David Varodayan
CS 109

Problem Set #2
January 17, 2020

Problem Set #2
Due: 1:00pm on Monday, January 27

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

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.

Written Problems

1. Let E and F be events defined on the same sample space S. Prove that:

$$P(EF) \geq P(E) + P(F) - 1$$

   (This formula is known as Bonferroni's Inequality.)

2. 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++?

3. A website wants to detect if a visitor is a robot. They give the visitor three 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 0.15. Assume all tests are independent.
   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. The percentage of visitors on the site that are robots is 5%. Suppose a visitor gets flagged. Using your answers from part (a), what is the probability that the visitor is a robot?

4. 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?

5. 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. Compute:
   a. $P(E \mid F)$
   b. $P(E \mid G)$

6. Two emails are received at a mail server. Suppose that each email is spam with probability 0.8 and that whether each email message is spam is an independent event from the other.
   a. Suppose that you are told that at least one of the two emails is spam. Compute the conditional probability that both emails are spam.
   b. Suppose now that one of the emails is randomly (accidentally) forwarded from the server to your account, and you see that this email is spam. What is the probability that both emails originally received by the server are spam in this case? Explain your answer.

7. 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.

8. Consider a hash table with 15 buckets, of which 9 are empty (have no strings hashed to them) and the other 6 buckets are non-empty (have at least one string hashed to each of them already). Now, 2 new strings are independently hashed into the table, where each string is equally likely to be hashed into any bucket. Later, another 2 strings are hashed into the table (again, independently and equally likely to get hashed to any bucket). What is the probability that both of the final 2 strings are each hashed to empty buckets in the table?

9. Five servers are located in a computer cluster. After one year, each server independently is still working with probability p, and otherwise fails (with probability $1 - p$).
   a. What is the probability that *at least* 1 server is still working after one year?
   b. What is the probability that *exactly* 2 servers are still working after one year?
   c. What is the probability that *at least* 2 servers are still working after one year?

10. 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?

11. A robot, which only has a camera as a sensor, can either be in one of two locations: L_1 (which does not have a window) or L_2 (which has a window). The robot doesn't know exactly where it is and it represents this uncertainty by keeping track of two probabilities: $P(L_1)$ and $P(L_2)$. Based on all past observations, the robot thinks that there is a 0.7 probability it is in L_1 and a 0.3 probability that it is in L_2.

    The robot then observes a window through its camera, and although there is only a window in L_2, it can't conclude with certainty that it is in fact in L_2, since its image recognition algorithm is not perfect. The probability of observing a window given there is no window at its location is 0.2, and the probability of observing a window given there is a window is 0.9. After incorporating the observation of a window, what are the robot's new probabilities for being in L_1 and L_2, respectively?

12. 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?

13. 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?

Continued on the next page…

Problem Set #2, Continued

Problem by Mehran Sahami, Chris Piech, and Lisa Yan with modifications by Anand Shankar.

14. **[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 that represent Boolean values:

- Boolean 0: Whether the 0th gene is expressed in the bat (G_0)
- Boolean 1: Whether the 1st gene is expressed in the bat (G_1)
- Boolean 2: Whether the 2nd gene is expressed in the bat (G_2)
- Boolean 3: Whether the 3rd gene is expressed in the bat (G_3)
- Boolean 4: Whether the 4th 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 answer questions in your PDF. For code-writing questions, follow the following constraints so your code works with our auto-grader:

- 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, and do not modify the name of that file.
- Make sure your return values are in the format we expect as described below.

A. **[Coding]** First, you will calculate the probability of the trait being expressed, namely $P(T)$, along with $P(G_i)$ for each gene *i*.

In this part, you'll write the function `part_a`. You should return a numpy array with shape `(6, )`. Mathematically, you can think of that as a row vector with 6 columns. The elements at indices 0 through 4 should be $P(G_0)$, $P(G_1)$, ... , $P(G_4)$ respectively, and the element at index 5 should be $P(T)$.

Important: You can test your code using our autograder on Gradescope to see if your solution matches ours. Our autograder calls your `part_a` function using the same `bats.csv` file that's 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;

[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!

you can resubmit as many times as you like, and we'll only grade your most recent submission. That being said, we might have hidden tests that test your code on a different `csv` file, though it would be in the same format described above (i.e. 1 row per bat, 6 columns with binary values, etc). Thus, make sure your code is general enough to handle any data file in the specified format. In other words, 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 `np.genfromtxt`:

```
    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 $P(T \mid G_i)$.

In this part, you'll write in the function `part_b`. You should return a numpy array with shape `(5, )`. The element at index *i* should be $P(T \mid G_i)$.

Just like in part (A), we've provided an autograder that runs on `bats.csv`, though we might have additional hidden tests on a data file that's different from the one we gave you. If we do, we promise the file will be in the same format that's described above.

Another Python tip to complement the ones listed above: check out the `np.where` 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 with the following code:

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

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

[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!)

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 provided a `part_c` function for convenience that gets called by the `main` function, but we won't grade any code for this part. You should write your answer in the PDF that you upload to Gradescope.

D. **[Written]** Give your best interpretation of the results from parts (a) through (c). Write your answer in the PDF you upload to Gradescope.

E. **[Written]** For extra credit, try and find conditional independence relationships between the genes and the trait. Incorporate this information to improve your hypothesis of how the five genes relate to whether or not a bat can carry Ebola. Write your answer in the PDF that you upload to Gradescope.