Homework 3: Conditional Probability

For each problem, remember you must briefly explain/justify how you obtained your answer, as correct answers without an explanation will not receive full credit. Moreover, in the event of an incorrect answer, we can still try to give you partial credit based on the explanation you provide.

In general, your goal in an explanation is to write enough that a student from class who has attended lecture, but not read the problem yet, could understand your approach, verify your reasoning, and believe your answer is correct. While we do not usually need to see arithmetic, you must include enough work that in principle one could rederive your answer with only a scientific calculator.

Unless a problem states otherwise, you should leave your answer in terms of factorials, combinations, etc., for instance 26^7 or $26!/7!$ or $26 \cdot \binom{26}{7}$ are all good forms for final answers.

Instructions as to how to upload your solutions to gradescope are on the course web page.

Remember that you must tag your written problems on Gradescope.

Submission: You must upload a **pdf** of your written solutions to Gradescope under "HW 3 [Written]". (Instructions as to how to upload your solutions to Gradescope are on the course web page.) The use of latex is *highly recommended*. (Note that if you want to hand-write your solutions, you'll need to scan them. We will take off points for hand-written solutions that are difficult to read due to poor handwriting and neatness.)

Your code will be submitted under "HW 3 [Coding]" as a file called `cse312_pset3_nb.py`.

Due Date: This assignment is due at 11:59 PM Wednesday July 16th.
You will submit the written problems as a PDF to gradescope. Please put each numbered problem on its own page of the pdf (this will make selecting pages easier when you submit), and ensure that your pdfs are oriented correctly (e.g. not upside-down or sideways). The coding problem will also be submitted to gradescope.

Academic Integrity: Please read the full academic integrity policy. If you work with others (and you should!), you must still write up your solution independently and name all of your collaborators in the separate question on gradescope.

1. Defense Against the Dark Arts MCQ [8 points]

You are taking a multiple-choice exam (at Hogwarts, of course). With probability p, you know the answer to the question (and always get it correct). With probability $1-p$, you don't know the answer and guess randomly among the 5 possible options (of which exactly one is correct).

(a) Calculate the probability you get a question correct. Please define events and state which rules/laws you are using to do the calculation. [3 points]

(b) Given that you got a question correct, what is the probability that you actually knew it (i.e., that you didn't get it correct by guessing)? Please define events and state which rules/laws you are using to do the calculation. [5 points]

2. Independence Day (Gradescope) [9 points]

Complete this problem on the gradescope assignment. Please check your answers as it will be autograded so we cannot give partial credit.

3. True or False (Gradescope) [6 points]

Complete this problem on the gradescope assignment. Please check your answers as it will be autograded so we cannot give partial credit.

4. ℙNA [15 points]

Biology background: Blood Types and the Human Genome

As you may remember from basic biology, the human A/B/O blood type system is controlled by one gene for which 3 variants ("alleles") are common in the human population – unsurprisingly called A, B, and O.

As with most genes, everyone has 2 copies of this gene, one inherited from the mother and the other from the father. Everyone passes a randomly selected copy to each of their children. This happens with probability 1/2 for each copy, independently for each child. Focusing only on A and O, people with AA or AO gene pairs have type A blood; those with OO have type O blood (A is "dominant", O is "recessive").

We call the alleles that one carries their *genotype*, and the outwardly observable characteristics their *phenotype*. Thus, if a person has the genotype AA or AO, they have the phenotype A. Likewise, if they have the genotype OO, they have the phenotype O.

Notation

You **must** use the following notation in your answers: Let $G_I = \#\#$ be the event that person I has the genotype $\#\#$, and $Ph_I = \#$ be the event that person I has the phenotype $\#$. For this problem, the set of possible genotypes is $\{AA, AO, OO\}$; the set of possible phenotypes is $\{A, O\}$. Use X, Y, Z, and C to refer to Xena, Yvonne, Zachary, and their Child respectively (you might not need to refer to all of them).

Give **exact answers as simplified fractions** and use the formulas of conditional probability to justify your reasoning for each of them (any combination of the definition of conditional probability, Bayes' Theorem, Law of Total Probability). **Answers that do not explicitly use the theorems will not receive any credit**. Carefully consider which theorems to use as some theorems may lead to simpler calculations than others.

The Problem

You know that Xena's parents have type A blood, but Xena's sister Yvonne has type O.

(a) What are the genotypes of Yvonne and their two parents, from the information above?

(b) The information you have from the previous part (and that information **only**) should be used to create your sample space (i.e., the set of all possibilities we should consider). Unless otherwise noted, future information should be analyzed by conditioning on it. Be sure to read carefully for what is part of the sample space and what to condition on to be sure your notation matches ours.

With that information in mind, what is the probability that Xena carries an O gene, supposing that Xena has type A blood?
Hint: To start you off and to get a feel for the notation, you are calculating $P(G_X = AO | Ph_X = A))$

(c) Xena marries Zachary, who has type O blood. You may consider Zachary's blood type to be part of the sample space (i.e., you don't need to condition on it).

Compute the probability that Xena and Zachary's first child will have type O blood, still supposing that Xena has type-A blood (this requirement should still be conditioning).

Make sure to represent the event using the required notation, and show all your work manipulating the equation to get your final result.

(d) Suppose that Xena and Zachary's first child had type A blood and Xena has type-A blood; conditioned on both of those pieces of information, what is the probability that Xena carries an O gene?

