Assignment 2: CS 215

Due: 2nd September before 11:55 pm, 100 points

All members of the group should work on all parts of the assignment. Copying across groups or from other sources is not allowed. We will adopt a zero-tolerance policy against any violation.

Submission instructions:

1. You should type out a report containing all the answers to the written problems in Word (with the equation editor) or using Latex, or write it neatly on paper and scan it. In either case, prepare a single pdf file.
2. The report should contain names and roll numbers of all group members on the first page as a header.
3. Put the pdf file and the code for the programming parts all in one zip file. The pdf should contain the names and ID numbers of all students in the group within the header. The pdf file should also contain instructions for running your code. Name the zip file as follows: A2-IdNumberOfFirstStudent-IdNumberOfSecondStudent-IdNumberofThirdStudent.zip. (If you are doing the assignment alone, the name of the zip file is A2-IdNumber.zip, if there are two students it should be A2-IdNumberOfFirstStudent-IdNumberOfSecondStudent.zip).
4. Upload the file on moodle BEFORE 11:55 pm on the due date. We will nevertheless allow and not penalize any submission until 10:00 am on the following day (i.e. 3rd September). No assignments will be accepted thereafter, and your group will get no scores for this assignment.
5. Note that only one student per group should upload their work on moodle, though all group members will receive grades.
6. Please preserve a copy of all your work until the end of the semester.

Questions: *In the following problems, you can use basic functions from MATLAB.*

1. Consider you are testing samples (e.g., blood, urine, saliva, etc. During the COVID19 pandemic, the most common method for testing was RT-PCR on nasal or throat mucus.) of n different subjects for a rare disease. Testing each subject independently requires n tests which can be expensive and time-consuming (and unjustifiable since most tests would report negative, i.e. healthy or not diseased). Pooled testing is a technique where you take small portions of the sample of a subset of people and mix them together to create a pool. This pool is tested instead of testing the individual samples. If the pool tests **negative** (i.e. not diseased), then you can declare all subjects participating in that pool to be healthy – assuming test results are accurate. Overall, this greatly saves on the number of tests, if the disease is rare. If the pool tests **positive** (i.e. one or more participating subjects is/are diseased), then more work needs to be done, as we shall see below. Typically, the number of pools is taken to be less than n to save on testing time and resources. We will use our knowledge of probability and statistics to analyze some pooled testing methods.
   (a) In a popular technique of group testing, we divide the n subjects into n/s disjoint subsets each containing s subjects. For simplicity, assume that n is divisible by s. In round 1, the n/s pools are tested. Participating subjects in a negative pool are declared healthy. Subjects that participated in a positive pool are tested individually in round 2. Let the prevalence rate of the disease (the proportion of people who have the disease) be p where $p \in [0, 1]$. Now do as directed. Assume throughout that the test results are accurate.

i. Prove that the expected total number of tests for this method (counting both round 1 and 2) is $T(s) := n/s + n(1-(1-p)^s)$.

ii. If p is very small, this equals $n/s + ps$. For what value of s, is this expected number the least? What is the expected number of tests in that case?

iii. What is the maximum value of p for which $T(s) < n$? For this last part, you may need to do some simple numerical computation in MATLAB. [3+(2.5+2.5)+4=12 points]

(b) In another method, round 1 involves testing of T_1 pools. A subject participates in a pool with probability π independent of the other items and other pools (so the pools may be overlapping in this case). Similar to the earlier method, all subjects that participated in a positive pool are individually tested in round 2. Let the prevalence rate of the disease be p. Do as directed. Throughout this exercise, assume that all tests are accurate. Also use the identity $(1-x)^a = e^{-xa}$ for small $x > 0$.

i. What is the probability that a genuinely healthy subject participates in a pool that tests negative?

ii. For what value of π (in terms of n and p) is this probability maximized? (Note that we would like this probability to be as high as possible.)

iii. Given this optimal π, what is the probability that all pools that a genuinely healthy subject participates in, test positive?

iv. Then what is the expected total number of tests (counting round 1 and round 2)? [Hint: In round 2, what is the expected number of tests of genuinely positive subjects? In round 2, what is the expected number of tests of genuinely negative subjects who happened to participate in all positive pools?]

v. For what value of T_1 is this expected number minimized? What is the expected number of tests in that case? [2 x 5 = 10 points]

