CS 177, Homework 6

Applications of Probability in Computer Science: Winter 2007

Due Date: Thursday March 1st

Suggested Reading

- Section 2.3, 2.4, 2.7, 4.3 (pages 274–279).

Note: You should hand in the homework below as hardcopy in class. In addition you should submit the MATLAB code for problem 3 electronically to EEE.

Problem 1

In this problem you will check how well the theoretical results we derived in class for the "gambler's ruin" problem match simulated results. You can modify the Markov chain simulation code you wrote in Homework 5 for this problem.

Specifically, there are two players A and B that play a sequence of games with independent outcomes where the winner of each game pays the other a dollar. A starts with i dollars, and B with $n-i$, so there are n dollars in total. The series ends when either A wins all n dollars or B wins all n dollars. The probability of A winning a single game is p, and the probability of B winning a single game is $q = 1 - p$.

Modify your MATLAB code to simulate the following cases:

1. $p = q = 0.5, n = 20, i = 4$.
2. $p = q = 0.5, n = 20, i = 10$.
3. $p = 0.55, q = 0.45, n = 20, i = 4$.
4. $p = 0.55, q = 0.45, n = 20, i = 10$.

For each of the 4 cases above, simulate 10,000 series and record how often out of the 10,000 that A wins the series. Compare this in each case to the theoretical answer and write some brief comments.

Using the same set of 10,000 series that you already simulated for each case, compute both (a) the average length of a series, and (b) the distribution of the lengths of the series and report this distribution as a histogram (you can use the function `hist.m`). Write a few comments on how you think the average length of a series depends on the parameters, i.e., on p, q, i, and n.

You do not need to hand in the MATLAB code or scripts that you write for this problem.

Problem 2

Let $f(x)$ be the probability density function for a real-valued variable X. We define the variance of X with respect to $f(x)$ as

$$Var(X) = E[(X - E[X])^2] = \int_{-\infty}^{\infty} (x - E[X])^2 f(x) dx.$$

This is a special case of the general rule that $E[g(X)] = \int_{-\infty}^{\infty} g(x) f(x) dx$, where here $g(X) = (X - E[X])^2$.

Prove that in general $Var(aX + b) = a^2 Var(X)$ where a and b are constants..

Problem 3

Let X have a Gaussian (Normal) density function $f(x)$ with parameters μ and σ. Write a MATLAB function called `gauss_plot.m` (using the template below) that plots $f(x)$ and cdf $F(x)$ over a range of x values for given values of μ and σ.

```
function gauss_plot(mu,sigma,x)
%
% function to plot a Gaussian/Normal pdf and cdf
%
%  INPUTS:
%     mu: mean parameter for the Gaussian pdf (a scalar)
%   sigma: standard deviation parameter for the Gaussian pdf (a scalar)
%     x: an n x 1 vector of values where the Gaussian pdf and cdf is to be evaluated
%        at, e.g., x = 0.0,0.1,0.2,.........19.8, 19.9, 20
%        (can be created in MATLAB by x = 0.0:0.1:20)
%
%                                               CS 177 Homework 6, 2007

% code follows below......
```

Turn in the following (the code should be turned in electronically via EEE to the Homework 6 folder, and the plots and numbers should be turned in with the hardcopy of your homework in class).

1. A copy of your documented MATLAB code for `gauss_plot.m`.

2. Graphs of $f(x)$ and $F(x)$ for (a) $\mu = 100, \sigma = 10$, (b) $\mu = 100, \sigma = 30$, with x ranging from 50 to 150 (plotting (say) 1000 values so that you get a smooth curve). You should hand in 4 separate graphs.

3. On each graph for $f(x)$, manually shade in the region of the graph that corresponds to the probability that x lies between 110 and 120.

4. For each of the two Gaussian density functions, calculate $P(110 \leq X \leq 120)$. To do this you can either numerically approximate the appropriate integral, or you can look up Normal probability tables for the cumulative Normal distribution (in Table A.1 in the text). Explain in detail how you calculated your answer.

Problem 4

A startup company wants to offer a free online photo-storage service to the public (the company will make money by selling advertising). In the first year of operation it plans to allow 1 million customers to sign up for its service. Assume that past studies of online photo usage has shown that the amount of disk space X_i that a randomly selected customer i uses in one year obeys a distribution with a mean of $\mu = 500$ Mbytes and a standard deviation $\sigma = 50$ Mbytes. Given this information, calculate the minimum amount of disk space the company should purchase for their photo-storage service in the first year of operation if the company wants to be 99.9% sure that it will be able handle all of the space requests by its 1 million customers.

Problem 5

In this problem you will verify the Central Limit Theorem via simulation. You do not need to hand in the MATLAB code or scripts that you write for this problem.

1. Generate 100,000 random numbers between 0 and 1 using the `rand` random number generator in MATLAB. Note that you may wish to call `rand` in vector mode (as discussed in class) to make this fast rather than looping over the function and calling it 100,000 times (see `help rand` in MATLAB for more information). Then generate a graph of a histogram between 0 and 1 showing how many of the random numbers fall into cells that are each of size 0.01 (i.e., a histogram between 0 and 1 with bins of size 0.01). Use the `hist.m` function to do this. Comment briefly on what you see in this histogram plot.

2. Let S_n be the sum of n random numbers, where each random number comes from a uniform distribution between 0 and 1. Generate n such numbers for each of the cases below. For example, in the first case you will have $K = 10$ different S_n values and each of the S_n numbers will itself be a sum of $n = 100,000$ random numbers.

(a) $K = 10$ and $n = 100,000$

(b) $K = 100,000$ and $n = 10$.

(c) $K = 100$ and $n = 100$

(d) $K = 10,000$ and $n = 1000$,

In each case, do the following:

(a) Calculate the theoretical mean (via the Central Limit Theorem) of the K "random sums" and also the "empirical mean" (which is the average of the sums that you generated, e.g., in the last case the the average of the $K = 100000$ sums). Comment on the differences between the two in each case.

(b) Plot a histogram of the values of the K sums for each case with about 10 to 30 equal-sized bins covering the data and centered around the "empirical mean". (So you will have 4 histograms). Use 5 bins for $K = 10$, 10 bins for $K = 100$, and use 30 bins for $K = 10000$ and $K = 100000$.

(c) Comment on the shape of the histograms. For example, do the histograms look Normal (Gaussian)? If not, why not? Comment on how these results relate to the theoretical result he Central Limit Theorem as discussed in class.