Juan Carlos Ventosa

IST 707 Data Analytics

Homework Assignment 1

Submit Date: 7/6/2020

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

Answer the exercise questions 1-3 in Textbook 1.7.

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.

This is not a data mining task. Dividing customers by gender is a trivial sorting task that does not extract any useful or previously unknown information from the data.

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

This is not a data mining task. Dividing customers by their profitability is mathematical task calculated using a defined value for profitability. However, predicting the profitability of a new customer is a data mining task requiring techniques such as clustering and classification to find patterns in the data that identify profitable customers.

* 1. Computing the total sales of a company.

This is not a data mining task. Calculating the total sales of a company is a fairly simple mathematical task that does not in itself provide useful information. Additional tasks such as comparing the total sales with previous time periods, predicting sales based on seasonal trends or identifying marketing initiatives that effected sales would be examples of data mining tasks.

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

This is not a data mining task. The act of sorting data based on likely chronological or randomly assigned identification numbers does not extract any useful or previously unknown information from the data.

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

This is not a data mining task. Calculating the probability of the outcomes of tossing a fair pair of dice is a known mathematical formula.

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

This is a data mining task. Discovering trends, correlations and anomalies from historical records may provide useful information towards predicting a company’s future stock price. The discoveries could also lead to key variables to be used in predictive models utilizing techniques such as regression or classification.

* 1. Monitoring the heartrate of a patient for abnormalities.

Monitoring a heartrate for abnormalities is a data mining task. Since abnormalities would have been previously defined in order to be detected data mining techniques such as classification can be used to identify known abnormalities as they occur.

* 1. Monitoring seismic waves for earthquake activities.

This is a data mining task. Data mining techniques such as classification can be used to identify seismic waves based on wavelength, depth, and source to measure an earthquake’s power on the Richter scale.

* 1. Extracting the frequencies of a sound wave.

This is not a data mining task. Extracting the frequencies of a sound wave would not provide previously unknown information. This task is a form of retrieving data much like a query from a database. No new discoveries are gained from the process.

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 a search engine company improve the relevance of its search results and improve the advertising performance of its clients. For example, clustering techniques can be applied to search algorithms to find similarities between websites and group them based on given search terms to produce relevant search results. Classification techniques can also be used to improve search results by categorizing the results’ relevancy as high, medium, or low to rank the order in which results are displayed. Association rule mining techniques can be used to improve search ad performance by determining how search terms may be associated with products, brands or services to create recommended key words to aid the relevancy of client ads with search terms. Anomaly detection techniques can be used to protect a search engine’s users by eliminating suspect, fraudulent or predatory websites from appearing in search results.

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

US Census data can provide an enormous amount of detailed information about the country’s population, but it is typically aggregated and anonymous in nature. This is not an important data privacy issue as long as the identity and information of individuals in the census remain unknown.

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

Although it is standard practice for websites to collect traffic information, the hosting website has a responsibility to inform its visitors that a record of their visit will be made. In addition the traffic information collected by the hosting website should not be available to the public; it must be considered an important privacy issue. Much like a telephone call will create a record through the phone company that hosts the call, access to the phone records should only be available to the parties involved in the call. A website works similarly where a user’s browser has a record of sites they visited and the host site has a record of visitors. Websites should clearly state in their privacy terms the security measures used to protect their visitors’ traffic information from being distributed or stolen.

1. Images from Earth-orbiting satellites.

Images from Earth-orbiting satellites can reveal important information that some may wish to remain private. Depending on the quality of the images, detailed information on activities done on private properties such as backyards, rooftops and driveways may create privacy issues for some property owners and individuals. A larger privacy concern would be using satellites to track movements via GPS or a live feed. These activities may require an individual’s permission to be tracked.

1. Names and addresses of people from the telephone book.

In the US, laws and regulations have required phone companies and phone directories to give people the option to not be listed in telephone books. People who choose not to optout are essentially providing permission for their names and addresses to be made public. This is not an important privacy issue as long as the right to optout remains.

