Task 1: review data mining concepts and tasks

Question 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
   No, this would be considered an information retrieval task.
b. dividing the customers of a company according to their profitability
   No, this would be considered an information retrieval task.
c. computing the total sales of a company
   No, this would be considered an information retrieval task.
d. sorting a student database based on student identification numbers.
   No, this would be considered an information retrieval task.
e. predicting the outcomes of tossing a (fair) pair of dice
   No, predicting the outcomes of rolling a fair dice is a probability problem.
f. Predicting the future stock price of a company using historical records
   Yes, it would fall under the predictive task category as a regression modeling task.
g. Monitoring the heart rate of a patient for abnormalities
   Yes, it would fall under the descriptive task category as anomaly detection.
h. Monitoring seismic waves for earthquake activities:
   Yes, it would fall under the descriptive task category as anomaly detection.
i. Extracting the frequencies of a sound wave:
   No, this would be considered a data collection task.

Question 2: Suppose that you are employed as a data mining consultant for a sports analytics 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.

Clustering could be used to predict rookie NFL performance. Depending on attributes such as height, weight, hand size, vertical jump, 40-yard dash time, or college stats, a particular player could be grouped with NFL players who came into the league with similar attributes. Based on that data and by using a nearest neighbors' model, perhaps, one could come up with a reasonable prediction about how the rookie will perform and a reasonable salary for a contract.

Classification could be used in the NFL to determine the likelihood of whether a player got concussed on the field when there is a stoppage due to suspected concussion. It would be difficult and would require high-quality data about what is happening on the field in real time. But if the model were accurate enough to be relied on by medical staff, it would ensure that players are only removed from the game if they need to be.

I believe that an interesting application of association rule mining in the NFL would be to see if there are any certain plays or combinations of plays that are frequently called together in drives or in certain game states. For example, what if a coach tended to call the same play whenever the situation is in 1st down, with 40 yards to go, and winning by 7. There are thousands of such

possible situations and there might be some patterns. I am not sure whether this would be the case, but as a consultant it might be worth considering since it could help with play calling.

Question 3: for each of the following data sets explain whether or not data privacy is an important issue.

j. Census data collected from 1900-1950

   Yes, data privacy is an issue. Although statistics can be reported, any personal identifiable information in census data is should be kept private according to US law.

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

   Yes, data privacy is an issue. IP addresses have been determined by law to be personally identifiable information and should be kept private.

l. Images from earth orbiting satellites

   Yes, data privacy is an issue. Some satellite images may be confidential. For example, there are certain areas that Google Earth must blur out.

m. Names and addresses of people from the telephone book

   No, data privacy is not an issue because the telephone book is publically accessible information.

n. Names and email addresses collected from the web

   No, data privacy is not an issue because the names and email addresses are publically accessible information.

Task 2: practice your critical thinking and writing

The first article attempts to deride Google GFT based on the finding of their estimate of influenza cases being highly overstated compared to the CDC estimate of influenza cases. As pointed out in the article, this was not the case for just one year but several years. Due to this discrepancy, the credibility of Google GFT as a reliable tool was brought into question.

The second article recognizes that Google GFT estimate of influenza cases is greater than the CDC estimate of influenza cases but explains why this is ok. Google GFT uses data from their search engine versus the CDC which uses data from the healthcare system. Furthermore, the team that developed Google GFT had met with the CDC to help shape the tool. There is more detail presented about how the Google GFT tool was designed to be compliment to the CDC and not a replacement.

In the first article, questioning Google's GFT data is understandable given how different it is from the CDC data. I agree that this merits some justification, but there is not enough consideration given to the fact that Google GFT was not designed to be a standalone tool. The second article provides more insight into this by citing conversations with leaders at Google who were part of the project. I think that Google GFT is an early example of looking at things differently in the new era of big data. In this new era, conventional belief systems are challenged by evidence to the contrary. There will usually always be some pushback, but this ends up being productive for society because it provokes rethinking about what the best way of doing things is.