Exercise 1.1 – part 1

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.

Answer: Dividing the customers of a company according to their gender is not a data mining task. We can accomplish this task using a filter in Excel or a SQL query.

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

Answer: Diving the customers of a company according to their profitability is not a data mining task. We can accomplish this task by calculating the profit in excel and then filter the data according to the profit value or a SQL query.

c. Computing the total sales of a company.

Answer: Since it is a computation task, it does not come under the paradigm of data mining. Computing a numeric value can be achieved in various other ways. Hence, it is not a data mining task.

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

Answer: Even though at first it seems to be a clustering analysis type task, but it is not a data mining task as it can be easily completed using a sorting query in SQL.

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

Answer: As a fair dice is used, we can easily find the probable outcomes and hence, this task is not a data mining task.

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

Answer: Yes, predicting the future price of a company's stock using historical records is a data mining task. As the price is a continuous value attribute, we leverage the regression technique of predictive modeling to predict the price value.

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

Answer: Yes, monitoring the heart rate of a patient for abnormalities is a data mining task. As the goal of the task is to detect abnormalities, we can leverage the anomaly detection technique of data mining. We can train a model to learn the normal behavior of the heart rate and whenever unusual behavior is sensed, it should raise an alarm.

h. Monitoring seismic waves for earthquake activities.

Answer: Yes, monitoring seismic waves for earthquake activities is a data mining task. The end goal of this task is to detect abnormalities, we can leverage the anomaly detection technique of data mining. We can train a model to learn the behavior of normal ecosystem and whenever unusual disturbance in the behavior is sensed, it should raise an alarm.

i. Extracting the frequencies of a sound wave.

Answer: Since the task is a simple data extraction task which does not involve any prediction, it can be said that this task is not a data mining task.

Exercise 1.1 – part 2

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.

Answer: Census data that was collected in 1900 to 1950 will not have any data privacy issue. As it is used to only produce statistics results.

b. IP addresses and visit times of web users who visit your website.

Answer: By accessing IP addresses and number of times the web users visit their website comes under the data privacy issue as recording the IP address (detail through which a user’s identity can be made) and recording web activity of users is not legal.

c. Images from Earth-orbiting satellites.

Answer: Considering the images from different satellites which orbit the Earth, there is no data privacy issue as that data comprises of only earth’s images at different angles.

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

Answer: As a general tendency, people opt-in and provide their names and addresses in the telephone book for some purpose and hence this data set will not have any data privacy issues.

e. Names and email addresses collected from the Web.

Answer: As a general tendency, people enter their names and email addresses over the web for the purpose of connection and hence this data set will not have data privacy issues.

Exercise 1.2 – part 1

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.

Answer: Generally speaking, time should be Continuous, Quantitative, Interval, however if we consider time only in terms of AM or PM, it should be Discrete, Qualitative, Ordinal.

b. Brightness as measured by a light meter.

Answer: Brightness measured by a light meter will be Continuous, Quantitative, Ratio as the meter will read continuous readings and differences and ratios are meaningful.

c. Brightness as measured by people's judgments.

Answer: Brightness measured by people's judgements will be Discrete, Qualitative, Ordinal as people will be able to distinguish the brightness just by few values and those values can be ordered.

d. Angles as measured in degrees between 0 and 360.

Answer: Angles measured in degrees will be Continuous, Quantitative, Ratio as the angles values will be continuous and differences and ratios are meaningful.

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

Answer: As the questions says, the values are Discrete, Qualitative, Ordinal as these can be ordered.

f. Height above sea level.

Answer: Height above sea level can be Continuous, Quantitative, Ratio or Continuous, Quantitative, Interval. Interval because the difference in the height is meaningful and Ratio because the difference and ratio both are meaningful and the height above sea level has a true zero point where zero height means no height or absence of height.

g. Number of patients in a hospital.

Answer: Number of patients in a hospital will be Discrete, Quantitative, Ratio as the difference and ratio are meaningful.

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

Answer: ISBN numbers for books are Discrete, Qualitative, Nominal as any book could have gotten any number. This example is just like employee ID.

i. Ability to pass light in terms of the following values: opaque, translucent, transparent.

Answer: Ability to pass light will be Discrete, Qualitative, Ordinal as there is an order between the three.

j. Military rank.

Answer: As Military rank consists of order, it will be Discrete, Qualitative, Ordinal.

k. Distance from the center of campus.

Answer: Distance from the center of campus will be Continuous, Quantitative, Interval or Continuous, Quantitative, Ratio. Interval because the difference in the difference is meaningful and Ratio because the difference and ratio both are meaningful and the distance from the center of campus has a true zero point where zero distance means no distance covered.

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

Answer: Density can take any number and hence will be Continuous, 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.)

Answer: The coat check number is just given for distinguish and identification purposes and hence it is Discrete, Qualitative, Nominal.

Exercise 1.2 – part 2

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 bestselling 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?

Answer: Just by keeping the track of number of customer complaints for the each of the product sold will not take into consideration about the quantity of each product sold. Hence, the boss of the marketing director is correct here as the marketing director overlooked the obvious thing about number of products sold. In order to fix the measure of satisfaction, we will have to take number of products sold into account and for a better customer satisfaction work according to these considerations and analyze the number of complaints in accordance with it.

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

Answer: The original product satisfaction attribute is ambiguous as we are not given any details about how the number of complaints relate to number of products and satisfaction. Considering if 100 units of product 1 were sold and had 10 complaints, while 20 units of product 2 sold had the same number of complaints as product 1. In our example, the product satisfaction is better in product 1 but to arrive at such conclusion for the whole scenario described in the question, we need some more details.

Exercise 1.2 – part 3

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

Answer: As per the definition of autocorrelation, if two measurements are closer in time, we consider those measurements as temporally autocorrelated. Per this definition, daily temperature will show more temporal autocorrelation than daily rainfall. The reason for this discussion is that at any time and location, the temperature will be similar in two closely located areas where as the same cannot be said for rainfall as even two closely located areas can have different amounts of rainfall.

Exercise 1.2 – part 4

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

a. Is noise ever interesting or desirable? Outliers?

Answer: The definition of noise itself mentions that noise is a random and unintended component of a measurement error. We always try to eliminate the noise and focus on robust algorithms in data mining as the process of eliminating the noise is difficult. On the other hand, outliers are part of the actual data, but they show unusual characteristics. On comparing noise and outliers, noise is the undesired component and outliers are the desired component.

b. Can noise objects be outliers?

Answer: Yes, as noise is the random distortion. In an actual data set, we can witness some outliers which in actual were the good data points but distorted by noise and hence the actual data of those data points is lost in noise and only noise remains for those data points.

c. Are noise objects always outliers?

Answer: No, it’s not necessarily true that the distortion of noise will always make the data point an outlier. Since, the distortion is random, a normal data point can also have noise component and still behave like a normal/noise data point.

d. Are outliers always noise objects?

Answer: No, it’s not necessarily true that all the outliers are noise objects. Even though noise can cause actual data point to behave like an outlier, but there can be actual outlier data points which show unusual characteristics naturally.

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

Answer: Since it is a random distortion, how the noise will affect the actual data will also be a random effect. The answer to our question is yes, noise can turn a typical value into an unusual one and also it can happen that an unusual data point after noise distorts it come up in the area where all the typical data points lie.