IST707 Homework 1

Kyle Walter

# Task1: Review data mining concepts and tasks

## Question 1 – Discuss whether each of the following activities is a data mining task.

1. Dividing the customer of a company according to their gender
   1. This is a preprocessing task of the knowledge discovery in databases. Maybe we want to look at habits of female customers and thus feature reduce the data set, but this is not the data mining process itself. Though confusing as the textbook argues in 1.3 that all KDD steps have come to be understood as under the umbrella of data mining. Traditionally this is not considered data mining.
2. Dividing the customers of a company according to their profitability
   1. This is a data mining task; it classifies the customers into groups based on their profitability (likely done through clustering techniques).
3. Computing the total sales of the company
   1. This is not a data mining concept or task. To turn it into one, the data could be used to look for a way to predict future sales for the company which would then turn it into a data mining task.
4. Sorting a student database based on the student identification numbers.
   1. This is a database management task that allows for the sorting a faster use of a database but is not a data mining technique.
5. Predicting the outcomes of tossing a (fair) pair of dice
   1. This is a pure statistic and thus a concept of utilized and put into application by data mining. The example in item E not a data mining task as described.
6. Predicting the future stock price of company using historical records
   1. This is data mining at its core. The historical information was analyzed for patterns and model designed based on those patterns. With input new input information, the model can predict the future price.
7. Monitoring the heart rate of a patient for abnormalities.
   1. Anomaly detection is defined as a data mining task.
8. Monitoring seismic waves for earthquake activities
   1. Along with item G is another example of anomaly detection. Looking for eco system disturbances that likely indicate when an earthquake will arrive or is arriving.
9. Extracting the frequencies of a sound wave
   1. This is likely a data mining task looking at patterns. If the oscillations has been recorded by a device data mining techniques could be applied to determine the frequency at which the sounds is occurring.

## Question 2 – Suppose you are employed as a data mining consultant for an internet searching 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.

For search engine we could help by doing several things. 1 would cluster your customers by the amount of usage they get on your site and which type of items that they often search. Then apply association ruling mining to their activity. This could help us identify places and sites that are relevant to these searches and suggest these searches to other users by using association rules. We could also utilize the information about the sites the users ultimately go to move up results in the engine. Lastly, we can use anomaly detection to determine when a site might be trying to inorganically move its result to a higher place in the ranking. This could help us increase ad revenue by offering a service to display the resulting site at the top certain searches and see if users really are interested in these sites, or if the site is trying to game the algorithm of results.

## Question 3 – For each of the following data sets, explain whether data privacy is an important issue.

1. Census data from 1900-1950
   1. Depends how detailed the records are and were collected from. The US Census site seems to only provide numbers by groups. However, the census does collect information from people such as names, addresses, age, and other demographics.
2. IP addresses and visit times of Web users who visit your website.
   1. IP addresses are information that need to be kept private. IP addresses can be linked to individuals and show their activity not just on the one site but across the internet.
3. Images of Earth-orbiting satellites
   1. These appear in the public domain and as such do not really have a privacy concern.
4. Names and addresses of people from the telephone book.
   1. Once again these are in the public domain anyone with a computer can get to this information.
5. Names and email addresses collected from the web
   1. These should be kept as the people who are signing up for services from a company likely had some expectation of privacy and leakage can have consequences.

# Task 2: Practice Critical Thinking and Writing

### Read the following two news articles. One criticized Google Flu Trend, and the other defended it. Write one paragraph to summarize the criticism, and another paragraph for the defense. Write the third paragraph to offer your own thought, e.g. is the criticism valid? Does the defense make sense? What other problems or benefit do you see in Google Flu Trend or similar big data applications?

The first of the two articles focus on criticizing Google's flu trends prediction for flu cases over the past few years. The article points out based on the prior flu seasons at the time the article was written that Google had overshot the number of predicted flu cases verses the actual reported by the CDC in 100 out of 108 weeks. In addition, after tweak the algorithm in the following season, Google’s prediction continued to overshoot the official number later reported by the CDC thirty percent of the time. The article also points out the academics who came out against google at the time suggested using more traditional data analysis techniques along with the CDC’s data would give a more accurate result.

The second article comes to the defense of Google’s Flu trends Big Data prediction. It points out in the paper the generated the first article and several others as well, but that mention of utilizing the CDC’s lagged data and Google Flu trends the CDC could get a better result that either of the two were providing alone. Thus, leading the article to conclude the if the CDC were aware of that they could combine the data form Google Flu Trends with their own and get a better epidemiological understand of the county. That in turn would mean the Google Flu Trends was a success based on its outlined purpose, namely, to provide data that would spur or successfully direct traditional research to areas that measures direct disease activity.

Among the two articles, what google outlined to the Google Flu Trends tool to be is a signal what was happening with the Flu around the world, rather than a stand-alone forecasting tool. In some ways google was victim of their own success. As the critical article pointed out it did a really good job at predicting flu outcomes in the 2008 and 2009 flu season. Then with up ticks in the flu related to the H1N1 pandemic, along with Google’s growing use by the larger public, the way that google was generating the results for Google Flu Trends needed to change. Google may have been doing this in the background, but it seems like the people who were mission critical had left the company as mentioned in the critical article. While publicly Google did admit to revisiting the algorithm in 2013, this was almost five years after the product first launched. Personally, I think as prediction tool, it is just that a prediction, which was publicly accessible. People may have used this information to reduce their chances of getting the flu, such as getting vaccinated, washing their hands more frequently, and searching the internet for other tips. The latter spiking googles predicted result. Overall, I think it was successful venture to get people thinking about what Big Data could do. However, due to the way in which is got shelved as a something that needed on going maintenance, Google Flu Trends suffered a huge blow reputationally and maybe to some extent the public’s confidence, for a time, in big data capability.