CS422 Data Mining

Assignment 1

1 Exercises (total points 3)

1.1 Tan, Chapter 1 (1 point divided evenly among the questions) Besides the lecture, make sure you read Chapter 1. After doing so, answer the following questions at the end of the chapter: 1, 3.

Question 1 *Discuss whether each of the following activities are data mining task*

a) Dividing the customers of a company according to their gender.
Ans: No. It's not a data mining task, it is possible by using sql with only one condition

b) Dividing the customers of a company according to their profitability.
Ans: We are not predicting anything by using current data, so it is not a Data Mining task.

c) Computing the total sales of a company.
Ans: We are computing here and not predicting anything so it's a mathematical problem and not Data Mining task.

d) Sorting a student database based on student identification numbers.
Ans: Not a Data Mining job, we can do it using simple sql query.

e) Predicting the outcomes of tossing a (fair) pair of dice.
Ans: If we use past data to predict the outcomes then it is going to be Data Mining task, even we can solve this problem using mathematics.

f) Predicting the future stock price of a company using historical records.
Ans: It is a perfect example of data mining task as we are using past data to predict outcomes of future event.

g) Monitoring the heart rate of patient for abnormalities.
Ans: Yes, it is a task of data mining, with the help of past data we can find the abnormalities in the graph

h) Monitoring the seismic waves for earthquake activities.
Ans: Yes, it is task of data mining, by using past data we can find the abnormalities in the waves to predict earthquake.

i) Extracting the frequencies of sound wave.
Ans: No it is a science problem which can be solved using simple mathematics.

Question 3 *For each of the following Data Sets explain whether data privacy is an important issue.*

a. Census data collected from 1900–1950.

Ans: It's a general type of data rather than personal data so data privacy is not important.

b. IP addresses and visit times of Web users who visit your Website.

Ans: It is possible to identify or hack systems using IP address, so data privacy is an important issue for given data set.

c. Images from Earth-orbiting satellites.

Ans: It's a general type of data so data privacy is not important in this case.

d. Names and addresses of people from the telephone book.

Ans: Telephone book is the public tool to get information about other people so data privacy wont matter in this case.

(e) Names and email addresses collected from the Web.

Ans: It depends on the web site from which you have collected the data, if it is healthcare domain or government official’s site then data privacy is an important issue.

1.2 Tan, Chapter 2 (1 point divided evenly among the questions) Besides the lecture, make sure you read Chapter 2, sections 2.1 – 2.3. After doing so, answer the following questions at the end of the chapter: 2, 3, 7, 12.

Question 2 *Classify the following attributes as binary, discrete, or continuous. Also classify them as qualitative (nominal or ordinal) or quantitative (interval or ratio). Some cases may have more than one interpretation, so briefly indicate your reasoning if you think there may be some ambiguity.*

Sr	Problem	Type	Reason
1	Time in terms of AM or PM.	Qualitative, Ordinal, Binary	a. Time as PM or AM so no number b. Order is important c. Possible values only two
2	Brightness as measured by a light meter.	Quantitative, Continuous, ratio	a. Number b. As quantity varies from one observation to another c. It can be fraction
3	Brightness as measured by people’s judgments	Qualitative, Ordinal, Discrete	a. Not a number b. Order exist c. Set is limited
4	Angles as measured in degrees between 0◦ and 360◦.	Quantitative, Continuous, ratio	a. Number to measure b. As quantity varies from one observation to another c. It can be fraction
5	Bronze, Silver, and Gold medals as awarded at the Olympics	Qualitative, Ordinal, Discrete	a. Not a number b. Order exist c. Set is limited
6	Height above sea level.	Quantitative, Continuous, ratio	a. Number to measure b. As quantity varies from one observation to another c. It can be fraction
7	Number of patients in a hospital.	Quantitative, ratio, Discrete	a. Number to measure b. It can be fraction c. Set is limited

8	ISBN numbers for books.	Qualitative, Nominal, Discrete	a. Not a number b. No order c. Set is limited
9	Ability to pass light in terms of the following values: opaque, translucent, transparent.	Qualitative, Ordinal, Discrete	a. Not a number b. Order exist c. Set is limited
10	Military rank.	Qualitative, Ordinal, Discrete	a. Not a number b. Order exist c. Set is limited
11	Distance from the center of campus.	Quantitative, ratio, Continuous	a. Number to measure b. It can be fraction c. As quantity varies from one observation to another
12	Density of a substance in grams per cubic centimeter.	Quantitative, ratio, Discrete	a. Number to measure b. It can be fraction c. Set is limited
13	Coat check number.	Qualitative, Nominal, Discrete	a. Not a number b. No order c. Set is limited

Question 3 *You are approached by the marketing director of a local company, who believes that he has devised a foolproof way to measure customer satisfaction. He explains his scheme as follows: "It's so simple that I can't believe that no one has thought of it before. I just keep track of the number of customer complaints for each product. I read in a data mining book that counts are ratio attributes, and so, my measure of product satisfaction must be a ratio attribute. But when I rated the products based on my new customer satisfaction measure and showed them to my boss, he told me that I had overlooked the obvious, and that my measure was worthless. I think that he was just mad because our best-selling product had the worst satisfaction since it had the most complaints. Could you help me set him straight?"*

A. *Who is right, the marketing director or his boss? If you answered, his boss, what would you do to fix the measure of satisfaction?*
   Ans: His boss is correct. Because we cannot consider number of complaints only to measure customer satisfaction, we must consider the quantity of products sold then only we can conclude the customer satisfaction. We can define customer satisfaction as number of complaints per total number of products purchased. Let's understand this using one example, suppose products Prod1 & Prod2 are manufactured by company and quantities are 30 & 100. Now number of complaints received for Prod1 & Prod2 are 5 & 10. If we consider number of complaints, then customers are more satisfied with product Prod1 but if we want to calculate actual measure of satisfaction then
   For Prod1 = 5/30 = 0.16
   Prod2 = 10/100 = 0.1

Which says measure of satisfaction for Prod2 is greater than Prod1, that's how we can conclude that marketing director is wrong.

