Lisa Yan and Jerry Cain
CS 109

Problem Set #2
September 25, 2020

Problem Set #2
Due: 1:00pm on Monday, October 5

With problems by Mehran Sahami, Chris Piech, Lisa Yan and Alex Tsun

- Submit on Gradescope by 1:00pm Pacific on Monday, October 5th, for a small, "on-time" bonus.
- All students have a pre-approved extension, or "grace period" that extends until Wednesday 1:00pm Pacific, when they can submit with no penalty. **The grace period expires on 1:00 Pacific on Wednesday, October 7th**, after which we cannot accept further late submissions.
- **Collaboration policy:** You are encouraged to discuss problem-solving strategies with each other as well as the course staff, but you must write up your own solutions and submit individual work. Please cite any collaboration at the top of your submission.
- **Tagging written problems**: When you submit your written PDF on Gradescope you must tag your PDF, meaning that you must assign pages of your PDF as answers to particular questions so that we can properly grade your submission. For problem sets, we are deducting **2 points** for any submissions that do not have all questions tagged.
- **For each problem, briefly explain/justify how you obtained your answer.** Brief explanations of your answer are necessary to get full credit for a problem even if you have the correct numerical answer. The explanations help us determine your understanding of the problem whether or not you got the correct answer. 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 don't need to calculate those all out to get a single numeric answer.
- If you handwrite your solutions, you are responsible for making sure that you can produce **clearly legible** scans of them for submission. You may also use any word processing software you like for writing up your solutions. On the CS109 webpage we provide a template file and tutorial for the LaTeX system, if you'd like to use it.

1 Written Problems

Submit your solutions to these written problems as a single pdf file on Gradescope.

1. Say in Silicon Valley, 35% of engineers program in Java and 28% of the engineers who program in Java also program in C++. Furthermore, 40% of engineers program in C++.
    a. What is the probability that a randomly selected engineer programs in Java and C++?
    b. What is the conditional probability that a randomly selected engineer programs in Java given that they program in C++?

2. A website wants to detect if a visitor is a robot or a human. They give the visitor five CAPTCHA tests that are hard for robots but easy for humans. If the visitor fails one of the tests, they are flagged as a robot. The probability that a human succeeds at a single test is 0.95, while a robot only succeeds with probability 0.3. Assume all tests are independent. The percentage of visitors on this website that are robots is 5%; all other visitors are human.
    a. If a visitor is actually a robot, what is the probability they get flagged (the probability they fail at least one test)?
    b. If a visitor is human, what is the probability they get flagged?
    c. Suppose a visitor gets flagged. Using your answers from part (a) and (b), what is the probability that the visitor is a robot?
    d. If a visitor is human, what is the probability that they pass exactly three of the five tests?
    e. Building off of your answer from part (d), what is the probability that a visitor with unknown identity passes exactly three of the five tests?

3. Say all computers either run operating system W or X. A computer running operating system W is twice as likely to get infected with a virus as a computer running operating system X. If 70% of all computers are running operating system W, what percentage of computers infected with a virus are running operating system W?

4. The Superbowl institutes a new way to determine which team receives the kickoff first. The referee chooses with equal probability one of three coins. Although the coins look identical, they have probability of heads 0.1, 0.5 and 0.9, respectively. Then the referee tosses the chosen coin 3 times. If more than half the tosses come up heads, one team will kick off; otherwise, the other team will kick off. If the tosses resulted in the sequence H, T, H, what is the probability that the fair coin was actually used?

5. After a long night of programming, you have built a powerful, but slightly buggy, email spam filter. When you don't encounter the bug, the filter works very well, always marking a spam email as SPAM and always marking a non-spam email as GOOD. Unfortunately, your code contains a bug that is encountered 10% of the time when the filter is run on an email. When the bug is encountered, the filter always marks the email as GOOD. As a result, emails that are actually spam will be erroneously marked as GOOD when the bug is encountered. Let p denote the probability that an email is actually non-spam, and let q denote the conditional probability that an email is non-spam given that it is marked as GOOD by the filter.
    a. Determine q in terms of p.
    b. Using your answer from part (a), explain mathematically whether q or p is greater. Also, provide an intuitive justification for your answer.

