CS 188 HW 1

1. **Instructions:**
    a. **Study Partners:**
        i. Arnav Garg (UID: 304-911-796)

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 population for the question is all LADT transit riders, so one issue is that the sample would not be data neutral. While the complex queries allowed by Twitter's API remove the overhead of a custom web scraper, data about how transit riders who don't use Twitter can be obtained only by other methods such as other social networks or by talking to them in person which is expensive. Another issue is the data could be biased because people on more extreme ends of the positive and negative spectrum are more inclined to post on Twitter about their experience.

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

Without more information, we cannot definitely say if this idea is good or bad. One factor that could make it a bad idea is that the UK and the US have quite similar populations, so the system might not generalize as well to populations from other parts of the world like Brazil. Another negative factor is that the system may not perform as well if my friend's data from Brazil is not collected on Hologic devices, which we don't know. However, a factor that could make it a good idea is that finding a way to leverage the system's ability to generalize across more populations from different countries provides hope for overcoming one of the biggest hurdles facing AI adoption in healthcare: obtaining enough data to be representative of the entire patient population.

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)?**

I'd try to gather the following features as input to my model: student's GPA, student's number of enrolled units, student's year in school, student's age, student's major, course, department of course, professor teaching course, start time of lecture, professor's number of years teaching, professor's rating by students, input date, and student's number of midterms during the week of the input date (calculated value based on input date and values for number of midterms in each week 1-10).

b. **How would you formulate your labels?**

I've used the interpretation that the problem we're trying to solve is predicting the probability that a student stopped attending a class after a certain date (and doesn't attend that class forever). This problem can be formulated as a binary classification problem. Given a feature vector, I'd formulate my labels as "Yes" or "No" the student stops attending after a given date and doesn't attend again. By making the assumption that the features are conditionally independent, we can use a Naïve Bayes Classifier to directly calculate the conditional probability of "Yes" or "No" for test data and then choose the label with the greater probability value. Pros of this method include being easy to implement and obtaining good results in most cases, while cons include loss of accuracy due to dependencies practically existing among the features.

c. **How could you source/obtain/gather the above data?**

Let the population be students in Management 108 at UCLA and let's assume attendance in lecture is mandatory and recorded by the professor in CCLE using iClickers. The following features can be obtained using the following methods/sources:

Features

- student's GPA, student's number of enrolled units, student's year in school, student's age, student's major, course, department of course, professor teaching course, start time of lecture
    - Obtained from UCLA Registrar
- professor's number of years teaching
    - Collected from administrative staff of professor's department at UCLA
- professor's rating by students
    - Collected from Bruinwalk.com
- input date

    - Input by user of model
- number of midterms in each week 1-10
    - Collected by surveying students

Label

- “Yes”/”No”
    - The ground truth label can be calculated using the value of the input date and attendance data recorded by the professor in CCLE using iClickers

5. **True or False**

Provide brief explanations for your answers.

a. **All data science investigations start with an existing dataset.**

False. While there may be prior knowledge from literature or human co-workers, a dataset for a given question or problem may not exist. In that case, a data scientist must decide how to collect or obtain the data.

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

False. Like programmers, different data scientists use a variety of programming languages and tools. For example, they may use R or SQL instead of Python and RStudio instead of Jupyter Notebooks.

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

False. As mentioned in lecture, a study suggests many data scientists spend the majority of their time on other tasks such as basic exploratory data analysis, conducting data analysis to answer research questions, communicating findings to business decision-makers, and data cleaning.

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

True. One of the major moral challenges facing the adoption of AI is that making predictions based on historical data can reinforce historical trends and 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.**

False. Data stored as integers doesn't necessarily mean they are quantitative, as can be seen when categorical data is encoded as integers before feeding it to a model. In this case, the integer values represent qualitative categories in a specific order.

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.

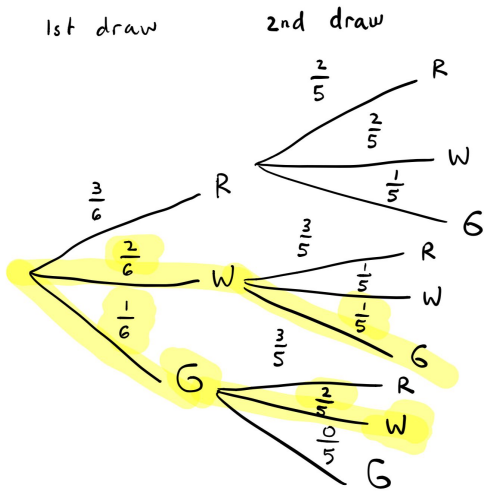
a. **What is P(X = 0)?**

In this case the events of drawing the 1st marble and the 2nd marble are independent, so we can use the multiplication rule.

$$P(X=0) = P(1st\ marble\ not\ red) \times P(2nd\ marble\ not\ red)$$
$$P(X=0) = \tfrac{3}{6} \times \tfrac{2}{5}$$
$$P(X=0) = \tfrac{6}{30}$$
$$P(X=0) = \tfrac{1}{5}$$

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

This is a bivariate distribution with two variables of discrete type. In this case, the events are dependent so we can use a tree diagram.



$$P(X=0,\ Y=1) = P(G\ 2nd\ draw \mid W\ 1st\ draw) + P(W\ 2nd\ draw \mid G\ 1st\ draw)$$
$$P(X=0,\ Y=1) = (\tfrac{1}{5})(\tfrac{2}{6}) + (\tfrac{2}{5})(\tfrac{1}{6})$$
$$P(X=0,\ Y=1) = (\tfrac{2}{30}) + (\tfrac{2}{30})$$
$$P(X=0,\ Y=1) = \tfrac{4}{30} = \tfrac{2}{15}$$

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.**

- Deletion of rows (Listwise Deletion) (while not technically imputation, this is one way to address null values)
    - Advantages: simple computation and fast
    - Disadvantages: produce biased parameters and estimates, loss of data
    - Effective on datasets: if the missing data is limited to a small number of observations
    - Inadvisable on datasets: most because assumptions of MCAR (Missing Completely at Random) do not typically hold in practice
- Substitute null values with median
    - Advantages: simple computation and fast
    - Disadvantages: doesn't factor the correlations between features
    - Effective on datasets: datasets with outliers
    - Inadvisable on datasets: datasets with a normal distribution
- Substitute null values with mean
    - Advantages: simple computation and fast
    - Disadvantages: reduces variance in the dataset
    - Effective on datasets: datasets with a normal distribution
    - Inadvisable on datasets: datasets with outliers
- Substitute with values predicted using KNN
    - Advantages: can be much more accurate than the imputing with mean, median
    - Disadvantages: computationally expensive, stores whole training dataset in memory
    - Effective on datasets: most datasets (except those with outliers) because KNN makes few assumptions about the distribution of the data
    - Inadvisable on datasets: KNN is sensitive to datasets with outliers