Student Name\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Student ID.NO\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**The exam is composed of 5 multiple choice questions (10 points), 20 short answer questions (50 points), and 2 case problems (40 points). The exam is worth 100 points. Please use the attached answer sheet to write down all the solutions and submit only the answer sheet by April 16th at 8:00 am. Good luck!**

1. In preparing categorical variables for analysis, it is usually best to (2pts)

A) convert the categories to numeric representations.

B) convert the categories to binary, dummy variables.

C) combine as many categories as possible.

D) let them remain categorical.

2. k-means clustering is the process of (2pts)

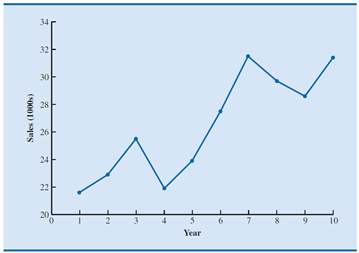
A) agglomerating observations into a series of nested groups based on a measure of similarity.

B) organizing observations into distinct groups based on a measure of similarity.

C) reducing the number of variables to consider in data-mining.

D) estimating the value of a continuous outcome variable.

3. A time series plot of a period of time (in years) versus sales (in thousands of dollars) is shown below. Which of the following data patterns best describes the scenario shown? (2pts)



A) Linear trend pattern

B) Nonlinear trend pattern

C) Seasonal pattern

D) Cyclical pattern

4. Data mining methods for classifying or estimating an outcome based on a set of input variables is referred to as (2pts)

A) supervised learning.

B) unsupervised learning.

C) dimension reduction.

D) data sampling.

5. A \_\_\_\_\_\_\_\_\_\_ classifies a categorical outcome variable by splitting observations into groups via a sequence of hierarchical rules. (2pts)

A) regression tree

B) scatter chart

C) classification tree

D) confusion matrix

6. What is the difference between moving averages and exponential smoothing? (2pts)

7. Using the concept of overfitting, explain why when a model is fit to training data, zero error with those data is not necessarily good. (2pts)

8. Write the R code that generates a sequence from -10 to 20 with an increment of 3. (2pts)

9. Write the R code that constructs a matrix with 5 rows that contain the number 1 up to 30. Your output matrix must be filled by the columns. (2pts)

10. Given the following confusion matrix, what is the accuracy? (2pts)

|  |  |  |
| --- | --- | --- |
| **Confusion Matrix** | | |
|  | **Predicted Class** | |
| **Actual Class** | **1** | **0** |
| 1 | 224 | 85 |
| **0** | 28 | 3,258 |

Use the Airfares.csv file to answer questions 11-15. Note, name the dataset as airfares.df

11. Write the R code that shows the data on the rows 10, 20, 30, 40, 50 and columns 1, 3, 5, 7, 9. (2pts)

12. Find the correlation coefficient between FARE and DISTANCE using R function. Output is necessary, R code is optional. (2pts)

13. Find the average DISTANCE of routes with a starting city of Chicago (S\_CITY) using R function. Output is necessary, R code is optional. (2pts)

14. Write the R code that conducts a principal components analysis of all the 10 numerical variables in the dataset airfares.df. Should the data be normalized? (2pts)

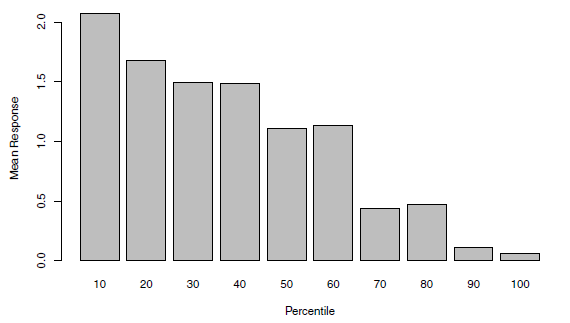
15. Create a side-by-side boxplot to measure the effect of Southwest Airlines’ presence (SW) on the average fare (FARE). Output is necessary, R code is optional. (2pts)

