Homework 3: Conditional Probability

Version 2: We clarified assumptions you can make about i, q in problem 5. The new assumptions are in blue.

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 April 21 (Seattle time, i.e. GMT-7).
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.

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

1. Miss Independent [15 points]

Alice has two urns. Urn A 10 purple balls and 5 gold balls. Urn B contains 5 purple balls and 10 Gold 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 gold and G_2 be the event that the second ball is gold.

(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 caclulations and the definition of independence, justify your assertion. [1 point]

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

2. Guessing Game [8 points]

You take a multiple choice exam. With probability p you know the answer to the question (and 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.

(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.

3. $\mathbb{P}$NA [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

Please 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

Suppose Xena and both of her parents have type A blood, but her sister Yvonne has type O.

(a) Explain what Yvonne's phenotype tells us about her and Xena's parents' genotype. With that in mind, what is the probability that Xena carries an O gene?
Hint: To start you off and to get a feel for the notation, you are calculating $P(G_X = AO|Ph_X = A))$

(b) Xena marries Zachary, who has type O blood. Compute the probability that their first child will have type O blood. Make sure to represent the event using the required notation, and show all your work maniulating the equation to get your final result.

(c) If Xena and Zachary's first child had type A blood, what is the probability that Xena carries an O gene?

4. Partitioning the Deck [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 $\mathbb{P}\left(A_1 \cap A_2 \cap A_3 \cap A_4\right)$.

(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 $\mathbb{P}\left(A_1 \cap A_2 \cap A_3 \cap A_4\right)$. Hint: you might want to use the chain rule! [8 points]

5. 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 a long 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. [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) Suppose that the best profile is at index i, and we match to the first profile q or after that is better than all the prior ones. What is the probability that you will match with profile i (Hint: use part b!) Unlike in the previous part, for this part you will also need to handle the case that $i < q$; you may still assume that $1 < q$. [5 points]

(d) We now set up a formula for the probability of selecting the best if we ignore everyone before an arbitary point q (so 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 admirer. 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 (you may wish to write a program or use graphing software for this part)

(f) **Extra Credit:** As $n \to \infty$, what will the best q be? Give a formula and prove that it is the maximizer of your function. Hint: our proof finds a function for which part of the formula is a Riemann sum and involves the function $-x \ln(x)$. To get any credit for this problem, you must have both the formula and a clear explanation.

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.

We have some slides that walk through the concepts needed to complete the assignment. We'll have video of a TA walking through these slides soon. This optional Ed lesson might help you understand the pieces that go into it.

You can also use these notes for help clarifying concepts, but beware that their implementation is slightly different than what we're looking for.

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

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`.