6. Two cards are randomly chosen without replacement from an ordinary deck of 52 cards. Let E be the event that both cards are Aces. Let F be the event that the Ace of Spades is one of the chosen cards, and let G be the event that at least one Ace is chosen.
    a. Compute $P(E \mid F)$.
    b. Are E and F independent? Justify your answer using your response to part (a).
    c. Compute $P(E \mid G)$.

7. Your colleagues in a comp-bio lab have sequenced DNA from a large population in order to understand how a gene (G) influences two particular traits (T_1 and T_2). They find that $P(G) = 0.6$, $P(T_1 \mid G) = 0.7$, and $P(T_2 \mid G) = 0.9$. They also observe that if a subject does not have the gene G, they express neither T_1 nor T_2. The probability of a patient having both T_1 and T_2 given that they have the gene G is 0.63.

   a. Are T_1 and T_2 conditionally independent given G?
   b. Are T_1 and T_2 conditionally independent given G^C?
   c. What is $P(T_1)$?
   d. What is $P(T_2)$?
   e. Are T_1 and T_2 independent?

8. The color of a person's eyes is determined by a pair of eye-color genes, as follows:

   - if both of the eye-color genes are blue-eyed genes, then the person will have blue eyes
   - if one or more of the genes is a brown-eyed gene, then the person will have brown eyes

   A newborn child independently receives one eye-color gene from each of its parents, and the gene it receives from a parent is equally likely to be either of the two eye-color genes of that parent. Suppose William and both of his parents have brown eyes, but William's sister (Claire) has blue eyes. (We assume that blue and brown are the only eye-color genes.)

   a. What is the probability that William possesses a blue-eyed gene?
   b. Suppose that William's wife has blue eyes. What is the probability that their first child will have blue eyes?

9. Consider the following algorithm for betting in roulette (`https://en.wikipedia.org/wiki/Roulette`). At each round ("spin"), you bet \$1 on a color ("red" or "black"). If that color comes up on the wheel, you keep your bet AND win \$1; otherwise, you lose your bet.

   i. Bet \$1 on "red"
   ii. If "red" comes up on the wheel (with probability 18/38), then you win \$1 (and keep your original \$1 bet) and you **immediately** quit (i.e., you do not do step (iii) below).
   iii. If "red" did not come up on the wheel (with probability 20/38), then you lose your initial \$1 bet. But, then you bet \$1 on "red" on *each* of the next **two** spins of the wheel. After those two spins, you quit (no matter what the outcome of the next two spins).

   Let X denote your "winnings" when you quit, i.e., the total amount of money won minus any amounts lost while playing. This value may be negative.

   a. Determine $P(X > 0)$.
   b. Determine $E[X]$. (Rhetorical question: Would you play this game?)

2 Coding

10. **[Coding + Written]** After the Ebola outbreak of 2015, there was an urgent need to learn more about the virus. You have been asked to uncover how a particular group of bat genes impact an important trait: whether the bat can carry Ebola. Nobody knows the underlying mechanism; it is up to you to hypothesize what is going on. For 100,000 independently sampled bats you have collected data of whether or not five genes are expressed, and whether

or not the bat can carry Ebola.[1] If a gene is expressed, it can affect both the probability of other genes being expressed and the probability of the trait being expressed. You can find the data in a file called `bats.csv`. A value of 1 denotes True, whereas a value of 0 denotes False. Each row in the file corresponds to **one bat** and has 6 columns representing Boolean values:

- Boolean 0: Whether the 1st gene is expressed in the bat (G_0)
- Boolean 1: Whether the 2nd gene is expressed in the bat (G_1)
- Boolean 2: Whether the 3rd gene is expressed in the bat (G_2)
- Boolean 3: Whether the 4th gene is expressed in the bat (G_3)
- Boolean 4: Whether the 5th gene is expressed in the bat (G_4)
- Boolean 5: Whether the trait is expressed in the bat; i.e., the bat can carry Ebola (T)

