Nicholas Lichtsinn

IST 707

1/19/2022

**Homework #1**

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

1. **Discuss whether or not each of the following activities is a data mining task.**
2. **Dividing the customers of a company according to their gender** - No, this is not a data mining task because you can easily sort data this way.
3. **Dividing the customers of a company according to their profitability** - No, this is not a data mining task because this is a simple sort of data that is already present.
4. **Computing the total sales of a company** - No, this is not a data mining task as it is simple accounting of already present data points
5. **Sorting a student database based on student id numbers** - No, this is sorting a database not using data mining techniques
6. **Predicting the outcomes of tossing a (fair) pair of dice** - No, since the dice are fair this is just a probability calculation.
7. **Predicting the future stock price of a company using historical records** - Yes, this is a data mining task because we could create a model to predict the value of the stock price based on historic value using predictive modeling.
8. **Monitoring the heart rate of a patient for abnormalities** - Yes, we could build a model of the ‘normal’ variation in heart rate to see if their heart rate falls in the normal model known as anomaly detection.
9. **Monitoring seismic waves for earthquake activities** - Yes, we could create a model of the different earthquake activities and sound the alarm when seismic waves that correspond with different earthquake activities present themselves.
10. **Extracting the frequencies of a sound wave** - No, the frequencies describe sound waves but no data mining task is present.

**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.**

1. As a consultant I would use classification to determine whether or not customers who click on ads that we place are going to purchase something from that ad, because this is a binary value to buy or not to buy. I would train a model to see what customers buy from ads and increase the ads they see. I would use clustering to segment my traffic to show them different ads based on their cluster, if some customers like biking but dislike trucks I would show them ads for environmentally friendly products for example. Association rule mining can be used to discover any patterns in how users interact with the website, are they clicking on certain websites together or in succession, should I place these next to each other or at the top of the page so they are featured together. Lastly, I would use anomaly detection to make sure there are no security breaches or anyone accessing the site via unacceptable methods compromising the site in any way.

**3. 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 because there are still people alive who were part of that census data.
2. **IP addresses and visit times of Web users who visit your Website**. - Yes, this information could allow hackers to steal their private information
3. **Images from Earth-orbiting satellites**. - At a high level no, but if they are taking clear close-up pictures of people then yes that would be a data privacy concern.
4. **Names and addresses of people from the telephone book**. - Yes, your name and address should only be included in a public setting if you allow access to it.
5. **Names and email addresses collected from the Web**.- Yes, same as the phone book, your name and address should only be included in a public setting if you allow access to it.

**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 benefits 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/>

In 2008 a team at Google created Google Flu Trends to use ‘Big Data’ to track flu outbreaks. The criticism in this article shows us how Google Flu Trends missed the mark, overestimating Flu cases in 100 of the 108 weeks starting August 2011. Media backlash was severe with most major news sources releasing an article condemning the new approach to Flu tracking. What most of these articles did not include was the original purpose of Google Flu Trends. This was intended to be a secondary tool to be combined with CDC data to give a better result than either source by themselves. They accomplished this task and in reality only failed in the court of public opinion.

The article from the NY Times defending Google Flu Trends expands on the premise that Google Flu Trends was never intended to be a magic way to monitor flu trends perfectly and was intended to be used as a real-time signal of flu outbreaks. We get more clarity on the Google Flu Trends mission and purpose. The author criticizes other media outlets for using Google Flu Trends to critique the use of big data and its drawbacks, while missing the reality that Google Flu Trends was only ever meant to be a ‘secondary signal’ to assist the CDC, not replace their findings.

The criticism and defense of Google Flu Trends both make good points. As a society we cannot just look at the data and take our results as fact. Falling into the trap of ‘big data hubris’ can be dangerous, just because you have a lot of data does not mean it is telling the correct story or that you are using it in the correct way. However Google Flu Trends is not an accurate representation of misused data, as it was only intended to be a secondary signal to assist in detecting outbreaks before the CDC could report flu rates. Even though they overestimated flu rates for years, Google Flu Trends did correctly give advanced indication of higher flu rates in the 07/08 season and for the 09 H1N1 outbreak. This proves to me that they have succeeded in their mission and no doubt this case study will be helpful for future sociotechnical systems.