DATA MINING ASSIGNMENT 1

Name : Pavani Rangineni

CWID: A20516359

1.1.1 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. No

(b) Dividing the customers of a company according to their profitability. No

(c) Computing the total sales of a company. No

(d) Sorting a student database based on student identification numbers. No

(e) Predicting the outcomes of tossing a (fair) pair of dice. No

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

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

(h) Monitoring seismic waves for earthquake activities. Yes

(i) Extracting the frequencies of a sound wave. No

1.1.3 For each of the following data sets, explain whether or not data privacy is an important issue.

(a) Census data collected from 1900-1950. No

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

(c) Images from Earth-orbiting satellites. No

(d) Names and addresses of people from the telephone book. (e) Names and email addresses collected from the Web. No

1.2.2 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.

(a) Time in terms of AM or PM.

Nominal, Binary, Qualitative

(b) Brightness as measured by a light meter.

Ratio, Continuous, Quantitative

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

Ordinal, Discrete, Qualitative

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

Ratio, Continuous, Quantitative

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

Ordinal, Discrete, Qualitative

(f) Height above sea level.

Interval, Continuous, Quantitative

(g) Number of patients in a hospital.

Ratio, Discrete, Quantitative

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

Nominal, Discrete, Qualitative

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

Ordinal, Discrete, Qualitative

(j) Military rank.

Ordinal, Discrete, Qualitative

(k) Distance from the center of campus.

Interval, Continuous, Quantitative

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

Ratio, Continuous, Quantitative

(m) Coat check number.

Nominal, Discrete, Qualitative

1.2.3 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 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?
The boss is right in this situation when the marketing manager overlooks the obvious. The number of complaints is meaningless without taking into account the number of products purchased. Another consideration to consider is the size of the minimum number of products sold to allow for accurate analysis. Therefore, the minimum number of products sold must be considered for an accurate analysis.

(b) What can you say about the attribute type of the original product satisfaction attribute?
The original Product Satisfaction attribute of Count, which is a related attribute, is the correct analysis. Since each complaint count is not based on the same scale, the datasets cannot be compared, but it does provide a systematic sampling dataset. This analysis is equivalent to having a sample set of temperatures measured in Celsius, Kelvin, and Fahrenheit, and returns only numeric temperatures without converting all measurements to a common scale range.

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

Daily temperature will show more temporal autocorrelation. Because the daily temperature will have more values while compared to temperature.

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

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

Noise is not an important attribute. It makes the data to be more unusual as it distorts original values. Outliers are the objects of data as it's the main task in some cases. Thus, outliers can be interesting or desirable. But noise is neither interesting or desirable.

(b) Can noise objects be outliers? Yes

(c) Are noise objects always outliers? No

(d) Are outliers always noise objects? No

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