5. Partitioning the Deck [12 points]

You have a non-standard deck of 5 suits of (Ace,2,3,4,5,6,7,8,9,10,Jack,Queen), for a total of 60 cards. You will deal the cards uniformly at random into 4 hands, each containing 15 of the cards (so each card ends up in exactly one hand). Let A_i be the event that hand i has exactly one of the five aces. In this problem, we'll calculate $\mathbb{P}(A_1 \cap A_2 \cap A_3)$.

(a) We might hope that the A_i are independent of each other (it would make the calculation easier...). Prove that A_1 and A_2 are **not** independent by appropriate calculations. [4 points]

(b) Calculate $\mathbb{P}(A_1 \cap A_2 \cap A_3)$. You must use the chain rule! [8 points]

6. Job applications [20 points]

Your company has recently opened up a new position and you're in charge of the hiring process! Applicants will be presented to you in a sequential list and for the sake of time, you can only traverse this list of applicants once. For each applicant, you get two actions: accept or reject. Furthermore, you cannot look ahead or go back to a previous applicant. If you accept an applicant, the position will be filled and the rest of the applicants will be automatically rejected. If you reject an applicant, you never get to see them again. Your goal is to maximize your chances of finding the best candidate out of the n applications.

You don't know much about the applicants (you can't estimate the chances that the current person is the best candidate), but you do know what you're looking for – you can immediately rank a new applicant *relative to those you have already seen*. You can assume that the n applicants will be presented in a uniformly random order.

A probability expert tells you that the optimal strategy is as follows:
Reject the first $q-1$ applicants you encounter (regardless of how good you think they are) for some number q.
Starting with applicant q, you will accept with the first applicant who is better than everyone you have seen so far.

In this problem, we'll compute the best value of q.

You may assume that $n \geq 1$.

(a) First, for a baseline, suppose your strategy were instead to accept the third applicant no matter what. What is your probability of accepting the best applicant among the n profiles? [3 points]

(b) Now, let's start analyzing our strategy. For two natural numbers $q \leq i$, compute the probability that the best applicant among the first $q-1$ is also the best applicant among the first $i-1$ (so the $\max[1,i) = \max[1,q)$). You may assume $1 < q \leq i \leq n$. [5 points]

(c) You accept the first applicant at index q or later that is better than all the prior applicants you have seen. Supposing that the best applicant is at index i, what is the probability that you will match with the best applicant? Unlike in the previous part, for this part you will also need to handle the case that $i < q$; you may still assume that $1 < q$. (Hint: use part(b)!) [5 points]

(d) We now set up a formula for the probability of selecting the best applicant if we ignore everyone before an arbitrary point q (i.e., we only start considering accepting someone if they are the q^{th} person we see or above). Use the Law of Total Probability to express the quantity as a summation over all possible placements of the best applicant. You will need to reason about the definition of our events to come up with the final result. Previous parts may be helpful here.

The final answer is not "pretty" for this problem (ours still has a summation in it, for example); simplify as far as you can, but don't expect a clean final answer. You also might need to have a separate formula for very small values of q or n (we have a special case when $q = 1$. If you have a separate case, you should explain where it comes from). To help you confirm if your answer is correct, when $n = 10$ and $q = 5$, the probability is approximately 0.3983, when $n = 10$ and $q = 4$ the probability approximately 0.3987. [5 points]

(e) If $n = 100$, what is the best value of q? If $n = 1000$, what is the best value of q? [2 points]
You do not need to provide an explanation for this part, but you may find it helpful to write a program or use graphing software for this part.

7. Naive Bayes [Coding, 20 points]

Use the Naive Bayes Classifier to implement a spam filter that learns word spam probabilities from our pre-labeled training data and then predicts the label (ham or spam) of a set of emails that it hasn't seen before.

Write your code for the following parts in the provided file: `cse312_pset3_nb.py`.

We have extra resources to help!

- We have slides and an optional video of a TA walking through the slides to introduce the assignment (video will go up on panopto by Friday).
- This optional Ed lesson might help you understand the pieces that go into it.
- You can also use these old notes for help clarifying concepts, but beware that their implementation is slightly different than what we're looking for.

Some notes and advice:

- Read about how to avoid floating point underflow using the log-trick in the notes.
- Make sure you understand how Laplace smoothing works.
- Remember to remove any debug statements that you are printing to the output.
- **Do not directly manipulate file paths or use hardcoded file paths.** A file path you have hardcoded into your program that works on your computer won't work on the computer we use to test your program.
- Needless to say, you should practice what you've learned in other courses: document your program, use good variable names, keep your code clean and straightforward, etc. Include comments outlining what your program does and how. We will not spend time trying to decipher obscure, contorted code. Your score on Gradescope is your final score, as you have unlimited attempts. **START EARLY**.
- We will evaluate your code on data you don't have access to, in addition to the data you are given.

Remember, it is not expected that Naive Bayes will classify every single test email correctly, but it should certainly do better than random chance! As this algorithm is deterministic, you should get a certain specific test accuracy around 90-95%, which we will be testing for to ensure your algorithm is correct. Note that we will run your code on a test dataset you haven't seen, but you will know immediately if you got full score.

(a) Implement the function `fit`.

(b) Implement the function `predict`.

8. Feedback + Collaboration [1 point]

Answer these questions on the separate Gradescope box for this question.

Please keep track of how much time you spend on this homework and answer the following questions. This can help us calibrate future assignments and future iterations of the course, and can help you identify which areas are most challenging for you.

- Which students did you collaborate with for this homework?
- Is the work that you are submitting your own and does not violate the academic integrity policy outlined in the syllabus?
- How many hours did you spend working on this assignment (excluding any extra credit questions, if applicable)? Report your estimate to the nearest hour.

- Which problem did you spend the most time on?
- Any other feedback for us?