CS 422 - Data Mining - Homework 1

Imaduddin Sheikh - isheikh@hawk.iit.edu

Due Date: **2/5/2022 11:59:59 PM**

1 Exercises

1.1. Tan, Chapter 1

Q.1)

a) This is not a data mining task as customers can be easily picked based on their gender using database queries.

b) It depends. This activity does not look like a data mining task but rather a financial/accounting calculation performed by the company, which is followed by the application of a threshold. However, it would have been a data mining task if a new customer's profitability had to be predicted.

c) This is clearly not a data mining task but rather a financial/ accounting task of the company. This result of this task can easily be obtained by querying through the database and applying the summation aggregation function.

d) Not a data mining task as it is a straightforward database query for sorting data based on the identification number of the students.

e) Since both the die are fair, mathematical calculation is needed to evaluate the probability for this task. There is not a data mining task.

f) This is a data mining task because we would be developing a model that predicts the continuous value of the stock price after training the model with the historical data. The area of data mining this task deals with is 'Predictive Modeling'.

g) This is a data mining task because a model is required that monitors the normal behavior of a heart beat and raises an alarm when it detects an abnormality while monitoring the heart rate. The area of data mining this task deals with is 'Anomaly Detection'.

h) This is a data mining task because a model is need to be built that monitors various types of seismic behavior associated with earthquakes and trigger a notification when one of the seismic activity is detected. This area of data mining can be 'Classification' (or 'Anomaly Detection').

i) This is not a data mining task. Extracting and producing frequency of sound waves are part of signals processing.

Q.3)
a) Yes and no because census data concerns with population in general and the relevant details at a particular time which helps the government with policy-making. However, census information contains information of an individual such as financial information, ethnicity, race, familial background etc, so government must protect it.

b) Yes because the IP addresses and number of visits performed by an IP address on that website can be used to identify a user's behavior on the internet. There are good examples when a website is sometimes hacked and the data can be utilized by hackers for malicious intent.

c) No because the images don't capture any individual data that can be used to identify a person, and that can be used for malicious intent. However there can be concerns of data privacy if the Earth-orbiting satellites capture more detailed data with improved resolution then these satellites can capture faces, person movement, etc. which can increase concerns for data privacy.

d) No because names and addresses of people on the telephone book are available publicly, and don't contain any confidential data.

e) No because names and email addresses of people on the web is such a data that users give willingly, and don't contain any confidential data. However, their email addresses can be a target for spammers.

1.2. Tan, Chapter 2

Q.2)
a) Binary, Qualitative, Ordinal.

b) Continuous, Quantitative, Ratio.

c) Discrete, Qualitative, Ordinal.

d) Continuous, Quantitative, Ratio.

e) Discrete, Qualitative, Ordinal.

f) Continuous, Quantitative, Ratio/Interval. It depends if sea level is considered as an arbitrary origin/zero or not.

g) Discrete, Quantitative, Ratio.

h) Discrete, Qualitative, Nominal.

i) Discrete, Qualitative, Ordinal.

j) Discrete, Qualitative, Ordinal.

k) Continuous, Quantitative, Interval. It depends on the calculation and how the variable value is expressed. However, the calculation requires that we take the difference between the location values(latitude and longitude) of the current distance and those of the center of campus to get the measurement.

l) Continuous, quantitative, Ratio.

m) Discrete, Qualitative, Nominal.

Q.3)

a) His boss is correct for pointing out that he has overlooked the obvious. The number of customer complaints filed have no essence if the director didn't take into account the number of sales of the product in question. One way to fix the satisfaction analysis is take into account the number of a product sold and the number of customer complaints filed on that particular product. One way to tackle this is to quantify the ratio of the number of complaints filed of a product to the total number of sales of that product.

However, considerations need to be taken when dealing with products with low number of product sales. For example, a product is only sold thrice, two out of three customers filed a complaint regarding that product. This predicament gives off a

66% complaint rate. Such measure is very unrealistic because three is not enough of a sample size to deduce that the product is in fact faulty. A solution to avoid this situation in the analysis is to include products that have meet a certain number-of-sales threshold.

b) The marketing director is correct about the product satisfaction attribute being a ratio attribute. However, the data set is biased because the attribute is not translated to a common scale domain.

Q.7)

Daily temperature will show better temporal auto correlation because the temperature in a particular location throughout the time is being recorded every unit time, and so has a continuous measure.
Same cannot be said for daily rainfall because rain is not continuous over the location, scattered being the other problem.

Q.12)

a) NO. Noise basically produces distortion in the signal, and makes the data appear more randomized. Hence, it is undesirable. On the other hand, outliers may be unusual and quite different from the rest of the data/signal. They might even be an error. But, it can be said that they are still worth considering for further analysis and may be desirable, unlike noise.

b) YES. Noise makes the data/signal more random than it really is. Noise points can are quite random, and can differ greatly from the true signal/data points to an extent that they can be considered as outliers.

c) NO. As noise objects are random, they don't necessarily differ greatly from the true data/signal. Noise objects can also be similar to the true data/signal itself. As a result, noise is not always an outlier.

d) NO. An outlier can never be considered as a noise object indefinitely because an outlier can be part of the true data/signal. An outlier may not seem to belong to a dataset/signal but it still has the potential to be desirable in some cases. Same cannot be said about noise.

e) YES. The source of noise makes some values in the dataset/signal more random or unusual. It transforms a true data object into an unusual or an object with different values through distortion. As it can distort a data point similar to the most of the other points in the dataset/signal to differ extensively(to be called an outlier). It can also distort a true outlier of the dataset/signal into a data point that is similar to the rest of the data points.