**Data Analytics and Visualization**

MET CS 555

@2020 Farshid Alizadeh-Shabdiz,

Hao Wu

MET CS 555 Assignment Logistic Regression – 20 points

**SUBMISSION REQUIREMENTS: Please submit a single document (word or PDF) for submission.  Your submission should contain a summary of your results (and answers to questions asked on the homework) as well as your R code used to generate your results (please append your R code to the end of your submission).**

**Homework submission filenames should take on the form LASTNAME - Homework #.doc(x)/.pdf.**

**No ZIP files please.**

**(10 points) Question 1**

1. In this problem, you will develop a model to predict whether a given car gets high or low gas mileage based on the Auto data set, which is included in the ISLR pachage.

(a) Create a binary variable, mpg01, that contains a 1 if mpg contains a value above its median, and a 0 if mpg contains a value below its median. You can compute the median using the median() function. Note you may find it helpful to use the data.frame() function to create a single data set containing both mpg01 and the other Auto variables.

Median is 22.75, using mutate method create binary variable mgp01.

(b) Explore the data graphically in order to investigate the association between mpg01 and the other features. Which of the other features seem most likely to be useful in predicting mpg01? Scatterplots and boxplots may be useful tools to answer this question. Describe your findings.

手机屏幕的截图

中度可信度描述已自动生成图表

描述已自动生成

For binary variable mgp01, we could saw except acceleration, year and origin variable, other variable has approach 1 or -1 correlation with mgp01; therefore, before we start do logic regression, we need cleaning this factor.

(c) Split the data into a training set and a test set with the 80 / 20 ratio.

dt = sort(sample(nrow(df1), nrow(df1)\*.8))

train<-df1[dt,]

test<-df1[-dt,]

(d) Perform logistic regression on the training data in order to predict mpg01 using the variables that seemed most associated with mpg01 in (b). What is the test error of the model obtained?

The test error rate of the model obtained is 1 - 0.8987342 = 0.1013 = 10.13%

**(10 points) Question 2**

This question should be answered using the Weekly data set, which is part of the ISLR package. This data contains 1,089 weekly returns for 21 years, from the beginning of 1990 to the end of 2010.

1. Produce some numerical and graphical summaries of the Weekly data. Do there appear to be any patterns?

文本, 表格

描述已自动生成

图表, 气泡图

描述已自动生成

We could saw all variable has a weak relationship. In Weekly dataset, we have a string variable “Direction” which is a binary variable contain “Down” and “Up”.

1. Use the full data set to perform a logistic regression with Direction as the response and the five lag variables plus Volume as predictors. Use the summary function to print the results. Do any of the predictors appear to be statistically significant? If so, which ones?

表格

描述已自动生成

We could saw only Lag2 variable appear to be statistically significant at 95% confidence interval; therefore, we will do logic regression between direction and Lag2 in next step.

1. Compute the confusion matrix and overall fraction of correct predictions. Explain what the confusion matrix is telling you about the types of mistakes made by logistic regression.

文本

描述已自动生成

After we compute the confusion matrix and overall fraction of correct prediction, we get 1-0.56 = 44% test error rate.

1. Now fit the logistic regression model using a training data period from 1990 to 2008, with Lag2 as the only predictor. Compute the confusion matrix and the overall fraction of correct predictions for the held out data (that is, the data from 2009 and 2010).

We try to use Lag2 as the only predictor test whether we can reduce the test error rate

文本

中度可信度描述已自动生成

The result is significant show us the test error rate reduce after we used Lag2 as only predictor