Nathan Bunch
Data Science I
Ryan Yates
9/15/2017

Homework I

1. Often data we will encounter as data scientists is derived from human individuals. Describe three challenges or dangers to encoding humans and human experiences as data.

There are many challenges/dangers that we face whenever we attempt to encode data from humans. Three that come to mind are: memory, communication, and biological error. In the memory error, while collection data from humans, a person may give details as to how something was and details of that observation and while doing so, may leave out some important details. This lack of data, because of the lack of remembrance of that data, is a serious error due to the fact that a single fact could change a significant section of the dataset being collected/compiled. In the second error, a person may fail to communicate the details of a subject being recorded. Since all humans are unique and therefore think uniquely, the way that every human communicates thought and details are all slightly different. Even though some may speak the same language, this does not mean that the language that each person shares is interpreted the same when it is communicated. The final error explored here, the biological error, is similar to the memory error, but slightly different in the way data is recalled and communicated. In the biological error, it is not the lack of data being presented, it is the error in the details that are presented. Because humans are always changing, their perception of events and details are also always changing. Memory is not always constant, and therefore when a detail changes from the original to something else, this results in a biological error. This error is probably the most dangerous as it is the error that is nearly impossible to correct and can completely destroy an accurate dataset.

2. At your job you are tasked with scheduling repairs to the roads in your city. The transportation department has a traffic app that helps drivers avoid highly congested areas. In addition this app records an entry in a log every time the accelerometer on the smart phone detects a bump that could be large enough to be a pot hole. Each log entry has a GPS location, speed, time, and user id. Describe some techniques you could use to turn a collection of these logs into data useful to your job. What fields or columns would you have in your final data?

The first step i would take is sorting all the log entries by the user id, then afterwards, sort each of those individual user id lists by time the accelerometer was triggered. Once this is done, I can correlate where the potholes are based on matching the GPS coordinates. I could potentially go further in analyzing the data and cleaning it up by removing the entries where they occur during rush hour, as cars will be breaking more suddenly (most likely) during those times. The final result of the collection and encoding of the data would result in a data frame with the columns entitle as follows: “userID”, “time”, “gpsLocation,” and “speed.”

3. Discuss at least three dangers to exclusively fixing roads with pot holes found from data from the app.

Three dangers made possible by this app could be: lack of some sort of pot hole size measurement, the entering of incorrect measurement data (stopping suddenly with a car, in theory, would produce a similar jerk that occurs when hitting a pothole), and since the app also helps drivers avoid where there is a lot of traffic, the places where the roads might be in worse condition may not even be getting recorded.

4. Your city traffic app communicates traffic information, describe how it might be extended based on the pot hole information you have derived.

The app could be extended where it would help people get to know which roads are more damaged and need repair than others. Over time, the app would get to know what roads need repair and which don't.

5. Download the following pothole data: potholes.csv. Determine how many likely potholes are in the data set. Which pothole has been noticed the most? Give your full reasoning for your answers.

To determine the total number of potholes in the dataset, we first take the latitude and longitude values we have been given, and compare them to each other. We do this by looping through both columns and dividing each by the following measured value in their respective columns. If the values have less than a 2% difference (both positive and negative), we count it as the same pothole and stare that count. Once we can determine how many potholes are in the set, we can then determine which pothole has been encountered multiple times. The reason we compared the values measured with a 2% difference was to account for any error made the cellphone of the user.

This first method did not work, maybe in theory…but not in practice.

In my second method to solve this analysis problem, one of my friends had suggested to graph the values (with X being the latitude and Y being the longitude). I attempted this and I was pleased to find that the data did indeed show where the potholes were, however, I still have a major issue: how many times had a certain pothole been crossed over in the dataset? After asking for a hint as to how I could do just that, I used the kmeans function to grab the count of the number of times a particular pothole was encountered. I have discovered that there is a particular pothole located at (42.42346, -78.15796) that is encountered the most (31 times).

Then I realized that the means function wasn't working for me.

I discovered in the last moment that I wasn't using the kmeans function properly. Rather than discovering right now as to how I wasn't using it properly, I decided to use a histogram to graph the frequency of the plots of the pothole coordinates. Now I know for certain that the data is being analyzed properly. Out of all the potholes, the one at 42.4265 latitude is by far the one that was encountered the most.