(c) For $n = 1000$ and $p \in \{10^{-4}, 5\times10^{-4}, 0.001, 0.005, 0.01, 0.02, 0.05, 0.08, 0.1, 0.2\}$, plot a graph of the expected number of tests for both methods versus n. For the first method, the expected number should be for optimal s and for the second method, it should be for optimal π and optimal T_1. Include this plot in your report and comment on it. [2 points]

2. Let X and Y be two independent random variables with PDFs $f_X(.)$ and $f_Y(.)$ respectively. Derive the PDF of the random variable $Z = XY$ in terms of f_X, f_Y. (Hint: Look at the discussion forum on moodle where the PDF of $X+Y$ was derived. Follow similar steps.) [16 points]

3. Consider a random variable X with PDF $f_X(.)$. Consider independent samples $\{x_i\}_{i=1}^n$ from this PDF. A student wants to estimate $E(X)$ given these samples. Let the estimated value be $\hat{x}$. Should (s)he consider $\hat{x} := \sum_{i=1}^n x_i/n$ or $\hat{x} := \sum_{i=1}^n f_X(x_i)x_i/n$ as the estimate for $E(X)$? Which one of these is correct and why? To answer the why part, you explain how the other option is incorrect (eg: maybe it is estimating the expectation of some other random variable). [10 points]

4. Read in the images T1.jpg and T2.jpg from the homework folder using the MATLAB function `imread` and cast them as a double array using the code

```
im = double(imread('T1.jpg');
```

These are magnetic resonance images of a portion of the human brain, acquired with different settings of the MRI machine. They both represent the same anatomical structures and are perfectly aligned (i.e. any pixel at location (x, y) in both images represents the exact same physical entity). Consider random variables I_1, I_2 which denote the pixel intensities from the two images respectively. Write a piece of MATLAB code to shift the second image along the X direction by t_x pixels where t_x is an integer ranging from -10 to +10. While doing so, assign a value of 0 to unoccupied pixels. For each shift, compute the following measures of dependence between the first image and the shifted version of the second image:

- the correlation coefficient ρ,
- a measure of dependence called quadratic mutual information (QMI) defined as $\sum_{i_1}\sum_{i_2}(p_{I_1I_2}(i_1,i_2) - p_{I_1}(i_1)p_{I_2}(i_2))^2$, where $p_{I_1I_2}(i_1,i_2)$ represents the normalized joint histogram (*i.e.*, joint pmf) of I_1 and I_2 ('normalized' means that the entries sum up to one),

- a measure of dependence called the mutual information (MI) defined as $E\left(\log\left(\frac{p_{I_1,I_2}}{p_{I_1}p_{I_2}}\right)\right)$ where the expectation is taken jointly over the values of I_1 and I_2.

For computing the joint histogram, use a bin-width of 10 in both I_1 and I_2. The joint histogram should be computed using only those pixels which are valid for both I_1 and I_2 (thus, the 'zeroed-out pixels' do not count). For computing the marginal histograms, you need to integrate the joint histogram along one of the two directions respectively. You should write your own joint histogram routine in MATLAB - do not use any inbuilt functions for it. Plot separate graphs of the values of ρ versus t_x, QMI versus t_x and MI versus t_x.
Repeat exactly the same steps when (i) the second image is a negative of the first image, i.e. $I_2 = 255 - I_1$, and (ii) the second image is obtained as $I_2 = 255 \times (I_1)^2/\max((I_1)^2) + 1$.
Comment on all the plots. In particular, what do you observe regarding the relationship between the dependence measures and the alignment between the two images? Your report should contain all plots labeled properly, and the comments on them as mentioned before. [25 points]

5. Given stuff you've learned in class, prove the following bounds: $P(X \geq x) \leq e^{-tx}\phi_X(t)$ for $t > 0$, and $P(X \leq x) \leq e^{-tx}\phi_X(t)$ for $t < 0$. Here $\phi_X(t)$ represents the MGF of random variable X for parameter t.
Now consider that X denotes the sum of n independent Bernoulli random variables $X_1, X_2, ..., X_n$ where $E(X_i) = p_i$. Let $\mu = \sum_{i=1}^{n} p_i$. Then show that $P(X > (1+\delta)\mu) \leq \frac{e^{\mu(e^t-1)}}{e^{(1+\delta)t\mu}}$ for any $t \geq 0, \delta > 0$. You may use the inequality $1 + x \leq e^x$. Further show how to tighten this bound by choosing an optimal value of t. [15 points]

6. Consider that there are n independent coin tosses with heads probability p. Let T be the trial number at which you get the first heads. Derive a formula for $E(T)$ in terms of p and/or n. [10 points]