Homework 6
IE 7275: Data Mining in Engineering

Task 1: Tutorial

- Practice R models presented in "R Code for Textbook Examples in Chap 11 12.pdf." For your convenience, the data sets referenced in the document are included the Homework 4 folder.

Chapter 11: Neural Nets

Problem 11.1

Car Sales. Consider the data on used cars (**ToyotaCorolla.csv**) with 1436 records and details on 38 attributes, including Price, Age, KM, HP, and other specifications. The goal is to predict the price of a used Toyota Corolla based on its specifications.

a. Fit a neural network model to the data. Use a single hidden layer with 2 nodes.
    - Use predictors Age_08_04, KM, Fuel_Type, HP, Automatic, Doors, Quarterly_Tax, Mfr_Guarantee, Guarantee_Period, Airco, Automatic_airco, CD_Player,Powered_Windows, Sport_Model, and Tow_Bar.
    - Remember to first scale the numerical predictor and outcome variables to a 0–1 scale (use function **preprocess**() with method = "range"—see Chapter 7) and convert categorical predictors to dummies.
b. Record the RMS error for the training data and the validation data. Repeat the process, changing the number of hidden layers and nodes to {single layer with 5 nodes}, {two layers, 5 nodes in each layer}.
    i. What happens to the RMS error for the training data as the number of layers and nodes increases?
    ii. What happens to the RMS error for the validation data?
    iii. Comment on the appropriate number of layers and nodes for this application.

Problem 11.2

Direct Mailing to Airline Customers. East-West Airlines has entered into a partnership with the wireless phone company Telcon to sell the latter's service via direct mail. The file **EastWestAirlinesNN.csv** contains a subset of a data sample of who has already received a test offer. About 13% accepted.

You are asked to develop a model to classify East–West customers as to whether they purchase a wireless phone service contract (outcome variable Phone_Sale). This model will be used to classify additional customers.

a. Run a neural net model on these data, using a single hidden layer with 5 nodes. Remember to first convert categorical variables into dummies and scale numerical predictor variables to a 0–1 (use function **preprocess**() with method = "range"—see Chapter 7). Generate a decile-wise lift chart for the training and validation sets. Interpret the meaning (in business terms) of the leftmost bar of the validation decile-wise lift chart.
b. Comment on the difference between the training and validation lift charts.
c. Run a second neural net model on the data, this time setting the number of hidden nodes to 1. Comment now on the difference between this model and the model you ran earlier, and how overfitting might have affected results.
d. What sort of information, if any, is provided about the effects of the various variables?

Chapter 12: Discriminant Analysis

Problem 12.1

Personal Loan Acceptance. Universal Bank is a relatively young bank growing rapidly in terms of overall customer acquisition. The majority of these customers are liability customers with varying sizes of relationship with the bank. The customer base of asset customers is quite small, and the bank is interested in expanding this base rapidly to bring in more loan business. In particular, it wants to explore ways of converting its liability customers to personal loan customers.

A campaign the bank ran for liability customers last year showed a healthy conversion rate of over 9% successes. This has encouraged the retail marketing department to devise smarter campaigns with better target marketing. The goal of our analysis is to model the previous campaign's customer behavior to analyze what combination of factors make a customer more likely to accept a personal loan. This will serve as the basis for the design of a new campaign.

The file **UniversalBank.csv** contains data on 5000 customers. The data include customer demographic information (e.g., age, income), the customer's relationship with the bank (e.g., mortgage, securities account), and the customer response to the last personal loan campaign (Personal Loan). Among these 5000 customers, only 480 (= 9.6%) accepted the personal loan that was offered to them in the previous campaign.

Partition the data (60% training and 40% validation) and then perform a discriminant analysis that models Personal Loan as a function of the remaining predictors (excluding zip code). Remember to turn categorical predictors with more than two categories into dummy variables first. Specify the success class as 1 (personal loan acceptance), and use the default cutoff value of 0.5.

