DSC 40A - Homework 7

Due: Friday, Dec 6th 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.

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.

Homework 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 to 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 51 points. The point value and difficulty of each problem or sub-problem is indicated by the number of avocados shown.

Note: For full credit, make sure to assign pages to questions when you upload your submission to Gradescope. You will lose points if you don't!

Problem 1. Reflection and Feedback Form

a) 🥑🥑 Make sure to fill out this Reflection and Feedback Form, linked here, for three points on this homework! This form is primarily for your benefit; research shows that reflecting and summarizing knowledge helps you understand and remember it.

b) 🥑🥑 Fill out the Student Evaluation of Teaching (SET) for DSC40A. Feedback on the course is very helpful to us personally and the DSC program in general to understand what is working well in the course and what changes can be made to improve for the future. Your feedback helps drive the course development and impacts future students taking the course. This is especially important in a relatively new program such as ours.

Problem 2. Independence and Conditional Independence

Consider the sample space $S = \{a, b, c, d, e, f\}$ with associated probabilities given in the table below.

outcome	a	b	c	d	e	f
probability	$\frac{4}{42}$	$\frac{4}{42}$	$\frac{2}{42}$	$\frac{10}{42}$	$\frac{16}{42}$	$\frac{6}{42}$

Let $X = \{a, b\}$ and $Y = \{b, e\}$. Remember to show your work for all calculations.

a) 🥑 Are X and Y independent?

Solution: Write your solution here.

b) 🥑🥑🥑🥑🥑🥑 In this problem, you will determine whether X and Y are conditionally independent given a third event, Z.

1. Suppose $Z_1 = \{a, b, d, e, f\}$. Are X and Y conditionally independent given Z_1?
2. Suppose $Z_2 = \{a, b, e\}$. Are X and Y conditionally independent given Z_2?
3. Suppose $Z_3 = \{c, d, e, f\}$ Are X and Y conditionally independent given Z_3?

Solution: Write your solution here.

Problem 3. Independence and Complements

a) 🥑🥑🥑 Let A and B be two independent events in the sample space S. Show that $\bar{A}$ and $\bar{B}$ must be independent of one another. Make sure your proof still works in the special case that one or more of A, B, $\bar{A}$, $\bar{B}$ has probability zero. **Hint:** It helps to draw a Venn diagram.

Solution: Write your solution here.

b) 🥑🥑🥑🥑 Let E and F be two events in a sample space S, with $0 < \mathbb{P}(F) < 1$. If $\mathbb{P}(E|F) = \mathbb{P}(E|\overline{F})$, must it be true that E and F are independent? Provide a proof of independence, or give a counterexample by specifying a sample space S and two dependent events E and F that satisfy the given conditions.

Solution: Write your solution here.

c) 🥑🥑🥑🥑 Let E and F be two events in a sample space S, with $0 < \mathbb{P}(F) < 1$. If $\mathbb{P}(E|F) = \mathbb{P}(\overline{E}|F)$, must it be true that E and F are independent? Provide a proof of independence, or give a counterexample by specifying a sample space S and two dependent events E and F that satisfy the given conditions.

Solution: Write your solution here.

Problem 4. Genetic test

Let's revisit the example we studied in class. 1% of the population have a certain genetic defect.

a) 🥑🥑🥑 A test has been developed such that 90% of administered tests accurately detect the gene (true positives). What needs to be the false positive probability so that the probability of a patient having the genetic defect given a positive result (the posterior) is 90%?

Solution: Write your solution here.

For subproblems (b)-(d), consider the following. A test has been developed such that 1% of administered tests are positive when the patient doesn't have the gene (false positives).

b) 🥑🥑🥑 What needs to be the true positive probability so that the probability of a patient having the genetic defect given a positive result (the posterior) is 50%?

Solution: Write your solution here.

c) 🥑🥑🥑 Show that there is no true positive probability such that the probability of a patient having the genetic defect given a positive result (the posterior) can be 90%. (Hint: remember a probability p needs to fulfill $0 \leq p \leq 1$.)

Solution: Write your solution here.

d) 🥑🥑🥑🥑 What is the highest *posterior* probability such a test can have? What is the corresponding true positive probability? (Hint: remember a probability p needs to fulfill $0 \leq p \leq 1$.)

Solution: Write your solution here.

Problem 5. Avi's music

Avi has a very unique music taste. Whenever he listens to a particular song, he would only enjoy it sometimes, depending on the artist of that song. Avi enjoyed songs by:

- Kendrick Lamar 90% of the time,
- Taylor Swift 75% of the time,
- Drake 45% of the time, and
- J Cole 50% of the time.

a) 🥑🥑🥑🥑 Avi played a song from one of the above four artists and he really enjoyed the song. You have no idea which of the four artists he listed to, so assume it was equally likely to be any of them. You can also assume that Avi only listens to those 4 artists.

Given that Avi enjoyed the song, what's the probability that it was a song from Kendrick Lamar? Taylor Swift? Drake? J Cole? Show your work.

Solution: Write your solution here.

b) 🥑🥑🥑🥑 Avi again listens to a song from one of the above 4 artists and enjoys the song. This time, instead of assuming that he's equally likely to listen to all four artists, suppose you know that Avi listens to:

- Kendrick Lamar 20% of the time,
- Taylor Swift 35% of the time,
- Drake 15% of the time, and
- J Cole 30% of the time.

Given that Avi enjoyed the song, what's the probability that the song was by Kendrick Lamar? Taylor Swift? Drake? J Cole? Show your work.

Solution: Write your solution here.

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.

Solution: Write your solution here.

Problem 6. 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. Do this part by hand, and show your work. You can check your answer using the code you'll write in part (b).

Solution: Write your solution here.

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".

Note: On datahub, you may need to install babypandas. To install, run '!pip install babypandas' at the begining of supplementary Jupyter notebook (linked).

Solution: Write your solution here.

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 different result than you got in part (a).

Solution: Write your solution here.

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