Chris Piech
CS109

PSet #4
Oct 25, 2017

Problem Set #4
Due: 3:00pm on Wednesday, Nov 8th

For each problem, **explain/justify how you obtained your answer** in order to obtain full credit. 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, 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:
   $f_{X,Y}(x,y) = c\frac{y}{x}$ where $0 < y < x < 1$
   a. What is the value of c in order for $f_{X,Y}(x,y)$ to be a valid probability density function?
   b. Are X and Y independent? Explain why or why not.
   c. What is the marginal density function of X?
   d. What is the marginal density function of Y?
   e. What is E[X]?
   f. What is E[Y]? Hint: At some point, integration by parts may be your friend on this problem. You may use Wolfram Alpha or a similar integration tool.

3. 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.

Deeper Questions

4. 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 has an ELO score of 5200. How many independent games would they have to play before the expected number of games that Ke wins is greater than or equal to 1?

Errata: this problem was changed on Oct 26th. In the previous version σ^2 was 200

5. Let X_i = 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?

6. 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. $X \sim Uni(a, b)$
   b. $X \sim N(\mu, \sigma^2)$
   c. $X \sim Exp(\lambda)$

7. 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 is true distance plus Gaussian noise with mean 0 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.

8. Choose a number *X* at random from the set of numbers {1, 2, 3, 4, 5, 6}. Now choose a number at random from the subset no larger than *X*, that is from {1, …, *X*}. Let *Y* denote the second number chosen.

   a. Determine the joint probability mass function of *X* and *Y*.
   b. Determine the conditional mass function $P(X = j \mid Y = i)$ as a function of *i* and *j*.
   c. Are *X* and *Y* independent? Justify your answer.

Try and do these before the midterm

And finish these after the midterm

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 = 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. Say we have an array of *n* doubles, `arr[n]` (indexed from 0 to *n* – 1), which contains uniformly generated *non-negative* real values (where each value in the array is unique). What is the expected number of times that "max update" (as noted by the comment in the code) is executed in the function below (assuming the function is passed the array `arr` and its size `n`). Give an expression (not a big-Oh running time) for the expectation, and explain how you derived your answer.

```
double max(double arr[], int n) {
   double max = -1;        // note: all elements in arr[] are > -1.
   for(int i = 0; i < n; i++) {
      if (arr[i] > max) {
         max = arr[i];     // max update: (max = arr[i])
      }
   }
   return max;
}
```

11. Say we have a coin with unknown probability X of coming up heads when flipped. However, we believe (subjectively) that the prior probability (before seeing the results of any flips of the coin) of X is a Beta distribution, where E[X] = 0.5 and Var(X) = 1/36 ≈ 0.02778.

    a. What are the values of the parameters *a* and *b* (where *a*, *b* > 1) of the prior Beta distribution for X?

    b. Now say we flip the coin 13 times, obtaining 8 heads and 5 tails. What is the form (and parameters) of the posterior distribution of (X | 13 flips resulting in 8 heads and 5 tails)?

    c. What is E[X | 12 flips resulting in 8 heads and 4 tails]?

    d. What is Var(X | 12 flips resulting in 8 heads and 4 tails)?

Titanic Probabilities

12. 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.

13. 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.