IST 565 Data Mining

Olivieri

HW 1

1.1

(a) Dividing customers of a company according to their gender does not seem like a data mining task – it can be done easily without clustering or associating any other variables.

(b) Dividing customers of a company according to their profitability is a data mining task as it can be done with clustering.

(c) Computing the total sales of a company is more of an accounting task than data mining.

(d) Sorting a student database based on student id numbers is a database administrator task, not data mining.

(e) Predicting the outcomes of tossing a pair of dice is statistical but not data mining. There will not be clustering, association, or anomalies in the results.

(f) Predicting the future stock price of a company using historical records is data mining:

clustering, and anomaly detection can be used in predictive models and applied to the problem.

(g) Monitoring the heart rate of a patient for abnormalities is not a data mining task.

(h) Monitoring seismic waves for earthquake activities is not a data mining task.

(i) Extracting the frequencies of a sound wave is not a data mining task.

1.2 Explore various online applications for data mining. Categorize what are the three applications where predictive tasks are required. Also try to identify the applications in daily routine where descriptive tasks are required.

* Spam detection, an automated chat bot for customer service, and financial analytics are applications where data mining, clustering, associating, and anomaly detection can be used in predictive models. Gmail uses spam detection. Automated chat bots for customer service are becoming increasingly popular with web applications. And financial analysis is used by large corporations to allocate funding to projects or investments.

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

(a) Census data collected from 1900-1950:

- No, census data is public.

(b) IP addresses and visit times of Web users who visit your Website:

- This is a privacy issue. A person has a reasonable expectation that these times will not be published publicly.

(c) Images from Earth-orbiting satellites:

- No, this is not a privacy issue. Nothing private can be collected or released from

satellites.

(d) Names and addresses of people from the phone book:

- No, this is public information already.

(e) Names and email addresses collected from the Web:

- This is definitely a privacy issue. Names and emails should not be collected and then distributed publicly.

2.1 Write one paragraph to summarize the criticism, and another paragraph for the defense.

* The first criticism of Google Flu Trends presented the fact that Google Flu Trends was inaccurate. Over many years, Google Flu Trends consistently posted higher predicted cases of the flu than actual reported values from the CDC; one year, more than 50% higher. Another criticism leveled at Google Flu Trends was that they favored higher-tech, big-data-style solutions over less fancy analysis.
* The defensive of Google Flu Trends rests on the sentiment that the tool was never designed as a replacement to current tracking strategies, but rather as a separate signal. As a separate signal, Google Flu Trends can be incorporated into analytical models to enhance the model. Researchers found this to be true: incorporating Google Flu Trends into existing models as well as combining Google Flu Trends results with other results painted a clearer picture than before it was invented.
* Both the criticism and defense make valid points. If Google Flu Trends is consistently inaccurate, it could be improved. However, Google Flu Trends, as the defense states, was not meant to, and should not, replace other statistical reporting and analysis tools. It should be used as yet another source of data rather than the only source of data.

The commotion over conflicting views on Google Flu Trends is similar to other conflicts over technology. Mainly, the hype of technology never lives up to the actual results of the technology; there’s always a let-down. However, typically, the creators of the technology are not the ones who are creating the hype. Or, if they’re encouraging popular acclaim of their nascent technology then that is often necessary to be noticed in a quickly-evolving landscape that favors new over reliable and requires large amounts of investment or profit to be sustainable. The real test of any technology’s practicality comes after the hype-cycle when the technology can be used reliably, either stand-alone or integrated into other a larger technological context.