that the Setosa species is differentiable on the basis of its small petal length and petal width and the other two species can be linearly classified on the basis of petal length and petal width.

2)We can also observe the strong correlation between petal length and petal width here.

3 Performance of the Logistic Regression Model

The Logistic Regression model works on multiclass classification using the One vs All algorithm.For classifying data into k classes,it designs k classifiers,one for each class,and to predict which class a data point belongs to,it finds the probability of the point being in each of the given classes using the k classifiers and assigns it the class with the most probability.

3.1 Confusion Matrix and Accuracy

Predicted Values in Columns
Actual Values in Rows

	Setosa	Versicolor	Virginica
Setosa	11	0	0
Versicolor	0	12	1
Virginica	0	0	6

The following five accuracies were measured based on the 5 fold cross validation: $\{96.6\%, 96.6\%, 93.3\%, 93.3\%, 100\%\}$ and the average accuracy is 95.96%.

3.2 Classification Report

	precision	recall	f1-score	support
class 0	1.00	1.00	1.00	11
class 1	1.00	0.92	0.96	13
class 2	0.86	1.00	0.92	6
accuracy			0.97	30
macro avg	0.95	0.97	0.96	30
weighted avg	0.97	0.97	0.97	30

4 Performance of the Gaussian Naive Bayes Model

Naive Bayes classifier is a probabilistic classifier which employs the Baye's theorem to find the probability of a data point lying in a particular class.The Gaussian version of this classifier assumes that points in each class belong to a particular Gaussian having some mean and some standard deviation.

4.1 Confusion Matrix and Accuracy

Predicted Values in Columns
Actual Values in Rows

	Setosa	Versicolor	Virginica
Setosa	11	0	0
Versicolor	0	12	1
Virginica	0	0	6

The following five accuracies were measured based on the 5 fold cross validation: $\{96.6\%, 96.6\%, 100\%, 90\%, 93.3\%\}$ and the average accuracy is 95.3%.

4.2 Classification Report

	precision	recall	f1-score	support
class 0	1.00	1.00	1.00	11
class 1	1.00	0.92	0.96	13
class 2	0.86	1.00	0.92	6
accuracy			0.97	30
macro avg	0.95	0.97	0.96	30
weighted avg	0.97	0.97	0.97	30

5 Performance of the K Means Unsupervised Learning Algorithm

This algorithm works by randomly selecting k clusters,and then labelling all points of the dataset to the nearest point,and then based on this labelling it again calculates the k centroids and repeats this process till the centroids don't change anymore or the maximum number of iterations has been reached.This process is also called the mean-shift algorithm.When this process ends the data has been divided into k clusters.

5.1 Confusion Matrix and Accuracy

Predicted Values in Columns
Actual Values in Rows

	Setosa	Versicolor	Virginica
Setosa	11	0	0
Versicolor	0	12	1
Virginica	0	0	6

The following five accuracies were measured based on the 5 fold cross validation: $\{96.6\%, 93.3\%, 100\%, 90\%, 93.3\%\}$ and the average accuracy is 95.3%.

5.2 Classification Report

	precision	recall	f1-score	support
class 0	1.00	1.00	1.00	11
class 1	1.00	0.92	0.96	13
class 2	0.86	1.00	0.92	6
accuracy			0.97	30
macro avg	0.95	0.97	0.96	30
weighted avg	0.97	0.97	0.97	30

6 Principal Component Analysis

Principal component analysis finds the features which explain most of the variance in the data,which are the primary features affecting the final label of that particular data point.Upon applying PCA on the Iris Dataset,we found that the total variance in the data is explained by the features of the data in this proportion:
Sepal Length:92.58%
Sepal Width:5.21%
Petal Length:1.65%
Petal Width:0.53%