Lisa Yan
CS 109

Problem Set #2
October 4, 2019

# Problem Set #2
# Due: 1:00pm on Monday, October 14th

With problems by Mehran Sahami and Chris Piech

**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.

## Warm-up

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. A bit string of length $n$ is generated randomly such that each bit is generated independently with probability $p$ that the bit is a 1 (and 0 otherwise). How large does $n$ need to be (in terms of $p$) so that the probability that there is at least one 1 in the string is at least 0.6?

5. The probability that a Netflix user likes a movie $M_i$ from the "Tearjerker" genre, given that they like the Tearjerker genre, is $p_i$. The probability that a user likes $M_i$, given that they **do not like** the Tearjerker genre, is $q_i$. Of all Netflix users, 60% like the Tearjerker genre. Assume that **conditioned** on knowing a user's preference for the genre (either liking the genre or not liking it), liking movie $M_1$, $M_2$, and $M_3$ are **independent** events. Express all of your answers to the following questions in terms of $q$'s and $p$'s. What is the probability that:

   a. A user likes all three movies $M_1$, $M_2$, **and** $M_3$, given that they like the Tearjerker genre?
   b. A user likes at least one movie $M_1$, $M_2$, **or** $M_3$, given that they like the genre?
   c. A user likes the Tearjerker genre, given that they like $M_1$, $M_2$, **and** $M_3$?

## System Design

6. 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.

7. 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?

8. 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?

*Continued on the next page. . .*

## Program Analysis

9. Suppose we want to write an algorithm `fairRandom` for randomly generating a 0 or a 1 with equal probability (= 0.5). Unfortunately, all we have available to us is a function:

   ```
   int unknownRandom();
   ```

   that randomly generates bits, where on each call a 1 is returned with some unknown probability $p$ that need not be equal to 0.5 (and a 0 is returned with probability $1 - p$).

   Consider the following algorithm for `fairRandom`:

```
int fairRandom() {
    int r1, r2;
    while (true) {
        r1 = unknownRandom();
        r2 = unknownRandom();
        if (r1 != r2) break;
    }
    return r2;
}
```

   a. Show mathematically that `fairRandom` does indeed return a 0 or a 1 with equal probability.

   b. Say we want to simplify the function, so we write the `simpleRandom` function below. Would the `simpleRandom` function also generate 0's and 1's with equal probability? Explain why or why not. Determine P(`simpleRandom` returns 1) in terms of $p$.

```
int simpleRandom() {
    int r1, r2;
    r1 = unknownRandom();
    while (true) {
        r2 = unknownRandom();
        if (r1 != r2) break;
    }
    return r2;
}
```

*Continued on the next page...*

## Localization

10. 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?

11. **[Coding]** Your cell phone is constantly trying to keep track of where you are. At any given point in time, for all nearby locations, your phone stores a probability that you are in that location. Right now your phone believes that you are in one of 16 different locations arranged in a grid with the following probabilities (see the figure on the left):

Prior belief of location

| | | | |
|---|---|---|---|
| 0.05 | 0.10 | 0.05 | 0.05 |
| 0.05 | 0.10 | 0.05 | 0.05 |
| 0.05 | 0.05 | 0.10 | 0.05 |
| 0.05 | 0.05 | 0.10 | 0.05 |

$P$(Observe two bars of signal | Location)

| | | | |
|---|---|---|---|
| 0.75 | 0.95 | 0.75 | 0.05 |
| 0.05 | 0.75 | 0.95 | 0.75 |
| 0.01 | 0.05 | 0.75 | 0.95 |
| 0.01 | 0.01 | 0.05 | 0.75 |

Your phone connects to a known cell tower and records two bars of signal. For each grid location $L_i$, you can calculate the probability of observing two bars from this particular tower, assuming that cell phone is in location $L_i$ (see the figure on the right). That calculation is based on knowledge of the dynamics of this particular cell tower and stochasticity of signal strength.

As an example: the value of 0.05 in the highlighted cell on the left figure means that you believed there was a 0.05 probability that the user was in the bottom right grid cell prior to observing the cell tower signal. The value of 0.75 in the highlighted cell on the right figure means that you think the probability of observing two bars, given the user was in the bottom right grid cell, is 0.75.

Write a program to calculate, for each of the 16 locations, the new probability that the user is in each location given the cell tower observation. The matrices are provided on the website on the Problem Set #2 page. The grid in the left figure is stored in a file called `prior.csv`, and the grid in the right figure is stored in a file called `conditional.csv`. Report the probabilities of all 16 cells and provide the code for your program.

## DNA

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?
    c. Still assuming that William's wife has blue eyes, if their first child had brown eyes, what is the probability that their next child will also have brown 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?

14. **[Coding]** 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`. Each row in the file corresponds to **one bat** and has 6 Booleans:

    - 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 ($G_3$)

[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!

- Boolean 4: Whether the 4th gene is expressed in the bat ($G_4$)
- Boolean 5: Whether the 5th gene is expressed in the bat ($G_5$)
- Boolean 6: Whether the trait is expressed in the bat; i.e., the bat can carry Ebola ($T$)

Write a program to analyze the data you have collected. Report the following:

a. What is the probability of the trait being expressed $P(T)$?

b. For each gene $i$ calculate and report $P(G_i)$.

c. 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 just estimates of the true probabilities.

d. For each gene $i$ that is not assumed to be independent of $T$, calculate $P(T \mid G_i)$.

e. Give your best interpretation of the results from (a) to (d).

f. 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.