Kendra Danforth Osburn

IST 707 | HW 1

January 16, 2019

*NOTE: Regarding formatting, I understand the instructions indicated a max 1” margin. However, I wasn’t sure how best to structure the document (including questions, numbers and bullets) to maintain clarity and readability, so I defaulted to what I’m familiar with. Please let me know if entirely left-aligned against the 1” margin is preferred in the future.*

**TASK 1: Review data mining concepts and tasks**

1. Discus whether or not each of the following activities is a data mining task

*NOTE: I could make a case for some of these being “data mining tasks,” but most of them are basic single-line SQL queries or fell into the “I could easily do this in excel” category which I -- possibly incorrectly -- categorize as “data* ***retrieval*** *tasks.” In short, working together, these tasks (clustering, indexing/sorting for example) make up parts of the whole of data mining. However, as the parts alone, these are simply data retrieval tasks. At least, that’s how I read the question, especially considering part 1.1 of the reading.*

* 1. Not a data mining task -- a data retrieval task (but could be considered clustering)
	2. Not a data mining task -- a data retrieval task (but could be considered clustering)
	3. Not a data mining task -- a data arithmetic task
	4. Not a data mining task -- a database indexing task
	5. Not a data mining task -- a data arithmetic task
	6. Not a data mining task -- a regression algorithm
	7. Not a data mining task -- put together with many other abnormalities in an attempt to better understand patterns in abnormalities is a data mining task
	8. Not a data mining task -- a monitoring and recording task
	9. Not a data mining task -- a data recording task

1. Suppose that you are employed as a data mining consultant for [insert ideal company here]. 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.

As a (fictional) employee of Google within their Google Photos division, I want to ensure consumers have an easy and enjoyable time sorting, organizing and sharing their many photos. Having (hypothetically) worked here for a few years, I’ve noticed some trends. People take photos for many reasons -- some are sentimental, some are not. Some contain people, some do not. Some are screenshots; some are selfies; some include cats. In the beginning, I start experimenting with clustering by “has face” and “does not have face,” and realize this isn’t nearly specific enough to help the majority of my clients organize their photos. Next, I cluster by date and categorize by capture method -- front-facing camera, back-facing camera, or screenshot. I then examine the data of routinely discarded photos and find some key similarities. Many discarded photos feature a negative brightness level, and are part of a series of photos taken in rapid succession. In this case, I use association rule mining to suss out other similar photos. Because many photos with negative brightness also are underexposed, I start suggesting underexposed photos to users as photos that might be appropriate to discard. Finally, I use anomaly detection to identify photos that are not part of a surrounding series. Previously, I would have categorized all of a user’s photos captured between 11-2-18 and 11-5-18 as belonging to a “Trip to Napa” series—but some of these photos are actually screenshots, and contain no camera data other than the time they were captured. As a result, I can safely guess that these screenshots should not be included in the “Trip to Napa” highlight reel that is automatically created and sent to this user.

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

*NOTE: As the textbook started to talk about in towards the end of section 1.6, privacy breaches bring up new challenges for the data mining community. Again, a case could be made on either side for each of the following examples. However, I believe that the purpose of this question was for us to arrive at the “data privacy is always important to consider” answer.*

* 1. Census data collected from 1900-1950
		1. Privacy important: Census data contains very personal information that could be used inappropriately in the wrong hands (e.g. by insurance or lending companies to bias against coverage or lending).
		2. Privacy not important: This data is largely irrelevant now.
	2. IP addresses and visit times of Web users who visit your Website
		1. Privacy important: As a web user, I’d like to keep my data private so I’m not aggressively targeted with ads.
		2. Privacy not important: As a marketer, I’d like to know information about who is coming to my site (and from where) so I can better understand how my product is being used and to optimize against this data. Additionally, as the internet is an open space that the user opted-into when coming to my site, I am free to take that data.
	3. Images from Earth-orbiting satellites
		1. Privacy important: What images are these satellites capturing? Does my right as a landowner grant me rights to the airspace above my property? Where does ownership begin and end?
		2. Privacy not important: When we were born on this planet, we unintentionally opted into being part of the greater earth and human species and this includes the satellites our government created to protect us and foster things like “the internet.”
	4. Names and addresses of people from the telephone book
		1. Privacy important: Having phone numbers and addresses out in the open is a terrifying prospect for many people -- someone could just show up at your house.
		2. Privacy not important: This information is already largely available in the anachronistic tombs we unwittingly receive on a bi-yearly basis.
	5. Names and email addresses collected from the web
		1. Privacy important: As a web user, I’d like to keep my email address private for only those people I know to utilize.
		2. Privacy not important: As a marketer, I’d like to have access to a potential customer’s email address so I can follow up with them with promotions and other offers.

**TASK 2: Practice critical thinking and writing**

**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 defence make sense? What other problems or benefits do you see in Google Flu Trend or similar big data applications?)**

In “Google Flu Trends: The Limits of Big Data," Steve Lohr criticizes Google Flu Data for overshooting flu predictions “by about 30 percent.” Lohr uses this discrepancy to highlight what he calls “big data hubris,” the tendency of 21st-century researchers to completely eschew “traditional data analysis” in favor of “big data sets.”

“In Defense of Google Flu Trends,” by contrast, comes to Google Flu’s defense. The author, Alexis C. Madrigal, claims that the only real failure of Google Flu Trends is in the “popular imagination,” where Google Flu Trends didn’t live up to the public’s unrealistic expectations of a product that was intended only to aid the CDC. Google Flu Trends, Madrigal maintains, never claimed to be able to accurately predict flu data, nor was Google Flu Trends ever intended to be used in a vacuum without other data points, such as those provided by the CDC.

I believe that Madrigal’s defense of Google Flu Trends is much more persuasive than Lohr’s critique. In the first article, Lohr treats Google Flu Trends as if it is the “end-all-be-all” solution in flu prediction. His conclusion that big data has failed is based on the premise that Google Flu Trends set out to replace or supplant the CDC’s existing dataset. In reality, Google Flu Trends merely sought to provide supplemental data that the CDC and other organizations could compare against their existing datasets to draw stronger conclusions. These articles show how important it is to view data in the proper context. After all, datum are not capable of hubris, but humans are. Data analyzed in a vacuum is largely unhelpful, and can be misinterpreted in either direction. However, when the intent behind a dataset is clear, and it is used in conjunction with other available datasets, we end up with a much clearer, more cohesive worldview.