DSC 40A - Homework 7

Due: Friday, March 4, 2022 at 11:59pm

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. You can use a slip day to extend the deadline by 24 hours.

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.

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 50 points. The point value of each problem or sub-problem is indicated by the number of avocados shown.

Note: For Problem 5 part (c), code your answer in the supplementary Jupyter notebook (linked). You'll need to turn in your completed Python file to Gradescope separately from the rest of this homework, in a file called `hw7code.py`. This part will be autograded, so no explanation is needed.

Problem 1. Probability Theory

Let E and F be two events in the sample space S. Assume $0 < P(F) < 1$.

a) In one sentence, or using a Venn diagram, explain why $P(E \cap F) + P(E \cap \overline{F}) = P(E)$.

b) If $P(E|F) = P(E|\overline{F})$, show that E and F are independent.

Hint: To show E and F are independent, you must show that $P(E \cap F) = P(E)P(F)$. You may need to use your result from part (a).

c) If E and F are independent, must it be true that $P(E|F) = P(E|\overline{F})$? If yes, explain why. If no, give a counterexample.

Problem 2. Unrelated Properties

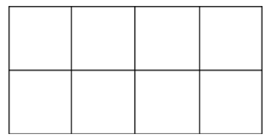
This problem will illustrate that independence and conditional independence are not related, in the sense that neither property implies the other.

Consider the sample space $S = \{1, 2, 3, 4, 5, 6, 7, 8\}$ with associated probability distribution $P(s) = \frac{1}{8}$ for each s in S. In other words, there are eight equally likely outcomes.

For each part below, define events A, B, C in this sample space that satisfy the given requirements. Make sure to choose events that are neither impossible nor certain, that is, $0 < P(A), P(B), P(C) < 1$. Define an event as a subset of S by listing which elements of the sample space it contains. For example, you might have $A = \{2, 4, 6\}$.

For each part, demonstrate that the given requirements are satisfied by computing the appropriate probabilities.

Hint: It might help to think of the sample space visually like the grid below.



a) A and B are not independent.
A and B are conditionally independent given C.

b) A and B are not independent.
A and B are not conditionally independent given C.

c) A and B are independent.
A and B are conditionally independent given C.

d) A and B are independent.
A and B are not conditionally independent given C.

Problem 3. Double Deck of Cards

A standard deck of cards contains 52 cards. There are 13 cards in each of 4 suits (hearts $\heartsuit$, spades $\spadesuit$, diamonds $\diamondsuit$, and clubs $\clubsuit$.) 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 Event A = 'you are dealt two Kings of Hearts' and Event B = 'some other player has a pair of Kings of Hearts'. Are these two events independent of each other?

b) Let Event A = 'you are dealt two Kings of Hearts' and Event C = 'some other player has a pair of Aces of Hearts'. Are these two events independent of each other?

c) Now suppose that you know you do not have any Aces of Hearts. Let Event A = 'you are dealt two Kings of Hearts' and Event C = 'some other player has a pair of Aces of Hearts'. Now that you know you don't have any Aces of Hearts, are these two events independent of each other?

d) Let Event A = 'you are dealt two Kings of Hearts' , Event C = 'some other player has a pair of Aces of Hearts', and Event D = 'you do not have any Aces of Hearts'. State the result of parts (b) and (c) in words, in terms of independence and conditional independence.

Problem 4. Picky Eating

Baby Luca is a picky eater. When he's offered a food, he'll eat it only sometimes, depending on the type of food. Baby Luca eats

- bananas 95% of the time,
- avocados 60% of the time,
- meat 45% of the time, and

- zucchini 30% of the time.

a) Baby Luca's grandpa gave him one of the above four foods and Baby Luca ate it all up. You have no idea which of the four foods it was, so assume it was equally likely to be any of them.

Given that Baby Luca ate the food, what's the probability that it was a banana? An avocado? Meat? Zucchini? Show your work.

b) As before, Baby Luca's grandpa gave him one of the above four foods and Baby Luca ate it all up. This time, suppose you know that Baby Luca's grandpa is particularly fond of green foods, and so he feeds Baby Luca

- bananas 15% of the time,
- avocados 40% of the time,
- meat 10% of the time, and
- zucchini 35% of the time.

Given that Baby Luca ate the food, what's the probability that it was a banana? An avocado? Meat? Zucchini? Show your work.

c) Compare your answers to part (a) and part (b) above. Identify which of the four probabilities you computed increased and which decreased, and explain why this makes sense intuitively.

Problem 5. Comic Characters

In this problem, we'll work with a dataset of characters from comic books. At first, we look at a limited set of 25 of the more popular comic characters. You can find this dataset on the last page of this assignment as well as on Datahub in the same directory as the supplementary Jupyter notebook (linked).

For each character, we have the following information.

