Assignment 1: CS 215

Due: 16th August 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. 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: A1-IdNumberOfFirstStudent-IdNumberOfSecondStudent.zip. (If you are doing the assignment alone, the name of the zip file is A1-IdNumber.zip).
3. Upload the file on moodle BEFORE 11:55 pm on the due date (i.e. 16th August). We will nevertheless allow and not penalize any submission until 6:00 am on the following day (i.e. 17th 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 n distinct values $\{x_i\}_{i=1}^n$ with mean μ and standard deviation σ, prove that for all i, we have $|x_i - \mu| \leq \sigma\sqrt{n-1}$. How does this inequality compare with Chebyshev's inequality as n increases? (give an informal answer) [7+3=10 points]

   Solution: By definition, we have $\sum_{i=1}^n (x_i - \mu)^2 = \sigma^2(n-1)$ and hence for any i, we have $(x_i - \mu)^2 \leq \sigma^2(n-1)$, i.e. $x_i - \mu \leq \sigma\sqrt{n-1}$.
   Comparison: Chebyshev's inequality provides a bound which is satisfied by some fraction of all the sample points in the dataset, not all of them, whereas this inequality is for all points. However, this inequality is much looser than Chebyshev's inequality, especially for large n. Note that Chebyshev's inequality is unaffected by the number of points. For small n, the bound in this question is better than Chebyshev's bound. For example, when $n = 10$, this inequality tell us that all values lie within $\pm 3\sigma$ of μ, whereas Chebyshev's inequality states that 99.4% of the values lie within $\pm 13\sigma$ of μ. When $n = 100$, this inequality tell us that all values lie within $\pm 9.95\sigma$ of μ, whereas Chebyshev's inequality states that 99.4% of the values lie within $\pm 13\sigma$ of μ. When $n = 1000$, this inequality tell us that all values lie within $\pm 31.6\sigma$ of μ, whereas Chebyshev's inequality states that 99.4% of the values lie within $\pm 13\sigma$ of μ. As n increases, you see Chebyshev's bound becoming better.
   Marking scheme: 7 points for the proof, 3 points for a reasonable comparison. A statement that 'Chebyshev's inequality provides a bound which is satisfied by some fraction of all the sample points in the dataset, not all of them, whereas this inequality is for all points.' is enough for 3 points.

2. Given n values $\{x_i\}_{i=1}^n$ having mean μ, median τ and standard deviation σ, prove that $|\mu - \tau| \leq \sigma$. Assume n is even. [10 points]

Solution: By the one-sided Chebyshev inequality, we have $P(X - \mu > \sigma) \leq 1/2$ and likewise $P(X - \mu < -\sigma) \leq 1/2$. Now, the median τ divides the dataset into two halves, one containing values less than equal to the median, and the other containing values greater than the median. The probability that X is greater than $\mu + \sigma$ is less than 0.5. Since the median is the 50 percentile, it cannot be greater than $\mu + \sigma$, and hence must be less than or equal to $\mu + \sigma$. Likewise the probability that X is less than $\mu - \sigma$ is less than 0.5. Again by similar logic, the median cannot be less than $\mu - \sigma$ and hence be greater than or equal to $\mu - \sigma$. Hence we must have $\tau \geq \mu - \sigma$ and $\tau \leq \mu + \sigma$. Hence $|\mu - \tau| \leq \sigma$.

3. In a certain town, there exist 100 rickshaws out of which 1 is red and 99 are blue. A person XYZ observes a serious accident caused by a rickshaw at night and remembers that the rickshaw was red in color. Hence, the police arrest the driver of the red rickshaw. The driver pleads innocence. Now, a lawyer decides to defend the hapless rickshaw driver in court. The lawyer ropes in an opthalmologist to test XYZ's ability to differentiate between the colors red and blue, under illumination conditions similar to those that existed that fateful night. The opthalmologist suggests that XYZ sees red objects as red 99% of the time and blue objects as red 2% of the time. What will be the main argument of the defense lawyer? (In other words, what is the probability that the rickshaw was really a red one, when XYZ observed it to be red?) [10 points]

