CS 188 HW 1

Homework Partners:

- Rosa Garza UID: 505315097

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?

- **Sample/Population** may not be the best - Is the population that tweets a good representative of the sample that actually uses LA public transport? Probably not, because not everyone goes to Twitter to tweet about it. If we want to stick to the scraping strategy, we might want to consider additional social media websites like Facebook and Instagram to see what people are really saying, and to cover a wider demographic (different age groups use different social platforms). May also need to perform manual surveys. May not target actual daily riders.
- **Bias in Tweet Distribution** - Might only tweet about negative things (very rare that people tweet about positive things) - can lead to a class imbalanced dataset (especially more so with neutral tweets). People rarely post about neutral things.
- **Chatting Language** - People use ‘chat’ or ‘texting’ language because of the 280 character limit on Twitter, resulting in a large bag of ambiguous words that could skew the results of the model while training.
- **Incorrect Hashtags** - People often end up using the wrong hashtag for their tweets, most often without realizing, which could cause random data not entirely associated with LA public transit to show up in the scraped data. This would be hard to filter for unless you make an abnormality classifier and remove those data points. Additionally, people who do tweet about LA transit may not actually use hashtags itself. Therefore, it becomes important to additionally monitor all posts the LA public transit twitter user is tagged in.
- **Manual data labeling** - Data is unstructured and is just in text form. It involves a lot of data cleaning and manual labeling of data into positive, neutral or negative sentiment. This can induce human bias of the person training the model. However, this could potentially be solved using unsupervised clustering or something of that sort - but the results will probably feel more random and less predictable.

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?

- **Reasons it is a bad idea:**
    - **Correlation does not equate to causation** - Just because the model trained on the UK dataset performs well on the USA dataset, first and foremostly, does not equate to the fact that a model trained on the USA dataset will do well on the UK dataset. By using data from two countries, we can't justify that this model will generalize well across all countries.
    - **Hologic may not be used in Brazil** - This was done using a specific device called Hologic (95%), General Electric (4%) and Siemens (1%). It makes sense to use the pre-trained model a majority of Brazilian hospitals also use this device, otherwise, the model can be skewed because of image quality (pixel density, image lighting, image sharpness, etc make a large difference in CNNs) from the Hologic device. In fact, Brazil's health care system is currently struggling to have as many good health care resources for women as those in the UK and US.
    - **Shouldn't be device-specific** - the main aim should be to generalize to mammogram images taken from other devices too. Therefore, the friend may be better off using aspects from this paper in training his/her own model from images in Brazil for higher specificity (low variance, high bias) in Brazil itself since he doesn't need a model that generalizes across countries.
    - **Data Sample:** We don't know what the model was trained on: are the woman the same age? different race? ethnicity? The model will only be as good as the distribution of the data. The generalizability of the model in the UK and US suggests the sample from the population is pretty good, but we still don't know for sure, and this could be vastly different from the demographic split in Brazil.
- **Reasons it is a good idea**:
    - Clearly, it has learned some non-trivial features that allow it to predict well on a dataset from a different country. This could suggest that:
        - Breast cancer has minimal variation across countries
        - The model has generalized very well.

      Therefore, this could be a good starting point for the Brazil specific model.
    - Especially good if the amount of Brazil specific training data is limited, because we can use transfer learning and fine-tuning to come up with a great model.

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)?

(As a side note, I think this is a great problem for a Naive Bayes Classifier)

- Student's year in school
- Professor teaching the course
- Current time in the quarter (either by day or more ideally, by week number)
- When the class is during the day (time)
- Which quarter is currently ongoing (maybe students stop attending lecture more in Spring quarter)
- Number of students 'enrolled' during each week historically in that quarter over the last few years
- Class Size - Number of students typically enrolled.
- Whether the course is an upper-division or lower-division course.
- Is attendance mandatory (i.e., is it a part of your grade?)
- Which building the class is in? (The further away it is, the higher the odds that a student will stop attending any particular lecture)
- Student's history of dropping classes/not attending a lecture.
- Number of classes (and time and date of those classes).
- Hours of sleep
- Amount of homework per class

(b) How would you formulate your labels?

- Since we are trying to predict a probability, this can be done either as a regression problem or as a classification problem.
- **Classification**: To do this as a classification problem, we can divide the *probability* into 3 buckets (thinking of it as X as a random variable with outcome space = {0,1,2}):
        - < 30% (will attend lecture)
        - 30%-70% (unsure/neutral)
        - > 70% (will stop attending lecture)
    - This ensures reasonable confidence intervals for each bucket and gives the model a chance to be more confident when making a prediction that a student will indeed stop attending a lecture.
- **Regression**: Try to get as close to the probability computed from previous years (also as a number), and then classify them into buckets later if you want. This is much harder because you need probability values computed for each student + class.

(c) How could you source/obtain/gather the above data?

- UCLA Registrar - Scrape registrar for student schedules class timings, locations, professor, class size/enrollment size.

