Ben Harwood

IST 707

January 5, 2020

Assignment 1

**Task 1**

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.

1. Discuss whether or not each of the following activities is a data mining task. (answers are italicized)
   1. Dividing the customers of a company according to their gender. *This is not a data mining task, simply a database query or spreadsheet sort.*
   2. Dividing the customers of a company according to their profitability. *This is not a data mining task. Assuming a threshold has been established this a simple check against that threshold.*
   3. Computing the total sales of a company.  *Again, this is just a straightforward calculation, and not a data mining task.*
   4. Sorting a student database based on the student identification numbers. *Similar to question a above, this is just a database query. Sorting a data set is not data mining.*
   5. Predicting the outcomes of tossing a (fair) pair of dice. *This is a straightforward (and well-known) probability calculation, not data mining.*
   6. Predicting the future stock price of a company using historical. *This is a data mining task because we would use the historical data to develop a model to predict the value of the stock price.*
   7. Monitoring the heart rate of a patient for abnormalities. *Assuming we had access to previous patients’ heart rates we could build a model of normal behavior to use to detect deviations or anomalies. So, this is a data mining task.*
   8. Monitoring seismic waves for earthquake activities. *This, too, is a data mining task. Using data from previous earthquakes we could build a model of the seismic waves before and during an earthquake and sound an alarm when the switchover is detected.*
   9. Extracting the frequencies of a sound wave. *I don’t think this is data mining. Isn’t this just a measurement?*
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.

Consider a normal retail environment. One of the key metrics many of them track is units per transaction, the idea being that more items on a specific ticket should lead to higher revenue from said ticket. Using this information, a retailer can use association rule analysis to see which items are frequently purchased together and use that information for both advertising purposes but also for training of their staff. A good example is noise-cancelling headphones, circa 2010. Almost every set available at that time required one or two AAA batteries to operate. While this is a perfectly logical item for stores to recommend, association analysis could indicate other items, such as phone chargers, movies, travel pillows, or rechargeable battery packs.

1. For each of the following data sets, explain whether or not data privacy is an important issue. (again, answers are italicized)
   1. Census data collected from 1900-1950. *Despite the still prevalent belief that the government uses Census data for nefarious deeds, there are no privacy concerns here.*
   2. IP address and visit times of Web users who visit your Website. *Definitely. IP addresses can lead hackers and other less-than-honest people to individual people’s location.*
   3. Images from Earth-orbiting satellites. *No, pictures do no paint a complete picture and are open to interpretation.*
   4. Names and addresses of people from the telephone book. *No, we are given the option of opting out of being included in the phone book.*
   5. Names and email addresses collected from the Web. *No, an email address itself cannot compromise an individual’s freedom, safety, security, finances, etc.*

**Task 2**

The article “Google Flu Trends: The Limits of Big Data” is a confirmation of the shellacking given to the Google Flu prediction model by the scholarly article “The Parable of Google Flu: traps in Big Data Analysis”. The general crux of the criticism is that Google Flu Trends had not only overestimated the number of flu cases in several of the years prior to 2013, but that they had grossly overestimated. Furthermore, the second article posits that simply using the C.D.C. data generated from doctor reporting would have been more accurate, and that the algorithm as a stand-alone monitor of flu cases was debatable. It is further argued that a combined approach of the algorithm and C.D.C. data is the best way to predict flu cases because the Google approach was too narrow.

Counter to this, the article “In Defense of Google Flu Trends” is a rousing support of Google Flu Trends. It cites the same article as that above, ironically pointing out the same point as the critical article did: that using both the algorithm and C.D.C. data is the best option. This article, however, clarifies that statement by noting that the combined approach would provide a better result than the algorithm or C.D.C. data could provide individually. The author of this article also points out in defense of Google Flu Trends that the intention of Google Flu Trends was to supplement, not replace, traditional methods of monitoring flu cases, and important distinction.

Personally, I think both arguments are a bit weak, but I agree more with the second article. The first article using the fact that the model overestimated the number of flu cases as the basis for the argument against using it glosses over the fact that the model was never intended to be the sole method of monitoring flu cases. The engineer that developed it himself said, as pointed out in both articles, that the system was designed to be complementary, so if the critical article points out that the designer himself said Google Flu Trends was not supposed to be the gospel when it comes to predicting the flu then the rest of the arguments are moot, because they all revolve around why it shouldn’t be used in the capacity of a primary monitoring signal. However, we could from learn Google Flu Trends consistent overestimations by continuing to refine the model as we get more and more data to use for training purposes.