   Solution: Let R_R, R_B be the events that the rickshaw was red, blue respectively. Let X_R, X_B be the events that XYZ perceived a rickshaw to be red, blue respectively. We have $P(X_R|R_R) = 0.99, R(X_R|R_B) = 0.02, P(R_R) = 0.01, P(R_B) = 0.99$. We need to evaluate $P(R_R|X_R) = P(X_R|R_R)P(R_R)/P(X_R)$. $P(X_R) = P(X_R|R_R)P(R_R) + P(X_R|R_B)P(R_B) = 0.99 \times 0.01 + 0.02 \times 0.99 = 0.99 \times 0.03$. Hence $P(R_R|X_R) = \dfrac{0.99 \times 0.01}{0.99 \times 0.03} = 1/3$. In other words, the probability that the rickshaw was red when XYZ observed it to be red is only 1/3. In other words, it is more probable that the rickshaw was a blue one, based on the available data!

4. A contestant is on a game show and is allowed to choose between three doors. Behind one of them lies a car, behind the other two there lies a stone. The contestant will be given whatever is behind the door that (s)he picked, and quite naturally (s)he wants the car. Suppose (s)he chooses the first door, and the host of the show who knows what is behind every door, opens (say) the third door, behind which there lies a stone (without opening the first door). The host now asks the contestant whether (s)he wishes to choose the second door instead of the first one. Your task is to determine whether switching the contestant's choice is going to increase his/her chance of winning the car. Remember that the host is intelligent: (s)he is always going to open a door not chosen by the contestant, and is also going to open a door behind which there is a stone. You should approach this problem only from the point of view of conditional probability as follows. To this end, let C_1, C_2, C_3 be events that the car is behind doors 1,2,3 respectively. Assume $P(C_i) = 1/3, i \in \{1, 2, 3\}$.

   (a) Let Z_1 be the event that the contestant chose door 1. Write down the value of $P(C_i|Z_1)$ for all $i \in \{1, 2, 3\}$.

   Solution: It is 1/3 as each event C_i is independent of Z_1.

   (b) Let H_3 be the event that the host opened door 3. Write down the value of $P(H_3|C_i, Z_1)$ for all $i \in \{1, 2, 3\}$.

   Solution: We have $P(H_3|C_1, Z_1) = 0.5$ as the host will open either door 2 or 3 (with equal probability) if the car is behind door 1, and the contestant chose door 1. Also we have $P(H_3|C_2, Z_1) = 1$ and $P(H_3|C_3, Z_1) = 0$ for the same reason.

   (c) Clearly the conditional probability of winning by switching is $P(C_2|H3, Z1)$. This is equal to $\dfrac{P(H_3|C_2, Z_1)P(C2, Z1)}{P(H_3, Z_1)}$. Evaluate this probability. Note that $P(A_1, A_2)$ denotes the joint probability of events A_1, A_2.

   Solution: $P(C_2|H_3, Z1) = \dfrac{P(H_3|C_2, Z_1)P(C_2, Z_1)}{P(H_3, Z_1)}$

$$= \frac{P(H_3|C_2, Z_1)P(C_2, Z1)}{P(H_3|C_1, Z_1)P(C_1, Z_1) + P(H_3|C_2, Z_1)P(C_2, Z_1) + P(H_3|C_3, Z_1)P(C_3, Z_1)}$$

$= \frac{P(H_3|C_2, Z_1)}{P(H_3|C_1, Z_1) + P(H_3|C_2, Z_1) + P(H_3|C_3, Z_1)}$ since $P(C_1, Z_1) = P(C_2, Z_1) = P(C_3, Z_1)$ as C_i and Z_1 are independent

$= \frac{1}{0.5 + 1 + 0} = \frac{2}{3}$.

(d) Likewise evaluate $P(C_1|H3, Z1)$.