- CCLE - Scrape CCLE for week number, relative amount of homework, classes.
- Hours of sleep - self-reported (but cannot be trusted)
- Can have professors report whether attendance is mandatory, OR, get it from CCLE using the syllabus and using OCR to extract the grade breakup (search for participation).

5. True or False

Provide brief explanations for your answers.

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

False: Most data science investigations actually start with formulating a question that is to be investigated, followed by determining how to gather/collect data to answer the question. Data comes in many forms - unstructured and structured (an existing dataset) and usually tends to be unstructured. A majority of the time, it must be gathered from a data stream/data lake, then run through an ETL pipeline, and finally stored in a database so that we can use it from there. This is not necessary, however.

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

False: Python is a tool to perform data science tasks (it only serves as a means to do it). Although the industry seems to prefer Python, many companies use R and Go. Managers in different companies who tend to call themselves ‘data scientists’ use Excel to make Pivot Tables for data analysis. The tools are decided based on the task and the end goal (deployment in a data pipeline? offline predictions?) etc, which the means just make the task easier (for example, some tasks such as plotting tend to be trickier in Python than in R).

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

False: About 80% of the time is spent on cleaning data, which involves gathering data, plotting data, analyzing correlations, understanding feature importance, feature engineering, etc. Developing new models and fine-tuning takes only 20% of the time.

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

True: This is true. While historical data is great in helping to analyze trends and forecasting future predictions, this still usually depends on the quality of the data. For example, if a human manually labeled a dataset in 2014 and 2015, and used it to make predictions in 2016, then time-series predictions for every subsequent year technically has the inherent human bias that has propagated through the years.

(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.

False: Once you essential ‘bin’ data, you’re actually converting numerical data to a categorical feature. Therefore, the income data in no longer quantitative, because the feature is actually being used a ‘label’ - we are no longer predicting a numerical value, but rather a label that is represented using numeric digits for different ‘bins’/classes.

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)?

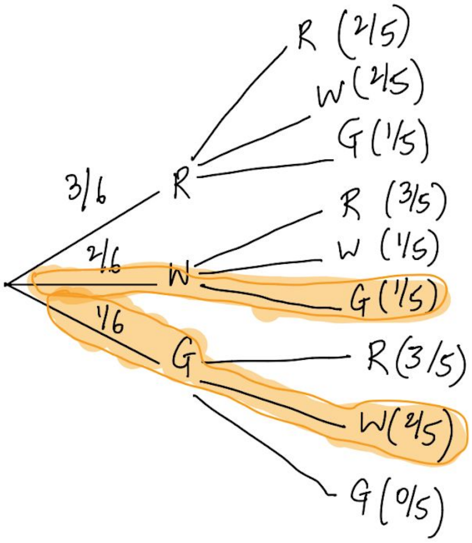
Since X=0, it means that we want to find the probability that no red marble is picked.

- When the first marble is picked, there are 3 marbles in the 6 that are not red.
- When the second marble is picked, there are 2 marbles remaining that are not red.

$$P(X = 0) = \frac{3}{6} * \frac{2}{5} = 0.2$$

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

This involves a joint probability distribution, where we can draw a tree diagram.



X → Number of red marbles drawn

Y → Number of green marbles drawn

$$\mathbb{P}(X=0, Y=1) = P(W, G) + \mathbb{P}(G, W)$$

$$= \frac{2}{6} \times \frac{1}{5} + \frac{1}{6} \times \frac{2}{5} = \frac{4}{30} = \boxed{\frac{2}{15}}$$

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.

- **Delete Rows**:
    - Advantages:
        - Removes outliers or things that can potentially skew the model, making the model more accurate.
        - Makes no assumptions about the data itself.
        - Works better for larger datasets.
    - Disadvantages:
        - For smaller datasets, this could significantly reduce the size which could decrease accuracy.
- **Replace with Median/Mean**:
    - Advantages:
        - For smaller datasets, you don't lose any data by dropping it.
        - Mean/Median are statistically significant and make a fair assumption about the possible value.
    - Disadvantages:
        - It creates outliers by assuming values for features that it doesn't know much about.
        - Adds a lot of variance to the model.
- **Predicting missing values**:
    - Advantages:
        - It makes no real assumptions about the distribution of data.
        - Predicts roughly accurate values, rather than standardizing all of them to the mean/median.
    - Disadvantages:
        - May work poorly if the feature has a very high variance.
        - A lot more work and could potentially difficult to train.
- **Assigning a unique category (for categorical variables):**
    - Advantages:
        - Easy, and makes no real assumption about the data itself.
        - Low variance - because the categorical variable is one hot-encoded, so adding one more column to the feature vector (especially for large categorical variables) will not affect the predictions much.
        - You don't lose any data.
    - Disadvantages:

- Lesser variance, so the model may overfit.
- If there’s a large distribution of null values (say greater than 30%), then this becomes a problem because it’ll make the feature important.