16. Demand for a product and the forecasting department’s forecast (naïve model) for a product are shown below. Compute the mean squared error. (3pts)

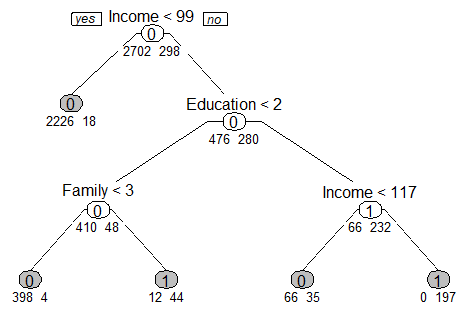
|  |  |  |
| --- | --- | --- |
| Period | Actual Demand | Forecasted Demand |
| 1 | 12 | - - |
| 2 | 15 | 12 |
| 3 | 14 | 15 |
| 4 | 18 | 16 |

17. Euclidean distance can be used to calculate the dissimilarity between two observations. Let *u* = (25, $350) correspond to a 25-year-old customer that spent $350 at Store A in the previous fiscal year. Let *v* = (53, $420) correspond to a 53-year-old customer that spent $420 at Store A in the previous fiscal year. Calculate the dissimilarity between these two observations using Euclidean distance. (3pts)

18. A firm that sells software services has been piloting a new product and has records of 500 customers who have either bought the services or decided not to. The target value is the estimated profit from each sale (excluding sales costs). The global mean is $2128. However, the cost of the sales effort is not cheap—the company figures it comes to $2500 for each of the 500 customers (whether they buy or not). The firm developed a predictive model in hopes of being able to identify the top spenders in the future. The lift decile charts for the validation set are shown below. Interpret the meaning of the first and second bars from the left. (3pts)



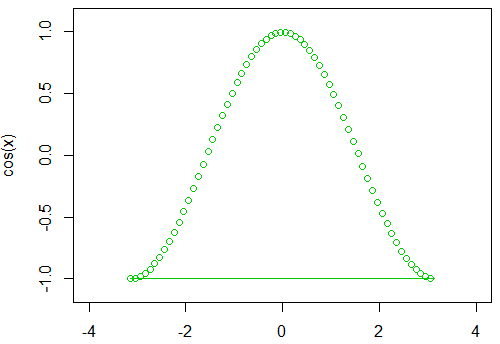
19. Universal Bank is interested in exploring ways of converting its liability customers to personal loan customers. A campaign the bank ran for liability customers showed a healthy conversion rate of over 9% successes. The bank’s retail marketing department wants to use these data to build a model that will classify customers as “accept” and “not accept”. This dataset includes data on 5000 customers. The data include customer demographic information (age, income, etc.), customer response to the last personal loan campaign (Personal Loan), and the customer’s relationship with the bank (mortgage, securities, account, etc.). The figure below shows the classification tree generated using the validation set. Write down the results in terms of rules. (3pts)



20. For each of the following variables, identify whether a variable is qualitative or quantitative by placing a check mark in the correct column: (3pts)

|  |  |  |
| --- | --- | --- |
| **Variable** | **Quantitative** | **Qualitative** |
| Your name |  |  |
| Your height |  |  |
| Your income |  |  |
| The month in which you were born |  |  |
| Your home address house number |  |  |
| The number of texts you send each day |  |  |
| The type of phone you have |  |  |
| Your satisfaction with your mobile service provider, measured on a scale of one to five |  |  |

21. Write the R code that generates the following figure. Hint: use cos() and lines() functions (3pts)



22. The following output is a generated logistic regression model for the financial condition of 20 banks. The outcome variable is Financial\_Condition which 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. 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.

Write the estimated equation that associates the financial condition of a bank with its two predictors in three formats: (3pts)

1. The logit as a function of the predictors
2. The odds as a function of the predictors
3. The probability as a function of the predictors

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -24.721 6.675 -2.205 0.0274 \*

TotExp.Assets 89.834 47.781 1.880 0.0601 .