Solution: $P(C_1|H_3, Z1) = \frac{P(H_3|C_1, Z_1)P(C_1, Z_1)}{P(H_3, Z_1)}$

$$= \frac{P(H_3|C_1, Z_1)P(C_1, Z_1)}{P(H_3|C_1, Z_1)P(C_1, Z_1) + P(H_3|C_2, Z_1)P(C_2, Z_1) + P(H_3|C_3, Z_1)P(C_3, Z_1)}$$

$= \frac{P(H_3|C_1, Z_1)}{P(H_3|C_1, Z_1) + P(H_3|C_2, Z_1) + P(H_3|C_3, Z_1)}$ since $P(C_1, Z_1) = P(C_2, Z_1) = P(C_3, Z_1)$ as C_i and Z_1 are independent

$= \frac{0.5}{0.5 + 1 + 0} = \frac{1}{3}$.

(e) Conclude whether switching is indeed beneficial.

Solution: Clearly switching choices is the better thing to do. This is indeed surprising, but you should bear in mind that these calculations explicitly account for the fact that the contestant was intelligent. Make sure you are convinced about this - see the problem statement again.

(f) Now let us suppose that the host were quite whimsical and decided to open one of the two doors not chosen by the contestant, with equal probability, not caring whether there was a car behind the door. In this case, repeat your calculations and determine whether or not it is beneficial for the contestant to switch choices. [2+2+5+5+1+5=20 points]

Solution: In this case, we have to simply re-evaluate certain probabilities, as the host just picks one of the two doors not chosen by the contestant with equal probability. Basically, we now have $P(H_3|C_1, Z_1) = 0.5 = P(H_2|C_1, Z_1)$. We also have $P(H_3|C_2, Z_1) = P(H_2|C_2, Z_1) = 0.5$, and $P(H_3|C_3, Z_1) = P(H_2|C_3, Z_1) = 0.5$.

With this in mind, we have: $P(C_2|H_3, Z_1) = \frac{P(H_3|C_2, Z_1)P(C_2, Z_1)}{P(H_3, Z_1)}$

$$= \frac{P(H_3|C_2, Z_1)P(C_2, Z1)}{P(H_3|C_1, Z_1)P(C_1, Z_1) + P(H_3|C_2, Z_1)P(C_2, Z_1) + P(H_3|C_3, Z_1)P(C_3, Z_1)}$$

$= \frac{P(H_3|C_2, Z_1)}{P(H_3|C_1, Z_1) + P(H_3|C_2, Z_1) + P(H_3|C_3, Z_1)}$ since $P(C_1, Z_1) = P(C_2, Z_1) = P(C_3, Z_1)$ as C_i and Z_1 are independent

$= \frac{0.5}{0.5 + 0.5 + 0.5} = \frac{1}{3}$.

We also have: $P(C_1|H_3, Z_1) = \frac{P(H_3|C_1, Z_1)P(C_1, Z_1)}{P(H_3, Z_1)}$

$$= \frac{P(H_3|C_1, Z_1)P(C_1, Z_1)}{P(H_3|C_1, Z_1)P(C_1, Z_1) + P(H_3|C_2, Z_1)P(C_2, Z_1) + P(H_3|C_3, Z_1)P(C_3, Z_1)}$$

$= \frac{P(H_3|C_1, Z_1)}{P(H_3|C_1, Z_1) + P(H_3|C_2, Z_1) + P(H_3|C_3, Z_1)}$ since $P(C_1, Z_1) = P(C_2, Z_1) = P(C_3, Z_1)$ as C_i and Z_1 are independent

$= \frac{0.5}{0.5 + 0.5 + 0.5} = \frac{1}{3}$.

Since the two probabilities are equal, there is no benefit in switching the choice (but it does not harm either).

In the following problems, you can use the mean, median and standard deviation functions from MATLAB.

5. Generate a sine wave in MATLAB of the form $y = 5\sin(2.2x + \pi/3)$ where x ranges from -3 to 3 in steps of 0.02. Now randomly select a fraction $f = 30\%$ of the values in the array y (using MATLAB function

'randperm') and corrupt them by adding random values from 100 to 120 using the MATLAB function 'rand'. This will generate a corrupted sine wave which we will denote as z. Now your job is to filter z using the following steps.

