Assignment 2: CS 215

Due: 22nd August before 11:55 pm

Remember the honor code while submitting this (and every other) assignment. All members of the group should work on all parts of the assignment. We will adopt a zero-tolerance policy against any violation.

Submission instructions:

1. You should type out 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 pdf file.
2. 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.zip. (If you are doing the assignment alone, the name of the zip file is A2-IdNumber.zip).
3. Upload the file on moodle BEFORE 11:55 pm on the due date (i.e. 22nd August). We will nevertheless allow and not penalize any submission until 6:00 am on the following day (i.e. 23rd August). No assignments will be accepted thereafter.
4. Note that only one student per group should upload their work on moodle.
5. Please preserve a copy of all your work until the end of the semester.

Questions:

1. Given random variables X and Y having probability density functions $f_X(x)$ and $f_Y(y)$ respectively and joint probability density function $f_{XY}(x, y)$, derive an expression for the probability density function of the random variable $Z = X + Y$ (i.e. the sum of X and Y) in terms of $f_X(x)$, $f_Y(y)$ and $f_{XY}(x, y)$. Also derive an expression for $P(X \leq Y)$. Refine the expressions if X and Y are independent. [3+3+2+2=10 points]
2. Let $X_1, X_2, ..., X_n$ be $n > 0$ independent identically distributed random variables with cdf $F_X(x)$ and pdf $f_X(x) = F'_X(x)$. Derive an expression for the cdf and pdf of $Y_1 = \max(X_1, X_2, ..., X_n)$ and $Y_2 = \min(X_1, X_2, ..., X_n)$ in terms of $F_X(x)$. [15 points]
3. Using Markov's inequality, prove the following one-sided version of Chebyshev's inequality for random variable X with mean μ and variance σ^2: $P(X - \mu \geq \tau) \leq \frac{\sigma^2}{\sigma^2 + \tau^2}$ if $\tau > 0$, and $P(X - \mu \geq \tau) \geq 1 - \frac{\sigma^2}{\sigma^2 + \tau^2}$ if $\tau < 0$. [15 points]
4. 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]

5. Basic statistics has interesting applications in what is called as 'group testing'. This is a problem to give you a taste for it. Consider that k people are to undergo a blood test for a certain disease. Suppose that each person independently of others has a probability $p \in (0, 1)$ of having the disease. One method is to test each of the k people separately, but this is time consuming as k tests are needed. Another method is to mix the blood samples of all k people and perform the test on the mixed sample. If the test on the mixture is negative, then clearly nobody has the disease, and in this case the total number of tests is 1. If the test is positive, the k people have to be tested separately, in which case the total number of tests is $k + 1$. Derive for what values of p (in terms of k), the second method has a smaller *expected* number of tests as compared to the first method. Given the formula, plot a graph of this expected number versus $k \in [2, 25]$ *for any two values of p for which the second test has a smaller expected number of tests.* Comment on the graph. In your report, show your derivation, include the plot and comments on it. [15 points]

6. 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. 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 $\sum_{i_1} \sum_{i_2} (p_{I_1 I_2}(i_1, i_2) - p_{I_1}(i_1) p_{I_2}(i_2))^2$, where $p_{I_1 I_2}(i_1, i_2)$ represents the *normalized* joint histogram (*i.e.*, joint pmf) of I_1 and I_2 ('normalized' means that the entries sum up to one).

   For computing the joint histogram, use a bin-width of 10 in both I_1 and I_2. For computing the marginal histogram, 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 a graph of the values of ρ versus t_x, and another graph of the values of QMI versus t_x.

   Repeat exactly the same steps when the second image is a negative of the first image, i.e. $I_2 = 255 - I_1$.

   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 four plots labelled properly, and the comments on them as mentioned before. [30 points]