Homework 3: Conditional Probability

In general, your goal in an explanation is to write enough that a student from class who has attended lecture, but not thought about 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.

Remember that you must "select pages" to identify which problem is on which page as part of your gradescope upload.

Submission: You must upload a **pdf** of your written solutions to Gradescope under "HW 3 [Written]". The use of $\LaTeX$ is *required*. 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 PM Wednesday July 12**.
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.

Note that there will be **no leeway** with the submission time. Any submission to Gradescope after 11:00:00 PM on the deadline day will be considered late. We recommend submitting a few minutes before the deadline in order to avoid technical issues with submissions.

Collaboration: Please read the full collaboration policy. If you work with others, you must still write up your solution independently and name all of your collaborators somewhere on your assignment.

Full submission instructions are in the course syllabus.

1. To Guess or Not to Guess [8 points]

You take a multiple choice exam. With probability $p = 0.3$, you know the answer to the question (and always get it correct). If you don't know the answer, you guess randomly among the 6 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. Do you ever urn? [16 points]

Like any normal teenager, Alice has two urns. Urn A contains 12 pink balls and 8 green balls. Urn B contains 5 pink balls and 15 green balls.

Alice will perform the following experiment: she flips a fair coin. If the coin is heads, she goes to urn A. If the coin is tails, she goes to urn B.

Then from whichever urn she is standing in front of, she draws two balls independently with replacement (that is, she will put the first ball back before drawing the second).

Let G_1 be the event that the first ball drawn is green and G_2 be the event that the second ball is green.

(a) Calculate $\mathbb{P}(G_1)$. Be sure to show your work, including starting with any formulas with symbols before plugging in numbers. [4 points]

(b) Calculate $\mathbb{P}(G_1 \cap G_2)$. Be sure to show your work, including starting with any formulas with symbols before plugging in numbers. [4 points]

(c) Calculate $\mathbb{P}(G_1|G_2)$. Be sure to show your work, including starting with any formulas with symbols before plugging in numbers. [4 points]

(d) Based on your calculations so far, are G_1 and G_2 independent? Using your calculations and the definition of independence, justify your assertion. [2 points]

(e) The result might be counter-intuitive. Explain the result **intuitively** (that is, you should not just refer directly to the numbers and the definition of independence; instead explain the result conceptually). [2 points]

3. Ace up my sleeve [12 points]

You have a standard deck of 52 cards. You will deal the cards into 4 hands, each containing 13 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 four aces. In this problem, we'll calculate $\mathbb{P}(A_1 \cap A_2 \cap A_3 \cap A_4)$.

(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) Use the chain rule to calculate $\mathbb{P}(A_1 \cap A_2 \cap A_3 \cap A_4)$. [8 points]

4. Secret Admirers [20 points]

Suppose you are using a dating app where you are presented with a list of n user profiles sequentially: you only get to traverse the list once, one user at a time. You do not get to look ahead or go back. For each user, you get two actions: match or pass. If you match, you can no longer look at other people's profiles and commit to your match. If you pass, you never get to see them again. The app guarantees that whoever you "match" with will agree to go out on a date with you. Your goal is to maximize your chances of finding the best date out the n people. You are torn: match too early, you miss out on someone better who may come along later. Match too late and you might let "the one" slip away. You don't know much about your potential dating pool (so you can't estimate the chances that the current person is the best), but you do know what you're looking for — you can immediately rank a new profile *relative to those you have already seen*. You also know that the app will show you the n profiles in a uniformly random order.

An oracle tells you that the optimal strategy is as follows:
reject the first $q-1$ admirers you encounter (regardless of how good you think they are) for some number q.
Starting with profile q, you will match with the first profile 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 match with the third profile no matter what. What is your probability of matching with your favorite among the n profiles. You may assume $n \geq 3$ for this part. [3 points]

(b) Now, let's start analyzing our strategy. For two natural numbers $q \leq i$, compute the probability that the best profile among the first $q-1$ is also the best profile 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 match with the first profile at index q or later that is better than all the prior profiles you have seen. Supposing that the best profile is at index i, what is the probability that you will match with the best profile?

