Homework 1

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. **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.

Question 1 (True or False)
Provide brief explanations for your answers.
(a) Correlation implies causation.
False, see lecture slides.
(b) Data scientists do most of their work in Python and are unlikely to use other tools.
False, there's other tools like R, SQL, Spark, etc.
(c) Most data scientists spend the majority of their time developing new models.
False, you should usually be using preexisting models.
(d) The use of historical data to make decisions about the future can reinforce historical biases.
True, bias is implicitly encoded in data, so if you use it to make predictions, the predictions will be biased too.
(e) All data science investigations start with an existing dataset.
False, some can start with a question and you have to find the dataset.
(f) Linear regression is an example of an unsupervised model.
False, linear regression is **supervised** learning.
(g) Linear regression can be used to confidently make predictions inside the range of data on which it was trained on (assume that linear regression works well as a model for this data).
True, this question is essentially asking about if interpolation is safe, which it is. Extrapolation should usually be avoided, and if it must be done, it should be done cautiously.

Note: The question states "assume linear regression **works well** as a model for this data," so we can deduce from this that a linear model does not overfit the data.

Question 2

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. Is this a good idea? Why or why not?

Not a good idea, there is the danger of extrapolation. To use it in Brazil, the model should have been trained on Brazilian data.

Question 3

Your friend working at UCLA housing has been given the task of determining how students feel about the new dorms. Your friend wants to accomplish this by scraping Twitter and news article comment sections for tweets containing keywords and hashtags related to the new dorm and then running them through a model that does sentiment analysis. This model contains an algorithm that says whether the text exhibits positive, neutral, or negative sentiment. What are some of the issues, if any, with what your friend proposes?

Potential issues:

- Not everyone uses Twitter (population not representative)
- People that tweet tend to feel strongly (either really like it or dislike it)
- There might be tweet about the dorm that are missed in the scraping process
- Sarcasm or other forms of confounding language might be used
- It can be hard to capture a spectrum of feelings (how to distinguish between really dislike and mildly dislike?)

Question 4

You would like to see if you can predict the probability that a given student will take a particular class in the upcoming quarter.

(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)?

Possible answers: student's major, year in school, what other courses they're planning to take, course time (morning/afternoon/night), professor rating, number of students in the class, class difficulty, ask directly if the student is planning to take it, etc.

(b) How would you formulate your labels? (Think about ordinal/categorical/numerical encoding, min/max/possible values, etc.)

For each feature mentioned in (a), make sure the encoding picked makes sense. Make sure the min/max/possible values given make sense too. For example, major would be OHE, year in school would be ordinal encoding, lecture time could be numerical or binned. GPA would be a min of 0 and max of 4, and stay numerical, etc...

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

Possible answers: survey the students, get data from the registrar, get data from departments.

Problem 5

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

Possible strategies: delete rows, delete columns, fill with default value, fill with median or mean, fill with value based on other values close to it (bootstrap). A good article about this can be found here.

(b) Outliers can also be problematic points in a dataset. Which of the strategies you mentioned in part (a) would also work for outliers, and why?

Any of the strategies mentioned above would work well. If replacing the value with mean/median/default/bootstrapped, even better if student mentions keeping the original value in a different column.

Problem 6

A jar contains 3 red, 2 yellow, 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 and Y represent the number of blue marbles drawn.

(a) What is P(X = 1)?

P(X = 1) = P(first red, second anything else) + P(first not red, second red)
= 3/6 * 3/5 + 3/6 * 3/5 = 2 * 9/30 = 9/15 = 3/5

(b) What is P(X = 0, Y = 1)?

P(X = 0, Y = 1) = P(no reds and 1 blue)
= P(first blue, second not red) + P(second not red nor blue, second blue)
= 1/6 * 2/5 + 2/6 * 1/5 = 2/30 + 2/30 = 2/15

(c) What is P(X = 1 | Y = 1)?

P(X = 1 | Y = 1) = P(1 red, given 1 blue was already drawn from the bag) = 3/5

Problem 7

Evaluate the following features and determine if you would one hot encode them. Justify your response:

(a) Walking pace, measured in miles per hour.

No, better left as numerical data.

(b) Class name (freshman, sophomore, junior, senior).

Depends, we'll take yes if the student mentions categories, or no if ordinal encoding is mentioned instead as a better option. It depends on what this attribute is being used for.

(c) City in the US (part of an address).

No, too many unique values.

(d) If someone has diabetes or not.

Yes, technically already one hot encoded, just need to convert false/true or no/yes into numerical 0/1.

Note: this question seemed to be unclear, so giving everyone full credit for it.

(e) Laptop brand.

Yes. However, one can also argue that there are many laptop brands, so no is also permissible.

(f) Interaction term of how much sleep one got the previous night (numerical value, in hours) and test difficulty (3 possible values: easy, medium, hard) when trying to predict test score.

Yes, we would bin amount of sleep (such as 0-3, 3-6, 6-9, 9+ hrs) and then there aren't too many values to one hot encode (3 * 4 = 12). Note that we're asking for the interaction term, not the 2 in isolation.