1. Names and email addresses collected from the web.

This is a potentially highly important privacy issue depending on how the names and email addresses are collected. If collected, distributed or even sold without permission from the subjects to whom the names and addresses belong, privacy has been broken to allow third parties to profit from the information and allow possible marketers, spammers or thieves to target the victims of the stolen data. If the data was legitimately collected with the subjects’ permission than high security measures should be used to protect such data from being stolen or distributed.

**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/

Google Flu Trends developed by Matt Mohebbi and Jeremy Gingsberg is a flu tracking service using search terms entered on Google to predict the incidence of flu infections in many countries around the world. In a paper published in the journal *Science* titled “The Parable of Google Flu: Traps in Big Data Analysis”, the four authors David Lazer and Alessandro Vespignani, professors from Northeastern University, along with Ryan Kennedy from the University of Houston and Gary King, the director of the Institute for Quantitative Social Science at Harvard University criticized Google Flu Trends’ performance. The authors specifically call attention to Google Flu Trends’ overestimated number of flu cases in the United States during 2011-2013 when compared to cases reported by the Center of Disease Control (CDC). Arguing big data sets does not always perform better than traditional data analysis, the authors found using flu trends from CDC reports provided more accurate predictions. Google Flu Trends’ focus on search results and correlation, the authors argue, does not account for a wide enough range of data and analysis tools to make accurate predictions. It fails as a stand-alone flu monitor. Google Flu Trends can improve its service by including data beyond the scope of Google search. To illustrate this, the authors’ analysis show combining Google Flu Trends data with CDC data provides better results for predicting flu cases in the United States.

*The Atlantic* article “In Defense of Google Flu Trends” by Alexis C. Madrigal suggest combining Google Flu Trends with other models is precisely the purpose of its developers and the success of its invention. Designed to provide an alternate model to current predictors on the prevalence of flu, Google Flu Trends offers a separate source of data for checks and balances with other sources. Google Flu Trends was not developed to replace other flu monitoring signals. It was developed to complement them. Along with the success of combining Google Flu Trends data with CDC data, a 2013 John Hopkins study using emergency room data also found incorporating Google Flu Trends data into its model improved results. Although flaws with Google Flu Trends’ predictions as a stand-alone monitor exists, the value it provides when used in tandem with other data sources and models was the success its developers intended.

Using the inaccuracy in Google Flu Trends’ predictions to demonstrate that volume in big data analysis does not supersede the importance of data quality and analysis tools was a valid criticism. Google Flu Trends did fail to provide better results than traditional predictions and its inaccurate estimates as a stand-alone monitor may have negative consequences on how the information is interpreted or used. However, the developers of Google Flu Trends intended it to be a complementary source or tool to improve other models. In this aspect Google Flu Trends did succeed and had contributed greatly to epidemiologists’ efforts in tracking influenza. By using its own separate data source to build its model, Flu Trends offered useful supplementary data to other predictor tools that was otherwise unavailable to them. The benefits from the distinctiveness of Flu Trends’ data allowed it to be a barometer to other models’ predictions and made it suitable for data integration without data replication from other sources. As a social science tool, Google Flu Trends likely suffered from the ever complicated nature on how people contributed to the model’s main data input source, search. True to many similar big data applications the analysis outputs of a model are only as good as data it receives. To be successful, predictive models require ongoing modifications to match the speed and fluidity of the changes it attempts to analyze. In areas affected by social science such as social media, marketing, epidemiology, etc., model predictions are difficult to build due to the rapid change in behaviors. In the final analysis Google Flu Trends succeeds as a complimentary or supplementary data source, but its stand-alone predictions should be viewed with a wary eye.

search and keeping other data sources out of the Google Flu Trends model was not designed by its developers to replace the CDC. It was designed as complimentary signal to other signals. It was not meant to incorporate CDC data into its model. It was designed to use a separate source of variables, search terms, to provide an alternate model to predict the prevalence of disease that could be used in tandem with other data sources and models. In addition to complementing the CDC data as mentioned in “The Parable of Google Flu:…” paper, a John Hopkins study in 2013 using emergency room data to build a flu monitoring model found GFT to