

CS177, Homework 2

Due Date: Wednesday, April 16th

Reading

Recommended reading for this week covers a couple more discrete distributions of interest as well as marginal and conditional probability, independence and Bayes' rule

- Olofsson Chapter 1, Sections 1.5-1.6.1 (p.29-53) : Conditional probability, independence and examples
- Olofsson Chapter 2, Sections 2.5.3 : Geometric distribution
- Course web page : Links to discussion of power-laws and scale-free graphs. These are optional background but worth taking a look at.

Problems

Please document clearly how you arrived at your answer. Making sure you understand the step-by-step process is as important as getting the final number correct (perhaps more important).

Problem 1: Marginals and Conditionals

Suppose we are investigating the joint pmf for a collection of 3 discrete random variables: $X \in \{x_1, \dots, x_L\}$, $Y \in \{y_1, \dots, y_M\}$, and $Z \in \{z_1, \dots, z_N\}$.

1. In general, how many numbers are required to specify the joint probability mass function $p(x, y, z) = P(X = x, Y = y, Z = z)$?
2. Suppose we are given the table of values for $p(x_i, y_j, z_k)$. Write down an equation that specifies the marginal distribution, $p(z)$, as a function of this table.
3. Now suppose we want to calculate the conditional distribution $p(z|x)$. Describe how to calculate this in a systematic manner, starting from the joint pmf. Write out a final equation that specifies this conditional as a function of the table.

Problem 2: Joint distributions

Consider a pair of random variables $X \in \{0, 1\}$ and $Y \in \{-1, 0, 1\}$ which have the following joint probability distribution

		Y		
		-1	0	1
X	0	0.1	0.4	0.2
	1	0.0	0.2	0.1

1. What is $P(X = 1|Y = 1)$?

2. What is the probability that $Y \geq 0$ given that $X = 0$?
3. Compute $E[Y]$.
4. What is the expected value of $(X + 1)(Y + 1)$?
5. Are X and Y independent? Show clearly how you arrive at your answer.
6. Suppose we add another variable W which is *independent* of Y and has marginal probabilities $P(W = 0) = 0.7$ and $P(W = 1) = 0.3$. Write out the table of joint probabilities for $P(W, Y)$.

Problem 3: RAID Arrays

One solution for dealing with hard disk failures is to distribute multiple copies of the data over different disks. It is possible to configure a RAID (Redundant Array of Identical Disks) system by using N disks to store the data along with an addition K disks which store parity information about the contents of the data disks. In this case, up to K disks can fail at once without losing any data but if more than K fails then the whole system fails and the data is lost. Suppose that in a 1 year period the probability of an individual disk failing is p .

1. Suppose we have a 3 disk array with $N = 2, K = 1$. What is the expected number of disk failures in one year? What is the probability that the whole system will continue to function without any data loss after one year as a function of p ?
2. What is the expected number of disk failures for a 5 disk array with $N = 3, K = 2$? What is the probability that this array will not suffer any data loss after one year as a function of p ?
3. Suppose $p = 0.1$. Which is more reliable, the disk array with 1 parity disk or 2 parity disks? How about if $p = 0.6$?

Problem 4: Dice

A 6 sided die is tossed three times where each roll is independent.

1. What is the probability that the first two rolls sum up to a number greater than 6 given that the first roll is a 3?
2. What is the probability that the first roll is a 1 given that the total of the 3 rolls is 5?
3. The first two dice rolled sum up to an even number, what is the probability that both dice show the same number?

Problem 5: Correlations

If the occurrence of the event B makes event A more likely, does the occurrence of event A make B or likely? Justify your answer in terms of conditional probabilities.

Problem 6: Ahead by 5

Suppose two people are shooting baskets. In a given round, they each take 1 shot worth 1 point which player A makes with probability p and player B makes with probability q . The first player to score 5 points wins the game.

1. Write out an expression for the expected number of rounds before someone wins as a function of p and q .
2. What is the expected number of rounds if $p = 0.4$ and $q = 0.4$?
3. How about if $p = 0.1, q = 0.9$?

Problem 7: Empirical session length data

In this problem and the next you will analyze a real data set collected from the ICS Web server. The data is contained in a file called `sessionlengths`, available from the class Web page. It contains 2524 integers. Each number records the number of Web sessions recorded over a few months on the ICS server. The first number is the number of sessions of length 1, the 2nd the number of sessions of length 2, and the 2524th number the number of sessions of length 2524 (there is only one of this length—this was the longest recorded session). In Web data analysis a "session" is defined as a series of page-requests from the same IP address such that there is no gap between page-requests of longer than 20 minutes.

Write a MATLAB script called `analyze_session_data.m` to print out the answers to the questions below. A script is a MATLAB `.m` file that just contains a sequence of commands (i.e. no function definitions). Your script should follow the structure of the template on the website and compute and print the following:

1. longest session length
2. shortest session length
3. most common session length
4. total number of sessions
5. total number of pages requested
6. empirical probability of a session of length k for $k = 1, 2, 3, 4, 5, 6$. This should be an estimate of the true probability based on the empirical frequency, i.e. taking the number of sessions of a given length divided by the total number of sessions.
7. μ , the mean or expected session length based on the empirical data Please submit a printout of your MATLAB function as well as uploading it electronically to the folder on EEE.

Problem 8: Models of session lengths

We would like to figure out a good model for session lengths. Write a MATLAB script called `plot_session_models.m`. You will find the template on the website to be helpful.

1. One possible model is the geometric distribution. Plot the pmf for the geometric distribution with values $n = 1 \dots 50$. To do this, you will need to choose the parameter p for the geometric distribution. One reasonable idea is to set p so that the expected value of our model $E[X] = 1/p$ is the same as the empirical mean value you estimated in problem 7, that is set $p = 1/\mu$.
2. Calculate the probability values for a power-law pmf for values $n = 1 \dots 50$. The power law pmf with parameter γ is defined as

$$p(n) = \frac{1}{C} n^{-\gamma}, \quad n \in \{1, 2, \dots, \infty\}$$

where C is a normalization constant which guarantees that $\sum p(n) = 1$. In this case, finding a good value for γ is a bit more tricky. For this data, use $\gamma = 1.98$ and $C = 1.6641$.

3. Generate three graphs. Each will have three curves corresponding to the empirical probability (from Problem 7), the geometric pmf, and the power-law pmf, but the three graphs will have different axes.
 - On the first graph plot the probability values for n ranging from 1 up to 50
 - On the second graph, use the MATLAB function `semilogy` in order to plot the probability (y-axis) on a logarithmic scale
 - On the third graph, use `loglog` in order to plot the three distributions with both the x-axis and y-axis on a logarithmic scale

In the template for `plot_session_models.m` you will find some code to get you started on the first graph.

4. Summarize the differences between the three probability mass functions. Which of the two pmfs most closely match the empirical data? Why is the geometric pdf a straight line on the second plot and the power-law pmf a straight line on the third plot? You should be able to explain this in terms of the equations for these two pmfs.

Please submit a hardcopy of the 3 graphs, a printout of your MATLAB function, and written discussion with the rest of your homework. Also upload an electronic copy of your script to the folder on EEE.