Jacob Dineen

IST 565 Data Mining

Homework 1

1.7

Task 1: Review Data Mining Concepts and Tasks

1.

1a) Dividing the customers of a company according to gender would be more of an example of descriptive statistics, and would not need to employ data mining techniques to find the answer. This would be more of an information retrieval task.

1b) Dividing the customers of a company according to profitability would also fall more under the descriptive branch, if we were simply focused on demographic data. This information is known by looking at the current data and segmenting the audience into different levels, or factors, of profitability. If we were looking at specific behavioral or economic variables in relation to profitability, that would probably be a data mining task.

1c) Computing the total sales of a company would also fall more under descriptive analysis. The data is already known, and simple arithmetic calculations would return our outcome.

1d) Sorting a student database based on ID numbers would fall more under the information retrieval definition, rather than data mining.

1e) Predicting the outcomes of tossing a (fair) pair of dice could fall under the data mining branch. We could be looking at the past X amounts of rolls in a classification problem. Employing data mining techniques could help to uncover patterns in previous sets of rolls, and help us to predict upcoming results.

1f) Predicting the future stock price of a company using historical records would be an example of using regression, as our dependent variable would be continuous in nature. Depending on the amount of variables that we have documented, we could identify patterns through data mining algorithms that would be unknown to the human eye.

1g)Monitoring the heart rate of a patient for abnormalities could be a cross between descriptive and data mining. It could be descriptive in regards to using something like a moving average, or something along those lines. But to counter that, we could use data mining techniques to cross-examine other patients with similar attributes to increase our sample size, and to better understand irregularities.

1h)Monitoring seismic waves for earthquake activities like utilizing both descriptive and data mining techniques. A lot of forecasting relies on predictive measures, rather than descriptive measures, to be able to stay ahead of the curve and issue warnings before it’s too late.

1i)Extracting the frequencies of a single sound wave would be more descriptive in nature. We wouldn’t necessarily be looking for patterns, because we are looking at an isolated event, and not trying to understand that event as part of a system.

2.

If I were employed by a search engine company, my main point of business likely falls underneath the advertising hierarchy, as that’s where our revenue generation is based. Having worked with adtech companies in the past, I understand that there’s uses for both classification and regression models. A lot of the classification models centered around user engagement for a particular ad, based on personal identifier information/cookies, as well as information about the ad itself. The regression models came into play when dealing with the bid prices for particular sets of inventory, as any kind of cost metric is continuous in nature. Audience clustering can come into play, for example, if we are selling ad space, or query results, for luxury cars. We’d want to cluster together a high income, potentially older, college educated audience to increase potential engagement- Or really just whatever the training data tells us is the cluster of user attributes more likely to engage with an ad. Showing an ad for a Mercedes to someone that is less affluent may result in lower engagement- Clustering can be used with association rules mining to determine socioeconomic clusters as they relate to our desired ‘right hand side’ outcome. Association Rule Mining is used for many recommendation engines. We could potentially even throw this into our classification/SVM model as well – For example, if a person visited this website, and that website, they are more likely to have a higher support level for clicking through this ad. Anomaly detection is useful when we have an outlier to the pattern. Let’s take the above example, about Mercedes. If we have a low income individual click through an ad, maybe it was a misclick, or maybe it was just for research purposes, we don’t necessarily want that users attributes to be included within the current model, as it could distort our target audience, and potentially lead to misidentification.

3.

I believe that many of these cases are subjective in nature, but here’s my best stab.

3a) Census data would not be subject to increased privacy because, as far as I know, we are dealing with aggregated numbers, and not individual records. Also, Census data is public.

3b) IP addresses and visit times of Web users is a case where data privacy is very important, although these types of data are often stored and tracked through cookies, and much of the time the data is sold or collected to use for marketing efforts.

3c) Images from Earth-orbiting satellites, I imagine, are subject to levels of confidentiality. There may be certain data/images that the general public are not disclosed. I think information sharing comes into play here as well- Data may not be disclosed to prevent foreign government interference.

3d) Names and Addresses from the telephone book, in my opinion, is semi-dangerous on a personal level, but efficient for business. You could opt out of appearing in the directory, for a fee, interestingly enough.

3e) Names and email addresses collected from the web would certainly seem to be an area where data privacy is an issue. At a business-level, exposing this information helps to promote brand awareness and ease the sales process, but this information at an individual-level could lead to unwanted attention/harassment/stalking, etc..

Task 2: Practice your critical thinking and writing

Summary of criticism

The topic in question is Google’s flu-tracking service, which attempts to predict/forecast the number of recipients impacted by the illness year over year. The article assesses that, out of 108 weeks of examination and prediction, Google predicted count was higher than the actual count, with a 50% error rate coming from the ’12-’13 flu season. The argument is that the algorithm Google uses for such a prediction is less accurate than simple analysis, and moving averages. Google countered by stating that the service was never meant to be used as a stand-alone tool, and would benefit from the infusion of third party data.

Summary of defense

Starting with the New York times article, there are clear defenses made for the technology towards the latter stages- “In the 2009 Nature paper, they showed that it had given that advance indication in the 2007-08 flu season — and that it would again during the 2009 H1N1 outbreak”. So even if the error was high, the signal appeared to work, at least during this timeframe. What was found during independent studies, as mentioned in the previous paragraph, was that when CDC data was merged with Google’s data, the predicted result was better than either stand-alone effort. If judged as an isolate, Google’s flu-tracking could be deemed a failure, but if it was judged as a part of the whole, an open source modeling effort, of sorts, it was anything but.

My own thought

To side with the criticism of Google’s flu-tracking service would be to side against innovation, in my opinion. They had a clearly defined goal that they were to act as a complimentary signal for identifying and counting flu cases over a time series. To say that they failed because their predicted outcomes were high would be to discount the impact they had in assisting more accurate predictions. I think we can look at Google search queries/results as part of the equation, and I’m sure they knew that those variables alone wouldn’t account for the variance in flu counts over a period of time, but as an accessory to more vast data from the CDC, you add statistically significant variables that increase the adjusted R2 over the model. What would be interesting is if Google, if they were to act as a standalone signal, were to switch their domain over to a classification problem identifying flu seasons. Rather than having a continuous dependent variable, have it be binary, and try to map specific time frames that are more susceptible to flu like symptoms across filtered geographical instances.