B. What can you say about the attribute type of the original product satisfaction attribute?
Ans: Attribute types can be of nominal, ordinal, ratio or interval. Attribute used here is a ratio a quantitative attribute which I think is correct as it is used to measure the customer satisfaction.

Question 7 *Which of the following quantities is likely to show more temporal autocorrelation: daily rainfall or daily temperature? Why?*

Ans: Relation between attributes is called as correlation, if attributes are more closely to each other at different time then it is going to be more temporal correlation between two attributes. As rainfall can vary drastically at given location time to time, but if we consider daily temperature which have very less variation at two different times. So, I think daily temperature has more temporal autocorrelation than daily rainfall.

Question 12 Distinguish between noise and outliers. Be sure to consider the following questions.

Points	Answers
Definition	Meaningless information is Noise whereas Observation that is away from other observations is known as Outliers
Is noise ever interesting or desirable? Outliers?	Noise is not interesting and desirable, and Outlier can be desirable
Can noise objects be outliers?	Yes, Simple distortion of data because of noise can be outlier
Are noise objects always outliers?	No, noise objects cannot be always outliers
Are outliers always noise objects?	No, outliers cannot be always noise objects
Can noise make a typical value into an unusual one, or vice versa?	Yes, noise can make typical value into unusual one

1.3 ISLR 7e (Gareth James, et al.) (1 point divided evenly among the questions) Section 3.7 (Exercises), page 120: Exercises 1, 3, 4-a.

Question 1 *Describe the null hypotheses to which the p-values given in Table 3.4 correspond. Explain what conclusions you can draw based on these p-values. Your explanation should be phrased in terms of sales, TV, radio, and newspaper, rather than in terms of the coefficients of the linear model.*

Ans:

Below is table 3.4

	Coefficient	Std. Error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	<0.0001
TV	0.046	0.0014	32.81	<0.0001
Radio	0.189	0.0086	21.89	<0.0001
Newspaper	-0.001	0.0059	-0.18	0.8599

According to the information given in the table, it can be inferred that p-values for TV and Radio are very low so probability or chances to become sale = 0 are very less. For every 1000 dollars spent on TV and Radio sale is increased by 46 and 189 items respectively. But in case of Newspaper p-value is nearly equals to one and for every 1000 dollars spent on Newspaper sale is increased by only one which shows there is no relationship between Newspaper and Sales.

Question 3 *Suppose we have a data set with five predictors, X1 = GPA, X2 = IQ, X3 = Gender (1 for Female and 0 for Male), X4 = Interaction between GPA and IQ, and X5 = Interaction between GPA and Gender. The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get βˆ0 = 50, βˆ1 =20, βˆ2 = 0.07, βˆ3 = 35, βˆ4 = 0.01, βˆ5 = −10.*

(a) Which answer is correct, and why?

i. For a fixed value of IQ and GPA, males earn more on average than females.

ii. For a fixed value of IQ and GPA, females earn more on average than males.

iii. For a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is high enough.

iv. For a fixed value of IQ and GPA, females earn more on average than males provided that the GPA is high enough.

(b) Predict the salary of a female with IQ of 110 and a GPA of 4.0.

(c) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

Ans: It is an example of multiple linear regression with response is starting salary after graduation with five predictors, so our linear equation is

Expected value of y = B0 + B1*X1 + B2*X2 + B3*X3 + B4*X1*X2 + B5*X1*X3

Value of X3 is 1 for female and 0 for male, so our new equations are

Expected Y for males = 50 + 20*X1 + 0.07*X2 +0.01*X1*X2 -------------------------------- 1

Expected Y for females = 50 + 20*X1 + 0.07*X2 + 35 + 0.01*X1*X2 – 10*X1

= 85 +10*X1 + 0.07*X2 + 0.01*X1*X2 -------------------------------- 2

A → As two terms are same in above two equations, so we are considering first two terms only to compare.

Expected Y for males = 50 + 20*X1

Expected Y for females = 85 +10*X1

1. False because put values X1=1 and 4 in above equations.
2. False because put values X1=1 and 4 in above equations.
3. True if put values of GPA high enough then it's true.

4. False it is not true for higher values of GPA.

B → Referring equation 2 to calculate salary of female student

Expected Y for females = 85 +10*X1 + 0.07*X2 + 0.01*X1*X2

Here X1 = 4, X2 = 110

Expected Salary = 85 + 10 * 4 + 0.07 * 110 + 0.01 * 4 * 110

= 85 + 40 + 7.7 + 4.4 = 137.1 *thousands of dollars*

C → False. Here coefficient value is 0.01 which is very low, but it can be possible that variance is very low for the given fit.

Question 4 4. *I collect a set of data (n = 100 observations) containing a single predictor and a quantitative response. I then fit a linear regression model to the data, as well as a separate cubic regression, i.e. Y = β0 + β1X + β2X2 + β3X3 + ǫ.*

(a) Suppose that the true relationship between X and Y is linear, i.e. Y = β0 + β1X + ǫ. Consider the training residual sum of squares (RSS) for the linear regression, and the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

Ans: Need more details to comment on it as we cannot decide RSS for linear and cubic as relationship between X and Y is linear, but we can say that RSS for linear is lower than cubic regression.