Nicholas Louis Brown

Professor Wallace

IST 707

September 3, 2018

Task 1:

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.

Not a data mining task, this would be a way of sorting data not mining.

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

Again, ‘dividing’ the data is only sorting it not mining.

1. Computing the total sales of a company.

Not a data mining task, I would use something like excel to do this.

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

Not a task for data mining. This would be easily done in a spreadsheet.

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

Not data mining because like flipping a coin we know all the possible outcomes and the probability for each of those outcomes.

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

We can use data mining for this problem, combine multiple different sources and types to create a predictive model

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

We can use data mining here too, detect anomalies in the patients heart.

1. Monitoring seismic waves for earthquake activities.

Yes data mining, build a model to categorize seismic activity by danger level.

1. Extracting the frequencies of a sound wave.

No, you would just use a microphone and some software

1. Suppose that you are employed as a data mining consultant for an Internet retail 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. First, I would advise the internet retail team to implement a-rules mining. This will allow new relationships between commonly searched terms to be uncovered. For example, an algorithm could be implemented using a-rules which could help to better place ads for recommended items on the page. If someone searches for a golf club perhaps a suggested item ad for sun block could appear on the results page. Other less obvious connections could be made using association rules mining leading to a more successful advertising platform and more sales.
   2. I would also try to use predictive modeling to try and forecast the hottest items for holidays so that our company can be sure to buy extra stock of those items before they become harder to source. We could even use this information to invest internally in companies we think will experience a lot of sales in the near future.
2. For each of the following data sets, explain whether or not data privacy is an important issue.
3. Census data collected from 1900-1950.

No census is public record

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

Yes, keeping this info private is important for all websites

1. Images from Earth-orbiting satellites.

Mostly no, perhaps some conflicting interests in the name of national security

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

No, publicly available information

1. Names and email addresses collected from the Web.

No, if available publicly on the web

Task 2:

Google Flu Trends was a project started at Google which used large scale, real-time data from the search engine to create a predictive model which would attempt to track the number of flu cases each year. While the project generated some interesting results, there was considerable criticism from the data science community: “Their technical criticism of Google Flu Trends is that it is not using a broader array of data analysis tools”. Professors at Northeastern University were able to improve the accuracy of Google’s model by including data from the CDC and other traditional sources. They criticize Google for over exaggerating the current capabilities of big data analysis and, for not utilizing more sources of data to better the prediction capabilities of the model.

A more recent article comes to the defense of Google. Flu Trends was a project started in 2008 “years before people started even talking about ‘big data’”. This early big data project, which the author argues to be a success, was in fact useful to the CDC and was able to be modified to have even greater prediction capabilities. The author defends GFT from the criticisms made by the professors at Northeastern University even going so far as to cite a study from Johns Hopkins University which found the data used in Google Flu Trends to be the only statistically significant source of forecasting data for predicting flu trends.

While the criticisms of Google Flu Trends are somewhat valid the authors seem to make certain assumptions to justify their points. Like the article which defends GFT I believe that Google Flu made no outrageous claims or unnecessary hype over their project. The fact that researchers were able to create a model which predicted the flu spread more accurately are not grounds to dismiss the project entirely as a failure. The largest problem I saw after reading both articles was the high expectations and over promising of new technology in the data science and big data fields. Somewhere down the road, big data and the possibilities contained within were over sold to the public. People expected technology capable of providing unparalleled insight given enough data without considering why we should exclude certain data. As is the case with GFT including certain data may improve the end result but muddle our understanding of the model and how it is working. Finally, as both articles stressed, something like GFT is best used in conjunction with other analytical methods to paint a more holistic picture of the subject that the data scientist is trying to study.