Course: IST-707
Name: Sharat Sripada
Homework #1
Date Submitted: 7/11/2020
Topic: Introduction to Data mining

Task 1: review data mining concepts and tasks

Exercise 1-7

Question-1

Data mining by definition is the process of automatically discovering useful information in large data repositories. Data-mining techniques are deployed to scour large databases in order to find novel and useful patterns that might otherwise remain unknown.

Further, looking up individual records using database management systems or web query are tasks related to information retrieval and do not classify as data-mining activities.

Classifying the following based on the definition:

a. **Dividing the customers of a company according to their gender**
This seems like a relatively simpler query to write to retrieve records from a table using a filter. It may not qualify as a data-mining task owing to its simplicity.

b. **Dividing the customers of a company according to their gender**
Not a data-mining task owing to its simplicity.

c. **Computing the total sales of a company**
Not a data-mining task owing to its simplicity.

d. **Sorting a student database based on student identification numbers**
Although this may need efficient sorting techniques (heap sort for instance is an efficient algorithm with time-complexity N log N) this would still not classify as a data-mining task.

e. **Predicting the outcomes of tossing a (fair) pair of dice**
With sufficient training data of outcomes for the pair of dice, it is possible to apply the steps of data mining and predict an outcome based on a model. This is a data-mining worthy activity.

f. **Predicting the future stock price of a company using historical records**
Ingesting historical records, visualizing patterns and building models to predict stock prices sounds a perfect recipe for data mining.

g. **Monitoring the heart rate of a patient for abnormalities**
Monitoring real time data gathered from pacemakers or probes on a patient by itself may not classify as a data mining activity. Using it as a means to build models and predict the likelihood of a patient suffering heart disease or heart attack is a useful application.

h. **Monitoring seismic waves for earthquake activities**
Not a data-mining activity as it is merely monitoring the seismic waves using certain probes or sensors watching for the slightest movements or shifts in tectonic plates (/seismic activity).

i. **Extracting the frequencies of a sound wave**
Not a data-mining activity.

Question-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.

The requirement for a search engine is to produce fast, accurate and relevant results constantly improving user experience (when context, location etc. is available); users of the search engine not having to:

- navigate through multiple pages of search results
- OR type multiple search strings for the same product (unless adding or modifying filters to be specific about their requirements)

Let's consider this via an example search string for a particular product – say, 'running shoes'

Clustering

A search engine would typically index several web pages at the backend. Extending the document clustering example in section 1.3 of the textbook, web pages could be clustered based on relevance while grading and ranking them by a probability of an arbitrary algorithm.

For this specific search string, web sites that sold shoes could clustered based on make, color, gender, type (running, casual, formal etc.), ongoing sales and brand popularity (say via a popularity index 1-5, 5 being most popular)

Could look like this:

Web site	Products	Relevant
1	Shoes: 100, Men shoes: 33, Women: 33, Children: 34, Color-1: 10, Color-2: 15, Sales: None, popularity-index: 3 and so on	Yes
2	Shoes: 100, Men shoes: 33, Women: 33, Children: 34, Color-1: 10, Color-2: 15, Sales: 2, popularity-index: 2 and so on	Yes
3	Shoes: 500, Men shoes: 166, Women: 166, Children: 168, Color-1: 50, Color-2: 75, Sales: 1, popularity index: 5 and so on	Yes
4	Electronics – Inventory or relevant for a store like say Best-buy	No

Table-1 – Clustering table

Classification

Next, once the data is clustered, classification can help identify the likelihood of the user buying a type of running shoe. This could be based on:

- customer data from past transactions
- location

- OR general trends if neither of the above is available

If relevant data as below across web sites and products be rendered, that can help determine the likelihood of a hit or miss (target variable – Bought). Further, this could translate to ranking a web page:

Web site	Gender- Male, Female, Child	Color-1	Color-2	Sale	Bought
1	M	1	0	0	0
1	M	0	1	0	1
2	M	1	0	1	1
2	F	1	0	0	0
3	C	0	1	1	1
3	C	0	1	1	1

Web site	Purchase probability	Rank
1	0.25	3
2	0.25	2
3	0.5	1

Table-2 – Classification and ranking

Association rule-mining
Association rule-mining can lead to holistic shopping experience for online web users – bundling frequently bought products means fewer clicks and screens. And, also higher revenue.

Extending the example in section 1.2 of the textbook to this specific use-case:

Web site	Transaction	Items frequently bought together
1	1	{Running shoes}
2	2	{Running shoes, socks, cap}
3	3	{Running shoes, socks}
3	4	{Running shoes, dri-fit clothing}

Table-3 – Deriving association rule-mining

The patterns also help web sites put out promotions on new combinations and get shoppers excited about trying different but relevant products. This could perhaps show up as ‘Promotion of the day’ or other like on the search page.

NOTE: The data likely coming from web sites needs to bubble up to the search engine so Ads or pages can be ranked accordingly. And, since Ads generate revenue, this is a win-win scenario for everyone in it.

Anomaly detection
Detecting outliers is essential to preventing a fraudulent transaction as seen in example 1.4 of the textbook.

In this context, a sudden spike in hits to a web site that is ranked lower by the engine (when the ranking already considered all factors) should raise an anomaly at the backend for further review.

Task 2: practice your critical thinking and writing

Google Flu Trends: The Limits of Big Data

The article talks about inaccuracy in Google Flu Trends' big-data engine that consistently over-estimated flu cases for several years prior to 2014 when the article was published – the estimates were high in 100 out of 108 weeks, a whopping 92% of the time. The four social scientists that co-authored the article, attack Google's approach that it did not consider conventional methods of data collection and analysis. And that they weren't against the idea of big-data or use of data-science but strongly opined that it be used alongside a broader range of existing data analysis tools. They proved through their study, that combining data from Google Flu Trends with C.D.C data was far more accurate.

In Defense of Google Flu Trends

The founders of Google Flu Trends while acknowledging the over-estimates, in their defense said it was never built as a standalone flu monitor but was meant to be complementary to existing signals. That is, it should not replace but indeed be used in the manner some researchers were already doing so, combining it with C.D.C data while making necessary adjustments. Yet, the belief was to keep it exclusive and use the best of both worlds when it mattered. The author reflects on the hype around new technology and how it led Google Flu Trends to fail in common public view. Finally, in defense he concludes that researchers had still found Google Flu Trends useful and this was seen through several citations across many fields.

My View

In my opinion, the criticism is far stronger than the defense. In the field of data related to disease (or flu in the context of this topic), it helps to have fewer parallel interpretations. It is clear in both articles, the public and to a larger extent the research community did not see the benefits of Google's standalone model although they were in agreement with the technology of big-data or data-science that backed it. While the idea at the outset was novel, the creators of Google Flu Trends should have considered working alongside C.D.C research, built more accurate models and extended the benefits of big-data to generate accurate predictions.