Tiffany Wong
Fall 2021: Homework 0

1 Exercises **(4 points)**

1.1 *Tan, Chapter 1 (2 points divided evenly among the questions)*
Besides the lecture, make sure you read Chapter 1. After doing so, answer the following questions at the end of the chapter: 1, 3.

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.

This is not a data mining task because there would already be a record of what the customer's gender is and a data mining task is either predicting the value of a certain attribute or deriving patterns from existing data. Dividing the customers of a company by their gender doesn't result in any meaningful data in that sense.

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

Same with exercise 1a, this isn't a data mining task because it seems like the customer's profitability is already determined, so the only task being done here is dividing up customers based on an attribute that already exists.

c. Computing the total sales of a company.

This is not a data mining task because the total sales of a company are dependent on the values of other departments and those values already exist. It's not a data mining task because it doesn't predict anything meaningful or produce any patterns.

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

This is not a data mining task because there is no predicting based on other attributes being done here and there aren't any patterns that can be used from sorting student ID numbers.

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

This is not a data mining task because it does not predict the value of a particular attribute based on the values of other attributes. In this case, predicting the outcome of tossing a pair of dice is based on probability and not on any other attributes.

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

This is a data mining task because it predicts the value of a particular attribute based on the values of other attributes. In this case, the attribute being predicted, aka the target variable, is the future stock price of a company and the attributes used for making the prediction, aka the explanatory variables, are the historical records of the company's stock prices. It's a predictive task.

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

This is a data mining task because it aims to derives patterns that can summarize the underlying relationships in data. In this case, the heart of a patient is being monitored to see if there are any abnormalities that can be observed. This is anomaly detection.

h. Monitoring seismic waves for earthquake activities.

This is a data mining task because it aims to derives patterns that can summarize the underlying relationships in data. In this case, the seismic waves are being monitored to see if there are any trends or patterns that can be observed. This is predictive modeling.

i. Extracting the frequencies of a soundwave.

This is not a data mining task because it's not predicting the future frequencies of a soundwave based on other attributes and it's also not trying to derive any patterns to observe an underlying relationship. It's simply retrieving the frequencies of a soundwave.

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, this is not a data privacy issue because census data is used to create statistics and general changes in policies to begin with.

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

Yes, this is a breach of privacy because it allows others to know of other people to know where they're accessing a site, and the time information would also hinder the privacy of the user.

c. Images from Earth-orbiting satellites.

No, satellite images are publicly released to the public to view and observe.

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

No, people register and sign up with their information to be displayed on the telephone book, this is the task to be done to knowingly put their information in the book to display for others to access and see.

e. Names and email addresses collected from the Web.

No, if a name and email address is already on the web, then it's meant to be a contact information for that person.

1.2 *Tan, Chapter 2 (2 points divided evenly among the questions)*
Besides the lecture, make sure you read Chapter 2, sections 2.1 – 2.3. After doing so, answer the following questions at the end of the chapter: 2, 3, 7, 12.

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. Result of either AM or PM.

b. Brightness as measured by a light meter.

Continuous, quantitative, interval. Light meter will read the brightness with a scientifically accurate unit of measurement.

c. Brightness as measured by people's judgments.

Discrete, qualitative, ordinal. People's judgement will most likely be on a scale of very bright to not bright.

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. It's like an employee ID, meant to identify which book is what, but not actually any numerical meaning.

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

Discrete, qualitative, ordinal

j. Military rank.

Discrete, qualitative, ordinal. Rank is on a scale of who has most authority to least, meant to describe someone.

k. Distance from the center of campus.

Continuous, quantitative, interval. The distance changes from every point and is numeric.

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

Discrete, quantitative, interval

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. It's like an employee ID, meant to identify who to give the coat to, but not actually any numerical meaning.

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?

In this situation, the boss is right because the marketing director's method doesn't take into account the total number of customers than bought the product. If a product has more complaints, it could also mean that they are bought more often and then their complaints are more due to the amount of people buying the product to begin with.

In order to fix the measure of satisfaction, a good way to utilize the customer complaint attribute is to use it in tandem with the total number of purchases made for that product. A way to do so is to use ratios. For example, we would compare the percentage of customer complaints to number of times the product was bought so that we can get a better idea of what is the proportion of satisfied to unsatisfied customers. Also, another thing to consider are products that are very unpopular but no complaints on the site. Those products would have a 100% approval rating, but in reality, only a small number of people would be buying it. There would need to be a better defined cutoff point for when a product is considered to be comaprable to the best selling products.

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

The original product satisfaction attribute is indeed a ratio attribute. For ratio attributes, both difference and ratios are meaningful. With customer satisfaction, it's valuable to know the difference in customer complaints but, it doesn't make sense to only look at those numbers because they wouldn't put products into perspective. The ratios of total sales to complaints would give a better idea of overall customer satisfaction for a product.

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 locations that are close in distance will most likely have similar temperatures due to their distance from the sun being basically the same. On the other hand, daily rainfall is more local, where states in the same timezone could have varying rainfall amounts because it's more dependent on the climate and surroundings of the location. For example, Midwest states like Illinois and Indiana have the same-ish daily temperatures, but they wouldn't have the same daily rainfall.

12. Distinguish between noise and outliers. Be sure to consider the following questions.
    a. Is noise ever interesting or desirable? Outliers?

Noise in attributes is usually undesirable because it serves to change how the original attribute value looks. On the otherhand, outliers can be interesting because it helps to view the overall spread of the data. So, outliers can be desirable in viewing trends in the data, but noise only distorts the data.

    b. Can noise objects be outliers?

Noise is the random component of a measurement error, so in attribute values, it can make the data look more randomized or unusual. In the same way, noise can indeed be outliers.

    c. Are noise objects always outliers?

Noise objects aren't always outliers because it can by intermixed with non-noise points. This means that noise object data points can be part of normal data, so noise objects are not always outliers.

    d. Are outliers always noise objects?

Outlier points in a dataset can be an important part of the dataset that shouldn't be disregarded, so they are not always noise objects.

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

Noice objects involve the distortion of a value in a dataset, and it randomly appears. So noise can definitely make a typical value into an unusual one, and vice versa. The randomness of noise appearing is what makes it possible to turn a certain value into another 'type' of value (meaning typical to unusual and the other way around as well).