DSC 40A - Homework 8

due Thursday, June 6th at 11:59PM (no slip days!)

Write your solutions to the following problems by either typing them up or handwriting them on another piece of paper. Homeworks are due to Gradescope by 11:59PM on the due date.

Homework will be evaluated not only on the correctness of your answers, but on your ability to present your ideas clearly and logically. You should **always explain and justify** your conclusions, using sound reasoning. Your goal should be to convince the reader of your assertions. If a question does not require explanation, it will be explicitly stated.

Homeworks should be written up and turned in by each student individually. You may talk to other students in the class about the problems and discuss solution strategies, but you should not share any written communication and you should not check answers with classmates. You can tell someone how to do a homework problem, but you cannot show them how to do it. We encourage you type your solutions in LaTeX, using the Overleaf template on the course website.

For each problem you submit, you should **cite your sources** by including a list of names of other students with whom you discussed the problem. Instructors do not need to be cited.

This homework will be graded out of 26 points. The point value of each problem or sub-problem is indicated by the number of avocados shown.

Notes:

- This homework is much shorter than earlier homeworks, though remember that all homeworks are worth the same amount in your overall grade.
- There is no Homework 8-specific survey; instead, we'll be releasing a more general End-of-Quarter Survey shortly. Stay tuned for more details.
- For full credit, make sure to **assign pages to questions** when you upload your submission to Gradescope.

Problem 1. Double Deck of Cards

A standard deck of cards contains 52 cards. There are 13 cards in each of 4 suits (hearts ♡, spades ♠, diamonds ♢, and clubs ♣.) Within a suit, the 13 cards each have a different rank. In ascending order, these ranks are 2, 3, 4, 5, 6, 7, 8, 9, 10, Jack, Queen, King, Ace.

You are playing a four-player card game using two regular decks of cards. **Each player will be dealt 26 cards** as follows:

- The first deck of cards will be randomly shuffled and dealt out, 13 cards to each of the four players.
- Then, the second deck of cards will be randomly shuffled and dealt out, 13 cards to each of the four players.

a) Let A be the event that you are dealt two Jacks of Diamonds and let B be the event that some other player is dealt two Jacks of Diamonds. Are events A and B independent?

b) Let A be the event that you are dealt two Jacks of Diamonds and let C be the event that some other player is dealt two 5s of Clubs. Are events A and C independent?

c) Let A be the event that you are dealt two Jacks of Diamonds and let C be the event that some other player is dealt two 5s of Clubs. Suppose you know for sure that you do not have any 5s of Clubs. Given this information, are events A and C independent?

d) Let:

- A be the event that you are dealt two Jacks of Diamonds,
- C be the event that some other player is dealt two 5s of Clubs, and
- D be the event that you do not have any 5s of Clubs.

State the result of parts (b) and (c) in words, in terms of independence and conditional independence.

Problem 2. Tips

Billy works as a waiter at Dirty Birds, the on-campus restaurant, on Thursdays, Fridays, Saturdays, and Sundays. He wrote down a few pieces of information for a random sample of customers that he served. That information is shown below.

	sex	day	time	tip_cat
0	Female	Sun	Dinner	Small
1	Female	Sun	Dinner	Small
2	Female	Sun	Dinner	Small
3	Female	Thur	Lunch	Small
4	Male	Sun	Dinner	Small
5	Male	Sat	Dinner	Small
6	Male	Fri	Dinner	Small
7	Female	Fri	Dinner	Medium
8	Male	Sun	Dinner	Medium
9	Male	Sun	Dinner	Medium
10	Male	Sun	Dinner	Medium
11	Male	Thur	Lunch	Medium
12	Female	Sun	Dinner	Large
13	Female	Thur	Lunch	Large
14	Female	Thur	Lunch	Large

Each row of Billy's dataset contains information about a single transaction. For each transaction, Billy kept track of:

- `"sex"`: The sex of the customer paying (in this case, either `"Male"` or `"Female"`).
- `"day"`: The day of the week (either `"Thur"`, `"Fri"`, `"Sat"`, or `"Sun"` — these are the only days that Billy works).
- `"time"`: Either `"Lunch"` or `"Dinner"`.
- `"tip_cat"`: The size, or category, of the tip that the customer leaves (either `"Small"`, `"Medium"`, or `"Large"`).

Billy wants to predict whether a customer will leave a small, medium, or large-sized tip, given their sex, the day of the week, and the time of day. He enlists you to help him, and you decide to use the Naive Bayes classifier that you learned about in class.

a) Using the Naive Bayes classifier and **no** smoothing, predict whether a male customer who comes to Dirty Birds on a Thursday for dinner will leave a small, medium, or large-sized tip.

You must show **all** of your steps in order to get full credit. Do not convert any probabilities to decimals; write your final results as fractions. You may use a calculator to simplify your fractions.

b) Using the Naive Bayes classifier **with smoothing**, predict whether a male customer who comes to Dirty Birds on a Thursday for dinner will leave a small, medium, or large-sized tip.

You must show **all** of your steps in order to get full credit. Do not convert any probabilities to decimals; write your final results as fractions. You may use a calculator to simplify your fractions.

c) **Moving forward, let's assume that we're using our results from part (b), i.e. that we're using smoothing.**

When deciding what to predict for a male customer coming for dinner on a Thursday, we only computed the numerators of $\mathbb{P}(\text{Small} \mid \text{Male, Thur, Dinner})$, $\mathbb{P}(\text{Medium} \mid \text{Male, Thur, Dinner})$, and $\mathbb{P}(\text{Large} \mid \text{Male, Thur, Dinner})$. This was because we weren't interested in the actual values of these three probabilities; rather, we're interested in which probability is the largest. Since all three have the same denominator, $\mathbb{P}(\text{Male, Thur, Dinner})$, the value of the denominator was irrelevant in making our prediction.

With that said, we do actually have enough information to compute the value of $\mathbb{P}(\text{Male, Thur, Dinner})$, which would help us compute the actual values of $\mathbb{P}(\text{Small} \mid \text{Male, Thur, Dinner})$,

$\mathbb{P}(\text{Medium} \mid \text{Male, Thur, Dinner})$, and $\mathbb{P}(\text{Large} \mid \text{Male, Thur, Dinner})$, not just their numerators. You'll see the benefit of doing this in the final part of this problem.

Using **the law of total probability** and your work from part (b), determine the values of the following three probabilities:

$$\mathbb{P}(\text{Small} \mid \text{Male, Thur, Dinner})$$
$$\mathbb{P}(\text{Medium} \mid \text{Male, Thur, Dinner})$$
$$\mathbb{P}(\text{Large} \mid \text{Male, Thur, Dinner})$$

You may leave your answers unsimplified.

Hint: Since a tip must be either small, medium, or large, the sum of your three answers must be 1. This part requires much less computation than the previous two!

d) Follow the instructions in the supplemental Jupyter Notebook, linked **here**. You **do not** need to submit this notebook anywhere; instead, include a screenshot of your code at the end of Problem 2(d) (before it says "For fun!") in your PDF submission for Homework 8.