Holden Herrell

IST 707 Data Analytics

18 January 2020

Homework 1

**Task 1:**

***1.* *Discuss whether or not each of the following activities is a data mining task.***

***a. Dividing the customers of a company according to their gender.***

This is not a data mining task. This is a categorization of customers. A data mining task would be attempting to predict the gender of a customer based on items purchased.

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

This is not a data mining task. This is a business calculation and categorization—attempting to predict profitability of a customer would be a data mining task.

***c. Computing the total sales of a company.***

This is not a data mining task. Similar to *b*, this is a business calculation.

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

This is not a data mining task—this is only a basic database query.

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

This is not a data mining task. This is a probability calculation due to the dice being fair. This would be a data mining task if the odds were not guaranteed for the dice and they had to be estimated.

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

This is a data mining task. A model could be developed using predictive modeling (i.e. regression or other time series methods).

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

This is a data mining task. Anomaly detection could be used to detect abnormal heartbeats by a model that establishes a “normal” heart rate.

***h. Monitoring seismic waves for earthquake activities.***

This is a data mining task. Similar to *g*, a model could be developed to monitor seismic activity, establish a benchmark (normal), and identify abnormalities that might indicate a potential earthquake.

***i. Extracting the frequencies of a sound wave.***

This is not a data mining task. This is just a measurement.

***2. Suppose that you are employed as a data mining consultant for a professional baseball team. 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.***

In the context of baseball, data mining can be very helpful to the organization in gaining a competitive advantage. Association rule mining can be used to determine the best positioning to shift the infielders defensively based on the batter tendency, runner position(s), and pitch count. This is done by aligning the defensive shift with the most commonly occurring attribute value conditions that result in an out (i.e. if 90% of the time the current batter grounds out in a 0-0 pitch count with no baserunners while the infield is shifted towards the first baseline—a shift to match that association rule would make sense).

***3. For each of the following data sets, explain whether or not data privacy is an important issue.***

***a. Census data collected from 1900-1950.***

This is not a data privacy concern. This data is collected voluntarily from participants who know that it will be publicized.

***b. IP addresses and visit times of Web users who visit your Website.***

This is a data privacy concern. This data was gathered not given and is potentially personally identifiable and/or something someone doesn’t want publicized.

***c. Images from Earth-orbiting satellites.***

This is not a data privacy concern.

***d. Names and addresses of people from the telephone book.***

This is not a data privacy concern. Similar to *a*, this data is volunteered with knowledge it will be published.

***e. Names and email addresses collected from the Web.***

This is not a data privacy concern. Again, this information was volunteered in/for a public forum.

**Task 2:**

In his article “Google Flu Trends: The Limits of Big Data,” Steve Lohr explores the critique of Google Flu Trends, a predictive model used to estimate cases of the flu. He references the reported inaccuracy of Google Flu Trends during the 2011-2014 flu seasons which ranged from 50% (2011) to 30% (2014) overestimates of flu cases. The article cites social scientist who criticize Google for its “big data hubris” and argue that traditional data collection and analysis is just as valuable as the use of big data. Additionally, these scientists question the algorithm behind Google Flu Trends primarily based on its inaccuracy despite improved overestimation rates. Lohr does include that members of Google admit that Flu Trends is not designed to replace traditional diagnoses and surveillance—rather it is meant as a barometer for potential flu outbreaks.

Alexis C. Madrigal offers a contrary perspective on Google Flu Trends in his article, “In Defense of Google Flu Trends.” He references the excitement about Google Flu Trends at its advent in 2008 both among the health professionals and the news media. At that time, Flu Trends was ground-breaking and worked very well based on the existing data. Madrigal also makes mention of the social science criticism of Flu Trends’ inaccuracy; however, he focuses more on the fact that it was meant as a supplemental signal to those from places like the Center for Disease Control (CDC). He highlights that the combination of Flu Trends and CDC data could result in better epidemiological understanding for the CDC. Furthermore, he emphasizes the value of Flu Trends data not being based on/merged with CDC data by referencing a study that improved its model using only Flu Trends data. His argument is that Google Flu Trends’ criticism is rooted in false expectations that overreach its original intention and that it has been useful in epidemiology and other fields.

While Lohr’s criticism of Google Flu Trends’ inaccuracy is a valid concern if one is relying solely on Flu Trends to signal flu outbreaks and cases—it is not valid if Flu Trends is being used as intended. Madrigal correctly calls attention to the true intention of Flu Trends as a complimentary source of flu case predictions. It cannot be reasonably expected that the model would predict flu cases with the same accuracy as CDC data based on confirmed cases (with 2 week lag). It is critical that Google Flu Trends and other big data applications be used in conjunction with traditional data analysis to result in the best modeling. Extreme care should be taken when attempting to use overarching application of big data to specific situations to prevent an inaccuracy rate such as the Google Flu Trends overestimation from having a detrimental impact.