Diego Valdes

IST 707

Mar. 30,2019

HW 1

Task 1

1. Data mining
   1. No, this can be accomplished with a query
   2. Yes, this will require exploration of the data and some analysis to make a determination
   3. No, this can be done with an aggregation query
   4. No, this can be done by sorting the id column
   5. No, this is a simple probability problem
   6. Yes, this will require exploration of historical data, analyzing trends through the years to make a prediction
   7. Yes, this requires examining the patterns in the heart beat and making a determination when there are outliers (anomalies)
   8. No, because scientists know what measure of seismic waves determine an earthquake, they would simply want all the readings that exceeded a certain value.
   9. No, because the frequency of waves is easily recorded and can simply be looked up
2. If employed at an insurance company many of the techniques in data mining would be useful
   1. Clustering – For something like workers comp claims, it would be interesting to plot all the claims by time of day of the accident and see if there are clusters in the times of day workers are getting hurt. This might reveal that there are times of the day where injuries are most likely to occur
   2. Classification – For personal insurance (home, auto, etc), analyzing current customers and the number of claims that have been filed and paid out could be used to classify them in risk factors. With this information, models could be built to asses risk classification of potential customers.
   3. Association Rules – Just like online retailers, mining for products a customer buys in the same basket would also work for an insurance company. Looking at all of the coverages that customers have and creating association rules can be used to make recommendations on what types of coverages to sell to new and existing customers.
   4. Anomaly Detection – Best use of anomaly detection would be in determining fraudulent claims.
3. Data privacy
   1. Yes, census data prior to 1950 is searchable online by anyone. It doesn’t just provide anonymous stats, it is possible to search for individuals, who could still be alive today.
   2. Yes, IP addresses can be used to derive geographic location, determine online habits, shopping habits, and everything a user does on the internet. In that manner, it is possible to know more about a house hold from the information derived from an IP address than knowing their physical address
   3. Depends on how good one can zoom into those images. There are many examples of google imagery containing private moments. If the imagery is good enough to capture a person doing something in the ‘privacy’ of their backyard—yes, but if the images are just a bird’s eye view of nations, no.
   4. Yes. The telephone book is now online and provides a lot more information than just names and address. It is possible to get a person’s residential history, family, various public records, simply by typing their name into the white pages’ website.
   5. Yes. Email addresses are a common form of logon credential. That along with the person’s name and many of the other ways to get easily accessible information that can cause harm.

Task 2

**Google Flu Trends: The Limits of Big Data**

The Google Flu Trends algorithm, which was designed to predict the number of flu cases, had a record of consistently mis-predicting the number of flu cases in a given year. While improvements had been made between the 2011/12 up to the 2013/14 flu seasons, the algorithm still overshot the actual number of cases by at least 30% during those years. Critics mainly point to the Google Flu Trends being limited in both data being used and the analysis techniques being applied. Using 2 weeks of C.D.C data would have performed better than Google Flu Trends. An amalgam of Google Flu Trends and C.D.C data with further tweaks seems to be the better approach in the matter, according to analysts examining the tools performance. Defenders of Google Trends points that the purpose of the tool is being misrepresented, as it was designed more as an early warning system, as it had shown in the past of being able to provide advanced notification about the 2007/08 flu season and the 2009 H1N1 outbreaks. Google Flu Trends continues to monitor multiple countries and is reviewed every year for improvements.

**In Defense of Google Flu Trends**

Google Flu Trends, a tool which was celebrated for its ability to predict flu prevalence, came under fire for continuing to greatly miss the mark on the actual number. In fact, it has been pointed out that using the C.D.C’s traditional analytical methods performed better than Google Flu Trends. Although Google Flu Trend’s misses have been used to herald the tool’s failure, two facts determine that is not the case. First, Google Flu Trend combined with C.D.C data was shown to be the best predictor. Second, Google Flu Trend was built as a compliment to what the C.D.C already had; it was never intended to replace their monitoring methods. When these two factors are considered, it can be seen that Google Flu Trends isn’t the great failure as it is being portrayed. It can be argued that since the combination of the C.D.C data and Google Flu Trends is the best predictor, there is some value in the tool. There is obviously room for improvement, to determine why the predictions from Google Flu Trends have become unreliable. The root problem with Google Flu Trends isn’t that its predictions have gone astray, but, the public and media’s perception of what the tool is supposed to do is flawed.

**My Response**

I find the criticism of the Google Flu Trends to have some validity, mainly in the turn that the application isn’t as accurate as it once was, but not in the way that it was presented in the NYT article. The main argument in the piece is that Google Flu Trends is a failure. That’s not true. Clearly, there is value in the knowledge that with more data, it can produce the best prediction of all current methods. But that isn’t what the C.D.C wanted. Because the C.D.C didn’t want a replacement for what they were using, just another tool to supplement their current methods, I am inclined to agree more with the defense of the technology. The C.D.C got what they wanted, and if they want a better prediction method, they know how to get that now as well. I think the bigger problem is when new technology that is ground breaking, like Google Flu Trends, is announced, the public’s perception of the technology may not represent what is the true purpose. Here, the main article is losing sight of the role of Google Flu Trends in a larger system to make a point about using big data/data analytics. There is a valid point, however, in questioning the application. The data being used is what is dictating the result, which can be seen that adding the C.D.C data provided different predictions. The application is only as good as the data and algorithms that power it.