

# CS 177, Homework 2

Applications of Probability in Computer Science: Winter 2007

Due Date: Tuesday January 23rd, in class

## Recommended Reading

- Pages 29 to 46 and 49 to 53 in Chapter 1.
- Section 2.5.3 on the geometric distribution in Chapter 2.
- The Web pages on the Power-Law and scale-free graphs (you can find these on the course Web page). These articles are optional background reading—they illustrate how power-laws and related ideas occur in a variety of different areas.

In all of the problems below, **clearly** show how you derived your solution, i.e., do not just state the answer but show how you arrived at the answer, step by step.

## Problem 1

Answer each of the following questions where  $A, B, C$  are each discrete-valued random variables. You can use lower-case values  $a, b, c$  to denote values of  $A, B, C$ .

1. if  $A, B, C$  can each take  $K$  possible values, how many numbers in total are required to specify the joint probability mass function  $P(a, b, c)$ ?
2. Say we know the joint pmf  $P(a, b, c)$  in the form of a 3-dimensional table or array of numbers. Write down an equation that precisely defines how we would calculate  $P(c)$  using this table.
3. Same as the last question, but now show (using multiple steps, and making it clear how each step is carried out) how we would calculate  $P(c|a)$ , starting from the joint pmf.

## Problem 2

Consider 2 random variables  $X$  (taking values 1, 2, or 3) and  $Y$  (taking values 0 or 1), with the following joint probability table:

		X		
		1	2	3
Y	0	0.1	0.4	0.2
	1	0.0	0.2	0.1

Answer the following questions.

1. What is the probability of  $P(X = 1|Y = 1)$ ?
2. What is the probability that  $X > 1$  given that  $Y = 0$ .
3. What is the value of  $E[X]$ ?
4. What is the value  $E[(X + 1)(Y + 2)]$ ?
5. Are  $X$  and  $Y$  independent? Clearly show how you arrive at your answer.
6. Now say we have a third variable  $Z$ . Say we know that  $Z$  is independent of  $X$  and that  $Z$  has probability  $P(Z = 0) = 0.4$  and  $P(Z = 1) = 0.6$ . Write out the joint probability table for  $P(X, Z)$ .

### Problem 3

A system has  $m$  components. Each component has a probability  $p$  of failing (assume this is over some fixed time-period such as a month). Components fail independently of each other. The system as a whole operates correctly if at least half of its components operate correctly (i.e., if at least half the components do not fail).

1. What is the probability that a system with 3 components will operate correctly?
2. What is the probability that a system with 4 components will operate correctly?
3. Which is more reliable, a system with 3 components or a system with 4 components, when  $p = 0.1$ ?

### Problem 4

If the occurrence of the event  $B$  makes the event  $A$  more likely, does the occurrence of  $A$  make  $B$  more likely? Justify your answer (an answer of just “yes” or “no” will not get many points).

### Problem 5

A die is cast tossed twice, with independent tosses, so we get 2 independently-generated numbers, each in the range 1 to 6. Calculate the conditional probability of each of the following events:

- that the first number is a 1 given that the total of the 2 numbers is 4;
- that the total of the two numbers is larger than 6, given that the first number is a 4;
- that the first number is a 6 given that the total of the two numbers is 8.

### Problem 6

Suppose 2 teams are playing a series of games, each of which is independently won by team A with probability  $p$  and by team B with probability  $1 - p$ . The winner is the first team to win 4 games. Find the expected number of games that are played as a function of  $p$ . What is this expected number of games if  $p = 0.5$ ? What is the value if  $p = 0.01$ ? if  $p = 0.99$ ?

### Problem 7

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 typically 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` to print out the answers to the questions below. A script is a MATLAB `.m` file that does not take any arguments. Your script should have the following structure:

```
% MATLAB script to analyze properties of Internet
% session length data.
%
%                               cs177, january 2007

load sessionlengths;

% compute the max length of any session
max_session_length = length(sessionlengths);
```

...your code goes here to solve the rest of this problem....

