**Homework-2**

CSCE 633: Machine Learning

Manav Gurumoorthy

830000011

Introduction:

In this HW, we looked at performing linear regression and logistic regression on a dataset containing traffic data of Sao Paulo, Brazil.

The dataset consisted of 17 features corresponding to some information collected regarding the traffic conditions. The outcome was quantified as Slowness in traffic in (%).

The first half of the HW was to make a linear regression model, using the ordinary least squares solution. The second half was to then perform logistic regression, in order to classify the data as either high traffic flow or low traffic flow.

Data Exploration:

In this part, we explored the data that we were going to model. The first step was to plot the histograms of al the data features to see how they were distributed.

The correlation of the data with respect to the outcome was visualised using scatter plots and quantified using Pearson’s ‘r’ metric. Pearson’s r was also computed between every pair of features as well.

Implementation of Linear Regression:

In this implementation of Linear Regression, we used the ordinary least squares solution, which is a closed form solution to compute a set of optimal weights for the model. In order to do this a data matrix was created from the dataset. The data matrix consisted of 18 columns, 17 columns corresponding to the 17 features of the dataset and a ‘1’ appended to it for the intercept.

Using this data matrix, the optimal weights were computed for the training dataset.

Testing the Linear Regression model:

With the set of optimal weights computed, the model was now fit. The next step was to use this model to make predictions on the test dataset. Along with this the RSS error and Pearson’s r between the actual and predicted outcome was calculated to measure the accuracy of the model. This gave an RSS error of 501.86 and r = 0.82.

Improving the model:

In order to improve the accuracy of the model, we played with the features being considered in the dataset. In this implementation, we dropped all features having an r < |0.1|, so we could have a more correlated dataset. The features dropped were “ 'Immobilized bus', 'Broken Truck', 'Running over', 'Occurrence involving freight', 'Incident involving dangerous freight', 'Fire', 'Manifestations' ”. This left us with a model having 11 features. Upon testing again, the RSS error dropped significantly to 3.96e-26 and r = 1. This meant that the data was perfectly predicted. While this seems like the model overfit for the data. The results reported here are on the test set, so it is unclear if the model was really overfit or if it was a simple dataset so the prediction was very accurate.

Logistic Regression:

As Logistic Regression is a classifier, the first step was to convert the outcomes into discrete classes. To do this the outcomes were split as class 0 and class 1 based on their value with respect to the mean of all outcomes. Once this was done, logistic regression was run on the dataset using sklearn’s inbuilt Logistic regression model.

Logistic Regression with regularisation:

In order to improve the accuracy of the model, we used regularisation. The regularisation strength is a hyper parament. Hence in order to figure out the optimal value of regularisation, we used cross-validation over the test set to find the hyper parameter giving the best cross validation accuracy. Then we run the model with this hyper parameter over the test set.

Results:

The final Linear Regression model had an RSS error of ~ 500 and r =0.82. The logistic Regression model with regularisation had an accuracy of 74%.