Chris Piech
CS 109

Problem Set #4
May 1, 2019

Problem Set #4
Due: 10am on Wednesday, May 15th

With problems by Mehran Sahami and Chris Piech

For each problem, briefly explain/justify how you obtained your answer. In fact, most of the credit for each problem will be given for the derivation/model used as opposed to the final answer. Make sure to describe the distribution and parameter values you used (e.g., Bin(10, 0.3)), where appropriate. Provide a numeric answer for all questions when possible.

Warmup

1. On average 5.5 users sign-up for an on-line social networking site each minute. What is the probability that:
   a. More than 7 users will sign-up for the social networking site in the next minute?
   b. More than 13 users will sign-up for the social networking site in the next 2 minutes?
   c. More than 15 users will sign-up for the social networking site in the next 3 minutes?

2. The joint probability density function of continuous random variables X and Y is given by: $f(X = x, Y = y) = \frac{4y}{x}$ where $0 < y < x < 1$
   a. What is the marginal density function of X?
   b. What is the marginal density function of Y?
   c. What is $E[X]$?

3. The median of a continuous random variable having cumulative distribution function F is the value m such that $F(m) = 0.5$. That is, a random variable is just as likely to be larger than its median as it is to be smaller. Find the median of X (in terms of the respective distribution parameters) in each case below.
   a. $X \sim \text{Uni}(a, b)$
   b. $X \sim \text{N}(\mu, \sigma^2)$
   c. $X \sim \text{Exp}(\lambda)$

4. You are tracking the distance to a satellite. An instrument reports that the satellite is 100 a.u. from Earth. Before you had observed the instrument reading, your belief distribution for the distance D of the satellite was a Gaussian $D \sim N(\mu = 98, \sigma^2 = 16)$. The instrument gives a reading that Gaussian with mean that is the true distance and variance 4.
   a What is the PDF of your prior belief of the true distance of the satellite?
   b What is the probability density of seeing an observation of 100 a.u. from your instrument, given that the true distance of the satellite is equal to t?
   c What is the PDF of your posterior belief (after observing the instrument reading) of the true distance of the satellite? You may leave a constant in your PDF and you do not need to simplify the PDF.

5. Let X, Y, and Z be independent random variables, where $X \sim N(\mu_1, \sigma_1^2)$, $Y \sim N(\mu_2, \sigma_2^2)$, and $Z \sim N(\mu_3, \sigma_3^2)$.
   a Let $A = X + Y$. What is the distribution of A?
   b Let $B = 5X + 2$. What is the distribution of B?
   c Let $C = aX - bY + c^2Z$, where a, b, and c are real-valued constants. What is the distribution (along with parameter values) for C? Show how you derived your answer.

Algorithmic Fairness

6. An artificial intelligence algorithm is being used to make a binary prediction (G for guess) for whether a person will repay a microloan. An important question to ask: is the algorithm fair with respect to a binary demographic (D for demographic)? To answer this question we are going to analyze the historical predictions of the algorithm and compare the predictions to the true outcome (T for truth) and the growing field of algorithmic fairness. Consider the following joint probability table from the history of the algorithm's predictions:

	D = 0		**D = 1**	
	G = 0	G = 1	G = 0	G = 1
T = 0	0.21	0.32	0.01	0.01
T = 1	0.07	0.28	0.02	0.08

D: is the demographic of an individual (binary)
G: is the prediction made by the algorithm (binary)
T: is the true result (binary)

Recall that cell in the joint table where $(A = i, C = j, Y = k)$ is the probability, $P(A = i, C = j, Y = k)$. For all questions, justify your answer. You may leave your answers with terms that could be input into a calculator.

a. What is $P(D = 1)$?
b. What is $P(G = 1|D = 1)$?
c. Fairness definition #1: Parity. An algorithm satisfies "parity" if the probability that the algorithm makes a *positive prediction* ($G = 1$) is the same regardless of the demographic variable. Does this algorithm satisfy parity?
d. Fairness definition #2: Calibration. An algorithm satisfies "calibration" if the probability that the algorithm is *correct* ($G = T$) is the same regardless of demographics. Does this algorithm satisfy calibration?
e. Fairness definition #3: Equality of odds. An algorithm satisfies "equality of odds" if the probability that the algorithm *predicts a positive outcome* ($G = 1$) is the same regardless of demographics *given* that the outcome will occur ($T = 1$). Does this algorithm satisfy equality of odds?

Try to do these before the midterm...

...and finish these after the midterm.

