Tania Soutonglang

CS 422-01

September 7, 2022

Homework 1

Chapter 1

1. Discuss whether or not each of the following activities is a data mining task.
   1. Dividing the customers of a company according to their gender.  
      It would not be considered a data mining task since this could be accomplished with a database query input of ‘if male’ or ‘if female’.
   2. Dividing the customers of a company according to their profitability.  
      Calculating the profitability of a customer would not be a data mining task since that is more calculating rather than predicting the profitability.
   3. Computing the total sales of a company.  
      Since computing the total sales of a company is just a simple calculation, it would not be considered a data mining task.
   4. Sorting a student database based on student identification numbers.  
      This would not be a data mining task since this could also be solved with a database query to sort the identification numbers.
   5. Predicting the outcomes of tossing a (fair) pair of dice.  
      This would not be a data mining task since this could be solved with a probability calculation.
   6. Predicting the future stock price of a company using historical records.  
      This is a data mining task since a model to predict stock prices can be made from historical data.
   7. Monitoring the heart rate of a patient for abnormalities.  
      This would be a data mining task since the normal heart rate of a patient would be considered the base line model and any abnormalities would be compared to the base line.
   8. Monitoring seismic waves for earthquake activities.  
      This would be a data mining task since models based on various seismic waves would be needed to be able to create an alarm when dangerous seismic waves are detected.
   9. Extracting the frequencies of a sound wave.  
      This would not be a data mining task since extracting the frequencies would be a signal processing task.
2. For each of the following data sets, explain whether or not data privacy is an important issue.
   1. Census data collected from 1900–1950.  
      Data privacy would not be an issue here since the data would be too old to have much relevance today.
   2. IP addresses and visit times of web users who visit your website.  
      This would be a data privacy issue as IP addresses could be used as the location of the user and is therefore personal data.
   3. Images from Earth-orbiting satellites.  
      This is not a data privacy issue as the images from Earth-orbiting satellites are published to the public and would not be personal data.
   4. Names and addresses of people from the telephone book.  
      The telephone book is a public domain so the names and addresses of the people in the book would not be treated as confidential data.
   5. Names and email addresses collected from the Web.  
      The Web would be considered a public domain so names and email addresses would not be considered confidential data and therefore is not a data privacy concern.

Chapter 2.1, 2.2, 2.3

1. 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
   1. Time in terms of AM or PM.  
      Binary, qualitative, nominal
   2. Brightness as measured by a light meter.  
      Discrete, qualitative, ordinal
   3. Brightness as measured by people’s judgments.  
      Continuous, qualitative, nominal
   4. Angles as measured in degrees between 0 and 360.  
      Discrete, quantitative, ratio
   5. Bronze, Silver, and Gold medals as awarded at the Olympics.  
      Discrete, qualitative, ordinal
   6. Height above sea level.  
      Discrete, quantitative, interval
   7. Number of patients in a hospital.  
      Discrete, quantitative, ratio
   8. ISBN numbers for books. (Look up the format on the Web.)  
      Discrete, qualitative, nominal
   9. Ability to pass light in terms of the following values: opaque, translucent, transparent.  
      Discrete, qualitative, ordinal
   10. Military rank.  
       Discrete, qualitative, ordinal
   11. Distance from the center of campus.  
       Discrete, quantitative, ratio
   12. Density of a substance in grams per cubic centimeter.  
       Discrete, quantitative, ratio
   13. 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, ordinal
2. 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?”
   1. 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 believe the boss is correct since the number of complaints is a bad measurement since it does not consider the sales of that product. To fix the measure you would need to compare the ratio of complaints to total sales.
   2. What can you say about the attribute type of the original product satisfaction attribute?  
      The attribute type is correct.

7. Which of the following quantities is likely to show more temporal autocorrelation: daily rainfall or daily temperature? Why?  
Daily temperature would portray more temporal autocorrelation since the daily temperature of an area would be similar between each hour, but daily rainfall could vary drastically during an hour and even between locations within the same radius of area.

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

a. Is noise ever interesting or desirable? Outliers?  
Noise would not be interesting or desirable since it would be distracting and misleading for the other data. Outliers could be interesting or desirable since it could show how a certain situation could result in that outlier.

b. Can noise objects be outliers?  
Noise could be outliers since noise could have relevant data mixed in.

c. Are noise objects always outliers?  
Noise objects are not always outliers.

d. Are outliers always noise objects?  
Outliers are not always noise objects.

e. Can noise make a typical value into an unusual one, or vice versa?  
It is possible to make a typical value into an unusual one and vice versa because the randomized data from noise could make data seem unusual or normal.