CSM148 Homework 1

Due date: Wednesday, April 14 before Midnight PST
Instructions: All work must be completed individually. If you consulted with any classmates for the homework, please note them on the first page.

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.

1 Data Collection and Analysis

Your friend working at the Los Angeles Department of Health has been given the task of determining how patients who visit public clinics in the city feel about Los Angeles's public health care systems. Your friend wants to accomplish this by scraping Twitter for tweets containing keywords and hashtags related to Los Angeles public health and running them through a model that does sentiment analysis (the algorithm will say whether a tweet contains positive, neutral, or negative sentiment). What are some of the issues, if any, with what your friend proposes?

2 Experimental Design

You would like to see if you can predict the probability that a given student will do well in the course.

(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?

3 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. 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 likely be the most effective approach, and what sorts of data would it be inadvisable for.

4 Utility of One-Hot-Encoding

One-Hot-Encoding is a process of converting a single categorical variable (with multiple discrete options) into a number of binary features, one for each possible value. This is often an incredibly important data pre-processing step, however there are times when it is inappropriate to employ one-hot-encoding. Please evaluate the following features and determine if you would one-hot-encode them. Justify your response:

(a) Heartrate (heartbeats per minute)

(b) A category for health where the values 1-5 represent health levels ranging from terminally ill, to completely healthy

(c) A list of fashion brands

(d) State (part of address)

(e) An Interaction term of time of day and direction-facing when trying to predict sun-exposure to a plot of land

5 True or False, Simple Explanations

Provide brief explanations for your answers.

(a) (T or F) A small p-value (< 0.05) provides evidence in support of the null hypothesis.

(b) (T or F) It is always advisable to augment your dataset with additional features.

(c) (T or F) All data science investigations start with an existing dataset.

(d) (T or F) One-hot encoding is good to use to encode categorical features with a high number of distinct values.

(e) (T or F) The use of historical data to make decisions about the future can reinforce historical biases.

6 Basic Probability

A jar contains 3 white, 2 red, and 1 blue 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 = 1)? (b) Let Y be the number of white marbles drawn. What is P (X = 0, Y = 1)?