Manas Gaur 001262723 Date : Feb 10, 2016

Solution to Data Mining HW-2

Solution to question 1

a. **No,** it is not a data mining task. This is because, the question ask regarding dividing the customers according to gender is similar to grouping of customers based on gender. This task can be easily be performed using “group by clause” in SQL. Hence a simple data base queries, such like the above, falls into the category of data analysis rather than data mining.
Example: select customer from company group by gender;

b. **Yes,** it is a data mining task. This is because, the question is considering the term “profitability” which is a relative term and is different for different individual. This key word “profitibility” of the customer is a deduced based on the analysis of the buying pattern of the customer either through online shopping or offline shopping. To provide an appropriate answer to this question, a data mining personnel has to uncover all the pattern using the historical data of the customer. **Hence it is truly a data mining problem as it might involve hypothesis testing, likelihood measurement and bayesian probability.**

c. **No**, it is database querying task. This is definitely not a data mining task as the questions asks for the total sales of the company. This is as simple as execution a “sum aggregation clause” on the data. The question above all is ambiguous as it does not mention whether the total sales are required for the current fiscal year or all the previous year including the current year. Still, it falls into data analysis problem rather data mining problem.
Example : **select sum(sales) from company;**

d. **No**, it is not data mining, rather it is a database querying task. In the question, it says sorting the student database based on student identification number is simple a data base operation as student identification number behave like a primary key and sorting it , is not really a data mining task as it is more close to data arrangement activity.
Example select student_id form student order by asc/desc;

e. **No**, it is not a data mining, as outcome is fixed and not dependent on previous outcomes. Since, the question aptly mention that the dice which is being tossed is a fair dice, hence it has a fixed and equally likely probability of {(1,1/6), (2,1/6), (3,1/6), (4,1/6), (5,1/6), (6,1/6)}. This ordered pair (number on the dice, probability) is fixed. There is no possibility of application of bayesian chain in the question, which would have been a possibility, if the dice would have been biased. Hence, it is a simple counting experiment rather than a data mining experiment.

f. **Yes,** it is definitely a data mining task. As mentioned in the question, that we need to predict the stock price of the company using the historical task is more close to a classification problem in data mining. This is because, the historical data can behave as a training, testing and validation data using cross-validation and hence we can train the classifier, which can predict future stock prices of the company.

g. **Yes**, it is a data mining task. This problem has a data mining task because of the key word “Abnormalities”. If we say, we are monitoring the heart rate of the patient, it is data collection task (ETL type task) but when we say “monitoring the heart rate of the patient for abnormalities” then it is a data mining task as it requires historical data and a classifier which will create a range of values of heart rate. If while monitoring the heart rate of the patient, if the

heart rate falls in that range, we can say that there is abnormality in the heart rate of the patient. This task can be extended to classify the patients as cardiac patients and non-cardiac patients.

h. **Yes**, it is a data mining task. Answer to this problem is similar to what has been said in Q1(g). In this question, if we are monitoring the seismic waves, we are merely collecting the data and it is ETL task. With this task we are preparing data for Mining techniques but if we are monitoring the seismic waves for earthquake activities, we are doing a data mining task. This is because, we assume that we have historical time series and seismic frequency data and we are applying a classifier on this data which will help us predict earth quake activities as it has been trained with historical data to a point of very small mean square prediction error. This process of training, validating and testing the classifier based on historical seismic waves data is called a Supervised data mining task.

i. **No**, it is not a data mining task as the objective is not illustrated, as to why we are extracting frequency from sound waves. The problem is merely to extract the frequencies of the sound wave which is a data collection and analysis task. If it would be have been extracting frequencies of sound wave which matches to that of terrorist voice frequency than it would have been a data mining task.

