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 activity can be a data mining task or not depending on the techniques used to discover the patterns between the data and the variable unknown (gender). If the gender of the customers is a known variable in the dataset, just sorting the dataset would solve it and it would not be considered data mining. However, when not knowing the gender, we can use a classical statistical approach, using simple regression models, considered to be data mining, or we could also use neural networks to discover patterns not so easy to see using traditional methods, considering this process to be data mining also.

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

- It is not considered data mining, as it can be solved using simple interactions with the database. relying only in traditional computer science techniques.

(c) Computing the total sales of a company.

- It is not considered Data mining, as this information can be obtained using simple operations from the database, such as computing the sum of sales in the period.

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

- It is not considered Data Mining, as It can be solved with a traditional sorting mechanism, such as merging.

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

- It can be considered a data mining task or not, depending on the process, if we try to predict the outcome theoretically, then it is a theoretic statistic task, however, if he have data about outcomes of tossing a pair of dice and other variables, we could discover how the different variables interact with the outcome, discovering new information in the process and it would be data mining.

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

- It is considered a Data Mining task, as we have to discover new information, not known before, related to our existing data, and using automatic information discovery mechanisms.

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

- It is considered Data Mining as it involves information discovery with automatic algorithms, in order to discover and predict the information, in this case patient abnormalities.

(h) Monitoring seismic waves for earthquake activities.

- It is considered Data Mining as it involves Information discovery algorithms that are not classical ones, and the relations between the input data and output are not known in advance.

(i) Extracting the frequencies of a sound wave.

- It is not considered Data mining, as it can be achieved with classical theory of waves and deterministic programming, we know how to define the frequencies and the impact that the data has in the frequency.

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.

- Data privacy is an important issue in this dataset, as the people that gave their information for the census, was not informed that it would be used for other purpose, however, if we use them anonymized, we could use this data without concerns.

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

- Data privacy is not an issue in this dataset, as people that enters a web page knows that they are leaving their IP address.

(c) Images from Earth-orbiting satellites.

- Data privacy is an issue for this dataset, as the information of what we see has not been asked for every person's property.

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

- Data privacy is an issue because these people were not asked if they wanted us to have their data for this purpose.

(e) Names and email addresses collected from the Web.

- Legally, I think that data privacy is not an issue, because these people must have been informed of the use we would give to their data when their addresses were collected. However, it is possible that they were not fully understanding what was going to be done with their data, so morally we should have some concerns.

Chapter 2, Data

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, interval

(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, ratio

3. What is aggregation? What are the motivations for aggregation? How are the values of attributes handled when aggregating data?

- Aggregation is the combining of two or more objects into a simple object. The motivations are mainly two, firstly, to get a broader picture of what is happening, as can be seen when aggregating daily sales into weekly sales, we get a bigger picture and reduce the relative noise. Secondly, when aggregating, we make smaller the quantity of the sample, allowing us to use algorithms that are more costly in terms of computing power.

The values of attributes are handled differently depending on the type of attribute, for example, quantitative attributes are summed or averaged, while qualitative are aggregated in groups.

7. A few months later, you are again approached by the same marketing director as in Exercise 6. This time, he has devised a better approach to measure the extent to which a customer prefers one product over other similar products. He explains, "When we develop new products, we typically create several variations and evaluate which one customers prefer. Our standard procedure is to give our test subjects all of the product variations at one time and then ask them to rank the product variations in order of preference. However, our test subjects are very indecisive, especially when there are more than two products. As a result, testing takes forever. I suggested that we perform the comparisons in pairs and then use these comparisons to get the rankings. Thus, if we have three product variations, we have the customers compare variations 1 and 2, then 2 and 3, and finally 3 and 1. Our testing time with my new procedure is a third of what it was for the old procedure, but the employees conducting the tests complain that they cannot come up with a consistent ranking from the results. And my boss wants the latest product evaluations, yesterday. I should also mention that he was the person who came up with the old product evaluation approach. Can you help me?"

(a) Is the marketing director in trouble? Will his approach work for generating an ordinal ranking of the product variations in terms of customer preference? Explain.

- The new approach is good, but as the testers don't know how to come up with a consistent ranking, the marketing director should ask them to compare the samples in pairs, and then fuse all the data from the different comparisons and create a ranking.

(b) Is there a way to fix the marketing director's approach? More generally, what can you say about trying to create an ordinal measurement scale based on pairwise comparisons?

- There is a way to fix the approach, which is to create the ordinal measurement variable. However, it is important that there are pairwise comparisons of any pair. This approach could result in a problem of circularity (A>B>C>A) in theory, but normally it doesn´t happen, one of the three should be better than the other two.

(c) For the original product evaluation scheme, the overall rankings of each product variation are found by computing its average over all test subjects. Comment on whether you think that this is a reasonable approach.

What other approaches might you take?

- It is not a reasonable approach, as people have different bias and precision in their measurements, some people would give a 0 to the worst and a 10 to the best, while other would give a 6 to the worst and 8 to the best, so the first person would count 5 times more than the second. I would recommend that they transformed the data to compensate for each person bias and precision, by changing each person's valuation of each product to the difference between their valuation and their minimum valuation and dividing it by the difference between their maximum and minimum valuations.

12. Discuss why a document-term matrix is an example of a data set that has asymmetric discrete or asymmetric continuous features.

- A document-term matrix is an example of asymmetric discrete features because only attributes that are non-zero are important (the words that are actually present in the text), and these attributes, being the number of times that word appears in the text are integers, which are discrete.