EXERCISES FROM CHAPTER 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.

- Even though this looks a lot like database query, it might be a descriptive data mining model called clustering. We separate data according to similarities. In this case, it's about genders.

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

- This is similar to (a). The task doesn't seem to be related to data mining. Again, it looks like a database query. But just like in (a), it could be the case that the data is clustered for data mining purposes.

c. Computing the total sales of a company.

- No, this is not a data mining task. This is just making calculations. Here we don't use the data for predicting anything.

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

- No. This is a database query task. It's not data mining. We just sort information but we are not using it for data mining purposes.

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

- This could be seen as a data mining activity. Even though it could be seen as just a mathematical problem that needs a simple algorithm. For instance, if we collect the results of 100 tosses and use this data to predict the 101st toss, then we could say that this is a data mining activity. However, we could also predict the outcomes of tossing a fair pair of dice without collecting any data by simply using an algorithm.

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

- Yes. This is a data mining activity. Here we use the continuous data (the price of the stock changes in time with no limits like binary values) collected from the historical records to make a prediction. This is predictive modelling called regression.

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

- If we collect the data of each heartbeat (and use the heartbeat records of previous patients) and use this data for detecting abnormalities, i.e., data that is very different from previous data, then yes, this is data mining. This is called anomaly detection and it is a type of classification.

h. Monitoring seismic waves for earthquake activities.

- Yes, this is data mining. We can use data records of previous seismic waves to predict whether an earthquake is coming (and when it is coming). Classification might be used here as we can model previous seismic waves that led to an earthquake.

i. Extracting the frequencies of a sound wave.

- No, this is just collecting data in database.

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.

- Census information contains information about the number of people living in an area and additional details about those people. These details may contain a sensitive information. Despite that, the information of the people is protected by law, hence, it should not represent a data privacy issue.

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

- This should not be considered a data privacy issue because IP addresses only contain a general information about people's geolocation (e.g. postal code). The number of visit times is definitely not a sensitive information. Hence, this is not an issue.

c. Images from Earth-orbiting satellites.

- The images from Earth-orbiting satellites do not contain any personal information. Thus, it's not an issue.

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

- Yes. This could definitely be considered a data privacy issue since it could be used for malicious purposes. For instance, I wouldn't want to be disturbed by anyone. I would only give my address and name to people that I trust.

e. Names and email addresses collected from the Web.

- This doesn't seem to be a data privacy issue. Unless they have the password for my email address, there shouldn't be a reason to be concerned about this data collection.

EXERCISES FROM CHAPTER 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.

Example: Age in years. **Answer:** Discrete, quantitative, ratio

a. Time in terms of AM or PM. → Discrete, Quantitative, Interval
b. Brightness as measured by a light meter. → Continuous, Quantitative, Ratio
c. Brightness as measured by people's judgments. → Discrete, Qualitative, Ordinal;
- In my experience, I have mainly heard people saying: "It's not bright at all/bright/very bright." Hence, I chose the classification to be Discrete, Qualitative, Ordinal.
d. Angles as measured in degrees between 0 and 360. → Discrete, Quantitative, Ratio; (I always think of angles as ratios in a unit circle or as ratios of another smaller/bigger angle);
e. Bronze, Silver, and Gold medals as awarded at the Olympics. → Discrete, Qualitative, Ordinal; Gold > Silver > Bronze
f. Height above sea level. → Continuous, Qualitative, Interval (from 0 to some end point); It's continuous because technically you can go as far up as you want from the sea level.
g. Number of patients in a hospital. → Discrete, Qualitative, Ratio/Interval; It's physically impossible to have an infinite number of patients in the hospital, hence, we classify it as a discrete attribute.
h. ISBN numbers for books. (Look up the format on the Web.) → Discrete, Qualitative, Nominal;
- ISBN is the unique number of a book used to separate it from the other books. Hence, it's used as a qualitative nominal attribute. (book1 != book2)
i. Ability to pass light in terms of the following values: opaque, translucent, transparent. → Discrete, Qualitative, Ordinal;
- It could be classified as nominal since it's used to just separate them. However, one could say that opaque, translucent, and transparent are characteristic that could be put in an order opaque (not transparent) < translucent (has some transparency) < transparent; I prefer to classify it as an ordinal attribute;
j. Military rank. → Discrete, Qualitative, Ordinal;
k. Distance from the center of campus. → Discrete, Quantitative, Ratio/Interval; Not infinite, starts from 0, thus – Discrete, Quantitative, Ratio/Interval;
l. Density of a substance in grams per cubic centimeter. → Continuous (there is no limit), 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;

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?

- I agree with the boss. In order to fix the measure of satisfaction, you need to take into account that the best-selling product would be used by the most people. Hence, more people would have an opinion and the probability of someone leaving a complaint is much higher than the worst selling product that has almost no customers. → We can use a ratio $\underline{complaints \div \#\ of\ buyers\ }$. The result would be from 0 to 1. If it’s very close to 0, it means that there are almost no complaints => the product is good. If it’s close to 1, that means that there are a lot of complaints => bad product.

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

- The original product satisfaction attribute offered by the marketing director is classified as continuous, quantitative, ratio/interval. It’s continuous because there is no set limit on how many complaints you can have. We are working with quantities, and finally, we can work with the data as a ratio or an interval (- it depends on how we are going to use the data. e.g.: product a has 20 complaints more than product b; product a has 20% more complaints that product b => same information used differently);

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

Daily temperature would show more temporal autocorrelation as it is much more likely that we notice similar temperatures in a couple of consecutive days than to notice a similar amount of rainfall. Temperature is just more consistent (seasonally) than rainfall.

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

a. Is noise ever interesting or desirable? Outliers?

- Noise is neither interesting, nor desirable. It can negatively impact the generalization of the data. Noisy data is considered wrong data. Hence, it's much better to do models without this data. However, outliers are different. They can be both important and not that important depending on the context. For instance, if you try to find the mean of the SAT scores of 30 students, if 29 of them have about 1350 and only one of them has 1550, it wouldn't change much. Nevertheless, if we consider another example, outliers can be much more important. Example: Healthcare users represent 1% of the population in the USA, however, they account for over 20% of the money spent on healthcare.

b. Can noise objects be outliers?

- The definition of an outlier is an object that doesn't fit the general case. Hence, it can be a real object or it could be a mistake (a.k.a. noise object). Thus, noise objects can be outliers.

c. Are noise objects always outliers?

- Noise objects can be anything. They could be in the general case or they could be an outlier. Noise objects simply represent wrong data.

d. Are outliers always noise objects?

- Outliers are not always noise objects. For instance, most people leave about 10% tip in restaurants. Nevertheless, we can have the case of a rich and generous person who decides to leave a tip of 150% tip because they really liked the service and they can afford it. In the 2-dimensional data (bill, tip) this would be an outlier, however, it's a legitimate object, not a noise object.

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

- Yes. Noise can make a typical value into an unusual one, and vice versa. We can consider the example given in a) but instead of having an outlier, we can consider having a noise object. Thus, this noise object can significantly affect the data and make the results unusual. The opposite case is true as well – it could be the case that the noise object makes unusual data look normal.