Assignment 2

Name: Preetesh Verma
EntryNo.:2018EEB1171

Foundations Of Data Science
Course Code:CS521

Abstract

This document demonstrates the various observations made from the second lab assignment given to us in this course.Our assignment was to implement Classification on the classic `Iris Dataset`.First we did the Exploratory Data Analysis of the dataset based on the various classes as well as on the complete dataset as a whole, and then we performed the task of classification via a five-fold cross validation method, using several algorithms such as Logistic Regression, Gaussian Naive Bayes Classifier and K Means algorithms and thereafter we had to plot the residuals sample points in a 3-D plot using the PCA (Principal Component Analysis) and also compare the results obtained from the different techniques using the Classification Report metric. The libraries used were **sklearn,Pandas,Matplotlib,Seaborn and Numpy**.

1 About Classification Problem

In machine learning and statistics, classification is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known. Examples are assigning a given email to the "spam" or "non-spam" class. Classification is a pattern recognition problem.

Classification predictive modeling is the task of approximating a mapping function (f) from input variables (X) to discrete output variables (y). The output variables are often called labels or categories. The mapping function predicts the class or category for a given observation. Classification can also be done using clustering analysis.

Thus classification task can be performed using both **Supervised** and **Unsupervised** Machine Learning Techniques. If we have only two classes then we term it as Binary Classification.Our problem is a multi class classification problem.By default sklearn package uses the ONE VS REST scheme for such problems,where each class's probability is calculated against all the remaining classes. The accuracy of the Classification problem is dependent on the type of dataset we have.If the distribution of the classes in the dataset is fairly even then Accuracy can be used as a parameter otherwise in case of imbalanced Classification problems we generally use Precision and Recall as metrics. Here in our problem we have computed the whole Classification Report which tells us all the three and since all the classes in the IRIS DATASET have equal frequencies so all the three metrics are valid.

1.1 EDA

From the data visualisation through the Matplotlib and sea born libraries the distribution of the features(petal length,petal width,sepal length,sepal width) is quite clear.Scatter plots,bar

plots,line plots have been used to explore the features with a clear cut distinction between the three classes('setosa', 'versicolor', 'virginica') in petal features while the sepal features does have a few overlapping sample points.The box plot graphs of the features also shows that the maximum number of outliers are present in the Sepal Width category with the trend in the other three features being that the setosa has smaller petals as compared to versicolor and virginica but sepal length being quite overlapped in the classes. Then we divided the data into Training and Test dataset respectively.

1.2 Logisitic Regression

It is Discriminative model which computes the Likelihood of sample being in a class. After the data has been split we first performed Logistic Regression using the K-Fold Cross Validation technique with k being 5 in this case.(In KFold cross validation the dataset is divided into k equal sets and k-1 sets are used for Training and the other one for Test.)For Logistic Regression it is clearly seen that using KFold method improves the accuracy of the method since in each step the algorithm performs better than the previous step and attains an overall accuracy of 96

1.3 Gaussian Naive Bayes

It is a Generative model which computes the probability distribution of each class and then classifies the samples. Similar to Logistic Regression in Gaussian Naive Bayes Method as well we use the KFold method and observe the increase in the accuracy of the algorithm with a slight dip towards the end which could be a result of overfitting.

Both the above methods were supervised machine learning methods.

1.4 K Means Clustering

It's a unsupervised machine learning algorithm where clusters are made based upon grouping of the features(using euclidean distance as the way of doing it) of the samples.Since we already know the number of classes here so we can use it as an input otherwise it serves as a hyper parameter and is estimated using the Elbow Method.The accuracy of the algorithm is relatively low as compared to supervised algorithms because of the overlapping of the sepal features in the three classes.

1.5 PCA and error analysis

PCA-Principal Component Analysis is a way of dimension reduction for datasets where the number of features is high.PDA is a good method to reduce the dimensions by keeping only the features which does convey a certain amount of the variance in the data.On performing the PDA of the iris dataset it is quite evident that the sepal width does little help to understand the Classification task and so it is not that important and can be ignored or we can get a new parameter using this parameter.To plot the 3-d plot of the misclassification samples we only use the remaining three features. From the Classification report of the three methods we can see clearly that the mistakes by the Supervised Machine Learning is in classifying the second and the third classes which is explainable from the box plot visualisation of the sepal width feature which has median of approximately equal magnitude in both the classes.Also an interesting fact is that Both the supervised learning algorithms misclassifies

Method	Accuracy
Logistic Regression	96
Gaussian Naive Bayes	96
K Means	89.33

Table 1: Results

the exactly same 6 samples wrongly.The unsupervised learning algorithm makes the wrong classification where the features of the classes are overlapping and is quite visible from the graphs plotted.

1.6 Conclusions and Learnings

The following are the conclusions:
1.The distribution plots show the distribution of the features vs the samples and the distribution of the Sepal width feature is the most overlapped one which also brings down our accuracy.

2.The petal features are generally distinguished for all the three classes showing a gradual increase in the size as we move from class 1 to another.

3.The k fold method is a good way to better fit our models as in most of the cases it improves our accuracy by training and testing on the complete dataset in an indirect way.It can also be useful when we have a small dataset like in our problem.
4.Elbow Graph for K means algorithm correctly predicts to use three classes as there are three and the graphs after three classes shows a horizontal movement forward.

5.PCA provides three dimensions which explain about 99.5 percent of the overall variance in the results and when predicted using these three features the accuracy is higher.

6.Gaussian Naive Bayes and Logistic Regression both have the overall same accuracy but different precision and recall again highlighting the importance of other performance metrics rather than just accuracy to judge a algorithm for it's performance.

The assignment exercise tells us that it is good to try several methods for a problem and which ever fits well should be used.It is quite possible to have both Supervised and Unsupervised methods working at high accuracy.EDA is amongst the most important steps to understand the data and should be done necessarily to get the insight from data.One vs All scheme is helpful in multi class classification problems and in case of Gaussian Naive Bayes the class with the highest probability is selected.