a. Compute summary statistics for the predictors separately for loan acceptors and nonacceptors. For continuous predictors, compute the mean and standard deviation. For categorical predictors, compute the percentages. Are there predictors where the two classes differ substantially?
b. Examine the model performance on the validation set.
    i. What is the accuracy rate?
    ii. Is one type of misclassification more likely than the other?
    iii. Select three customers who were misclassified as acceptors and three who were misclassified as nonacceptors. The goal is to determine why they are misclassified. First, examine their probability of being classified as acceptors: is it close to the threshold of 0.5? If not, compare their predictor values to the summary statistics of the two classes to determine why they were misclassified.
c. As in many marketing campaigns, it is more important to identify customers who will accept the offer rather than customers who will not accept it. Therefore, a good model should be especially accurate at detecting acceptors. Examine the lift chart and decile-wise lift chart for the validation set and interpret them in light of this ranking goal.
d. Compare the results from the discriminant analysis with those from a logistic regression (both with cutoff 0.5 and the same predictors). Examine the confusion matrices, the lift charts, and the decile charts. Which method performs better on your validation set in detecting the acceptors?
e. The bank is planning to continue its campaign by sending its offer to 1000 additional customers. Suppose that the cost of sending the offer is $1 and the profit from an accepted offer is $50. What is the expected profitability of this campaign?
f. The cost of misclassifying a loan acceptor customer as a nonacceptor is much higher than the opposite misclassification cost. To minimize the expected cost of misclassification, should the cutoff value for classification (which is currently at 0.5) be increased or decreased?

Problem 12.2

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 administrators. They are stored in the file **SystemAdministrators.csv**.

Using these data, the consultant performs a discriminant analysis. The variable Experience measures months of full time system administrator experience, while Training measures number of relevant training credits. The dependent variable Completed is either Yes or No, according to whether or not the administrator completed the tasks.

a. Create a scatter plot of Experience vs. Training using color or symbol to differentiate administrators who completed the tasks from those who did not complete them. See if you can identify a line that separates the two classes with minimum misclassification.
b. Run a discriminant analysis with both predictors using the entire dataset as training data. Among those who completed the tasks, what is the percentage of administrators who are classified incorrectly as failing to complete the tasks?
c. Compute the two classification scores for an administrator with 4 months of experience and 6 credits of training. Based on these, how would you classify this administrator?
d. How much experience must be accumulated by an administrator with 4 training credits before his or her estimated probability of completing the tasks exceeds 0.5?
e. Compare the classification accuracy of this model to that resulting from a logistic regression with cutoff 0.5.

Problem 12.3

Detecting Spam E-mail (from the UCI Machine Learning Repository). A team at Hewlett-Packard collected data on a large number of e-mail messages from their postmaster and personal e-mail for the purpose of finding a classifier that can separate e-mail messages that are spam vs. nonspam (a.k.a. "ham"). The spam concept is diverse: It includes advertisements for products or websites, "make money fast" schemes, chain letters, pornography, and so on. The definition used here is "unsolicited commercial e-mail." The file **Spambase.csv** contains information on 4601 e-mail messages, among which 1813 are tagged "spam." The predictors include 57 attributes, most of them are the average number of times a certain word (e.g., mail, George) or symbol (e.g., #, !) appears in the e-mail. A few predictors are related to the number and length of capitalized words.

a. To reduce the number of predictors to a manageable size, examine how each predictor differs between the spam and nonspam e-mails by comparing the spam-class average and nonspam-class average. Which are the 11 predictors that

appear to vary the most between spam and nonspam e-mails? From these 11, which words or signs occur more often in spam?

b. Partition the data into training and validation sets, then perform a discriminant analysis on the training data using only the 11 predictors.
c. If we are interested mainly in detecting spam messages, is this model useful? Use the confusion matrix, lift chart, and decile chart for the validation set for the evaluation.
d. In the sample, almost 40% of the e-mail messages were tagged as spam. However, suppose that the actual proportion of spam messages in these e-mail accounts is 10%. Compute the constants of the classification functions to account for this information.
e. A spam filter that is based on your model is used, so that only messages that are classified as nonspam are delivered, while messages that are classified as spam are quarantined. In this case, misclassifying a nonspam e-mail (as spam) has much heftier results. Suppose that the cost of quarantining a nonspam e-mail is 20 times that of not detecting a spam message. Compute the constants of the classification functions to account for these costs (assume that the proportion of spam is reflected correctly by the sample proportion).

Files Included in the Folder:

1. **Homework 6.pdf**
2. **R Code for Textbook Examples in Chap 11 12.pdf**
3. **Accidents.csv**
4. **EastWestAirlinesNN.csv**
5. **RidingMowers.csv**
6. **Spambase.csv**
7. **SystemAdministrators.csv**
8. **Tinydata.csv**
9. **ToyotaCorolla.csv**
10. **UniversalBank.csv**