CSE 312: Foundations of Computing II — Winter 2021

Problem Set 3 (due Wednesday, January 27, 11:59pm)

Directions:
***Answers**: For each problem, remember you must briefly explain/justify how you obtained your answer, as correct answers without an explanation will 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. It is fine for your answers to include summations, products, factorials, exponentials, or combinations; you dont need to calculate those all out to get a single numeric answer, for instance 26^7 or $26!/7!$ or $26 \cdot \binom{26}{7}$.*

Your solutions need to be concise and clear. We will take off points for lack of clarity or for excess verbosity. Please see section worksheet solutions (posted on the course website) to gauge the level of detail we are expecting.

Please clearly indicate your final answer, in such a way as to distinguish it from the rest of your explanation.

***Groups:** This homework is to be completed in groups of 1 or 2. Specific guidelines about collaboration are available on the syllabus, but every group will be submitting their own submission. Please cite any collaboration at the top of your submission. Instructions are available here as to how to add groupmates to your submission.*

Before you start your homework, write down the list of people you collaborated with. Remember, you can only collaborate outside your group by discussing the problems at a high level. Provide names and email addresses.

***Submission**: You must upload a **pdf** of your written solutions to Gradescope under "PSet 3 [Written]". Problem 8 is a coding problem, so under "Pset 3 [Coding]" you will be uploading a .py file called cse312_pset3_nb.py. (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.)

Please cite any collaboration at the top of your submission (beyond your group members, who should already be listed).

1. Testing: 1, 2, 3 (12 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?

2. Pharmaceutical trials (16 points)

A pharmaceutical company proudly publishes results from a trial of its new test for a certain genetic disorder. The false negative rate is small: the test returns a negative result for only 4% of patients with the disorder. The false positive rate is also small: the test returns a positive result for only 12% of participants that do not have the disorder. Assume that 0.5% (that is, the fraction 0.005) of the population has the disorder. Lets see how good a test this will be and what a test result would mean to you as a patient. Calculate your answers to 2 significant digits.

(a) [4 Points] What is the probability of having the disorder if you have a negative test result? (Seeing your answer, how reassured should you be if you were the one that had a negative test result?)

(b) [4 Points] What is the probability of having the disorder if you have a positive test result? (Seeing your answer, how anxious should you be if you were the one that had a positive test result?)

(c) [4 Points] Repeat part (a) assuming that 15% of the population has the disorder.

(d) [4 Points] Repeat part (b) assuming that 15% of the population has the disorder.

3. Blood types (15 points)

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, and everyone passes a randomly selected copy to each of their children (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".) Suppose Xena and both of her parents have type A blood, but her sister Yvonne has type O. Give exact answers as simplified fractions and provide a 1-2 sentence to explain your reasoning for each of them.

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.

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 $\#$. 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).

To start you off and to get a feel for the notation, what you are trying to find in the first part is

$$P(G_X = AO | Ph_X = A))$$

(a) What is the probability that Xena carries an O gene?

(b) Xena marries Zachary, who has type O blood. What is the probability that their first child will have type O blood?

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

4. Balls (12 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 independent 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. Aces (10 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.

6. Doggone, Doggtwo, Doggthree... (10 points)

A hunter has two hunting dogs. One day, on the trail of some animal, the hunter comes to a place where the road diverges into two paths. She knows that each dog, independent of the other, will choose the correct path with probability p. The hunter decides to let each dog choose a path, and if they agree, take that one, and if they disagree, to randomly pick a path. What is the probability that she ends up taking the correct path?

Hint: Use the law of total probability, partitioning based on whether the dogs choose the same path or different paths.

7. Conditional probability and probability spaces (12 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,

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

8. 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. See the slides from section 3 for details on implementation and notes from Alex's book. 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.