Student: Alireza Zarrinmehr

1. A. Dividing the customers of a company according to their gender

No. The objective of this task is neither predicting the value of a particular attribute based on another attribute nor driving a pattern that summarizes the underlying relationship between data.

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

No. This is an accounting computation that is then put into practice by applying a threshold. Data mining, however, would be required to forecast the profitability of a new client.

1. C. Computing the total sale of a company

No. The objective of this task is neither predicting the value of a particular attribute based on another attribute nor driving a pattern that summarizes the underlying relationship between data.

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

No. The objective of this task is neither predicting the value of a particular attribute based on another attribute nor driving a pattern that summarizes the underlying relationship between data.

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

No. This is an estimation of probabilities because the die is fair. This is more in line with the issues that data mining takes into consideration if the dice were not fair, and we had to estimate the odds of each event from the data. However, in this situation, mathematicians came up with answers to the issue.

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

Yes. We would try to develop a model that can forecast the continuous stock price value. This is an illustration of predictive modeling, a branch of data mining. Although scholars in numerous domains have created a wide range of strategies for forecasting time series, we may use regression for this modeling.

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

Yes. We would create a model of the heart rate's typical behavior.
and sound an alarm if any strange heart behavior was seen. By doing so,
involve the anomaly detection branch of data mining.

1. H. Monitoring seismic waves for earthquake activities.

Yes. In this scenario, we would create a model of the many seismic wave behavior patterns connected to earthquake activity, and we would sound an alarm whenever one of these patterns of seismic activity was noticed.

1. I. Extracting the frequencies of a sound wave.

No. The objective of this task is neither predicting the value of a particular attribute based on another attribute nor driving a pattern that summarizes the underlying relationship between data. This may be signal processing.

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?

Clustering: In addition to showing results for the keyword " iPhone XS cover" the search engine should also provide results for "iPhone XS LCD protection, “iPhone XS charger", "Protective Cases" and "Apple Care" when the user types in "iPhone XS cover"

Classification: The search engine might offer news linked to a keyword. This is accomplished by applying a classification algorithm to the keyword after applying classification rules or decision trees.

Association rule: A person looking to purchase a laptop online might also be considering a new bag for carrying the laptop. The search engine should also return results for bags that are made to carry laptops.

Anomaly detection: if a website like a company's website always takes seconds to load but today it is taking more than usual, the website may be under maintenance. Therefore, the search engine may want to stop showing the results for that website and alert the website owner that there is a loading speed problem.

3. For each of the following data sets, explain whether data privacy is important issue.

(a) Census data collected from 1900-1950.

Yes. Since someone born in 1950 would be 72 years old now, census data could pose a privacy concern. The census data includes financial details that some people would feel awkward disclosing on websites. It also includes information on medical issues. Your father’s name is one example of additional census information that could be utilized to get around security questions.

(b) IP addresses and visit times of web users who visit your website.

Yes, there may be privacy issues with this information. There are many instances of websites being compromised and user data being leaked, which hackers can use to blackmail individuals.

(c) Images from Earth-orbiting satellites.

No, this is not a problem right now. However, this can be a problem if more in-depth information becomes available. For instance, photographs of license plates, faces, etc. should be hidden in Google Maps street view.

(d) Names and addresses of people from the telephone book.

The telephone directory doesn't pose any data privacy problems for names and addresses. information has been made publicly available. However, the use of this information for calling people and spamming them may cause issues.

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

Just like the last example names and email addresses on the web are also not a data privacy issue. However, these data should not get used for spamming people.

Task 2. 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?

In addition to overestimating the number of flu cases in the United States during the 2012–2013 flu season, Google's flu-tracking service has also frequently overshot in recent years. This is one of the criticisms of Google's flu-tracking service. Critics also debated the algorithm's effectiveness as a standalone flu monitor. However, As Matt Mohebi, co-inventor of Google Flu Trends, pointed out; a lot of the current criticism ignores the fact that the service was always meant to be a "complementary signal" rather than a standalone forecasting tool. Investigation demonstrates that the greatest results can be achieved by combining Google Flu Trends with C.D.C. data and using a few adjusting strategies which means Google's flu-tracking service was a helpful effort and did add value.

Google Flu Trends is not a miraculous device that uses an algorithm to take the role of the CDC. The folks who created it, however, had no idea that it would be. If it failed, it most likely failed in the public's perception and the desires of superficial Big Data adherents. GFT data are readily available, allowing individuals to create these more complex models, which they did. Consider the research from Johns Hopkins. In this research, a team looked at how to create a better influenza model, starting with clinical data from emergency rooms and attempting to add any additional information that might be useful, such as variables like "GFT, meteorological data, and temporal variables. They discovered that the only external data source that produced statistically meaningful forecast enhancements over the original model was Flu Trend data. Google Flu Trends and its methodologies have shown to be valuable and pertinent to researchers in epidemiology and beyond. Over 1,000 citations in a variety of disciplines have been made to the original Nature paper that described the experiment.

It is undeniable that businesses need to be more forthcoming about outlining the dangers and potential problems associated with new data mining tools like Google Flu Trends. The first service was not designed to forecast flu cases, serve as a substitute for established surveillance systems, or eliminate the need for laboratory-based diagnosis and surveillance. GFT intended to offer such a nearly real-time signal, and it has demonstrated that it is useful. It's important to remember that Google's big data analysis was just the beginning of what is conceivable. Therefore, expecting the ideal result is unreasonable. Even though GFT and other comparable big data applications might not be the ideal solution for a specific issue, they might extend the researcher's perspective with current experience and offer insights for resolving challenges in the future.