Solution to Q2. Being a data mining consultant in an internet search engine company, my job will be towards improved and search based improvement of results based on past searches made by the customer through recommendation, customized links and improved search time through cookies management

Firstly solving the data mining problem using clustering. Clustering is a data mining technique used in unsupervised domain that is the data does not have any labelled attribute. similar techniques like K-means clustering ,collaborative filtering, cosine similarity and jacquard similarity are some of the popular clustering technique used by an search engine company. For instance, we are taking over youtube.com, the famous site for videos, a frequent visitor to this site, can view videos from different genres which can helps us making a video-mix for the customer consisting of all the videos he liked or visited previously. This example, aptly uses clustering technique.

Secondly, considering classification, which is a data mining task used in supervised learning. Suppose, the customer, who is a daily visitor of the youtube.com, views videos from different genres such as TV-shows, Hollywood movies, Hollywood songs, Sport videos or educational videos. All these genres are discrete and independent but based on trend of viewing videos from different genres, we can customized the page of the visitor under different genres which is a classification problem.

Thirdly, considering association rule mining (ARM) which is a type of rules base clustering technique has been used immensely in market basket analysis, stock market portfolio development and enhancing the sales of shops by grouping items based on confidence and support metrics of association rules. This data mining technique tend to establish relation between discrete entities by calculating support and confidence. The confidence value should be greater than or equals to the support for establishing a rule. It has a very nice involvement in search engines as ARM improves search queries. For instance, when we type a name of a place say San diego on google, then we see links for tourism, universities, population and organizations. Considering youtube.com, based on the surfing of the visitor, we did clustering for making video-mix, classification for customizing the page and ARM can be used to formulate the

order in which videos can be seen. For instance, base on previous surfing data about the visitor, we can create one rule such as Songs->TV-Shows->Open-Courseware.
Lastly, Anomaly DetectionAnomaly detection is when an input variable is very dissimilar from other variables (or events) contained in the data base. This is a helpful tool to insure that only pertinent information is included in search results. Anomaly detection can be useful in the first pharmacology example to insure that the only information relayed to the user is about related drugs, rather than drugs associated with treating unrelated symptoms.

Solution to Q3 : log files provide information about User Name, IP Address, Time Stamp, Access Request, Bytes Transferred, Result status, URL Referred and User Agent. Analyzing these log files, gives us a neat idea about the user. Web log data analysis using frequent items mining can also gives an idea about the buying or surfing patterns of the user. Since, it has a timestamp attribute and URL attribute, it is temporal dependent data.

Solution to Q4: nucleotide data is a type of biological data consisting of discrete sequences or strings of strands. Different ordering of these sequences provides details about the characteristic features of proteins. The structure of proteins are highly complex and evolving with time, hence, nucleotide biological data is temporal data or time-series data. Moreover, we are talking about finding patterns in terms of biological properties, which we can do based on the dependency between different nucleotide structures. Hence, it is a multivariate discrete sequences dependent data.

Solution to Q5 : Classification of customers based on demo-graphic data is a decision tree data mining problem as the attributes are categorical such as Single, Committed, Married, etc. Hence this problem is a classification type data mining problem.

Solution to Q6 : Since it is given that merchant has information about someone has bought the widget or not, the data corresponding to that given information can behave as a training data. With respect to this training data we can define a classifier for classification. This problem is a classification problem as there is supervised learning involved. Since, the problem is more of binary one (yes/No) hence a decision tree classifier will do the job.

Solution to Q7 : The problem is asking to find set of items that are often bought together. This problem is mainly concerned with finding sets or clusters of items, which makes this problem, a clustering task. Since, we are to find frequent item sets, we will be using Frequent ItemSet mining that is FP-Tree formulation or Enumeration trees. This data mining technique is a more advanced version of A-Priori Mining or Association rule mining.

Solution to Q8 : Since given that small number of customers lie about their demographic profile, which resulted in mismatch between buying behavior and demographic profile, these entries in the data are termed as outliers in the data as because of the presence of such tuples, there are chances of overfitting in the curve. Hence the supervised learning problem as per answer Q7 is transformed to outlier detection and removal problem. Moreover, as said in Q6, we are about to use decision tree classifier, which will provide a confusion matrix, which helps us identifying false position, true positive, false negative and true positive in the data and hence the outliers can be localized in the data.