Jeff Levesque

Professor Gates

IST 565

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.
   1. Dividing the customers of a company according to their gender.
      1. False, this is a trivial operation.
   2. Dividing the customers of a company according to their profitability.
      1. False, this is a trivial operation.
   3. Computing the total sales of a company.
      1. False, this is a trivial calculation.
   4. Sorting a student database based on student identification numbers.
      1. False, this is a trivial database query.
   5. Predicting the outcomes of tossing a (fair) pair of dice.
      1. False, since the probability is constant for all possibility, this is a trivial known prediction probability.
   6. Predicting the future stock price of a company using historical records.
      1. True, stock prices are non-trivial since unlike tossing a fair dice, the probability of predicting stock prices is unknown, with varying level of confidence and error. Therefore, a model will need to be generated, to allow predictive analysis.
   7. Monitoring the heart rate of a patient for abnormalities.
      1. True, human heart rate is not exactly the same between one person and another, and even may change with a given individual. Additionally, when a model is created, anomaly detection can be implemented to find noise within the heart rate process.
   8. Monitoring seismic waves for earthquake activities.
      1. True, this is similar to the heart rate analysis. However, classification techniques can also be implemented. Specifically, past data can be used to train a model to help predict whether future attributes can be classified as an earthquake.
   9. Extracting the frequencies of a sound wave.
      1. False, extracting frequency of a sound wave does not involve generating a predictive model, since it is a simple extraction process.
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. Clustering: word phrases (or singular words) can be grouped by geographic location. This will allow the search engine to know what groups of words are used, for a given geographic location. Additionally, word frequency filter can be implemented. More specifically, clustering can be implemented by the top 100 most searched words. This will allow search engine to have the top 100 searched words to be aggregated into groups, per geographic location.
   2. Classification: once the exploratory analysis (i.e. clustering has completed), data scientist such as myself can label each cluster into a particular category, then train a corresponding model. On future instances, we can attempt to predict if certain words are used, what category the group of words belong to.
   3. Association rule mining: just like the implementation of rule mining with groceries, and supermarkets, a search engine can attempt to predict if particular words are used, what other words will be searched for. The combination of predicting what word, or words will be searched for, will allow suggestion techniques to be implemented by the search engine. For example, google provides similar alternative search options, while users are typing in the search bar.
   4. Anomaly detection: search engines can develop profiles of individual search tendencies. This will allow the corresponding model to develop conditional bounds of expectable search tendencies. However, if users search patterns exceed these bounds (perhaps particular category is search more frequently), then previous models (classification, association rule mining) will need to be adjusted. This places great importance for the anomaly detection, since it will signal when user patterns may be changing and may signal when previous models need to be updated.
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.
      1. No, the data is likely outdated.
   2. IP addresses and visit times of Web users who visit your Website.
      1. Yes, this information will likely be important with respect to marketing strategies. Therefore, for business purposed, this information should not be provided to competitors. Also, if the users require authentication, then confidentiality of who uses a particular system, is certainly a data privacy issue.
   3. Images from Earth-orbiting satellites.
      1. Yes, if the images are related to intelligence information. Otherwise, google images are available to the public.
   4. Names and addresses of people from the telephone book.
      1. No, this information is available to anyone.
   5. Names and email addresses collected from the Web.
      1. No, anyone can harvest public email addresses on the web.

**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/>

The google flu criticism, was premised on the fact that algorithms employed by google, consistently overestimated the flu trends later reported by the CDC. This problem was first seen in 2011-2012 (a few years after initial deployment). Specifically, the estimates reported by google was 50% higher than the actual values reported by the CDC. This motivated the data scientists at google to better engineer their algorithms. Later in 2013-2014 flu season, the reported values were 30% higher than the actual values found by the CDC. Though the improvements were significant, the overestimate was still significantly high. Analysts providing the report of the predictability of the google flu algorithms, reported that combining multiple techniques in conjunction with Google’s algorithm, significantly improved the flu trend predictability. Similarly, the founder of the google flu algorithms stated that the intention of the algorithms was to provide a general signal of potential trends.

Founders of the Google Flu Trends (GFT), found much criticism, as a team led by Northeastern’s David Lazer, reported that the GFT consistently overestimated predictions. However, what many have likely not considered, was that the GFT method was not a fully self-contained prediction tool. Instead, the prediction methodology was designed to be a general tool, complementary of the CDC analysis. The reasoning for this, was to allow an independent perspective, providing professionals multiple viewpoints of the Flu Trends. In a way, this allows medical professionals to triangulate, and use multiple techniques, to verify and understand whether multiple signals were reporting the same story. This process encouraged many others to take this base methodology and enhance the techniques themselves. Overall the implementation of the GFT has provided many, the understanding of how to combine multiple signals, to verify what is happening on a larger scale.

Algorithms are seldom perfect. However, the approaches, and techniques are many times invaluable, and more important than the exact predictability. As the latter article stated, sometimes new techniques can generate a significant amount of hype and expectation difficult to fulfill. During the release of the GFT, the hype generated expectations far beyond the scope of the products intent. Those not satisfied by the hype turned towards criticism. What most had failed to remember, or take note of, was that the GFT was an independent model, used to help signal potential flu outbreaks. Since this provided an alternative approach and provided the basis for others to pioneer significant improvements, the GFT was successful endeavor.