- Create a new array y_{median} to store the filtered sine wave.
- For a value at index i in z, consider a neighborhood $N(i)$ consisting of $z(i)$, 8 values to its right and 8 values to its left. For indices near the left or right end of the array, you may not have 8 neighbors in one of the directions. In such a case, the neighborhood will contain fewer values.
- Set $y_{median}(i)$ to the median of all the values in $N(i)$. Repeat this for every i.

This process is called as 'moving median filtering', and will produce a filtered signal in the end. Repeat the entire procedure described here using the arithmetic mean instead of the median. This is called as 'moving average filtering'. Repeat the entire procedure described here using the first quaritle (25 percentile) instead of the median. This is called as 'moving quartile filtering'. Plot the original (i.e. clean) sine wave y, the corrupted sine wave z and the filtered sine wave using each of the three methods on the same figure in different colors. Introduce a legend on the plot (find out how to do this in MATLAB). Include an image of the plot in your report. Now compute and print the relative mean squared error between each result and the original clean sine wave. The relative mean squared error between y and its estimate $\hat{y}$ (i.e. the filtered signal - by any one of the different methods) is defined as $\dfrac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i y_i^2}$.
Now repeat all the steps above using $f = 60\%$, and include the plot of the sine waves in your report, and write down the relative mean square error values.
Which of these methods (median/quartile/arithmetic mean) produced better relative mean squared error? Why? Explain in your report. [5+5+4+3+3=20 points]

Solution: See code in the homework folder. The median-based method is better than the mean-based method. The quartile-based method outperforms the median-based method when the number of corrupted points is more.

6. Suppose that you have computed the mean, median and standard deviation of a set of n numbers stored in array A where n is very large. Now, you decide to add another number to A. Write a MATLAB function to update the previously computed mean, another MATLAB function to update the previously computed median, and yet another MATLAB function to update the previously computed standard deviation. Note that you are not allowed to simply recompute the mean, median or standard deviation by looping through all the data. You may need to derive formulae for this. Include the formulae and their derivation in your report. Note that your MATLAB functions should be of the following form

```
function newMean = UpdateMean (OldMean, NewDataValue, n),
function newMedian = UpdateMedian (oldMedian, NewDataValue, A, n),
function newStd = UpdateStd (OldMean, OldStd, NewMean, NewDataValue, n).
```

Also explain, how would you update the histogram of A, if you received a new value to be added to A? (Only explain, no need to write code.) **Note:** For updating the median, you may assume that the array A is sorted in ascending order, that the numbers are all unique. For sorted arrays with a even number of elements, MATLAB returns the answer as $(A(N/2) + A(N/2 + 1))/2$. You may use MATLAB's convention though it is not strictly required. [4+5+5+1 = 15 points]

Solution: The derivations to update the mean and standard deviation can be found in an accompanying pdf. The code is also provided on the homework folder.
Let n be the number of elements before insertion of the new value x. To update the median, you can maintain a sorted array and maintain a record of the location of the median. In case of odd-numbered arrays with unique values, the median is a single number $v = A[floor(n/2)+1]$. In case of even-numbered arrays, the median is not unique and can be any value within an interval (v_1, v_2) where $v_1 = A[n/2], v_2 = A[n/2+1]$. Now, each time a new value x arrives, you can insert x into the correct location in the array A to maintain the sorted

order. The update of the median is as follows (from now on, using indices to reference elements in the updated array): (1) If n is odd and $x < v$, then the new median is the interval $[A[floor(n/2)+1], A[floor(n/2)+2]]$. If $x > v$, then the new median is the interval $[A[(n+1)/2], A[(n+1)/2+1]]$. (2) If n is even and x falls in the median interval (v_1, v_2), then the new median is x. If $x < v_1$, then $v_1 = A[n/2+1]$ is the new median. If $x > v_2$, then $v_2 = A[(n+2)/2]$ is the new median.

7. Determine using a mathematical formula and a computer algorithm the smallest number n of people such that the probability that at least two of them share their birthday is at least p where $p \in \{5, 10, 15, 20, 30, 40, 50, 60, 70, 80, 90, 95, 99, 99.99, 99.9999, 100\}\%$. Plot a graph of n on Y axis versus p on X axis. The algorithm is to be implemented in MATLAB. [15 points]

   Solution: See code in the homework folder.