Follow the instructions in each subpart of this question to either write code or submit written answers in your PDF. For code-writing questions, follow the below guidelines so that your code works with our autograder:

- Do not modify the name or signature of any function we ask you to write, though you may use helper functions if you wish.
- You'll write code in the file `cs109_pset2.py`, which you can download from the course website. Submit only that file to Gradescope, and do not modify the name of that file. Do not submit a zip file.
- Make sure that your return values are in the format we expect as described in each part.
- Do not use global variables.

a. **[Coding]** First, calculate the probability of the trait being expressed, namely $P(T)$, along with $P(G_i)$ for each gene i. In this part, you will implement the function `calculate_probs`. Your function should return a numpy array with shape `(6,)`. Mathematically, you can think of your return array as a row vector with 6 columns: the elements at indices 0 through 4 will be $P(G_0), P(G_1), \ldots, P(G_4)$, respectively, and the element at index 5 will be $P(T)$.

Important: You can test your code using our autograder on Gradescope to verify that your solution matches ours. Our autograder calls your `calculate_probs` function using the same `bats.csv` file that is provided to you, then calls our hidden reference solution, and finally checks whether your answers match ours. Feel free to use this to your advantage—you can submit as many times as you like, and we will only grade your most recent submission. That being said, we have tests that test your code on a different `csv` file following the same format (i.e., 1 row per bat, 6 columns with boolean values, etc.), so make sure that your code is general enough to handle any data file in the specified format. In short: don't hard-code specific probabilities.

Here are some Python tips that you might find useful:

- We strongly recommend using numpy in this (and subsequent) questions.[2] You can load a `csv` file into a numpy array named `data` by using the function

[1] Humane note: bats can carry Ebola, but it causes them no harm. No fake bats were hurt in the making of this problem. Why are bats immune to the harmful effects? Open question!

[2] numpy is significantly faster than writing loops over files, arrays, etc. You might not notice the speed bump on this assignment, but it'll become very noticeable later when we talk about machine learning. Exciting!

`np.genfromtext` [3]:

```
data = np.genfromtxt(filename, delimiter=',')
```

- You can get the `i`-th column from data using slicing by writing `data[:,i]`. That returns an array of shape `(n, )`, where `n` is the number of rows in the `csv` file.
- You can take the mean of a numpy array using the `np.mean` method. Example:

```
arr = np.array([1, 2, 3])
print(arr.shape)              # Output: (3, )
print(np.mean(arr))           # Output: 2.0
```

- **If you leverage the `axis` parameter in `np.mean`**, your function will be just a few lines long.

b. **[Coding]** For each gene i, calculate $P(T|G_i)$. In this part, you'll implement the function `calculate_cond_probs`. Your function will return a numpy array with shape `(5, )`, where the element at index i is $P(T \mid G_i)$.

Just like in part (a), we've provided an autograder that runs on `bats.csv`, though we may have additional tests on a different data file in the specified format.

Some more Python tips to complement the ones listed above:

- Check out the `np.where` [4] method. Specifically, if your `csv` is stored in a numpy array named `data`, you can store a subset of the rows where the `i`-th column is 1:

```
subset = data[np.where(data[:, i] == 1)]
```

- Check out the `np.logical_and` method [5].

c. **[Written]** For each gene i, decide whether or not you think that is would be reasonable to assume that G_i is independent of T. Support your argument with numbers. Remember that our probabilities are based on 100,000 bats, not infinite bats, and are therefore only estimates of the true probabilities.

If you need to write code, we have provided an optional `sandbox` function for convenience that gets called by the `main` function, but we won't grade any code for this part. You should include your writeup to this answer in the PDF that you upload to Gradescope.

d. **[Written]** Give your best interpretation of the results from (a) to (c). You should include your writeup to this answer in the PDF that you upload to Gradescope.

[3] https://docs.scipy.org/doc/numpy/reference/generated/numpy.genfromtxt.html
[4] https://thispointer.com/find-the-index-of-a-value-in-numpy-array/
[5] https://numpy.org/doc/stable/reference/generated/numpy.logical_and.html