CS 188 HW 1

Due on Friday, January 24 at 11:59PM

1 Instructions:

You may form small groups (e.g. of up to four people) to work on this assignment, but you must write up all solutions by yourself. List your study partners for the homework on the first page, or "none" if you had no partners.

Keep all responses brief, a few sentences at most. Show all work for full credit.

Start each problem on a new page, and be sure to clearly label where each problem and subproblem begins. All problems must be submitted in order (all of P1 before P2, etc.).

No late homeworks will be accepted. This is not out of a desire to be harsh, but rather out of fairness to all students in this large course.

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?

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?

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?

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.

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 $P(X = 0, Y = 1)$?

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.