ISYE 7406 Homework 3

**Solution**

1. **Introduction**

The purpose of this assignment to analyze and explore the data, create a classification variable and see which all features would impact the classification. We will need to do thorough EDA to find out relationship between dependent and independent variables. Once we know which variables are important for classification, we will use various classification techniques and determine which one would give best results.

This dataset is taken from statlib which is maintained by CMU, below is description of each variable in dataset

|  |  |  |
| --- | --- | --- |
| Column Name | Type | Comment |
| Mpg | Continuous | Miles per gallon |
| Cylinders | Categorical | Number of cylinders in vehicle |
| Horsepower | Continuous |  |
| Weight | Continuous |  |
| Acceleration | Continuous |  |
| Year | Discreate |  |
| Origin | Categorical | 1 is a car made in America, 2 in Europe and 3 in Asia or other part of the world (Categorical variable) |

In this dataset, as an output, mpg feature was transformed into binary variable(0-1) based on the median of mpg. We will be creating a classification model to predict this new variable mpg01

1. **Exploratory data analysis**

Let us start with creating a correlation heatmap matrix using python seaborn library

A picture containing graphical user interface

Description automatically generated

If we check mpg01 correlation with other variables it shows high correlation with cyclinders, displacement, weight and horsepower (all of them highly negatively correlated)

We also plotted pair plots with hue as mpg01 to check how 2 variables interact with each other and how mpg01 is related to same

Diagram

Description automatically generated

We also created box plots to see how mpg01 is related to each parameter and how it is distributed

Chart, box and whisker chart

Description automatically generated

Using pandas\_profiling library we ran profile report on dataframe and output of same was provided in html file. Html file is attached to assignment as well. Report is quite exhaustive and shows overview, variables, interaction and correlations

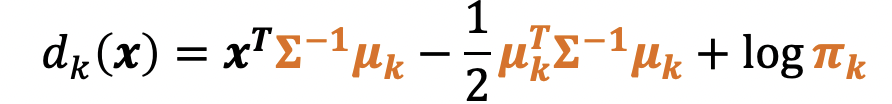
From the heatmap, our attention is immediately drawn to variables, siri, and density. Especially, these variables are shown to have a strong relationship with the dependent one with its absolute term close to 1. On the other hand, we also have two variables (i.e., free and height) which display a weak relationship, eventually leading to a less explanatory power of a model.

1. **Methods**

Let us try all techniques suggested in assignment, before that data needs to be divided into training test and validation. We divided the data in training and test using sklearn train\_test\_split package. Based on EDA done in previous section we can assume displacement,horsepower, weight and acceleration would be most relevant feature to predict mpg01

1. Linear discriminant analysis

We will use sklearn package LinearDiscriminantAnalysis. As per the lecture notes LDA uses linear discriminant function as below



Validation or test data set created from original dataset would be used to check the performance of model apart from running on training dataset

1. QDA

Again we will use sklearn package and validation would be done on test dataset. Assumption as per lecture notes

Text, letter

Description automatically generated

1. Naïve Bayes

Again we will use sklearn package and validation would be done on test dataset. Assumption as per lecture notes

Diagram, text

Description automatically generated with medium confidence

1. Logistic regression

Assumption: Logistic regression uses probability of each case to build models. It utilizes logit function g(πi) = log( πi/(1−πi ).

1. KNN

KNN is memory-based, and does not need a model to be fit. It uses training examples to find k examples that are closet to a given point. That is, di = |xi − x0| where di is distance between xi and x0

In this classification model we will have to scale the data so that all values are at same scale which will help determining closest neighbor better. We will calculate error rate for different values of k, lowest error rate value would finalize the value of k for model

1. **Results**

Below are results of each method trained using train and test dataset

|  |  |  |
| --- | --- | --- |
| Classification Method | Training Error | Testing Error |
| LDA | 0.08394 | 0.16101 |
| QDA | 0.08394 | 0.169491 |
| Gaussian Naïve Bayes | 0.098540 | 0.169491 |
| Logistic Regression | 0.105839 | 0.14406 |

There is hardly any difference in errors between various methods. Logistic regression is slightly better on testing dataset compare to other models.

Next lets explore KNN method, for KNN method we need to find K value which gives least error. We ran KNN for values between 1 to 10. Below is error corresponding to each k value

1 0.11864406779661017

2 0.09322033898305085

3 0.11864406779661017

4 0.11864406779661017

5 0.1271186440677966

6 0.13559322033898305

7 0.1440677966101695

8 0.1440677966101695

9 0.1440677966101695

10 0.13559322033898305

It is easily noticeable that error was least for k=2 and then error value is increasing . Hence we would suggest to use K=2 for the algorithm. If we use K=2 and evaluate the confusion matrix below is what it would look like

Square

Description automatically generated with medium confidence

This is quite surprising as it has not misclassified one class which is really unexpected

1. **Findings**

Below are findings for the entire exercise

We did EDA on dataset and determined how each feature interact with mpg01, it also helped us determine important features. Using important features we created various classification model using techniques like LDA, QDA, Gaussian Naïve bayes, logistic regression and KNN.

We realized that LDA, QDA and Naïve Bayes have very similar results on test dataset (same range of error). However, logistic regression was slightly better results. Explanation for this could be that logistic regression is based on probability while LDA, QDA and Naïve bayes are based on Bayes.

KNN seems to give good result when value of k was chosen to be 2. K=2 value model was quite good and didn’t mis-classified one class.

However this was very small datasets, just 392 records if sample size was huge or if number of features were high these results would have been drastically different.