**Foundations of Intelligent Systems**

**Homework #2 Report**

**Created by Pooja Ketan Shah**

**ABSTRACT:**

Machine learning and regression are important concepts which can be used to analyze the data. The transformer statistics is used to predict the health index of transformers using linear regression and logistic regression. Four models are built up using linear and logistic regression. In linear regression, two models are built up: One model is built for all the dataset and other model is built by selecting the highly correlated features based on Interfacial Tension attribute. In logistic regression, two models are built up: One model is built for all the dataset and other model is built by selecting the highly correlated features based on Class attribute.

**INTRODUCTION:**

**Machine learning:** Machine learning is a type of artificial intelligence which gives computers the ability to learn without explicitly being programmed [1]. It involves writing an algorithm that can learn from and make future predictions of the data. Machine learning is used in the fields of data mining and statistics.

Machine learning is divided into three categories:

* Supervised Learning
* Unsupervised Learning
* Reinforcement Learning

**Supervised Learning:** In supervised learning, the expected output is known. It can infer the function from labeled training data [2]. A supervised learning algorithm takes a known set of input data and known responses to the data which trains the model to predict the outcome of the new data [3].

Known input data

Known responses

Model

Expected outcome (Predicted)

New data

Figure 1: Supervised Learning Model

Supervised learning is divided into two categories which are as follows:

* Classification: In machine learning, classification is the technique which is used to identify which category the new data belongs to from a given set of categories based on the training set of data which contains the set of past observations. For example, the illness of a patient can be predicted based on the symptoms of the patient and other patient history which is stored in the database.
* **Regression**: Regression is a statistical method which helps to determine the relationship between a dependent variable and one or more independent variables [4]. Regression is used in various prediction and forecasting applications. For example, the weather of a particular city can be forecasted using regression. There are various types of regression which are as follows:
  + Linear Regression
  + Logistic Regression
* **Linear Regression**: Linear regression is a regression model which is used when the dependent variable is continuous or numerical value. Linear regression model helps to determine the relationship between a dependent variable and one or more independent variables [5]. A simple linear regression determines the relationship between a dependent variable and one independent variable. A simple linear regression is represented by the formula:

A multiple linear regression determines the relationship between a dependent variable and many independent variables. A multiple linear regression is represented by the formula:

* **Logistic Regression**: Logistic regression is a regression model which is used when the dependent variable is categorical value [6]. Logistic regression determines the relationship between a categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function [6]. A logistic regression is represented by the formula:

**STEPS OF REGRESSION MODEL IMPLEMENTATION:**

**Linear Model:** Two models are created for linear model. One linear model is built to predict the relationship between Interfacial tension and all independent variables. Other linear model is built to predict the relationship between Interfacial tension and highly correlated independent variables.

Below are the steps to build a linear model which predicts relationship between Interfacial tension and all independent variables:

1. Read the csv file
2. Exclude the columns ‘year’ and ‘Class’
3. Split the data into training data (70%) and testing data (30%)
4. Build a linear regression model using ‘lm’ function where Interfacial tension is dependent variable and other columns are independent variables on training data.
5. Predict the testing data on the model that was built.
6. Calculate the RMSE (Root Mean Square Error) which will measure the error between predicted and actual value.

Below are the steps to build a linear model which predicts relationship between Interfacial tension and highly correlated independent variables:

1. Read the csv file
2. Exclude the columns ‘year’ and ‘Class’
3. Find the highly correlated independent variables from correlation matrix where the absolute of correlation value between Interfacial tension and independent variables is greater than 0.5
4. Split the data into training data (70%) and testing data (30%)
5. Build a linear regression model using ‘lm’ function where Interfacial tension is dependent variable and highly correlated columns are independent variables on training data.
6. Predict the testing data on the model that was built.
7. Calculate the RMSE (Root Mean Square Error) which will measure the error between predicted and actual value.

**Logistic Regression**: Two models are created for logistic model. One logistic model is built to predict the relationship between Class and all independent variables. Other linear model is built to predict the relationship between Class and highly correlated independent variables.

Below are the steps to build a logistic model which predicts relationship between Class and all independent variables:

1. Read the csv file
2. Exclude the columns ‘year’ and ‘Interfacial tension’
3. Add 4 columns to the data to convert the categorical Class data into numerical data.
4. Split the data into training data (70%) and testing data (30%)
5. Build 4 different logistic models for each class using ‘glm’ function on training data.
6. Predict the testing data on the model that was built for each class. Add these predicted values to 4 columns in testing data.
7. From the 4 columns that where added which includes the predicted values of each class, find the maximum of these 4 columns and consider the max of these 4 columns as a class which was predicted.
8. Build a confusion matrix on predicted class and actual class columns.
9. Plot ROC for each predicted class and actual class columns.

Below are the steps to build a logistic model which predicts relationship between Class and highly correlated independent variables:

1. Read the csv file
2. Exclude the columns ‘year’ and ‘Interfacial tension’
3. Using randomForest library, identify the variables that are highly correlated with the Class variable.
4. Add 4 columns to the data to convert the categorical Class data into numerical data.
5. Split the data into training data (70%) and testing data (30%)
6. Build 4 different logistic models for each class using ‘glm’ function on training data.
7. Predict the testing data on the model that was built for each class. Add these predicted values to 4 columns in testing data.
8. From the 4 columns that where added which includes the predicted values of each class, find the maximum of these 4 columns and consider the max of these 4 columns as a class which was predicted.
9. Build a confusion matrix on predicted class and actual class columns.
10. Plot ROC for each predicted class and actual class columns.