Compute and print (to the screen) the following values:

1. the longest session length
2. the shortest session length
3. the most common session length
4. the total number of sessions
5. the total number of page-requests
6. the *empirical probability* of session length  $k$ , where  $k = 1, 2, 3, \dots, 9, 10$ . You can use the “frequency-based” method of estimating these empirical probabilities, e.g., the probability of a session of length 4 is the number of sessions of length 4 divided by the total number of sessions. The probabilities are called *empirical* because they come from data and not from a model.
7. the mean or expected session length.

In your hardcopy you should submit the values that your script computes, plus a copy of the script you wrote. You should also submit an electronic copy of your script to EEE under the Homework 2 folder.

## Problem 8

Write a single MATLAB script called `plot_session_data` to do the problem below. Your script should have the following structure:

```
% MATLAB script to fit probability models to Internet
% session length data.
%
%                               cs177, january 2007

load sessionlengths;

% calculate the parameter p for the geometric distribution
... your code here....

% define the values for gamma and the normalization constant
% for the power-law distribution
```

```

....your code here....

kvalues = 1:50;
% note below we use a for-loop over each value of k to define
% the pmfs - you could also do this by "vectorization" if you
% prefer, it will be faster in general
for k=1:length(kvalues)
    pmf_empirical(k) = ....
    pmf_geometric(k) = ....
    pmf_power_law(k) = ....
end

% plot the 3 graphs requested
.... your code here.....

```

1. calculate the probability values for a geometric pmf for values 1 to 50. For the parameter  $p$  use a best-fit to the session-length data in the previous problem. Do this by estimating the mean session length  $E[X]$  (as you did in the last problem), then setting  $p = 1/E[X]$  (you should try to figure out why this is a good idea), and then calculating the geometric pmf with this value for  $p$ .
2. now calculate the probability values for a power-law pmf for values 1 to 50. The power-law pmf is defined as:

$$P(k) = \frac{1}{C} k^{-\gamma}, \quad k = 1, 2, \dots, \infty$$

where  $C$  is a normalization constant and  $\gamma$  is a parameter of the model. For the parameter  $\gamma$  you again use a best-fit to the session-length data in the previous problem, i.e., we search for the value of  $\gamma$  that produces a pmf that is closest to the empirical data. This involves ideas from statistics that are beyond the scope of this class so you can just use the result that the best-fit  $\gamma$  for this data set turns out to be  $\gamma = 1.98$  (and  $C = 1.6641$ ).

3. Generate 3 graphs, and on each graph show 3 curves, one each for the empirical probabilities, and the geometric pmf, the power-law pmf.
  - On the first graph plot the probability values (for each of the 3 types of probabilities) for  $k = 1, 2, \dots, 50$ .
  - On the second graph, put the probability (y-axis) on a logarithmic scale versus  $k$  (which is still on a standard linear scale). You can use the MATLAB command `semilogy`.

- On the third graph, put both the x-axis and y-axis on a logarithmic scale: you can use the MATLAB command `loglog`.

To help you to generate these plots, below is some MATLAB code to produce the first of the graphs above:

```
figure; n = 50; plot(kvalues(1:n),p(1:n),'.'); hold on;
plot(kvalues(1:n),pmf_geometric(1:n),'r:');
plot(kvalues(1:n),pmf_power_law(1:n),'g-');
legend('empirical','geometric','power-law');
xlabel('k'); ylabel('PROBABILITY, P(k)');
title('PROBABILITY MASS FUNCTIONS FOR INTERNET SESSION LENGTH DATA');
```

4. Briefly explain the differences between the 3 probability mass functions. Which of the two pmfs fit the empirical data the best? Explain why on the 2nd plot the geometric pmf is a straight line and on the 3rd plot the power-law pmf is a straight line (you should be able to define 2 equations for your explanation).

In the hardcopy version of your homework submit the 3 graphs requested, the explanation for the last part of the problem, and a printed copy of your script function. Please also submit an electronic copy of your script to the Homework 2 folder in EEE.