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?

Solution: Goal is for student to discuss bias, as Twitter users tweeting about transit won't be representative of LA as a whole, or even other Twitter users. Users may be more likely to tweet about transit when they're upset than when they're satisfied, skewing the results. Additionally, the method used for collecting tweets doesn't guarantee that a tweet will be solely about transit.

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?

Solution: Goal is for students to mention some of the following:

(a) Different mammogram machines may be used in Brazil, resulting in images that the system has difficulty with

(b) Brazil has a different distribution of ethnic groups than the UK or USA, and the original model may not generalize to them

(c) Clinical procedures may be different in Brazil, resulting in patients getting screened at different times than in the UK or USA

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)?

Solution: Goal is to get interesting features like time of lecture, lecture location, student GPA, distance from home to lecture.

(b) How would you formulate your labels?

Solution: Labels anything that makes sense, like "student has an uninterrupted streak of at least one absence at the end of the quarter" or "student overall attendance drops below 80

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

Solution: Sourcing could be anything including direct from the campus/registrar to surveying students and data gathering via smartphone app

5 True or False

Provide brief explanations for your answers.

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

Solution: False. In many settings a data scientist is tasked with a question or problem and must decide how to collect or obtain data to answer the question or solve the problem.

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

Solution: False. Data scientists use many programming languages and tools. In class we discussed surveys that suggested that SQL and then R are the most commonly used languages.

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

Solution: False. Sadly, data suggests that most data scientists spend the majority of their time collecting and cleaning data and doing exploratory data analysis.

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

Solution: True. A key ethical challenge of data driven decision making is that we tend to reinforce trends in our data.

(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.

Solution: False. Although stored as integers, these values represent ordered categories so they are qualitative.

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 $P(X = 0)$? **Solution:** The event that $X = 0$ is the same as the event that no red marbles are drawn. We can use a counting as follows: there are $\binom{6}{2} = \frac{6!}{2!4!} = 15$ ways to draw a subset of two marbles. Of those, the number of subsets with no red marbles is $\binom{3}{2} = \frac{3!}{2!1!} = 3$, so the proportion of draws without red marbles is $\frac{3}{15} = 1/5$.

Alternatively, we can use conditional probability. The chance no red marbles are drawn is the same as the event that the first draw isn't red and the second draw isn't red.

$$p = P(\texttt{1st draw not red and 2nd draw not red})$$

This is equivalent to

$$P(\texttt{1st draw not red}) * P(\texttt{2nd draw not red given 1st is not red}) = \frac{3}{6} * \frac{2}{5} = \frac{1}{5}$$

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

Solution: For X to be 0 and Y to be 1, means that we drew 1 green and 1 white ball. We can draw green first and then white, which has chance $\frac{1}{6} * \frac{2}{5}$ or white first and green second, which has chance $\frac{2}{6} * \frac{1}{5}$. The combined probability is $\frac{4}{30} = \frac{2}{15}$. This approach is using conditional probability, i.e. $P(X = 1, Y = 1) = P(X = 0) * P(Y = 1|X = 0)$. We found $P(X = 0)$ above to be $\frac{1}{5}$. For the conditional probability, if we know $X = 0$ then we know that we are drawing from the 2 white and 1 green marbles. There are 3 possible ways to draw 2 marbles from these 3 and 2 of the possibilities give us 1 green and 1 white. Putting these together we have $\frac{1}{5} * \frac{2}{3} = \frac{2}{15}$.

Alternatively, we can count the number of subsets that have 1 green and 1 white marble, which is 2, and divide by the number of ways to choose 2 marbles out of 6 (which we calculated above to be 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.

Solution: Some possible responses include imputing with the median, mean, or mode, as well as with a fixed value (0, -1). Other responses also accepted.