|  |  |
| --- | --- |
| Part 1 (36) |  |
| Part 2 (8) |  |
| Part 3 (56) |  |
| Total (100) |  |
| Do not mark - for grading purposes only! | |

**ITEC 620**

AMERICAN UNIVERSITY

MIDTERM EXAM (**PRACTICE**)

INSTRUCTIONS

There are three parts of the exam, totaling 100 points:

9 Multiple Choice questions (4 points each)

1 Short Essay question (8 points)

4 Problems (8, 20, 8, and 20 points)

You may use one double-sided sheet of notes only. All electronic devices are prohibited, except for a basic calculator (scientific or graphing calculator is fine).

Exams are randomized in several different ways. You will not have exactly the same exam as the people around you.

Good luck!

***Please note that this practice exam is NOT intended to be a carbon copy of the actual midterm. Its purpose is to give you a sense of the length, scope, and format to expect.***

**Part 1: Multiple Choice Questions**

1. Which of the following is NOT part of the analytics cycle?

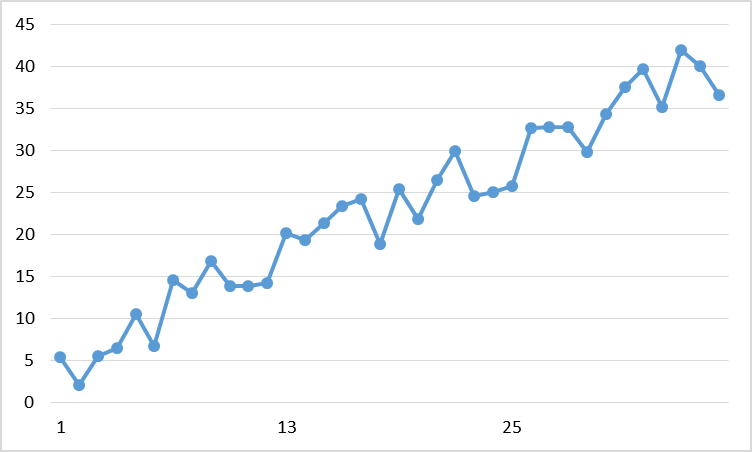
A) Formulate the question/problem  
B) Identify and collect the available data related to the question  
C) Determine what outcome you hope to find  
D) Write up conclusions and recommendations

**C**. The other three are all steps in the analytics cycle, and important parts of any analysis that involves learning from data.

2. Which of the following is an example of descriptive analytics?

A) Determining the correlation between daily website traffic and number of server outages  
B) Finding the best possible server configuration  
C) Forecasting the number of website visitors between 12-4 AM on Thursday  
D) Predicting which banner ad a user will be the most likely to click

**A**. B is prescriptive, C and D are predictive.

3. Which of the following would be the most appropriate forecasting method for the time series shown in the chart below? (The time series contains monthly data.)  
  


A) Single exponential smoothing  
B) Double exponential smoothing  
C) Holt-Winters  
D) Moving average

**B**. A and D do not capture the clear increasing trend in the data. C also includes seasonality, which does not appear to be present in the data.

4. Which of the following is generally true about large data sets (say, with *n*=8000)?

A) They cannot be analyzed without cloud computing and professional versions of software.  
B) Because the sample size is large, we can assume that no biases exist in the data.  
C) Traditional descriptive statistics can detect all relevant patterns and relationships in the data.  
D) Some analytics methods are more useful with these datasets than they would be with less data.

**D**. A data set of this size is easily manageable on a personal computer in Excel/R/other free software. Bias is inherent to how the data points are measured/collected/recorded, and is not mitigated by sample size. Traditional descriptive statistics methods were not designed for very large data sets; they are still useful, but not comprehensive. D is correct; this is part of the reason for the increased demand for analytics!

5. X and Y have a correlation of 0.8. Which of the following is true?

A) When X is higher than average, Y tends to be higher than average  
B) When X is higher than average, Y tends to be lower than average  
C) X and Y are not related in any substantial way  
D) Changing X will cause Y to change as well

**A**. This is exactly what positive correlation means. B describes negative correlation, C would imply ~0 correlation, and D describes causality, which can never be concluded from correlation.

6. Which of the following is true of outliers?

A) Every data set has at least one outlier  
B) An outlier is usually a data error  
C) We can remove outliers without affecting the result of analytics methods  
D) Outliers affect the mean of a variable more than the median

**D**. Data sets do not have to have outliers, but often do, even when all data points are correct. Outliers are a challenge specifically because they can unduly influence results.

7. What does the command: “Salaries[8]” do in R? (You can assume the Salaries object contains a vector of salaries.)

A) Display the first eight salaries  
B) Display the eighth salary  
C) Delete all salaries after the eighth one  
D) Set the value of Salaries to eight

**B**. That’s what the square bracket notation means in R.

8. Which of the following is true of lift ratios in association rules?

A) A lift ratio close to zero indicates no noticeable relationship between the premise and the conclusion  
B) A lift ratio of 0.8 means that, given the premise, the probability of the conclusion is 0.8  
C) A lift ratio of 3 means that the premise makes the conclusion 3 times more likely  
D) A lift ratio of 1 means that the conclusion is true for all of the data points

**C**. A lift ratio close to 1 indicates no relationship. B is describing confidence (not lift ratio).

9. Which of the following is NOT a benefit of partitioning a time series to compare forecasting models?

A) We can compute a measure of predictive accuracy for each model  
B) It makes it less likely that we will end up making predictions with an overfit model  
C) It increases the number of data points that we can use to build each model  
D) We can evaluate the models based on data points that were not used to create them

**C**. The only real downside of partitioning is that fewer data points are used to build each model. (For large data sets, that is typically an acceptable sacrifice.)

