**HW1 A20402686**

**Tan Chapter 1 -**

**Question 1 - Discuss whether or not each of the following activities is a data mining task.**

a) Dividing the customers of a company according to their gender.

Ans – It is not a data mining task because it is a just database query.

(b) Dividing the customers of a company according to their profitability.

Ans - It is not a data mining task because this is a mathematical problem.

(c) Computing the total sales of a company.

Ans - It is not a data mining task because this is a mathematical problem.

(d) Sorting a student database based on student identification numbers.

Ans - It is not a data mining task because it is a just database query.

(e) Predicting the outcomes of tossing a (fair) pair of dice.

Ans - It is not a data mining task because, this is a probability calculation as the die is fair. If the die wouldn’t have been fair, then this will be data mining task.

(f) Predicting the future stock price of a company using historical records.

Ans – Yes, it is a data mining task because it will need to create a model that can predict the value of the stock price.

(g) Monitoring the heart rate of a patient for abnormalities.

Ans – Yes, it is a data mining task because this will need a model which monitors behavior of heart rate and indicate us when an unusual heart behavior occurred. This is also called as anomaly detection.

(h) Monitoring seismic waves for earthquake activities.

Ans – Yes, it is a data mining task because, this will need a model which monitors behavior of seismic wave and indicate us when an unusual behavior of seismic wave occurred.

(i) Extracting the frequencies of a sound wave.

Ans - It is not a data mining task because it comes under signal processing field.

**Question 3 - For each of the following data sets, explain whether or not data privacy is an important issue.**

(a) Census data collected from 1900–1950.

Ans – No.

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

Ans – Yes.

(c) Images from Earth-orbiting satellites.

Ans – No.

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

Ans – No.

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

Ans – No.

**Tan chapter 2 –**

**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.**

**Example:** Age in years. **Answer:** Discrete, quantitative, ratio

(a) Time in terms of AM or PM.

Ans - Binary, qualitative, ordinal

(b) Brightness as measured by a light meter.

Ans - Continuous, quantitative, ratio

(c) Brightness as measured by people’s judgments.

Ans - Discrete, qualitative, ordinal

(d) Angles as measured in degrees between 0*◦* and 360.

Ans - Continuous, quantitative, ratio

(e) Bronze, Silver, and Gold medals as awarded at the Olympics.

Ans - Discrete, qualitative, ordinal

(f) Height above sea level.

Ans - Continuous, quantitative, interval/ratio (depends on whether sea level is regarded as an arbitrary origin)

(g) Number of patients in a hospital.

Ans - Discrete, quantitative, ratio

(h) ISBN numbers for books. (Look up the format on the Web.)

Ans - Discrete, qualitative, nominal (ISBN numbers do have order information, though)

(i) Ability to pass light in terms of the following values:

Ans - opaque, translucent, transparent. Discrete, qualitative, ordinal

(j) Military rank.

Ans - Discrete, qualitative, ordinal

(k) Distance from the center of campus.

Ans - Continuous, quantitative, interval/ratio (depends)

(l) Density of a substance in grams per cubic centimeter.

Ans - Discrete, quantitative, ratio

(m) Coat check number. (When you attend an event, you can often give

your coat to someone who, in turn, gives you a number that you can

use to claim your coat when you leave.)

Ans - Discrete, qualitative, nominal

**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?

In this case, **the boss is right**.

We have formula for measure of satisfaction as follows,

Satisfaction(product) = number of complaints for the product / total number of sales for the product*.*

(b) What can you say about the attribute type of the original product satisfaction attribute?

We cannot say about the attribute type of the original measure. Because two products having the same numbers of complaints may have different level of customer satisfaction and vice-versa.

**Question 7 - Which of the following quantities is likely to show more temporal autocorrelation:**

**daily rainfall or daily temperature? Why?**

Generally, temperatures are similar for geographically close locations, but rainfall is not because it can be localized, and this is the feature of spatial autocorrelation. Hence, daily temperature shows more spatial autocorrelation then daily rainfall.

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

(a) Is noise ever interesting or desirable? Outliers?

Ans – No.

(b) Can noise objects be outliers?

Ans – Yes, noise objects can be outliers because, random distortion of the data is responsible for outliers.

(c) Are noise objects always outliers?

Ans - No.

(d) Are outliers always noise objects?

Ans - No.

(e) Can noise make a typical value into an unusual one, or vice versa?

Ans - Yes.

**ISLR 7e Section 3.7 –**

**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 –**

The null hypothesis for TV –

In the presence of radio ads and newspaper ads, TV ads have no effect on sales.

The null hypothesis for radio –

In the presence of TV and newspaper ads, radio ads have no effect on sales.

The null hypothesis for newspapers –

In the presence of TV and radio ads, newspaper ads have no effect on sales.

Null hypotheses are false for TV and radio because of low p-values.

Null hypotheses are true for newspaper because of high p-values.

**Question 3 - Suppose we have a data set with five predictors, *X*1 =GPA, *X*2 = IQ, *X*3 = Gender (1 for Female and 0 for Male), *X*4 = Interaction between GPA and IQ, and *X*5 = 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 ˆ*β*0 = 50*,* ˆ*β*1 =20*,* ˆ*β*2 = 0*.*07*,* ˆ*β*3 = 35*,* ˆ*β*4 = 0*.*01*,* ˆ*β*5 = *−*10.**

**Ans –**

Y = 50 + 20 + 0.07 + 35 + 0.01 - 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.

Y = 50 + 20 k\_1 + 0.07 k\_2 + 35 + 0.01 - 10

male: 50 + 20 + 0.07 + 0.01

female: 50 + 20 + 0.07 + 35 + 0.01 - 10

Ans - Males earn more on average after reaching GPA high. (iii)

(b) Predict the salary of a female with IQ of 110 and a GPA of 4*.*0

Y = 50 + 20 \* 4 + 0.07 \* 110 + 35 + 0.01 - 10 \* 4 = 137.1

(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 - False. p-value of the regression coefficient must be examined.

**Question 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* = *β*0 + *β*1*X* + *β*2*X*2 + *β*3*X*3 + *\_*.**

**(a) Suppose that the true relationship between X and Y is linear, i.e. *Y* = *β*0 + *β*1*X* + *\_*. Consider the training residual sum of squares (RSS) for the linear regression, and also 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 -** Polynomial regression should be lower training RSS than the linear regression since it could have been made a more fit than data that matched with a wider irreducible error.