Variable	Definition
`ID`	The identity status of the character ("Public Identity" or "Secret Identity")
`SEX`	Sex of the character ("Male Characters" or "Female Characters")
`ALIVE`	Whether the character is alive ("Living Characters" or "Deceased Characters")
`COMPANY`	The comic company that created the character ("DC" or "Marvel")
`ALIGN`	The alignment of the character ("Bad Characters", "Good Characters", or "Neutral Characters")

a) Use the set of 25 popular characters to make a prediction using Naive Bayes (without smoothing) for the following character's alignment:

- "Secret Identity"
- "Male Characters"
- "Living Characters"
- "Marvel"

Your prediction should be "Bad Characters", "Good Characters", or "Neutral Characters", whichever is most likely according to Naive Bayes.

b) In the supplementary Jupyter notebook (linked), we've provided not only the dataset of 25 popular characters, but also a much larger dataset of over ten thousand comic characters. For this part, you will be implementing a Naive Bayes classifier (without smoothing) to predict the alignment of any character based on their features (ID, SEX, ALIVE, COMPANY), using the larger dataset.

Complete the function `predict_align`, which takes in a DataFrame of comic characters and the features of one particular character, and returns the predicted alignment for that character according to Naive Bayes without smoothing, using the input DataFrame of characters.

Your function should return a string, either "Bad Characters", "Good Characters", or "Neutral Characters".

c) Use your `predict_align` function to predict the alignment for the same character as you did in part (a), this time using the larger dataset of comic characters. As a reminder, the features were

- "Secret Identity"
- "Male Characters"
- "Living Characters"
- "Marvel"

You should get a different prediction than you got in part (a), when you used only the 25 popular characters. Why do you get different results? What specific difference between the two datasets explains why your predictions are different?

Hint: If you're stuck, try printing out each term in the products that you calculate in `predict_align`. Compare these terms when you input the DataFrame of 25 popular characters and when you input the larger DataFrame.

Hint: Another way to get started on this is to try other combinations of features as input to `predict_align`. When using the DataFrame of 25 popular characters, try to find a combination of features that leads to a diffrent result than you got in part (a).

name	ID	SEX	ALIVE	COMPANY	ALIGN
Wilson Fisk (Earth-616)	Public Identity	Male Characters	Living Characters	Marvel	Bad Characters
Joker (New Earth)	Secret Identity	Male Characters	Living Characters	DC	Bad Characters
Kent Nelson (New Earth)	Secret Identity	Male Characters	Deceased Characters	DC	Good Characters
Timothy Dugan (Earth-616)	Public Identity	Male Characters	Deceased Characters	Marvel	Good Characters
Samuel Wilson (Earth-616)	Public Identity	Male Characters	Living Characters	Marvel	Good Characters
Hulk (Robert Bruce Banner)	Public Identity	Male Characters	Living Characters	Marvel	Good Characters
Reed Richards (Earth-616)	Public Identity	Male Characters	Living Characters	Marvel	Good Characters
Benjamin Grimm (Earth-616)	Public Identity	Male Characters	Living Characters	Marvel	Good Characters
Iron Man (Anthony \"Tony\" Stark)	Public Identity	Male Characters	Living Characters	Marvel	Good Characters
Jessica Drew (Earth-616)	Secret Identity	Female Characters	Living Characters	Marvel	Good Characters
Johnathon Blaze (Earth-616)	Secret Identity	Male Characters	Living Characters	Marvel	Good Characters
Brian Braddock (Earth-616)	Public Identity	Male Characters	Living Characters	Marvel	Good Characters
Dane Whitman (Earth-616)	Secret Identity	Male Characters	Living Characters	Marvel	Good Characters
Captain America (Steven Rogers)	Public Identity	Male Characters	Living Characters	Marvel	Good Characters
Crystalia Amaquelin (Earth-616)	Public Identity	Female Characters	Living Characters	Marvel	Good Characters
Superman (Clark Kent)	Secret Identity	Male Characters	Living Characters	DC	Good Characters
Batman (Bruce Wayne)	Secret Identity	Male Characters	Living Characters	DC	Good Characters
Franklin Rock (New Earth)	Public Identity	Male Characters	Living Characters	DC	Good Characters
Cassandra Sandsmark (New Earth)	Public Identity	Female Characters	Living Characters	DC	Good Characters
Garth (New Earth)	Public Identity	Male Characters	Deceased Characters	DC	Good Characters
James Rhodes (Earth-616)	Secret Identity	Male Characters	Living Characters	Marvel	Good Characters
Deadpool (Wade Wilson)	Secret Identity	Male Characters	Living Characters	Marvel	Neutral Characters
Odin Borson (Earth-616)	Public Identity	Male Characters	Living Characters	Marvel	Neutral Characters
Otto Octavius (Earth-616)	Secret Identity	Male Characters	Deceased Characters	Marvel	Neutral Characters
Wolverine (James \"Logan\" Howlett)	Public Identity	Male Characters	Living Characters	Marvel	Neutral Characters