IST 407/707 Data Analytics & Data Mining

HW1

**Task 1: review data mining concepts and tasks**

Answer the exercise questions 1-3 in Textbook 1.7. For Question 2, feel free to change the question scenario from “an Internet search engine company” to any organization that you would like to think of. It can be a company, government office, NGO, etc.

See questions 1-3 as below:

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

No, we cannot consider this a data mining task.

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

No, we cannot consider this a data mining task as well.

* 1. Computing the total sales of a company

No, this is not a data mining task.

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

No, this is not a data mining task. This is a sorting task.

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

No, this is not a data mining task. This is a simple probability task.

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

Yes, this is a data mining task where we are predicting future stock prices.

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

Yes, this a data mining task. There could be two possible ways to interpret this, one where, one is just observing the heart rate which is not a data mining task. But if one is creating a model to detect abnormalities in them then it is a data mining task.

* 1. Monitoring seismic waves of earthquake activities

Yes, this is a data mining task. This is a similar situation as the previous one. We are predicting earthquake activities by monitoring seismic waves. We will have to build a model to do that.

* 1. Extracting the frequencies of a sound wave

No, this is not a data mining task. I think this is physics.

1. 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.

Data mining can help the company in the following ways –

* We can cluster similar type of users by their querying activity which would enable us to show them similar ads running on our engine.
* We can also use these clusters to display results of their queries.
* We can find bots, malicious activity using anomaly detection techniques. This could be done by using the browsing activity of users.
* We could use classify our users based on their querying activity. This could help us to say if they would buy our premium version or not.

1. For each of the following data sets, explain whether or not data privacy is an important issue.
   1. Census data collected from 1900-1950

Yes, the data collected should be kept private. One can draw inferences and display mean, median results of income, race, gender etc. But if personal information is present in the collected data then it should not. But this data seems to be quite old. So, I guess there shouldn’t be any issue, but laws could defer by country.

* 1. IP addresses and visit times of Web users who visit your website

Yes, the IP addresses of a user could be an important to secure. One could track the location of the user using the IP address

* 1. Images from Earth-orbiting satellites  
     No, I guess this is also made public by some websites. If this is for a specific research or task which one would not like to share with everybody until complete, then yes.
  2. Names and addresses of people from the telephone book

No, the telephone book is anyways available to all of us. People have enrolled in it to be made available

* 1. Names and email addresses collected from the Web

No, this is not a data privacy issue. This is because if the name and email address of a person is on the Web, it means it is available for the entire world to see anyways.

**Task 2: practice your 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?

<http://bits.blogs.nytimes.com/2014/03/28/google-flu-trends-the-limits-of-big-data/>

<http://www.theatlantic.com/technology/archive/2014/03/in-defense-of-google-flu-trends/359688/>

Criticism:

The Google Flu Trend was criticized by a group of researchers. They said that Google Flu Trend overestimated the Flu. Their results were overestimating the results for almost 90% of the weeks. They said that using Big data and all these tools and technology was of no use if the prediction was going to be so wrong. They criticized the hype about big data. They also said that the data available by CDC was able to predict very accurately compared to the Google Flu Trend.

Defense:

In their defense Google said that this tool was only meant to be a support for the currently running system. It was not meant to replace it. It tried something different and took a different perspective on how to predict Flu. Even the CDC mentioned that it was great to have such a tool which could assist them. They also proved that when both the systems were used in unison, they produced results better than either one of them.

My View:

I this the criticism was too harsh. The fact that Google accepted that their system was not made to replace the current system but assist the current one then its fine. They also made efforts to improve consistently. But trying something new is always cool and you do not know when this would turn to be super beneficial, so it was worth a try. Criticizing them for using new technology to solve a real-world problem is harsh. The world would never improve if this was the case.

Yes, overestimating the flu by a lot could cause panic amongst the people. This could be a point where Google went wrong. But this is not what the critics intended to say.