Will Monroe
CS 109

Problem Set #2
July 5, 2017

Problem Set #2
Due: 12:30pm on Wednesday, July 12th

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.

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 he/she programs in C++?

3. A games website has determined that they have a problem with players cheating using automated playing tools (robots). They decide to deploy random "bonus games" that are hard for robots, so they function as CAPTCHA-like tests of whether players are actually robots.

   Each player is given three bonus games. If the player fails in one of the games, they are flagged as a possible robot. The probability that a human succeeds at a single bonus game is 0.95, while a robot only succeeds with probability 0.3.
   a. If a player is actually a robot, what is the probability they get flagged (the probability they fail at least one game)? If a player is human, what is the probability they get flagged? Assume players (both humans and robots) succeed or fail in each of the three games independently.
   b. The fraction of players on the site that are cheating with robots is 1/10. Suppose a player gets flagged. Using your answers from part (a), what is the probability that player is a robot?

4. 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.

5. Consider a hash table with 5 buckets, where the probability of a string getting hashed to bucket i is given by p_i (where $\sum_{i=1}^{5} p_i = 1$). Now, 6 strings are hashed into the hash table.
   a. Determine the probability that *each of* the first 4 buckets has at least 1 string hashed to each of them. Explicitly expand your answer in terms of p_i's, so that it does not include any summations.
   b. Assuming $p_1 = 0.1, p_2 = 0.25, p_3 = 0.3, p_4 = 0.25, p_5 = 0.1$, explicitly compute your answer to part (a) as a numeric value.

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

7. 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?

8. 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?

9. 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?

10. 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;
}
```

11. 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?

12. A robot, which only has a camera as a sensor, can either be in one of two locations: L_1 or L_2. 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 values for $P(L_1)$ and $P(L_2)$?

13. **[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.