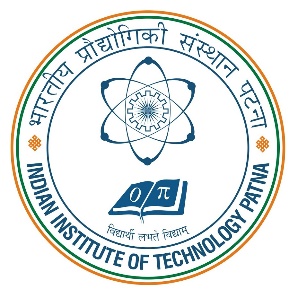
**Foundation of Machine Learning**

**(CS564)**

Assignment Report on Linear Regression, Regularized Liner Regression and Logistic Regression

Anindya Sundar Das (1811MC02)



Introduction:

The purpose of this assignment was to extract and analyse features of a given dataset of instances and predict attribute values and calculate the performance different machine learning models on the same dataset for different parameter values. Machine learning models such as Linear Regression and Regularized Linear regression was used to address regression problem and Logistic Regression Model addresses the classification problem.

Problem Definition:

The winequality-red dataset has the following features:

|  |  |  |  |
| --- | --- | --- | --- |
| Number of instances | 1599 | Attribute Characteristic | Real |
| Number of attributes | 12  [fixed acidity, volatile acidity,  Citric acid,  Residual Sugar,  Chlorides,  Free sulphur dioxide,  Total sulphur dioxide,  Density,  pH,  Sulphates,  Alcohol,  Quality] | Dataset Characteristic | Multivariate |

Dataset Download Link: https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/

1. The Task is to train a linear regression model and predict the ‘Alcohol Variable’ and to report average Residual Sum of Square error over 5 folds.
2. The same dataset is to be used for Regularized Regression model on ‘Alcohol Variable’ and report the residual error of each fold and to determine the value of λ(penalty term) which provides best outcome.
3. Multiclass Logistic regression is to be performed on variable quality and to report per-class precision, recall, and f-measure for each fold and to calculate the average values.

Experiment Evaluation:

1. Linear Regression:
   1. Methodology: The dataset was divided into independent and dependent variable datasets. The dependent variable dataset contains ‘Alcohol variable’ while the independent variable dataset contains rest of the attributes. Using Sklearn KFold module the independent and dependent datasets were broken down into k=5 folds.

After the dataset preparation was done, the linear regression model was trained using test and train data for each fold and independent test data was used to predict the dependent variable. The Residual sum of square was calculated for each folds for this linear regression model.

* 1. Results:

Detail output:



Summary:

Fold : 1

RSS Value : 130.45978122223067

Fold : 2

RSS Value : 220.17131114573982

Fold : 3

RSS Value : 98.08035976774121

Fold : 4

RSS Value : 92.61662465333796

Fold : 5

RSS Value : 102.24399756026314

^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^

Average RSS value 128.71441486986254

1. Regularised Linear Regression:
   1. Methodology: Same method has been followed to extract the independent and dependent variable datasets and to prepare train and test data for each of five folds each of train and test datasets as mentioned in 1.1. The train data were used to fit the ridge model and independent test data was used to predict the dependent variable.

The KFold ridge regression was performed for range of λ(penalty term)

values .Based on average Residual Sum of square value the best model(λ) was reported.

* 1. Results:

Detail Output:



Summary:

alpha = 0 Average rss value = 128.71441486986498

alpha = 1e-05 Average rss value = 128.6573718787976

alpha = 2e-05 Average rss value = 128.61369616432526

alpha = 4e-05 Average rss value = 128.56438938498442

alpha = 5e-05 Average rss value = 128.55778007256936

alpha = 6.000000000000001e-05 Average rss value = 128.56257978659826

alpha = 7.000000000000001e-05 Average rss value = 128.57834587211505

alpha = 8e-05 Average rss value = 128.60465282867563

alpha = 9e-05 Average rss value = 128.64109157395524

alpha = 0.0001 Average rss value = 128.6872687428144

alpha = 0.00011 Average rss value = 128.74280601987635

alpha = 0.00012000000000000002 Average rss value = 128.80733950384894

alpha = 0.00013000000000000002 Average rss value = 128.8805191019059

alpha = 0.00014000000000000001 Average rss value = 128.96200795258503

alpha = 0.00015000000000000001 Average rss value = 129.05148187556367

alpha = 0.00016 Average rss value = 129.14862884716769

alpha = 0.00017 Average rss value = 129.25314849996042

alpha = 0.00018 Average rss value = 129.36475164546852

alpha = 0.00019 Average rss value = 129.48315981869288

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

**alpha = 5e-05 gives best fit , with average RSS = 128.55778007256936**

1. Logistic Regression:
   1. Methodology:

The input dataset was divided into dependent variable (variable quality) and independent variable( all features except variable quality)

dataset . The dataset was further divided into k=5 folds using module StratifiedKFold, each folds having train and test dataset.

For each fold the train dataset was used to fit the logistic regression model and train dataset was used predict the output variable. The predicted and test outputs were used to determine confusion matrix, per-class precision, recall , f-measure using the modules from sklearn.metrics for each folds. Trace of the confusion matrix determines the correctly classified instances, while the sum of the off-diagonal elements describes number of misclassified instances.

* 1. Results:

Detail results:



Summary:

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Average precision [0. 0. 0.6470982 0.51490857 0.36135338 0. ]

Average recall [0. 0. 0.7427866 0.61183563 0.095 0. ]

Average f-measure [0. 0. 0.68224313 0.54955022 0.14407908 0. ]

Conclusion:

In this assignment we have fitted the our train data to Linear Regression model using five-fold cross validation and calculated the error. The linear model predicts the output with average Residual sum of square (RSS) 128.71441486986254

The Ridge regression model gives the best fit on the same dataset which minimizes the RSS for λ(penalty term)= **5e-05**

The multiclass logistic regression successfully identifies 6 classes as present in the input dataset and reports different metrics of the model as shown in the results.