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 match if we ignore everyone before an arbitrary point q (i.e., we only start considering matching with 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 match. 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.

5. Real-World: Bayes Theorem [12 points]

The tools of this class are useful to computer scientists, but many of them are useful beyond just "classic" computer science. In this assignment you'll consider an application of Bayes' Rule in the real-world.

We will consider the use of DNA evidence and witnesses in criminal trials. We acknowledge that a full discussion of DNA evidence and witnesses would require a discussion of many other issues [1] – for this problem, we are going to limit ourselves to just how information should be communicated to juries, from a probabilistic viewpoint.

5.1. Case 1

DNA evidence has been used in court cases for decades. Over time some common patterns of (dubious) argumentation have emerged, which you'll analyze in this problem.

Consider the following scenario:

A crime is committed somewhere in Seattle. No witnesses were at the crime, but there was blood left at the scene which had DNA extracted from it. The DNA was run against the 13 million DNA samples on file with the FBI. There was one match: a person who lived in Tacoma at the time of the crime.

You know the following facts about the DNA test that was run:

- The false positive rate of the test is only $\frac{1}{10{,}000{,}000}$.
- The false negative rate of the test is $\frac{1}{100{,}000{,}000}$.

5.1.1. The Prosecution

The prosecutor argues as follows

> The DNA match with the blood on the scene is strong. There is only a $\frac{1}{10{,}000{,}000}$ chance that the defendant is innocent (after all, the test only has a $\frac{1}{10{,}000{,}000}$ rate of failure) – certainly not a reasonable amount of doubt. You must vote to convict.

Let T be the event of a positive test, and G be the event that the defendant is guilty.

(a) In terms of G and T What probability or conditional probability is the prosecutor describing with their phrase "the chance the defendant is innocent, knowing about the test"? [1 point]

[1] Among others: under what circumstances DNA samples be taken from people and/or stored in databases.

(b) What probability or conditional probability does the $\frac{1}{10{,}000{,}000}$ come from? [1 point]

(c) Describe the prosecutor's error concisely (2-3 sentences). [2 points]

5.1.2. The Defense

The defense attorney argues as follows:

> The test isn't as good as it sounds. If we ran the test on all 330,000,000 people in the country, we'd have 33 innocent people come up with positive tests. The true probability of my client being guilty is only about $1/34$.

Recreate the Bayes' Rule application that the defense attorney is using

(a) What prior is being used? Recall the "prior" is the probability you're hoping to analyze *prior to* running the test. Your answer here should include both a number and where it came from. [2 points]

(b) Now use Bayes' rule to confirm that (starting from that prior), the calculation is correct. [2 points]

5.2. Case 2

In a separate court case, a witness to a crime claims that that they saw that the criminal wore a blue shirt (in a world where we wear the same color shirts every day...). The witness is able to correctly identify whether a person is wearing a blue shirt or not 95% of the time. 7% of the population wears blue shirts. The jury determines that since the witness can identify blue-shirt-wearers with 95% accuracy, there is a 95% chance the suspect was wearing a blue shirt. Let B be the event that the suspect wore a blue shirt, and W be the event that the witness says the suspect wore a blue shirt.

(a) In terms of B and W, what probability or conditional probability is the jury considering to hold 95% of the time as referred to by 'there is a 95% chance the suspect was wearing a blue shirt'? [1 point]

(b) In terms of B and W, suggest a probability or conditional probability that the jury should consider for a better estimate of whether the suspect was wearing a blue shirt. [1 point]

(c) Calculate the probability you suggested in part (b). Round to the fourth decimal. [2 points]

6. 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 to introduce the assignment.
- The textbook has a description of the background for this assignment under topic **9.3**.

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`.