2 Data Collection With Transit Tweets

Your friend working at the Los Angeles Department of Transportation has been given the task of determining how transit riders feel about Los Angeles's public transit systems. Your friend wants to accomplish this by scraping Twitter for tweets containing keywords and hashtags related to Los Angeles public transit and running them through a model that does sentiment analysis (the algorithm will say whether a tweet contains positive, neutral, or negative sentiment).

(a) What are some of the issues, if any, with what your friend proposes?

The problem here is that your data is somewhat biased because all your data from Twitter users. Another problem is that when people feel okay about public transit, they don't just go Twitter and tweet about it. There are more important things to do in life. But when people have a bad experience using the transit, they will complain and more likely to spend extra effort and tweet about it. Thus, many of the tweet data you collect would probably have negative sentiment, which doesn't represent the actual distribution of transit riders' feelings.

3 Model Extensibility

You recently learned about Google's new system for detecting breast cancer in mammograms (https://www.nature.com/articles/s41586-019-1799-6). The system was trained on a large dataset of annotated mammogram images from the UK and the USA, most of which were acquired on devices made by Hologic. The paper shows that the model can be trained on the UK dataset and still perform well on the USA dataset. Your friend finds the work exciting, and would like to use Google's pre-trained model to detect breast cancer in Brazil.

(a) Is this a good idea? Why or why not?

No it is not because the model was trained on images gathered from people living in UK and USA. The model performing well on UK dataset is not very surprising because the model was trained on both UK and USA dataset. Therefore, we can't assume it will perform well on Brazil dataset because the model was never trained with images from Brazil, and people in UK and USA might have different trends/features than people in Brazil in terms of breast cancer.

4 Experiment Design

You would like to see if you can predict the probability that a given student will stop attending any particular lecture.

(a) What are some features you would try to gather to investigate this problem (e.g. student's year in school, professor teaching the course)?
(b) How would you formulate your labels?
(c) How could you source/obtain/gather the above data?

(a) Professor's rating on Bruinwalk, number of quarters attended, number of lecture stopped attending, average number of units taken per quarter

(b) Taking a few lectures off does not count as stop attending a lecture. So I would formulate my label as follows: A student stops attending a lecture if and only if he/she missed more than 90% of the lecture days.

(c) I could open a poll online with some incentives (e.g. money) to ask students to fill out a form where I ask questions from which I can extract information mentioned in part (a). Note the data might be unreliable because students could lie when answering questions.

5 True or False

Provide brief explanations for your answers.

(a) All data science investigations start with an existing dataset.

(b) Data scientists do most of their work in Python and are unlikely to use other tools.

(c) Most data scientists spend the majority of their time developing new models.

(d) The use of historical data to make decisions about the future can reinforce historical biases.

(e) If you have a dataset where data on income are stored as integers, with 1 standing for the range under $50k, 2 for $50k to $80k and 3 for over $80k, the income data is quantitative.

(a) No. We probably need to create new dataset if it is very domain specific. For example, if I want to train a model to predict hours of studying per day of a UCLA student, I would need to ask students questions and gather data. There might be a dataset for university students in general, but probably there isn't one for a specific university.

(b) No. As we learned in class, Python is popular but not necessarily the only tool for data science. For example, you could use R or Java.

(c) No. As we learned in class, most of the time is spent on cleaning dataset.

(d) Yes. Because when we use data to make predictions about new unseen data, we essentially find a pattern in the new data and see if we observe similar patterns in historical data, then we make a prediction based on that.

(e) Yes. The data are numbers and those numbers have meanings (i.e. The greater the number, the greater the income).

6 Probability

A jar contains 3 red, 2 white, and 1 green marble. Aside from color, the marbles are indistinguishable. Two marbles are drawn at random without replacement from the jar. Let X represent the number of red marbles drawn.

(a) What is P(X = 0)?

(b) Let Y be the number of green marbles drawn. What is P (X = 0, Y = 1)?

(a) C(3,2) / C(6,2) = 1/5

(b) C(1,1) * C(2,1) / C(6,2) = 2/15

7 Imputation

In Project 1 you learned about imputing data, the step a data scientist must take to deal with missing or null values in a dataset.

(a) List four different strategies you could reasonably use to address null values. For each, clarify what the advantages and disadvantages to it are. Additionally, for each strategy speculate on what sorts of datasets it would be the most effective, as well as what types of data it is inadvisable for.

- Mean: replace null values with mean of that column. If the dataset is very close to normal distribution, for example height of university student, then it makes sense to use mean value to replace null. On the other hand, If the dataset is skewed, for example, 90% of data has income below 100k and 1% is above 10000k, then mean is not a good idea to fill missing values.
- Median: replace null values with median of that column. If the dataset is skewed, for example, 90% of data has income below 100k and 1% has above 10000k, then median makes more sense. On the other hand, using median value does not make sense when missing means 0. For example, if a dataset has a movie with null reviews, then it doesn't make sense to fill median values there.
- Most Frequent: replace null values with most frequent data of that column. If the dataset is consists of people's answer to the question "what's the color of an apple?", then most of the data will be red. When there is a missing data, it makes sense to fill red. Using most frequent data is inappropriate when data is continuous. For example, it doesn't make sense to use most frequent income to fill the missing values because most of the data in the table will have a frequency of 1.
- Constant: replace null values with a constant. This method works well when null means 0. For example, if the number of movie reviews is null, then it means no one has reviewed it yet, and thus makes sense to put 0. On the other hand, when computing missing income of a person, for example, it's inappropriate to arbitrary pick a number and use it for imputation.