Exercises

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.

-Data mining task because it gives the company the ability to make adjustments to whether focus on having services/products that is centered on the gender that is most frequent or the other way around to balance all genders.

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

-Not a data mining task because it doesn’t give a useful enough information as the profitability of customers are affected by reasons such as their budgets and their needs and simply dividing the customers by the profitability is not enough to have an idea on what improvements to make.

1. Computing the total sales of a company.

-Not a data mining task because the total sales of a company only give the general idea of how the company is doing and not specifically the reasons that contributes to the total sale which is better because it gives a clear idea on what to improve on and can be useful to predict the future total sales in the future.

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

-Not a data mining task because it doesn’t yield useful data aside from the identity of the student of that ID number.

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

-Not a data mining task because since its only fair, it won’t have any factors that will affect its outcome and the probability of the outcome has a fixed amount. If tossing the pair of dice are not fair then it will become a data mining task because it will have other variables that it will dependent on.

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

-Data mining task since the future stock will be relying on the historical records and history records is a strong variable since it can repeat the outcomes of the variables that is dependent on it.

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

-Data mining task because vital statistics like heart rate are major factors when diagnosing abnormalities. Doctors can predict what possible abnormalities can occur with the use of heart rate.

1. Monitoring seismic waves for earthquake activities.

-Data mining task since an earthquake activity is a dependent variable and is dependent on the seismic waves.

1. Extracting the frequencies of a sound wave.

-Data mining task because the frequencies of sound waves give information about the sound wave such as predicting on what causes it and sometimes its location. Sound wave is dependent on its frequency.

2. Suppose that you are employed as a data mining consultant for an Internet search engine company. Describe how data mining can help the company by giving specific examples of how techniques, such as clustering, classification, association rule mining, and anomaly detection can be applied.

Search engines can be optimized in many ways and with the help of the data mining techniques. Clustering can be used in search engine optimization by clustering all of the possible results that depends on how well it matches on what the user searched and will be then be served to the user. For the classification, it can be used to the search engine to sort all of the data depending on the classes that will make it easy to search engine to give back to the user of they searched for a specific class or category. Association rule mining is the use of machine learning to study the patterns of how the data is being processed and can be used to search engine as a way to evolve further and predict more scenarios such as what the user would search in the future based on what he has been searching previously or simply optimizing the performance of the search engine by memorizing the patterns so that it does not go through that patterns again which cause performance issue in the future. Anomaly detection on the other hand is useful for search engines since it filters the data that has anomalies in it with the worst case of it having a malware which will negatively affect both the user and the company itself. Anomaly detection takes care of that process and removes all of data that is acting weird beforehand.

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.

-Yes, because we are talking about the data of a person so whether or not that person is still alive or not, data privacy is still important because it is their personal data unless the inquirer is given permission by its relatives.

(b) IP addresses and visit times of Web users who visit your Website.

-Yes, all of data across the internet especially the IP addresses are sensitive data that can be used to gain private data from the users of that IP address. Data privacy is important here since it gives assurance and trust that the website will not be leaking the user’s IP address.

(c) Images from Earth-orbiting satellites.

-No because the images from the satellites are often used to monitor the state of the earth. Data privacy will only be important here if the images contain sensitive/confidential content that are for the users and it can be solved by censoring those content.

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

-Data privacy here is only important for the addresses since a telephone book is used to search for contacts so the names do not require data privacy. On the other hand, the address is a private information and it is obvious that people would not want to reveal their address for the sake of their safety and privacy.

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

-Yes, both names and emails require data privacy because almost every user does not want to reveal their identity in the internet because of the internet can easily spread data around which is not good for the users.