Chris Piech
CS109

PSet #2
April 14, 2017

Problem Set #2

Due: 10:30am on Monday, April 24th

For each problem, briefly explain/justify how you obtained your answer in order to obtain full credit. Your explanations will 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, unless you are specifically asked for a single numeric answer. Parts of this problem set are brand new. We will post errata if we update any problems. We hope you enjoy the new problems! Some questions are extra credit (and are optional) and some questions require coding.

Warmup

1. Say in Silicon Valley, 36% of engineers program in Java and 24% of the engineers who program in Java also program in C++. Furthermore, 33% 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++?

2. 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 | F)
   b. P(E | G)

3. 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, given a user's preference for the genre (either liking the genre or not liking it), liking movie M_i and M_j are **conditionally independent** events. Express all your answers in terms of qs and ps. What is the probability:
   a. That a user likes all three movies M_1, M_2 **and** M_3 given that they like the Tearjerker genre?
   b. That they like at least one movie M_1, M_2 **or** M_3 given that they like the Tearjerker genre?
   c. That they like the Tearjerker genre given that they like M_1, M_2 **and** M_3?

4. 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* 3 servers are still working after one year?
   c. What is the probability that *at least* 3 servers are still working after one year?

Errata: we updated the wording to question 3 to make it clear that a user's preference for movies is also conditionally independent given that the user does not like the genre.

5. 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 0.7?
6. 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.

Program Analysis

7. 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`**:

```
def fairRandom():
   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() {
   r1 = unknownRandom()
   while True:
      r2 = unknownRandom()
      if (r1 != r2): break
      r1 = r2
   return r2
```

   c. **[Extra Credit Coding]** Consider a game that uses a generator which produces independent random numbers between 1 and 100, inclusive, where each outcome is equally likely. The game starts with a sum S = 0. The first player adds random numbers from the generator to S until S > 100 and records her last random number 'x'. The second player, continues adding random numbers from the generator to S until S > 200 and records her last random number 'y'. The player with the highest number wins, i.e. if y > x the second player wins. Write a program to simulate 100,000 games. What is the probability estimate, based on your

simulations, that the second player wins? Give your answer rounded to 3 places behind the decimal. For even more extra credit, calculate the probability analytically. This question is optional, but a good way to get your head around data.

Localization

In this multi part problem you are going to solve key components for "Localization" which is the computer science problem of tracking the location of objects when there is uncertainty.

8. A robot, which only has a camera as a sensor, can either be in one of two locations: L1 or L2. The robot doesn't know exactly where it is and it represents this uncertainty by keeping track of two probabilities: P(L1) and P(L2). Based on all past observations, the robot thinks that there is a 0.8 probability it is in L1 and a 0.2 probability that it is in L2.

   The robot then observes a window through its camera, and although there is only a window in L2, it can't conclude that it is in fact in L2: 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 is the robot's new values for P(L1) and P(L2)?

9. **[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.

For each of the 16 location position, calculate the new probability that the user is in each location given the cell tower observation. Write a program to calculate the probabilities. The matrices are provided on the website on the problem set #2 page. Report the probabilities of all 16 cells and write a short explanation of your program. The grid in the left figure is stored in a file called "prior.csv" the grid in the right figure is stored in a file called "conditional.csv"

DNA

10. 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?

11. 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 $P(G) = 0.6$, $P(T_1|G) = 0.8$ and $P(T_2|G) = 0.9$. They also observe that if a subject does not have the gene, 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 is 0.72
    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?

12. **[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)

[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 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 5th gene is expressed in the bat (G_5)
- Boolean 6: Whether the trait is expressed in the bat, eg 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 *Gi* 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 | 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.