CIS8695

Homework I

What to Turn In:

Please write your answers and paste required results/reports in **a MS Word file**. **Note: Please do NOT submit any other file format, because it will cause grading inconveniency. Failing to submit in correct file format will cause the loss of homework grades! You can use screenshots if it is convenient for you.**

Problems

Homework problems are all from the textbook. For your convenience, the pictures of the problems are also attached in the document at the end. Datasets are provided on iCollege.

Problem 1. (Linear Regression). Look at Problem 6.1 in Chapter 6 of the Textbook, page 169. (A picture of the problem is attached below). Complete (b), (c) and (d.iii) of Problem 6.1.

Problem 2. (Logistic Regression). Look at Problem 10.1 in Chapter 10 of the Textbook, page 268. (A picture of the problem is attached below). Complete (b), (d) and (e) of Problem 10.1. Note for part (b), please write down your calculation equations.

(Optional) You are also encouraged to try Ridge Regression and Lasso for both the two problems.

6.1 **Predicting Boston Housing Prices.** The file *BostonHousing.csv* contains information collected by the US Bureau of the Census concerning housing in the area of Boston, Massachusetts. The dataset includes information on 506 census housing tracts in the Boston area. The goal is to predict the median house price in new tracts based on information such as crime rate, pollution, and number of rooms. The dataset contains 13 predictors, and the response is the median house price (MEDV). Table 6.9 describes each of the predictors and the response.

TABLE 6.9 DESCRIPTION OF VARIABLES FOR BOSTON HOUSING EXAMPLE

CRIM	Per capita crime rate by town
ZN	Proportion of residential land zoned for lots over 25,000 ft^2
INDUS	Proportion of nonretail business acres per town
CHAS	Charles River dummy variable (= 1 if tract bounds river; = 0 otherwise)
NOX	Nitric oxide concentration (parts per 10 million)
RM	Average number of rooms per dwelling
AGE	Proportion of owner-occupied units built prior to 1940
DIS	Weighted distances to five Boston employment centers
RAD	Index of accessibility to radial highways
TAX	Full-value property-tax rate per $10,000
PTRATIO	Pupil/teacher ratio by town
LSTAT	Percentage lower status of the population
MEDV	Median value of owner-occupied homes in $1000s

a. Why should the data be partitioned into training and validation sets? What will the training set be used for? What will the validation set be used for?

b. Fit a multiple linear regression model to the median house price (MEDV) as a function of CRIM, CHAS, and RM. Write the equation for predicting the median house price from the predictors in the model.

c. Using the estimated regression model, what median house price is predicted for a tract in the Boston area that does not bound the Charles River, has a crime rate of 0.1, and where the average number of rooms per house is 6? What is the prediction error?

d. Reduce the number of predictors:

i. Which predictors are likely to be measuring the same thing among the 13 predictors? Discuss the relationships among INDUS, NOX, and TAX.

ii. Compute the correlation table for the 12 numerical predictors and search for highly correlated pairs. These have potential redundancy and can cause multicollinearity. Choose which ones to remove based on this table.

iii. Use stepwise regression with the three options (*backward, forward, both*) to reduce the remaining predictors as follows: Run stepwise on the training set. Choose the top model from each stepwise run. Then use each of these models separately to predict the validation set. Compare RMSE, MAPE, and mean error, as well as lift charts. Finally, describe the best model.

6.2 **Predicting Software Reselling Profits.** Tayko Software is a software catalog firm that sell... ...tional software. It started out as a software manufacturer

PROBLEMS

10.1 Financial Condition of Banks. The file *Banks.csv* includes data on a sample of 20 banks. The "Financial Condition" column records the judgment of an expert on the financial condition of each bank. This outcome variable takes one of two possible values—*weak* or *strong*—according to the financial condition of the bank. The predictors are two ratios used in the financial analysis of banks: TotLns&Lses/Assets is the ratio of total loans and leases to total assets and TotExp/Assets is the ratio of total expenses to total assets. The target is to use the two ratios for classifying the financial condition of a new bank.

Run a logistic regression model (on the entire dataset) that models the status of a bank as a function of the two financial measures provided. Specify the *success* class as *weak* (this is similar to creating a dummy that is 1 for financially weak banks and 0 otherwise), and use the default cutoff value of 0.5.

a. Write the estimated equation that associates the financial condition of a bank with its two predictors in three formats:

i. The logit as a function of the predictors

ii. The odds as a function of the predictors

iii. The probability as a function of the predictors

b. Consider a new bank whose total loans and leases/assets ratio = 0.6 and total expenses/assets ratio = 0.11. From your logistic regression model, estimate the following four quantities for this bank (use R to do all the intermediate calculations; show your final answers to four decimal places): the logit, the odds, the probability of being financially weak, and the classification of the bank (use cutoff = 0.5).

c. The cutoff value of 0.5 is used in conjunction with the probability of being financially weak. Compute the threshold that should be used if we want to make a classification based on the odds of being financially weak, and the threshold for the corresponding logit.

d. Interpret the estimated coefficient for the total loans & leases to total assets ratio (TotLns&Lses/Assets) in terms of the odds of being financially weak.

e. When a bank that is in poor financial condition is misclassified as financially strong, the misclassification cost is much higher than when a financially strong bank is misclassified as weak. To minimize the expected cost of misclassification, should the cutoff value for classification (which is currently at 0.5) be increased or decreased?

10.2 Identifying Good System Administrators. A management consultant is studying the roles played by experience and training in a system administrator's ability to complete a set of tasks in a specified amount of time. In particular, she is interested in discriminating between administrators who are able to complete given tasks within a specified time and those who are not. Data are collected on the performance of 75 randomly selected administr