**RESULTS ANALYSIS**:

**Types of graphs:**

* **Residuals Vs Fitted Graph**: Residual is calculated by subtracting the predicted values from the actual values of the dependent variable [7]. Residual Vs Fitted graph provides the best estimate of the error term from the regression model [7].
* **Normal Q-Q Graph**: A normal quantile-quantile plot of residuals should follow a straight line which shows that the errors of the model follow a normal distribution [7]. If the deviations are away from straight line, the errors do not follow normal distribution [7].
* **Scale-Location Graph**: It is similar to residuals vs fitted graph but uses square root of standardized residuals [7]. It determines whether it follows the pattern of homoscedasticity where homoscedasticity means that the variance of residual is same for all values of X [8].
* **Residuals Vs Leverage Graph**: Leverage is a measure which helps to determine how much each data point influences the regression [9]. Cook’s distance measures how much the regression would change if a point was deleted [9].

**Linear Regression Model Graphs:**

* **Residuals Vs Fitted Graph**: As shown in the linear regression graph for all dataset and highly correlated variables, there are three outliers which are mentioned in numbers. Linear Regression Model graph for all dataset and highly correlated variables are attached below:



* **Normal Q-Q Graph**: As shown in linear regression graph for all dataset and highly correlated variables, Normal Q-Q graph follows symmetric with fat tails distribution.

****

* **Scale-Location Graph**: As shown in linear regression graph for all dataset and highly correlated variables, scale-location graph follows the pattern of homoscedasticity.



* **Residuals Vs Leverage Graph**: As in linear regression graphs for all and highly correlated variables, the leverage is high.



**Logistic Regression Model Graphs:**

* **Residuals Vs Fitted Graph**: Logistic Regression Model graph for all dataset and highly correlated variables for each class are attached below:

Logistic Regression with all datasets:

Class B Class G Class M Class N



Logistic Regression with highly correlated variables:

Class B Class G Class M Class N



* **Normal Q-Q Graph**: For logistic regression, normal q-q graph is S shaped graph. Logistic Regression Model graph for all dataset and highly correlated variables for each class are attached below:

Logistic Regression with all datasets:

Class B Class G Class M Class N



Logistic Regression with highly correlated variables:

Class B Class G Class M Class N

****

* **Scale-Location Graph**: Logistic Regression Model graph for all dataset and highly correlated variables for each class are attached below:

Logistic Regression with all datasets:

Class B Class G Class M Class N



Logistic Regression with highly correlated variables:

Class B Class G Class M Class N



* **Residuals Vs Leverage Graph**: Logistic Regression Model graph for all dataset and highly correlated variables for each class are attached below:

Logistic Regression with all datasets:

Class B Class G Class M Class N



Logistic Regression with highly correlated variables:

Class B Class G Class M Class N



**Testing Data with Predicted Values**:

**Linear Regression with predicted Interfacial Tension for all datasets:**



**Linear Regression with predicted Interfacial Tension for highly correlated variables:**



**Logistic Regression with predicted Class for all dataset:**



**Logistic Regression with predicted Class for highly correlated variables:**



**Root Mean Square Error (RMSE) for Linear Regression:**

RMSE is calculated by using the formula:

* RMSE for Linear Regression with all dataset: 3.72676
* RMSE for Linear Regression with highly correlated variables: 3.74375

**Confusion matrix**:

**Confusion matrix of logistic regression for all dataset:**



**Accuracy** of logistic regression for all dataset can be calculated as:

(Number of correctly predicted Class B + Number of correctly predicted Class G + Number of correctly predicted Class M + Number of correctly predicted Class N) / total number of rows in testing data

= (22 + 170 + 2) / 220 = **88.18%**

**Confusion matrix of logistic regression for highly correlated variables:**



**Accuracy** of logistic regression for highly correlated variables can be calculated as:

(Number of correctly predicted Class B + Number of correctly predicted Class G + Number of correctly predicted Class M + Number of correctly predicted Class N) / total number of rows in testing data

= (22 + 172 + 2) / 220 = **89.09%**

**ROC**:

**Logistic Regression ROC for all datasets for Class B:**



**Logistic Regression ROC for highly correlated variables for Class B:**



**Logistic Regression ROC for all datasets for Class G:**



**Logistic Regression ROC for highly correlated variables for Class G:**



**Logistic Regression ROC for all datasets for Class M:**



**Logistic Regression ROC for highly correlated variables for Class M:**



**Logistic Regression ROC for all datasets for Class N:**



**Logistic Regression ROC for highly correlated variables for Class N:**



**REFERENCES:**

[1] [*https://en.wikipedia.org/wiki/Machine\_learning*](https://en.wikipedia.org/wiki/Machine_learning)

[2] [*https://en.wikipedia.org/wiki/Supervised\_learning*](https://en.wikipedia.org/wiki/Supervised_learning)

[3] [*http://www.mathworks.com/help/stats/supervised-learning-machine-learning-workflow-and-algorithms.html*](http://www.mathworks.com/help/stats/supervised-learning-machine-learning-workflow-and-algorithms.html)

[4] [*https://en.wikipedia.org/wiki/Regression\_analysis*](https://en.wikipedia.org/wiki/Regression_analysis)

[5] [*https://en.wikipedia.org/wiki/Linear\_regression*](https://en.wikipedia.org/wiki/Linear_regression)

[6] [*https://en.wikipedia.org/wiki/Logistic\_regression*](https://en.wikipedia.org/wiki/Logistic_regression)

[7] [*http://www.stat.berkeley.edu/~spector/s133/Lr0.html*](http://www.stat.berkeley.edu/~spector/s133/Lr0.html)

[8] [*http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/R/R5\_Correlation-Regression/R5\_Correlation-Regression7.html*](http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/R/R5_Correlation-Regression/R5_Correlation-Regression7.html)

[9] [*http://strata.uga.edu/6370/rtips/regressionPlots.html*](http://strata.uga.edu/6370/rtips/regressionPlots.html)