TotLns.Lses.Assets 9.371 5.779 1.449 0.1474

Use the Utilities.csv file to answer questions 23-25. Note, name the dataset as utilities.df

23. Remove the column of Company in the dataset, normalize the data, and rename the row names of the data frame be the corresponding company names. Please name you final data frame as utilities.df.norm. Write the R code. (3pts)

24. Apply hierarchical clustering with Complete Linkage method to utilities.df.norm. Plot the dendrogram. Write the R code and show the generated dentrogram. (3pts)

25. From the dendrogram in Question 24, how many clusters will be created if we choose a cutoff distance of 5.7. Compare different clusters using a heatmap and describe each cluster in this context (e.g., utilities with high fixed charge and fuel cost, low cost and RoR). No need to write the R code. (3pts)

26. A company that manufactures riding mowers wants to identify the best sales prospects for an intensive sales campaign. In particular, the manufacturer is interested in classifying households as prospective owners or nonowners on the basis of Income (in $1000s) and Lot Size (in 1000 ft2). The marketing expert looked at a random sample of 24 households, given in the file RidingMowers.csv. For questions a-f, include the R output, don’t include the R code. Write down your entire R code to answer question g. (15pts)

1. Create a scatter plot of Income vs. Lot Size using color or symbol to distinguish owners from nonowners. From the scatter plot, which class seems to have a higher average income, owners or nonowners? (2pts)
2. Use all the data to fit a logistic regression of ownership on the two predictors. Interpret the meaning of the coefficient for Income. (2pts)
3. Create a confusion matrix to evaluate model performance. Among nonowners, what is the percentage of households classified correctly? (2pts)
4. To increase the percentage of correctly classified nonowners, should the cutoff probability be increased or decreased? (2pts)
5. What are the odds that a household with a $60K income and a lot size of 20,000 ft2 is an owner? What is the classification of a household with a $60K income and a lot size of 20,000 ft2 using cutoff = 0.5. (2pts)
6. What is the minimum income that a household with 16,000 ft2 lot size should have before it is classified as an owner? (2pts)
7. Write down your entire R code here. (3pts)

27. The Research and Innovative Technology Administration’s Bureau of Transportation Statistics conducted a study to evaluate the impact of the September 11, 2001 terrorist attack on US transportation. The goal of the study is to provide a greater understanding of the passenger travel behavior patterns of persons making long distance trips before and after 9/11.

The study analyzes monthly passenger movement data between January 1990 and May 2004. Data on three monthly time series are given in file Sept11Travel.csv for this period: (1) Actual airline revenue passenger miles (Air), (2) rail passenger miles (Rail), and (3) vehicle miles traveled (Car).

In order to assess the impact of September 11, BTS took the following approach: using data before September 11, they forecasted future data (under the assumption of no terrorist attack). Then, they compared the forecasted series with the actual data to assess the impact of the event. Our first step, therefore, is to split each of the time series into two parts: pre- and post September 11. We now concentrate only on the earlier time series. (25pts)

1. Partition the dataset into training (pre-event) and validation sets. Use the recent 32 records as validation set. Plot the pre-event AIR time series. What time series components appear? Hint: create a time series object for each mode of transportation (Air, Rail, Car). (2pts)
2. Fit a regression model with exponential trend and seasonality for the AIR series. Remember to use only pre-event data. Write down the summary of output. (3pts)
3. Interpret the coefficients for February and August. Are they statistically significant? (2pts)
4. Create an ACF (autocorrelation) plot of the regression residuals. What does the ACF plot tell us about the regression model’s forecasts? How can we improve the model? (3pts)
5. Fit a regression models with quadratic trend and seasonality for the Rail series. Remember to use only pre-event data. Write down the summary of output. (3pts)
6. Fit a regression models with linear trend and seasonality for the Car series. Remember to use only pre-event data. Write down the summary of output. (3pts)
7. For each series (Air, Rail, Auto), plot the complete pre-event and post-event actual series overlayed with the predicted series. (3pts)
8. What can be said about the effect of the September 11 terrorist attack on the three modes of transportation? Discuss the magnitude of the effect, its time span, and any other relevant aspects. (2pts)
9. Write down your entire R code here. (4pts)