Assignment 1
Daniel Prusinski
CIS 435

1) Discuss whether or not the following activities are data mining tasks:
In general, we have two types of data mining tasks.

- Predict the value of a particular, not yet been seen, attribute based on the values of other attributes. Data mining tasks of this kind are called *labeled* and are known as *predictive* (supervised) learning.
- Derive a pattern that summarizes the underlying relationships in data. Data mining tasks that do not have any specially designated attributes are called *unlabelled* and are known as *descriptive* (unsupervised) learning.
    - Classification: used to predict group membership for data instances. Ie, Direct Marketing
    - Clustering: Groups of closely related observations belonging to the same cluster more similar to each other than observations belonging to other clusters. Ie, data points more similar and less similar.
    - Association Rule Discovery: predict the occurrence of an item based on occurrences of other items, given a set of records, each of which contain some number of items from a given collection.
    - Sequential Pattern Discovery: Rules predict sequential dependencies among different events using a given set of objects. Each object is associated with its own timeline of events.
    - Regression: Utilizes a data set in which the target values are known.
    - Deviation Detection: Detects significant deviations from normal behavior.

a. Dividing the customers of a company according to their gender
At first glance, this task is not a data mining task given that it revolves around information retrieval. But, if gender was not a known attribute it could be determined through cluster analysis which is a data mining task.

b. Dividing the customers of a company according to their profitability
I see this as an information retrieval based on threshold, and not a data mining task.

c. Computing the total sales of a company
This situation should be classified as finance or accounting and not data mining.

d. Sorting a student database based on student identification numbers
This task is not a data mining task given that it revolves around information retrieval, and is almost identical to circumstance “a”.

e. Predicting the outcomes of tossing a pair of dice
Tossing a pair of dice is an exercise in probability, and not a data mining task.

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

Predictive learning is being utilized through which target values are known and an unknown value is being predicted. Regression, time series, and logistic regression could be methods.

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

This question is very ambiguous, and I am going to break it into two parts.

If I was monitoring the heart rate of a patient for abnormalities:

This would be a trivial task given that I do not know what an abnormality would be. Furthermore, I would not have previous data established to assess an abnormality and diagnose the situation.

A professional monitoring the heart rate of a patient for abnormalities:

This clearly would be deviation detection for the following reasons -

- A professional would have past experience making the experience non-trivial, and thus a data set would exist for normal data and it would be easy to detect anomalies.

It could also be classification if an extensive data set exists on heart rates, and one could predict from monitoring the heart.

h. Monitoring seismic waves for earthquake activities

Situation "g" is very similar to this task. This clearly would be deviation detection for the following reasons -

- A professional would have past experience making the experience non-trivial, and thus a data set would exist for normal data and it would be easy to detect anomalies.

It could also be classification if an extensive data set exists on past seismic waves in association with an earthquake, and one could predict from monitoring the heart.

i. Extracting the frequencies of a sound wave

This situation should be classified as information retrieval and not data mining.

j. Using regression formula to predict the student's potential at admission office

Predicting the student's potential at the admission office based on a regression model utilizing past records and key performance indicators would be utilizing data mining techniques classification and regression depending on the model.

k. Detecting the purchase behavior of certain customers to find the abnormal purchasing

This would be anomaly detection, a data mining technique. I used this at US Bank to fight fraud.

l. Associating the relationship between frequently purchased items and preferred customers

In this instance, classification is demonstrated since preferred customer status is predicted through frequently purchased items for data instances. It should be noted that the

relationship between frequently purchased items (FPI) is already established but assigning the status of preferred customer (PC) is not yet done. The classification would look like this: FPI→PC From this association, a customer is classified based on frequently purchased items.

m. Predicting the species of a flower based on its characteristics

This could also fall under classification, and it sounds like decision tree would be appropriate.

2) Distinguish between noise and outliers by answering the following questions:

a. Is noise ever interesting or desirable?

The definition of noise is, "the random component of a measurement error and involves the distortion of a value or the addition of spurious objects (ITDM page 37)." From an analysts perspective, noise is not desirable seeing that it makes clear data opaque. Noise might have an interesting aspect to an analyst, but this would be an anomaly as opposed to the norm.

b. Is outlier ever interesting or desirable? – Page 40

An outlier is desirable or interesting for an analyst since it is different from most of the other data objects. The difference between an outlier and noise is an outlier can be an actual data object or value, thus it is valued.

c. Can noise objects be outliers?

A noise object can be in the form of an outlier, but upon further investigation it would be revealed that the object does not actually represent a real data object.

d. Are noise objects always outliers?

Noise objects are not always in the form of an outlier they can represent any form of an object, but noise is always a measurement error.

e. Are outliers always noise objects?

As discussed earlier, outliers can represent actual data objects and therefore are not always noise.

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

Seeing that noise is a measurement error, it can make a typical value into an unusual one, and vice versa. It is quite overwhelming to realize the distortions noise can have on data.

3) Classify the following attributes as binary, discrete, or continuous. Also, classify them as qualitative (normal or ordinal) or quantitative (interval or ratio). For example, age in years. Answers should be discrete, quantitative, and a ratio.

1. Binary, discrete, or continuous
2. Qualitative Quantitative
3. (normal or ordinal) (interval or ratio)

a. Time in terms of a.m. or p.m. (Binary, qualitative, ordinal)

b. Brightness as measured by a light meter - Continuous, quantitative, ratio

c. Brightness as measure by people's own judgment, such as bright, dim, dark –Discrete, qualitative, ordinal
d. Angles as measured in degrees between 0 and 360 –Continuous, quantitative, ratio
e. Olympics medals, such as bronze, silver, and gold – Discrete, Qualitative, ordinal
f. Height above sea level –Continuous, Quantitative, Interval (0 in this case is seen as an arbitrary point_
g. Continuous, Quantitative, Interval
h. Discrete, Quantitative, Ratio
i. ISBN of book – Discrete, Qualitative, Nominal
j. Military rank - Discrete, Qualitative, Ordinal
k Your feeling of pain using the scale of 1 to 10 - Discrete, Qualitative, Ordinal
l. Your IQ score - Discrete, Quantitative, Interval, Based on this definition "When current IQ tests are developed, the median raw score of the norming sample is defined as IQ 100 and scores each standard deviation (SD) up or down are defined as 15 IQ points greater or less, although this was not always so historically.[1] By this definition, approximately 95 percent of the population scores an IQ between 70 and 130, which is within two standard deviations of the median – Wikipedia"

m. Your score on standardized test, such as SAT, ACT, GRE – These are all graded differently according to my research. I will focus on the ACT –

- Students earn 1 point for each correct answer and neither lose nor gain points for each omitted or incorrect answer.
- A student's raw score for a section is calculated by determining the number of questions answered correctly in that section. Example: If a student answered 60 questions correctly in the English section, his English raw score would be 60.
- A student's raw score for a section is converted to a scaled score, which ranges between a 1 and a 36, with 36 being the highest possible score. Students receive a scaled score for each of the four multiple-choice test sections (English, Math, Reading, and Science).
- A student's composite ACT score is the average of the student's scaled scores for the four multiple-choice test sections. Example: If a student scored a 24 English, 28 Math, 26 Reading, and 23 Science, his composite ACT score would be (24 + 28 + 26 + 23)/4 = 25.25, which is rounded down to a 25.

This would be Discrete, Qualitative, Ordinal