**Part 2: Short Essay Questions**

1. (8 pts.) Choose any real-world organization. Describe clearly and specifically two ways in which this organization could use descriptive analytics methods to perform more efficiently or effectively.

(Almost any organization would work, and there are countless ways in which descriptive analytics could be used. The following is an example of an answer that would receive full credit.)

Target could use descriptive analytics in many different ways. I provide two examples here. First, given a database of transactions, they could offer specific coupons to customers who have purchased an item that is the premise of an association rule with a very high lift ratio. Second, they could cluster their frequent customers based on purchasing habits, and develop different sets of promotional materials and discounts tailored individually to each cluster.

**Part 3: Problems**

**1**. (8 pts.) We have observed the following time series (in chronological order):

18, 26, 26, 16, 30, 27, 30

a) (4 pts.) Using a 3-period moving average, what is the prediction for the next value in the time series?

**29**. (30 + 27 + 30) / 3

b) (4 pts.) Using a single exponential smoothing model with alpha=0.5, the prediction for the most recent period was 26. What is the prediction for next period?

(26 + 30) / 2 = **28**

**2**. (19 pts.) The following is a portion of association rules output in R for transactions at a local coffee shop:

lhs rhs support confidence lift

[1] {Sandwich} => {Drip Coffee} 0.02 0.20 0.80

[2] {Muffin} => {Drip Coffee} 0.08 0.40 1.60

[3] {Latte} => {Sandwich} 0.01 0.05 0.50

[4] {Drip Coffee} => {Muffin} 0.08 0.32 1.60

[5] {Muffin} => {Latte} 0.06 0.30 1.50

a) (4 pts.) If a customer bought a sandwich, what is the probability that that customer also bought a drip coffee?

**0.20**. (The confidence of the “If Sandwich, then Drip Coffee” rule)

b) (4 pts.) Is the overall proportion of customers who bought drip coffee higher or lower than the answer to part a?

**Higher**. The lift ratio of the “If Sandwich, then Drip Coffee” rule is less than 1, meaning a customer who bought a sandwich is less likely than an average customer to buy a drip coffee.

c) (4 pts.) What is the support of “Latte”? That is, what proportion of overall customers bought a latte? (Hint: use the last rule in the table)

**0.20**. The question is asking for P(latte). The confidence of the last rule in the table tells us that P(latte | muffin) = 0.30. The lift ratio of that rule tells us that P(latte | muffin) = 1.5\*P(latte). That is, 0.30 = 1.5\*P(latte), or P(latte) = 0.30 / 1.5 = 0.20.

d) (4 pts.) While most items can be made quickly, lattes and sandwiches each require substantial individual attention from an employee. Why is the low lift ratio of the third association rule in the table helpful, from the perspective of a manager running the coffee shop? (Please limit your answer to 40 words or less.)

It shows that despite sandwiches and lattes being reasonably popular individually, they are very rarely ordered together. That means very few orders will require more than one time-consuming item.

e) (4 pts.) If there are 1000 transactions in the data set, how many of these transactions include both a muffin and a latte?

**60**. The support of the 5th rule is 0.06, meaning 6% of transactions include both the premise and the conclusion. 1000\*0.06 = 60.

**3**. (8 pts.) The following output shows a centroid table produced by running k-means clustering in R. Each data point is a company for which we have three variables: the total number of employees at the company, the valuation (estimated total $ value) of the company, and the number of countries in which the company has at least one office.

Number of Employees Valuation ($B) Global Presence (# of countries)

1 3300 0.61 8.32

2 51700 50.37 137.85

3 46500 48.24 16.94

4 17200 16.22 46.61

a) (4 pts.) Briefly (1-2 sentences), describe the type of company that would be in Cluster 2.

Cluster 2 contains very large global companies. They have many employees, are valued very highly, and have a presence in most countries.

b) (4 pts.) American Conglomerates is one of the largest companies in the data set, but has no offices outside North America. Which cluster is it in?

**Cluster 3**. That’s the cluster with lots of employees and high valuations, but very low global presence.

**4**. (20 pts.) Each of the following blocks of R code is not functioning as intended due to one line with a mistake. For each one, identify the line of code with a mistake, and rewrite it correctly.

a) (5 pts.) Running k-means clustering with *k*=4 on a dataset of athletic conferences:

Conferences.norm <- scale(Conferences)

set.seed(12345)

Conferences.kmclusters <- kmeans(Conferences, 4, nstart=10)

Conferences.kmclusters <- kmeans(Conferences.norm, 4, nstart=10)

(The kmeans function was not using the normalized data.)

b) (5 pts.) Calculating the mean of the first five columns of a customer dataset:

for (i in 1:5) {

col\_mean <- mean(Customers[i,])

print(paste(“The mean of column”,i,”is”,col\_mean))

}

col\_mean <- mean(Customers[,i])

(i should refer to the column, not the row.)

c) (5 pts.) Creating and sorting a set of association rules for a recipes dataset:

rules <- apriori(recipes.binary)

rules.sorted <- sort(rules, by=”lift”)

inspect(recipes.binary)

inspect(rules.sorted)

(The inspect function should be used on the rules, not on the dataset itself)

d) (5 pts.) Creating a single exponential smoothing model for a GDP time series of annual data starting in 1950:

GDP\_ts <- ts(GDP,start=1950)  
GDP\_ts\_SES <- HoltWinters(GDP\_ts)

GDP\_ts\_SES <- HoltWinters(GDP\_ts, beta=FALSE, gamma=FALSE)

(The Holt-Winters function needs to be told not to use the beta and gamma  
parameters if we want to do single exponential smoothing.)