Longer Questions

7. Recall the example of zero sum games for teams with "ELO" scores S_1 and S_2. When a game is played between the two teams they each sample an ability (A_1 and A_2 respectively) from a normal distribution with mean equal to the team's ELO score and constant variance. The variance is different for different types of games. For this problem we will use the GO rating variance of $\sigma^2 = (2000/7)^2$. In lecture we talked about how to calculate the probability that a team wins via sampling. In this problem we will work out a closed form calculation.

   a What is the probability distribution for the difference between A_1 and A_2?

   b A team wins if their sampled ability is larger. Come up with a closed form expression for the probability that team one wins.

   c The best human GO player in the world is Ke Jie with an ELO score of 3670. Alpha GO Zero is a computer with an ELO score of 5200. How many independent games would they have to play before the expected number of games that Ke wins is ≥ 1?

8. Let X_i be the number of weekly visitors to a web site in week i, where $X_i \sim N(2200, 52900)$ for all i. Assume that all X_i are independent of each other.

   a. What is the probability that the total number of visitors to the web site in the next two weeks exceeds 5000?

   b. What is the probability that the weekly number of visitors exceeds 2000 in at least 2 of the next 3 weeks?

9. A robot is located at the center of a square world that is 10 kilometers on each side. A package is dropped off in the robot's world at a point (x, y) that is uniformly (continuously) distributed in the square. If the robot's starting location is designated to be $(0, 0)$ and the robot can only move up/down/left/right parallel to the sides of the square, the distance the robot must travel to get to the package at point (x, y) is $|x| + |y|$. Let D be the distance the robot travels to get to the package. Compute $E[D]$. *The distance calculation used in this problem is often called the "L1 Norm" and is a common metric for many problems.*

10. You roll 6 dice. How much more likely is a roll with: [one 1, one 2, one 3, one 4, one 5, one 6] than a roll with six 6s? Think of your dice roll as a multinomial.

Algorithmic Analysis

11. Consider the following function, which simulates repeatedly rolling a 6-sided die (where each integer value from 1 to 6 is equally likely to be "rolled") until a value ≥ 3 is "rolled".

```
def roll():
  total = 0;
  while (True):
    # equally likely to return 1,...,6
    roll = randomInteger(1, 6)
    total += roll
    # exit condition
    if (roll >= 3) break
  return total
```

    a. Let X be the value returned by the function roll(). What is E[X]?

    b. Let Y be the number of times that the die is "rolled" (i.e., the number of times that randomInteger(1, 6) is called) in the function roll(). What is E[Y]?

12. Our ability to fight contageous diseases depends on our ability to model them. One person is exposed to llama-flu. The method below returns the number of individuals who will get infected.

```
# Get number of people infected by one individual
def numInfected():
   # most people are immune to llama-flu
   immune = bernoulli(p = 0.99)
   if immune: return 0

   # people who are not immune, spread the disease far
   spread = 0

   # they make contact with k people (up to 100)
   k = binomial(n = 100, p = 0.25)
   for i in range(k)
     spread += numInfected():

   # total infections should include this individual
   return spread + 1
```

What is the expected return value of `numInfected`?

Titanic Probabilities

13. On April 15, 1912, the largest passenger liner ever made collided with an iceberg during her maiden voyage. When the Titanic sank it killed 1502 out of 2224 passengers and crew. This sensational tragedy shocked the international community and led to better safety regulations for ships. One of the reasons that the shipwreck resulted in such loss of life was that there were not enough lifeboats for the passengers and crew. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others. Write a program that reads the data file and finds the answers to the questions on the webpage: `cs109.stanford.edu/titanic.html`

Biometric Keystrokes

14. Did you know that computers can know who you are not, just by what you write, but also by how you write it? Coursera uses Biometric Keystroke signatures for plagiarism detection. If you can't write a sentence with the same statistical distribution of key press timings as in your previous work, they assume that it is not you who is sitting behind the computer. In this problem we provide you with three files:

    - personKeyTimingA.txt has keystroke timing information for a user A writing a passage. The first column is the time in milliseconds (***since the start of writing***) when the user hit each key. The second column is the key that the user hit.
    - personKeyTimingB.txt has keystroke timing information for a second user (user B) writing the same passage as the user A. Even though the content of the passage is the same *the timing* of how the second user wrote the passage is different.
    - email.txt has keystroke timing information for an unknown user. We would like to know if the author of the email was user A or user B.

    Let X and Y be random variables for the duration of time, in milliseconds, for users A and B (respectively) to type a key. Assume that each keystroke from a user has a duration that is an independent random variable with the same distribution.

    a. Estimate $E[X]$ and $E[Y]$
    b. Estimate $E[X^2]$ and $E[Y^2]$
    c. Use your answers to part (a) and (b) and approximate X and Y as Normals with mean and variance that match their biometric data. Report both distributions.
    d. Calculate the ratio of the probability that user A wrote the email over the probability that user B wrote the email. You don't need to submit code, but you should include the formula that you attempted to calculate and a few sentence description of how your code works.