CS 422 Homework0

Xingli Li, Illinois Institute of Technology

1. Exercises

Question 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,this is a simple database query.
```

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

```
No, this is a simple numerical calculation.
```

c. Computing the total sales of a company.

```
No, this is a simple numerical calculation.
```

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

```
No,this is a simple sort of data.
```

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

```
No, this is a probability problem.
```

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,this is signal processing.
```

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.

```
For example, the age of the population, the company served, the home address,
etc., which can provide a lot of important information.
```

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

```
We don't particularly know the people who visit the website, but we can mine more
valuable information through the location corresponding to the IP address, the
number of visits to the website, the length of stay, and the content of the
website itself.
```

c. Images from Earth-orbiting satellites.

```
In fact, earth orbiting satellites can collect geographic information of some
countries and use it for the bad side in the military and political field.
```

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

```
Through the names in the phone book, we can infer the user's gender, even the
person's age of birth. The address information is related to the house price
information. Indirectly, we can infer the income level and social status. Through
the mobile phone number, we can retrieve the user's social account and other
information from the Internet big data. These information are actually very
important and sensitive.
```

e. Names and email addresses collected from the Web.

```
Using the Internet search engine, we can search out more relation people's names
and e-mail address information, leave traces in the social network, and there may
be some messages and articles from users to indirectly infer people's interests
and hobbies.
```

Question 1.2

Exercise 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.

```
binary，qualitative，ordinal
```

b. Brightness as measured by a light meter.

```
continuous，quantitative，ratio
```

c. Brightness as measured by people’s judgments.

```
discrete， qualitative，ordinal
```

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

```
continuous，quantitative，ratio
```

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

```
discrete，qualitative，ordinal
```

f. Height above sea level.

```
continuous，quantitative，interval
```

g. Number of patients in a hospital.

```
discrete，quantitative，ratio
```

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

```
discrete，qualitative，nominal
```

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

```
discrete，qualitative，ordinal
```

j. Military rank.

```
discrete, qualitative, ordinal
```

k. Distance from the center of campus.

```
continuous, quantitative, interval/ratio
(this is relative distance.from other observation point to the center point,
which is interval. The comparison from each observation point to the center point
can be the ratio.)
```

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

```
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.)

```
discrete,  qualitative, nominal
```

Exercise 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?

```
Boss is right.
Because with the increase of product sales, the custmers complaint rate will
gradually appear and increase. We can quantitatively measure the product
experience with the satisfaction (percentage) of products in many aspects, which
is also convenient for subsidiaries to improve product design and quality.
```

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

```
I think it's worthless. Products sell well,and have more custmers and higher
exposure. To meet the needs of different groups of customers, there may be more
dissatisfied custmers.
```

Exercise 7

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

```
Daily temperature is more autocorrelated.
Because the temperature changes regularly every day, with the sun rising in the
East and setting in the West; Daily rainfall is seasonally related and has great
uncertainty.
```

Exercise 12

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

a. Is noise ever interesting or desirable? Outliers?

```
No, noise is not interesting, but outliers are different personalities, and we
may be more interested in research.
```

b. Can noise objects be outliers?

```
Maybe, noise is usually external interference, and data distortion may be caused
by outliers.
```

c. Are noise objects always outliers?

```
No.
```

d. Are outliers always noise objects?

```
No. Outliers do not necessarily cause interference to the system, but noise will.
```

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

```
Both, the specific value is uncertain.
```