CSE 312: Foundations of Computing II Summer 2022

Problem Set 3 (due Friday, July 15, 11:59pm)

Directions: *For each problem, explain/justify how you obtained your answer, as correct answers without an explanation may receive* ***no credit****. Moreover, in the event of an incorrect answer, we can still try to give you partial credit based on the explanation you provide. Unless you are asked to, you should leave your answer in terms of factorials, combinations, etc., for instance* 26^7 *or* $26!/7!$ *or* $26 \cdot \binom{26}{7}$.

Submission*: You must upload a* ***pdf*** *of your solutions to Gradescope under "Pset 3 [Written]". The use of LaTex is highly recommended, and we have provided a template. (Note that if you want to hand-write your solutions, you'll need to scan them. If we cannot make out your writing, your work may be ungradable, so make sure it is legible.) Your code will be submitted as a .py file under "Pset 3 [Coding]".*

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, or you will potentially receive ***no credit*** *as mentioned in the syllabus. Please put each numbered problem on its own page in 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). As stated above, the coding problem will also be submitted to Gradescope.*

Collaboration*: This pset must be submitted* ***individually****. You are welcome and encouraged to discuss approaches with your fellow students, but everyone* ***must write up their own solutions.*** *Failure to do so is an instance of academic dishonesty.*

1. Conditional Support (16 points)

Consider a probability space $(\Omega, \Pr(\cdot))$ and suppose that F is an event in this space where $\Pr(F) > 0$. Verify that $(\Omega, \Pr(\cdot \mid F))$ is a valid probability space, i.e., that it satisfies the following required three axioms:

(a) $\Pr(E \mid F) \geq 0$ for all events $E \subseteq \Omega$.

(b) $\Pr(\Omega \mid F) = 1$.

(c) For any two mutually exclusive events G and H in Ω,

$$\Pr(G \cup H \mid F) = \Pr(G \mid F) + \Pr(H \mid F)\,.$$

2. Testing: 1, 2, 3 (16 points)

You are taking a multiple choice test that has 4 answer choices for each question. In answering a question on this test, the probability you know the correct answer (and choose it) is p. If you don't know the correct answer, you choose one (uniformly) at random. What is the probability that you knew the correct answer to a question, given that you answered it correctly?

3. Aces (16 points)

Suppose that an ordinary deck of 52 cards (which contains 4 aces) is randomly divided into 4 hands of 13 cards each. We are interested in determining p, the probability that each hand has an ace. Let E_i be the event that the i-th hand has exactly one ace. Determine

$$p = \Pr(E_1 \cap E_2 \cap E_3 \cap E_4)$$

using the chain rule.

4. Balls (20 points)

Consider an urn containing 12 balls, of which 8 are white and the rest are black. A sample of size 4 is to be drawn (a) with replacement, and (b) without replacement. What is the conditional probability (in each case) that the first and third balls drawn will be white given that the sample drawn contains exactly 3 white balls?

Note that drawing balls *with replacement* means that after a ball is drawn (uniformly at random from the balls in the bin) it is put back into the urn before the next independent draw. If the balls are drawn *without replacement*, the ball drawn at each step (uniformly at random from the balls in the bin) is not put back into the urn before the next draw.

Please use the following notation in your answer: Let W_i be the event that the i^{th} ball drawn is white. Let B_i be the event that that the i^{th} ball drawn is black, and let F be the event that exactly 3 white balls are drawn.

5. Naive Bayes [Coding] (16 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. See the slides from section 3 for details on implementation and Section 9.3 in the textbook.
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`.

6. Real World Vaccine Data (16 points)

Given the numerous discussions around vaccines and clinical trials that we have had this year, we would like you to get a deeper understanding of what some of these terms mean. We will look at a simplified version of vaccine trial information and try to understand what vaccine efficacy really means. There is much more that can be extrapolated from vaccine trial data, but we will be limiting ourselves to a small subset.
The following is a simplified version of Table 3 from the paper that discusses the efficacy of BNT162b2, the Pfizer-made vaccine for COVID-19:

Efficacy Endpoint Subgroup	Number of COVID Cases in Vaccine Group (N=17,397)	Number of COVID Cases in Placebo Group (N=17,498)	Vaccine Efficacy, %
Overall	8	162	95.0
Age Group			
16-55 years	5 (N=9897)	114 (N=9955)	95.6
>55 years	3 (N=7500)	48 (N=7543)	93.7
>65 years	1 (N=3848)	19 (N=3880)	94.7
>75 years	0 (N=774)	5 (N=785)	100.0

Table 1: Trial data for the Pfizer made vaccine for COVID-19. Source: Link

(1) Starting with the basics (4 Points)

Answer the questions using table 1. Let the event V denote that a random participant was in the vaccine group, and the event S denote that the participant got sick.

(a) Whats the probability that a random participant gets sick, regardless of what group they were in? What about for each participant group (vaccine and placebo)? Clearly state the symbols you use for each event and the conditional and unconditional probability.

(b) Given that a participant got sick, what is the probability that they were in the vaccine group? This should be a direct Bayes Rule calculation.

(c) Is this enough data to convince you to get the vaccine? Is there anything important that this table is missing?

(2) So what is Efficacy anyway? (6 Points)

(a) "Vaccine Efficacy %" is defined as

$$100 * \left(1 - \frac{\mathbb{P}(S \mid V)}{\mathbb{P}(S \mid V^C)}\right)$$

where S is the event that a random patient gets sick, and V is the event that a random patient is in the vaccine group. Unfortunately, most people confuse efficacy this with

$$1 - \mathbb{P}(S \mid V)$$

What is the difference between these two equations?

(b) Why do you think we use vaccine efficacy % rather than $1 - \mathbb{P}(S \mid V)$ to measure how well a vaccine is protecting against a disease? Can you come up with a scenario that illustrates why the former metric is more useful than the latter? If not, in your own words explain why efficacy is calculated as

$$100 * \left(1 - \frac{\mathbb{P}(S \mid V)}{\mathbb{P}(S \mid V^C)}\right)$$

and not

$$1 - \mathbb{P}(S \mid V)$$

(3) Debunking Myths (6 Points)

Table 2 is from the same paper about the Pfizer vaccine. It concerns the occurrence of "adverse events" e.g., negative reactions after a vaccine. Note that the number of people in the trial here is higher because some people in this table dropped out mid-way through the trials.

(a) Some people have argued that taking a vaccine is not worth it because the chance of developing a severe adverse event after a vaccine $\frac{240}{21,621}$ is larger than the chance of actually getting COVID-19 without a

Event Type	Vaccine Group (N=21,621)	Placebo (N=21,631)
Any (including mild reactions)	5770	2638
Severe	240	139
Life-threatening	21	24

Table 2: Vaccine side effects data from clinical trials. Note: The data is cumulative, so "Life-threatening" is included in the "Severe" and "Any" categories and "Severe" is included in the "Any" category.

vaccine $\frac{162}{17,498}$. Using the symbols, V and S above, as well as A_s to denote the event of a participant getting a severe adverse event, what two conditional probabilities are we comparing here? State both the conditional probabilities as $\mathbb{P}(A \mid B)$ clearing stating the events A and B. Hint: you can use events you have already defined in the previous parts.

(b) What is a logical flaw in the argument above? How would you argue that getting a vaccine is still worth it? Note: there are several answers to this question! Be creative!